Mathematical Patterns Associated with Genetic Recombination in HIV-1 and SARS-CoV-2 via Explainability in CNNs. Relationship of Interpretability Scoremaps with Viral Phylogeny and the Role of the Third Codon Nucleotide

Ana Guerrero Tamayo

Faculty of Engineering
University of Deusto

A thesis submitted for the degree of
Doctor of Engineering

Abstract

To research genetic recombination among viruses, we applied genomic spectrograms and pre-trained CNNs via Transfer Learning, followed by explainability tools. Our approach aimed to understand recombination mechanisms and characterize genomic sequences. Via this methodology, we detected mathematical patterns that indicated genetic recombination in complete HIV-1 sequences. The main hot zone was near $f = 1/3$ in 5'LTR, extending to the first half of *gag*. LTRs regulate gene expression and virus replication. Nucleotides A and T in polyA and polyT stretches make these regions prone to polymerase skipping, aiding genetic recombination. The high Adenine content in HIV-1 explains the prominent role of A and, secondly, T.

We found a high correlation between hot zone similarity and phylogenetic proximity, with subtypes U, G, and O showing unique patterns. The CNN's behavior aligned with the biological traits of these subtypes. Our phylogenetic analysis and literature review confirmed the microbiological reality for subtype U. For subtype G, our results supported its potential recombinant origin and provided data for future research.

Applying Two-Stage Transfer Learning and Three-Step Explainability to complete SARS-CoV-2 genomic spectrograms, we detected a clear mathematical signature associated with the recombinant feature at $f = 1/6$ in the Spike protein, which is crucial for SARS-CoV-2 evolution and with multiple double-crossover genetic recombination events. By this methodology, we also identified clear mathematical signatures for non-recombinant SARS-CoV-2 variants: pre-VOC, Alpha, Delta, and Omicron.

Analyzing recombinant and non-recombinant SARS-CoV-2 sequences, we found a high Thymine content exceeding 45% in the third codon nucleotides at Spike protein. This phenomenon is statistically significant considering the low conservation degree of nucleotides in Spike (0.23%).

We found evidence suggesting that third codon nucleotides, less significant for amino acid coding, may contain additional information characterizing SARS-CoV-2. Periodicities of 3 and their multiples were linked to recombinant features in HIV-1 and SARS-CoV-2, and to specific variants in SARS-CoV-2. This suggests that third-position nucleotides may hold encrypted information affecting viral classification. Validating this hypothesis could reshape our understanding of the genetic code and spur new research in molecular biology and viral evolution.

Mathematical Patterns Associated with Genetic Recombination in HIV-1 and SARS-CoV-2 via Explainability in CNNs. Relationship of Interpretability Scoremaps with Viral Phylogeny and the Role of the Third Codon Nucleotide



Ana Guerrero Tamayo

Faculty of Engineering

University of Deusto

A thesis submitted for the degree of

Doctor of Engineering

Mathematical Patterns Associated with Genetic Recombination in HIV-1 and SARS-CoV-2 via Explainability in CNNs. Relationship of Interpretability Scoremaps with Viral Phylogeny and the Role of the Third Codon Nucleotide



Ana Guerrero Tamayo

Faculty of Engineering
University of Deusto

A Dissertation submitted for the degree of *Doctor of Engineering*

Supervised by Prof. Dr. Iker Pastor López and
Prof. Dr. María Dolores Moragues Tosantos

Doctoral candidate	Dissertation supervisor	Dissertation supervisor
GUERRERO TAMAYO, ANA (FIRMA) Firmado digitalmente por GUERRERO TAMAYO, ANA (FIRMA) Nombre de reconocimiento (DN): c=ES, serialNumber=78884726Q, sn=GUERRERO, givenName=ANA, cn=GUERRERO TAMAYO, ANA (FIRMA) Fecha: 2024.09.10 19:05:08 +02'00'	PASTOR LOPEZ, IKER (FIRMA) Firmado digitalmente por PASTOR LOPEZ, IKER (FIRMA) Fecha: 2024.09.11 10:39:55 +02'00'	MORAGUES TOSANTOS, MARIA DOLORES (FIRMA) Firmado digitalmente por MORAGUES TOSANTOS, MARIA DOLORES (FIRMA) Fecha: 2024.09.11 11:14:59 +02'00'
Ana Guerrero Tamayo, P.E.	Prof. Dr. Iker Pastor López	Prof. Dr. María Dolores Moragues Tosantos

Bilbao, September 7, 2024

A mis padres...

Acknowledgements

Soy una mujer afortunada.

Estos tres últimos años han sido un recordatorio constante de este hecho. He recibido tanto apoyo y tanto cariño que sólo puedo decir:

GRACIAS.

Quiero hablaros de mis padres. Antonio Guerrero Guerrero y Ana María Tamayo González. Unos héroes que rompieron varios moldes en sus vidas. Las personas a las que me quiero parecer. Y quiero hacerles un pequeño homenaje. Se conocieron siendo él Director Técnico y ella Directora Financiera en la misma empresa. Incansables trabajadores y personas entregadas a su cometido. Desde que tengo uso de razón me enseñaron lo importante que es esforzarse en todo momento. Vivimos momentos económicos muy duros en casa. La empresa en la que trabajaban cerró y mis padres tuvieron que reinventarse. Y lo hicieron. Vaya si lo hicieron... Así aprendí que nunca es tarde, que nunca eres demasiado mayor y que todo es posible.

Nunca, jamás, me enteré de nada de lo que estaba pasando. Viví en el mismo hogar feliz de siempre. Pero de algún modo me enseñaron que hay que ser prudente en el comportamiento, que hoy estás arriba y mañana estás abajo.

Mi padre resultó amenazado por ETA por el mero hecho de que en su empresa le mandaron a trabajar a la central de Lemoiz. Nunca nos dijo nada. Me enteré a los 30 años. Así aprendí la importancia de proteger a mis seres queridos. Y la valentía que recubre algunos silencios.

Mi padre ha sido tan querido por tantísima gente... Toda la vida le he visto ayudando a alguien. Y siempre con una sonrisa.

Papá, Don Antonio, no sabes la cantidad de gente que vino a tu funeral. Fue impresionante. Me dejaste hace ya 15 años y el dolor es como si todo hubiera sucedido ayer. Tienes que saber que, aún hoy, cualquiera que habla de ti siempre empieza diciendo *"Cuánto me ayudó tu padre..."*. Hace 15 días un buen amigo tuyo aún lloraba recordándote. Eras (eres) una persona muy especial, de esos que rompen el molde. Único por tantas cosas... Dedicaré mi vida a parecerme a ti. Este es mi objetivo vital más importante. Gracias papá...

Y mi madre... Doña Ana. Mi Ana Mari... Una carrera profesional admirable, un ejemplo para todas las mujeres. Una Directora Financiera en los años 60, cuando las mujeres difícilmente desarrollaban carreras en el mundo de la empresa. Y sin

darte importancia de ningún tipo. Una mujer divertidísima, con un sentido del humor agudo, inteligente y contagioso. Una mujer respetada y valorada, cuya opinión siempre era escuchada con alto interés. Mamá, siempre has estado ahí en los momentos realmente importantes. Recuerdo esa cara seria que se te ponía y ese tono de voz cuando te pedía consejo y tú me dabas el mejor de los posibles. Llevo en mi cartera como un tesoro ese papel en el que escribíste *"Investigadora Universidad de Deusto"* para poder decirles a todas tus amigas mi nuevo trabajo. Qué orgullosa estabas y qué contenta... Acertaste cuando dijiste que no ibas a vivir para verme Doctora. Yo me reía y te decía que no dijeras bobadas. Qué dolor siento en este momento... Seguro que estás ahí, mirando desde algún lado con papá. Mamá, gracias... Te quiero mucho...

Perdonad por favor todas mis torpezas, gracias por quererme tanto y por enseñarme tanto. Y gracias por mostrarme el mejor ejemplo de cómo debe ser una buena persona.

Os echo tanto de menos...

Os quiero tanto...

Estos son mis padres. Los mejores. Mis referencias.

Mis tíos, unos segundos padres... Mis tíos Jesús, Merche, Maria Luisa, Manolo, José Miguel. Quiero hacer una mención especial a mi tío D. Jesús Cartón Ibeas, gran intelectual y miembro de esta Comunidad Deustense a la que entro a formar parte. Tío, voy a ser Doctora, como tú. Ojalá estuvieras aquí. Y mi tia Merche, otra mujer rompedora sin darse importancia, tan culta y tan cariñosa...

Cómo duelen las ausencias...

Al amor de mi vida, a mi caro sposo, a mi amor eterno, a la razón de todo lo bueno que pueda hacer, a mi ancla, mi faro, mi vida. Izaskun Kintana Arkoitza. Amor mío, este Doctorado es nuestro. Si juntásemos todas las bibliotecas del mundo, no habría espacio para escribir todo lo que te mereces leer. Me debo conformar con estas palabras. Gracias por hacer que mi vida sea maravillosa. Eres un sueño hecho realidad. Luchadora incansable, un barco en medio de la tempestad que nunca se hunde, la mayor intelectual holística que he conocido. Lo entiendes todo... La mejor persona que he conocido, no hay nadie como tú, nadie. Me conoces como nadie, me comprendes como nadie, me haces reír como nadie. Sacas lo mejor de mí. Contigo a mi lado soy mi mejor versión. Tu luz es tan grande, tan bonita y tan intensa que ilumina un cielo con luna nueva. Tu energía me cura, tu presencia me sosiega, tu voz me mece, tu mente me alimenta. Tu alma me embelesa. Gracias por enseñarme a pedir perdón, a comprender, a abrir la mente, a levantarme cada vez que me caigo, a tener paciencia, a no ver rojo... Me has enseñado tantas cosas... Gracias porque cada vez que te veo, me emociono. Gracias por este matrimonio perfecto. Gracias porque me has regalado la historia de amor más bonita jamás contada. Gracias por quererme...

Como podéis ver, soy la mujer más afortunada del mundo. Y ella es la razón principal. El amor de mi vida. Cada día descubro nuevos detalles en esos ojos verdes como no hay otros en el universo, esos ojos que todo lo ven y todo lo entienden. Y cada día me enamoro más y más de ti. Te amo...

A mi querida María Ángeles Arkoitza Marina. Mi segunda madre. La mujer que me ha acogido como una hija más, siempre sonriente, siempre tranquila, siempre soportando todo sin una queja, sin una mala cara. Puedes con todo, te has puesto el mundo por montera y has sacado adelante a esa hija maravillosa. Gracias por cuidarme tanto y por darme tanto amor. Qué familia más bonita...

Mis angelitos, mis regalos del Universo: Umi, Izatz, Yoru, Sora, Noa, Serkan, Eda, Juno, Io. No se conoce el amor verdadero hasta que un animal te lo enseña. Cuánto tenemos que aprender de vosotros... Gracias por elegirme, os quiero con toda mi alma...

A mis primos y sobrinos. Elena, Gonzalo, Elsa, Michelle, Inara, Gorka, sois mis hermanos mayores, os quiero mucho. A todos mis sobrinos: sois maravillosos, os quiero mucho.

A Begoña y Eunate, mis tímidas amigas del colegio, por ser esas amigas que son familia, por estar ahí siempre con una entrega total, por venir cuando todo el mundo se va. Más de 30 años de amistad y cada día os quiero más. Sois las mejores. Gracias por sacar fotos de patos, gracias por buscar cómo ir a estudiar las belugas a Francia, gracias por apoyarme, por hacerme reír, por enseñarme a hablar en portugués, por llamarme regularmente para cuidarme, gracias por acogerme en vuestra casa, por esa postal en Diciembre que me reconcilia con la Navidad. Gracias por tantas y tantas cosas... A Iker y María, sabed que siempre estaré ahí para vosotros.

A toda mi familia, a todos mis amigos y a todas las personas que se han cruzado en mi vida y me han regalado un poquito de su bondad.

Héctor Archilla Segade, Doctor y estimado Profesor en la Universidad de Badajoz. Otro gran intelectual que no se da importancia. Gracias por aparecer en un mal momento, hablarme con sosiego y confianza, tranquilizarme y solucionarme la papeleta. Eres un amigo de los que es difícil encontrar...

Mi querida Alauda Ruíz de Azúa, la gran contadora de historias... Gracias por esos cafés, esas charlas, esas risas que nos hemos echado en estos 35 años de amistad. Gracias por ganar ese Goya que tanto te mereces. Pero sobre todo, gracias por recordarme que los sueños se cumplen. Por favor, sigue contándonos historias así de bonito.

A mi mentora profesional y personal Ana Macarulla Arenaza. Ay Ana, cuántas anécdotas contigo... Empezaste dándome clase, seguiste siendo mi Directora de Proyecto de Fin de Carrera y terminaste siendo mi referencia y mi amiga. Esta Facultad es el reflejo de tu obra. Tiene tu capacidad infinita de trabajo, tu sentido

de familia, tu cultura del cuidado mutuo, tu exhaustiva metódica y tu locura. Gracias por construirnos una casita tan preciosa. Me enseñaste a ser ingeniera y me enseñaste a ser persona. Eres mi referencia a un montón de niveles, mi consejera, mi soporte, MI AMIGA. Has estado tan presente en mi vida en tantos momentos... Tienes un sentido de la moral, de la ética, de los valores... Y eres una de las personas más divertidas que conozco. Gracias por hacerme llorar de la risa tantas veces. Admiro lo fuerte y lo valiente que eres. Siempre te pediré consejo para hacer las cosas bien y para ser mejor persona. Cuánto aprendo contigo... Un cerebro privilegiado envuelto en miles de virtudes. Te quiero mucho Ana, muchísimo. Gracias por iniciar toda esta aventura y ser el primer paso hacia la vida que siempre he querido llevar. Tenías que ser tú, no podía ser de otra forma.

A esa roca que siempre está ahí para que te sujetes en medio de un río de aguas bravas. Isabel Muñoz San Ildefonso. Otro de los cimientos de esta Facultad de Ingeniería. Otra de las responsables de esa casita de la que hablaba en el párrafo anterior. Siempre ahí para todo, siempre dispuesta, siempre invitando a aprender con una sonrisa. Isabel, sigo pensando que eres una gran pérdida para la Facultad de Ingeniería. Te necesitamos aquí. Pero es que te necesitamos en todas partes y eso no es posible, lo entiendo. Gracias por atender mis miles de dudas de lo más variopintas. Anda que te he preguntado cada cosa... Gracias por darme siempre la mejor referencia, el mejor consejo, gracias por ese café cuando murió mi madre. La mujer que resuelve todo, otra referencia de trabajo y valores. Gracias por tu amistad y ojalá algún día pueda ayudarte en cualquier cosa que necesites. Te la debo.

Fijaos si soy afortunada que he tenido la suerte de conocer a dos investigadoras brillantes que trabajan nada más y nada menos que en el Centro Nacional de Microbiología - Instituto de Salud Carlos III (ISCIII). Cuántas vidas se habrán salvado por su trabajo... No quiero extenderme en sus carreras porque necesitaría otro documento del tamaño de esta Tesis. Lo único que es aún mayor que su trayectoria profesional y su intelectualidad es su bondad. Isabel Olivares Zarco, Concha Casado Herrero, gracias por confiar en mí sin conocerme de nada, habéis sido mi gran suerte de este Doctorado y de mi vida. Sin vuestro apoyo esto no hubiera funcionado. Gracias por esas llamadas de seguimiento, gracias por estar ahí siempre, por ser tan solícitas con los plazos, por revisar todo al detalle, por estar siempre a mi disposición (anda que no tendréis otras cosas más importantes que hacer...). Concha, siempre llevaré conmigo esa frase que me hace levantarme como un resorte cuando decaigo: SEGUIMOS. Isabel, mi mentora, la que con cada llamada me ha sacado una sonrisa que me ha durado todo el día. Sois un modelo de trabajo, de efectividad, de cómo se hacen las cosas, sois garantes de nuestra salud. Y, en mi caso, sois las que me habéis dado la oportunidad de dedicarme a un trabajo maravilloso. Gracias...

Mi querida Lola... Lola Moragues Tosantos. Mi amatxu en La Boheme... Fíjate las vueltas que ha dado la vida desde que me tirabas de la oreja allá por los años 90 en el Arriaga. Una amistad llena de música y la mejor Directora que podía soñar. Gracias por tirarte a la piscina, por ayudarme sin preguntar nada, sin ningún tipo de duda. *"Me tienes para lo que haga falta"*, me dijiste. Mira que estos años han sido duros... Y siempre has estado ahí, siempre disponible, siempre a tiempo, siempre sonriente, siempre sin prisa ninguna. Gracias por tu dedicación, por el trabajazo que te has pegado, por ser exigente, por ser exhaustiva, por ser metódica, por hacerme mejor profesional y mejor persona. Eres mi referente, tienes una trayectoria impresionante y no nos la cuentas, me tengo que enterar por Google. Y eres una persona de las que no hay. Lola, qué suerte la mia... Seguiremos haciendo música como buenas borealers y combatiremos bichitos poquito a poco. Eres una de mis grandes fortunas.

Dani, por fin te puedo poner en los agradecimientos. Gracias por cuidar todo, gracias por hacerme reír cada vez que oía *"pessssada"* por el teléfono. Ay Dani, me he reído mucho. Gracias por toda tu chispa, tus chistes, tus ánimos, gracias por tu ayuda y tu apoyo absoluto en todas mis iniciativas. Dani, eres muy grande...

Y ahí estuvieron confiando en mí Borja Sanz Urquijo e Iker Pastor López. Borja, gracias por ser el primero que estuvo ahí, fuera de horas de trabajo y a cambio de nada. Gracias por esa primera oportunidad. Eres un compañero de viaje protector, un buen amigo y esa persona que siempre aparece para dar soporte cuando hay un problema. Admiro tu capacidad de gestión en medio del caos, tu sosiego. Estás ahí aunque estés reventado por dentro. Sonríes y ayudas. Gracias, qué suerte encontrarte... Iker, mi otro mejor Director que se puede tener. Cuántas cosas haces en tan poco tiempo... Contigo he aprendido a que es posible sacar adelante todo cuando parece imposible. Tengo en el corazón esos paseos por la Universidad, gracias por aguantarme, gracias por enseñarme tanto, gracias por confiar en mí. Gracias por hacerme reír hasta sacarme las lágrimas. Ay, qué risas me he echado contigo, madre mía... Gracias por tener la palabra adecuada en cada momento, por aparecer en el momento justo, por tu apoyo, por el cariño, por llegar y hacer tu magia... Es que son mil cosas. Gracias por tu inmensa generosidad, eres un mentor generoso como nadie. Gracias por todas las oportunidades, Iker, gracias por contar conmigo y por hacer todo lo que haga falta para que me vaya lo mejor posible. Eres una persona tan buena... Eres un buen amigo Iker, tengo mucha suerte. Gracias. No dejéis nunca de hacer vuestra magia.

Mi querida Miren Arretxe Zamakona... Nuestra amatxu de DeustoTech... No por edad, que conste, sino por ese cuidado y ese cariño que nos has dado a todos. Miren, te quiero mucho. Estuviste ahí en pleno proceso final de mi ama, cuidándome en "presente" y en "a distancia". Tengo todos y cada uno de tus mails y mensajes en el

corazón. No sabes cuánto me ayudaste en ese momento Miren, no te haces a la idea. Tampoco creo que sepas lo mucho que te queremos todos los que hemos tenido la suerte de encontrarnos contigo. Eras la paz en DeustoTech, la protección, el cariño. Si Miren está, todo está bien. Traías el color y la alegría a ese pasillo. Cuando estabas, todo era mejor. Eres una persona increíble, un ejempo de trabajo, de constancia, de cómo resolverlo todo. Sin darte importancia ninguna. Y con una sonrisa eterna. Tienes un sitio muy especial en mi corazón, Miren. Eskerrik asko danagatik...

Enrique Pallarés Molins, un sabio, un gran intelectual al que he tenido la gran suerte de conocer. Un gran amigo, un hombre simplemente espectacular. Querido Enrique, gracias por acogernos en tu lista de amistades. Eres muy especial. Gracias por esa ayuda tan pragmática y tan cariñosa en un momento tan malo de mi vida... No he conocido a nadie como tú. Enrique, me has alegrado tantos desayunos escuchando tu podcast y leyendo tus artículos. Eres un maestro. Un hombre excepcional. Siempre diré con orgullo que he tenido la suerte de conocerte y de ser tu amiga.

Mi querido Fernando José Sadio Ramos... Te diría tantas cosas... Fernando, gracias por acogerme, gracias por cruzarte en mi camino, gracias por darme todo y nunca pedirme nada. Eres un sabio. Tienes una forma maravillosa de ver la vida. Gracias por descrubirme las maravillas de la Filosofía. Gracias por enseñarme tanto de cómo diseñar metodologías de investigación basándote en los grandes pensadores. Qué lecciones tan impresionantes me has dado... Gracias por compartir tu inmensa cultura conmigo. Podría estar horas escuchándote, hablando de lo que sea. Porque todo lo que dices es sabio, todo lo que dices es inteligente, todo lo que dices es parte de tu bondad inmensa. Y mil gracias por esas fotos diarias de ese león precioso que es Mario, que me alegra todas las tardes. Ese pequeño Mario que es un angelito... Mi querida María Angustias Ortíz Molina. No me extraña que dediques tu vida y tu experiencia a cuidar del Patrimonio de la Humanidad porque tu forma de ser te convierte en garante de lo mejor del espíritu humano. Qué buena persona eres... Os llevo en el corazón y os llevo cerquita siempre, aunque no os vea todo el tiempo que me gustaría. Dicen que Mahatma Gandhi dijo una vez *"La grandeza de una nación y su progreso moral pueden ser juzgados por la forma en que se trata a sus animales"*. Y estoy de acuerdo. Y os miro y pienso que toda vuestra bondad, toda vuestra espiritualidad se refleja en vuestro amor a los animales. Gracias por existir.

Gracias a los Professores Doutores João Paulo Martins Gouveia y Teresa Cristina Melo Fragoso. Mil gracias por ese apoyo incondicional hasta el punto de organizar semejante ciclo de conferencias en una institución tan prestigiosa como es el Instituto Politécnico de Coimbra (IPC). Gracias por ese Workshop tan espectacular. Estoy tan orgullosa de haber inaugurado semejante encuentro... Gracias por vuestras palabras, por vuestras sonrisas, por vuestra amabilidad infinita, por tener siempre

la mejor muestra de cariño posible. El Atlántico nunca fue tan pequeño, gracias por abrirme las puertas del mundo. Trabajaremos juntos toda la vida.

A Juan Mari Collantes e Ibone Lizarraga, nunca olvidaré ese recibimiento en la Facultad de Física de la UPV/EHU. Nunca olvidaré aquella frase... *"Un espectrograma es la evolución de la FFT a lo largo del tiempo"* (algo así). Y de repente lo entendí todo. Vuestra inteligencia es bella, vuestra generosidad es grande. Y vuestra amistad es fortuna.

Mi querido Nacho Angulo Martínez. No me sorprende que absolutamente todo el mundo te quiera, te respete, te valore. Todas las personas que trabajan aquí, absolutamente todas, hablan maravillas de ti. Y no me extraña. Te tengo que agradecer tantas cosas... La confianza que has depositado en mí es la fuerza que me lleva a hacer todo lo mejor posible. Ahora y siempre. Nunca olvidaré todo lo que has hecho por mí. Siempre con esa sonrisa indeleble que te acompaña y transmite tanta tranquilidad, apoyo, cariño... Eres mi modelo de jefe. Pero no sólo eso, eres mi modelo de persona, lo que quiero ser. Nacho, eres el alma de esta maravillosa Facultad de Ingeniería de la Universidad de Deusto. Hay una pátina tuya en tantas cosas, en tantos detalles... Maravillosa persona, Nacho, maravillosa persona...

Mis queridos compañeros de esa fantástica Primera Promoción, mis colegas, mis amigos... Amaia, Alberto, no os digo nada pero os lo digo todo...

Mi querida Nekane Ione Sainz Bedoya, mi mentora en esta nueva aventura. Tu sola presencia inunda todo de paz y sosiego, tu sonrisa cura el estrés, tu conversación acaricia el alma y tu soporte la fortalece. Gracias por dedicarme un tiempo que no tienes y por todo tu apoyo desde el primer dia. Eres otra de mis fortunas...

Ay Iñaki García Quintanilla... Casi todos mis mejores recuerdos de la Universidad están en el Laboratorio de Electrónica de Potencia. Contigo. Y gracias a ti. Ahí nos tenías a esa panda de garrulos metiendo ruido, manejando el eje lineal infinitas horas, o tal vez simplemente pasando la tarde con amigos que tenían que hacer cosas, e incluso dándote sin querer con el láser en el ojo (perdona, por favor, aún no lo hemos superado). Nuestro hermano mayor con una paciencia infinita. Y una sonrisa eterna y contagiosa como ninguna. Iñaki, qué alegría volverte a ver, qué alegría tenerte cerquita y qué agradecimiento más profundo siento...

Arantza Mugica Arrien... Ay Arantza, mezclo el cariño de juventud con el cariño de la madurez. Y esa mezcla es tan bonita... Eres uno de los pilares de esta Facultad. Una persona que levantó de cero toda esta gran familia, la profesora a la que recordamos con tanto cariño... Entre todos generásteis un sentimiento común de que éramos (somos) una gran familia. Alumnos y profesores. Todos juntos. Esa magia la asentásteis tú y todos vosotros. Y esa magia nos acompaña toda la vida. Gracias por estar ahí, siempre, gracias por el apoyo, por la oportunidad, por todo tu cariño, por estar pendiente, por ser como una hermana mayor que, si ve

que te vas a caer, se adelanta y te sujeta para que eso no ocurra. Una inmensa inteligencia rodeada de una bondad infinita.

Begoña Garcia Zapirain... Otra parte fundamental de nuestros recuerdos... Recuerdo con todo el cariño aquel micro-espía láser que resultó un fracaso. Nos viste las caras y, sin mediar palabra, nos llevaste a un laboratorio y nos enseñaste que realmente no lo era, que funcionaba. Nos hablaste, nos llenaste de motivación, nos devolviste la sonrisa. Y la autoestima. Recuerdo que siempre brindabas con nosotros en la Txampanada. No bebías nada pero levantabas la copa, te echabas unas risas con nosotros, nos devolvías la copa y nosotros te dábamos un fuerte aplauso. Profesora cercana, enrrollada, exigente, entregada y metódica. Querida Begoña, gracias por todo.

Queridos profesores todos, todos vosotros... Todos nosotros, ex-estudiantes, cada vez que pasamos por el puente de Deusto, dedicamos varios segundos a mirar la Universidad y, por un momento, recordamos de golpe los mejores años de nuestra vida. Sonreímos con la mirada perdida y continuamos caminando con esa sonrisa que no termina de desaparecer. Esto es gracias a todos vosotros. Sois tan importantes en nuestra vida... Gracias porque sin vosotros, sereíamos peores.

Además, soy una mujer afortunada porque he tenido la gran suerte de ser uno de los tres miembros de la Banda de DeustoTech (el nombre no lo pusimos nosotros, que conste). Armando, Tony y yo. Siempre juntos, para siempre. Armando, Tony y yo, un trío inseparable forjado con una sólida amistad que hace que me sienta tan afortunada...

Mi queridísimo Armando Aguayo Mendoza... Querido amigo, qué puedo decir... Agradezco a la vida que nos haya presentado. Mi GRAN AMIGO. Mi GRAN SUERTE... Soy tan afortunada por haberte conocido... Eres una persona maravillosa, siempre ayudando a todo el mundo, siempre con la palabra perfecta, con la mejor sonrisa... Nos haces reír tanto a todos con ese humor tan inteligente, tan preciso y tan elaborado. Sin ti, esta Universidad es un sitio más gris, más feo. Todo el mundo te conoce, todo el mundo te quiere. Eres muy grande amigo mio... Tu grandeza sólo es inferior a mi suerte por haberte encontrado. Desde el momento en que nos conocimos hasta el infinito, estamos unidos por una amistad eterna. No hay palabras para recopilar todo lo vivido, te quiero un chingo mi querido Armando, te quiero un chingo... Eres familia...

Mi Tony Castillo Calzadilla, un amigo de los que no hay. Tony, nuestro profesor en esta aventura. Eres una persona tan generosa, tan cariñosa, tan divertida, tan incondicional... Guardo como un tesoro toda la ayuda que me has dado siempre, cada momento vivido, cada abrazo infinito. Eres grande, amigo mio. Tu presencia es fuerte, tu alma es bella. Querido amigo mio, hemos vivido tantas cosas... Brillas, querido amigo, brillas y arropas. Gracias por introducir en mi lista de fortunas

a Oda y a esa niña tan maravillosa, Madeleine... Qué niña mas maravillosa, con una sonrisa que podría terminar con todas las guerras...

En esta lista de suertes no puede faltar Ignacio Fidalgo Astorquia. Ignacio, querido amigo mio, no sabes la suerte que tengo de haberte conocido. Eres una persona maravillosa. El mejor cerebro de la Facultad y el mejor y más grande corazón. Eres una de mis suertes, amigo mío. Estoy muy orgullosa de ser tu amiga. Eres muy grande, como esos antiguos sabios que lo eran y no se daban ninguna importancia porque no hacia falta. El que sabe mirar, sabe ver. Y tú, querido amigo, eres excepcional a todos los niveles. No cambies nunca.

Mis amigos de DeustoTech:

Ainhoa Alonso Vicario y Pili Elejoste Larrucea, hacéis que todo funcione. Gracias por todas esas cosas que hacéis por nuestro bien y que nadie se entera de que las hacéis.

Ainhoa, gracias por acogerme desde el primer día. Gracias por tus consejos, por tu apoyo, por animarme cuando parece imposible conseguirlo. Pues tú lo consigues. Siempre.

Pili, el alma de DeustoTech y la que nos alegra la vida con las canciones de Mecano. Ay Pili, eres muy especial...

Alex, mi querida niña, llenabas el laboratorio de colores y tu fortaleza me dió una lección de vida que nunca olvidaré. Te quiero mucho... Elisa, con una paciencia infinita aguantando mis protestas y acogiéndome el primer día como una más con todo tu cariño. Te echo de menos... Salomey, esa mujer tan divertida y tan friolera, la alegría de DeustoTech. Erik, mi niño, el galán del laboratorio y ese buen amigo que siempre está ahí con mil detalles que te alegran la vida. Maite, esa amiga que lo deja todo para ayudarte, la que siempre está ahí. Nerea, mi compi del alma. Te tengo al lado y sé que todo va bien. La risa más contagiosa de DeustoTech que te cambia el humor en tres segundos. Alcides, un amigo divertido como nadie, pragmático, resolutivo y pendiente. Marcos, gracias por agacharte cada vez que me hablas, gracias por estar pendiente y por sonreir así de bonito.

Te debo una ronda Mikel, compañero, no se me va a olvidar. Cuando nos veamos de nuevo, lo primero que voy a hacer. Y nos tomaremos ese café que la vida nos ha arrebatado.

Gracias a todos los que me habéis ayudado a ser mejor...

Soy una mujer afortunada...

Contents

List of Figures **xxv**

List of Tables **xxxi**

List of Abbreviations **xli**

1 Introduction **1**
- 1.1 May I share a brief story with you? . . . 1
- 1.2 Why SARS-CoV-2 + FCoV? . . . 5
- 1.3 Difficulties in Predicting Genetic Recombination Between Two Viruses with the Current State of the Art . . . 9
- 1.4 Applicability Limitations in the Case Study of Genetic Recombination between SARS-CoV-2 and FCoV . . . 12
- 1.5 Redefinition of the Research Line . . . 14
- 1.6 Reformulation of the Virus Case Study . . . 14
- 1.7 Determination of the Specimen Virus for Genetic Recombination Study 15
- 1.8 Deep Learning Tools of Choice . . . 15
- 1.9 Conversion of Genomic Sequences into Images . . . 17
- 1.10 Interpretation of Results . . . 19
- 1.11 Research Hypothesis . . . 20
- 1.12 Research Objectives . . . 20
- 1.13 Basic Methodology . . . 21
- 1.14 Methodical Process . . . 22
- 1.15 Final Hypothesis . . . 25

2 Microbiological Considerations **27**
- 2.1 Introduction . . . 27
- 2.2 Genome. Definition and Concepts. . . . 28
- 2.3 Viral Diversification Mechanisms . . . 34
  - 2.3.1 Mutation . . . 34
  - 2.3.2 Genetic Recombination . . . 36
- 2.4 Multiple Sequence Alignment . . . 38
- 2.5 Phylogeny. Definition and Concepts . . . 41

3 Technological Background **47**
3.1 Introduction . . . 48
3.2 Introduction to the Concept of Machine Learning . . . 49
3.3 Introduction to Deep Learning Architectures . . . 51
3.4 Introduction to Convolutional Neural Networks . . . 52
3.5 Introduction to Transfer Learning . . . 57
3.6 Introduction to Explainability Tools . . . 60
3.6.1 Gradient-Weighted Class Activation Mapping (Grad-CAM) . 61
3.6.2 Local Interpretable Model-Agnostic Explanations (LIME) . . 62
3.6.3 Gradient Attribution . . . 64
3.7 Objective of this State-of-the-Art Review . . . 65
3.8 Methodology . . . 65
3.9 Search Topics and Preliminary Article Detection . . . 68
3.10 Summary Diagram of Technological Background Review Methodology 68
3.11 Genomic Spectrograms. State of the Art. . . . 71
3.12 Deep Learning Applied to Biosanitary Field. State of the Art . . . 75
3.13 Deep Learning Applied to Genomics. State of the Art . . . 82
3.13.1 Nature of Multi-omic Data . . . 82
3.13.2 Genome Processing and Algorithms . . . 84
3.13.3 Identification and Analysis . . . 87
3.13.4 Prediction . . . 93
3.13.5 Specific Deep Learning Applications in Genetic Recombination 95
3.13.6 Metagenomics . . . 96
3.13.7 Study of Diseases . . . 99
3.13.8 Precision Medicine . . . 103
3.13.9 Crop Improvement . . . 105
3.13.10 Improvement of Existing Technology . . . 106
3.14 Interpretability of Biosanitary Deep Learning. State of the Art . . . 109
3.15 Critical Analysis of the State of the Art . . . 113
3.15.1 Multiple Sequence Alignment . . . 113
3.15.2 Deep Learning Applied to Biology and Medicine . . . 119
3.16 Conclusions . . . 133

4 Genomic Spectrogram **137**
4.1 Introduction . . . 137
4.2 Brief Background . . . 138
4.3 Concept of Genomic Spectrogram . . . 141
4.4 Genomic Spectrograms vs. Multiple Sequence Alignment . . . 142
4.5 Types of Genomic Spectrogram . . . 144
4.6 Biomathematical Considerations about Genomic Spectrograms . . . 151
4.7 Conclusions . . . 152

5 HIV-1 Genetic Recombination Analysis 155
5.1 Introduction . . . 156
5.2 HIV-1 Microbiological Considerations . . . 159
5.3 Equipment . . . 163
5.4 Methodology Block Diagram . . . 163
5.5 HIV-1 Complete Genomic Sequences Compendium . . . 164
5.6 HIV-1 Dataset . . . 169
5.7 Spectrogram Generation . . . 169
5.8 Pre-trained CNN Selection Criteria . . . 170
5.9 Best Performance Based Decision Criteria . . . 171
5.10 Robustness Criteria . . . 174
5.11 Explainability Tools' Result Generation . . . 175
5.12 Phylogenetic Tree and Comparison with Explainability Tools' Patterns 176
5.13 Raw Results . . . 179
5.13.1 Concatenated Spectrograms . . . 180
5.13.2 Superposed Spectrograms . . . 183
5.14 Results Processing . . . 186
5.14.1 CNN Selection . . . 186
5.14.2 Robustness . . . 189
5.14.3 Hot Zones Determination . . . 190
5.15 Biological Consistency Analysis . . . 194
5.15.1 Analysis of Successful Classifications . . . 195
5.15.2 Similarity between Sequences of Different Subtypes . . . 225
5.15.3 Analysis of Failed Classifications . . . 237
5.16 Conclusions . . . 246
5.17 Future Studies . . . 249

6 Analysis of Total Hot Zones in HIV-1 by Subtype 253
6.1 Introduction . . . 253
6.2 Equipment . . . 254
6.3 Methods . . . 255
6.4 HIV-1 Classification Error Rates per Subtype . . . 257
6.5 Total Hot Zones per Category . . . 259
6.6 HIV-1 Total Hot Zones per Subtype . . . 261
6.6.1 Non-recombinant Subtypes . . . 262
6.6.2 Recombinant Subtypes . . . 274
6.7 Conclusions . . . 290
6.8 Future Studies . . . 293

7 A Brief Analysis of HIV-1 Subtypes F1, CRF22_01A1, and O 295
7.1 Introduction . . . 295
7.2 Equipment . . . 296
7.3 Methods . . . 296
7.4 Subtype F1 . . . 297
7.5 Subtype CRF22_01A1 . . . 326
7.6 Subtype O . . . 341
7.7 Conclusions . . . 360
7.8 Future Studies . . . 362

8 Analysis of HIV-1 Subtype U 363
8.1 Introduction . . . 364
8.2 Equipment . . . 365
8.3 Methods . . . 366
8.4 Global Results of the Experiment . . . 368
8.5 Analysis of Phylogeny and Explainability Patterns . . . 370
8.5.1 AB485664, AB485665 and KY392773 . . . 370
8.5.2 AY046058, KY392780, MH078559 and MK095236 . . . 379
8.5.3 FJ388921, EF029066, EF029067, EF029068, EF02906, and KY392775 . . . 386
8.5.4 HM215249 and HM215251 . . . 391
8.5.5 JF683772 . . . 399
8.5.6 JX140679 . . . 405
8.5.7 KP718920, CRF11_cpx, CRF13_cpx, and CRF27_cpx . . . 413
8.5.8 MN153489, A6 and A1 . . . 422
8.5.9 MN153492, F1, CRF12_BF, and F2 . . . 427
8.5.10 MN736701 and MN736702 . . . 431
8.5.11 Preliminary Conclusions . . . 440
8.6 Conclusive Biological Coherence . . . 441
8.7 Non-Conclusive Biological Coherence . . . 457
8.8 Another Similarities . . . 469
8.9 Conclusions . . . 470

9 Analysis of HIV-1 Subtype G 475
9.1 Introduction . . . 475
9.2 Microbiological Considerations and Recombinant Origin of Subtype G 476
9.3 Equipment . . . 477
9.4 Methods . . . 478
9.5 Global Results of the Experiment . . . 478
9.6 Analysis of Phylogeny and Explainability Patterns . . . 480

9.6.1 Subtype G Philogenetic Group 1 . . . 480
9.6.2 Subtype G Philogenetic Group 2 . . . 499
9.6.3 Subtype G Philogenetic Group 3 . . . 509
9.7 Discussion . . . 512
9.7.1 Comments on Subtype G Phylogeny . . . 512
9.7.2 Analysis of Hot Zones in Intra-patient Sequences . . . 516
9.7.3 Prevalence Analysis of Recombinant Subtype G Sequences . 521
9.7.4 Analysis of Similarity Patterns among Subtype G Hot Zones 522
9.7.5 Influence of dual infections . . . 540
9.8 Conclusions . . . 542

10 Comprehensive Analysis of the Genomic Elements of SARS-CoV-2 545
10.1 Introduction . . . 545
10.2 Genomic Protein Analysis in SARS-CoV-2 . . . 546
10.3 Equipment . . . 561
10.4 Methods . . . 562
10.4.1 Sequence Preprocessing . . . 562
10.4.2 Sequence Filtering . . . 565
10.4.3 Fragmentation of Filtered Sequences . . . 566
10.4.4 Determination of Consistency Ratios . . . 567
10.4.5 Calculation of Consistency Ratios . . . 571
10.5 Results . . . 572
10.5.1 Filtering Results . . . 572
10.5.2 Consistency Level Results . . . 579
10.6 Discussion . . . 581
10.7 Conclusions . . . 587

11 SARS-CoV-2 Genetic Recombination Analysis 589
11.1 Introduction . . . 590
11.2 Coronaviruses Microbiological Considerations . . . 592
11.3 Materials and Methods . . . 599
11.3.1 SARS-CoV-2 complete genomic sequences compendium . . . 599
11.3.2 Dataset Design . . . 601
11.3.3 Generation of Genomic Spectrograms . . . 602
11.3.4 Two-Stage Transfer Learning . . . 604
11.3.5 Test Bench . . . 605
11.3.6 Performance Measurement . . . 605
11.3.7 Explainability Analysis . . . 605
11.4 Results . . . 609
11.4.1 Results per Subsampling . . . 609

11.4.2 Results per Number of Epochs . . . 611
11.5 Discussion . . . 612
11.5.1 Optimal Configuration Selection . . . 614
11.5.2 Reference Sequence . . . 614
11.5.3 Analysis of Non-Recombinant Results . . . 616
11.5.4 Analysis of Recombinant Results . . . 617
11.6 Conclusions . . . 621

12 Analysis of SARS-CoV-2 Non-recombinant Variants 625
12.1 Introduction . . . 625
12.2 Equipment . . . 627
12.3 Methods . . . 628
12.3.1 Spectrogram Generation . . . 628
12.3.2 Dataset of Non-recombinant SARSCoV-2 Variants . . . 628
12.3.3 Two-Stage Transfer Learning . . . 631
12.3.4 Best Performance Based Decision Criteria . . . 633
12.3.5 Explainability Tools' Result Generation . . . 633
12.4 Results and Discussion . . . 635
12.4.1 Training Curves . . . 635
12.4.2 Performance Measurement Ratios . . . 639
12.4.3 Total Hot Zones per Variant . . . 642
12.5 Conclusions . . . 651

13 Exploring the Influence of the Third Codon Nucleotide on Genetic Recombination in the SARS-CoV-2 Spike Protein 655
13.1 Outcome Hypothesis Analysis . . . 656
13.2 Equipment . . . 657
13.3 Methods . . . 657
13.3.1 Selection of Study Subsampling . . . 657
13.3.2 Selection of Study Sequences . . . 658
13.3.3 Download of EMBL Files . . . 658
13.3.4 Extraction of Relevant Protein Genomes . . . 659
13.3.5 Preliminary Identification of Visual Patterns in Arithmetic Sequences of 6 Nucleotides in S . . . 660
13.3.6 Calculation of Nucleotide Content . . . 661
13.3.7 Search for Common Nucleotides in S among Sequences of the Same Category . . . 662
13.3.8 Biological Interpretation of Results . . . 662
13.4 Results and Discussion . . . 663
13.4.1 Subsampling Under Study . . . 663

13.4.2 Sequences Under Study 663
13.4.3 Preliminary Verification of the Actual Position in S of the First Sixth Nucleotide 667
13.4.4 Preliminary Identification of Visual Patterns in Arithmetic Sequences of 6 Nucleotides in S 670
13.4.5 Nucleotide Content 672
13.4.6 Common Nucleotides 678
13.4.7 Microbiological Analysis 686
13.5 Conclusions 687
13.6 Future Research 691

14 Is There a Role Associated with the Third Codon Nucleotide in the Characterization of a Genomic Sequence? 693
14.1 Introduction 694
14.2 Compilation of Evidence 696
14.2.1 Recombinant Feature Hot Zones HIV-1 and SARS-CoV-2 (Superposed spectrograms) 696
14.2.2 Genetic Recombination in HIV-1 698
14.2.3 HIV-1 Hot Zones per Subtype 700
14.2.4 Genetic Recombination in SARS-CoV-2 709
14.2.5 SARS-CoV-2 Variant Hot Zones 710
14.3 Microbiological and Genomic Analysis 712
14.3.1 Differences between HIV-1 and SARS-CoV-2 712
14.3.2 The Role of Nucleotides A and T in Periodicities of 3 and their Multiples 715
14.3.3 LTRs of HIV-1 718
14.3.4 Differences between the LTRs in HIV-1 and the UTRs in SARS-CoV-2 720
14.3.5 The Spike Protein in SARS-CoV-2 722
14.4 Hypotheses 726

15 Final Conclusions 729
15.1 Introduction 730
15.2 Some Numbers 732
15.3 Main Contributions 733
15.3.1 Chapter 3. Technological Background 734
15.3.2 Chapter 4. Genomic Spectrogram 735
15.3.3 Chapter 5. HIV-1 Genetic Recombination Analysis 736
15.3.4 Chapter 6. Analysis of Total Hot Zones in HIV-1 by Subtype 743

15.3.5 Chapter 7. A Brief Analysis of HIV-1 Subtypes F1, CRF22_01A1, and O 749
15.3.6 Chapter 8. Analysis of HIV-1 Subtype U 752
15.3.7 Chapter 9. Analysis of HIV-1 Subtype G 754
15.3.8 Chapter 10. Exhaustive analysis of SARS-CoV-2 Genomic Elements 758
15.3.9 Chapter 11. SARS-CoV-2 Genetic Recombination Analysis . 762
15.3.10 Chapter 12. Analysis of SARSCoV-2 Non-recombinant Variants 764
15.3.11 Chapter 13. Exploring the Influence of the Third Codon Nucleotide on Genetic Recombination in the SARS-CoV-2 Spike Protein 767
15.3.12 Chapter 14. Is There a Role Associated with the Third Codon Nucleotide in the Characterization of a Genomic Sequence? The Case of HIV-1 and SARS-CoV-2 771
15.4 Fulfillment of Research Objectives 778
15.5 Main Conclusions 784
15.6 Implications 786

16 Future Research Lines **787**
16.1 Introduction 788
16.2 Synthetic Sequence Generation 789
16.2.1 Preliminary Considerations 789
16.2.2 Potential Architecture 794
16.2.3 Technological Challenges 802
16.2.4 Some advantages 803
16.3 Genomic Size Delimitation for the Effective Application of Genomic Spectrograms 804
16.4 Deeper Understanding of the Detected Mathematical Signatures . . 806
16.5 Identification of Mathematical Signatures Associated with Antifungal Resistance 807
16.6 Identification of Mathematical Signatures Associated with Antibiotic Resistance 811
16.6.1 *Mycobacterium tuberculosis* 813
16.6.2 *Clostridium difficile* 815
16.6.3 Vancomycin-Resistant Enterococci 818
16.6.4 Methicillin-Resistant *Staphylococcus aureus* 821
16.6.5 *Neisseria gonorrhoeae* 823
16.6.6 Carbapenem-Resistant Enterobacteriaceae (*Escherichia coli*, *Klebsiella pneumoniae*,) 826

16.7 Bio-Mathematical Correlation between the HIV-1 Genome and the Host Genome 828
16.8 Deepening the Mathematical Signatures Characterizing the Phylogeny among Different Pathogens 830

Appendices

A Compendium of Subtype F1 CNN Classification Results and Grad-CAM Hot Zones, in Alphanumeric Order 837
A.1 Subtype F1 Sequences 837

B Compendium of Subtype CRF22_01A1 CNN Classification Results and Grad-CAM Hot Zones, in Alphanumeric Order 847
B.1 Subtype CRF22_01A1 Sequences 847

C Compendium of Subtype O CNN Classification Results and Grad-CAM Hot Zones, in Alphanumeric Order 851
C.1 Subtype O Sequences 851

D Compendium of Subtype U Sequences and their Phylogenetically Related Sequences, in Alphanumeric Order 859
D.1 Subtype U Sequences 859
D.2 Subtype A1 Sequences 862
D.3 Subtype A2 Sequences 865
D.4 Subtype A6 Sequences 865
D.5 Subtype A7 Sequences 867
D.6 Subtype F1 Sequences 867
D.7 Subtype F2 Sequences 868
D.8 Subtype G Sequences 869
D.9 Subtype H Sequences 870
D.10 Subtype K Sequences 871
D.11 Subtype CRF02_AG Sequences 872
D.12 Subtype CRF04_cpx Sequences 876
D.13 Subtype CRF06_cpx Sequences 877
D.14 Subtype CRF09_cpx Sequences 877
D.15 Subtype CRF11_cpx Sequences 878
D.16 Subtype CRF12_BF Sequences 881
D.17 Subtype CRF13_cpx Sequences 881
D.18 Subtype CRF16_A2D Sequences 882
D.19 Subtype CRF18_cpx Sequences 883
D.20 Subtype CRF20_BG Sequences 884

D.21 Subtype CRF22_01A1 Sequences 884
D.22 Subtype CRF26_A5U Sequences 885
D.23 Subtype CRF27_cpx Sequences 885
D.24 Subtype CRF37_cpx Sequences 886
D.25 Subtype CRF45_cpx Sequences 887
D.26 Subtype CRF63_02A6 Sequences 888

E Compendium of Subtype G Sequences in Alphanumeric Order 889
E.1 Complete Subtype G Interpretability Patterns 889

F SARS-CoV-2 Complete Results per Subsampling 901

G SARS-CoV-2 Complete Results per Number of Epochs 913

H SARS-CoV-2 Recombinant Sequences Study Case. Statistical Data on the Types of Nucleotides Present in the Arithmetic Series (Arithmetic Ratio = 6) at S Protein 919

I SARS-CoV-2 Recombinant Sequences Study Case. Statistical Data on the Types of Nucleotides Present in the Arithmetic Series (Arithmetic Ratio = 6) at M Protein 927

J SARS-CoV-2 Recombinant Sequences Study Case. Statistical Data on the Types of Nucleotides Present in the Arithmetic Series (Arithmetic Ratio = 6) at N Protein 935

K SARS-CoV-2 Non-Recombinant Sequences Study Case. Statistical Data on the Types of Nucleotides Present in the Arithmetic Series (Arithmetic Ratio = 6) at S Protein 943

L Compendium of Hot Zones in SARS-CoV-2 Sequences Under Analysis 951
L.1 Recombinant Sequences Group 951
L.2 Non-Recombinant Sequences 954

M Scientific Evidence 957

References 1057

List of Figures

1.1 Number of Complete Sequences of Various Viruses as of March 2023 13
1.2 Basic Methodology 21
1.3 Research Progression 23

2.1 Codon Wheel 31
2.2 Three-Dimensional Visualization of a Protein and the Corresponding Genomic Sequence 35
2.3 CRF23_BG Genome Map 37
2.4 CRF18_cpx Genome Map 38
2.5 HIV-1 Multiple Sequence Alignment 39

3.1 Machine Learning vs. Conventional Programming 49
3.2 Temporal progression of AI 50
3.3 Typical elements of a CNN layer 54
3.4 ResNet-based Architecture 55
3.5 Inception V3 Architecture 56
3.6 VGG-16 Architecture Schematic 57
3.7 Diagram of the Technological Background Review Methodology . . 70
3.8 HIV-1 Complete Consensus Multiple Sequence Alignment Excerpt 1 115
3.9 HIV-1 Complete Consensus Multiple Sequence Alignment Excerpt 2 116
3.10 Visual example of scenarios contained in a dataset 120

4.1 Procedure Comparison 145
4.2 Spectrograms Schemes 147
4.3 Jet Color Map 148

5.1 HIV-1 Gene Map 159
5.2 Schematic Representation of HIV-1 Virion 160
5.3 Methodology Block Diagram 165
5.4 Confusion Matrix 172
5.5 ROC Curve 174
5.6 Phylogenetic Tree of HIV-1 178
5.7 Confusion Matrix Scheme 180

5.8 Hot Zones Interpreted by Grad-CAM and LIME on Concatenated Spectrograms . . . 191
5.9 Hot Zones Interpreted by Grad-CAM and LIME on Superposed Spectrograms . . . 193
5.10 Partial Phylogenetic Tree Including Subtype O MH705149 and KU168283 Sequences . . . 199
5.11 Partial Phylogenetic Tree Including Subtype CRF22_01A1 AY371165 and JN864058 Sequences . . . 201
5.12 Partial Phylogenetic Tree Including Subtype CRF01_AE MH078560 and MH327763 Sequences . . . 203
5.13 Partial Phylogenetic Tree Including Subtype CRF106_cpx MT276998 and MT277002 Sequences . . . 204
5.14 Partial Phylogenetic Tree Including Subtype C DQ207941 and DQ351235 Sequences . . . 208
5.15 Partial Phylogenetic Tree Including Subtype O AB485666 and MH705149 Sequences . . . 210
5.16 Partial Phylogenetic Tree Including Subtype CRF01_AE AY358044 and AY358061 Sequences . . . 212
5.17 Partial Phylogenetic Tree Including Subtype CRF20_BG KT276270 and MW062981 Sequences . . . 214
5.18 Partial Phylogenetic Tree Including Subtype CRF58_01B KC522031 and KC522033 Sequences . . . 219
5.19 Partial Phylogenetic Tree Including Subtype CRF02_AG AY371129 and AY371130 Sequences . . . 224
5.20 Partial Phylogenetic Tree Including Subtype CRF01_AE JX960631 and Subtype CRF02_AG KF561435 Sequences . . . 228
5.21 Genome Map of Recombinant Subtype CRF01_AE . . . 228
5.22 Genome Map of Recombinant Subtype CRF02_AG . . . 229
5.23 Partial Phylogenetic Tree Including Subtype CRF01_AE AB253423 and Subtype CRF69_01B AB845349 Sequences . . . 230
5.24 Genome Map of Recombinant Subtype CRF69_01B . . . 230
5.25 Genome Map of Recombinant Subtype CRF55_01B . . . 232
5.26 Partial Phylogenetic Tree Including Subtype CRF01_AE JX448059 and Subtype CRF55_01B JX574661 Sequences . . . 232
5.27 Genome Map of Recombinant Subtype CRF53_01B . . . 235
5.28 Genome Map of Recombinant Subtype CRF52_01B . . . 235
5.29 Partial Phylogenetic Tree Including Subtype CRF53_01B DQ366663 and Subtype CRF52_01B DQ366664 Sequences . . . 236
5.30 Partial Phylogenetic Tree Including Subtype G FJ389363 and Subtype CRF30_0206 AJ508596 Sequences . . . 241

5.31 Partial Phylogenetic Tree Including Subtype G FJ389363 and Subtype CRF53_01B DQ366663 Sequences 244

6.1 Total Hot Zones per Category (Concatenated Spectrograms) 259
6.2 Total Hot Zones per Category (Superposed Spectrograms) 260

8.1 Total Explainability Patterns by Category (Superposed Spectrograms) 369
8.2 Partial Phylogenetic Tree Including Subtype U AB485665 and KY392773 Sequences 371
8.3 Genome Map of Recombinant Subtype CRF04_cpx 373
8.4 Genome Map of Recombinant Subtype CRF18_cpx 374
8.5 Partial Phylogenetic Tree Including Subtype U AY046058, KY392780, MH078559 and MK095236 Sequences 380
8.6 Genome Map of Recombinant Subtype CRF45_cpx 384
8.7 Partial Phylogenetic Tree Including Subtype U FJ388921, EF029066, EF029067, EF029068, EF029069 and KY392775 Sequences 387
8.8 Partial Phylogenetic Tree Including Subtype U HM215249 and HM215251 Sequences 392
8.9 Genome Map of Recombinant Subtype CRF26_A5U 396
8.10 Genome Map of Recombinant Subtype CRF16_A2D 399
8.11 Partial Phylogenetic Tree Including Subtype U JF683772 Sequence 400
8.12 Genome Map of Recombinant Subtype CRF02_AG 404
8.13 Genome Map of Recombinant Subtype CRF63_02A6 404
8.14 Partial Phylogenetic Tree Including Subtype U JX140679 Sequence 406
8.15 Genome Map of Recombinant Subtype CRF37_cpx 411
8.16 Genome Map of Recombinant Subtype CRF01_AE 412
8.17 Partial Phylogenetic Tree Including Subtype U KP718920 Sequence 414
8.18 Genome Map of Recombinant Subtype CRF11_cpx 418
8.19 Genome Map of Recombinant Subtype CRF13_cpx 422
8.20 Partial Phylogenetic Tree Including Subtype U MN153489 Sequence 423
8.21 Partial Phylogenetic Tree Including Subtype U MN153492 Sequence 428
8.22 Genome Map of Recombinant Subtype CRF12_BF 431
8.23 Partial Phylogenetic Tree Including Subtype U MN736701 and MN736702 Sequences 432
8.24 Genome Map of Recombinant Subtype CRF20_BG 434
8.25 Genome Map of Recombinant Subtype CRF22_01A1 435
8.26 Genome Map of Recombinant Subtype CRF06_cpx 436
8.27 Genome Map of Recombinant Subtype CRF27_cpx 437
8.28 Genome Map of Recombinant Subtype CRF56_cpx 454
8.29 Genome Map of Recombinant Subtype CRF36_cpx 469

9.1 Total Explainability Patterns by Category (Superposed Spectrograms) 479
9.2 Subtype G Phylogenetic Group 1 481
9.3 Subtype G Phylogenetic Group 1 Clade 1 481
9.4 Subtype G Phylogenetic Group 1 Clade 2 484
9.5 Subtype G Phylogenetic Group 1 Clade 3 486
9.6 Subtype G Phylogenetic Group 1 Clade 4 490
9.7 Subtype G Phylogenetic Group 1 Clade 5 496
9.8 Subtype G Phylogenetic Group 2 500
9.9 Subtype G Phylogenetic Group 2 Clade 1 501
9.10 Subtype G Phylogenetic Group 2 Clade 2 504
9.11 Subtype G Phylogenetic Group 3 511

10.1 Schematic Representation of SARS-CoV-2 Virion 547
10.2 SARS-CoV-2 Genomic Element Analysis Block Diagram 562

11.1 *Nidovirales* Taxonomy 592
11.2 Genomic Spectrogram Schemes 603
11.3 Two-Stage Transfer Learning Methodology 604
11.4 Three-step Explainability Scheme 606
11.5 Total Hot Zones (SUB_05, 200 Epochs) 607
11.6 Confusion Matrix Scheme 609
11.7 Statistical Graphics 615
11.8 Identification of Total Hot Zones (Step 3) in Non-recombinants for 200 Epochs 617
11.9 Localization of Hot Zones in Non-recombinant Sequences 618
11.10 Identification of Total Hot Zones (Step 3) in Recombinants for 200 Epochs 619
11.11 Localization of Hot Zones in Recombinant Sequences 620
11.12 Vertical Positioning at 200 Epochs Configuration Subsampling 06 . 621

12.1 Two-Stage Transfer Learning Methodology Non-Recombinant Variant Detection 632
12.2 Two-Step Explainability 634
12.3 VARIANT_SUBSAMPLING_01 Training Curve 637
12.4 VARIANT_SUBSAMPLING_02 Training Curve 638
12.5 VARIANT_SUBSAMPLING_01 Confusion Matrix 641
12.6 VARIANT_SUBSAMPLING_02 Confusion Matrix 641
12.7 Total Hot Zones per Variant (VARIANT_SUBSAMPLING_01) . . 643
12.8 Total Hot Zones per Variant (VARIANT SUBSAMPLING 02) . . . 646

13.1 Graphic Representation of the Nucleotide Excerpt According to Arithmetic Series (Frames) with a Ratio of 6 671
13.2 Schematic representation of the SARS-CoV-2 Spike Protein Sub-elements 680

14.1 Total Hot Zones Recombinant Feature of Superposed Spectrograms for HIV-1 and SARS-CoV-2 696
14.2 Total Hot Zones per Category (Concatenated Spectrograms) of HIV-1 699
14.3 Total Hot Zones per Category (Superposed Spectrograms) of HIV-1 699
14.4 Identification of Total Hot Zones (Step 3) in SARS-CoV-2 Sequences for 200 Epochs (Superposed Spectrograms) 709

16.1 Generation of Genomic Segments Individually 790
16.2 Simultaneous Generation of the Entire Genome 793

List of Tables

1.1 Percentage of similarity with human (*Homo sapiens*) ACE2, in % . 7

2.1 DNA Notation 30

3.1 Results of Article Identification by Search Topic 69

5.1 Non-Recombinant HIV-1 Sequences up to 2020 166
5.2 Recombinant HIV-1 Sequences up to 2020 167
5.3 Dataset Structure 169
5.4 Test Bench 171
5.5 VGG-16 Concatenated Spectrograms Results 181
5.6 ResNET-101 Results Concatenated Spectrograms 182
5.7 Inception V3 Results Concatenated Spectrograms 183
5.8 VGG-16 Results Superposed Spectrograms 184
5.9 ResNET-101 V3 Results Superposed Spectrograms 185
5.10 Inception V3 Results Superposed Spectrograms 186
5.11 Summary of Configurations that Achieved the Best Results 187
5.12 Total Distances 189
5.13 Test Results of Three Replicates - Concatenated Spectrograms . . . 189
5.14 Test Results of Three Replicates - Superposed Spectrograms 190
5.15 Comparison of Explainability of Non-Recombinant Concatenated Spectrograms 196
5.16 Comparison of Explainability Recombinant Concatenated Spectrograms 200
5.17 Comparison of Explainability of Non-Recombinant Superposed Spectrograms 205
5.18 Comparison of Explainability of Recombinant Superposed Spectrograms 211
5.19 Comparison of Explainability of HIV-1 Phylogenetically Close Non-Recombinant Sequences in Concatenated and Superposed Spectrograms Case 1 216
5.20 Comparison of Explainability of HIV-1 Phylogenetically Close Non-Recombinant Sequences in Concatenated and Superposed Spectrograms Case 2 217

5.21 Comparison of Explainability of HIV-1 Phylogenetically Close Recombinant Sequences in Concatenated and Superposed Spectrograms Case 1 218
5.22 Comparison of Explainability of HIV-1 Phylogenetically Close Recombinant Sequences in Concatenated and Superposed Spectrograms Case 2 220
5.23 Comparison of Explainability of HIV-1 Phylogenetically Close Recombinant Sequences in Concatenated and Superposed Spectrograms Case 3 222
5.24 Comparison of Explainability of HIV-1 Phylogenetically Close Recombinant Sequences in Concatenated and Superposed Spectrograms Case 4 223
5.25 Similarity between Sequences of Different Subtypes 226
5.26 Similarity between CRF53_01B DQ366663 and CRF52_01B DQ366664 234
5.27 Non-recombinant Concatenated Failed Results 237
5.28 Recombinant Concatenated Failed Results 237
5.29 Non-Recombinant Superposed Failed Results 238
5.30 Recombinant Superposed Failed Results 238
5.31 Non-Recombinant FJ389363 (subtype G) Erroneously Classified as Recombinant (Concatenated Spectrograms) 240
5.32 Recombinant JN864049 (subtype CRF22_01A1) Erroneously Classified as Non-Recombinant (Concatenated Spectrograms) 242
5.33 Non-Recombinant FJ389363 (subtype G) Erroneously Classified as Recombinant (Superposed Spectrograms) 243
5.34 Recombinant KX907417 (subtype CRF41_CD) Erroneously Classified as Non-Recombinant (Superposed Spectrograms) 245

6.1 HIV-1 Non-recombinant Error Rates per Subtype 257
6.2 HIV-1 Recombinant Error Rates per Subtype 258
6.3 Total Hot Zones per Non-Recombinant Category and Representative Subtypes 262
6.4 Total Hot Zones per Recombinant Subtype 275

7.1 Total Hot Zones per Category and Subtype F1 299
7.2 Subtype F1 Sequences Classified by the CNN as Non-Recombinant, Confidence Score $\geq 90\%$ in Concatenated Spectrograms 300
7.3 Subtype F1 Sequences Classified by the CNN as Non-Recombinant, $70\% \leq$ Confidence Score $< 90\%$ in Concatenated Spectrograms . . . 306
7.4 Subtype F1 Sequences Classified by the CNN as Non-Recombinant, $50\% \leq$ Confidence Score $< 70\%$ in Concatenated Spectrograms . . . 307

7.5 Subtype F1 Sequences Misclassified by the CNN as Recombinant in Concatenated Spectrograms (Confidence Score $< 50\%$) 310
7.6 Subtype F1 Sequences Classified by the CNN as Non-Recombinant, Confidence Score $\geq 90\%$ in Superposed Spectrograms 311
7.7 Subtype F1 Sequences Classified by the CNN as Non-Recombinant, $70\% \leq$ Confidence Score $< 90\%$ in Superposed Spectrograms 316
7.8 Subtype F1 Sequences Classified by the CNN as Non-Recombinant, $50\% \leq$ Confidence Score $< 70\%$ in Superposed Spectrograms 318
7.9 Subtype F1 Sequences Misclassified by the CNN as Recombinant in Superposed Spectrograms (Confidence Score $< 50\%$) 320
7.10 Subtype F1 Misclassified Sequences as Recombinant in One Type of Spectrogram 322
7.11 Subtype F1 Misclassified Sequences as Recombinant in Both Types of Spectrogram 325
7.12 Total Hot Zones per Category and Subtype CRF22_01A1 328
7.13 Subtype CRF22_01A1 Sequences Classified by the CNN as Recombinant, Confidence Score $\geq 90\%$ in Concatenated Spectrograms . . 330
7.14 Subtype CRF22_01A1 Sequences Classified by the CNN as Recombinant, $50\% \leq$ Confidence Score $< 90\%$ in Concatenated Spectrograms 332
7.15 Subtype CRF22_01A1 Sequences Misclassified by the CNN as Non-Recombinant in Concatenated Spectrograms (Confidence Score $< 50\%$) 333
7.16 Subtype CRF22_01A1 Sequences Classified by the CNN as Recombinant, Confidence Score $\geq 90\%$ in Superposed Spectrograms 335
7.17 Subtype CRF22_01A1 Sequences Classified by the CNN as Recombinant, $70\% \leq$ Confidence Score $< 90\%$ in Superposed Spectrograms 337
7.18 Subtype CRF22_01A1 Sequences Classified by the CNN as Recombinant, $50\% \leq$ Confidence Score $< 70\%$ in Superposed Spectrograms 337
7.19 Subtype CRF22_01A1 Sequences Misclassified by the CNN as Non-Recombinant in Superposed Spectrograms (Confidence Score $< 50\%$) 338
7.20 Subtype CRF22_01A1 Misclassified Sequences as Non-Recombinant in One Type of Spectrogram 339
7.21 Subtype CRF22_01A1 Misclassified Sequences as Non-Recombinant in Both Types of Spectrogram 340
7.22 Total Hot Zones per Category and Subtype O 342
7.23 Subtype O Sequences with a Grad-CAM Hot Zone in the Low Frequency Range of the Medium Area of Concatenated Spectrograms 344
7.24 Subtype O Sequences with a Grad-CAM Hot Zone Different of the One in the Low Frequency Range of the Medium Area of Concatenated Spectrograms 350

7.25 Subtype O Sequences with a Grad-CAM Hot Zone in the High Frequency Range Not in 5'LTR of Superposed Spectrograms 352
7.26 Subtype O Sequences with a Grad-CAM Hot Zone in the Low Frequency Range at 3'LTR of Superposed Spectrograms 356
7.27 Subtype O Sequences with a Grad-CAM Hot Zone Different from the Previous Ones of Superposed Spectrograms 357
7.28 Subtype O Misclassified Sequence in Both Types of Spectrogram . . 360

8.1 CNN Error Rates of Classification 368
8.2 KY392773 (Subtype U) and CRF04_cpx Explainability Patterns . . 373
8.3 AB585664 and AB585665 (Subtype U) and CRF18_cpx Explainability Patterns 375
8.4 Subtype H Explainability Patterns 377
8.5 AY046058, KY392780, MH078559 and MK095236 (Subtype U) Explainability Patterns 381
8.6 Subtype A7 Sequences Explainability Patterns 382
8.7 Subtype CRF45_cpx Sequences Explainability Patterns 383
8.8 Subtype CRF02_AG and Subtype CRF09_cpx Sequences Explainability Patterns 385
8.9 FJ388921, EF029066, EF029067, EF029068, EF029069, and KY392775 (Subtype U), and Subtype K Explainability Patterns 388
8.10 Subtype F2 Explainability Patterns 390
8.11 HM215249 and HM215251 (Subtype U) Explainability Patterns . . 393
8.12 Subtype CRF26_A5U Sequences Explainability Patterns 395
8.13 Subtype A2 Sequences Explainability Patterns 397
8.14 Subtype CRF16_A2D Sequences Explainability Patterns 398
8.15 JF683772 (Subtype U), CRF63_02A6 and CRF02_AG Sequences Explainability Patterns 401
8.16 JX140679 (Subtype U) and Subtype CRF02_AG Sequences Explainability Patterns 407
8.17 Subtype A1 Sequences Explainability Patterns 408
8.18 Subtype CRF37_cpx Sequences Explainability Patterns 411
8.19 KP718920 (Subtype U) and Subtype CRF11_cpx Sequences Explainability Patterns 415
8.20 Subtype CRF27_cpx Sequences Explainability Patterns 419
8.21 Subtype CRF13_cpx Sequences Explainability Patterns 420
8.22 MN153489 (Subtype U) Explainability Patterns 424
8.23 Subtype A1 Sequences Explainability Patterns 424
8.24 Subtype A6 Sequences Explainability Patterns 425

8.25 MN153492 (Subtype U) and Subtypes F2, F1 and CRF12_BF Sequences Explainability Patterns 429
8.26 MN736701 and MN736702 (Subtype U) Explainability Patterns . . 433
8.27 Subtypes G, CRF20_BG, CRF22_01A1 Sequences Explainability Patterns 434
8.28 Subtypes CRF18_cpx, CRF06_cpx, CRF27_cpx Sequences Explainability Patterns 436
8.29 Subtype CRF02_AG Sequences Explainability Patterns 438
8.30 FJ388921, EF029066, EF029067, EF029068, EF029069 and KY392775 Results 442
8.31 EF029067, EF029068 and EF029069 Most Analogous Explainability Patterns 444
8.32 EF029066 Most Analogous Explainability Patterns 445
8.33 HM215249 and HM215251 Results 446
8.34 JF683772 Results and Most Analogous Explainability Patterns . . . 448
8.35 JX140679 Most Analogous Explainability Patterns 449
8.36 MN153489 Results and Most Analogous Explainability Patterns . . 450
8.37 MN153492 Results and Most Analogous Explainability Patterns . . 451
8.38 Phylogenetically-Close-to-MN153492 F1 Sequences, classified as Recombinant 453
8.39 MN736701 and MN736702 Results 453
8.40 MN736701 Most Analogous Explainability Patterns 455
8.41 MN736702 Most Analogous Explainability Patterns 456
8.42 AB485664 and AB485665 Results 458
8.43 AB485665 Most Analogous Explainability Patterns 459
8.44 KY392773 Results and Most Analogous Explainability Patterns . . 461
8.45 Comparison between KY392774 and JX140679 Explainability Patterns 463
8.46 MK095236, KY392780, AY046058 and MH078559 Results 463
8.47 MK095236 Most Analogous Explainability Patterns 465
8.48 AY046058, KY392780 and MH078559 Most Analogous Explainability Patterns 466
8.49 KP718920 Results and Most Analogous Explainability Patterns . . 467
8.50 Similarities among AF005496 (Subtype H), FN392877 (CRF45_cpx) and subtype CRF02_AG Sequences AF063223 and AY371140 . . . 470
8.51 Similarities between the Explainability Patterns of AF190128 (Subtype H) and MN791873 (Subtype A1) 471
8.52 EF029066, EF029067, EF029068, and EF029069 Explainability Patterns 472

9.1 Subtype Phylogenetic Group 1 Clade 1 Results 482
9.2 Subtype G Phylogenetic Group 1 Clade 2 Results 485

9.3 Subtype G Phylogenetic Group 1 Clade 3 Results 487
9.4 Subtype G Phylogenetic Group 1 Clade 4 Results 490
9.5 Subtype G Phylogenetic Group 1 Clade 5 Results 497
9.6 Subtype G Phylogenetic Group 2 Clade 1 Results 501
9.7 Subtype G Phylogenetic Group 2 Clade 2 Results 505
9.8 Subtype G Group 3 Results 510
9.9 AB231893, AB287003, and AB287004 517
9.10 AB485662, AB485663, AF061640, and AF061641 518
9.11 JX140674, KX228812, and EU786670 519
9.12 JX140675 and KX228801 520
9.13 MF614607 and MF614606 520
9.14 Subtype G Recombinant Feature Prevalence 522
9.15 Subtype G Explainability Patterns Cluster 1 523
9.16 Subtype G Explainability Patterns Cluster 2 529
9.17 Subtype G Explainability Patterns Cluster 3 530
9.18 Subtype G Explainability Patterns Cluster 4 535
9.19 Subtype G Explainability Patterns Cluster 5 536
9.20 Subtype G Explainability Patterns Cluster 6 537
9.21 Subtype G Explainability Patterns Cluster 7 539
9.22 Subtype G Explainability Patterns Cluster 8 540

10.1 Breakdown of Genomic Elements in the SARS-CoV2 Wuhan-Hu-1 Reference Genome Sequence (NC_045512.2) 548
10.2 Implementation Process of the Generated Final Sequence - Step 1 . 568
10.3 Implementation Process of the Generated Final Sequence - Step 2 . 568
10.4 Implementation Process of the Generated Final Sequence - Step 3 . 569
10.5 Most Common First and Last Triplets of Genomic Elements of SARS-CoV-2 Sequences and their Number of Occurrences 573
10.6 Number of Sequences Detected Without Labeling Errors 577
10.7 Measures of Consistency Level by Genomic Element 579

11.1 Non-recombinant SARS-CoV-2 sequences compendium 600
11.2 Recombinant SARS-CoV-2 sequences compendium 601
11.3 Dataset structure 602
11.4 Best results per subsampling 610
11.5 Best Results per Number of Epochs 612
11.6 Summary Table of Total Hot Zones Step Two 613
11.7 Total Hot Zones per Configuration 614
11.8 Location of protein coding regions in the SARS-CoV2 Wuhan-Hu-1 reference genome sequence (NC_045512.2) 615

12.1 Clustering of Complete Non-Recombinant SARS-CoV-2 Sequences . 628
12.2 Clustering of Complete Non-Recombinant SARS-CoV-2 Sequences After Excluding Minority Variants 630
12.3 Number of Sequences per Category per Subsampling 631
12.4 Dataset Composition per Subsampling 631
12.5 Performance Measurement Ratios of Non-Recombinant VARIANT SUBSAMPLING 01 639
12.6 Performance Measurement Ratios Non-Recombinant VARIANT SUBSAMPLING 02 640
12.7 Comparative Table of Total Hot Zones in Both Subsamplings 649

13.1 Sequences under Study and Their Lineages 663
13.2 Lineage Distribution of Recombinant Sequences under Study 665
13.3 Lineage Distribution of Non-Recombinant Sequences under Study . 666
13.4 Variant Distribution of Non-Recombinant Sequences under Study . 667
13.5 Sixth Nucleotide Position for the Selected 39 Successfully Identified Recombinant Sequences Using the Pretrained CNN 668
13.6 SARS-CoV-2 Recombinant Sequences Study Case. Incidence of each nucleotide in the Frames with a Ratio of 6 (S Protein) 673
13.7 SARS-CoV-2 Recombinant Sequences Study Case. Statistical Results Concerning Each Frame of 6 Nucleotides (M Protein) 674
13.8 SARS-CoV-2 Recombinant Sequences Study Case. Statistical Results Concerning Each Frame of 6 Nucleotides (N Protein) 675
13.9 SARS-CoV-2 Non-Recombinant Sequences Study Case. Statistical analysis of the incidence of each nucleotide in the Frames with a Ratio of 6 (S Protein) 676
13.10Comparison of Average Nucleotide Contents between Recombinant and Non-Recombinant Sequences 677
13.11Common Nucleotides of S Gene in the 39 Analyzed Recombinant Sequences 678
13.12Common Nucleotides of S Gene in the 39 Analyzed Non-Recombinant Sequences 681
13.13Common Nucleotides in the S Protein Sequences, Shared between Recombinants and Non-Recombinants 683
13.14Common Nucleotides in the S Protein Sequences, Specific to Recombinants 684

14.1 Compilation of HIV-1 Non-Recombinant Subtypes with Hot Zones Associated with Nucleotide Periodicities of 3 or Their Multiples . . 701

14.2 Compilation of HIV-1 Recombinant Subtypes with Hot Zones Associated with Nucleotide Periodicities of 3 or Their Multiples 703
14.3 Comparative Table of Total Hot Zones of the Main Non-Recombinant Variants of SARS-CoV-2, Divided into Two Subsamplings 711

15.1 Optimal Hyperparameter Configurations in Concatenated Spectrograms 737
15.2 Optimal Hyperparameter Configurations in Superposed Spectrograms 737
15.3 Most Relevant Results of VGG-16 on Concatenated Spectrograms . 738
15.4 Most Relevant Results of VGG-16 on Superposed Spectrograms . . 738
15.5 Optimal Hyperparameter Configurations (Superposed Spectrograms) 765

A.1 Subtype F1 CNN Classification Results and Interpretability Patterns 837

B.1 Subtype CRF22_01A1 CNN Classification Results and Interpretability Patterns 847

C.1 Subtype O CNN Classification Results and Interpretability Patterns 851

D.1 Subtype U Interpretability Patterns 859
D.2 Subtype A1 Interpretability Patterns 862
D.3 Subtype A2 Interpretability Patterns 865
D.4 Subtype A6 Interpretability Patterns 865
D.5 Subtype A7 Interpretability Patterns 867
D.6 Subtype F1 Interpretability Patterns 867
D.7 Subtype F2 Interpretability Patterns 868
D.8 Subtype G Interpretability Patterns 869
D.9 Subtype H Interpretability Patterns 870
D.10 Subtype K Interpretability Patterns 871
D.11 Subtype CRF02_AG Interpretability Patterns 872
D.12 Subtype CRF04_cpx Interpretability Patterns 876
D.13 Subtype CRF06_cpx Interpretability Patterns 877
D.14 Subtype CRF09_cpx Interpretability Patterns 877
D.15 Subtype CRF11_cpx Interpretability Patterns 878
D.16 Subtype CRF12_BF Interpretability Patterns 881
D.17 Subtype CRF13_cpx Interpretability Patterns 881
D.18 Subtype CRF16_A2D Interpretability Patterns 882
D.19 Subtype CRF18_cpx Interpretability Patterns 883
D.20 Subtype CRF20_BG Interpretability Patterns 884
D.21 Subtype CRF22_01A1 Interpretability Patterns 884
D.22 Subtype CRF26_A5U Interpretability Patterns 885
D.23 Subtype CRF27_cpx Interpretability Patterns 885

D.24 Subtype CRF37_cpx Interpretability Patterns 886
D.25 Subtype CRF45_cpx Interpretability Patterns 887
D.26 Subtype CRF63_02A6 Interpretability Patterns 888

E.1 Subtype G Interpretability Patterns 889

F.1 Results in Subsampling 01 by the Number of Epochs 902
F.2 Results in Subsampling 02 by the Number of Epochs 903
F.3 Results in Subsampling 03 by the Number of Epochs 904
F.4 Results in Subsampling 04 by the Number of Epochs 905
F.5 Results in Subsampling 05 by the Number of Epochs 906
F.6 Results in Subsampling 06 by the Number of Epochs 907
F.7 Results in Subsampling 07 by the Number of Epochs 908
F.8 Results in Subsampling 08 by the Number of Epochs 909
F.9 Results in Subsampling 09 by the Number of Epochs 910
F.10 Results in Subsampling 10 by the Number of Epochs 911

G.1 Results at 10 Epochs 913
G.2 Results at 25 Epochs 914
G.3 Results at 50 Epochs 914
G.4 Results at 75 Epochs 915
G.5 Results at 100 Epochs 915
G.6 Results at 150 Epochs 916
G.7 Results at 200 Epochs 916
G.8 Results at 250 Epochs 917

H.1 Frame 6, 12, 18, ... 919
H.2 Frame 7, 13, 19, ... 921
H.3 Frame 8, 14, 20, ... 922
H.4 Frame 9, 15, 21, ... 923
H.5 Frame 10, 16, 22, ... 924
H.6 Frame 11, 17, 23, ... 925

I.1 Frame 6, 12, 18, ... 927
I.2 Frame 7, 13, 19, ... 929
I.3 Frame 8, 14, 20, ... 930
I.4 Frame 9, 15, 21, ... 931
I.5 Frame 10, 16, 22, ... 932
I.6 Frame 11, 17, 23, ... 933

J.1 Frame 6, 12, 18, ... 935
J.2 Frame Series 7, 13, 19, ... 937

J.3 Frame 8, 14, 20, ... 938
J.4 Frame 9, 15, 21, ... 939
J.5 Frame 10, 16, 22, ... 940
J.6 Frame 11, 17, 23, ... 941

K.1 Frame 6, 12, 18, ... 943
K.2 Frame 7, 13, 19, ... 945
K.3 Frame 8, 14, 20, ... 946
K.4 Frame 9, 15, 21, ... 947
K.5 Frame 10, 16, 22, ... 948
K.6 Frame 11, 17, 23, ... 949

L.1 Case Study of Recombinant Sequences Group 951
L.2 Case Study of Non-Recombinant Sequences Group 954

List of Abbreviations

4acC . . . . . N4-Acetyldeoxycytosine.

6-HB . . . . . Six-Helix Bundle.

A . . . . . Adenine.

ABC . . . . . ATP-Binding Cassette.

ACE2 . . . . . Angiotensin-Converting Enzyme 2.

AD . . . . . Activation Domain.

ADHD . . . . Attention-Deficit/Hyperactivity Disorder.

AI . . . . . Artificial Intelligence.

AIDS . . . . . Acquired Immunodeficiency Syndrome.

ALS . . . . . Amyotrophic Lateral Sclerosis.

AML . . . . . Acute Myeloid Leukemia.

AMP . . . . . Antimicrobial peptide.

ARDS . . . . . Acute Respiratory Distress Syndrome.

ARG . . . . . Antibiotic Resistance Gene.

ART . . . . . Antiretroviral Treatment.

ASD . . . . . Autism Spectrum Disorder.

AUC . . . . . Area Under the Curve.

BAM . . . . . Binary Sequence Alignment/Map Format.

BCE . . . . . B-Cell Epitopes.

BCI . . . . . Brain-Computer Interface.

BCoV . . . . . Bovine Coronavirus.

BEAST . . . . Bayesian Evolutionary Analysis Sampling Trees.

BSA . . . . . Bulked Segregant Analysis.

BtCoV . . . . Bat Coronavirus.

BtEoCoV . . . Roussetus Bat Coronavirus.

BtHpCoV-ZJ13 Bat Hp-betacoronavirus Zhejiang2013.

BtKYNL63 . . NL63-Related Bat Coronavirus Strain BtKYNL63-9b.

BtMiCoV . . . Miniopterus Bat Coronavirus.

BtMiCoV-1 . Miniopterus Bat Coronavirus 1.

BtMy-Sax11 . Myotis Ricketti Alphacoronavirus Sax-2011.

BtNy-Sc13 . . Nyctalus Velutinus Alphacoronavirus SC-2013.

BtPiCoV . . . Pipistrellus Bat Coronavirus.

BtRfCoV . . . Rhinolophus Ferrumeguinum Alphacoronavirus.

BtRhCoV . . . Rhinolophus Bat Coronavirus.

BtRoCoV . . . Rousettus Bat Coronavirus.

BtScCoV-512 Scotophilus Bat Coronavirus 512.

BtTyCoV . . . Tylonycteris Bat Coronavirus.

BuCoV-HKU11 Bulbul coronavirus HKU11.

BWCoV-SW1 Beluga Whale Coronavirus SW1.

C . . . . . Cytosine.

CA . . . . . Community-Acquired.

CAGE . . . . . Cap Analysis of Gene Expression.

CCoV . . . . . Canine Coronavirus.

CCoV Type I Canine Coronavirus Type I.

CCoV Type II Canine Coronavirus Type II.

cDNA . . . . . Circular DNA.

cDNA . . . . . Complementary DNA.

CDS . . . . . Coding Sequence.

ChIP . . . . . Chromatin Immunoprecipitation.

CITE-seq . . . Cellular Indexing of Transcriptomes and Epitopes by Sequencing.

CGR . . . . . Chaos Game Representation.

CJD . . . . . CreutzfeldtJakob Disease.

Clustal . . . . CLUSTer ALignment.

CMCoV-HKU21 Common Moorhen Coronavirus HKU21.

CNA . . . . . Copy Number Alteration.

CNN . . . . . Convolutional Neural Network.

CNV . . . . . Copy Number Variation.

COVID . . . . Coronavirus Disease.

COVID-19 . . Coronavirus Disease 2019.

CPIs . . . . . CompoundProtein Interactions.

CPX . . . . . Complex. In the context of HIV-1 subtype nomenclature, the emergence of these initials indicates that the CRF (HIV-1 Circulating Recombinant Form) is a mosaic of more than 3 subtypes.

CRCoV . . . . Canine Respiratory Coronavirus.

CRF . . . . . HIV-1 Circulating Recombinant Form. CRFs are a specific type of HIV-1 subtype that results from the recombination of two or more different subtypes of HIV-1.

CRISP . . . . Clustered Regularly Interspaced Short Palindromic Repeats.

CT . . . . . Computed Tomography.

CTCF . . . . . CCCTC-Binding Factor.

CVD . . . . . Cardiovascular Disease.

CVMLUB . . Computer Vision and Machine Learning at the University of Barcelona.

CYP450 . . . . Cytochrome P450.

DDR . . . . . DNA Damage Repair.

DFT . . . . . Discrete Fourier Transform.

DGE . . . . . Digital Gene Expression.

DHS . . . . . DNase I Hypersensitive Site.

DL . . . . . Deep Learning.

DMR . . . . . DNA Methylated Region.

DNA . . . . . Deoxyribonucleic Acid.

DNA-Seq . . . DNA Sequencing.

DNN . . . . . Deep Neural Network.

DOT . . . . . Directly Observed Therapy.

DSP . . . . . Digital Signal Processing.

DTW . . . . . Dynamic Time Warping.

DWT . . . . . Dyadic wavelet transform.

E . . . . . Envelope Protein.

eccDNA . . . . Extrachromosomal Circular DNA.

ECG . . . . . Electrocardiogram.

ECoV . . . . . Equine Coronavirus.

EEG . . . . . Electroencephalogram.

EMBL . . . . . European Molecular Biology Laboratory.

EMR . . . . . Electronic Medical Records.

ENA . . . . . European Nucleotide Archive.

env . . . . . Envelope.

EPI . . . . . Enhancer-Promoter Interaction.

ER . . . . . Endoplasmic Reticulum.

EriCoV-1 . . . Hedgehog Coronavirus 1.

ESC . . . . . Embryonic Stem Cell.

ESCC . . . . . Esophageal Squamous Cell Carcinoma.

FASTA . . . . Fast All.

FCGS . . . . . Frequency Chaos Game Signal.

FCoV . . . . . Feline Coronavirus.

FCoV Type I Feline Coronavirus Type I.

FCoV Type II Feline Coronavirus Type II.

FECoV . . . . Feline Enteric Coronavirus.

FFT . . . . . Fast Fourier Transform.

FIPV . . . . . Feline Infectious Peritonitis Virus.

FL . . . . . Federated Learning.

FMT . . . . . Fecal Microbiota Transplantation.

FP . . . . . Fusion Peptide.

FPM . . . . . Fourier Ptychographic Microscopy.

FRCoV . . . . Ferret Coronavirus.

FS . . . . . Functional Similarity.

FTIR . . . . . Fourier Transform Infrared Spectroscopy.

G . . . . . Guanine.

G4 . . . . . G-Quadruplex.

G-E . . . . . GeneEnvironment.

gag . . . . . Group-Specific Antigen.

GAN . . . . . Generative Adversarial Network.

GEP . . . . . Gene Expression Profiles.

GLProbs . . . Genome-wide Likelihood-based Probabilistic Inference.

GNN . . . . . Graph Neural Network.

GPCR . . . . . G Protein-Coupled Receptors.

GPT . . . . . Generative Pre-Trained Transformer.

Grad-CAM . . Gradient-Weighted Class Activation Mapping.

GRN . . . . . Gene Regulatory Network.

gRNA . . . . . Guide RNA.

GRU . . . . . Gated Recurrent Unit.

GWAS . . . . . Genome-Wide Association Studies.

HCC . . . . . Hepatocellular Carcinoma.

HCoV . . . . . Human Respiratory Coronavirus.

HCoV-229E . Human Coronavirus 229E.

HCoV-HKU1 Human Coronavirus HKU1.

HCoV-NL63 . Human Coronavirus NL63.

HCoV-OC43 . Human Coronavirus OC43.

HE . . . . . Hemagglutinin-Esterase.

HFpEF . . . . Heart Failure with Preserved Ejection Fraction.

HGSOC . . . . High-Grade Serous Ovarian Cancer.

HGT . . . . . Horizontal Gene Transfer.

HIV-1 . . . . . Human Immunodeficiency Virus Type 1.

HNSCC . . . . Head and Neck Squamous Cell Carcinoma.

HR1 . . . . . Heptad Repeat 1.

HR2 . . . . . Heptad Repeat 2.

IBV . . . . . Avian Infectious Bronchitis Virus.

IC . . . . . Intracellular Region.

ICTV . . . . . International Committee on Taxonomy of Viruses.

IIR . . . . . Infinite Impulse Response.

ILC . . . . . Innate Lymphoid Cell.

IMP . . . . . Imipenemase Metallo-β-lactamase.

IN . . . . . Integrase.

IPC . . . . . Instituto Politécnico de Coimbra.

ISCIII . . . . . Instituto de Salud Carlos III.

IVF . . . . . In Vitro Fertilization.

KPC . . . . . Klebsiella Pneumoniae Carbapenemase.

LANL . . . . . Los Alamos National Laboratory.

LIME . . . . . Local Interpretable Model-Agnostic Explanations.

lncRNA . . . . Long Non-Coding RNA.

LRNV . . . . . Lucheng Rn Rat Coronavirus.

LSA . . . . . Language Sample Analysis.

LSTM . . . . . Long Short-Term Memory.

LTR . . . . . Long Terminal Repeat.

M . . . . . Membrane Glycoprotein.

m5U . . . . . 5-methyluridine.

MAFFT . . . . Multiple Alignment using Fast Fourier Transform.

MAG . . . . . Metagenome-Assembled Genomes.

MASE . . . . . Multiple Aligned Sequence Editor.

MCA . . . . . Middle Cerebral Artery.

MCC . . . . . Matthews Correlation Coefficient

MCoV . . . . . Mink Coronavirus I.

MDD . . . . . Major Depressive Disorder.

item[MDR] Multidrug-Resistant.

ME . . . . . Macular Edema.

MERS-CoV . Middle East Respiratory Syndrome Coronavirus.

MERSr-CoV . Middle East Respiratory Syndrome-Related Coronavirus.

MHV . . . . . Mouse Hepatitis Virus or Murine Hepatitis Virus.

MIC . . . . . Minimum Inhibitory Concentration.

miRNA . . . . MicroRNA.

ML . . . . . Maximum Likelihood.

ML . . . . . Machine Learning.

MLeV . . . . . Microhyla Letovirus 1.

MM . . . . . Multiple Mieloma.

mRNA . . . . Messenger RNA.

MRSA . . . . . Methicillin-Resistant *Staphylococcus Aureus*.

MPR . . . . . Multidrug Resistance-Associated Superfamilies.

MSA . . . . . Multiple Sequence Alignment.

MSAProbs . . Multiple Sequence Alignment Probabilistic Consistency-Based.

MSI . . . . . Microsatellite Instability.

MuCoV-HKU13 Munia Coronavirus HKU13.

MUCSI . . . . Master's Degree in Computing and Intelligent Systems (University of Deusto).

MUSCLE . . . MUltiple Sequence Comparison by Log-Expectation.

N . . . . . Nucleocapsid Phosphoprotein.

NBSLRR . . . Nucleotide-Binding Site-Leucine-Rich Repeat.

NCBI . . . . . National Center for Biotechnology Information.

ncRNA . . . . Non-Coding RNA.

NC-IUB . . . . Nomenclature Committee of the International Union of Biochemistry.

NDM . . . . . New Delhi Metallo-β-lactamase.

NGS . . . . . Next-Generation Sequencing.

NHCoV-HKU19 Night Heron Coronavirus HKU19.

NLP . . . . . Natural Language Processing.

NOWS . . . . Neonatal Opioid Withdrawal Syndrome.

NP-hard . . . Non-Deterministic Polynomial-Time Hard.

NSCLC . . . . Non-Small Cell Lung Cancer.

nsp . . . . . Non-Structural Protein.

NTD . . . . . N-Terminal Domain.

nts . . . . . Nucleotides.

OCT . . . . . Optical coherence tomography.

OOD . . . . . Out-Of-Distribution.

ORF . . . . . Open Reading Frame.

OSA . . . . . Obstructive Sleep Apnea.

Pangolin-CoV Pangolin Coronavirus.

PAUP . . . . . Phylogenetic Analysis Using Parsimony and Other Methods.

PAS . . . . . Polyadenylation Signal.

PBP . . . . . Penicillin-Binding Proteins.

PBP2a . . . . Penicillin-Binding Protein 2a.

PD-L1 . . . . . Programmed Death-Ligand 1.

PEDV . . . . . Porcine Epidemic Diarrehoea Virus.

pH . . . . . Potential of Hydrogen.

PHEV . . . . . Porcine Hemagglutinating Encephalomyelitis Virus.

PhyML . . . . Phylogenetic Estimation Using Maximum Likelihood.

PoCoV-HKU15 Coronavirus HKU15.

polyA . . . . . Polyadenylation Sites.

pp . . . . . Polyprotein.

PPT . . . . . Polypurine-Tract.

PR . . . . . Protease.

PRCoV . . . . Porcine Respiratory Coronavirus.

PRISMA . . . Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

ProbCons . . . Probabilistic Consistency-Based Multiple Sequence Alignment.

PRS . . . . . Polygenic Risk Score.

psd . . . . . Power Spectral Density.

PSG . . . . . Polysomnograms.

PTM . . . . . Post-Translational Modification.

R . . . . . Repeated.

RAxML . . . . Randomized Axelerated Maximum Likelihood.

RbCoV HKU14 Rabbit Coronavirus HKU14.

RBD . . . . . Receptor Binding Domain.

RBN . . . . . RNA-Binding Protein.

RdRp . . . . . RNA-Dependent RNA Polymerase.

RFE . . . . . Recursive Feature Elimination.

rG4 . . . . . RNA G-Quadruplex.

RNA . . . . . Ribonucleic Acid.

RNA-Seq . . . RNA sequencing.

RNN . . . . . Recurrent Neural Network.

ROC . . . . . Receiver Operating Characteristic.

RT . . . . . Reverse Transcriptase.

RtCoV-HKU24 China Rattus Coronavirus HKU24.

RT-PCR . . . Reverse Transcription Polymerase Chain Reaction.

S . . . . . Spike Protein or Surface Glycoprotein.

s2m . . . . . Stem-Loop II-Like Motif.

SARS-CoV . . Severe Acute Respiratory Syndrome Coronavirus.

SARS-CoV-2 . Severe Acute Respiratory Syndrome Coronavirus 2.

SAV . . . . . Single Amino Acid Variant.

SBDD . . . . . Structure-Based Drug Design.

scChIP-seq . . Single-Cell Chromatin ImmunoPrecipitation DNA-Sequencing.

SCCmec . . . Staphylococcal Chromosomal Cassette mec.

scMulti-omics Single-Cell Multi-omics.

scRNA-seq . . Single-Cell RNA Sequencing.

SD . . . . . Standard Deviation.

SD-OCT . . . Spectral-Domain Optical Coherence Tomography.

sgRNA . . . . Single-Guide RNA.

SHAP . . . . . Shapley Additive Explanations.

siRNA . . . . . Small Interfering RNA.

SMF . . . . . Serum Metabolic Fingerprint.

SMILES . . . . Simplified Molecular Input Line Entry System.

SN . . . . . Sensitivity.

SNP . . . . . Single Nucleotide Polymorphism.

SNV . . . . . Single Nucleotide Variant.

SP . . . . . Specificity.

SPECT . . . . Single Photon Emission Computed Tomography.

SPECT . . . . Spatial Transcriptomics.

STED . . . . . Stimulated Emission Depletion.

STFT . . . . . Short-Time Fourier Transform.

STI . . . . . Sexually Transmitted Infection.

SV . . . . . Structural Variant.

T . . . . . Thymine.

T3SS . . . . . Type III Secretion System.

T4SS . . . . . Type IV Secretion System.

TAR . . . . . Trans-Activation Responsive.

TAR . . . . . Tuberculosis.

T-Coffee . . . Tree-based Consistency Objective Function For alignmEnt Evaluation.

TCR . . . . . T-Cell Receptor.

TE . . . . . Transposable Element.

TF . . . . . Transcription Factor.

TFBS . . . . . Transcription Factor Binding Site.

TGEV . . . . . Transmissible Gastroenteritis Virus.

TILs . . . . . Tumor Infiltrating Lymphocytes.

TIME . . . . . Tumor Immune Microenvironment.

TIS . . . . . Translation Initiation Site.

TM . . . . . Transmembrane Domain.

TME . . . . . Tumor Microenvironment.

ToA . . . . . Theater of Activity.

TP . . . . . Transport Protein.

tRNA . . . . . Transfer RNA.

TSS . . . . . Transcription Start Site.

U . . . . . Uracile.

U3 . . . . . Unique 3'.

U5 . . . . . Unique 5'.

UNAIDS . . . United Nations Programme on HIV/AIDS.

Unespar . . . . Universidade Estadual do Paraná.

UPGMA . . . Unweighted Pair Group Method with Arithmetic Mean.

UPS . . . . . Ubiquitin-Proteasome System.

UPV/EHU . . Universidad del País Vasco / Euskal Herriko Unibertsitatea

URF . . . . . Unique Recombinant Forms.

UTR . . . . . Untranslated Region.

VAE . . . . . Variational Autoencoder.

VAF . . . . . Variant Allele Frequencies.

VGG . . . . . Visual Geometry Group.

VIM . . . . . Verona Integron-Encoded Metallo-β-lactamase.

VIS . . . . . Virus Integration Site.

VISTA . . . . V-Domain Ig Suppressor of T Cell Activation.

vsiRNA . . . . Virus-Derived Small Interfering RNA.

VOC . . . . . Variant Of Concern.

VRE . . . . . Vancomycin-Resistant Enterococci.

WECoV-HKU13 White-Eye Coronavirus HKU13.

WHO . . . . . World Health Organization.

WiCoV-HKU20 Wigeon Coronavirus HKU20.

XAI . . . . . Explainable Artificial Intelligence.

XDR . . . . . Extensively Drug-Resistant.

XXDR . . . . Extremely Drug-Resistant.

If I were not a physicist, I would probably be a musician. I often think in music. I live my daydreams in music. I see my life in terms of music.

— Albert Einstein, *one of the most influential scientists in history, a talented violinist, dedicated pianist, and a loyal admirer of Mozart and Bach.*

1 Introduction

Contents

1.1 May I share a brief story with you? 1
1.2 Why SARS-CoV-2 + FCoV? 5
1.3 Difficulties in Predicting Genetic Recombination Between Two Viruses with the Current State of the Art 9
1.4 Applicability Limitations in the Case Study of Genetic Recombination between SARS-CoV-2 and FCoV . . . 12
1.5 Redefinition of the Research Line 14
1.6 Reformulation of the Virus Case Study 14
1.7 Determination of the Specimen Virus for Genetic Recombination Study 15
1.8 Deep Learning Tools of Choice 15
1.9 Conversion of Genomic Sequences into Images 17
1.10 Interpretation of Results 19
1.11 Research Hypothesis 20
1.12 Research Objectives 20
1.13 Basic Methodology 21
1.14 Methodical Process 22
1.15 Final Hypothesis 25

1.1 May I share a brief story with you?

Probably this research started in an unorthodox way. It did not begin with exploring a field and ending up posing a research hypothesis. The start was directly the

hypothesis. At a dramatic moment, moreover. Let me tell you the story.

I live with several cats rescued from the street. People who live with rescued street cats usually deal with the typical pathologies of these animals. And we know their symptoms, consequences, and possible treatments quite well. Unfortunately, it is common for these animals to suffer from multiple viral infections, usually severe, that we have to identify and address.

In early 2020, I had a very intense cold. I do not recall ever having such a severe and persistent cold. Two to three days after my symptoms began, my cats started sneezing. They were a bit lethargic. It surprised me because I have never transmitted any type of illness to them, let alone a cold. I didn't think much of it.

In early March, my cats were perfectly healthy, without any sneezing. I still had some lingering effects from that intense cold, but nothing out of the ordinary.

On March 10, 2020, I was notified by my company: two colleagues had "COVID". I immediately self-isolated as a precaution. On March 14, 2020, the entire population was put under lockdown. Fortunately, everything ended well for everyone.

The days of lockdown passed. It was late March, and one day I heard on the news "cytokine storm". At that moment, they were discussing a possible explanation for pneumonia caused by Coronavirus Disease 2019 (COVID-19). It was not confirmed yet, but it seemed plausible that this cytokine storm was the reason behind that pneumonia.

At that moment, a clear association of ideas came to my mind like a spring: Feline Coronavirus (FCoV). In this coronavirus, the cytokine storm produces an absolutely lethal pneumonia in kittens and geriatric cats [1, 2]. In fact, it leads to irreversible multiorgan failure once the phenomenon begins, practically in 100% of cases [3, 4].

In a naive manner, I wrote a dossier to various authorities stating this circumstance and indicating that in cats, the treatment of choice is high doses of cortisone (immunosuppressive) to reduce the inflammatory process as much as possible. No one responded. Until I wrote to the National Microbiology Center - Instituto de Salud Carlos III (ISCIII), where they sent me a so kind thank-you email for the collaboration.

It was early April 2020. I remembered that cold in February, remembered those sneezes from my cats... And immediately, I raised that hypothesis from which everything starts: What if cats can contract COVID-19? What would happen if Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) interacts with FCoV? The vast majority of colony cats are carriers [5–7]. And if symptoms start, FCoV presents a mortality rate in kittens close to 100% [8]. If an equivalent occurred in humans, the scenario would be absolutely terrifying.

I began to research, and delving even further, I discovered that the feline coronavirus has already recombined, specifically with the canine coronavirus [9, 10]. And that there are multiple factors that could pose some type of zoonotic risk to humans (Section 1.2).

My next question was: What can I do? I am an engineer in electronics and automation. My professional life was focused on the electrical sector and on strategic management and internationalization. Very far from the field of virology. My only experience in this area was just caring for stray cats.

As I do not give up easily, I thought, analyzed, and finally, I put a first name to what I wanted to do: Predictive Modeling of the Emergence of New Variants of the SARS-CoV-2 Virus as a Result of its Genetic Recombination with the Feline Coronavirus (FCoV), using Artificial Intelligence.

I spoke with my dear professor and eternal mentor, Ana Macarulla, who supported and guided me. This hypothesis deserves a Ph.D. I contacted again the leading entity at the national level that kindly assisted me, Instituto de Salud Carlos III - National Center for Microbiology. I exposed my idea, and I was fortunate that they showed a high interest in the case study. That is how I met my dear and admired Isabel Olivares Zarco and Concha Casado Herrero, both of whom are leading principal researchers with such impressive careers, who offered me their support and co-supervision to start this doctorate. Even today, I can not believe how lucky I was... Incredible professionals and wonderful people.

With Ana Macarulla's guidance, I presented the case at D4k Research Team, where my dear Iker Pastor and Borja Sanz, professors in computing at the University

of Deusto - Faculty of Engineering, committed to the engineering part with me. Without knowing me at all, they were firm and dedicated. The second of my great strokes of luck.

As I needed a thesis supervisor in the microbiology field, I contacted my dear Lola Moragues Tosantos, a prominent principal researcher at the Faculty of Medicine and Nursing at the Universidad del País Vasco / Euskal Herriko Unibertsitatea (UPV/EHU). An admirable researcher and an unparalleled person and friend. Lola, without hesitation, accepted the offer unconditionally. My third stroke of great luck.

We are all here. I am with the best.

The tool of choice for this project was Artificial Intelligence, AI. At that time I called it that because I did not know how else to name it. Although at that time I had no idea what AI was, I had a strong intuition that it could be a very promising tool, as well as innovative.

If we are able to develop algorithms that accurately predict the evolution of existing viruses, perhaps we can anticipate a SARS-CoV-3. And pre-design vaccines (messenger RNA, this is, mRNA, etc.), treatments. And if it is not possible, at least we will be better prepared, with tools that allow us to minimize (even more) the development timelines of all kinds of weapons against whatever may come.

The case I presented was one that I considered to be of sufficient interest and relevance after analysis. But there are many viruses with zoonotic capabilities. Therefore, there are many developments and many lines of research to pursue.

Before starting my doctoral studies, I took advantage of the lockdown to train in two fields completely unrelated to my previous professional trajectory, which was mainly focused on management and business leadership.

I would like to think that I learned something. And I will be happy if this whole adventure can contribute even just a little bit.

In any case, I take away a great lesson from this experience. At 45 years old, with a professional background focused on management and business leadership, I have been able to immerse myself in the world of microbiology, virology, and computing and AI.

And I have come this far. And if I have been able to reach here, it means that anything is possible. We can all do it.

Moreover, I have the best possible team by my side.

1.2 Why SARS-CoV-2 + FCoV?

As I discuss in the following paragraphs, there are several indications that lead to considering genetic recombination between both viruses as an interesting case study. These indications are related to sufficient parental proximity, common aspects that make co-infection possible, and potential harmful consequences in the event of recombination between them.

The SARS-CoV-2 is a virus belonging to the Coronaviridae family (coronaviruses), Betacoronavirus genus, and Sarbecovirus subgenus, responsible for the pandemic called Coronavirus Disease 2019 (COVID-19) [11]. The COVID-19 pandemic had its first signs in late 2019 in the city of Wuhan, China, when cases of a mysterious pneumonia were reported. On December 31, 2019, Chinese authorities informed the World Health Organization (WHO) about several cases of an unknown respiratory illness. In January 2020, Chinese scientists isolated and sequenced the virus, identifying it as a new coronavirus, named SARS-CoV-2. The rapid global spread of the virus led to the declaration of COVID-19 as a pandemic on March 11, 2020, marking a milestone in the history of modern public health [12, 13]. Multiple research studies indicated that SARS-CoV-2 has a zoonotic origin, meaning it was transmitted from an animal host to humans [14–16]. As of March 2023, 1,541,293 complete genomes of SARS-CoV-2 were identified [17], and globally, 761.4 million cases were diagnosed, with a total number of deaths of 6.9 million people [18].

The Feline Coronavirus (FCoV) is a virus belonging to the Coronaviridae family (coronaviruses), Alphacoronavirus genus, and Tegacovirus subgenus [19]. The FCoV is classified into two serotypes: serotype I, which causes the majority of natural infections in Europe and America (80-95%), and serotype II, described only in Asia, likely resulting from a double homologous recombination of serotype I FCoV and Canine Coronavirus (CCoV) [20]. The FCoV, in either of its two serotypes,

mainly affects kittens and geriatric cats, with a mortality rate close to 100% when the cytokine storm process is triggered [3].

Both SARS-CoV-2 and FCoV are single-stranded RNA viruses with positive polarity and an envelope [20, 21].

There is strong evidence of the ability of coronaviruses to jump from one species to another. We have recent examples of Severe Acute Respiratory Syndrome Coronavirus, SARS-CoV (first detected in 2002 in Guangdong, China) and Middle East Respiratory Syndrome Coronavirus, MERS-CoV (first detected in 2012 in Saudi Arabia). In both cases, the origin was traced back to a bat coronavirus that jumps to humans using an intermediate host as an amplifier (possibly civets in the case of SARS-CoV and camels in the case of MERS-CoV) [22].

In November 2020, a variant of SARS-CoV-2 originating from farmed minks was detected in Denmark, with four mutations concentrated in the S protein [23]. Likewise, it was evidenced that cats and ferrets have a high susceptibility to SARS-CoV-2 infection [24].

On April 20, 2021, the results of an investigation were published, suggesting the possible animal origin of the B.1.1.7 lineage of SARS-CoV-2 (commonly known as the Alpha variant), probably from dogs. The research aimed to locate the predecessor of this variant, considering not only genomes identified in humans but also in dogs, minks, tigers, and cats. The assumption of these researchers was that these viruses, so similar to B.1.1.7, could have evolved through transmission among animals before passing back to humans. And the closest genome (early specimen of this variant) was precisely located in a dog [25].

SARS-CoV-2 has recombined. There is even evidence of community transmission of recombinants. The XE variant is precisely a recombinant of the Omicron variant, lineages BA.1 and BA.2, with most of its genome (including the spike protein) belonging to BA.2 [26].

Based on the previous considerations and given the ease demonstrated by SARS-CoV-2 to perform interspecies jumps [27], it seems logical to consider the possibility that both new variants and even new coronaviruses could in the future affect

different animal species, and more specifically humans. One possible mechanism for this is genetic recombination between coronaviruses from different animal species. Considering that cats are highly susceptible to SARS-CoV-2 infection, it is interesting to consider the possibility of genetic recombination between this virus and FCoV if they infect the same animal simultaneously [28].

Delving further into the aforementioned:

- Genetic recombination is a common phenomenon in the coronavirus genus, leading to the emergence of new strains or new coronaviruses (SARS-CoV, MERS-CoV, or SARS-CoV-2 itself) [29, 30].

- The serotype II of FCoV is, in fact, the product of a double homologous recombination between serotype I of FCoV and Canine Coronavirus (CCoV). Therefore, there is evidence of previous recombinations involving FCoV [9].

- SARS-CoV-2 has a strong affinity with the receptor called Angiotensin-Converting Enzyme 2 (ACE2), which represents the main pathway for the virus to enter the interior of the cells it infects [31]. After chimpanzees and macaques, the cat is the species whose ACE2 is most similar to humans (at 85.2%) [32]. This degree of similarity partly explains the cat's susceptibility to infection and its potential transmission to other species, including humans, given the strong interaction between both species. See the identity percentages between the ACE2 of different animal species and the human ACE2 in Table 1.1.

Table 1.1: Percentage of similarity with human (*Homo sapiens*) ACE2, in % [32].

Species	**Similarity (%)**
Pan troglodytes (chimpanzee)	99.0
Macaca mulatta (macaque)	94.9
Felix catus (cat)	85.2
Manis javanica (malayan pangolin)	84.8
Oryctolagus cuniculus (rabbit)	84.8
Nyctereutes procyonoides (racoon dog)	83.9

One of the fundamental conditions for genetic recombination to occur between two pathogens is that both infect the same cells. This condition is highly probable in the case of SARS-CoV-2 and FCoV since both viruses have been detected to

be expressed in enterocytes, that is, the digestive epithelial cells (cells lining the intestine) [33, 34].

Multiple technological tools were developed in the last 20 years for detecting genetic recombination events [35, 36].

Certain Artificial Intelligence (AI) algorithms have proven to be a powerful tool in the healthcare field, particularly in the fight against SARS-CoV-2 [37]. The application of such algorithms has been mainly aimed at providing diagnostic tools and, on a predictive level, determining risk groups, the evolution of the disease itself based on certain factors, etc. Detailed information is available at Chapter 3.

However, the application of AI is interesting not only in predictive methodologies for diagnosis, evolution, and detection of risk groups, but also in predicting variations, strains, and even future zoonoses. In this specific case, the generation of predictive genomes resulting from genomic recombination between SARS-CoV-2 and FCoV. Knowing these "future" genomes as much as possible will also make it feasible to pre-design vaccines, treatments, etc. Thus, the potential of AI in this area of predictive medicine is enormous and of high added value.

Consequently, this research initially set two objectives:

1. To mathematically anticipate possible recombinations derived from the coinfection of FCoV and SARS-CoV-2 in cats. These could give rise to new variants of SARS-CoV-2 with implications for pathogenesis, infectivity, or resistance to treatments or vaccines.

2. As a generalization of the previous objective, the application of a Deep Learning algorithm for the comparability and genomic modeling of possible interactions between two pathogens. In a first stage, mathematical anticipation of recombinations derived from coinfection between two RNA viruses.

1.3 Difficulties in Predicting Genetic Recombination Between Two Viruses with the Current State of the Art

Predicting genetic recombination using Deep Learning techniques presents several complex challenges with the current state of the art.

First, the genome is a vast and complex instruction manual, of which we have only read and understood a few pages. Our current level of understanding is reduced to merely the tip of the iceberg.

Most of the genome of any organism remains largely unknown for several reasons, among which the following stand out:

1. Unidentified Genes and Functions. We have yet to identify all the genes present in the genome, nor do we fully understand the functions of many of the genes we have identified. Much of the genome is composed of non-coding regions, whose function and biological relevance are scarcely known.

2. Complex Genetic Regulation. Gene expression regulation is extremely complex and not fully understood. Epigenetic mechanisms, transcription factors, and other forms of genetic regulation act together in ways we are just beginning to decipher.

3. Repetitive Elements. The genome contains a large number of repetitive sequences and transposons, whose function and impact on genome evolution and functioning are not fully clear. These sequences can pose significant challenges in genomic analysis.

4. Genomic Interactions. Interactions between different parts of the genome and between the genome and other cellular components are complex and not fully mapped. Gene and protein interaction networks are vast and challenging to unravel.

5. Genetic Variation. Genetic variation among individuals and populations is immense, and its impact on health, diseases, and other phenotypic characteristics is only partially understood. Many genetic variants have not yet been correlated with specific phenotypes.

6. Genomes of Other Organisms. While the human genome has been the focus of much research, the genomes of many other organisms are even less understood. Genetic diversity in nature is immense, and each organism presents a unique set of challenges for sequencing and analysis.

7. Tools and Methods. Current tools and methods for genomic analysis, while powerful, have limitations. Technology continues to evolve, and new techniques are constantly being developed to address questions that were previously inaccessible.

With such limited understanding of the genome, its functions, and the implications of any modification therein, the difficulty of predicting genomic sequences resulting from viral evolution is evident.

Considering genetic recombination as a complex process of viral evolution, wherein two genomes merge, generating a viable recombinant mixture of their parental viruses, several specific difficulties arise:

1. Biological and Genetic Complexity.

- Limited Biological Knowledge. Recombination hotspots (regions where recombination is more frequent) are not fully understood. Although some have been identified, their behavior is not completely predictable.

- Stochastic Nature of Recombination. Genetic recombination is a highly stochastic and variable process, making exact prediction difficult. This process can occur at multiple points in the genome and is influenced by various genetic and environmental factors.

- Genetic Diversity. Genetic variability among virus families, different viruses, and among variants and lineages within the same viruses adds additional layers of complexity. Models must be generalizable and adaptable to different genetic contexts.

2. Data Limitations.

- Data Availability. Data on recombination events are limited and, in many cases, difficult to obtain. Recombination studies require detailed sequencing and tracking of successive generations.

- Data Quality and Noise. Genetic data may contain errors and noise, making it challenging to train accurate models. Genomic sequences may be incomplete or contain sequencing errors.

- Data Size. Although large amounts of genomic data are generated, specific and well-annotated recombination events are not as abundant, limiting the models' ability to learn relevant patterns.

3. Modeling and Neural Network Architecture.

- Designing Suitable Architectures. Deep Learning architectures must be specifically designed to capture the complexities of genetic recombination. This requires deep knowledge of both biology and machine learning.

- Explainability. Deep neural networks are often seen as black boxes, making it difficult to interpret how decisions are made and whether models are learning biologically relevant patterns.

- Generalization. Ensuring that models not only perform well on specific training datasets but also generalize well to unseen data is a considerable challenge.

4. Computing and Resources.

- Computational Requirements. Training Deep Learning models on genomic data requires a large amount of computational resources due to the size and complexity of the data. This includes memory, processing power, and storage.

- Parameter Optimization. Tuning and optimizing the hyperparameters of Deep Learning models to achieve the best performance is a complex and time-consuming task.

5. Integration of Biological Knowledge.

- Incorporation of Prior Knowledge. Integrating previous biological knowledge about recombination into Deep Learning models can enhance their performance, but finding the right way to do so is a challenge.

- Hybrid Models. Developing models that combine traditional computational biology approaches with Deep Learning techniques can be a solution, but this requires innovation and experimentation in model design.

6. Evolution and Adaptation.

- Continuous Evolution. Recombination patterns can change over time due to evolution, meaning that models must be continually updated and adapted to new data.

- Genetic Plasticity. Genome plasticity, including mutations and structural variations, adds an additional layer of complexity to recombination prediction.

Despite these difficulties practically rendering reliable Deep Learning development for effectively predicting genetic recombination between two viruses nearly impossible, these very challenges motivate the need to advance research in this field in order to develop new technologies and analysis methods.

1.4 Applicability Limitations in the Case Study of Genetic Recombination between SARS-CoV-2 and FCoV

The effective application of Deep Learning tools requires [38]:

- Using a dataset composed of THOUSANDS OF DATA points to learn complex patterns.

- Ensuring that they are sufficiently balanced.

- Being representative in the application domain. In other words, they should represent a sufficiently large population sample.

- Adequately including all scenarios. To avoid biases, it is necessary to consider even scenarios that are not applicable to the study, but exist.

Given these basic requirements, the first major obstacle in my research was determined by the notorious scarcity of data regarding FCoV. As of March 2023, the total number of complete FCoV sequences identified was only 72. I downloaded them from the European Nucleotide Archive (ENA) public repository [39]. Additionally, I downloaded the total number of complete sequences of SARS-CoV-2 (1,541,293) available as of that date from the public repository National Center for Biotechnology Information, Virus Database (NCBI Virus Database) [40].

Therefore, I discovered several factors that penalized the initially proposed case study. The limited sequencing of FCoV results in insufficient data for the

implementation of a robust model. Additionally, considering that FCoV was first isolated in the 1970s, we lack representative data on its evolution, and we likely do not know all possible variants. The imbalance with SARS-CoV-2 is extremely large, with the consequent risk of bias or model fragility. Given that FCoV is the priority case study, as its mortality rates are higher than those of SARS-CoV-2, the fact that we have so little information adds complexity, especially in terms of validating the model results.

Regarding the quantity of registered sequences, the case of SARS-CoV-2 is completely anomalous (unfortunately) in terms of the number of complete genomic sequences recorded in public repositories. As seen in Figure 1.1, as of March 2023, the NCBI repository registered up to 1,541,293 complete sequences of SARS-CoV-2. However, the quantity of sequences for other viruses, in general terms, is considerably lower and incomplete. We only had 13,554 sequences of HIV-1 and 72 sequences of FCoV. The entire compilation of remaining coronaviruses (Alpha, Beta, Delta, and Gamma Coronaviruses) encompassed only 6,486 sequences.

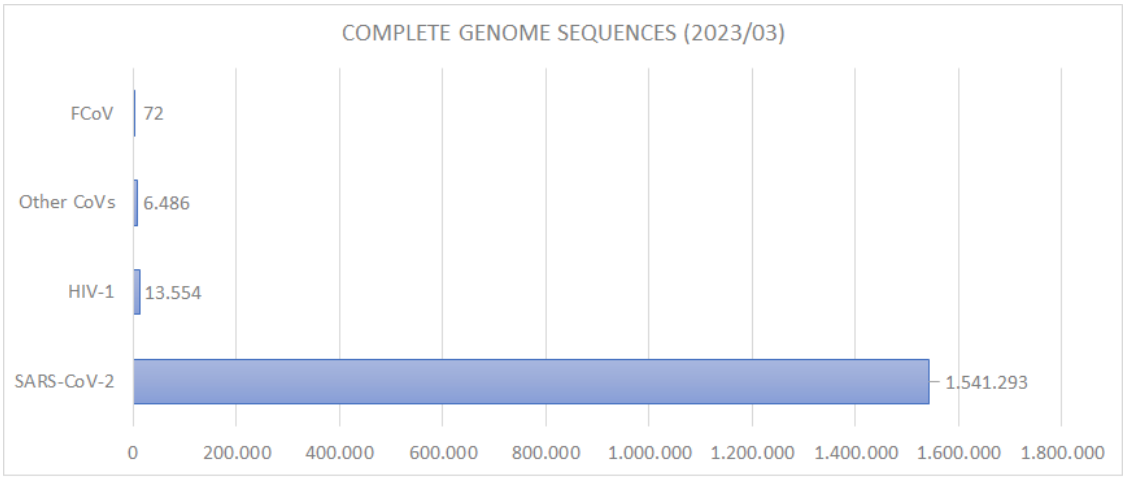


Figure 1.1: Number of Complete Sequences of Various Viruses as of March 2023. I downloaded the sequences of SARS-CoV-2 from the public repository NCBI Virus Database [40]. I downloaded the sequences of HIV-1 from the public repository of the Los Alamos National Laboratory HIV Database [41]. I downloaded the rest of the sequences from the public repository European Nucleotide Archive (ENA) [39].

Therefore, given the current sequencing status of FCoV, the robust application of Deep Learning tools is unfeasible, especially considering the extreme imbalance

between FCoV and SARS-CoV-2.

None of the current techniques for addressing sequence imbalance is efficiently applicable to such a degree of imbalance and such a limited number of FCoV sequences.

1.5 Redefinition of the Research Line

The limitations outlined in Sections 1.3 and 1.4 mean that, on the one hand, the potential results of any Deep Learning application aimed at predicting genetic recombination between two viruses contain an unacceptable level of uncertainty. On the other hand, the insurmountable imbalance between SARS-CoV-2 and FCoV makes the efficient application of Deep Learning tools unfeasible with minimal scientific and methodological guarantees.

Therefore, given the existing conditions, I considered a revision of the research approach necessary.

I decided to focus my research on applying Deep Learning technologies to increase our understanding of the genome, especially regarding the mechanisms that facilitate genetic recombination between two viruses.

1.6 Reformulation of the Virus Case Study

Once the decision was made to redirect the research towards a deeper understanding of genetic recombination, in such a way that the minimum requirements necessary for an effective application of Deep Learning tools are met, the next step was to redefine the case study on SARS-CoV-2.

However, genetic recombination phenomena in SARS-CoV-2 have been limited. Of the 1,541,293 complete sequences downloaded, only 1,565 belonged to recombinant subtypes. Therefore, I still had an imbalance between recombinant sequences and non-recombinant sequences.

Following a principle of prudence, I considered it appropriate to initially train the Deep Learning tools with data related to a well-documented case of genetic recombination, with sufficient and as balanced as possible data, to then extrapolate

that learning to the case of SARS-CoV-2 and minimize the imbalance between categories detected in it as much as possible.

1.7 Determination of the Specimen Virus for Genetic Recombination Study

According to the conclusion obtained in Section 1.6, first, I needed a significant specimen regarding genetic recombination phenomena. This is the case of the Human Immunodeficiency Virus Type 1 (HIV-1). HIV-1 caused one of the worst pandemics in recent history, the Acquired Immunodeficiency Syndrome (AIDS) [42]. Therefore, we currently have a significant number of fully sequenced genomes, even though this pandemic began in the 1980s, when the degree of development of sequencing techniques was still limited [43]. The first cases of what would later be identified as AIDS were reported in the United States of America in 1981 [44].

Furthermore, HIV-1 is characterized by high genetic variability due to the error-prone nature of its reverse transcriptase enzyme, leading to frequent errors, mutations, frameshifts, and subsequent recombination events within its genome [45]. Therefore, we have many perfectly documented cases of genetic recombination, making it an excellent candidate to start the knowledge exploration about genetic recombination.

The next step was to analyze genetic recombination in SARS-CoV-2, given its high number of sequences. As explained in Chapter 5, I trained a CNN (previously trained on HIV-1) with SARS-CoV-2 sequences. I assessed its performance and, with explainability tools, determine the relevant hot zones in the recombinant feature.

1.8 Deep Learning Tools of Choice

To deepen our understanding of genetic recombination, I considered a fundamental first step to determine if there are patterns in the genome that trigger the recombination phenomenon or identify a sequence as recombinant.

In the field of Deep Learning, Convolutional Neural Networks (CNNs) emerged as formidable tools for pattern detection, especially in the field of computer vision, where they have demonstrated exceptional robustness and effectiveness.

CNNs have excelled in automatically extracting relevant features from images, being successfully employed in tasks ranging from facial recognition to object classification in photographs. This level of performance has been attributed to the inherent ability of CNNs to learn hierarchies of features, detecting complex and discriminative patterns [46].

Genomic sequences exhibit a linear structure representing the arrangement of nucleotides in Deoxyribonucleic Acid (DNA). Converting such genomic sequences into an image format would enable the direct application of established CNN architectures, leveraging their ability to learn local and global patterns. Additionally, this visual representation could facilitate the interpretation of results and the identification of key regions in the genome related to specific characteristics or phenotypes. In my case, the characterization of a sequence as recombinant.

Therefore, I considered CNNs to be the most reliable tool of choice, and within them, pretrained CNNs, which already have prior knowledge from extensive training with all kinds of images that we can leverage. To apply them, I needed to convert genomic sequences into an image format.

Some advantages of using pretrained CNNs are as follows [47]:

1. Pretrained on Large Datasets. Pretrained CNNs are initially trained on large datasets such as ImageNet, which contain millions of images spanning a wide variety of categories.

2. Extraction of General Features. During training on massive datasets, these networks learn to extract general and hierarchical features from images, making them experts in identifying complex visual patterns.

3. Knowledge Transfer. The idea behind pretrained CNNs is that the knowledge acquired from learning visual representations on a large dataset can be transferred to smaller or different specific tasks.

To apply the knowledge from these pre-trained convolutional neural networks (CNNs) to new domains, a technique called Transfer Learning is used. This technique involves taking a model that has been previously trained on a large dataset and transferring its knowledge to solve a different, yet related problem. Instead of training a model from scratch, Transfer Learning allows for adapting an already trained model to new, specific tasks, which can save time and improve performance, especially when there is limited data available for the new task.

For optimal utilization of CNN capabilities, the use of pretrained CNNs and Transfer Learning are key concepts in the field of Deep Learning. The Transfer Learning technique allows [48]:

1. Application in Specific Tasks. Transfer Learning involves taking a pretrained CNN and adapting it to a specific task or a smaller dataset related to the task of interest.

2. Pretrained Layers and Custom Layers. Pretrained layers of the network can be used, and only the final layers can be retrained for the specific task, or custom layers can be added to the existing architecture to fit the requirements of the new problem.

Transfer Learning offers significant advantages, such as the ability to train effective models with smaller datasets, as it leverages previously learned representations. This is especially useful when facing data limitations.

It also reduces the time and resources required to train a model from scratch since the network has already learned useful features on large datasets.

Therefore, pretrained CNNs and Transfer Learning allow capitalizing on the knowledge acquired from large datasets and accelerate the development of accurate models for specific tasks, even when limited datasets are available.

1.9 Conversion of Genomic Sequences into Images

To apply pretrained CNNs and fully harness their pattern detection potential in this new domain, it is imperative to convert genomic sequences into image format.

Transforming genomes into images poses its own challenges, such as selecting optimal visual representations and efficiently managing genomic information in a two-dimensional format.

Analytically, I needed to preserve as much information as possible in this transformation to work securely. To this end, I required a representation with minimal filtering and processing to minimize information loss and retain seemingly secondary patterns that may be relevant and/or of unknown importance. Additionally, the image representation should be manageable in terms of size and sufficiently sharp, allowing for work with complete genomes with reasonable computational load.

From a purely deductive standpoint, I must avoid image conversions based on random decisions. For instance, a mere conversion of each nucleotide type to a color. The resulting image will be heavily influenced by arbitrary decisions such as whether I want a square or rectangular image, etc. This could lead to a rearrangement of genomic elements that could plausibly alter the biological coherence of the genome.

However, I wanted to conduct a series of tentative tests with square and rectangular images of different sizes and apply transfer learning to several pretrained CNNs. In all cases, Training, Validation, and Test Accuracy indicated that the CNNs essentially classified the images randomly. This reaffirmed my deductive process.

Therefore, I require a genome-to-image transformation that is biologically coherent. My transformation of choice was the conversion of genomic sequences into spectrograms.

Genomic spectrograms are visual representations derived from the mathematical decomposition of genetic sequences in the frequency domain. This transformation converts nucleotide sequences into images that encapsulate relevant frequency information, allowing for detailed analysis of their intrinsic patterns. Since these images have an inherent mathematical basis, results obtained through explainability techniques in Deep Learning models provide insights into the underlying mathematical reasons for genome classification and characterization. The identifiability of regions in the spectrograms that are crucial for model predictions reinforces the understanding that these areas reflect specific mathematical characteristics of

genomic sequences. Consequently, explainability not only offers a clearer biological insight but also guides us towards a profound mathematical understanding of recombination and genomic variability processes.

Obtaining positive and coherent results will signify the mathematical nature related to nucleotide periodicity in the genome. If this holds true in the case of genetic recombination, the existence of mathematical patterns related to other viral characteristics would not be dismissed.

1.10 Interpretation of Results

Given that this is a multidisciplinary research at the intersection of microbiology/virology and computer science, we cannot solely rely on validating the results based on metrics that evaluate the performance of the CNN.

My objective goes beyond mere computational ratio verification. My aim is to identify if there are inherent elements in the genome that are crucial for detecting genetic recombination.

Given the multidisciplinary nature of my research, I sought to corroborate these potential findings with scientific literature in the fields of microbiology, virology, and genomics.

Therefore, I needed means to identify which areas of the genomic spectrogram are crucial for the CNN to classify a sequence as recombinant or non-recombinant. Subsequently, I must verify if these areas correspond to scientific evidence in the field. I do believe that this biology-centric explainability is the best validation.

Explainability tools allow to locate where the CNN looks to make a decision. So, these tools allow us to identify where the CNN is looking to classify an image. We do not know what the CNN sees, but we know where it looks. Therefore, the use of these tools is very powerful for extracting information and interpreting the results, based on expert literature.

1.11 Research Hypothesis

The application of advanced Deep Learning techniques and mathematical analysis to genomic spectrograms of viruses will enable not only the effective characterization of viral genomes, but also the identification of encrypted mathematical signatures. These signatures can facilitate the understanding of viral genotypic and phylogenetic relationships, providing a solid foundation for the design of generative architectures aimed at predicting genomic evolution in viruses.

1.12 Research Objectives

I. Evaluate genomic spectrograms as a method for representing genomic sequences.

II. Develop and evaluate Convolutional Neural Network (CNN) models for the categorization of genomic sequences.

III. Develop and evaluate explainable models to detect hidden mathematical signatures in the genome associated with the genotypic characterization of a sequence.

IV. Develop and evaluate a methodology for the identification of mechanisms associated with genetic recombination from a computational and bio-mathematical perspective.

V. Provide descriptive information on encrypted mathematical signatures in the genome associated with the genotypic characterization of sequences.

VI. Validate the developed methodology (genomic spectrograms + CNN + explainability) in applied cases for HIV-1 and SARS-CoV-2 viruses.

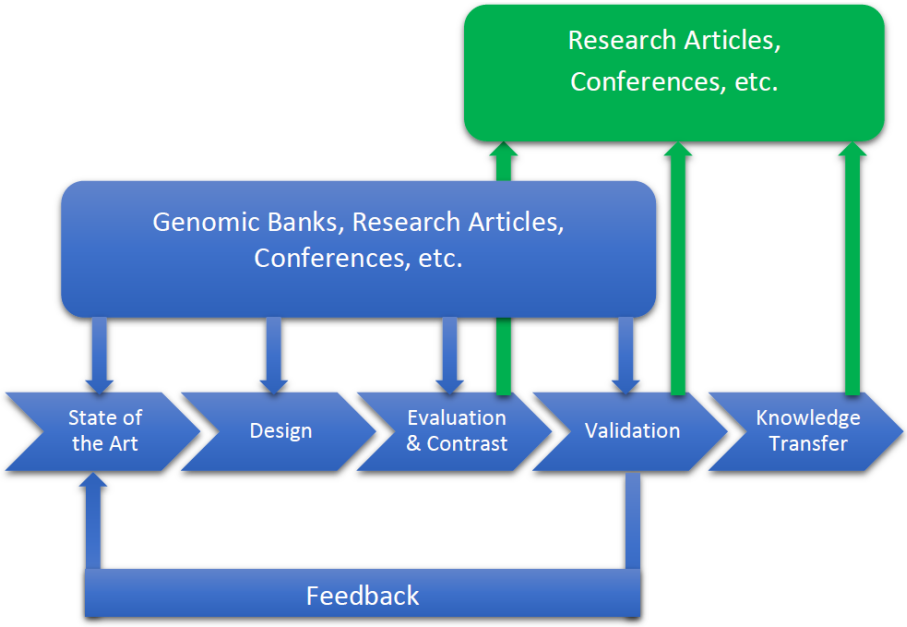


Figure 1.2: Basic Methodology

1.13 Basic Methodology

The basic methodology I followed throughout the experimentation is depicted in Figure 1.2.

Given that this research merges two scientific branches with exponential growth, especially considering the current circumstances, an agile methodology and constant updating with the state of the art are necessary. A continuous and permanent review and feedback cycle is essential to adapt the project as fast as possible to new advances that may occur in either of the two branches that make up this research.

1. Knowledge Update.

- Continuous research on new technological developments, emerging solutions, new genomic sequences, mutations, emerging strains, and recombination events.

- Gathering information from various sources, including databases, public genomic repositories, scientific literature, advanced courses, and participation in relevant conferences and symposiums.

2. Design of Technological Solution and Experimental Strategy.

- Selection of technologies, architectures, and algorithms.

- Design and planning of experimental test benches considering technical limitations.

3. Evaluation and Contrast.

- Systematic assessment of performance ratios.

- Systematic evaluation of the coherence level of the results obtained.

4. Bio-centric Validation.

- Comprehensive interpretation of the results obtained to determine the viability of the findings.

- Detailed characterization of the resulting genome to understand its biological significance and relevance.

5. Knowledge Transfer to the Scientific Community and Society.

- Communication of findings through publications in scientific journals, presentations at conferences, and other scientific events.

- Dissemination of research results to society at large through the publication of accessible summaries, informative talks, and awareness-raising activities.

6. Feedback and Continuous Update.

- Collection of feedback and input from the scientific community and other stakeholders.

- Adaptation and refinement of the approach based on new data and advances in the field, ensuring continuous improvement of knowledge and methodology employed.

1.14 Methodical Process

Figure 1.3 represents the progression of my research graphically and schematically.

Once the initial research hypothesis was formulated and a preliminary exploration conducted, as described in this Chapter, I analyzed in detail a series of microbiological considerations (Chapter 2) and the technological background of application (Chapter 3). With all this information, I conducted an analysis, established the design of the proposed methodology and its corresponding techniques (genomic spectrograms + CNN + explainability).

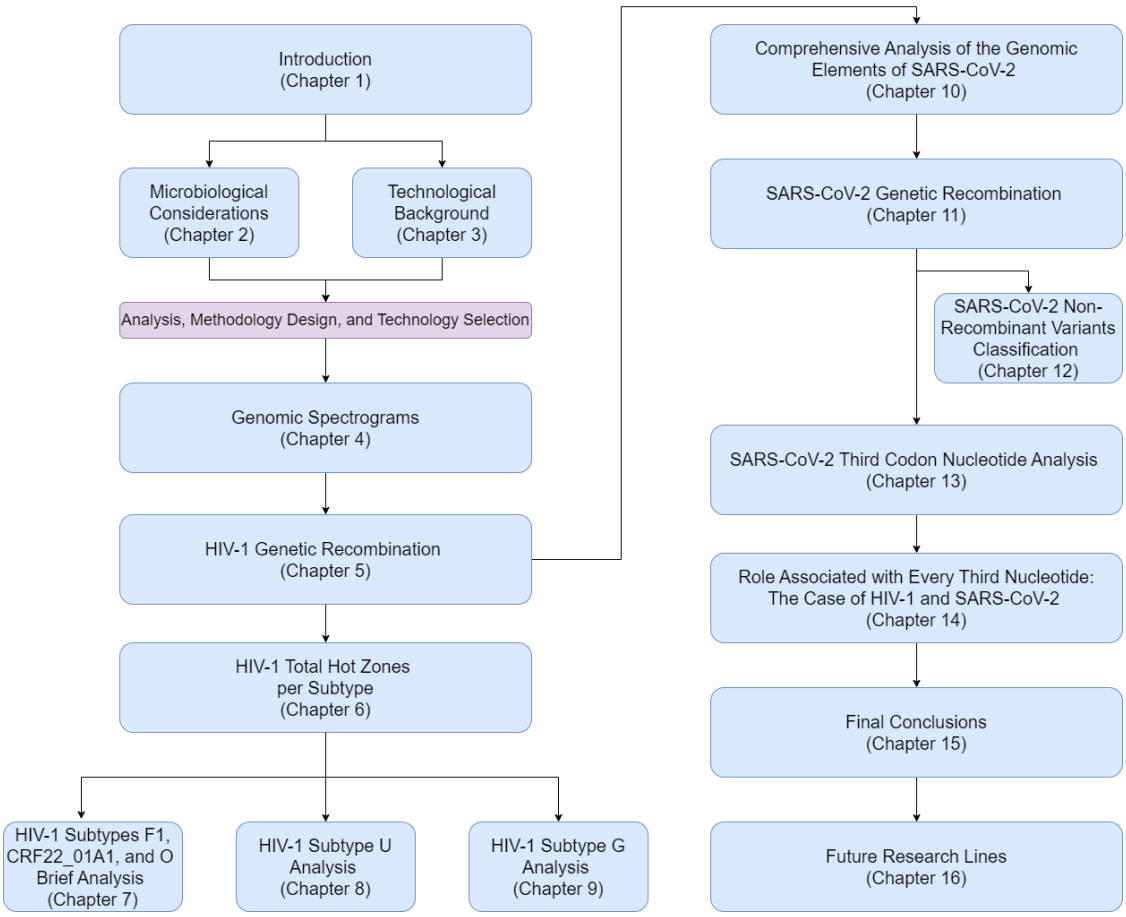


Figure 1.3: Research Progression

The next step was the design of spectrographic representations of genomic sequences in two complementary modalities: concatenated and superposed genomic spectrograms (Chapter 4).

Then, I applied the proposed methodology to the specimen virus for genetic recombination: HIV-1. In Chapter 5, I applied this methodology to the classification of complete HIV-1 sequences as recombinant or non-recombinant and conducted a thorough analysis to determine the degree of homology between the explainability results and the phylogenetic proximity. To this end, I generated the phylogenetic tree of all sequences composing the HIV-1 Complete Consensus Multiple Sequence Alignment from the Los Alamos National Laboratory HIV-1 Database (until 2020).

In light of the results obtained, I opened an additional line of research where I delved deeper into the understanding of the main recombinant and non-recombinant subtypes (Chapter 6).

From the results of this subtype analysis, I detected several subtypes with particular characteristics in my findings. These were subtypes U, G, F1, CRF22_01A1, and O (outlier).

I conducted an analysis of the hot zones of the sequences belonging to subtypes F1, CRF22_01A1, and O, establishing a comparison of their explainability results with those corresponding to each subtype and its respective category. The results of this analysis are detailed in Chapter 7.

The results obtained for Subtypes U and G closely aligned with the microbiological and phylogenetic reality. Therefore, I decided to further investigate these specific subtypes. The first one was denominated U (unknown). The results of my investigation are detailed in Chapter 8. And the second one was subtype G, with a possible recombinant origin according to a large part of the expert HIV-1 scientific community. The results of my analysis are detailed in Chapter 9. Therefore, this additional research is reflected in Chapters 6, 7, 8, and 9.

With the degree of knowledge already acquired and prior to moving on to the study of genetic recombination in SARS-CoV-2, I wanted to analyze in detail the degree of conservation of each genomic region of SARS-CoV-2 from a computational and mathematical perspective. This analysis is detailed in Chapter 10.

Subsequently, once the knowledge extracted by the CNN regarding genetic recombination in HIV-1 trained with genomic spectrograms was obtained, I replicated the experiment with that acquired knowledge in SARS-CoV-2 (Chapter 11).

In light of the results obtained, I wanted to open a new additional line of research, analyzing the capability of my methodology in classifying complete SARS-CoV-2 sequences into the four main variants of SARS-CoV-2: pre-VOC, Alpha, Delta, and Omicron. This additional line is detailed in Chapter 12.

Returning again to the main research line, and having obtained a striking result regarding the detection of the recombinant feature in the S protein of SARS-CoV-2, I wanted to deepen the bio-mathematical knowledge of those results (Chapter 13). To this end, I conducted several computational analyses on the S protein

that revealed specific patterns in this protein regarding its third codon nucleotides in the analyzed study groups.

Throughout Chapters 5 - 13, the periodicities of 3 and its multiples were a constant, playing a relevant role in genetic recombination in HIV-1, genetic recombination in SARS-CoV-2, and the classification of SARS-CoV-2 sequences into variants. In Chapter 14, I compiled the main indications of my research related to the third codon positions.

This document concludes with the final conclusions drawn from all the research conducted (Chapter 15), as well as some future research lines derived from it (Chapter 16).

1.15 Final Hypothesis

The entire course of the research, all the results, and all the clues found inevitably led me to conclude this document with an exciting hypothesis:

The third codon position contains encrypted information that influences the classification of certain viral characteristics. Furthermore, I raise the possibility that this role of nucleotides in the third codon position may be common in different microorganisms, suggesting a degree of universality that needs to be further evaluated.

As for the origin of this whole adventure (remember?), perhaps today we have not been able to anticipate the outcome of genetic recombination between SARS-CoV-2 and FCoV to prepare for a future pandemic resulting from a similar event. But I firmly believe that with all of our contributions, we are closer to achieving it.

It may not be today, but it will be.

I'm just a link in a chain made up of every scientist who published something that helped us keep moving forward.

— Katalin Karikó, *savior of millions of lives thanks to her discoveries that enabled the development of effective mRNA vaccines.*

2 Microbiological Considerations

Contents

2.1 Introduction 27
2.2 Genome. Definition and Concepts. 28
2.3 Viral Diversification Mechanisms 34
2.3.1 Mutation 34
2.3.2 Genetic Recombination 36
2.4 Multiple Sequence Alignment 38
2.5 Phylogeny. Definition and Concepts 41

2.1 Introduction

The primary purpose of this Chapter is to provide a strong foundation of microbiological and virological knowledge, which is crucial for a comprehensive and smooth understanding of the content of this doctoral thesis. Although this research spans various disciplines (computer science, microbiology, virology, and applied mathematics), the fundamental basis of this project is of an engineering nature. Therefore, it is essential that, as a reader, you possess a solid understanding of certain essential concepts that will be recurring throughout this document.

Section 2.2 provides the definition of genome and some related concepts. In Section 2.3 I explain some of the main mechanisms of viral diversification. Section

2.4 covers the definition and related concepts of Multiple Sequence Alignment, as well as some of its main algorithms. Finally, in Section 2.5, I present the definition and related concepts of phylogeny, as well as the main algorithms for generating phylogenetic trees.

2.2 Genome. Definition and Concepts.

The genome is a comprehensive compendium of genetic instructions encoded in Deoxyribonucleic Acid (DNA). It constitutes the fundamental instructions that govern the development, functioning, and inheritance of an organism [49]. In simple terms, the genome can be understood as the manufacturing manual of an organism.

The genome consists primarily of genes, which are specific sequences of DNA capable of encoding proteins or Ribonucleic Acid (RNA) molecules. Proteins play a vital role in virtually all biological processes, and the genome contains the precise instructions for their synthesis. In addition to genes, the genome also includes non-coding regions that are essential for the regulation of gene expression and other cellular processes [50]. In the following paragraphs, I will delve into the different types of genomic regions.

Nucleotides are the basic components of DNA and RNA molecules, playing a crucial role in storing and transmitting genetic information. They consist of three main components [51]:

i) A phosphate group. It is a chemical group containing phosphorus and oxygen. The phosphate group is located on the outer part of the nucleotide and connects with other nucleotides through phosphodiester bonds to form nucleic acid chains.

ii) A sugar. In the case of DNA, the sugar is deoxyribose, while in RNA, it is ribose. This sugar is located in the center of the nucleotide and forms the backbone of the nucleic acid chain.

iii) A nitrogenous base. Nitrogenous bases are molecules containing nitrogen and carbon atoms and are responsible for encoding genetic information.

Nucleotides are joined together by phosphodiester bonds to form long chains, which in turn make up nucleic acid molecules. The sequence of nitrogenous bases

in the nucleotides is what encodes genetic information in DNA and RNA, playing a fundamental role in inheritance and protein synthesis [52]

These building blocks consist of four different types: Adenine (A), Thymine (T), Guanine (G), and Cytosine (C). The precise sequence in which these nucleotides are arranged in a DNA chain encodes the unique genetic information of an organism [49].

In the case of RNA, Thymine (T) is replaced by Uracil (U). This difference in nucleotide composition between DNA and RNA is crucial for the synthesis and function of these molecules in biological processes [53].

In notation, there are other symbols used in nucleotide sequences when the exact nucleotide at that position is not known with certainty, or when the sequence contains mixtures of nucleotides at that position. For example, W means that at the position where this letter is found in a sequence, there could be an Adenine (A) or a Thymine (T), but it is not specified which of the two nucleotides is present at that position. Table 2.1 is the listing of the different possibilities for representing the content at each nucleotide position.

The genome of an organism consists of different types of regions, each with specific functions. The main types of regions in the genome include [55]:

i) Coding Regions. These regions contain genes, which are segments of DNA that encode information for the synthesis of proteins or functional RNA molecules. Genes are made up of exons, which are the parts of the gene that are translated into proteins or functional RNA. Proteins are key elements in most cellular functions, and the genetic information contained in the coding regions is essential for the structure and function of the organism [56].

The coding regions of the genome are directly related to codons in the translation process, which is the synthesis of proteins from the genetic information contained in messenger RNA (mRNA) [57]. Codons, also known as triplet bases, are sequences of three nucleotides in mRNA that code for a specific amino acid. Each codon acts as a nucleotide triplet that functions as a lexical unit in the genetic code and establishes a unique association with a specific amino acid [58]. For example, the codon AUG codes for the amino acid methionine and serves as the start codon

Table 2.1: DNA Notation. Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984. Nomenclature Committee of the International Union of Biochemistry (NC-IUB) [54].

Symbol	Description	Represented Bases				Complement
A	Adenine	A				T (or U)
C	Cytosine		C			G
G	Gunanine			G		C
T	Thymine				T	A
U	Uracil				U	A
W	Weak	A			T	W
S	Strong		C	G		S
M	aMino	A	C			K
K	Keto			G	T	M
R	puRine	A		G		Y
Y	pYrimidine		C		T	R
B	not A (B comes after A)		C	G	T	V
D	not C (D comes after C)	A		G	T	H
H	not G (H comes after G)	A	C		T	D
V	not T (V comes after T)	A	C	G		B
N	any Nucleotide (not a gap)	A	C	G	T	N
Z	Zero					Z

in translation [59]. In the process of translation, mRNA is sequentially read by ribosomes, which are cellular structures responsible for protein synthesis. When a ribosome encounters a codon in the mRNA, it seeks the complementary anticodon on a transfer RNA (tRNA) molecule that brings the corresponding amino acid to be attached to the nascent protein. This process is repeated as the ribosome moves along the mRNA [60].

In the coding regions of the genome, nucleotides are grouped into triplets (codons) because each set of three nucleotides encodes a specific amino acid during the process of translation [61]. This is a fundamental aspect of the structure of the genetic code. However, in non-coding regions, there is no restriction on grouping into triplets, as they are not translated into proteins and may contain nucleotide sequences with various functions, such as regulatory regions or introns. Therefore, in non-coding regions, the same rule of triplet grouping followed in

coding regions does not apply [62].

Figure 2.1 is a representation in circular table format that displays all possible combinations of three nucleotides (codons) and the amino acids they encode [61].

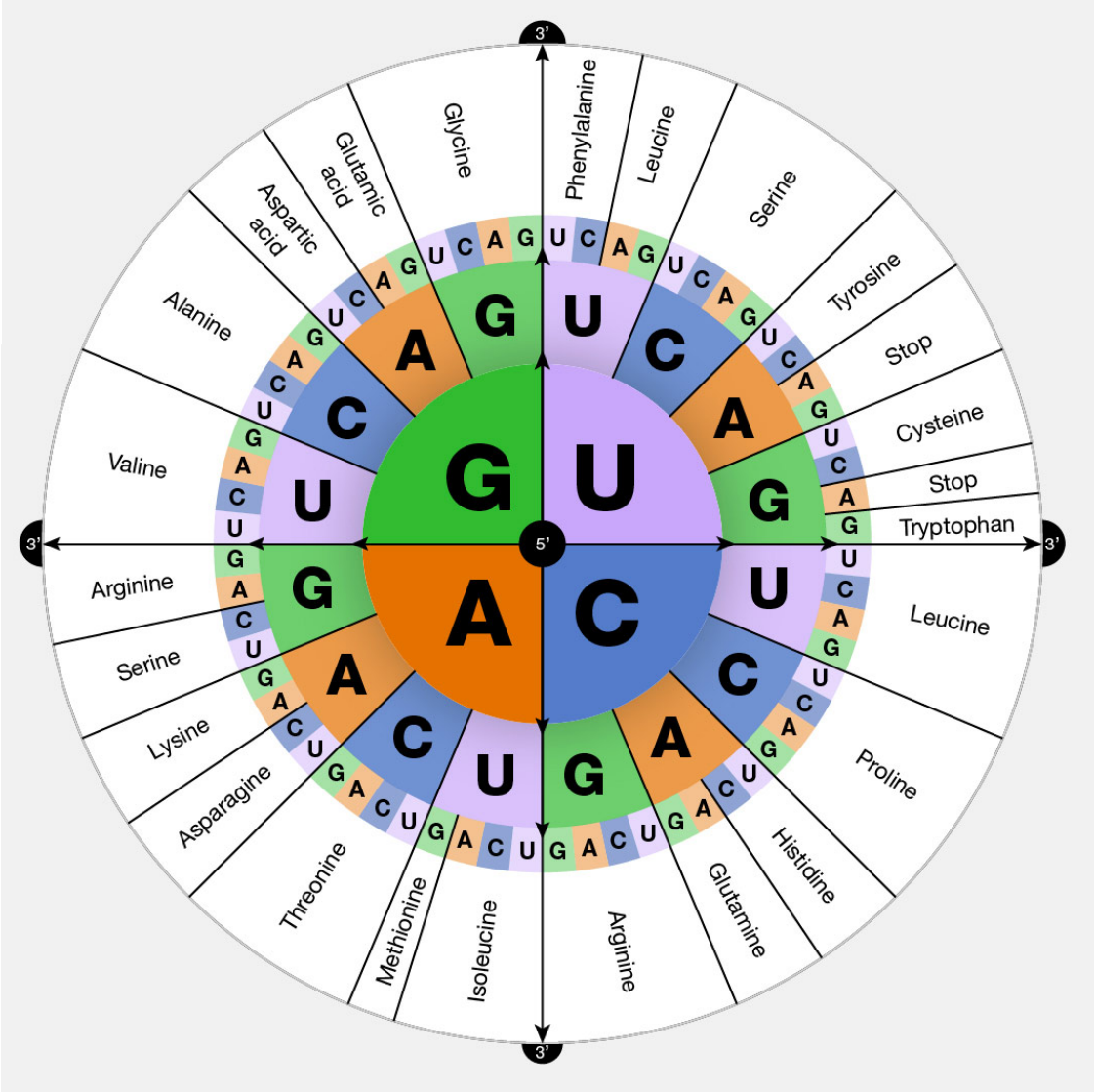


Figure 2.1: Codon Wheel. This circular table is read by starting from the center and moving outward. Each codon is represented as a sequence of three letters, corresponding to the nitrogenous bases present in messenger RNA (Adenine, Cytosine, Guanine, and Uracil). Along the circumference of the wheel, all possible combinations of these bases are shown. Image extracted from National Human Genome Research Institute [61].

ii) Non-Coding Regions. They do not encode proteins. In the case of eukaryotic organisms, these regions constitute the majority of the genome. However, in the case of viruses, these regions are minor. Despite their lack of encoding,

non-coding regions fulfill critical functions. Some of these functions include gene expression regulation (regulatory regions) [63], the organization of the chromosome structure (structural regions) [64], and the protection of coding genes against harmful mutations [65, 66]. MicroRNAs and other types of non-coding RNAs that play an important role in gene regulation and cellular processes have also been identified. The functions of these fragments are largely unknown [67].

The genome of a virus is quite different from the genome of eukaryotes. Typical features of viral genomes include [68, 69]:

- Small Size. The viral genomes are typically much smaller than those of cellular organisms. They can range from about 5,300 base pairs (as in the case of the bacteriophage φX174 [70]), to around 41,100 base pairs [71] (as in the case of nidoviruses, and among them, SARS-CoV-2 ranges 26,000 - 32,000 base pairs [72]), but are considerably shorter compared to cellular genomes [73].

- Simpler Structure. Viral genomes tend to have a simpler structure compared to cellular genomes. They often consist of a single strand of DNA or RNA and can be linear or circular. Genetic information in viruses is stored more compactly, as they tend to carry only the genes essential for their replication and function [74].

- More Coding Regions. In comparison to their size, viral genomes tend to have a higher proportion of coding regions (genes) rather than non-coding sequences. This is because viruses need to efficiently utilize their limited genetic material to carry out essential functions, such as replication and viral protein production [75].

- Overlapping Genes. Some viruses have overlapping genes, which means they share part of their sequence in different reading frames. This is a strategy to maximize the encoding of genetic information in a limited space [76, 77].

- Diverse Organization. The organization of viral genomes can vary significantly. Some viruses have linear genomes with cohesive ends or redundant terminals, while others have circular genomes. Some viruses may also have separate segments of DNA or RNA that are packaged together in the viral particle [78].

In viruses, non-coding regions can also have important functions, although their role may vary depending on the type of virus and its life cycle. Some

common types of non-coding regions in viral genomes include regulatory elements, such as promoters and enhancers, as well as other elements which control viral transcription and replication:

a) Regulatory Regions. Viruses can have non-coding regions that act as:

a.1) Promoters. They are located near the beginning of a gene and serve as the initial binding site for transcription factors and transcription machinery. The interaction in the promoter region determines whether a gene will be transcribed or not. These regions are crucial for regulating the expression of viral genes and their replication cycle [79, 80].

a.2) Enhancers: These genomic elements increase promoter activity, resulting in increased gene transcription. They are often found at a certain distance from the gene they regulate [81].

b) Initiation and termination sequences. Non-coding regions may contain sequences for the initiation and termination of viral transcription and translation. These sequences are essential for marking where viral protein synthesis begins and ends [82].

c) Splicing Elements. In some viruses, non-coding sequences may contain splice sequences that are important for processing viral mRNA during replication and gene expression [83].

d) Replication Origin. In DNA viruses, non-coding sequences often contain the replication origin, which is necessary to initiate viral genome replication [84].

e) Terminal Repeat Regions. Some viruses, such as retroviruses, have repetitive non-coding regions at their ends that are essential for the integration of the viral genome into the host genome [85].

f) Latency-Associated Regulatory Regions. They are regulatory elements that perform functions similar to silencers, meaning they can negatively control gene expression. They are commonly found in DNA viruses and some retroviruses. For example, some DNA viruses like herpesviruses have sequences that control viral latency, allowing the virus to remain dormant in the host genome for long

periods before reactivation [86]. Retroviruses can also have regulatory sequences that control viral transcription and their ability to remain latent in host cells [87].

It is important to note that the importance and function of non-coding regions in viral genomes can vary significantly among different types of viruses [88–90]. Understanding these non-coding regions is fundamental to comprehending gene expression regulation and other genetic and epigenetic processes [91]. Therefore, it is imperative to identify non-coding regions and unequivocally establish their function to understand the biology and pathogenicity of viruses.

Figure 2.2 shows an example of a genomic sequence (HIV-1 *env* gene) and its encoded protein.

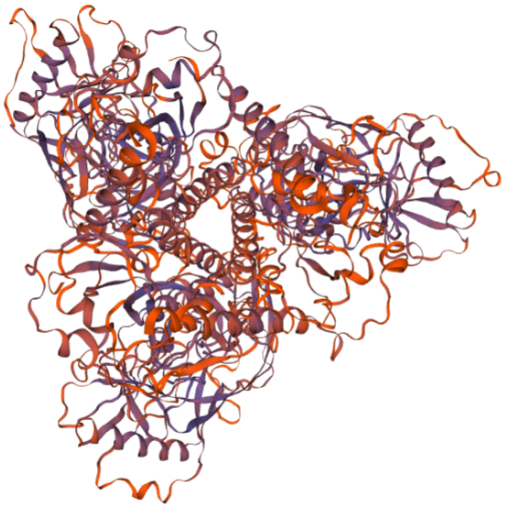
2.3 Viral Diversification Mechanisms

The mechanisms of viral diversification are crucial biological processes that allow viruses to adapt and evolve in response to changes in their environment and the selective pressures imposed by the immune system and therapeutic interventions [93]. These mechanisms are essential for the survival and spread of viruses, as they allow them to generate genetic variants that may be more effective in infecting hosts and, in some cases, evade the immune system defenses [94]. Throughout this section, I will explore some of the key mechanisms of viral diversification.

2.3.1 Mutation

The capacity of viruses to vary severely complicates health care and prevention policies [95]. Mutation is one of the main mechanisms underlying viral diversity. It consists of introducing random errors during the replication process. These errors may result in differences in some viral characteristics [96].

Mutation is a permanent alteration or change in the DNA sequence of an organism's genome. Mutations can occur spontaneously or be induced by various factors such as environmental agents, radiation, chemicals, or errors during DNA replication or repair processes [97].



(a) HIV-1 *env* gp160 3D Protein

```
ATGAGAGTGAAGGAGAAGTATCAGCACTTGTGGAGATGGGGGTGGAAATGGGGCACCATG
CTCCTTGGGATATTGATGATCTGTAGTGCTACAGAAAAATTGTGGGTCACAGTCTATTAT
GGGGTACCTGTGTGGAAGGAAACAACCACCACTCTATTTTGTGCATCAGATGCTAAAGCA
TATGATACAGAGGTACATAATGTTTGGGCCACACATGCCTGTGTACCCACAGACCCCAAC
CCACAAGAAGTAGTATTGGTAAATGTGACAGAAAATTTTAACATGTGGAAAAATGACATG
GTAGAACAGATGCATGAGGATATAATCAGTTTATGGGATCAAAGCCTAAAGCCATGTGTA
AAATTAACCCCACTCTGTGTTAGTTTAAAGTGCACTGATTTGAAGAATGATACTAATACC
AATAGTAGTAGCGGGAGAATGATAATGGAGAAAGGAGAGATAAAAAACTGCTCTTTCAAT
ATAGGCACAAGCATAAGAGATAAGGTGCAGAAAGAATATGCATTCTTTTATAAACTGGAT
ATAGTACCAATAGATAATACCAGCTATAGGTTGATAAGTTGTAACACCTCAGTCATTACA
CAGGCCTGTCCAAAGGTAACCTTTGAGCCAATTCCCATACATTATTGTGCCCCGGCTGGT
TTTGCGATTCTAAAATGTAATAATAAGACGTTCAATGGAACAGGACCATGTACAAATGTC
AGCACAGTACAATGTACACATGGAATCAGGCCAGTAGTATCAACTCAACTGCTGTTAAAT
GGCAGTCTAGCAGAAGAAGATGTAGTAATTAGATCTGCCAATTTCACAGACAATGCTAAA
ACCATAATAGTACAGCTGAACACATCTGTAGAAATTAATTGTACAAGACCCTACAACAAT
AAAAGAAAAAGTATCCGTATCCAGAGGGGACCAGGGAGAGCATTTGTTACAATAGGAAAA
ATAGGAAATATGAGACAAGCACATTGTAACATTAGTAGAGCAAAATGGAATGCCACTTTA
AAACAGATAGCTAGCAAATTAAGAGAACAATTTGGAAATAATAAAACAATAATCTTTAAG
CAATCCTCAGGAGGGGACCCAGAAATTGTAACGCACAGTTTTAATTGTGGAGGGGAATTT
TTCTACTGTAATTCAACACAACTGTTTAATAGTACTTGGTTTAATAGTACTTGGAGTACT
GAAGAGTCAAATAGCACTGAAGGAAGTGACACAATCACACTCCCATGCAGAATAAAACAA
TTTATAAATATGTGGCAGAAAGTAGGAAAAGCAATGTATGCCCCTCCCATCAGTGGACAA
ATTAGATGTTCATCAAATATTATTGGGCTGCTATTAACAAGAGATGGTGGTAATAACAAC
AATGGGTCCGAGATCTTCAGACCTGGAGGAGGCGATATGAGGGACAATTGGAGAAGTGAA
TTATATAAATATAAAGTAGTAAAAATTGAACCATTAGGAGTAGCACCCACCAAGGCAAAG
AGAAGAGTGGTGCAGAGAGAAAAAAGAGCAGTGGGAATAGGAGCTTTGTTCCTTGGGTTC
TTGGGAGCAGCAGGAAGCACTATGGGCGCACGGTCAATGACGCTGACGGTACAGGCCAGA
CAATTATTGTCTGGTATAGTGCATCAGCAGAACAATTTGCTGAGGGCTATTGAGGCGCAA
CAGCATCTGTTGCAACTCACAGTCTGGGGCATCAAACAGCTCCAGGCAAGAATCCTGGCT
GTGGAAAGATACCTAAAGGATCAACAGCTCCTGGGGATTTGGGGTTGCTCTGGAAAACTC
ATTTGCACCACTGCTGTGCCTTGGAATGCTAGTTGGAGTAATAAATCTCTGGAACAGATT
TGGAATAACATGACCTGGATGGAGTGGGACAGAGAAATTAACAATTACACAAGCTTAATA
CACTCCTTAATTGAAGAATCGCAAAACCAGCAAGAAGAGAATGAACAAGAATTATTGGAA
TTAGATAAATGGGCAAGTTTGTGGAATTGGTTTAACATAACAAATTGGCTGTGGTATATA
AAATTATTCATAATAATAGTAGGAGGCTTGGTAGGTTTAAGAATAGTTTTTGCTGTAGTT
TCTATAGTGAATAGAGTCAGGCAGGGATATTCACTATTATCGTTTCAGACCCACCTCCCA
ATCCCGAGGGGACCCGACAGGCCCGAAGGAATAGAAGAAGAAGGTGGAGAGAGAGACAGA
GGCAGATCCATTCGATTAGTGAACGGATCCTTAGCACTTATCTGGGACGATCTGCGGAGC
CTGTGCCTCTTCAGCTACCACCGCTTGAGAGACTTACTCTTGATTGTAACGAGGATTGTG
GAACTTCTGGGACGCAGGGGGTGGGAAGCCCTCAAATATTGGTGGAACCTCCTACAATAT
TGGAGTCAGGAACTAAAGAATAGTGCTGTTAACTTGCTCAATGCCACAGCCATAGCAGTA
GCTGAGGGGAAAGATAGGGTTATAGAAGTATTACAAGCAGCTTATAGAGCTATTCGCCAC
ATACCTAGAAGAATAAGACAGGGCTTGGAAAGGATTTTGCTATAA
```

(b) HIV-1 *env* gp160 FASTA File

Figure 2.2: Three-Dimensional Visualization of a Protein and the Corresponding Genomic Sequence. Figure 2.2a shows the 3D visual representation of the protein gp160 encoded by the *env* gene. Q79670 (ENV_HV1MV) Human Immunodeficiency Virus Type 1 group O (isolate MVP5180) (HIV-1) Envelope glycoprotein gp160. See SWISS-MODEL [92]. Figure 2.2b shows the FASTA file containing the genomic sequence of the HIV-1 *env* gene, complete cds (ID. L42371).

Mutations can have diverse effects on an organism, ranging from no apparent impact to causing genetic disorders or contributing to evolutionary changes [98].

Types of mutation:

i) Point mutation. It is an alteration in a single pair of DNA bases. This can include the substitution of one base for another (base substitution mutation). This means that it changes one letter in the genetic code. They are usually the most common ones [99].

ii) Insertion mutation. It involves the addition of one or more bases into the DNA sequence. This can shift the reading frame of the codons (three

nucleotides in messenger RNA) and alter the way the DNA sequence is translated into a protein [100].

iii) Deletion mutation. It occurs when one or more bases are deleted from the DNA sequence. Similar to an insertion mutation, it can affect the reading frame and have consequences on the translation of DNA into proteins [101].

iv) Duplication mutation. It occurs when an additional copy of a DNA fragment is produced. This can result in an increased quantity of certain genes or regions of the genome [102].

v) Inversion mutation. It involves the reversal of the order of a section of DNA. This can alter the way the genetic sequence is transcribed and translated [103].

vi) Translocation mutation. It occurs when a portion of DNA is transferred from one location to another in the genome. This can have implications for gene expression and chromosomal structure [104].

The size of genetic mutations can vary widely, from very small changes in the DNA sequence to larger insertions or deletions (major genomic rearrangements) [105]. In HIV-1, deletions or insertions affecting several hundred base pairs have been observed. These mutations can have a significant impact on the function of viral proteins and, in some cases, may contribute to resistance to antiretroviral drugs [106, 107].

2.3.2 Genetic Recombination

Genetic recombination is another frequent phenomenon. This happens when two different strains of a virus, or two different viruses, co-infect a cell at the same time and the result of this replication is a new hybrid genome, called recombinant [108]. Recombinant viruses are usually life-incompatible. However, in some cases, they can produce new pandemic viruses. This is the case of SARSCoV-2, MERS-CoV, Influenza Virus or HIV-1 [109].

Viruses resulting from genetic recombination can cause pandemics with new and potentially devastating effects [110–113].

Recombination in RNA viruses is an exchange of genetic information between genomic RNA molecules [114], which occurs during viral replication usually by a mechanism known as copy-choice which consists of a switch of the growing RNA from the genomic RNA template of a virus to the genomic RNA of another, resulting in a hybrid viral genome containing RNA fragments from both parental genomes [115]. Recombinant viruses have been detected with high frequency in various RNA viruses, such as picornaviruses [116] and coronaviruses [117]. But until now, HIV-1 is one of the viruses with the most recombinant sequences characterized, and a large amount of information on whole genome sequences is found in sequence databases.

In some cases, genetic recombination can be relatively straightforward. Figure 2.3 shows example of genetic recombination. It corresponds to the Circulating Recombinant Form (CRF) subtype of HIV-1 CRF23_BG [118].

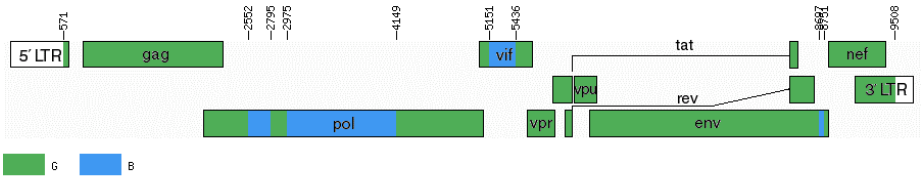


Figure 2.3: CRF23_BG Genome Map. Image sourced from the Los Alamos National Laboratory HIV-1 Database HIV Circulating Recombinant Forms (CRFs) [41].

In this case, the genome of the CRF23_BG subtype, as implicitly indicated by its own nomenclature, is a mixture of the pure subtypes G and B. The majority of the CRF23_BG genome comes from subtype G, except for two fragments in *pol* and the central region of *vif*, which come from subtype B.

Sometimes, genetic recombination phenomena are complex mosaics of several subtypes, as is the case with the subtype CRF18_cpx, composed of fragments from up to six different subtypes [119] (Figure 2.4). If the recombination involves more than three subtypes, the CRF nomenclature includes the distinction Complex (CPX) [120].

In the case of CRF18_cpx subtype, the majority of *gag* comes from A, except for the last fragment, which comes from G. The *pol* gene consists of fragments from G, F,

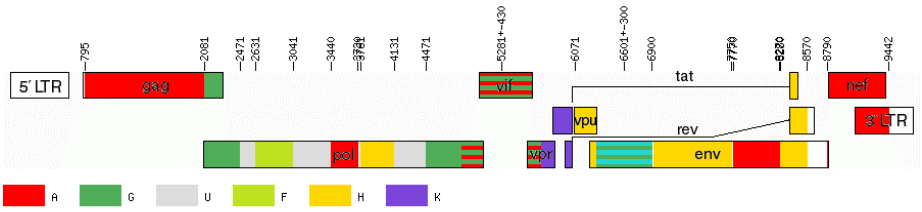


Figure 2.4: CRF18_cpx Genome Map. Image sourced from the Los Alamos National Laboratory HIV-1 Database HIV Circulating Recombinant Forms (CRFs) [41].

A, H, from an indissoluble mixture of A and G, as well as certain regions categorized as U (unknown). The *vif* gene is entirely composed of this mixture between A and G. The *vpr* gene is composed in its first half of this mixture and the second half comes from K. The *vpu* gene consists of a first half from K and a second from H. The *env* gene is composed of fragments from H, A, and a mixture related to G.

2.4 Multiple Sequence Alignment

Traditional methodologies for the analysis and comparison of genome sequences are based on multiple sequence alignment. This requires introducing gaps to place each nucleotide, letter by letter, in the same position and then make a direct comparison of the aligned sequences. A similar process is applied for protein sequence alignments.

The alignment process involves placing sequences such that homologous positions are aligned with each other while minimizing differences and penalizing insertions or deletions (gaps).

Some features of Multiple Sequence Alignment (MSA) algorithms are as follows:

- Identification of similarities and differences. MSA allows for the identification of conserved and variable regions between sequences, which can provide information about evolution, biological function, and phylogenetic relationships among the sequences.

- Resolution of insertions and deletions. Multiple sequence alignment helps identify regions where nucleotides (or amino acids in the case of protein alignment) have been inserted or deleted relative to a reference sequence. These regions may have significant functional or evolutionary implications.

- Establishment of reading frames. In the case of DNA sequences, MSA can help establish the correct reading frames for translation into proteins, which is crucial for accurate gene prediction and interpretation of protein function.

- Comparison of secondary structures and conserved domains. Sequence alignment can facilitate the comparison of secondary structures and the identification of conserved domains among proteins, which can help infer functions and structural relationships.

- Visualization and comparative analysis. MSAs are often graphically represented to facilitate the visualization and comparative analysis of aligned sequences. This allows researchers to identify patterns, conserved regions, and specific features in the sequences.

In Figure 2.5, I present a basic example of multiple alignment in a coding region of HIV-1, where the insertion of two triplets is required to save a gap in one of the sequences to achieve proper alignment.

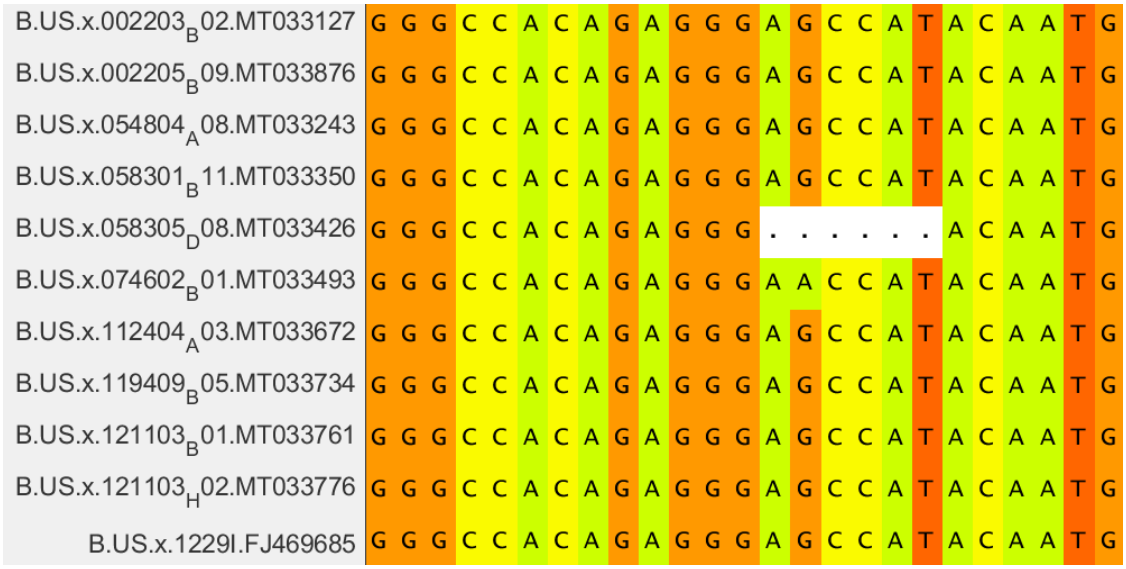


Figure 2.5: HIV-1 Multiple Sequence Alignment. By introducing gaps (·), MSA algorithms search the better position for each nucleotide of each sequence so a direct comparison is possible. This Figure was generated using MATLAB2021b Computational Biology Apps.

To perform these multiple alignments, various specific algorithms have been designed, each with different performance in terms of speed, accuracy, and specific capabilities.

In general, sequence alignment algorithms can be classified into two main categories:

i) Pair-based Algorithms. Pair-based algorithms compare pairs of sequences and gradually extend the alignment to multiple sequences. Examples of these algorithms include the Needleman-Wunsch algorithm for global alignment [121] and the Smith-Waterman algorithm for local alignment [122].

ii) Progressive Algorithms. Progressive algorithms build multiple alignments iteratively, starting with related sequences and gradually adding other sequences. These algorithms include the progressive alignment method used in tools such as ClustalW [123] and MAFFT [124].

The main goal of these algorithms is to identify similar regions among the sequences and align them based on their similarities, facilitating the identification of conserved patterns and the inference of evolutionary relationships among the sequences [125]. Some of the most well-known algorithms for multiple sequence alignment include:

ClustalW and Clustal Omega. Clustal (short for CLUSTer ALignment) is a widely used family of algorithms. ClustalW is one of the earliest multiple sequence alignment algorithms, and Clustal Omega is its latest version, which is faster and more accurate [126]. These programs use a tree-based approach to guide the alignment process [127].

MUSCLE (MUltiple Sequence Comparison by Log-Expectation). MUSCLE is another popular algorithm that uses an iterative technique to progressively improve alignments. It is known for being fast and accurate, and is widely used for aligning protein sequences [128].

MAFFT (Multiple Alignment using Fast Fourier Transform). MAFFT is an algorithm that uses the Fast Fourier Transform to perform multiple sequence alignments. It provides options for aligning large sequences and has several alignment strategies to accommodate different needs [124].

T-Coffee (Tree-based Consistency Objective Function For alignmEnt Evaluation). T-Coffee combines information from multiple alignment methods

to improve accuracy. It can align sequences with varying degrees of similarity and is useful for enhancing alignments in highly conserved regions [129].

ProbCons (Probabilistic Consistency-Based Multiple Sequence Alignment). This algorithm uses a probabilistic approach for multiple alignment and relies on statistical models to enhance accuracy, especially in regions of low similarity [130].

MSAProbs (Multiple Sequence Alignment Probabilistic Consistency-Based). It is based on a probabilistic approach for multiple alignment and employs statistical models to enhance alignment accuracy. This algorithm is renowned for its ability to handle highly divergent biological data and regions of low similarity in sequences. Its probability-based approach allows for accounting for uncertainty in alignment and enhancing accuracy, particularly in challenging alignment regions [131].

GLProbs (Genome-wide Likelihood-based Probabilistic Inference). It uses an adaptive approach based on the degree of similarity between the sequences. When dealing with input sequences that exhibit a high degree of similarity, it opts for a global alignment approach, considering the entire sequences. Conversely, for sequences showing moderate to low similarity, we may choose to disregard the flanking regions and instead perform local alignments [132].

2.5 Phylogeny. Definition and Concepts

A phylogenetic tree is an acyclic graphical representation that models the evolutionary relationships among biological entities, such as species, genes, or DNA sequences, based on the similarity or divergence of their inherited characteristics. These relationships are established through the hierarchical grouping of entities into clades, where clades represent monophyletic groups that share a common exclusive ancestor.

The basic steps for constructing a phylogenetic tree are [133]:

i) Selection of Sequences. Select a set of genetically related sequences to be used for constructing the phylogenetic tree. These sequences can be DNA, RNA, or protein sequences.

ii) Sequence Alignment. Align the sequences to identify similar regions among them. This involves aligning the sequences so that homologous characters are in the same position in all sequences. This can be done using sequence alignment programs such as ClustalW, MAFFT, or MUSCLE.

iii) Construction of the Distance or Similarity Matrix. Calculate the distance or similarity between each pair of aligned sequences. This matrix is used to determine how related the sequences are to each other.

iv) Selection of the Tree Construction Method. There are several methods for constructing a phylogenetic tree, such as the Neighbor-Joining method, the Maximum Likelihood method, and the Maximum Parsimony method. Each method has its own assumptions and advantages.

v) Tree Construction. Use the selected method to build the phylogenetic tree from the distance or similarity matrix. This step involves assembling the sequences into a tree structure that reflects the evolutionary relationships among them.

Some concepts related to phylogeny are as follows [133]:

- Nodes. A phylogenetic tree consists of nodes and branches. The nodes are points on the tree where the branches split. Each node represents the common ancestor of the species or groups found in the branches that split at that node.

- Branches. The branches represent the evolutionary relationships between species or groups. The closer two species are on a branch, the more recent their common ancestor, and vice versa. The length of the branches is often interpreted as a measure of the amount of time that has passed since the divergence between species.

- Root. The root of the tree, typically located at the bottom, represents the oldest common ancestor of all the species depicted in the tree.

- Clades. Clades are monophyletic groups, meaning a common ancestor and all its descendants. In a phylogenetic tree, clades are represented as groups of species that share a common ancestor, and all species within a clade are more closely related to each other than to any other species outside that clade.

- Species or Groups. At the tips of the tree branches, the species or groups of organisms being compared are placed. Each species or group is positioned on the tree based on its evolutionary relationships with others.

In virology, determining the common ancestor or viral phylogeny can be more complicated compared to other groups of organisms due to several reasons [134]:

a) High Mutation Rate. Viruses tend to have very high mutation rates. This means that their genetic sequences can change rapidly over time, making it challenging to identify a common ancestor. Mutations can accumulate significantly in a short period, complicating the tracking of evolutionary relationships [135].

b) Genetic Recombination. Viruses often can exchange genetic material with other related viruses. This recombination process can mix different segments of viral DNA or RNA, further complicating the determination of phylogenetic relationships. New viruses can emerge from the recombination of different strains, making it difficult to identify a single common ancestor [136].

c) Rapid Evolution. Viruses can evolve quickly in response to selective pressure, such as the host immune response or changing environmental conditions. This rapid evolution can make phylogenetic relationships difficult to trace over time [137].

d) Lack of Fossil Records. Unlike other groups of organisms, viruses do not leave fossil records, which would make it possible to trace evolution through fossils. Therefore, data on virus evolution must be based on genetic and molecular analyses, which may not provide a complete picture of phylogenetic relationships [138].

Despite these challenges, virologists use genetic sequencing techniques, phylogenetic analysis, and advanced bioinformatics tools to infer the evolutionary relationships among viruses. Although determining an exact common ancestor may be difficult, it is possible to construct phylogenetic trees that provide a representation of how related viruses cluster based on genetic similarities and shared evolution. These trees are valuable tools for understanding the epidemiology and evolution of viruses, which is crucial in the prevention and control of viral diseases.

There are several algorithms and methods for constructing phylogenetic trees from molecular or genetic data. Some of the most commonly used ones are:

- Neighbor-Joining. This is a heuristic method that constructs phylogenetic trees from a distance matrix between sequences. It is fast and efficient for large datasets [139].

- Maximum Likelihood. Maximum Likelihood (ML) methods aim to find the tree that maximizes the probability of observing the given data given the evolutionary model. RAxML (Randomized Axelerated Maximum Likelihood) is an example of a widely used ML-based program [140, 141].

- Maximum Parsimony. The principle of parsimony seeks the tree that requires the fewest evolutionary changes to explain the observations [142]. PAUP (Phylogenetic Analysis Using Parsimony and Other Methods) is a program that implements this approach [143].

- Bayesian Inference. Bayesian methods combine prior information with observed data to estimate the posterior probability distribution of parameters and the phylogenetic tree [144]. BEAST (Bayesian Evolutionary Analysis Sampling Trees) is a software that utilizes this approach [145].

- UPGMA (Unweighted Pair Group Method with Arithmetic Mean). Similar to the Neighbor-Joining method, UPGMA constructs phylogenetic trees from a distance matrix. It is simple but may not be suitable for data with variable rates of evolution [146].

- FastTree. This is a fast method that uses heuristics to construct phylogenetic trees from large datasets. It is useful when computational efficiency is required [147].

- PhyML (Phylogenetic estimation using Maximum Likelihood). It is a program that implements Maximum Likelihood methods and allows for the search of optimal phylogenetic trees under different evolutionary models [148].

- MrBayes. This is a program that uses Bayesian inference to construct phylogenetic trees and model the evolution of sequences [149].

The choice of algorithm will be highly influenced by the following factors (among others):

- Dataset Size. Some algorithms may be more efficient than others for large datasets. In general terms, algorithms such as Neighbor-Joining [150] and UPGMA [151] tend to be more efficient for large datasets due to their simplicity and lower computational demand. On the other hand, algorithms such as Maximum Likelihood [152] and Bayesian methods [153] may be less efficient for large datasets due to their greater complexity and need for more intensive calculations.

- Evolutionary Model Complexity. Some algorithms may be better suited to handle complex and heterogeneous evolutionary models. This is the case of Maximum Likelihood [154] and Bayesian methods [155], as they have the capability to incorporate detailed and flexible evolutionary models in phylogenetic analysis.

On the other hand, simpler algorithms like Neighbor-Joining [150] and UPGMA [156] may be less suitable for handling complex evolutionary models, as they have more simplified assumptions about evolution and may not adequately capture the heterogeneity in the data.

- Computational Requirements. Some algorithms may require more computational resources than others, so the availability of computational resources may also influence the choice of algorithm. The algorithms that usually require more computational resources are those that perform intensive calculations, such as Maximum Likelihood [152] and Bayesian methods [157], due to their complexity and exhaustive search algorithms. On the other hand, simpler algorithms like Neighbor-Joining [158] and UPGMA [151] tend to require fewer computational resources due to their more straightforward and less computation-intensive approach.

- Presence of Missing Data or Errors. Some algorithms may be more robust against missing data or errors in the data than others, because they can effectively handle data variability and compensate for the lack of information. Some robust algorithms in this regard may include Neighbor-Joining and maximum parsimony methods [159, 160], as they tend to ignore sites with missing or ambiguous data in phylogenetic analysis.

On the other hand, algorithms that heavily rely on the availability of complete and accurate data, such as Maximum Likelihood and Bayesian methods [161, 162], may be less robust when faced with missing data or errors, as the absence of information can significantly affect the accuracy of the results.

- Specific Analysis Objectives. Depending on the objectives of the phylogenetic analysis (e.g., inferring precise evolutionary relationships, estimating divergence dates, etc.), certain algorithms may be more suitable than others [163].

In summary, the choice of algorithm will depend on a combination of factors, including the nature of the data, available computational resources, and the objectives of the phylogenetic analysis.

In any case, currently the generation of large phylogenetic trees is still a technological challenge [164].

The most technologically efficient machine that man has ever invented is the book.

— Northrop Frye, *philosopher, linguist, and theologian. Honorary Doctor from around thirty European and American universities. During his first year at Oxford, Frye forgoes the purchase of a typewriter in favor of a piano, which he has delivered to his room at Merton College.*

3 Technological Background

Contents

3.1 Introduction 48
3.2 Introduction to the Concept of Machine Learning . . . 49
3.3 Introduction to Deep Learning Architectures 51
3.4 Introduction to Convolutional Neural Networks 52
3.5 Introduction to Transfer Learning 57
3.6 Introduction to Explainability Tools 60
3.6.1 Gradient-Weighted Class Activation Mapping (Grad-CAM) 61
3.6.2 Local Interpretable Model-Agnostic Explanations (LIME) 62
3.6.3 Gradient Attribution 64
3.7 Objective of this State-of-the-Art Review 65
3.8 Methodology 65
3.9 Search Topics and Preliminary Article Detection . . . 68
3.10 Summary Diagram of Technological Background Review Methodology 68
3.11 Genomic Spectrograms. State of the Art. 71
3.12 Deep Learning Applied to Biosanitary Field. State of the Art 75
3.13 Deep Learning Applied to Genomics. State of the Art 82
3.13.1 Nature of Multi-omic Data 82
3.13.2 Genome Processing and Algorithms 84
3.13.3 Identification and Analysis 87
3.13.4 Prediction 93
3.13.5 Specific Deep Learning Applications in Genetic Recombination 95
3.13.6 Metagenomics 96
3.13.7 Study of Diseases 99
3.13.8 Precision Medicine 103
3.13.9 Crop Improvement 105

3.13.10 Improvement of Existing Technology 106
3.14 Interpretability of Biosanitary Deep Learning. State of the Art 109
3.15 Critical Analysis of the State of the Art 113
3.15.1 Multiple Sequence Alignment 113
3.15.2 Deep Learning Applied to Biology and Medicine 119
3.16 Conclusions 133

3.1 Introduction

As I mentioned in Chapter 1, this research combined two disciplines: microbiology+virology and computer science. In Chapter 2, I introduced a series of microbiological concepts necessary for understanding this document. Similarly, in this Chapter, I briefly outlined concepts related to Machine Learning (ML) and Deep Learning (DL) and provided a brief analysis of the current state of the art.

This Chapter begins with several introductory sections to various concepts: Machine Learning (Section 3.2), Deep Learning (Section 3.3), Convolutional Neural Networks (Section 3.4), Transfer Learning (Section 3.5), and explainability (Section 3.6).

In Section 3.7, I present the objective of the State-of-the-Art Review conducted. Section 3.8 details the methodology based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses method (PRISMA). In Section 3.9, I present the established search topics and the results of a Preliminary Article Detection.

In the following sections, I present the results of the analysis conducted for each search topic: genomic spectrograms (Section 3.11), Deep Learning applied to biosanitary field (Section 3.12), Deep Learning applied to genomics (Section 3.13), and interpretability of biosanitary Deep Learning (Section 3.14).

Subsequently, I conducted a critical analysis of the state of the art, where I discuss various limitations and controversial aspects found in Multiple Sequence Alignment and in Deep Learning applications in biology and medicine. I detail the results in Section 3.15.

This Chapter concludes with the findings from the analysis (Section 3.16).

3.2 Introduction to the Concept of Machine Learning

ML and DL techniques essentially focus on the development of algorithms that enable machines to learn and improve by themselves through experience and data, rather than being explicitly programmed to perform specific tasks. This involves the use of mathematical models including linear algebra, calculus, statistics and probability, etc., that allow computer systems to enhance their performance as they are exposed to more data [38]. ML algorithms have a more data-centric approach [165, 166]. Figure 3.1 provides an easily understandable representation of the paradigm shift compared to conventional programming. In Section 3.10, , I provide a diagram that visually aids in the interpretation of the methodology followed.

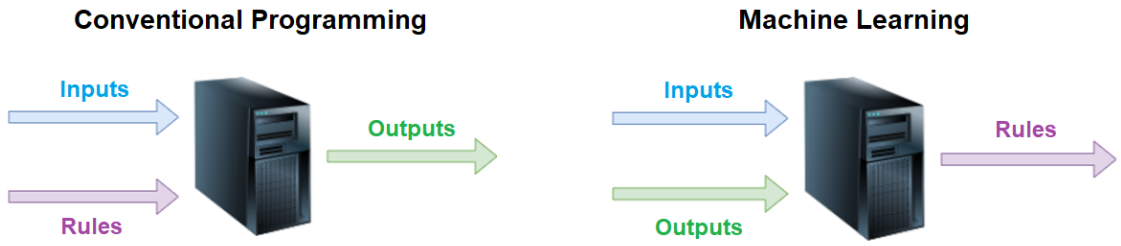


Figure 3.1: Machine Learning vs. Conventional Programming. In conventional programming, the machine is fed with input data along with corresponding instructions and operating rules. The machine processes this data according to the provided rules and produces output data as a result. The development of Machine Learning (ML) algorithms marks a paradigm shift in programming. In ML (and subsequently in Deep Learning, DL), the machine receives both input and output data. From these, it detects patterns and learns the underlying rules. Subsequently, the machine returns these rules governing the relationships between the input and output data.

Deep Learning algorithms constitute an advanced branch within Machine Learning that revolutionized the capability of machines to perform complex tasks. It relies on the use of deep artificial neural networks, which consist of multiple layers of interconnected computational units, enabling the learning of hierarchical data representations. These learned representations are capable of capturing abstract and complex features of the input data, enabling the system to perform more sophisticated tasks.

The takeoff of Deep Learning began in 2006. The ability of deep neural networks to learn complex hierarchical representations from large datasets, driven by advances in hardware and algorithms, enabled remarkable performance in tasks such as computer vision, natural language processing, generative models, etc., consolidating DL as a fundamental pillar in the current landscape of the concept known as Artificial Intelligence (AI) [38]. Figure 3.2 schematically represents the temporal progression of AI tools.

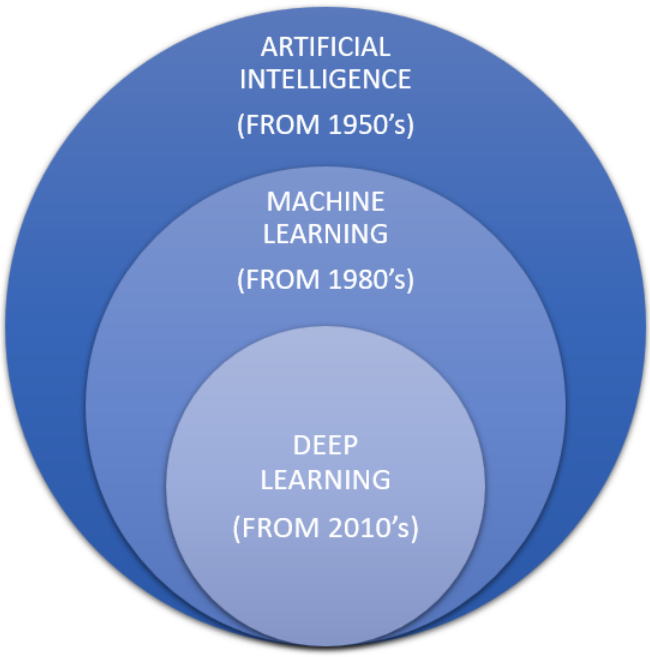


Figure 3.2: Temporal progression of AI. The first milestone represents the initial conception of AI in the 1950s, reflecting rule-based and symbolic approaches. The second milestone, in the 1980s, marks the advent of machine learning algorithms, emerging as a distinctive component of AI, which enable machines to learn patterns from data. The third milestone is the consolidation of deep learning tools in the 2010s. These algorithms stand out as a specialized form of machine learning, characterized by deep neural networks and their ability to learn complex hierarchical representations.

In summary, the evolution from initial AI tools to the significant rise of DL algorithms has been a technical process marked by advances in algorithms and computational capabilities. Initially, AI relied on logical rules and symbols, evolving towards more data-centric approaches with the advent of machine learning. The application of deep neural networks enabled the processing of complex data, enhancing modeling capability and accuracy in specific tasks [38]. These advancements,

supported by large datasets and hardware improvements, redefined the ability of machines to tackle complex problems across various fields.

3.3 Introduction to Deep Learning Architectures

Some of the main DL architectures are:

1. Convolutional Neural Networks (CNNs). Widely used for computer vision tasks such as image recognition and semantic segmentation. They are composed of convolutional layers that apply filters to detect features in images, followed by pooling layers to reduce dimensionality, and fully connected layers for classification [167].

2. Recurrent Neural Networks (RNNs). They are suitable for tasks involving sequences of data [168], such as Natural Language Processing (NLP) [169] and automatic translation [170]. They have feedback connections that allow them to maintain information over time. However, they can have issues with gradients vanishing or exploding during training, which has led to the development of variants such as Long Short-Term Memory (LSTM) [171] and Gated Recurrent Unit (GRU) [172], to alleviate these limitations.

3. Long Short-Term Memory Neural Networks (LSTM). They are a variant of RNNs. They are designed to address the long-term forgetting problem in sequence training. They incorporate gates that control the flow of information, allowing them to learn long-term dependencies and have been shown to be effective in sequence modeling tasks [173].

4. Autoencoders. They are used to learn efficient representations of input data. They consist of two main parts: an encoder, which converts input data into a compressed representation (called latent code), and a decoder, which tries to reconstruct the input data from this compressed representation. Autoencoders are useful in dimensionality reduction applications, noise removal in images, data generation, and information compression [174].

5. Transformers. The main novelty of these models is that they introduce the so-called attention mechanism. This mechanism allows capturing long-distance

dependency relationships. Therefore, it is a very effective tool in natural language processing tasks, such as automatic translation and text generation. Unlike RNNs, transformers do not have recurrent connections and rely entirely on attention to capture relationships between different parts of the input sequence [175].

6. Generative Adversarial Networks (GANs). They are generative models, meaning they are oriented towards generating new data, primarily realistic images. They consist of two neural networks: a generator that creates new data and a discriminator that tries to distinguish between real and fake data. These networks compete against each other in an adversarial learning process until the generator produces data that is difficult to distinguish from real data [176].

In the context of Deep Learning, generative models are implemented using various neural network architectures, such as GANs [177], Variational Autoencoders (VAEs) [178], and language models based on Transformers [179].

3.4 Introduction to Convolutional Neural Networks

CNNs have several important advantages:

- They have demonstrated solid robustness in pattern identification in images, which has remained over time despite the emergence of new architectures [180].

- They are computationally efficient [181].

- There are multiple pre-trained models that are highly powerful, with their high performance extensively demonstrated in the existing literature [182].

- They allow for reliable application of Transfer Learning. Transfer Learning is a technique in Machine Learning, where a model trained on one task is adapted or reused for another related task. Instead of training a model from scratch for each specific task, the learned knowledge by the pre-trained model on one task is leveraged, and its learned representations are transferred to a new task. This often results in more efficient and faster learning for the new task, especially when there are limitations in the amount of available data [183].

- They allow for reliable and visually understandable application of explainability tools [184].

These neural networks apply the mathematical operation known as convolution. This operation involves sliding a filter (also known as a kernel) over the input image and calculating the dot product between the elements of the filter and the corresponding region of the image at each position. This process generates a feature map that highlights specific patterns in the image, such as edges, textures, or shapes. The application of this operation preserves the spatial structure of the data, meaning the spatial relationship between the pixels. The convolutional layers perform this operation of running filter to detect local patterns in the image [185].

The discrete convolution operation is performed by multiplying an input matrix by a kernel matrix (filter), which is a smaller and localized version of the original image. The mathematical formula for discrete convolution for a two-dimensional image I and a kernel K is:

$$(I * K)(i, j) = \sum_m \sum_n I(m, n) \cdot K(i - m, j - n) \tag{3.1}$$

Where:

- $I * K$ is the result of the convolution,
- (i, j) are the coordinates of the resulting pixel,
- m and n are the relative coordinates within the kernel,
- $I(m, n)$ is the value of the pixel of the image at the position (m, n),
- $K(i - m, j - n)$ is the value of the pixel of the kernel shifted at the position $(i - m, j - n)$,
- and the summation is performed over all values of m and n within the limits of the kernel and the image respectively.

This formula represents the 2D convolution for a single position (i, j) of the result. In practice, this operation is performed for each position of the result to obtain the complete feature map.

Furthermore, in CNNs, pooling layers are typically applied after convolutional layers. These layers reduce the dimensionality of the feature maps obtained through convolution, preserving the most relevant features while reducing the number of parameters and improving computational efficiency. Although the pooling layer reduces the spatial resolution of the data, the general spatial information is preserved thanks to the preceding convolution process [186–188].

A graphical representation of the basic operation of a convolutional module is depicted in Figure 3.3 [38].

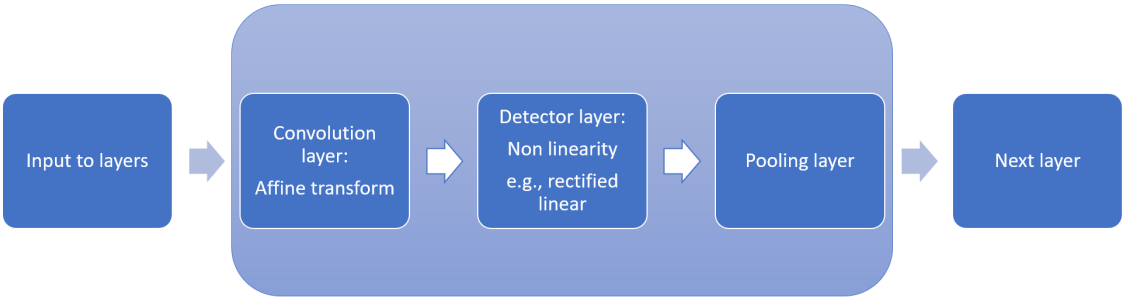


Figure 3.3: Typical elements of a CNN layer. When convolution is applied in a CNN layer, filters are used to detect specific features in the input. These filters capture patterns such as edges, textures, or shapes in the image. Then, the output of the convolution is passed through a nonlinear activation function, such as ReLU, which introduces nonlinearities in the network and helps to learn more complex and nonlinear representations of the data. Finally, the pooling layer is responsible for reducing the dimensionality of the feature maps while preserving the most important features.

A pre-trained model of a CNN is a model that has been previously trained on a large dataset, extensively and comprehensively. These pre-trainings are generally oriented towards an image classification task, such as ImageNet [189]. During this training process, the model learns to recognize a wide variety of features and patterns in the images, enabling it to accurately classify them into various categories.

Applying Transfer Learning to these pre-trained models makes it possible to leverage their potential in a new task, leveraging that prior knowledge. The previously

learned features by the model are adapted and refined for this new technique.

Within the catalog of existing pre-trained models, notable architectures include VGG-16, ResNET-101, and Inception V3. Their main features are as follows.

1. ResNet-101 [190, 191]:

- Residual connections. ResNet-101 introduces the concept of residual connections, which help mitigate the problem of performance degradation in very deep neural networks. This allows for the construction of even deeper and more easily trainable architectures.

- Depth and accuracy. ResNet-101 is one of the deeper variants of the ResNet model family, enabling it to capture more complex and accurate representations of images. This is especially useful for tasks requiring high precision.

- Transfer Learning. Due to its depth and effectiveness, ResNet-101 is a popular choice for Transfer Learning, where previously learned knowledge from ImageNet can be adapted to specific tasks with smaller datasets.

Figure 3.4 depicts a comprehensive diagram of the ResNet architecture [192].

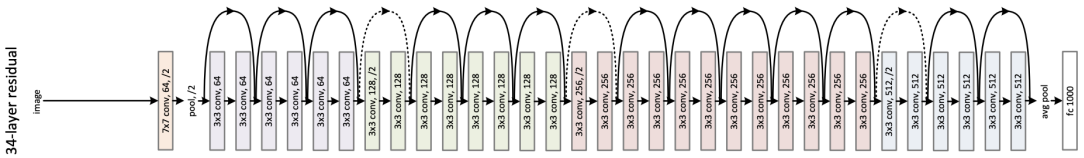


Figure 3.4: ResNet-based Architecture. This diagram depicts a residual network with 34 parameter layers. ResNet-101 consists of a total of 101 layers, including convolutional, normalization, activation, and pooling layers. These layers are organized into residual blocks.

2. Inception V3 [193, 194]:

- Filter Diversity. Inception V3 employs filters of different sizes (1x1, 3x3, 5x5) in parallel in some layers, allowing it to capture features at different scales and levels of abstraction in images.

- Computational Efficiency. Despite its more complex architecture, Inception V3 has been designed to be computationally efficient, making it suitable for implementation on devices with limited resources, such as mobile devices.

- Regularization. Inception V3 incorporates regularization techniques such as kernel size reduction and L2 regularization, which help prevent overfitting and improve model generalization.

Figure 3.5 shows a detailed scheme of Inception V3 architecture [195] (extracted from Inception V3 Advanced Guide by Google [196]):

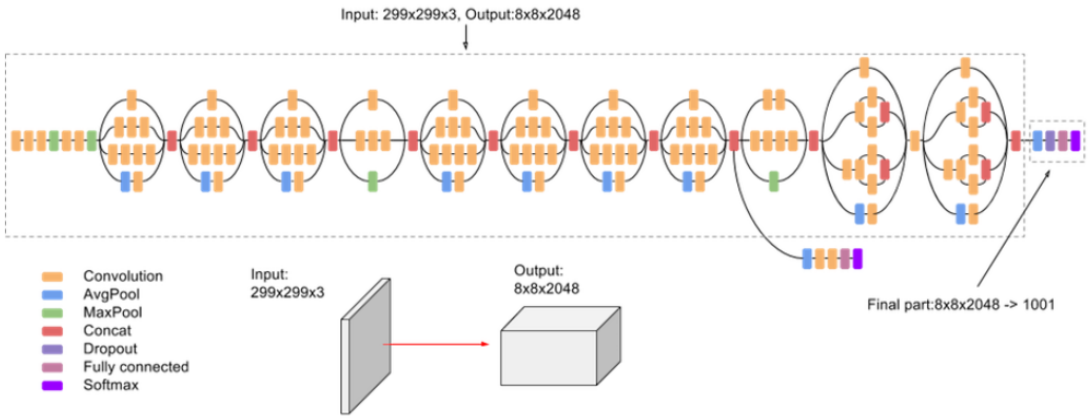


Figure 3.5: Inception V3 Architecture. This architecture is characterized by the incorporation of Inception modules to capture features at different scales and levels of abstraction. The integration of filters of multiple sizes and the use of regularization techniques contribute to improving the efficiency and performance of the model.

3. VGG-16 [197, 198]:

- Simplicity and effectiveness. One of the main characteristics of VGG-16 is its simple and effective architecture, which consists of deep convolutional blocks with small-sized filters (3x3) and max pooling layers. This simplicity makes it easy to understand and implement.

- Competitive results. Despite its simplicity, VGG-16 has shown competitive performance in image classification tasks on various datasets, including ImageNet.

- Wide adoption. Due to its simplicity and good performance, VGG-16 has been widely adopted and used as a basis for many other models and research projects in the field of computer vision.

Figure 3.6 displays the architecture of VGG-16. This figure is based on the diagram drawn by Kenneth Leung, which illustrates VGG-16 architecture [199].

As seen in Chapter 5, the architecture of choice was VGG-16 as it yielded the best results in my experimental benchmark.

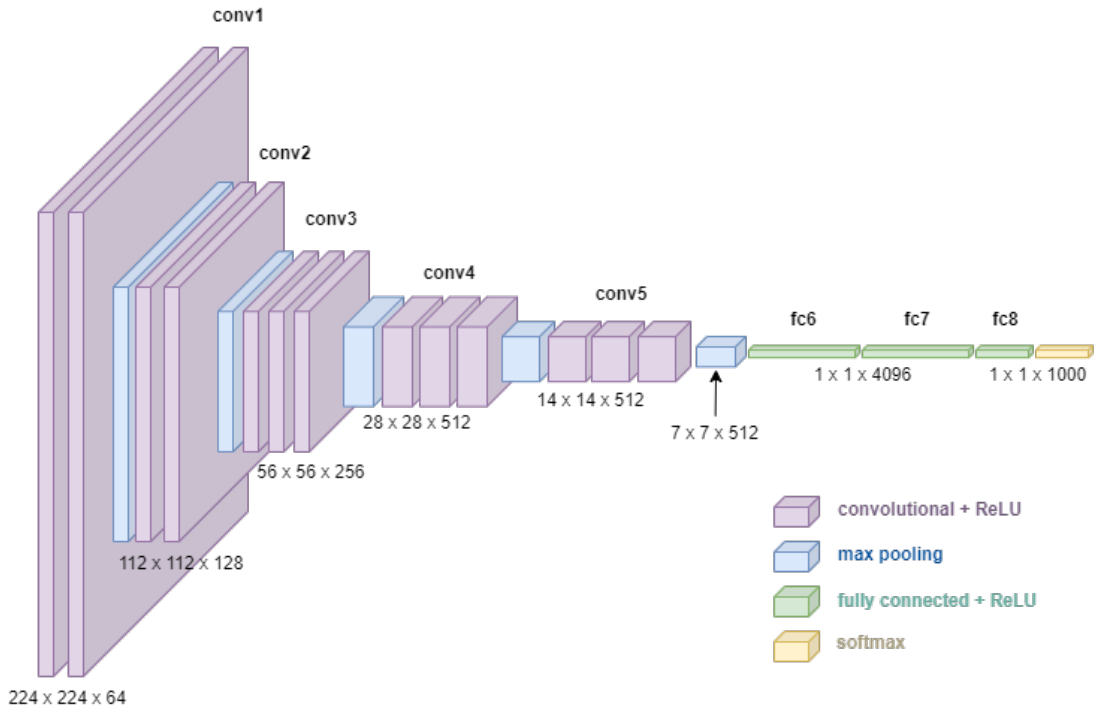


Figure 3.6: VGG-16 Architecture Schematic. The VGG-16 architecture is distinguished by its structure of deep convolutional blocks, where small-sized filters (3x3) are repeatedly used in successive layers to capture features at different scales and levels of complexity. This arrangement of filters enables the neural network to learn increasingly abstract representations of the images as it goes deeper into the layers. Additionally, regularization through techniques such as data augmentation and kernel size reduction helps prevent overfitting and improve model generalization.

3.5 Introduction to Transfer Learning

As indicated in Section 3.4, one of the advantages of utilizing pre-trained CNNs is the application of Transfer Learning to harness this acquired knowledge.

Transfer Learning is a technique in Machine Learning that reuses a model previously trained on one task (source task) for a new task (target task). This technique is especially useful when the target task has limited labeled data, as it leverages the knowledge acquired in the source task to improve performance in the target task [48].

Basic Concepts [200, 201]:

- Domain ($\mathcal{D}$). Defined by a feature space $\mathcal{X}$ and a marginal distribution $P(X)$, where $X \in \mathcal{X}$. For example, in image classification, $\mathcal{X}$ may represent the

pixel space and $P(X)$ the distribution of images.

- Task ($\mathcal{T}$). Consists of a label space $\mathcal{Y}$ and an objective function $f : \mathcal{X} \to \mathcal{Y}$. The task involves learning f from pairs (x, y) where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

Specific Definitions [202]:

- Source Task ($\mathcal{T}_S$) and Source Domain ($\mathcal{D}_S$). Represent the task and domain in which the model is initially trained.

- Target Task ($\mathcal{T}_T$) and Target Domain ($\mathcal{D}_T$). Represent the new task and domain to which the knowledge is transferred.

Transfer Learning Categories:

1. Inductive Transfer Learning [203]:

The target task ($\mathcal{T}_T$) is different from the source task ($\mathcal{T}_S$), although the domains ($\mathcal{D}_S$ and $\mathcal{D}_T$) may be similar or different. For example, transferring a model trained on ImageNet for image classification to the task of medical image classification.

2. Transductive Transfer Learning [204]:

The tasks ($\mathcal{T}_S$ and $\mathcal{T}_T$) are the same, but the domains ($\mathcal{D}_S$ and $\mathcal{D}_T$) are different. Such as using a voice recognition model trained in English for voice recognition in Spanish.

3. Unsupervised Transfer Learning [205]:

Both the tasks and domains differ between the source task and the target task, and there are no labels available in the target domain. For instance, utilizing a pre-trained model to generate text data features and applying them to document clustering in an unlabeled domain.

Transfer Learning Procedure [206–208]:

1. Pre-trained Model:

1.1. Initial Training:

- Source Domain: A Deep model (CNN, RNN, Transformers) is trained on a large dataset.

- Feature Learning: The initial layers of the network learn general features, while later layers learn more specific features.

1.2. Feature Extraction:

- The early layers of the network capture general features (e.g., edges, textures), useful across many different tasks.

2. Model Adaptation:

2.1. Layer Freezing:

- Definition: Freezing certain layers involves keeping their weights fixed during training on the target task.

- Application: Typically, early layers are frozen, while later layers are kept trainable or replaced with new layers adapted to the target task.

2.2. Fine-Tuning:

- Definition Adjusting the weights of the pre-trained model on the target task.

- Procedure: The model is trained on the target task with a low learning rate to avoid large updates to the already learned weights.

Mathematical Formalization [209–212]:

1. Pre-training on the Source Task:

Given a model M_S with parameters θ_S, the objective is to minimize the loss function $\mathcal{L}_S$:

$$\theta_S = \arg\min_{\theta} \mathcal{L}_S(\theta; \mathcal{D}_S, \mathcal{T}_S)$$

2. Fine-Tuning on the Target Task:

To adjust the model for the target task, a new loss function $\mathcal{L}_T$ is minimized, which includes a regularization term to penalize large deviations from θ_S:

$$\theta_T = \arg\min_{\theta} \left[\mathcal{L}_T(\theta; \mathcal{D}_T, \mathcal{T}_T) + \lambda \|\theta - \theta_S\|^2\right]$$

where λ is a hyperparameter controlling the regularization.

Some of the advantages provided by the use of Transfer Learning are as follows:

1. Reduction in Training Time. Leveraging pre-trained models accelerates the training process for the target task.

2. Improved Performance. Transfer Learning enhances performance on the target task, especially with limited data.

3. Data Efficiency. Transfer Learning enables working with smaller datasets since the model has learned useful general features.

The use of Transfer Learning involves analyzing the following factors to ensure successful application [213]:

1. Domain Mismatch. If the source and target domains are very different, transfer may not be effective.

2. Computational Overhead. Pre-trained models are usually large and require significant computational resources.

3. Hyperparameter Tuning. Finding the optimal settings for learning rate and other hyperparameters can be complex.

Therefore, through Transfer Learning, it is possible to develop advanced applications in highly specific domains by leveraging knowledge acquired from pre-trained networks. The main application areas include:

1. Computer Vision. Image classification, transfer of models trained on large datasets (e.g., ImageNet) to specific tasks such as disease classification in medical images [214–216].

2. Natural Language Processing (NLP). As an example of text classification, adaptation of language models like BERT or GPT for specific tasks such as sentiment analysis or email categorization [217, 218].

3. Speech and Audio Analysis. Transfer of speech recognition models trained in one language to other languages or adaptation of models to detect emotions in voice recordings across different contexts [219–221].

3.6 Introduction to Explainability Tools

The nature of neural networks as "black boxes" is widely recognized in the scientific literature [222] (see Section 3.15). Visual explainability tools help obtain information about the understanding of decisions made by the neural network, highlighting areas of special interest on the images [223].

Explainability techniques are based on the mathematical generation of score maps where the areas of the image relevant to the neural network's decision-making

are discriminated. Subsequently, a color scale is applied to the original image so that we can directly visualize on the image where the network looked to classify it. We don't know what it sees, but at least we know where it looks [224].

Three of the main explainability tools are Grad-CAM, LIME, and Gradient Attribution.

3.6.1 Gradient-Weighted Class Activation Mapping (Grad-CAM)

Grad-CAM stands for Gradient-weighted Class Activation Mapping. The main advantage of Grad-CAM is its ability to generate heatmaps that highlight specific regions of an image that the model considers most relevant for making a classification decision. These heatmaps provide a visual interpretation of the model's decision-making process [225].

The Grad-CAM technique works by backpropagating the gradient of the output class score with respect to the activations of the last convolutional layer of the neural network. This allows calculating the relative importance of each pixel in the image with respect to the class of interest. Pixels with higher gradients indicate regions of the image that significantly contribute to the model's classification decision.

Once the gradients are calculated, Grad-CAM produces a heat map that highlights the most important regions of the image. These regions correspond to areas where changes in pixels have the greatest impact on the output of the neural network, providing an intuitive visual understanding of how the model is interpreting the image [226].

Suppose that a convolutional neural network (CNN) produces a class output y from an input image X. We aim to comprehend which parts of the image are significant for predicting class c.

1. Activations of the last convolutional layer. Denote the activations of the last convolutional layer before the output layer as A. These activations represent high-level features extracted from the image by the CNN.

2. Gradient of the class score with respect to the activations of the last convolutional layer. We calculate the gradient of the class score y_c with respect to the activations A. This gives us a measure of how much the class score y_c changes when the activations A change.

3. Grad-CAM heatmap. Finally, we multiply each channel of the activations A by the gradient calculated in the previous step and take the spatial average to obtain the Grad-CAM heatmap. Mathematically, this is expressed as:

$$\text{Grad-CAM}(X, c) = \text{ReLU}\left(\sum_i \sum_j \frac{\partial y_c}{\partial A_{ij}} A_{ij}\right) \tag{3.2}$$

where A_{ij} represents the activation at position (i, j) of the last convolutional layer, $\frac{\partial y_c}{\partial A_{ij}}$ is the gradient of the class score with respect to that activation, and ReLU is the rectified linear activation function, which takes the maximum value between 0 and the value inside the parentheses.

This formula is the essence of Grad-CAM and is used to generate the heatmap that highlights important regions of an image for a specific class [227, 228]. This heatmap provides a visual interpretation of how the Deep Learning model is making predictions on images [229].

3.6.2 Local Interpretable Model-Agnostic Explanations (LIME)

LIME (Local Interpretable Model-Agnostic Explanations) is a technique used to provide interpretable explanations for the predictions of machine learning models, regardless of their complexity or architecture [230]. It works by locally approximating the original machine learning model in the neighborhood of the point of interest and evaluating the relative importance of each input feature for the model prediction [231].

The operational process of LIME can be described in the following steps [232, 233]:

1. Generation of local models:

LIME operates by generating a locally interpretable model that approximates the original Machine Learning model in the vicinity of the point of interest. This

local model is constructed by randomly or weightedly selecting samples near the point of interest in the feature space.

2. Feature Perturbation:

Once the neighborhood of the point of interest has been selected, LIME modifies the input features of these samples using perturbation techniques, such as feature deletion, random noise, or domain-dependent semantic perturbations.

3. Evaluation of Feature Importance:

After feature perturbation, the changes in the predictions of the local model when modifying the input features are evaluated. This involves constructing an interpretable model, such as linear regression, that fits the perturbed data and assessing its ability to predict class labels.

4. Interpretable Explanations:

Based on the evaluation of feature importance, LIME provides interpretable explanations on how each input feature contributes to the model output. These explanations can be presented in the form of relative weight or contribution of each feature, allowing understanding of which features are more influential in the model prediction.

Therefore, LIME is a technique that focuses on generating interpretable explanations for the predictions of machine learning models. Although LIME does not rely on a single mathematical formula like Grad-CAM, its operation involves the construction of local models and the evaluation of feature importance [234]. The LIME process can be summarized with the following pseudo-code:

1. Generation of local models:

- Select a point of interest.

- Sample or generate a set of data points neighboring the point of interest.

- Weight the neighboring data points to account for their proximity to the point of interest.

2. Feature Perturbation:

- For each neighboring data point, randomly modify the input features to generate a perturbed version of the original data point.

3. Construction of the interpretable model:

- For each perturbed version of the neighboring data points, make a prediction using the original Machine Learning model.

- Use an interpretable model, such as linear regression, to fit the perturbed data and predict class labels.

4. Evaluation of Feature Importance:

- Evaluate how the predictions of the interpretable model change when modifying the input features.

- Calculate the relative importance of each input feature based on how it affects the prediction of the interpretable model.

3.6.3 Gradient Attribution

The gradient attribution methods explored in this example generate high-resolution pixel maps that reveal which pixels are most significant for network classification. These methods compute the gradient of the class score with respect to input pixels. Intuitively, the map identifies which pixels have a greater impact on the class score when altered. While these methods provide more detailed maps than Grad-CAM or occlusion sensitivity, they tend to be noisier. This is because a well-trained deep network does not heavily rely on the exact value of specific pixels. Gradient attribution techniques are useful for identifying broad areas of an image that are relevant for classification [235].

The most basic method of gradient attribution is to compute the gradient of the class score for the predicted class with respect to each pixel in the input image [236]. This approach reveals which pixels have the greatest impact on the class score, thus indicating the most influential pixels for classification. In this example, the use of gradient attribution alongside two extended methods is presented: guided backpropagation [237], and integrated gradients [238]. However, the use of these techniques is under debate, as it has not been clearly established how much additional information they can provide about the model [239].

The gradient attribution techniques rely on finding the gradient of the prediction score with respect to the input image. The gradient attribution map is calculated using the following formula:

$$W_{cxy} = \frac{\partial S_c}{\partial I_{xy}} \tag{3.3}$$

where W_{cxy} represents the importance of the pixel at location (x, y) to the prediction of class c, S_c is the softmax score for that class, and I_{xy} is the image at pixel location (x, y).

3.7 Objective of this State-of-the-Art Review

In the current context of rapid and widespread technological advancements, society finds itself immersed in an environment where computational models are experiencing significant growth and development. This rapid evolution has led to a proliferation of new algorithms at an unprecedented pace. The inherent versatility of these tools has driven their adoption across virtually all domains of human knowledge.

I focused on the application of advanced DL methods in the biosanitary field. This sector, characterized by its particular sensitivity and relevance in the current landscape, demands innovative and effective solutions to address a wide range of challenges. However, the vast amount of scientific literature available in this field is overwhelming and challenging to cover in its entirety.

Therefore, my goal in this Chapter could not be a comprehensive analysis of the entire state of the art in both subjects. I aimed to provide a sufficient approach with the purpose of evaluating the latest trends, developments, and advancements, as well as the most significant contributions, and to use the results of this survey as support in selecting algorithms, technologies, and mathematical data transformations.

3.8 Methodology

I briefly based the current state-of-the-art analysis methodology on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses method (PRISMA)

[240]. The main phases of this method were:

1) Definition of search topics.

All of them belong to Deep Learning applied in the biosanitary field. In this review, I specifically delimited the search topics, focusing them narrowly on aspects directly related to my specific experiments. This precise approach aimed to avoid the inclusion of articles that do not bear substantial relevance to my research objectives, thus minimizing the presence of non-pertinent information that could introduce noise and confusion into this review.

By restricting the searches to highly specific topics aligned with my experiments, I sought to ensure the relevance and coherence of the reviewed studies, optimizing the usefulness of the gathered information for my specific objectives. Nevertheless, I included some articles seemingly further from my research framework in order to analyze and, if necessary, corroborate the performance demonstrated by my application techniques and technologies. I also included some articles as illustrative examples of specific aspects.

2) Identification and screening of bibliographic sources.

I expressly excluded doctoral theses, bachelor's or master's theses, as well as unofficial publications or those not subjected to a peer review process. Therefore, all articles in preprint format are expressly excluded. I also did not consider articles from non-indexed journals.

3) Quality assessment and data extraction.

From the data collected during the initial search, I selected those that are thematically coherent and meet the specific criteria established for the research.

4) Critical analysis of the state of the art.

I conducted a review of the methodology in the application of DL algorithms. If the application of these algorithms is incorrect or improper, these articles were excluded from this analysis and became part of a specific section in this chapter, called "Critical Analysis of the State of the Art". See Section 3.15.

Therefore, I screened the articles between "problematic" and "non-problematic" ones. The articles classified as "problematic" were those where the methodology for

applying DL technologies did not meet the appropriate standards for them. Therefore, the results are at high risk of containing biases and being inherently incorrect. In this case, I identified the following factors that make an article "problematic":

1. Dataset with less than 1,000 original samples. Original samples are defined as those that have not been subjected to any data augmentation or derivation process. The application of data augmentation to an original sample does not result in the creation of new samples per se. When referring to an original sample, I specifically mean the initial sample and not the samples generated from it through data augmentation.

For instance, let us consider an original dataset comprising 10 X-ray images, each corresponding to a single patient. While augmenting this dataset may yield an expanded data collection, it is crucial to acknowledge that there remain only 10 original X-ray images. Despite the augmentation resulting in a larger dataset, the genuine diversity of the samples does not undergo a substantial increase.

It is crucial to understand that the data augmentation technique is designed to generate new instances of data under certain specific conditions. However, in the biosanitary field, where data diversity and representativeness are of paramount importance, this technique does not necessarily contribute to increasing such diversity.

Therefore, in this context, the indiscriminate application of data augmentation may not be appropriate, as it does not necessarily guarantee the inclusion of a significant variety of data in the expanded dataset. It is essential to carefully consider the selection and diversity of the original data to ensure adequate representativeness in the dataset used for analysis and modeling in the biosanitary field.

Some studies rely on synthetically generated data. I consider this approach unorthodox as, especially in the field of genomics, verifying the coherence of generated data is complex or impossible. Therefore, I do not consider the uncertainty introduced by this technique suitable for the application of DL technologies. The optimal (and appropriate) strategy is to utilize a broad and representative dataset covering all existing categories, particularly when seeking generalization or universality of results.

2. Omission of categories in the dataset targeted for classification. For example, in the case of a CNN designed to classify medical images as cancer/non-cancer, it is crucial to possess data representing both categories. The absence of data corresponding to any of the classes invalidates the approach.

The inclusion of an article as "problematic" led to its inclusion in Section 3.15, where the main methodological issues detected were analyzed.

5) Processing, analysis, and study of selected articles.

6) Drawing conclusions.

3.9 Search Topics and Preliminary Article Detection

Table 3.1 provides a summarized overview of the search outcomes.

Therefore, in the brief review of the state of the art, I processed a total of 3,548 articles. As a notable fact, it is worth mentioning that only 7 articles correspond to the use of genomic spectrograms. The scarcity of articles referring to the interpretability and explainability of DL models is also striking. There are only 115 articles in the preliminary search compared to a total of 2,723 in the "Deep Learning applied to genomics" category.

Given the need to design interpretable systems that mitigate the "black box" nature of neural networks, the fact that there are few actual developments denotes, on one hand, the intrinsic difficulty in achieving interpretable models, and therefore, the necessity to continue researching in the design and implementation of effective explainability and interpretability tools.

3.10 Summary Diagram of Technological Background Review Methodology

Figure 3.7 shows the methodology diagram for conducting this review, as well as the articles identified at each step.

Table 3.1: Results of Article Identification by Search Topic. These topics are organized into properly numbered categories to facilitate their classification into themes. The column "Total number of identified articles" indicates the total number of raw articles located for each of the categories.

1.	**Genomic spectrograms**	
	Search Topic	**Total number of identified articles**
1.1.	Genome digital signal processing	247
1.2.	Genome Fast Fourier Transform FFT	10
1.3.	Genome spectrogram	27
1.4.	Genomic spectrogram	21
	Subtotal:	**305**
2.	**Deep Learning applied to Biosanitary Field**	
	Search Topic	**Total number of identified articles**
2.1.	Deep Learning microbiology	280
2.2.	Deep Learning virology	57
2.3.	Deep Learning genetic recombination	24
2.4.	Deep Learning viral evolution	44
	Subtotal:	**405**
3.	**Deep Learning applied to genomics**	
	Search Topic	**Total number of identified articles**
3.1.	Deep Learning genomic sequences	668
3.2.	Deep Learning genomics	1,230
3.3.	Deep Learning omics	768
3.4.	Deep Learning metagenomics	50
3.5.	Deep Learning genome spectrogram	7
	Subtotal:	**2,723**
4.	**Interpretability of biosanitary Deep Learning**	
	Search Topic	**Total number of identified articles**
4.1.	Interpretability Deep Learning genome	92
4.2.	Explainability Deep Learning genome	21
4.3.	Interpretability Deep Learning microbiology	1
4.4.	Interpretability Deep Learning virology	1
	Subtotal:	**115**
Total number of articles:		**3,548**

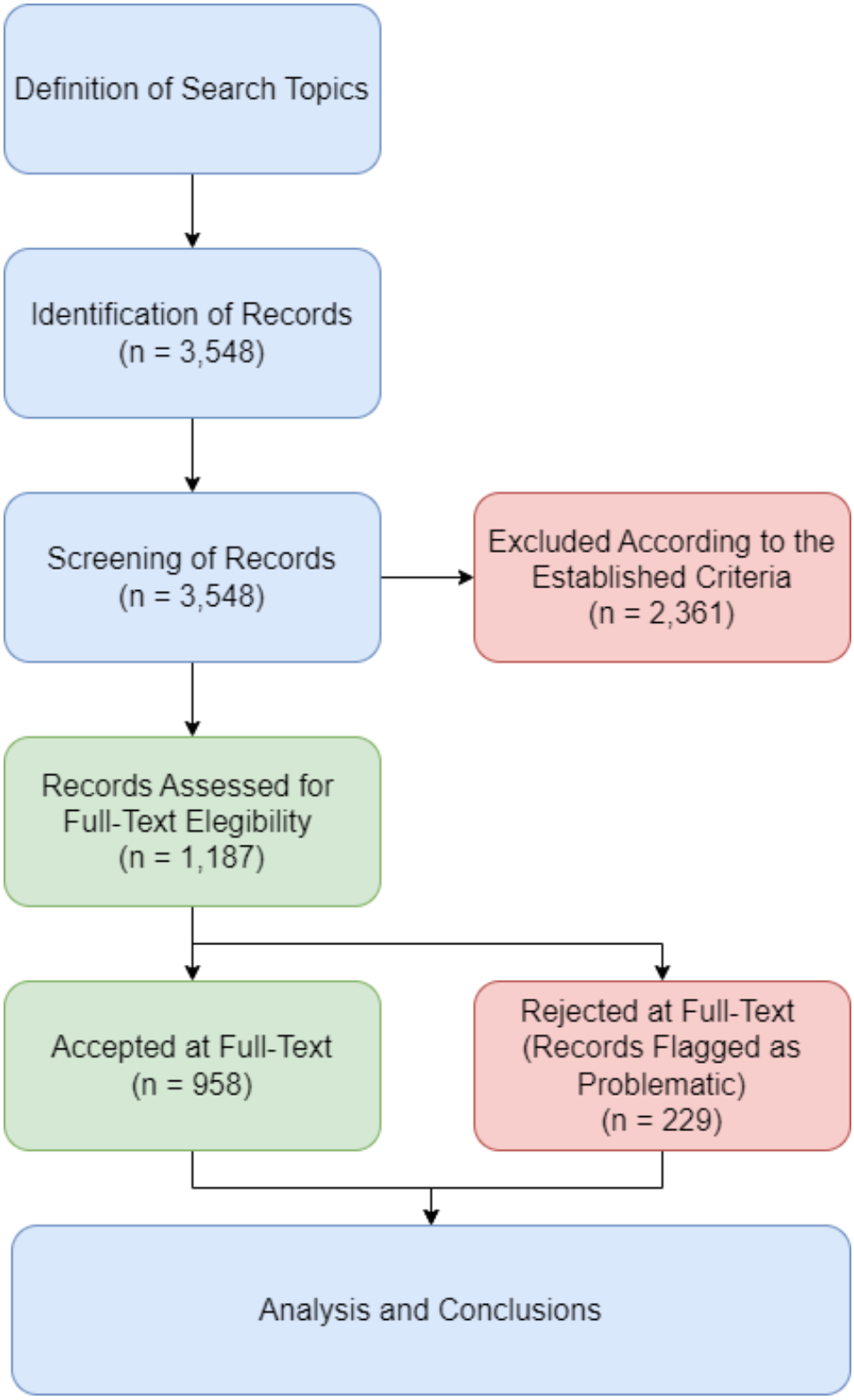


Figure 3.7: Diagram of the Technological Background Review Methodology

3.11 Genomic Spectrograms. State of the Art.

Although the application of Deep Learning Tools in health is still controversial [241], multiple studies demonstrated that the performance of Deep Learning Tools using spectrograms renders good performances in the interpretation of medical results [242–247].

Below, I aim to outline some of the successful applications of frequency domain analysis using spectrograms in the biosanitary field.

Traditionally, spectrograms have been used for the characterization of medical temporal signals. Among these temporal signals are electrocardiograms (ECGs) [248–252] , electroencephalograms (EEGs) [253–259], lung bioacoustics and biomedical signal fingerprinting [260], cardiograms [261, 262], ultrasound imaging [263–265], doppler ultrasound [266], and phonocardiograms [267–269], among others. Applications for the analysis and interpretation of circadian rhythms have also been developed [270, 271], or for the processing of piezoelectric sensors related to swallowing [272], .

However, its application has extended beyond the realm of time series. It has also been successfully applied to the analysis of infrared microspectroscopy [273, 274], or the analysis of biomedical polymers [275]. Also used in the analysis of biomedical hydroxyapatite powders and hydroxyapatite coatings on medical metal implants [276], applications in gas chromatography [277], advanced microscopy techniques [278], FTIR spectroscopic (Fourier Transform Infrared Spectroscopy) imaging applied to tissue and cell culture analysis [279], advanced computational microscopy technique Fourier Ptychographic Microscopy (FPM) [280], application in infrared spectroscopy [281], improvements in Optical Coherence Tomography (OCT) [282–285], mass spectrometry improvement [286]. There is also research aimed at predicting disease development through time series analysis [287].

Therefore, given the wide range of applications beyond time series, the use of spectrograms as a data representation model for medical and microbiological data enjoys strong support.

In the context of genomics, the conversion of genomic sequences into spectrograms is at an earlier stage than in other fields within the biosanitary domain. However, there is a constant increase in research related to genome analysis in the frequency domain, with highly interesting and inspiring investigations.

I must note that I excluded from the analysis those articles on Digital Gene Expression (DGE) because it is a DNA sequencing technique and, therefore, it falls outside of my scope.

Several research studies are related to the application of multiple Digital Signal Processing (DSP) technologies applied to genomics [288–301]. Some studies extract frequency features from the genome [302]. Some studies specifically determine how to convert genomes into digital signals [303–306].

Other studies establish how to apply digital signal processing to determine the presence of amino acids in digital signals [307]. Multiple research studies are aimed at detecting various genomic elements, such as coding and non-coding regions [308], contig overlaps [309], transcription factor regulatory targets [310, 311], or microarrays [312, 313]. Genomic spectrograms are additionally employed for predicting the subcellular localization of proteins [314], and for genomic data compression [315].

I found applications related to the tetrahedral composition of nucleotides (a tetrahedron where each vertex represents one of the four RNA nucleotidesadenine, uracil, cytosine, and guanineand the edges and faces of the tetrahedron represent the codons formed by combinations of these nucleotides) [316, 317].

There are also studies that relate the digital information of the genome to analog aspects such as thermodynamic stability [318].

There are DL applications that, using digital signal processing, remove noise in Cap Analysis of Gene Expression (CAGE), minimizing the occurrence of false positives in transcription start sites [319]

Regarding the transform employed within the frequency domain, there exist various alternatives. The Complex Morlet wavelet has been used as an analysis technique and the Frequency Chaos Game Signal (FCGS) as a digital dataset

[320]. Other transformations in the frequency domain have also been applied, such as Wavelet Transform [321], Gabor Wavelet Transform [322], LombScargle periodograms [323], a Gauss-wavelet-based analysis [324], or Huffmanbased encoding technique [325, 326]. There are studies that applied Karhunen-Loeve like transforms for the interpretation of DNA microarray data for gene expression [327]. Other frequency transforms utilized are Infinite Impulse Response (IIR), inverse Chepyshev [328], Cepstrum coefficients [329], Smoothed Fourier Transform [330], or Short-Time Ramanujan Fourier Transform [331].

There are specific developments of transforms whose functions are to prevent spectral leakage and reduce background noise, which manifests with high amplitude in the Fourier power spectrum [332]. Other research is specifically aimed at eliminating the noise produced by microarrays [333]. Non-parametric filters applied to spectrograms have also been designed [334], or also bidirectional filters to enhance digital genome processing [335].

There are several applications in genomics of time series analysis algorithms, for instance, Dynamic Time Warping (DTW) [336], or Dyadic Wavelet Transform (DWT) in conjunction with DTW [337]. There are even applications for progressive sequence alignment [338].

Other digital signal processing algorithms have been applied in domains such as solving the haplotype assembly problem [339].

Frequency-based tools have been developed to detect 3-periodicity fitted to small DNA fragments [340]

Focusing on the Fourier transform, particularly on spectrograms, there is a growing interest in their application as a tool for analyzing genomic sequences. This genomic visualization method facilitates the detection of novel and relevant patterns [306, 341–354]. Several research studies are focused on analyzing the horizontal line at $f = 1/3$ [327, 354, 355], and its relationship with Open Reading Frame (ORF) detection [356]. Fourier transforms are also used for the identification of coding and non-coding regions [357–359].

The applications of genomic spectrograms include, among others, the comparison of protein sequences [360], the localization of coding regions [361], more specifically, protein coding regions [362], gene prediction [363, 364], localization of exons and introns [365–376]. There are also applications using genomic spectrograms for mutation detection [377, 378], and even for the detection of cancer and tumor genes in the genome [379, 380].

There are multiple research studies that analyze complex genomic elements via genomic spectrograms, such as CpG islands [381] or genomic islands [382], detection of tandem repeats [383–386], palindromic repeats and Clustered Regularly Interspaced Short Palindromic Repeats (CRISP) [387–389], alpha satellites [390, 391], GC skew measurements [392], o crystals structural parameters [393].

The FFT has been demonstrated to be useful for reducing noise and more easily locating replication termini in bacteria [394], and in trasncriptome profiles [395]

Genomic spectrograms were applied to genomic sequences of various microorganisms, such as bacteria [396], eukariotic [397] and prokariotic [398]. Some research analyzed microscopic worms [399], parasites [400–402], and even insects [403].

Genomic spectrograms were used to support the identification of new organisms such as bacteriophages [404], viruses [405–408], and human hemagglutinin [409].

This research examines the genome of crop plants from the perspective of the Discrete Fourier Transform (DFT) to optimize crop yield [410].

The FFT was also being applied in detecting features in multi-omic profiles [411].

This study analyzed the parameterization of the spectrogram based on the organism under study [412]. There are explanatory articles on fine-tuning the parameterization of the spectrogram [413], refinement in the representation of spectrograms [414], filters for better noise removal [415, 416], and polyphase filters to eliminate noise in the spectrum [417].

Alignment algorithms for spectrograms were also developed [418, 419]

Several studies endorse the use of this genome representation system along with ML algorithms [420–422] and DL [423, 424]. Multiple articles recommend

generating datasets of genomic spectrograms to be analyzed with CNNs, given their potential in detecting patterns not visible to the human eye in images [425–427].

The numerous and varied successful applications documented endorse that frequency domain analysis of the genome is a relevant and promising approach.

3.12 Deep Learning Applied to Biosanitary Field. State of the Art

Machine Learning and Deep Learning tools emerge as highly promising resources to support professionals in the biosanitary field, thanks to their ability to efficiently and accurately process massive volumes of data [428–435].

Currently, ML and DL applications in the biosanitary field are more oriented towards the processing and management of other types of data, primarily medical imaging [436–441]. Some studies even affirm that DL systems have surpassed other existing technologies in medical imaging analysis [442].

DL applications in medical imaging are predominantly focused on detecting various types of cancer. There are multiple tools with seemingly validated results, such as those for classifying brain tumors [443–448], for liver tumors detection [449–451], lung tumors [452–454], skin tumors [455], gastric intestinal tumors [452, 456–458], colon cancer [459], ovaric cancer [460], breast cancer [461–463], bone cancer [464], cervical cancer [465, 466], skin lesions, skin cancer, melanoma, etc. [467–470], acral melanoma [471], or bladder cancer [472].

Another major goal of applying DL algorithms for early cancer detection is the development of tools capable of accurately identifying different subtypes of the same type of cancer. For instance, Tumor Microenvironment (TME) analysis [473], and detection of the Tumor Immune Microenvironment (TIME) with histopathological images [474]. Also, research has been conducted on determining breast cancer subtypes with Electronic Medical Records (EMR) [475]. An especially interesting case study is the analysis of cancer with an unknown primary tumor [476].

Studies focused on early cancer detection are not limited solely to medical imaging. Increasingly, research tends to effectively integrate omics data. For

example, this study early predicts breast cancer by integrating gene expression, DNA methylation, and microRNA (miRNA) expression [477].

There are also various applications for biomarker localization [478–483], even for the detection of Microsatellite Instability (MSI) from histopathology images [484].

Within DL applications in medical imaging, the goal is not only cancer detection but also predicting its progression. For example, I found developments aimed at detecting Tumor Infiltrating Lymphocytes (TILs) [485], early detection of metastasis [486], and tools for distinguishing between primary and metastatic tumors [487]. There are applications aimed at classifying cells as cancerous/non-cancerous [488], or predicting the response to chemotherapy with Single Photon Emission Computed Tomography (SPECT) and radiomic features [489].

Of course, DL applications in the healthcare field are not limited solely to tumor and cancer detection; there are also applications aimed at diagnosing a variety of additional diseases and injuries [490–524]. I want to highlight sophisticated studies like this one for Macular Edema (ME) classification on Spectral-Domain Optical Coherence Tomography (SD-OCT) images [525], or this research aimed at analyzing Hemodynamic parameters of the Middle Cerebral Artery (MCA) with ultrasonic images [526]. There are areas where the effective introduction of DL models is still in an early stage, for example, in matters of myopia detection and quantification [527].

Medical image analysis is not limited to two-dimensional images; it even extends to the analysis of three-dimensional images, such as the detection of Obstructive Sleep Apnea (OSA) in 3D craniofacial scans [528].

Similarly to cancer research, there are also applications oriented towards predicting the progression of diseases based on medical imaging [529], or other medical data [530].

The application of DL in the biosanitary area is not limited solely to medical imaging. There are powerful applications that utilize data of various types [531]. For example, this study predicts acute episodes from the analysis of clinical notes [532]. There are developments to predict the transmissibility of diseases [533, 534], Or even for early detection of postoperative wound complications [535].

The application of DL models dedicated to the analysis of time series data is common [536], for example, for the analysis of respiratory data [537, 538], or data related to stress measurement [539]. There are studies aimed at predicting the duration of antibody presence [540], estimating age and mortality risk through polysomnograms (PSGs) [541], or even predicting epidemic evolution [542].

Other DL applications are oriented towards predicting sex and age based on brain shape [543], or to predict age from over 50 omics data features [544]

One of the major applications of DL tools in the biosanitary field is in precision medicine. Precision medicine is a medical paradigm that relies on the integration of genomic, molecular, and clinical data to provide a highly individualized approach in the diagnosis, treatment, and prevention of diseases. By utilizing specific information about each patient's molecular and genetic biology, as well as environmental and lifestyle factors, it is possible to customize medical interventions to maximize efficacy and minimize risks. This approach allows treatments to be tailored to the unique characteristics of each patient, with the aim of improving clinical outcomes and quality of life [545].

There are multiple studies aimed at predicting the effectiveness of treatments based on the evolution of multiple medical data [546], and omic data [547]. Several studies investigate treatment design [548, 549], seek to identify drug candidates [550–553], predict antimicrobial resistance [554–559], or detection of antibiotic resistances [560, 561]. There are research studies related to the analysis of body composition from tomography scans [562]. Others seek to predict Compound-Protein Interactions (CPIs) [563], identify novel antimicrobial peptides [564–567], or predicting anticancer hyperfoods [568]. There are also semi-supervised learning tools that analyze causal inferences in microbiology and medicine. For example, applications to determine if a treatment causes alterations in gut flora and also influences metabolism [569].

Within the field of microbiology, one of the most common applications is the identification of microorganisms in microscopic images or spectrometric data [570–

592], or analysis of cytometry data [593]. There are also research studies aimed at detecting biofilms (complex microbial communities) in microscopy images [594, 595].

These applications are not limited exclusively to two-dimensional (2D) analysis but also encompass three-dimensional (3D) analysis. For example, analyzing the 3D morphology of cells using GAN networks [596], or studying events of aggregation of microscopic particles in 3D [597]

There are some advancements focused on image segmentation [598–600]. These advancements include specific applications, such as individual cell segmentation [601], or segmenting the neuronal microscopy images [602].

There are purely classification applications, for example, for identifying microorganisms in images [603, 604], for cell type identification [605], for labeling molecules in histopathology images [606], or even for protein identification in mass spectrometry [607]. There are also trained models for cell classification from single-cell mass cytometry data [608], for determining analysis sample outcomes from images in petri dishes [609], for interpreting Gram stain results [610], or for identifying molecules in DNA molecule images [611]

I found several specialized developments in the counting of microorganisms in colonies [612–617] and in monitoring and modeling the growth of microorganisms [618–620].

There are more advanced developments capable of predicting rare cell populations [621]. There are also algorithms specialized in predicting the response of different types of cells to perturbations [622], or in predicting collective cell migration in microscopic images [623]. This interesting research determines the optimal growth temperature of pathogens using CNNs [624]. This algorithm focuses on predicting species interaction networks [625]. Within metagenomics, there is notable research aimed at detecting relations between human gut microbes and diseases [626].

Taxonomic analysis is the focus of several studies, both in terms of taxonomic classification [627] and morphological profiling [628, 629].

There are numerous advancements enabling comprehensive analysis of various facets of microbiology. Applications in this field are highly diversified, characterized

primarily by their specificity and a notable level of complexity. Among this wide range of applications, the following stand out: Fiducial marker detection in tomographies [630], β-Hemolysis Detection [631], retrotransposons classification in plant genomes [632–634] and in the genome of the human nervous system [635], crystal centering for macromolecular crystallography [636], detection of transcriptome signatures of cancer [637], pyroptosis detection [638], interpretation of Raman spectroscopy to identify molecular fingerprints [639], detection of Copy Number Variation (CNV) [640–642], interpretation of conventional glass Gram stain microscopy slides [643], analysis of protein-RNA interactions [644, 645], prediction of protein-protein interactions [646–654], including human-virus [655], .

There are also applications aimed at molecular dynamics identification, analysis and simulation [656, 657], identification and classification of phage virion proteins [658], detection of defects in pre-mRNA (pre-messenger RNA) splicing [659], analysis of causal interactions in gene-regulatory networks to determine relationships between transcription factors and their target genes [660], analysis of virus morphogenesis [661] or virion morphology [662], analysis of secretomes [663, 664].

This development is aimed at determining the threshold cycle to interpret Reverse Transcription Polymerase Chain Reaction (RT-PCR) data [665]. As for transcriptome analysis, Deep Learning tools for spatially resolved transcriptomics are being developed [666]. There are also certain studies that introduce bots as support tools for transcriptome analysis [667].

There are also multiple applications in the agro-sanitary sector. For instance, predicting food expiration based on various data sources such as image analysis (presence of pathogens, amount of water, packaging condition, etc.) and time series, including the evolution of parameters like pH, temperature (cold chain), etc. [668, 669]; also, to monitor the quality of crops [670, 671], or the detection of crop diseases via imaging [672].

There are also Deep Learning applications that not only integrate computer science and microbiology but also incorporate other branches of knowledge such as physics. For example, applications aimed at analyzing kinetic models, that is,

particle dynamics, such as current models of airborne transmission of respiratory viruses [673]. DL was also applied in the study of algebraic topology. For example, to determine infectivity based on algebraic topologies [674].

The application of Deep Learning algorithms is not limited solely to biomedical research; it is also applied in enhancing existing techniques and technologies. There are applications aimed at improving the resolution of medical images [675], as well as enhancing the long-term visualization of dynamic interactions between intracellular structures, in other words, improving the resolution of microscopic images [676, 677].

Improvements were developed in microscopic imaging technologies aimed at a more effective application of DL tools [678, 679]. Enhanced imaging technologies supported by DL technologies were also developed, such as in structured illumination microscopy [680]. Also, tools for reducing noise in technical interferences caused by systematic errors (for example, temperature, reagent concentration, and well location), which always blend with real biological signals in microscopic images, have been developed [681]. Noise removal in Scanning Transmission Electron Microscopy (STEM) [682]. Noise removal in Electroencephalography (EEG) signals [683].

There are also support tools for managing healthcare resources, predicting recovery periods, medical needs such as transfusions, etc. [684–687], and medical complications [688].

This article presents a web server that integrates different Deep Learning models to predict the Type III Secretion Systems (T3SSs) at the genomic level of a bacteria of interest [689].

Tools for extracting data from biomedical texts were also developed [690–692], with natural language processing being a very promising tool.

Incipiently, DL technologies are beginning to be applied to surgery, mainly classifying images and predicting medical data, currently, in small non-invasive interventions [693–697]. There is even a documented case of a total mesorectal excision supported by an AI system [698].

There is a DL application for surgical robot calibration [699].

An interesting application is the use of DL systems for surgical training [700], or tools to support surgeons [701, 702].

However, the application of DL tools in surgery is extremely complex. This, combined with the importance of the field, makes it still a very nascent area. Studies like the one conducted by Kitaguchi et al. demonstrate the limited performance of deep learning tools in tasks such as surgical instrument segmentation [703].

Other interesting applications include the automatic generation of implants [704]

This DL application aims to improve the training of a robotic RL system based on a Brain-Computer Interface (BCI) [705].

With regard to data processing, multiple conversions were also performed to optimize the use of specific architectures. I want to highlight this study, in which they work with SMILES (Simplified Molecular Input Line Entry System) strings, converting molecular structures into character strings [706].

The most commonly used technology in medical image classification is undoubtedly the CNN architecture [707]. However, Visual Transformers are demonstrating extremely interesting performances [708–711]. Several studies use transformers for cancer survival prediction [712, 713].

Several studies in the field of biomedicine endorse the high capacity of CNNs in detecting patterns not visible to the human eye in images [714–717]. Multiple studies endorse the application of Transfer Learning on a pre-trained CNN to extend its potential to new types of images and new domains such as medical diagnostics. For example, for prediction of multiple diseases from chest X-ray images [718], classification of abdominal ultrasound images [719], diabetic retinopathy images [720], cellular morphological changes to identify cellular phenotypes (incipient cancer process and so on) [721], or detection of Major Depressive Disorder (MDD) by observing electroencephalogram (EEG) signals transformed to spectrograms [722].

As evidenced by the preliminary analysis conducted, the use of DL techniques in the biosanitary field has experienced exponential growth, largely driven by the COVID-19 pandemic crisis. This development has diversified significantly,

encompassing virtually all disciplines within the field, even venturing into highly specific and specialized analyses.

There are numerous research studies employing a wide variety of architectures. Despite the considerable success observed in new architectures such as transformers and Natural Language Processing techniques, Convolutional Neural Networks (CNNs) remain the predominant choice in most image-focused research. Additionally, the persistent number of successful applications using CNNs over the years further reinforces the robustness of these systems. The additional advantage of techniques like Transfer Learning facilitates the extension of CNN capabilities into new application domains with greater confidence

All these advantages, especially the endorsement from existing literature regarding their correct application in multiple sectors of the biosanitary field, lead me to select CNNs as the tool of choice.

3.13 Deep Learning Applied to Genomics. State of the Art

In this section, I included those articles closely related to the specific field of genomics.

3.13.1 Nature of Multi-omic Data

Multi-omic data refers to sets of biological information obtained through the integration of multiple omics technologies, such as genomics, transcriptomics, proteomics, metabolomics, and epigenomics. These data sets provide a holistic and multidimensional view. By combining data from different "omics", researchers can deepen their understanding of complex biological processes and molecular interactions underlying phenomena such as development, disease, and response to treatments, among others [723].

DL applications in the biosanitary field mostly work with multi-omic data (genomic data, medical imaging, and other medical data such as biomarkers, clinical, and environmental data...) [724–731]. These data structures primarily seek relational

factors between sets of data that are more easily interpretable and can be contrasted, such as medical imaging; and data with incipient and mostly unknown interpretation, as is the case with genomic data (especially if non-coding regions are involved). DL algorithms, given their ability to extract high-dimensional features, are promising for handling this type of data [732–745]. It is even claimed that unsupervised DL is a promising analytical approach for integrative analysis of multi-omic data [746, 747], including temporal data [748].

There are studies comparing ML and statistical methods, suggesting that these tools are particularly useful for handling data with large variables but a small sample size. ML also outperforms classical statistical methods in handling complex datasets [749–751]. There are studies that argue that, in general terms, applications that integrate multimodal data instead of unimodal data tend to exhibit better performance [752] .

It is common to develop models that integrate genomic data and medical imaging [753, 754], including hyperspectral images [755]. Other studies integrate genomes with Computed Tomography (CT) images [756, 757]. This study identifies strains of *Pseudomonas aeruginosa* by jointly analyzing the genome and morphological features in images, such as sizes, shapes, edges, textures, and degrees of opacity and color [758]. There are studies that work with genomic sequences and 3D images of the encoded proteins, for example, to analyze the role of hydrophobic nodes in the dynamics of Class A β-Lactamases [759]. Another example of cases where genomes and molecular structures are used to identify biophysical interactions between proteins and peptides [760].

Although other types of data are also integrated, such as histopathological data like biopsy results with molecular data or other omics data [531, 761–769], radiogenomics [770], metadata [771], transcriptomic and glycomic data [772], spatial transcriptomics (ST) [773–775], transcriptomic and chromatin information [776–779] multisource single-cell transcriptome data [780], and biochemical properties [781].

This study analyzes gene prognosis by integrating gene expression profile, underlying DNA sequence, and three-dimensional protein structures [782].

This development integrates five multi-scale and multi-omic modalities (RNA-Seq, miRNA-Seq, whole-slide images, copy number variation, and DNA methylation) using a late fusion strategy and machine learning techniques [783].

I identified several studies that analyze Gene-Environment (G-E) interaction analysis, meaning they work with multi-omic data to determine relationships between genotype, phenotype, and environment [784–790]. This study even identifies complex motifs in massive omics data [791]

In terms of architectures, various studies suggest that autoencoders are effective for integrating multi-omic data [792–798], similarly to Graph Neural Networks (GNNs) [799–801]. Some studies indicate that GNNs have the potential to investigate microbial metabolism due to the graph-based nature of microbial metabolomics datasets and microbe-metabolite associations [802].

3.13.2 Genome Processing and Algorithms

When applying ML and DL in the field of genomics, one of the first steps is to process the files containing the genomic sequences (character strings) and transform them into the format that best fits the architecture to be applied.

Many studies converted genomes into numerical format, for instance, by transforming those sequences into arrays of numbers, and subsequently, they may undergo a conversion into image format (or not). There are multiple conversion procedures. In this development, they first converted the genomic sequence into a tensor, and then they transformed the obtained tensor into an image [803]. This other study converted the genome into an encoded matrix [804]. In this research, a numerical conversion of the genome was also performed [805]. This one followed quantitative coding procedures, for example, performing conversions from amino acid to numerical code, aiming to compress genomic sequences [806]. This research converted genomes into images based on a cartographic strategy relying on entropy to transform high-dimensional gene expression data into a configured image format [807]. This application generated a dataset of genetic distance matrices from a

set of sequences, and using a CNN, predicted the evolution of active outbreaks, in this case, of the HIV pandemic [808]

A commonly used graphical representation involves converting genomic sequences into k-mer images, where k represents a specific length (number of nucleotides) [809, 810]. The choice between k-mer representation and spectrogram to describe a genome depends on the context of the specific task and the objectives of the analysis. Both representations have advantages and disadvantages, and the appropriate choice may depend on factors such as the nature of the data, the purpose of the analysis, and the available computational resources. Below, I describe some potential disadvantages of the k-mer representation compared to the spectrogram representation:

1. Loss of spatial information. The k-mer representation captures information about the presence and frequency of specific sequences of length k, but it does not retain information about the position of those sequences in the genome [811]. On the other hand, a spectrogram can represent the frequency distribution along the sequence, retaining spatial information [812].

2. Feature space size. As the value of k increases in the k-mer representation, the feature space can become large and sparse, which can increase computational and memory requirements [813]. Spectrograms can have a more manageable size depending on the resolution and size of the input sequence [814].

3. Sensitivity to errors and variability. The k-mer representation can be sensitive to small variations in the sequence, such as sequencing errors or point mutations, as these can significantly affect the presence or absence of specific k-mers [815]. Spectrograms can be more robust against a certain level of variability, as they represent more general patterns [816].

4. Context of information. The k-mer representation treats each k-mer independently, without considering its context or relationship with other k-mers in the sequence [817]. In contrast, a spectrogram can capture more complex patterns and relationships between different parts of the sequence [818].

5. Biological interpretation. Directly interpreting biological information from the k-mer representation can be more challenging, as it consists of specific

sequences without an explicit relationship with biological functions [819]. By representing the frequency distribution, spectrograms can provide a more intuitive and information-rich insight into the structural and functional characteristics of the sequence [355].

Other genome representations are used for applying ML, such as Chaos Game Representation (CGR) [820, 821]. As in the previous case, each system has its own advantages and disadvantages. Some specific disadvantages of genome representation with CGR compared to spectrograms are:

1. Limited Information. CGR representation has limited information about interactions between nucleotide bases and DNA secondary structure. It is a visual and geometric representation but does not provide direct information about specific DNA properties, such as the presence of specific motifs or chromatin structure [822]. Genomic spectrograms can capture more complex patterns and provide detailed information about the sequence, such as the presence of coding and non-coding regions, as well as structural features [425].

2. Sensitivity to Sequence Length. CGR technique can be sensitive to sequence length and does not scale well for very long sequences [823]. Genomic spectrograms are more robust in terms of handling sequences of different lengths and can be applied to entire genomes [412].

3. Direct Interpretation. CGR requires extensive expertise and specific knowledge to interpret the geometric features and symmetries present in the representation [824]. Genomic spectrograms allow for a more direct interpretation, as visual features (such as peaks and patterns) can be intuitively associated with specific biological properties [343].

4. Requires Additional Preprocessing. CGR may require additional preprocessing to handle DNA sequences containing additional information, such as specific gene annotations [822]. Genomic spectrograms can be more easily integrated with machine learning techniques and quantitative analysis without the need for additional preprocessing [354].

Regarding the algorithms applied, virtually all existing architectures were utilized. Some notable examples are the following. For genome processing, These studies applied LSTM architectures for bacterial genome detection [825], or for pan-cancer analysis (search for common patterns among different types of cancer) [826]. Federated Learning (FL) was also used, for example, for protein function analysis [827]. This study utilized graph transformers for drug response analysis [828].

The prevalence of research in the field of image processing relies primarily on the use of Convolutional Neural Networks (CNNs). This choice is grounded in the CNNs' ability to capture and learn relevant features from images, thereby enabling robust and effective analysis across a wide range of applications.

3.13.3 Identification and Analysis

In genomics, the application of ML and DL tools is more incipient and mostly oriented towards the search for profiles [425, 829–857].

The main objective of these applications was to determine phenotype associations [858–878]. Other objectives included identifying functional effects of DNA sequences [879],identifying epistatic feature interactions [880], genome association with specific aspects (association of mutations with diseases, high risk, etc.) [881, 882], labeling of genomic elements [883, 884], or calculation of genomic parameters and constants [885].

Research was conducted for species classification [886], including microorganisms [887–895], even engaged in the detection of novel microorganisms [896, 897].

Various studies are focused on investigating evolutionary biology and taxonomic classification. I found specialized applications in taxonomic classification [898–900], analysis of virus taxonomy [901].

An important subject of study is the analysis of genomic diversity and evolution, for example, determining function conservation across species [902], analysis of natural selection [903, 904], detection of cross-species transfer of regulatory code signals [905], analysis of dispersal inference from population genetic variation [906]. There are also research studies aimed at establishing phylogenetic proximities

between organisms [907–910]. Other applications clustered bacterial families based on the shapes of phylogenetic trees [911], or imputed distance matrices for calculating the phylogenetic tree [912]. Other developments were aimed at visualizing complex hierarchical structures [913], or for lineage tree reconstruction [914].

This article pointed out that CNNs outperformed traditional statistical methods in determining Population Genetic Inference [915].

Numerous DL applications were dedicated to expanding specific knowledge about genes. These applications focused on identifying specific patterns and features in genetic sequences, allowing for a deeper understanding of gene function, structure, and regulation, as well as their relevance in various biological processes. I found multiple studies on gene identification and analysis [916–923], gene Ontology analysis [924], evaluation of Functional Similarity (FS) of genes [925], identification of partial sweeps [926], detection of gene expression and its regulatory elements [927–941], analysis of Gene Expression Profiles (GEP) [942–946], analysis of gene interactions [947].

Other studies delved into understanding gene regulation [948–950], detecting and analyzing Gene Regulatory Networks (GRNs) [951, 952], reconstruction of GRNs [953], analysis of quantification of gene expression and surface proteins via single-cell Cellular Indexing of Transcriptomes and Epitopes (CITE-seq) [954], estimation de the expression of specific genes based on the expression of characteristic genes [955], identification of functional genomic elements [956, 957], analysis of genetic variants [958–962], analysis of Introgression [963], or detecting structural genomic rearrangements [964].

Also, there are algorithms applied to Bulked Segregant Analysis (BSA) [965], identification of Nucleotide-Binding Site-Leucine-Rich Repeat (NBSLRR) gene family [966], detection of multipartite genes [967], detection and analysis of Horizontal Gene Transfer (HGT) [968, 969], and analysis of metrics to detect and classify Transposable Elements (TEs) [970].

Another field of interest encompasses the study of mutations. Within this area, the analysis of mutational effects stands out [971, 972]. There are also developments

focused on the analysis of rare or de novo variants [973], genotypes associated with mutations [974], detection and analysis of Structural Variants (SVs) [975–978], identification of mutation patterns [979–981], epigenetic modifications [982–985], differentiation of Single Amino acid Variants (SAVs) [986], discrimination between driver and passenger mutations [987], analysis of transposon mutagenesis [988, 989], or scoring the deleterious effects of noncoding mutations [990].

A considerable number of research studies were dedicated to the analysis of cells and their associated processes, such as cell segregation [991], cell differentiation [992], Ki67 index counting [993], estimation of the cell fractions from bulk gene expression profiles [994], inference of the trajectories of cellular changes and identification of their transcriptional programs [995], detection of germline mutations and quantification of their rates [996, 997], mapping of cellular features [998], analysis of cellular decision-making [999], identification of cellular antiviral responses [1000], classification of large sample single-cell RNA-seq data [1001], clustering analysis of single-cell multimodal data [1002], analysis of single-cell immunoblots [1003], analysis of Single-Cell Chromatin ImmunoPrecipitation DNA-Sequencing (scChIP-seq) [1004], analysis of biological network inference from Single-cell Multi-Omics (scMulti-omics) [1005], analysis of single-cell chromatin accessibility in the human brain [1006], analysis of single-cell transcriptomics [1007], or analysis of T-Cell Receptor (TCR) sequencing data [1008].

Another cornerstone of knowledge is the study of proteins and their intrinsic functions. I located research on the identification of protein complexes [1009], the search of homologous proteins [1010], protein subcellular localization [1011], analysis of protein disordered regions [1012], detection of intrinsically disordered or unstructured proteins [1013], end-to-end protein domain analysis [1014], classification of adaptor proteins [1015], analysis of the subcellular location of a protein [1016], metalloproteins detection [1017], kinase-substrate prediction from phosphoproteomics data [1018], analysis of sequence-function mapping of the protein [1019], proteome mapping [1020, 1021], protease inhibitor resistance analysis,

coreceptor utilization, and body-site identification [1022], analysis of Ubiquitin-Proteasome System (UPS) [1023], identification of proteinprotein interactions [1024, 1025], or DNA-protein interactions [1026].

Several DL applications focused on the study of molecules and their biochemical interactions. There are applications aimed at identifying molecular subtypes [1027], molecular dynamics identification, analysis and simulation from genomic sequences [1028], or extraction of molecular entities and their interactions [1029].

I located ML applications for the search of specific parts of molecular pathways associated with differences in observable characteristics (phenotypes) between different groups [1030], or for Identification of B-Cell Epitopes (BCEs) [1031].

There are studies related to chromosome recognition [1032], to the detection of the presence of short Extrachromosomal Circular DNAs (eccDNAs) [1033], the analysis of allelic architecture of complex traits [1034], the detection of Variant Allele Frequencies (VAF) [1035]. Another research areas were the reconstruction of the haplotypes of an organism [1036], the identification and analysis of nucleosomes and nucleosome position [1037–1040], or the detection of ribosome stalling [1041].

The diversity of DL applications in studying very specific and advanced aspects in the genomics field is remarkable. Numerous research endeavors aim at identifying highly specific segments of DNA or RNA, which involves the precise and discriminative localization of nucleotide sequences within complex genomic or transcriptomic structures.

Numerous studies focused on the identification of promoter and enhancer sequences [1042–1061], promoters analysis [1062–1064], identification of Enhancer-Promoter Interactions (EPIs) [1065–1071], ARN editing sites detection [1072], localization of DNA-DNA replication sites [1073, 1074], replication forks detection [1075], analysis of DNA Damage Repair (DDR) [1076], and DNA repair mutation detection [1077].

Multiple models were also applied to various phenomena related to splicing, such as analysis of splicing events [1078, 1079], analysis of pre-mRNA splicing process [1080], alternative splicing [1081–1084], exon splicing [1085], splicing of

genomic exons into mRNAs [1086, 1087], and prediction of splice sites, splice junctions [1088–1095].

Various applications focused on the analysis of transcriptional regulation [1096], Histone Post-Translational Modifications (PTMs) analysis [1097], Transcription Factors (TFs) analysis [1098–1106], annotation of transcription start sites (TSSs) [1107, 1108], localization of transcriptional aberrations [1109], and aberrant splicing [1110].

Multiple algorithms were applied to locate and analyze regulatory regions [1111–1116] and predict their effects [1117], locate cis-regulatory regions [1118–1123], to analyze how representations of regulatory genomic sequences are constructed [1124], to identify cis-elements [1125], to detect Single Nucleotide Polymorphisms (SNPs) [1126–1133], and Single-Nucleotide Variants (SNVs) [1134], to filtering small genomic variants in short-read DNA sequence data [1135], feature analysis of single-cell RNA sequencing (scRNA-seq) [1136], to detect doublets formed during single-cell RNA sequencing (scRNA-seq) [1137], analysis of Long Non-Coding RNAs (lncRNAs) [1138–1154], to determine of non-coding regions [1155, 1156], to establish the relationships between non coding RNA (ncRNA) - protein interactions [1157]. There are applications focused on various aspects such as identifying lncRNAPCG associations [1158], prospection of Long Non-Coding RNA (lncRNA) - disease with omics data [1159], analysis of non-coding regulatory areas [1160], distinction of novel long non-coding RNAs (lncRNAs) from protein-coding messenger RNAs (mRNAs) [1161, 1162], circular RNA identification [1163, 1164]. Multiple applications were dedicated to the detection of specific elements such as ring boundaries [1165], stem-loops [1166], codons [1167], genetic loci [1168], binding residues [1169], CCCTC-binding factor (CTCF) [1170], acidic transcription Activation Domains (ADs) [1171], N4-methylcytosine [1172–1176], DNA 5-methylcytosine detection [1177], RNA N6-methyladenosine sites [1178–1184], G-quadruplexes (G4s) [1185, 1186], RNA G-quadruplex (rG4) [1187], malonylation [1188], regression of genome-wide DNA metyhlation [1189, 1190], patterns at DNA N6-methyladenine sites [1191–1193], DNase I Hypersensitive Sites (DHSs) [1194, 1195], and Z-DNA regions [1196].

A relevant study area was the detection and analysis of Polyadenylation sites (polyA) [1197–1204], analysis of variants in 5' and 3' Untranslated Regions (UTR) [1205], and detection of 5' splice site GT>GC or +2T>C variants [1206]. Detection and analysis of microarrays, microRNA (miRNAs) was also relevant [1207–1212] and of pre-microRNAs [1213–1216], as well as the relationship between microRNA and protein expression [1217], and the analysis of miRNAs (microRNAs) and lncRNAs (long non-coding RNAs) associations [1218].

Several studies were dedicated to analyze chromatin features and chromatin interactions [1219–1223], and chromatin loops [1224–1226], including CTCF-mediated [1227].

I found promising results in the detection of complex genetic elements, such as certain pyrimidine modifications, like 5-methyluridine (m5U) [1228].

An area of growing interest is the study of epigenomics, a discipline of biology that focuses on the study of chemical and structural modifications in DNA and the proteins associated with DNA. I found several investigations falling within this area of knowledge, dedicated to discovering combinatorial patterns of epigenetic marks [1229], classification of epigenome marks [1230].

Of particular interest is the application of DL in three-dimensional data in the field, such as analysis of 3D genome architecture [1231–1236], analysis of DNA interactions and 3D chromosome folding at the genome-wide scale [1237–1239], and identification of 3D structures in cryo-electron micrographs [1240].

Finally, I found particularly fascinating this research, which explicitly explores the concept of "regulatory grammar of DNA" [1241].

In light of the exposition in this section, although the application of DL algorithms in genomics is still in its early stages, I identified multiple developments in various aspects, identifying and analyzing highly specific and complex functions, relationships, genomic elements, etc. The ability of these algorithms to process vast amounts of data, even complex ones, suggests that these algorithms can provide efficient tools for deepening our understanding of the genome.

3.13.4 Prediction

In addition to the identification and analysis of multiple and diverse aspects of the genome, the other major application block of DL tools is oriented towards predicting genomic phenomena and gene expression.

Similar to section 3.13.3, the casuistry is very diverse and highly specific, even finding research aimed at predicting inter-omics interactions [1242].

Regarding the prediction of specific gene elements, there are models to build sequence-to-expression models [1243], to predict gene functions [1244, 1245], and to predict essential genes in diverse organisms, such as *Drosophila melanogaster* [1246], or even in the human being [1247].

These predictive models are also applied in the cellular domain, for example, for predicting cellular dynamics [1248], depletion of CD8+ T cells with omics data [1249], T-Cell Receptors (TCRs) binding specificities of the neoantigensand T cell antigens in general [1250], surface protein abundance from single-cell multimodal omics data [1251], or subcellular localization of Gram-negative bacterial proteins [1252].

There are numerous studies dedicated to the prediction of various characteristics and properties of proteins. For example, there are application developments to predict protein functions [1253–1256], including structures [1257–1260], sequence-based protein crystallization prediction [1261, 1262], analysis and prediction of non-structural proteins (nsp) [1263].

Another elements of prediction are Transport Proteins (TPs) [1264], protein mitochondrial localization [1265], ubiquitylation and sumoylation sites [1266], protein succinylation sites [1267], protein nitration and nitrosylation sites [1268], histone modifications [1269], cleaved peptides [1270], TCR-peptide binding [1271], via attention mechanisms [1272], or epitopes [1273].

This model predicted whether a chemical structure leads to a desired or adverse biological effect using multi-omic data [1274]. These inspiring studies predicted genome-wide chemical-protein interactions for still very unknown proteins [1275, 1276].

One of the most prominent examples in protein folding prediction is the case of AlphaFold and AlphaFold 2. It was one of the most successful applications, a development that enables predicting the three-dimensional structure of proteins from their amino acid sequence [1277, 1278].

There are numerous studies where both AlphaFold and AlphaFold 2 were used to predict the protein structure [1279], or the quaternary structure and stoichiometry of a protein complex [1280]. There are multiple research studies that utilized AlphaFold, for example, to analyze stress regulation in plant genomes [1281, 1282], or to analyze the antibiotic resistance of *Vibrio vulnificus* by analyzing its 3D structure [1283]. There are also studies that provide curated datasets via AlphaFold [1284], or applications aimed at optimizing its use [1285, 1286].

Furthermore, I found several alternative developments to AlphaFold [1287–1290].

Another focus of interest in the application of DL tools in genomics (especially generative models) is the prediction of genetic evolution, for example, prediction and interpretation of large-scale mutagenesis data [1291], prediction of the effects of genetic variants on splicing [1292–1295], isoform functions [1296], viral mutations [1297], or damaging missense variants [1298].

This interesting research predicts chromosomal flexibility, that is, the ability of chromosomes to adapt and change their structure in response to different environmental conditions or internal stimuli [1299].

Numerous studies are focused on predicting various specific segments of DNA, as well as their functions and multiple characteristics:.prediction of promoters [1300, 1301], fusion breakpoints, fusion events [1302, 1303], Polyadenylation Signals (PAS) and Translation Initiation Sites (TIS) [1304–1308], functions at the mRNA Isoform Level [1309], Post-Translational Modifications (PTMs) [1310], chromatin contacts [1311], chromatin accessibility [1312–1317], base-resolution chromatin immunoprecipitation (ChIP)nexus binding profiles of pluripotency Transcription Factors (TFs) [1318], integration sites [1319–1321], binding sites [1322–1335], proteinglycan interactions [1336], splice sites [1337], RNA-protein sequence and structure binding preferences [1338], RNA-binding Proteins (RBNs) [1339–1342], Transcription Factor

Binding Sites (TFBS) [1343], DNA methylation [1344–1354], N4-acetyldeoxycytosine (4acC) [1355], N4-acetylcytidine sites [1356, 1357], pre-mRNA back-splicing sites [1358], single-guide RNA (sgRNA) activity [1359], efficiency of Small interfering RNA (siRNA) [1360, 1361], lncRNA-protein interactions [1362], lncRNA-miRNA interactions [1363], G-quadruplex propensity [1364], CRISPR-Cas9 Off-target [1365, 1366], or metabolite abundances from microbe abundances [1367].

This interesting article aims to predict mRNA abundance using information from DNA sequences as well as epigenetic modifications within genes and their cis-regulatory regions [1368].

Furthermore, certain developments address other aspects of interest, such as predicting asymptomatic host adaptation [1369] or even age prediction [1370].

This research focused on the Generation of Paired Single-Cell Multiomics Profiles was highly promising [1371].

The total amount of articles focused on prediction was lower than those developments centered on identification and analysis. However, the areas of interest were similar. The applications are also varied and highly specialized.

3.13.5 Specific Deep Learning Applications in Genetic Recombination

It is noteworthy the smaller number of articles describing research related to the application of DL algorithms to the detection of genetic recombination events and/or to the analysis of this phenomenon. This model applied RNNs to detect genetic recombination events [1372].

The aim of many studies dedicated to investigating genetic recombination is the detection of hot spots. To this end, they applied multiple architectures, such as using Deep Neural Networks (DNNs) [1373], even combining genomic sequences with physio-chemical properties [1374]. 1-D convolutional layers [1375], and autoencoders were also utilized [1376].

I want to highlight this application for detecting hotspots using a phylogenetic hidden Markov model-based approach [1377].

Based on the review conducted, although there are specific applications in genetic recombination using DL, their number is limited compared to other areas of research in genomics. The phenomenon of genetic recombination is an extremely complex process involving the transfer of genetic material between different organisms or within the same organism. This process can be subject to a variety of variables and considerations, such as DNA structure, gene expression, and epigenetic regulation, which may hinder its prediction and modeling with DL techniques. Additionally, the limited availability of high-quality datasets and the challenge of interpreting results may also contribute to the scarcity of specific deep learning applications in this field.

This suggests an area of opportunity for future developments and research in this particular field.

3.13.6 Metagenomics

Metagenomics is a branch of genomics that focuses on the study of collective genetic material. That is, instead of analyzing the genome of an individual organism, metagenomics examines the genetic diversity of complete microbial communities, allowing for an understanding of the composition and function of these communities in their natural environment [1378].

The development of metagenomic research has been a direct consequence of the new next-generation sequencing technologies. The power in processing large volumes of data provided by DL has led to multiple studies integrating DL with metagenomic data. For example, for predicting next-generation sequencing depth [1379].

Metagenomics provides a new perspective on the diversity and functions of microbial communities, revealing previously unknown species and metabolic pathways [1380–1383].

ML and DL algorithms are promising tools for the analysis of metagenomic data [1384]. Specialized tools were developed for this field, highly focused on analyzing microbiomes using DL [1385–1390].

Most research on metagenomic data is highly focused on the classification or clustering of genomic fragments [1391–1397]. Within this classification or

identification, the object of study in these research studies is diverse and specific. Among the different identifications, the identification of microorganisms stands out [1398, 1399], also eukariotics [1400], and virus [1401–1404], for example, the search of short viral genomes [1405], and phage-specific proteins from metagenomic sequencing data [1406]. Interesting applications analyze Deep-Sea Sediments [1407], gut flora [1408], treatment resistances from metagenomic data [1409], or identification of phages and plasmids from metagenomic fragments [1410], new enzymes [1411], genes [1412], protein analysis [1413], Genome Disorder Prediction [1414], or even Taxonomic classification [1415, 1416], and research to extend phylogenetic trees and analyze microbial community structures [1417].

The application of Deep Learning tools in microbiome research has yielded promising results, for instance, in the analysis of plant microbiomes [1418] and human microbiomes [1419]. Multiple studies prospect the prediction of diseases and health status based on human microbiome analysis [1420–1423].

Similarly to the observations made in preceding sections, I could verify the wide variety of research aspects, also in metagenomics, such as the analysis of virophage diversity in metagenomic data [1424], biocorrosion analysis (produced by bacteria) [1425], prediction of the "Theater of Activity" (ToA) of a microbiome [1426], analysis of association between microbiome and plant phenotype [1427], identification of type II polyketides [1428], genome binning of viral entities [1429, 1430], determination of the predicted nitrogen biochemical network-related enzymes from metagenomic samples [1431], analysis of high-throughput transcriptomics data [1432], prediction of Drosha and Dicer cleavage sites [1433], analysis of Dynamic Intron Retention (IR) [1434], detection of functional elements in the genome [1435], species-specific Succinylation Site Prediction [1436], binning long reads in metagenomic samples [1437].

I identified compelling applications aimed at studying cancer using metagenomic data, such as the analysis of the heterogeneity of prostate cancer cells from the tumor microenvironment [1438].

There are also applications in precision medicine that work with metagenomic data, to characterize antimicrobial resistance using metagenomic data [1439], or antibiotic resistance [1440].

Multiple studies detected diseases through the analysis of metagenomic samples (disease gene prediction and prognosis) [1441–1451]. There are also developments characterizing neuropsychiatric disorders [1452]. Or research aimed at predicting disease progression, mortality, etc. [1453–1455].

The high dimensionality and complexity of metagenomic data currently require tools that enable the advancement of existing technologies for the efficient application of Deep Learning tools. To this end, specific algorithms were designed to optimize metagenome assembly hyperparameters [1456], applications to enhance the understanding of metagenomic data [1457, 1458], dimensionality reduction systems in metagenomic data [1459, 1460], and multi-omics [1461–1465], For data mining in omics data [1466]. This article describes a design to attempt to solve the class imbalance problem [1467]. Several studies are aimed at, for example, minimizing pruning and quantization losses in the training of compressed models to reduce computational complexity [1468], recovery of draft genomes (so called Metagenome-Assembled Genomes, or MAG) [1469], or normalization of multi-batch untargeted metabolomics data [1470]. Moreover, this study proposes an interpretability based on visualization of Self-Organizing Maps, especially tailored to metagenomic data [1471].

I want to highlight the significant amount of articles falling within the scope of metagenomics, despite it being a nascent and complex field due to the high dimensionality and variability of the data handled. There are numerous research efforts aimed at enabling the application of Deep Learning algorithms in metagenomics, especially to address the challenge of high dimensionality versus small sizes of available datasets. This implies, on one hand, that the pattern detection power of these tools, as well as their ability to process complex data, make them particularly valuable in the study of metagenomic samples. Metagenomics is still an emerging area within genomics, so the sequencing of metagenomic samples is

still limited, and therefore, the compilation of datasets can sometimes be scarce in the current state of the art. The evolution of Deep Learning tools specifically designed for application in metagenomics will likely grow proportionally to the advancement of this area of knowledge, and it is expected that in the coming years, there will be truly interesting advancements in the field.

3.13.7 Study of Diseases

Clearly, one of the main objectives of genome study is to acquire greater knowledge about various diseases. Therefore, the application of Deep Learning models to the genome also pursues this goal in numerous research endeavors. Many of them are aimed at identifying genomic elements related to the development of diseases, for example, miRNAdisease associations [1472–1476], prediction of disease-gene associations [1477–1479], measurement of polygenic risk [1480], prediction of the pathogenicity of human genome structural variants [1481, 1482], and missense variants [1483], biomarker detection [1484–1492] disease subnetwork detection [1493], association of genetic factors with disease prognosis [1494, 1495], and prediction of disease progression [1496–1501]. Some studies specifically focused on disease control through the analysis of genomic data [1502].

Especially interesting is the study of epimutations and their relationship with exposure to environmental toxicants through the increase in differential DNA Methylated Regions (DMRs) [1503].

Once again, cancer research motivates numerous studies, mainly working with multi-omics data. Primarily, identifying the relationship between genetic mutations and cancer development [1504–1507], or detecting cancer breakpoints [1508], molecular characterization of tumors [1509], characterization of tumor heterozygosity from the genome [1510], identification of cancer driver genes [1511], or detection of cancer dependencies via omic data [1512].

The usual goal is early cancer detection through predictive analysis [753, 1513–1515].

I identified developments that investigate various types of cancer in diverse aspects, for example, radiogenomic classification of brain tumors [1516], prevention and diagnosis of carcinomas [1517], prediction of disease progression in patients with Head and Neck Squamous Cell Carcinoma (HNSCC) [1518], detection of Hepatocellular Carcinoma (HCC) [1519–1521], identification of the risk factors en Esophageal Squamous Cell Carcinoma (ESCC) using omic data [1522], monitoring colorectal adenocarcinoma cells gene expression and drug-response [1523], identification of genomic variations in melanomas [1524, 1525], breast cancer analysis [1526–1530], early detection of ovarian cancer [1531, 1532] and ovarian carcinoma susceptible genes [1533], lung cancer identification [1534], detection of future Acute Myeloid Leukemia (AML) [1535], early diagnosis of Non-Small Cell Lung Cancer (NSCLC) [1536], identification of the genomic features responsible for the progression of Multiple Myeloma (MM) cancer [1537], kidney cancer analysis processing jointly histopathology imaging and proteomics [1538], analysis of cancer with unknown primary tumor [1539, 1540], even unclassified tumor types [1541].

The study of mutations associated with cancer is the subject of multiple Deep Learning applications, such as the identification and analysis of somatic mutations [1542–1553], or detection of cancer mutations [1554].

The study of metastasis is another major area of application, both for metastasis detection [1555–1558], and for distinguishing, from the genome, whether they are primary or metastatic cells through their mutational patterns [1559, 1560], predicting primary and metastatic tumor types [1561], or predicting lymph node metastasis [1562].

The identification of cancer subtypes is the purpose of numerous developments [1563–1587].

There are several studies aimed at understanding its progression [1588, 1589], or survival prediction [1590–1594]. Other factors such as measuring the risk of cancer recurrence with omics data were also studied [1595], identification of cancer-related G Protein-Coupled Receptors (GPCRs) using multi-omics data [1596], prediction of Innate Lymphoid Cells (ILCs) by transcriptional profiles [1597], cancer MiRNA

biomarker classification [1598], or analysis of the relationship between endocrine disorders and exacerbation of other diseases such as cancer, using omics data [1599].

Even this interesting research analyzes the mechanisms by which tumors evade cellular defenses using omics data [1600].

The multi-omic information for training these DL models is varied and heterogeneous. Several studies are highly interesting and promising, such as this study, which integrates three omics data, namely gene expression, Copy Number Alteration (CNA), and DNA methylation data, for breast cancer stage prediction [1601]. This application detects High-Grade Serous Ovarian Cancer (HGSOC) by integrating miRNA expression, DNA methylation, and copy number variation (CNV), which are integrated with mRNA expression data to construct a multi-omic feature space [1602].

In addition to the study of cancer, various diseases, pathologies, or syndromes were also investigated:

- Cardiovascular Diseases. Detection of Heart Failure with Preserved Ejection Fraction (HFpEF) using omics data [1603], genetic detection of Cardiovascular Disease (CVD) [1604, 1605], early detection of stroke by analyzing Serum Metabolic Fingerprints (SMFs) [1606], detection of aortic aneurysm using omics data [1607], risk prediction of cardiovascular events [1608], and prediction of intracranial aneurysm rupture risk based on multi-omics factor [1609].

- Respiratory Diseases. Detection of biomarkers associated with eosinophilic asthma [1610], and quantification of the impact of lung function on radiation pneumonitis prediction using a dual-omics analysis method [1611]

- Neurological Diseases and Syndromes. Detection of Alzheimer's disease [1612–1617], detection of diseases like Alzheimer's through the evaluation of connectivity between genes and the physicochemical properties of genes [1618, 1619], prioritization of susceptibility variants for mental disorders [1620, 1621], detection of bipolar disorder [1622, 1623], diagnosis and prognosis of Autism Spectrum Disorder (ASD) [1624, 1625], analysis of Schizophrenia Genome-Wide Association Studies (GWAS) [1626], analysis of CreutzfeldtJakob Disease (CJD) [1627],

Neonatal Opioid Withdrawal Syndrome (NOWS) detection [1628], identification of Amyotrophic Lateral Sclerosis (ALS) [1629], prediction of DNA methylation regulatory variants to detect psychiatric disorders [1630], analysis of the presence of Attention-Deficit/Hyperactivity Disorder (ADHD) [1631], and also early detection of cognitive impairment [1632]. An interesting fact is that sleep research is suitable for the application of computational approaches that can handle high-dimensional data [1633].

- Diabetes. Diabetes mellitus identification [1634], prediction of regulatory variants in pancreatic islets and refine type 2 diabetes association signals [1635], and detection of type 2 diabetes using genomic spectrograms [426].

- Reproductive Disorders. Identification of the causal genes of male infertility [1636], early detection of female reproductive disorders with omics data [1637], and prediction of In Vitro Fertilization (IVF) results with omics data [1638].

- Metabolic disorders. Identification of biomarker panels for metabolic syndrome [1639], analysis of metabolic dysfunctions [1640], and prediction of dyslipidemia using gene mutations [1641].

- Oral health conditions. Analysis of Periodontitis with omics data [1642].

Predicting disease progression is a common application of Deep Learning in genomics [1643–1654], as well as the analysis of survival-associated interactions [1655–1658], and even analysis of Polygenic Risk Scores (PRSs) [1659].

Specifically in virology, Deep Learning tools have been developed to predict the antigenic evolution of a virus [1660–1662] or variants of its proteins [1663, 1664], to predict the potential integration of a virus into a new host [1665], or to detect hosts [1666, 1667].

An interesting area of research is the analysis between a viral infection and the associated immune response, for instance, prediction of oncogenic virus integration sites (VISs) in the human genome [1668, 1669], and closely related, detection of viral genome in the host genome [1670], and detection of Virus-derived Small Interfering RNAs (vsiRNAs) [1671].

Especially noteworthy for its potential is Bacteriophage Genetic Edition [1672].

Studying diseases from a genomic perspective will undoubtedly be a significant application of Deep Learning algorithms, as already demonstrated by the numerous research studies conducted to date.

3.13.8 Precision Medicine

Many studies address advances in precision medicine with genomic data. DL algorithms, with proper development and training, can be interesting supportive tools for precision medicine [1673–1684]. Several studies indicate that ML and DL tools are promising in pharmacology [1685, 1686], and in pharmacogenomics (the analysis of genetic influence on individual response to drugs) [1687].

Instead of adopting a one-size-fits-all approach, precision medicine acknowledges the genetic and molecular diversity among individuals, allowing for the identification of specific treatments tailored to the unique characteristics of each patient. One of the fundamental pillars of precision medicine is the design of personalized treatments [1688]. Numerous studies apply DL to predict treatment and vaccine response [1689–1713]. A promising area of research is the prediction of the response to immunotherapy [1714], and the relationship between gene response to specific perturbation stresses induced by treatments [1715].

There are highly specialized researches involving highly multidimensional datasets, such as prediction of clinicopathological features [1716], prediction of Human Cytochrome P450 (CYP450) enzyme inhibition using omic data [1717], elucidation of the relationship between insulin resistance and multiomic characteristics, including microbiome data [1718], prediction of the response to intravenous and subcutaneous immunoglobulin in inflammatory myopathies [1719], drug response prediction integrating somatic mutation, copy number aberration and gene expression data [1720], identification of pathobiology and drug repurposing in Alzheimers disease con multi-omic data [1721, 1722], Circular RNA (cRNA) drug sensitivity association identification [1723], and identification and functionally assessment of pharmacovariants in whole exome sequencing data [1724].

The development of new and improved personalized treatments is one of the most promising and heavily researched applications. In this area of investigation, I found multiple DL applications in the analysis of drug-protein interactions [1725], prediction of drug-target interactions [1726–1737], drug response [1738–1745], drug-drug interactions [1746], multi-drug synergies [1747, 1748], analysis of drug combinations [1749, 1750], drug federation [1751–1754], drug resistance prediction [1755–1762], even applied in HIV-1 [1763], and toxicity prediction [1764].

The design of personalized treatments to combat a complex disease like cancer is a priority area of research, for example, in identifying cancer treatments based on the analysis of genomic sequences [1765–1767], the classification of multidimensional genomics data into their similar annotation group for medication dosage control in cancer patients [1768], prediction of phenotypic drug response on cancer cell lines [1769, 1770], analysis of 3D architectures of structured omics data for cancer treatment design [1771, 1772], and the design of molecular structures with a desired pharmacophore, integrating three-dimensional molecular information [1773].

Several studies describe the exploration of new findings, such as discovering de novo drug-like molecules with reinforcement learning [1774], discovery of putative neoantigens [1775], and discovery of new Antimicrobial Peptides (AMPs) [1776].

In my opinion, one particularly promising area is the study of an individual's immune response. Especially interesting research areas include antibody discovery [1777], optimization of therapeutic antibodies [1778, 1779], and Structure-Based Drug Design (SBDD) targeting nucleic acid macromolecules, particularly RNA [1780].

Another particularly promising (and prioritized) area is the study of resistance to chemotherapeutic agents, such as antibiotic-resistant bacteria. There are genuinely interesting DL applications in the detection of Type IV Secretion Systems (T4SS) [1781], characterization of bacteriophages [1782], analysis of the transcriptional/-translational regulatory network of Hfq in the bacterium *Pseudomonas fluorescens* to analyze its adaptation to the environment, with omics data [1783], and analysis of *Alternaria*-host interactions [1784].

Also of high interest is the prediction of Minimum Inhibitory Concentrations (MIC) of antimicrobials [1785], or the prediction of Antimicrobial Peptide Generation and Activity, jointly analyzing sequences and experimental Minimum Inhibitory Concentration (MIC) data [1786].

Other high-value added applications include the study of the relationships between histone modifications and Embryonic Stem Cells (ESC) fate decision [1787].

Also within precision medicine, the identification of robust disease and risk factor signatures from omics data stands out [1788], or the determination of relationships between cis-regulatory elements and addictions [1789].

There are also research efforts aimed at improving medical management through genomic and multi-omics data, for example in pharmacogenomics, to address discrepancies in the input-output space between preclinical and clinical datasets [1790], patient stratification [1791–1795], analysis of epidemiological association patterns [1796], prediction of disease progression [1797, 1798], etc.

And gene silencing analysis is also very interesting [1799], with broad potential applications in gene therapy.

In light of the state of the art in DL developments in precision medicine, the high interest of the field itself is evident. Furthermore, given the complexity and multidimensionality of the data involved, this field may enable optimal utilization of the potentials of these algorithms.

3.13.9 Crop Improvement

An interesting Deep Learning application popular in the field of genomics and transcriptome analysis is its use for crop improvement, starting from the study of plant genomes [1800–1803]. Successful developments in this field can enable detailed study of the genetic profiles and gene expression of botanical species. This can provide valuable information about desirable agronomic traits, such as disease resistance, tolerance to adverse soil or climate conditions, crop yield, and nutritional quality. This, in turn, can enable the development of more precise and efficient genetic improvement strategies, leading to the production of better

adapted and more productive plant varieties, which can have a significant impact on food security and agricultural sustainability.

I came across very innovative studies specialized in measuring abiotic stresses in plants [1804, 1805].

Probably, given the importance of the field in terms of climate change and increasing food demand, advances in this field will be notable in the coming years.

3.13.10 Improvement of Existing Technology

In addition to all these applications aimed at deepening biomedical knowledge through ML and DL models, multiple research efforts are clearly oriented towards improving existing technology.

There are various applications aimed at optimizing the integration of multi-omic data. Given the importance of integrating heterogeneous high-dimensional data, there are multiple studies proposing integration strategies [1806–1816]. This article proposes a dynamic integration framework mindful of uncertainty for multi-omic classification [1817]. This study integrates various multi-omic data from different databases and identifies the most suitable optimal design for the available data [1818]. I identified tools for multimodal genomic data analysis [1819]. There are DL applications that even suggest and locate patterns in the genome, for example, regarding elements encoded within it [1820]. There are algorithms specifically designed to extract intrinsic biomedical information from high-dimensional omics data [1821], also to streamline multi-omics data processing technologies [1822–1825], and algorithms for compressing genomic sequences [1826, 1827], proposals for DNA compaction mechanisms [1828], improvement of prime editing efficiency [1829], improvement of accuracy, robustness, and speed of scRNA-Seq data processing [1830], and development of noncoding variants functional prioritization methods [1831].

There are studies capable of predicting phenotypes from compressed genomic and omics data [1832, 1833]. There are genome conversion systems to vectorized representations of genomic features [1834].

Multiple algorithms for reducing the dimensionality of omics data were developed [1835–1837] or for identifying high-dimensional mediators, especially useful in the utilization of multiple omics data [1838]. This interesting development specialized in predictive model building for omics data, improving the issue of the small number of samples (n) versus the large number of features (p) [1839].

A framework for novel autoencoder-assisted cancer subtyping is presented, which utilizes the compressed latent space of a sparse autoencoder for multi-omics clustering [1840].

There are also DL algorithms applied to improving base editing in genomic editing [1841], algorithms for noise removal of attribution [1842], algorithms for denoising, imputation, and embedding ability in omics data [1843], and in scRNA-seq [1844].

This development allows for the determination of the quality of mass spectrometry images [1845]. This one is aimed at detecting sequencing errors [1846–1848], and this research proposes a system for reducing variability [1849].

There are studies that conclude that neural network models are less susceptible to the presence of missing data, which is a notable advantage compared to other technologies [1850]. However, although it is an advantage in principle, the issue of missing data is still not fully resolved. For this reason, there are research efforts aimed at addressing this problem. This study proposes an imputation method to solve the issue of missing points [1851, 1852]. This one proposes a methodology to reduce the impact of missing discrete data imputation in the genome [1853]. Several DL models attempt to complete those missing data [1854], even using GAN networks [1855].

Repeatedly, I located multiple studies attempting to solve very specific issues in the genomics field, such as resolving gaps in draft genome assemblies [1856], predicting missing amino acids in the scaffold in the forward and reverse directions [1857], batch correction model for scRNA-seq data [1858], batch effect removal when working con metabolimic data [1859], sequence-specific bias correction [1860], controlling quality of Next-Generation Sequencing (NGS) data [1861], reduction of false positives caused by Out-Of-Distribution (OOD) sequences in metagenomic

sequence classification [1862], adjustment of cycling patterns in the unspliced-spliced RNA space of cell cycle-related genes [1863], optimization of C-to-G base editors [1864, 1865], tandem mass spectrum prediction tool (it allowed for the detection of unidentified peptides better than mass spectrometry and its spectral libraries) [1866], proposal of a multimodal analysis framework for single cells [1867], a computational method for inferring drug-induced proteomic patterns [1868], design of CRISPR gRNAs (CRISPR guide RNA) [1869, 1870], and measurement tools for uncertainty estimation to determine the reliability of the results obtained from the application of Deep Learning algorithms [1871].

There are also studies aimed at improving the application of DL algorithms to multi-omics data themselves. For example, this one analyzed the benefits of using stratified loss functions to work with heterogeneous data [1872]. This one proposed methods of DL and ML with Recursive Feature Elimination (RFE) [1873]. This article proposed a simplification in the training of custom CNN or hybrid CNN-RNN models on genomic data [1874]. Algorithms specially designed for unlabeled genomic datasets, even at large scale, were also developed [1875, 1876].

Specific libraries were developed to apply deep learning to genomic data [1877–1879], highlighting a library for motif discovery in large sets of biopolymeric sequence data, optimizing the required computation [1880]. This also enabled the handling of multi-omics data achieving high computational speeds [1881]

There are also web platforms to facilitate the application of DL and ML with genomic data [1882], or to locate omics data [1883]. Specific DL algorithms for integrating genomic databases are available [1884–1887]. Curated datasets were prepared for genomic data research [1888, 1889], there are also protocols for transferring genomic data via DL [1890, 1891]. I found interesting the implementation of algorithms guided by scientific literature for application [1892].

And particularly interesting is the application of spatial omic and pathomic technologies to obtain information about the structural compositions of such data [1893, 1894], as well as deciphering patterns in the structural and functional anatomy of genes [1895].

Although the application of DL in genomics is still in its early stages and, as I outlined in Section 3.15, the current nature of DL algorithms poses challenges for their efficient application in genomics, advances in the development of techniques, methodologies, and tools that address these obstacles will minimize or at least reduce the current limitations of these tools for efficient application in this field.

3.14 Interpretability of Biosanitary Deep Learning. State of the Art

Interpretability in Deep Learning can be defined as the ability to understand and determine why a DL model makes a particular decision. This is especially important when interpreting the results obtained from the model. In the case of the biosanitary field, given the importance of the sector and the potential consequences involved, it is even more critical.

One of the major weaknesses of DL systems is the lack of transparency in decision-making [1896]. Many studies validate that interpretability tools can provide information on areas that are determinants of subsequent classification [223, 1897, 1898]. However, the current state of the art does not allow for a complete understanding of the reasons behind DL model decision-making. We can only obtain partial information. Technically, we only know where it looks but we don't know what it sees. Derived from the need to address this deficiency, specifications have been prescribed where the requirements that a system must meet to be considered interpretable are defined [1899].

I found some examples of interpretability in medical images, for instance, in tumor detection/classification [1900], for subject classification and pathogeny identification with an interpretability system [1901], for distinguishing between images of healthy individuals or those with different pathologies [1902], or by applying interpretability systems to segment medical images to calculate interpretable features [1903]. Grad-CAM tools are popular in discriminating relevant features in images [1904, 1905].

Explicit explainability tools are being developed for handling biomedical data [1906].

Added to the above, if the application of DL in genomics is still in its infancy, even more so is the effective application of interpretability tools, despite the importance of these tools for extracting knowledge [1907].

Therefore, advanced interpretability tools are necessary to determine the decision-making process of these systems [1908–1910].

The literature on interpretability in DL systems applied to the biosanitary field is very scarce. I found several research studies working with heat map calculations [1911–1914], or with correlation maps [1915]. This article describes a specific explainability system for graph convolutional neural networks in biomedical applications [1916].

The peculiarities of genomics, especially regarding the special typology of the data, certain particularities of the genome itself, and the limited degree of existing knowledge, lead to the need to develop interpretability tools specifically for this field. All of them with relative success regarding the purpose for which they are designed. There are studies designing explainability systems for genomic deep neural networks [1917–1920]. This one proposes an explainability methodology applied to the detection of relevant genes [1921]. This research contributes interpretability by constructing morphological genes [1922]. This one provides methods that compare the activation of each neuron with a reference activation and assigns scores based on the difference between both activations [1923]. This development allows visualizing the inferred characteristics and compares them with gene expression patterns determined by immunohistochemistry [1924]. This research addresses interpretability through the use of a gradient boosting approach [1925]. This development claims to enhance the original Integrated Gradients method by utilizing RNA splicing prediction [1926]. There are highly specific interpretability developments; this study proposes an interpretability system for epistasis detection based on sparse transformers and assigning attention scores [1927]. Several studies apply Shapley Additive Explanations (SHAP) [1928, 1929]. This one utilizes SHAP values as a

unified measure of feature importance, which are calculated using game theory [1930]. This combines automatic feature extraction with analysis of significant motifs and an analysis of filter activities and feature activations within the model [1931].

Several developments obtain interpretability data through embedded attention mechanisms [1423, 1932–1934]. This development applied an attention maps-based approach [1935]. Here, the researchers propose a method of induced self-attention that provides information about which specific features are contributing to the classification of an instance [1936].

There are studies that even employ multiple interpretability tools as a measure of redundancy and to obtain additional information. For example, this study analyzes reaction coordinates of alanine dipeptide isomerization and applies interpretability tools such as LIME and SHAP [1937].

In this implementation, researchers develop an express neural network for interpretability where each node has a molecular equivalent, such as a protein or a gene, and each edge has a mechanistic interpretation, such as a regulatory interaction along a signaling pathway [1938]. Other developments use autoencoders to enhance interpretability [1939].

In high-dimensional datasets, achieving effective interpretability is an even greater challenge. I identified several studies working on the development of these tools, integrating heterogeneous data. This one combines multimodal imaging and genomics to analyze Mechanisms of Brain Cognition and generates interpretable activation maps to quantify pixel-level contributions of input features [1940].

Explainable models for cancer classification are emerging [1941]. This model integrates feature selection, causal inference, and genetic-imaging data analysis and shows hotspots in brain images [1942].

As I already mentioned in Section 3.6, within the widely studied field of interpretability in images, three tools stand out that have shown solid results in detecting the relevant areas of the image for the decision-making of the neural network:

1) Grad-CAM (Gradient-weighted Class Activation Mapping) is a technique for detecting the most important areas of an image so that the Deep Learning

tool can classify it into a certain category. Using gradients of the classification score with respect to the final convolutional feature map [225, 226]. Grad-CAM is intuitive, but its resolution is low.

2) LIME (Local Interpretable Model-agnostic Explanations) is an interpretability tool that computes feature importance segmenting an image into several features and generating "artificial" observations by randomly including or excluding these features. If a concrete feature is excluded, the pixels corresponding to this feature are replaced with the value of the average image pixel. The network classifies these synthetic observations, and uses both the absence or presence of a feature and its corresponding scores for the predicted class to train a regression problem with a simpler model [1943, 1944]. Its accuracy can vary but it exhibits good visual interpretability.

3) Gradient Attribution points out with high accuracy the specific pixels of the image which are critical for the decision making of the Deep Learning tool. Returning a pixel map that highlights the most affecting pixels of the class score when changed [1945, 1946]. Its accuracy is very high; however, visually, it is very noisy.

In terms of accuracy and ease of use for the human eye, the advantages and disadvantages of these three tools make them particularly interesting to work with simultaneously, enabling the comparison of results and obtaining complementary data on features and patterns.

There are review studies and comparative analyses of interpretability in medical imaging that claim Grad-CAM showed the highest validity in medical image interpretability, as well as good results in robustness [1947]. This fact is important for the proper selection of the appropriate tool for the specific target task. This article validates Grad-CAM as a valid technique for interpreting results in genomics [1948].

Although the "black box" nature of neural networks is one of the main impediments to their reliable application in the biosanitary field, it has not yet been possible to develop a complete and robust system that unequivocally explains the entire decision-making process of these networks. The lower number of identified articles that explicitly focus on interpretability or explainability compared to other search topics is a symptom of this fact. And this underscores the need to actively

seek solutions to the complex problem of understanding how these neural networks work. If we do not overcome this hurdle, the contribution of Deep Learning tools in the scientific field will be severely compromised.

3.15 Critical Analysis of the State of the Art

In this Section, I present the results of the identification and treatment process of all articles identified as problematic according to the criteria established in section 3.8. Additionally, conclusions derived from various critical reviews on the application of Machine Learning (ML) and Deep Learning (DL) algorithms in the biosanitary field, including genomics, are included. From these analyses, relevant conclusions are drawn that may be useful for the development of DL applications recognized by the scientific community.

I also analyze the current limitations of Multiple Sequence Alignment (MSA) algorithms in processing large-scale genomic data, either due to a high number of sequences, a large number of base pairs, or both.

3.15.1 Multiple Sequence Alignment

In recent years, advances in sequencing technologies and the rise of metagenomics have revolutionized the field of genomics, enabling the rapid and cost-effective generation of vast amounts of genomic data [1949, 1950]. These technologies have a significant impact on the field of metagenomics, which involves the study of genetic material extracted directly from environmental samples [1951].

The recent major technological advancements in genomics are aimed at the massive acquisition of genomic data. However, at this time, there are no effective tools for the management and processing of large volumes of information.

Multiple Sequence Alignment (MSA) algorithms are complex, computationally expensive, and do not allow the handling of large amounts of data; therefore, they provide suboptimal solutions [1952].

MSA methods and algorithms involve heuristics, approximate methods and/or complex statistical approximations. So they often require intricate manual adjustments performed by experts [1953]. Typically, well-known multiple sequence alignment algorithms like ClustalW or MUSCLE exhibit a time complexity proportional to $\mathcal{O}(N^2L)$, where N represents the number of sequences and L denotes the average sequence length. In simpler terms, this means that the computational time required for alignment grows quadratically with both the number of sequences (N) and the average length of the sequences (L). For example, aligning 40 sequences, each with an average length of one million nucleotides, can consume several hours or even days to complete.

Multiple alignment is an NP-hard problem (Non-deterministic Polynomial-Time Hard), which means that there is no known efficient algorithm that can find the optimal solution in a reasonable time for all instances.So, due to the computational complexity of the problem, multiple sequence alignment algorithms often provide suboptimal solutions. To achieve the optimal solution, it is essential at this point to perform manual adjustments letter by letter. These adjustments are based on expert knowledge.

The different alignment algorithms have different penalties for introducing gaps since adding those gaps implies introducing information with no added value [1954]. These algorithms exhibit varying degrees of permissiveness regarding gap introduction. For instance, the Smith-Waterman alignment algorithm is more flexible in terms of the amount of gaps you can introduce compared to the Needleman-Wunsch algorithm [1955].

For a better understanding of the potential impact of introducing gaps, two excerpts from the HIV-1 Complete Consensus Multiple Sequence Alignment from the Los Alamos National Laboratory HIV-1 Database (until 2020) are shown [41] (Figures 3.8 and 3.9).

In Figure 3.8, it is visible that, being coding regions, to comply with the triplet rule, gaps are introduced every 3 positions. This fragment shows a fairly



Figure 3.8: HIV-1 Complete Consensus Multiple Sequence Alignment Excerpt 1. This excerpt is located in a region around the end of the *gag* gene and the beginning of the *pol* gene. This is a clearly coding region. Therefore, gaps are introduced in triplets, given the conversion of 3 nucleotides into an amino acid.

homogeneous genomic region, where the total number of gaps is lower than the nucleotide count.

However, as seen in Figure 3.9, the number of gaps can exceed the actual content of the genomic sequence.

The introduction of gaps in aligned sequences involves the insertion of empty positions, which disrupts the congruence with the actual nucleotide content. This phenomenon results in an increase in the size of the resulting dataset from the alignment, leading to greater computational demand for its manipulation and analysis.

Additionally, the insertion of gaps contributes to amplifying the noise in the dataset. When gaps are introduced into aligned sequences, empty positions are added, which can introduce a certain level of uncertainty or noise into the data. This is because each gap represents a lack of information about the sequence at that specific position, which can affect the alignment accuracy.

Each gap introduces additional uncertainty about the homologous relationship between the sequences. The appearance of a gap at a specific position can generate

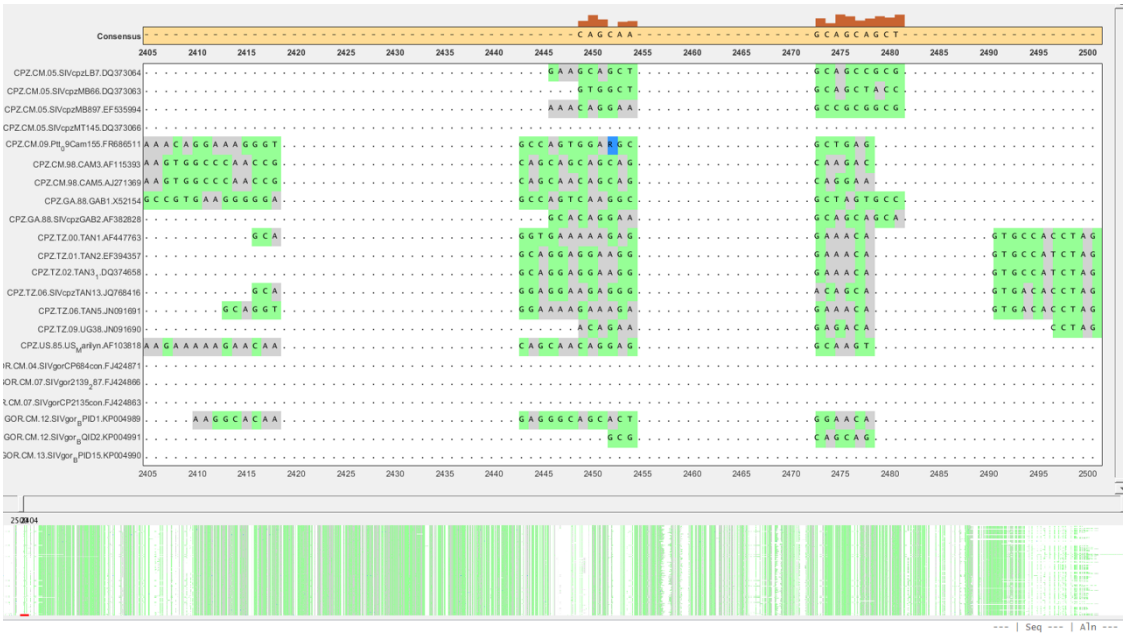


Figure 3.9: HIV-1 Complete Consensus Multiple Sequence Alignment Excerpt 2. This excerpt is located in the vicinity of the 5'LTR or the beginning of the *gag* gene.

uncertainty about the evolutionary relationship between the aligned sequences in that region since it is not possible to unequivocally determine if that gap is the result of a true insertion or deletion of nucleotides/amino acids at that position or if it is simply a random coincidence.

In cases where the number of gaps exceeds the nucleotide content, these issues are compounded to the extent that, as can be seen in Figure 3.9, it becomes extremely complex to even visualize the results.

As the number of sequences subject to alignment increases, there is an increase in the number of gaps introduced, and consequently, an increase in the likelihood of encountering the aforementioned issues.

The development of sequencing technologies has enabled the sequencing of large amounts of genomic data massively and obtaining sequences with increasingly longer lengths.

Multiple sequence alignment, which involves the simultaneous comparison of multiple genomic sequences, becomes increasingly complex as the quantity and length of sequences to be aligned increase. This complexity can be particularly challenging when working with massive sequencing data, as the volume of data

can be overwhelming and the required computational processing capacity can be prohibitive. Moreover, the increased length of sequences can raise the likelihood of alignment errors and the introduction of incorrect gaps.

Therefore, the computational demands of biotechnological algorithms and the current limitations of commonly accepted biotechnological applications often force researchers to work with few sequences and/or fragments thereof.

This entails an inherent loss of information in a still unexplored environment. That is, relinquishing or selectively choosing data in an environment where we are unaware of their importance leads us to work with biased, incomplete information, to the likely loss of crucial information, to the generation of probably biased results, and to the need for repeated experimentation and processing to address the aforementioned shortcomings.

However, the state of the art is not enough to face the current challenges, like the interpretation of the complete human genome [1956] or the sequencing of almost 2 million SARS-CoV2 sequences [1957]. With the increasing amount of sequencing data, new challenges emerge in the field of bioinformatics, particularly in the area of multiple sequence alignment [1958]. Assembling and aligning large and complex genomes, as well as metagenomic datasets, is a computationally intensive task that requires sophisticated algorithms and high-performance computing resources. Moreover, these datasets often contain errors and variations that make it challenging to achieve accurate alignments.

Multiple alignment techniques and algorithms are statistical approximations that require complex expert manual adjustments [1953]. In general, the most common multiple sequence alignment algorithms, such as ClustalW or MUSCLE, tend to have a time complexity on the order of $\mathcal{O}(N^2L)$, where N is the number of sequences and L is the average sequence length. That is, the execution time will increase quadratically with N and L. This means that a multiple sequence alignment of 40 sequences with an average length of one million nucleotides can take hours or even days to complete.

There is still no efficient solution for the analysis of large numbers of sequences and no objective method has been approved by the entire scientific community. The best approaches are to apply several algorithms, select their common results, and then perform manual adjustments based on expert microbiological knowledge.

This study even proposes the application of DL models to perform multiple sequence alignments [1959].

While multiple sequence alignment remains a fundamental tool in bioinformatics for comparing genomic sequences and revealing evolutionary relationships, it has certain limitations. These limitations include its computational inefficiency in handling large datasets and sequences of considerable length, as well as its susceptibility to alignment errors in regions of highly variable sequences or with large insertions or deletions.

Considering the increasingly larger size of current sequencing, especially if it involves complete genomes, it is evident that the current limitations of multiple alignment systems will increasingly hinder effective sequence comparison. Therefore, it is imperative to urgently overcome the current limitations in this area. The entire human genome is composed of approximately 3 billion DNA nucleotides. The genome of the long-tailed lizard *Anolis carolinensis*, which is a species commonly used in research, has approximately 1.8 billion DNA nucleotides. The genome of *Saccharomyces cerevisiae*, which is a fungus commonly used in research and food production, has approximately 12 million DNA nucleotides. The bacterium *Salmonella enterica* subsp. Enterica Serovar Typhimurium strain, which is a common cause of food poisoning, has a circular genome approximately 4.9 million nucleotides in length [39].

To address these limitations, alignment-free methods have been developed, allowing for a faster and more efficient comparison of genomic sequences. Additionally, alignment-free methods may be more robust in analyzing highly divergent or fragmented sequences.

3.15.2 Deep Learning Applied to Biology and Medicine

Apparently, the emergence of ML and DL in the biosanitary field is abundant and successful. There are even studies that consider DL as a valuable tool for post-Genome-Wide Association Studies (post-GWAS) analysis [1960]. However, analyzing in detail from an expert point of view, it is not as promising as it seems. There are more critical studies that highlight the risk of biases in apparently successful DL applications in the biosanitary field. For example, in a critical study of 731 applied models, 545 had a high risk of biases. The main causes were inadequate sample sizes and inappropriate or incomplete model performance evaluation [37].

In my opinion, in order to effectively apply DL algorithms, a series of unavoidable premises must be met. These algorithms only function properly when trained with thousands of data points. Additionally, in my view, the data should be balanced, with sufficient samples from all categories under study. Again, in my opinion, these algorithms should be trained with all applicable scenarios related to the case study, even if they are not directly related. This is important to ensure the absence of biases in training and to have reliability in the results. It is essential to deeply understand the representativeness of all the data used. Working with data solely from certain targeted categories is not advisable.

Figure 3.10 shows a very simple yet very illustrative example of the different scenarios that must be considered in the development of ML and DL applications.

Let's consider a DL development dedicated to bird identification. In the dataset design itself, we must include sufficient data of all possible postures and appearances of a bird, including images where nothing more than a more secondary, less representative part of its body is visible. This way, it will be more likely to ensure the correct functioning of the model for objective performance. Ultimately, the result will be more robust and reliable.

Therefore, if these conditions are not met (thousands of balanced data points including all scenarios), we cannot guarantee the proper functioning of these DL algorithms, and clearly, I must advise against their use.



(a) This is a bird **(b)** This is also a bird

Figure 3.10: Visual example of scenarios contained in a dataset. Photographs taken by Ana Guerrero Tamayo [1961].

I endorse these statements and complement them with the findings indicated in the study conducted by Whalen et al. [1962]. In their analysis, this research team detected errors in various ML applications within the genomics field. These researchers warn about everything previously mentioned and also add certain additional flaws such as errors derived from dependent variables. In ML and DL, each category is presumed to be independent of the others. In genomics, this may not be the case. Simply put, we do not know. Therefore, this can lead to fundamental errors in training that may remain hidden beneath seemingly successful results. The current lack of understanding of most genome mechanisms and meanings will likely result in the omission of variables whose function is relevant to the case study in training. This can introduce statistical errors, especially if a variable influencing the relationship between the independent variable (cause) and the dependent variable (effect) is not considered or controlled.

In this review, I identified multiple studies that suffer precisely from all of this. In Section 3.8, I outlined the criteria under which I considered a study problematic. Applying them to the selection of identified raw articles, I detected hundreds of articles with suspicious or directly incorrect methodologies. To graphically demonstrate the impact of so many articles, I want to explicitly cite, case by

case, each and every article that suffers from an excessively small dataset (less than 1,000 samples) to convey even graphically the magnitude of the problem. By presenting them individually, a more comprehensive and detailed appreciation of the quantity and nature of these articles is achieved, as opposed to merely mentioning their existence in a general manner. Thus, the reader can fully comprehend the volume of questionable sources encountered.

In this case, 980 samples [1963]. In this one, 939 samples [1964]. In this one, 935 data points [1965]. In this other one, 925 samples [1966]. This one conducted experiments on 915 patients [1967]. These two studies worked on 913 patients [1968, 1969]. This one worked with 878 images [1970]. This one with 875 samples [1971]. This one with 870 samples [1972]. This one had a dataset of 842 samples [1973]. This one analyzed 839 subjects [1974]. This one, 839 individuals [1975]. This one operated with 836 patients [1976]. This one dealt with 828 pathogen samples[1977]. This one, with 817 patients [1978]. This one, with 770 samples [1979]. This other one, with 769 patients [1980]. This one, with 768 instances [1981]. This other research processed 761 patients [1982]. This model worked with 746 images [1983]. And this one, with 745 sequences [1984]. This study was conducted with 741 images [1985]. This one, with 733 images [1986]. This model worked with 730 microscope images [1987]. This one, with 724 data points [1988]. And this one, with 720 [1989]. This other study handled 702 samples [1990]. And this one, 702 images [1991]. In this case, researchers analyzed 696 subjects [1992]. This one used 696 records [1993]. This research processed 681 data points [1994]. This model worked with 679 images [1995]. This one with 664 patients (survival prediction in ovarian cancer) [1996]. In this case, the model learned from 660 images with 33 different genera and species of bacteria [1997]. This one, with 653 images [1998]. This one, with 650 [1999]. This other model, with 634 data points [2000]. This one with 628 Metagenome-Assembled Genomes (MAGs), it was used to identify phages carrying Antibiotic Resistance Genes (ARGs) and predict their hosts [2001]. This one, with 603 samples [2002]. And this one, with 600 data points [2003]. This research worked with 583 patients [2004]. This one, with 581 samples [2005]. This

model had a training set of 550 samples [2006]. The dataset used in this research consisted of 549 images [2007]. This one processed 548 molecules [2008]. This one, with 540 sequences [2009]. This one, with 536 data points [2010]. This study analyzed 553 sequences [2011]. This one, 528 patients [2012]. This study worked with 524 subjects [2013]. This one, with 522 patients [2014]. This one, with 516 patients [2015]. This algorithm trained with a dataset of 510 data points [2016]. This other one, with 506 HSI images from 20 pigs, annotated with a total of 19 classes [2017]. This one, with 505 samples [2018]. This one, with 502 diagnostic slides [2019]. This study processed around 500 images per class [2020]. This one worked with 500 images [2021]. This one, with 499 patients [2022]. And this one, with 494 patients [2023]. The dataset of this study consisted of 480 samples [2024]. This one analyzed 469 subjects [2025]. This one, with a total of 450 samples [2026]. And this one, with 442 data points [2027]. This one worked with 437 subjects [2028]. This one, with 434 human viral genomes from 187 species [2029]. This one, with 424 mitochondrial proteins [2030]. This one analyzed 423 colorectal cancer patients (to predict metastasis) [2031]. This study processed 420 samples [2032]. This one, 420 samples [2033]. This other one, 417 patients [2034]. This model operated with a dataset of a total of 415 photos of primary melanoma tumors from different patients [2035]. This one worked with 409 samples [2036]. This one, with 405 subjects [2037]. In this case, the dataset contained a total of 404 samples [2038]. This one, with 401 subjects [2039]. In this case, the training set contained 399 samples [2040]. This study to predict microbiomes was trained with 384 individuals [2041]. This one, with 378 samples [2042]. This other one, with 373 samples [2043]. And this one, with 368 samples [2044]. This research worked with 352 patients [2045]. This one, with 351 patients [2046]. This one, with 351 patients [2047]. And this one, with 346 patients [2048]. This model had images from 345 patients [2049]. This study handled 333 metagenomic samples [2050]. This one analyzed 326 gastric cancer patients [2051]. This algorithm worked with 322 images [2052]. This one, with 321 echocardiograms [2053]. This study worked with 316 patients [2054]. This one, with a total of 300 genomic sequences, 100 per category [2055]. This one, with 300

patients [2056]. This one, with 295 data points [2057]. This model operated with a total of 294 images [2058]. This one, with 288 patients [2059]. This one, with 286 samples [2060]. This study analyzed 274 human subjects [2061]. The dataset of this model consisted of 272 labeled samples [2062]. In this case, 260 samples for metagenomic data classification [2063]. And in this case, 253 [2064]. In this case, the training dataset contained 250 bp reads [2065]. This study analyzed 248 patients [2066]. This processed around 246 samples [2067]. The dataset of this model consisted of 233 images [2068]. This study processed 233 genomes [2069]. This one worked with 232 patients [2070]. This one, with 230 samples [2071]. In this case, the dataset consisted of 228 complete genomes, 76 of each of these three viruses: SARS-CoV-2, SARS-CoV, and MERS-CoV [2072]. This one worked with 226 samples [2073]. This one, with 224 patients [2074]. This one, with 224 data points [2075]. And this one, with 221 subjects [2076]. This one handled a total of 220 data points [2077]. This one analyzed 220 patients [2078]. This research, 218 patients [2079]. And this one, 217 cancer patients [2080]. The dataset of this study contained 212 images [2081]. This consisted of 211 Computed Tomographies (CT scans) [2082]. In this one, the dataset consisted of 210 images [2083]. This one processed 210 samples [2084]. This research used 209 Stimulated Emission Depletion (STED) microscopy images [2085]. This one, 208 samples [2086]. And this other, 207 images [2087]. This algorithm analyzed a total of 200 images [2088]. This one operated with 197 genomes [2089]. And this one, with 188 samples [2090]. This study analyzed 186 participants [2091]. This one worked with 179 samples [2092]. The training set applied to this model consisted of 170 images [2093]. This one worked with 169 cases [2094]. In this other study, the dataset consisted of 160 videos [2095]. This one processed 160 samples [2096]. This one, 157 samples [2097]. This one, 154 samples [2098]. This model processed 151 cases [2099]. This other one analyzed 146 patients [2100]. In this one, the dataset consisted of 144 images [2101]. The function of this model is to predict the evolution of a glioma. It was trained with 142 glioma patients [2102]. This study was conducted with 140 patients [2103]. This one, with 131 images [2104]. This other one also worked with 131 data points

[2105]. This one, with 129 samples [2106]. This one, with 123 samples [2107]. This one processed 123 cases [2108]. This one, 122 images [2109]. In this study for the detection of multidrug-resistant *Klebsiella pneumoniae*, only 121 strains were worked with [2110]. This development processed a dataset of 121 genomic data [2111]. This one worked with 120 complete genomes from metagenomic samples [2112]. The dataset of this other model consisted of information from 117 patients [2113]. This one worked with 116 cases [2114]. This one, with 115 patients [2115]. This one, with 114 samples [2116]. This one with a set of 133 experimentally tested functional and non-functional microRNA-gene interactions [2117]. This one, with 112 data points [2118]. This one processed 111 videos [2119]. This one, 111 patients [2120]. This experiment involved 110 patients [2121]. This other one, 105 patients [2122]. This study classified the presence of cancer. Its dataset consisted of 50 cancer-positive genomes and 50 cancer-negative genomes, totaling 100 genomes [2123]. This one processed 99 genomes [2124]. This one worked with a dataset containing images of 98 patients [2125]. This one worked with a dataset of 90 cases (72 cases as the training set and 18 as the test set) [2126]. In this case, the dataset consisted of 90 images [2127]. In this case, the training set was composed of 83 observations [2128]. This research encompassed a total of 78 patients (55 patients and 23 subjects with Parkinson's syndromes) [2129]. This one worked with 74 patients [2130]. These two studies worked respectively with 73 images [2131, 2132]. These, with 72 patients [2133, 2134]. This study analyzed 65 patients [2135], all of them were hospitalized, thus there was no data available on mild cases. This one worked with 64 samples [2136]. This one, with 62 bone marrow samples [2137]. In this case, the dataset consisted of images from 60 patients [2138]. In this case, the training set consisted of 60 images [2139]. This model processed 59 samples [2140]. This one, 52 cases [2141]. And this one, 52 patients [2142]. This study consisted of 51 samples [2143]. In this case, the dataset was composed of 50 images [2144]. These two studies worked with 45 patients respectively [2145, 2146]. This research works with a total of 50 data points (a training set of 40 data points and a test set of 10 data points) [2147], in a Deep Learning application in surgery. This study

handled a total of 34 samples [2148]. This one, 32 gene features [2149], this one, 32 patients [2150], and this one, 32 pathogens [2151]. The study group for this DL research consisted of 27 patients [2152]. This application was developed with a dataset of 26 samples (10 training samples and 13 samples for testing) [2153]. This one had 21 samples [2154]. This research worked with data consisting of 40 BAM (Binary sequence Alignment/Map format) files, encompassing 20 individuals across chromosomes 1 through 22 [2155]. This utilized 19 metagenomic samples [2156]. This one, 15 subjects [2157], and this another one, 15 DNA samples [2158]. This one worked with 11 complete 3D images of a plant [2159]. This one involved sequencing data from 6 patients with Acute Myeloid Leukemia (AML) and 4 healthy individuals [2160]. This one, with 8 clinical isolates [2161].

This study utilized genomic spectrograms to detect liver cancer. For this purpose, a neural network was trained with a dataset containing 4 spectrograms with cancer and 4 without cancer [2162].

The smallest dataset found in this analysis consisted of only 7 transcriptome samples [2163]

Regarding the data imbalance issue, I found several investigations conducted with such an imbalance that it becomes imperative to question the actual quality of the results obtained. This one was trained with an unbalanced dataset (46 cases vs 946 controls) [2164]. This one also worked with an imbalanced dataset, with 3,802 samples from one category and 184 from the other [2165]. This one was also seriously imbalanced, comprising 453 ESM variants and 4,511 non-ESM variants [2166].

At this point, I want to emphasize the need to operate with a dataset that is truly representative in the application domain. ML and DL tools are absolutely interdisciplinary, meaning they can be applied in any field of knowledge if the necessary premises for their correct application are met. Therefore, the methodology of experimentation and analysis, as well as the validation of the results obtained, should be done from the perspective of the application domain, not merely from that of computation.

For example, this study on AI application in anesthesia calculations works with 20,000 samples. This is correct from a computational standpoint. However, these 20,000 samples come from only 10 patients [2167], posing a serious risk of bias regarding the patient profile it entails. I detected another case where the dataset consisted of thousands of samples from only 54 patients [2168].

Cases like these necessarily involve a strong presence of biases and a lack of representativeness of the general population due to insufficient diversity in the sample, as well as reduced statistical power. All of this questions the validity of these studies, even though computationally the training methodology may be correct. Not so much the real utility in the application domain, which should be the ultimate goal of such developments.

I found studies with poor and unbalanced data in relevant aspects of biomedicine. This study handled 80 samples of Programmed Death-Ligand 1 (PD-L1) and 800 samples of non-PD-L1 binding [2169]. The objective of this article was to design monoclonal antibody drugs, so given the importance of the topic and the consequences of incorrect developments, it would have been absolutely imperative to work with many more, better, and more balanced data.

Below, I present a representative example of a dataset formed by few data points and, moreover, highly biased. This study for lung cancer detection had a dataset containing only 311 images [2170], with the aggravating factor that the 311 images corresponded solely to 311 early-stage Non-Small Cell Lung Cancer (NSCLC) patients receiving surgical treatment at Massachusetts General Hospital. Therefore, the information was absolutely biased. The network was only trained with a very specific category of a very specific type of lung cancer, and likely with patients from a very specific geographical area as well.

Utilizing data augmentation to augment the number of training samples for the algorithm can be tempting. However, the critical content of the images remains the same, with all the biases that this fact entails [2171]. So, the validity of this method, at least in the biosanitary field, is more than doubtful. In my opinion, it is not advisable to use this technique, especially considering the importance

of the field of application and the harmful consequences that may arise from incorrect research methodology.

Generative models can also be applied to produce synthetic samples, synthetic data. However, generative models are still in their early stages, so contrasting results still discourage their unrestricted application in significant domains such as biomedicine. Synthetic samples obtained may have questionable quality and/or inherent biases, especially in domains with limited knowledge, such as genomics. Therefore, I strongly advise against these initiatives. Synthetic samples should not replace real samples. In my opinion, the validity of models trained in this way is questionable.

Despite all of the aforementioned, there are numerous studies that, given their small dataset, applied data augmentation or generative AI to fabricate synthetic data [2172–2180]. In this case, researchers had a dataset of 504 samples and expanded it with data augmentation [2181]. In this other case, they had 12 samples and applied data augmentation [2182]. From a computational standpoint, they did indeed have thousands of images. However, from a strictly scientific perspective, they only worked with 12 samples. Moreover, the limited number of samples necessarily implies a dataset deeply biased by the omission of various scenarios.

In one of the analyzed studies, the 50% of the data is synthetically generated, with the aggravating factor that all the generated samples belong to one of the two training categories. I quote verbatim: "We used the 741 human enhancers of the V-domain Ig Suppressor of T cell Activation (VISTA) as positive enhancers, and generated 741 negatives by randomly selecting 741 genomic background regions of similar length and chromosome distribution to the positives" [2183].

I found some developments specifically dedicated to the generation of synthetic data [2184–2187]. Indeed, there are studies that design data augmentation algorithms for genomic data [2188], even in genomic aspects as specific as de novo synthetic regulatory DNA generation [2189]. Given the current lack of scientific knowledge in the field of genomics, I consider this practice to pose a significant risk and to be inconsistent with the necessary standards of scientific rigor.

The widespread or excessive application of these techniques to artificially increase the dataset introduces biases and diminishes its quality. When applying data augmentation to a specific image, the critical content of the image remains unchanged. Therefore, considering these modified images as new is incorrect. I find questionable the validity of research conducted with datasets substantially augmented using these techniques.

Another risky dynamic is the one carried out in this study, where autonomous systems have been designed for virus segmentation, with the aim of automating the labeling process [2190]. The automation of the labeling process through autonomous systems can lead to incorrect data labeling, which in turn can affect the quality of model training and, ultimately, the accuracy of system predictions. Furthermore, the lack of proper human supervision can result in erroneous or even dangerous decisions.

The Transfer Learning technique has proven to be useful and promising for leveraging knowledge from previously acquired patterns [2191, 2192]. In my opinion, faced with a poor and unbalanced dataset, we should not simply rely on the application of Transfer Learning. Transfer Learning does not constitute the solution for an improperly sized dataset. Obtaining good results with inadequate datasets actually translates into obtaining poor results. This apparent improvement is not genuine and, therefore, is not valid.

In the face of an inadequate dataset, the only valid solution for rigorous research is to dismiss the application of ML and DL algorithms. This is because the quality and integrity of the results obtained largely depend on the quality of the data used to train these algorithms. Using a poor dataset can lead to erroneous or biased conclusions, compromising the validity of any study or analysis conducted. Therefore, it is essential to prioritize the quality of the dataset and, if it does not meet the necessary standards, seek alternatives or reconsider the methodological approach to ensure the validity and reliability of the results obtained.

In addition to the aforementioned, I found studies whose performance metrics are questionable. For instance, in a study analyzing glioblastomas, an AUC value of 0.718 and a Validation and Test Accuracy of 0.477 were obtained [2193]. This

development achieved an efficiency of 68% in sensitivity (SN), 82% in specificity (SP), and a Matthew Correlation Coefficient (MCC) of 0.51, respectively [2194].

In this analysis, I observed that as omics data became more complex, the dataset size was often smaller. It is likely that these studies were unable to obtain more robust data. The clear conclusion is that it is imperative to increase the amount of data, conduct more sequencing, and document medical tests and records more thoroughly and accurately.

I also found questionable articles in terms of interpretability. Some studies describe their architecture as self-explanatory. Although the results focus on explainability, they deliberately overlook their quality (they explicitly admit that the quality is poor) [2195]. In this case, the model cannot be considered valid for real-world application, no matter how explainable it is.

Thus far, the results of the critical analysis I conducted. Next, I aim to reflect on some of the criticisms found in the scientific literature regarding the implementation of deep learning in the biosanitary field.

Numerous studies reported on the intrinsic limitations of DL that make it difficult to replace traditional research and detection methods [2196–2201]. Some studies asserted that DL should not be overestimated either in academia or in industry, as it actually has many technical issues to address due to its inherent nature [2202].

Several studies question the actual performance of complex architectures for genome analysis [2203–2212], as well as their real-world application for clinical studies [2213–2224]. Some publications directly called for more transparency and honesty in the results [2225].

Some researchers warned of the existence of serious flaws in multiple research articles [2226, 2227]. For example, this article exposed that all ML and DL applications in diabetic retinopathy detection are incorrect [2228].

At the clinical level, the application of ML and DL tools is still in its early stages [2229, 2230]. Some studies called for more standardized research aligned with clinical study procedures [2231–2238]. This article even advocated for the priority of clinically evaluating ML models [2239].

Additionally, I identified multiple studies that suggested designing algorithms more tailored to genomics, not solely focused on achieving the highest possible prediction accuracy, but integrating genomic peculiarities efficiently, with improved user interfaces [2240–2242], and take into account the peculiarities of the data, such as proper weighting of features and the complexity of such data [2243]. Currently, there are some applications in this direction [2244].

It is striking how many articles stressed that in order to have DL systems that work properly, it is necessary to generate more data, of better quality and always ensure that the datasets are sufficient, complete and balanced [1067, 1119, 2214, 2216, 2245–2261]. There are studies that underlined the importance of biases in the data [2262–2264]. Biases can influence the results of analyses and data-driven decisions. They can distort the representation of reality and lead to erroneous conclusions if not properly identified and addressed. Moreover, biases can introduce inequities and discrimination in models and algorithms.

Having quality data will significantly facilitate the development of new technologies, such as single-cell sequencing, spatial transcriptomics, and multiplexed imaging, making them more accessible [2265]

It is imperative to sequence more, as several studies demand [2266–2272]. This is especially urgent in the case of metagenomic studies, where many studies still work with the few existing samples, exacerbated by the fact that the integration and analysis of metagenomic data remains a challenge [2273, 2274]

In the case of rare diseases, the limitations of ML and DL are greater due to the sometimes impossible absence of sufficient datasets. However, there are promising studies regarding the possible application of future algorithms in the field [2275].

It is evident that for an effective application of DL models, it is necessary to sequence more, in order to feed the models with all the data they need. We need to sequence more and gather as much supplementary data as possible so that certain studies can work with less incomplete data structures. Especially when working with multi-omic data, it is necessary to generate more and more data and databases, standardize formats, and share them [2276–2282].

Furthermore, it is essential to precisely understand the representativeness of the data used. In other words, there should be sufficient data from all possible categories, not just from certain targeted categories. Additionally, a comprehensive analysis of existing biases is necessary before determining the validity of these models [2283].

Another current challenge is the effective design of strategies for integrating multi-omic data [2038, 2284–2295]. The main problem is that most biological data have high-dimensional characteristics and a small sample size [2296–2302], even considering the spatio-temporal heterogeneity of certain types of biomedical data, such as certain characteristics of cancerous tissue [2303]. Therefore, some studies emphasized the importance of developing appropriate dimensionality reduction systems for DL applied to omics data [2304].

In genomics, the limited availability of reliable data and the complexity of intrinsic genome patterns exacerbate this issue, resulting in slower development in this field compared to others, such as oncology [2305].

Furthermore, there are still types of multi-omic data that are difficult to integrate, such as radiomic data. Large datasets with high-quality data still do not exist today. This, coupled with poor standardization of methodology, the black-box nature of DL, and lack of reproducibility, hinder an effective clinical implementation of radiomics with DL [2306]. Some studies demonstrated that models working with multi-omic data are still not sufficiently robust to noise [2307].

Some studies pointed out the need to optimize computational capacity when working with multidimensional data [2308].

It is also essential to standardize further the data formats in repositories [2309, 2310]. Effective integration of transcriptomic data remains a challenge [2311]. It is imperative to integrate, in addition to transcriptomic and genomic data, others such as metabolomic, environmental data, etc., as well as to make new data sources available to the research community, such as the UK BioBank and DriverDBv3 [2312].

Another pending challenge is optimizing model generalization, especially considering the high variability of training data and the significance of less common cases [2313, 2314].

Several experts lament the underutilization of existing documentary resources. We could have access to an enormous amount of clinical information globally, but currently, it is dispersed, poorly documented, and unstructured [2315]. This study even proposed the implementation of a decentralized federated learning model to create a single training model [2316].

Regarding the interpretability of DL models, a large number of studies emphasized the need for DL systems applied to genomics to be interpretable [1675, 2317–2340]. Another noted deficiency is the need for improvement in medical image segmentation [2341].

Furthermore, there are still challenges such as using qualitative and quantitative metrics, the challenge of causal inference [2342, 2343], strategies for leveraging unpaired data [2344], boundary conditions [2345], etc.

In the design of interpretability tools within the biosanitary context, it is essential to involve clinical experts in the algorithm design process and a need to build inherently explainable deep learning algorithms. Most explanation techniques are designed for general use and do not incorporate context [2346].

This insightful article identified the need for a biocentric interpretability, enhancing the biological plausibility of the obtained results [2347]. If the effective implementation of interpretable systems was achieved, there would be an increased willingness to use them in the medical field [2348].

Even Alphafold did not escape critical analysis. Several studies showed deficiencies in the results obtained [2349–2351].

In light of all the above, in the face of all the criticism from so many experts, I found some applications that I consider somewhat daring, excessively ambitious, especially considering the current state of the art. Just an example: this study proposes virtual healthcare centers, where AI systems diagnose based on symptomatology data and medical imaging [2352]. Taking into account the limitations

of AI regarding biases and poorly dimensioned, balanced, and labeled datasets, I consider this application a bold move.

I would like to conclude this section by encouraging readers to delve into an article exploring the potential of AI to replace radiologists [2353].

3.16 Conclusions

The growing interest in the biosanitary field is due to several converging factors. Firstly, technological advances have enabled the massive collection and analysis of biological and medical data, leading to an avalanche of information available for study. This includes genomic data, medical imaging, electronic health records, and more, providing a detailed insight into human biology and diseases.

Furthermore, ML and, especially, DL have advanced significantly over the last two decades, opening up new possibilities for interpreting and applying this data in the healthcare domain. DL algorithms can identify complex patterns in large datasets, aiding in disease prediction, medical diagnosis, treatment personalization, and clinical process optimization.

The growing concern for public health, accentuated since the COVID-19 pandemic, has propelled attention towards innovation in the biosanitary field. There is an increasing demand for solutions that enhance the efficiency and quality of healthcare, and improve access to healthcare services.

All this translates into a large number of research efforts focused on the application of DL tools in the biosanitary field, many of which address very specific and specialized areas of study.

Although genomics represents a field where the incorporation of ML and DL tools is still in its early stages compared to the progress in other areas like medical imaging, I found a wide variety of approaches for its implementation, as well as models applied to the study of very specific and complex aspects. Applications were developed that cover virtually all available architectures, along with various genomic processing methods.

The interest in applying ML and DL algorithms in the biosanitary field extends to various specialized areas, among which metagenomics and precision medicine stand out.

In the case of metagenomics, the capacity for analysis and processing of large biological datasets offers new opportunities to understand microbial diversity in various environments, such as soil, water, and the human body. AI and ML algorithms can assist in identifying and characterizing microorganisms present in environmental or biological samples.

On the other hand, precision medicine greatly benefits from the predictive capabilities of ML and DL. These algorithms can analyze genomic, gene expression, and clinical data to identify unique patterns in individual patients. This allows for more precise customization of medical treatments, taking into account the genetic and molecular variability of each patient. Additionally, these models can help predict the effectiveness of different therapies and medications, contributing to more personalized and effective healthcare.

Despite the potential of ML and DL algorithms, their effective application in the biosanitary field still requires many advancements. Reliable applications are needed, closer to the standards in biomedicine, not just focused on computing standards. This way, the results obtained in the application of ML and DL provide confidence to the end users of these tools.

For us, it was imperative to integrate everything discussed in Section 3.15 as constructive criticism when designing the compendium of experiments. This way, I aimed to avoid conceptual errors that could have led me to inherently flawed results under an apparent (and mistaken) success rate.

Many researchers advocate for effective multidisciplinarity among engineers, biochemists, microbiologists, computer scientists, mathematicians, statisticians, and physicists, among others [2354, 2355]. Following this premise, my determination was to conduct "de facto" multidisciplinary research. The entire design of the research methodology is consistent not only with standards in computing but also with standards in the biosanitary sector.

From the outset, I maintained the premise that Deep Learning (DL) tools should be applied with a defined purpose, recognizing that they are not an end in themselves. My primary goal was to contribute to the advancement of knowledge, particularly in the realm of its final application, specifically in understanding the genome and viruses in this case. My aim was to develop a reliable application for:

- Genomic data processing.
- DL.
- Interpretability.

It is imperative to provide reliable DL applications for biosanitary specialists. To achieve this, it is urgent to offer efficient, robust, and thoroughly validated developments. To this end, my choice was to "sacrifice" novelty in the computational part in favor of the biosanitary aspect by applying mature tools with multiple proven successes in the biosanitary field. Especially regarding the handling of CNNs, whose ability in pattern detection in images is amply demonstrated. I aimed to provide an implementation aligned with clinical standards that can be a further step towards the safe and multidisciplinary introduction of these DL tools in the biosanitary sector.

This prioritization of the biosanitary field is yet another factor to rely on inherently more reliable results. I applied techniques that we know work, that have worked repeatedly. Individually. However, as of the writing of this document, I did not find any research article that applies these techniques in the same way that I did.

The innovation does not lie so much in the techniques used themselves; that was not my purpose. What was truly innovative was applying them successfully to the genomes of different viruses and even for different research objectives.

In my research, therefore, the emphasis should not have been so much on the application of the DL algorithm, but on interpretability. It was there, once the other phases have been successfully overcome, where the contribution of relevant scientific knowledge resided.

And, as good multidimensional research, I did not just limit myself to merely applying these interpretability tools, but I compared the results of that interpretability with the existing literature to conduct a thorough study [2356]. I implemented

a direct application of that concept of biocentric interpretability indicated in Section 3.15.2. This concept was deeply inspiring to me because this biocentric interpretability allows for the assurance of the truthfulness and coherence of the results obtained. And that is exactly what I did.

Therefore, all explainability application results were subject to additional validation from the biomedical literature.

In this way, my aim was to develop research from the field of computer science that is reliable within the biosanitary domain. Thus, I could contribute accurate and robust scientific knowledge, bridging two seemingly distant areas of knowledge that greatly complement each other.

Thus, I utilized genomic spectrograms, an advantageous alignment-free system compared to processing genomic sequences using multiple sequence alignment, both for efficiently handling more sequences and for working with larger sequences. Additionally, the use of genomic spectrograms allowed me to harness the power of CNNs in image analysis. These networks are well-established technologies, which contributes to the reliability of the experiment. Moreover, I applied explainability not only to achieve good efficiency ratios but also to extract the maximum possible knowledge from the CNN's operation. Explainability was not limited to merely obtaining score maps and heatmaps. It was imperative for me to validate the results of explainability tools with existing literature, seeking a biocentric interpretation of the obtained results.

The deep study of nature is the most fruitful source of mathematical discoveries.

— Jean Baptiste Joseph Fourier, *a mathematician and a soldier. Unintentionally, he bequeathed mathematical tools that, decades later, would astonishingly revolutionize medicine and communication technologies.*

4 Genomic Spectrogram

Contents

4.1 Introduction 137
4.2 Brief Background 138
4.3 Concept of Genomic Spectrogram 141
4.4 Genomic Spectrograms vs. Multiple Sequence Alignment 142
4.5 Types of Genomic Spectrogram 144
4.6 Biomathematical Considerations about Genomic Spectrograms 151
4.7 Conclusions 152

4.1 Introduction

As introduced in Chapter 1, Section 1.9, to optimize the power of Convolutional Neural Networks (CNNs) in pattern detection, I need to convert genomic sequences into two-dimensional images.

These images must be biologically coherent, meaning they adhere to the fundamental principles of microbiology and genomics, regardless of whether these principles are known or unknown.

Therefore, the chosen transformation was the generation of genomic spectrograms. That is, generating the spectrogram for each of the four digital signals corresponding

to each of the types of nucleotides. In an immediately subsequent phase, I combined these four spectrograms into a single two-dimensional image that will have a unique correspondence with each genome. In this Chapter, I described in detail this conversion performed and I introduced the concept of genomic spectrogram.

I begin with a brief background on the use of spectrograms in various fields of science, as well as their suitability for use with CNNs (Section 4.2). Section 4.3 provides an in-depth description of the concept of the genomic spectrogram. In Section 4.5, I describe the two different types of genomic spectrograms used in my research (concatenated and superposed), as well as the complementarity of using both types of spectrogram representations. In Section 4.6, I present some biomathematical considerations. The chapter concludes with some conclusions (Section 4.7).

4.2 Brief Background

In the field of genomic research, mathematical models emerged as essential tools to address the inherent complexity of the genome [2357–2359]. The increasing availability of genomic sequencing data and the development of bioinformatics drove the creation of models applied in a variety of contexts, including genetic evolution [2360], molecular phylogeny [2361–2363], gene expression [2364, 2365], DNA structure [2366, 2367], and prediction of mutation effects [2368–2370]. Numerous studies and scientific publications highlighted the applicability and relevance of these models in the field of genomics, supporting their role in genetic research and understanding molecular biology.

The Fast Fourier Transform (FFT) is a computational algorithm that efficiently computes the Discrete Fourier Transform (DFT) of a sequence, which represents the Fourier coefficients of a time-domain signal [2371]. It is widely applied in signal processing, data analysis, and related fields to analyze and manipulate signals in the frequency domain. By recursively dividing the DFT into smaller sub-problems, the FFT algorithm employs a divide-and-conquer strategy to compute the DFT using elementary operations. This leads to a significant reduction in computation time compared to directly calculating the DFT, especially for larger input sizes.

Spectrograms represent the time evolution of the FFT. This mathematical representation is commonly used in various fields:

- Acoustics. Spectrograms are widely used in acoustics to analyze and visualize sound signals, such as speech, music, and animal vocalizations [2372]. They can reveal important information about the frequency content, timing, and intensity of sounds, which can be used for various purposes, such as speech recognition, music analysis, and bioacoustics research [2373, 2374].

- Signal processing. Spectrograms are used in signal processing to analyze and manipulate various types of signals, such as vibration signals [2375], electromagnetic signals [2376], and radar signals [2377]. They can help identify patterns, anomalies, and other characteristics of signals that are difficult to detect in the time domain.

- Astrophysics. Spectrograms are used in astrophysics to analyze and classify gravitational waves from core-collapse supernova [2378].

- Seismology. Spectrograms are used in seismology to analyze and visualize seismic signals, such as avalanches [2379] or seismic events [2380]. They can provide information about the location, magnitude, and source of seismic events, which can help seismologists monitor and predict seismic activity.

- Medical imaging. Spectrograms are used in medical imaging to analyze and visualize physiological signals, such as electrocardiograms (ECGs) [2381] and electroencephalograms (EEGs). They can help diagnose and monitor various medical conditions, such as obstructive sleep apnea [2382] or autism spectrum disorders [2383]. Although the application of Deep Learning Tools in health is still controversial [241], multiple studies have demonstrated that the performance of Deep Learning Tools using spectrograms renders good results in the interpretation of medical results [242–247]. Applying Transfer Learning on a pre-trained CNN allows extending its potential to new types of images and new domains such as medical diagnostics: prediction of multiple diseases from chest X-ray images [718], classification of abdominal ultrasound images [719], diabetic retinopathy images [720], cellular morphological changes to identify cellular phenotypes (incipient

cancer process and so on) [721] or detection of Major Depressive Disorder (MDD) by observing electroencephalogram (EEG) signals transformed to spectrograms [722].

Its successful application in such diverse fields validates the extrapolation of this mathematical transform to new areas of knowledge where its use is not as widespread. There is a growing trend in research to apply genomic spectrograms, which are closely linked with machine learning and Deep Learning techniques [343, 425, 2384].

In Chapter 3, Sections 3.11 and 3.13.2, I analyzed the degree of implementation of genomic spectrograms in the biosanitary field. I also compared this graphical representation of genomic sequences with other representation systems also used (k-mer images and Chaos Game Representation, CGR).

In view of the results of the analysis conducted, in my opinion, this method of conversion to image was appropriate. The genomic spectrogram is the least processed type of transformation, more raw. I did not apply any underlying filtering to the mathematical processing of the data. Although the result of the absence of filtering may be a noisier image than other mathematical representations of genomes, it has an advantage. Filtering signals potentially implies the loss of information of which we do not know the degree of importance. Therefore, I was interested in handling the rawest information possible.

Furthermore, this noisy aspect did not pose a problem for CNNs, whose potential for detection of patterns not visible to the human eye is well demonstrated in the scientific literature. I wanted to verify that the CNN located patterns in the raw data, without additional processing that could imply, in my particular case, a loss of information that the CNN could detect, even if the human eye did not.

As additional endorsement to this decision, I identified several scientific studies that affirm that genome frequency-domain analysis is a promising graphical representation for detecting biological patterns. Calculation of the spectrogram of a genomic sequence makes it possible to obtain a graphic representation of the significance of each periodicity (frequency) at each position (assimilating the concept of time to position) of the genome [399, 2385]. There is even literature

that states that the optimal use of CNNs' potential requires biologically consistent graphical conversion of genomic sequences [2386].

Therefore, for the purpose I seek, representing genomic sequences as genomic spectrograms is the most appropriate image conversion system.

4.3 Concept of Genomic Spectrogram

Genome frequency-domain analysis is a promising graphical representation for detecting biological patterns [343, 399, 2385].

The application of signal processing tools to genomic sequences of different organisms has shed light on the relationship between nucleotide periodicity and various features [418], such as the relationship between coding areas and sensitivity at frequency $f = 1/3$ [2387], non-coding RNA molecules location [327], GpC islands detection or micro-satellites identification [355].

In addition, these tools have the advantage of not requiring Multiple Sequence Alignment (MSA) to compare different genomes. This reduces the computational resources needed and eliminates the useless information of aligned sequences [412].

Specifically, spectrogram imaging, as a graphical representation of the FFT, is increasingly used to analyze genomic sequences [425]. The first step in generating the spectrogram of a sequence was to convert that sequence into four digital signals (U_a, U_g, U_c, U_t), which are activated by the presence of a nucleotide type (A, G, C, T) at each position.

$$U_\alpha(x_i) = \begin{cases} 1, & \text{if } x_i = \alpha \\ 0, & \text{otherwise} \end{cases} \tag{4.1}$$

The next step was to calculate the spectrogram for each of the four binary signals obtained from this decomposition.

The FFT corresponding to each of these signals was performed, and this is a computationally efficient method for the calculation of the Discrete Fourier Transform (DFT) [2388].

$$X(k) = \sum_{n=0}^{N-1} U_\alpha(x_i) W_N^{kn} \qquad 0 \leq k \leq N-1 \tag{4.2}$$

Where:

$$W_N = e^{-j2\pi/N} \tag{4.3}$$

FFT algorithms decompose an N-point DFT into smaller DFTs [2389] using sliding windows.

Calculation of the spectrogram of a genomic sequence makes it possible to obtain a graphic representation of the significance of each periodicity (frequency) at each position (assimilating the concept of time to position) of the genome.

The x-axis represents the position of each nucleotide. The y-axis is the frequency range and the z-axis is the value of the FFT. The spectrogram is a two dimensional representation obtained by replacing the z-axis with a color palette according to the amplitude of the FFT of the position of a nucleotide (x-axis) at a given frequency (y-axis) [414].

The spectrogram is a visual tool that represents the spectral composition of a signal as a function of time using the Fast Fourier Transform. This image contains a significant mathematical component.

4.4 Genomic Spectrograms vs. Multiple Sequence Alignment

A major advantage of working with genomic spectrograms is that this is an alignment-free processing system.

In Chapter 2, Section 2.4, and in Chapter 3, Section 3.15.1, I discussed the current limitations of Multiple Sequence Alignment (MSA) systems, summarized as the introduction of additional noise into the data (gaps), consideration of these algorithms as NP-hard problems, and suboptimal solutions obtained that often require manual adjustments letter by letter.

These limitations severely hinder the widespread application of these algorithms to comparative analyses involving a large number of sequences or cases where the

length of these sequences is large. Depending on the type and complexity of the organism, the length of its genome can be extremely extensive, making it difficult to handle and compare. The entire human genome is composed of approximately 3.2 billion DNA nucleotides [2390]. The genome of the long-tailed lizard *Anolis carolinensis*, which is a species commonly used in research, has approximately 1.8 billion DNA nucleotides [2391]. The genome of *Saccharomyces cerevisiae*, which is a fungus commonly used in research and food production, has approximately 12 million DNA nucleotides [2392]. The bacterium *Salmonella enterica* subsp. Enterica Serovar Typhimurium strain, which is a common cause of food poisoning, has a circular genome approximately 4.9 million nucleotides in length [39, 2393]. The DNA of all these organisms is too long for a graphic representation of their complete genome.

However, the genome size of viruses, which are the smallest "quasi-living" organisms, is much smaller and presents with a more simple structure. Its genome is designed to maximize the encoding of genetic information, given the limited space available in their genomes [2394].

Advancements in high-throughput sequencing technologies and the development of scientific disciplines such as metagenomics exacerbate these limitations. Therefore, it is necessary to develop efficient and robust genomic sequence processing systems that are alignment-free. In this regard, the generation of genomic spectrograms represents a significant advantage.

It should be noted that the conversion of genomic sequence to genomic spectrogram entails a loss of information regarding biological aspects. The generation of the spectrogram inherently involves a simplification of the structure and biological function inherent in the original genomic sequence. By representing genomic information in a spectrogram, the precise encoding of nucleotide sequences is lost. Additionally, the transformation to a frequency space entails a loss of contextual and sequential information, which may hinder the interpretation of the biological function of genomic sequences, obscure relevant details, and limit the complete understanding of the underlying biology.

Furthermore, the localization of genomic regions of interest is not as precise as in a multiple sequence alignment.

In my opinion, processing genomic sequences using genomic spectrograms cannot completely replace MSA algorithms. However, genomic spectrogram analysis can be a valuable complementary technique, especially for identifying genomic areas relevant to the study object. For example, spectrograms can highlight conserved or variable regions, as well as the presence of repetitive elements or unique sequences. This visual representation can also facilitate the identification of similarities and differences in genetic structure and composition. The complementarity of both techniques can provide a deeper level of knowledge in the field of genomics.

Nevertheless, working with genomic spectrograms has an additional advantage, which is the use of Deep Learning tools. Certain DL algorithms such as CNNs have demonstrated good and robust performance in multiple domains, with the advantage of being more scalable and efficient algorithms in handling large volumes of genomic data.

4.5 Types of Genomic Spectrogram

As a continuation and consequence of the issues discussed in Section 4.4, one of the limitations observed in Chapter 3, Section 3.15.1, is that most research works with fragments of genomic sequences, typically coding regions, instead of working with complete genomes. This is motivated by computational limitations that prevent processing entire genomes, especially when working with organisms whose genome is relatively large (millions of nucleotides).

Working with incomplete genomes can lead to the loss of biological information contained in non-coding regions or other auxiliary elements. I consider it important to work with complete genomes to avoid losing information in a field where it is not possible to know exactly which elements are relevant and which are not, or how genomic elements interrelate with each other.

Working with viruses allowed to process complete genomes without being affected by computational limitations. However, I must note the limitations of this system

regarding visualization if we intend to work with longer and more complex organisms. A future line of research will be precisely to delineate the maximum lengths of complete genomes with which one can work correctly and, therefore, to which organisms the scope of application of genomic spectrograms may be limited.

In order to minimize the impact of the weaknesses involved in processing genomic spectrograms, I employed two different representations of the genomic spectrogram. Working with two different graphical representations of the genomic spectrogram enabled me to obtain complementary information. In this way, I aimed to minimize the loss of information involved in this mathematical transformation.

As explained in Section 4.3, I generated the spectrogram for each digital signal corresponding to each nucleotide. From there, I employed two different types of genomic spectrogram representation.

The first one, called **Concatenated Spectrogram**, displays the concatenation of the spectrogram for each of the four types of nucleotides on the x-axis. Figure 4.1 is a graphical representation of the process of generating concatenated spectrograms vs. processing each genomic spectrogram by nucleotide type independently.

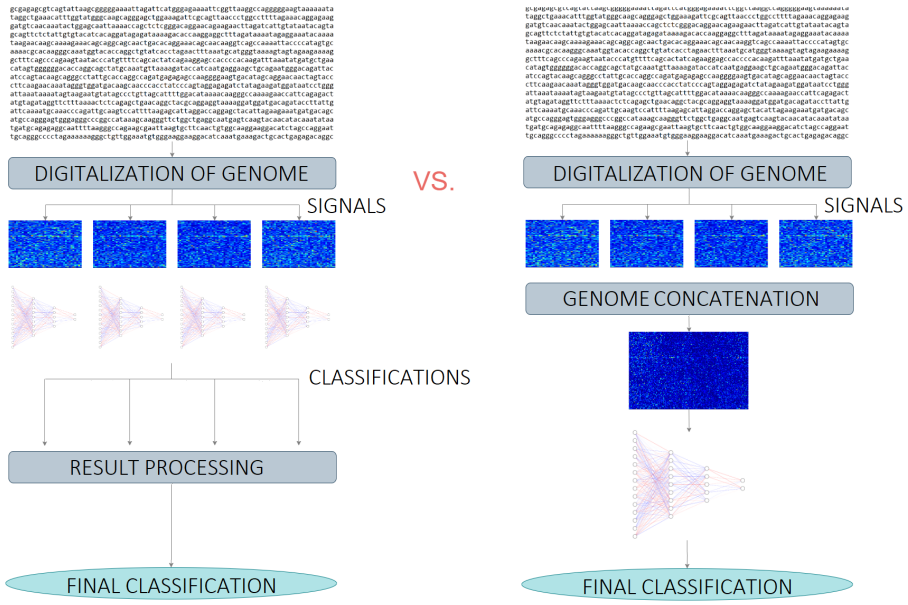


Figure 4.1: Procedure Comparison. The left graphical representation illustrates the procedure of classifying genomic spectrograms by each nucleotide type independently. The right representation outlines the procedure of classifying concatenated genomic spectrograms.

In Figure 4.1, the scheme on the left consists of a first phase of genome conversion (sequence primarily consisting of the letters A, C, G, and T) into four genomic spectrograms, one for each type of nucleotide. Next, I would require four neural networks, one for each type of genomic spectrogram (for A, C, G, and T, respectively). Therefore, I would obtain four independent classification results. A result processing phase would be necessary with the four classifications, one for each type of genomic spectrogram, after which I would arrive at a final classification based on the four previous results.

However, it is possible to operate with a genomic spectrogram that includes all four spectrograms per nucleotide type. As shown in the scheme on the right in Figure 4.1, after the same phase of converting the genome to genomic spectrograms by nucleotide type, I concatenated these four figures, these four spectrograms. This way, I obtained in a single image the information contained in the four images from the previous phase. Therefore, I only needed a neural network to classify that single concatenated spectrogram, and additional result processing from four independent classifications was not necessary.

The advantages of using concatenated spectrograms instead of four independent spectrograms are grouped as follows:

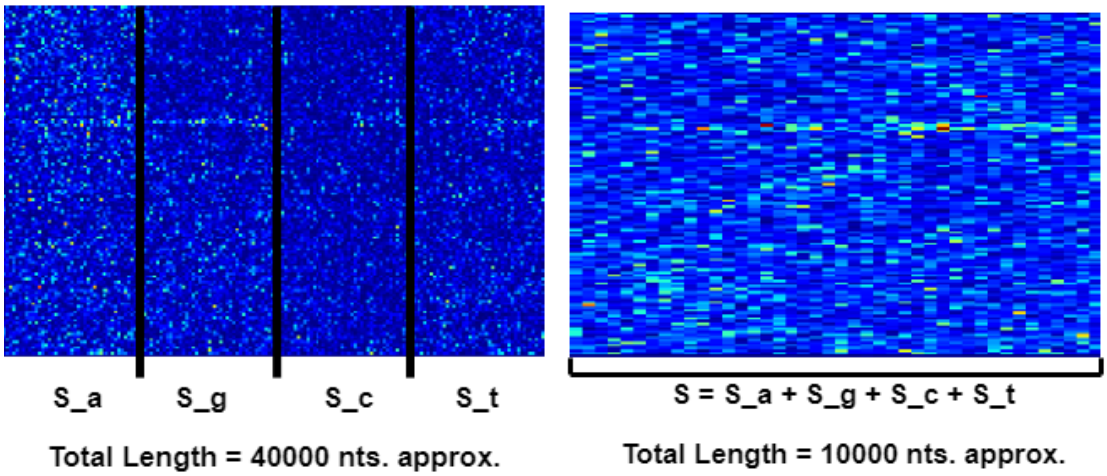
- Greater computational efficiency. By generating only one result, computational load is reduced compared to processing four results. This can result in shorter processing times and more efficient use of available computational resources.

- Faster processing. By reducing the number of results generated.

- Simplification of data analysis and interpretation.

- Reduced likelihood of errors. By reducing the number of results generated, the likelihood of errors in data processing or interpretation is decreased. This can improve the accuracy and reliability of the results obtained.

I employed two different and complementary graphical representations of genomic spectrograms (Figure 4.2). A schematic representation of a HIV-1 concatenated spectrogram is shown in Figure 4.2a.



(a) Concatenated **(b)** Superposed

Figure 4.2: Spectrograms Schemes. The x-axis represents the position of each nucleotide. Therefore, in Figure 4.2a the length is four times the sequence length owing to the concatenation of the four spectrograms. In Figure 4.2b the length of x-axis matches the genome length. In both cases, the y-axis represents a frequency range from 0 to 0.5. The z-axis is the spectrogram calculation.

This representation system can determine the most influential nucleotide types for the characterization of a virus as recombinant. However, it is difficult to determine accurately the involved area of the genome.

It should also be noted that this system may only be viable for analyzing small to medium genomes.

The second one is called **Superposed Spectrogram**. In this representation, the z-axis is the arithmetic sum of the values along the z-axis for each of the four nucleotide types:

$$S = S_a + S_g + S_c + S_t \tag{4.4}$$

Figure 4.2b shows the graphical concept of this type of representation, which allows to more effectively specify the influential zones of the genome.

To work with two-dimensional spectrograms, the z-axis is assimilated to a "jet" color palette (Figure 4.3). Using the matplotlib.pyplot.pcolormesh function, I mapped the minimum value of the FFT to deep blue (0) and the maximum value to deep red (1), automatically generating color assignments for each intermediate

value of the FFT. By default, the data range is mapped to the colorbar range using linear scaling.



Figure 4.3: Jet Color Map. Blue color corresponds to lowest sensitivities and red to the highest ones [2395].

Each spectrogram is identified by the reference of the corresponding sequence.

I generated spectrograms of both datasets using Python, the Scipy library, scipy.signal.spectrogram. To avoid biases and bad practices during CNN training, I omitted axes, margins and any other elements that may distort the operation of the network. Therefore, the image is limited to the spectrogram only [2396, 2397].

The parameters for generating genomic spectrograms were set to their default values [2398]. This includes:

- Sampling frequency of the x-axis (time series) (fs). This parameter refers to the sampling frequency in the time series [2399]. Default value = 1.0. This means that sampling is done nucleotide by nucleotide.

- Window (window). This parameter indicates the desired window to use. It is a mathematical function applied to each segment of the signal before calculating its STFT, in order to improve the representation of frequencies present in the signal and control the resolution in both the frequency and time domains. Defaults to a Tukey window with a shape parameter of 0.25. The Tukey window is also known as a tapered cosine window. By applying this window, smoothing of the periodogram estimate is achieved thanks to the smooth transition using a tapered cosine function to smooth the edges. This helps to reduce artifacts caused by discontinuities at the edges of signal segments, thus improving the quality of the resulting spectrogram. The shape parameter determines the proportion of the segment length where the smooth transition between unitary values in the center of the window and attenuated values at the edges is applied [2389].

- Length of each segment (nperseg). This is the number of data points used to compute the Short-Time Fourier Transform (STFT) in each segment of the input signal, i.e., the length of each segment [2400]. The default value is 256.

- Number of overlapping points between successive segments (noverlap). In the calculation of the STFT, the input signal is divided into overlapping segments of a certain length, a number of segment points. Subsequently, a window (such as the Tukey window, Hanning, etc.) is applied to each segment before performing the STFT. This parameter controls how many points overlap between adjacent segments [2401]. The default value is $noverlap = nperseg//8$.

- Length of the FFT used (nfft). This refers to the number of data points used in the calculation of the Short-Time Fourier Transform (STFT) and affects the frequency domain resolution of the resulting spectrogram. A higher value of "nfft" will result in greater frequency domain resolution of the spectrogram, meaning it will be able to distinguish finer frequencies. However, it may also require more computational resources and can generate spectrograms that are more detailed but slower to compute [2402]. In this case, the default value is $nfft = nperseg$, where nperseg is 256.

- Detrend (detrend). This parameter is used to remove or reduce unwanted trends or low-frequency components from a signal or data series. In many cases, signals may contain trend components that are unrelated to the phenomenon of interest, such as slow fluctuations or baseline shifts, and removing these trends can be useful for focusing on the signal's features of interest [2403]. The default value is "constant", to remove a constant trend.

- Return a one-sided spectrum for real data (return_onesided). In signal analysis, especially in real signals (which exhibit Hermitian symmetry in the frequency domain), half of the values in the STFT are redundant, as the other half is symmetric about the center frequency. Therefore, to save memory and computation time, one can choose to return only half of the results and conserve computational resources. This parameter is boolean (True or False). If set to True, the function will return only half of the values, typically the positive or right half, whereas if set to False, it will return all computed values [2404]. By default, its value is True, thus returning the positive half of the values.

- Scaling (scaling). In the calculation of the spectrogram, this parameter refers to choosing between two options: "density" and "spectrum". When set to "density", the spectral density of power per unit frequency is computed. When set to "spectrum", the power spectrum is calculated. In other words, the power at each frequency is represented without considering power density [2405]. It defaults to "density", so I computed spectral density of frequency.

- Axis along which the spectrogram is computed (axis). The default value is "over the last axis".

- Mode. Defines what kind of return values are expected [2406]. Options are "psd", "complex", "magnitude", "angle", and "phase". Each of these options means:

i) Power Spectral Density (psd). It indicates that the function is expected to return the power spectral density. This means that the return values will represent the power density of the signal in the frequency domain.

ii) Complex. It indicates that the function is expected to return the values of the complex Fourier transform. This will include both the real and imaginary parts of the frequency components of the signal.

iii) Magnitude. It indicates that the function is expected to return the magnitude of the Fourier transform. This will represent the amplitude or the absolute value of the frequency components of the signal.

iv) Angle. It indicates that the function is expected to return the angle of the Fourier transform. This will represent the phase or angle of the frequency components of the signal.

v) Phase. It indicates that the function is expected to return the phase of the Fourier transform. This can be useful for analyzing the phase relationship between different frequency components of the signal.

The default value of the Mode parameter is "psd", so the function returns the power spectral density.

These are the default values with which I generated the concatenated and the superposed spectrograms in all experiments conducted.

4.6 Biomathematical Considerations about Genomic Spectrograms

The mathematical nature of the genomic spectrogram inherently entails that it is an image with biological coherence. This means it is not subject to individual decisions regarding graphical representation.

For example, if we graphically represent a genome by simply replacing each nucleotide with a pixel of a specific color, we will obtain an array of colors whose length matches that of the genome itself. If we require a two-dimensional representation, we must decide on the size and group the lines accordingly. In this process, we lose the biological sense of the images, approximating distant genomic regions without direct relationship and distancing closely related genomic regions. For instance, fragments of the same gene or non-coding regions. All this simply due to adjusting the chain to the selected size.

This loss of biological coherence is confirmed by several tentative experiments I conducted with pre-trained CNNs. These networks yielded poor results, processing images resulting from a direct nucleotide-to-color conversion based on the appropriate size for that network.

On the contrary, the genomic spectrogram does maintain the biological coherence of the represented genome. A mathematical transformation is performed on the concatenation of nucleotides. And the resulting graphical representation of this transformation is generated automatically, according to the corresponding FFT value.

The mathematical nature of the spectrogram implies that genomic data is represented using mathematical techniques related to the frequency domain. By using this transform, we can identify and characterize patterns such as peaks, valleys, and waveforms, which will reflect different aspects of the genetic composition. We can also reveal correlations and complex relationships between genomic components, providing a deeper understanding of the structure and function of genomes [414, 2407].

In an initial approach to understanding these mathematical patterns, the first notable fact in the generation of a genomic spectrogram is the presence of a horizontal line at the frequency $f = 1/3$. Several studies relate this phenomenon to the fact

that 3 nucleotides form an amino acid in the coding regions of the genome [355, 2387]. Indeed, t is easily verifiable that, in the coding regions of the genome, the presence of this horizontal line intensifies. However, in the non-coding regions, this line practically disappears. Therefore, a first verification that the genomic spectrograms have been generated correctly is the identification of this horizontal line and its coincidence with the coding regions (genes).

There is abundant scientific evidence that analyzes the presence of mathematical patterns related to the periodicity of nucleotide occurrence in genomic sequences [316, 426, 2357, 2408, 2409]. These mathematical patterns can be more or less complex. These signatures may be related to multiple features and even unknown correlations among genomic elements.

Recent research has not yet unequivocally identified these mathematical patterns embedded in the genome. However, the results of each and every one of these articles reinforce, through multiple avenues, that there is encrypted information in the genome that determines or indicates certain characteristics, in this case, viral features such as presence of genetic recombination.

I consider it of high interest, therefore, to deepen our understanding of the frequency spectrum associated with genomic sequences.

4.7 Conclusions

Therefore, and in view of the supporting scientific literature, the use of genomic spectrograms allows for the identification of influential regions in the genome, providing valuable information for sequence analysis and characterization. For example, genomic spectrograms can highlight conserved or variable regions and the presence of repetitive elements.

Furthermore, a genomic spectrogram is a biologically coherent two-dimensional image. This preservation of the biological coherence of the original genome allows for the conservation of biological identity in the mathematical patterns embedded in this type of representation, thus enabling reliable and robust results.

Another additional advantage of genomic spectrograms is that, being biologically coherent image representations, they enable the application of CNNs. These DL algorithms have demonstrated very high power in detecting patterns not visible to the human eye. The results of multiple CNN applications have proven to be solid, robust, and highly efficient.

Likewise, the application of explainability tools allows to identify genomic areas relevant to the study of a particular biological feature.

Although genomic spectrograms cannot fully replace Multiple Sequence Alignment (MSA) algorithms, they offer valuable complementarity, especially in guiding relevant genomic regions.

In order to minimize the loss of biological information inherent in the transformation from genomic sequence to genomic spectrogram, I worked with two different types of genomic spectrograms: concatenated and superposed spectrograms. Additionally, I obtained complementary information derived from each type of spectrogram. Concatenated spectrograms provided us with information about relevant nucleotide types, while superposed spectrograms allowed us to more accurately determine relevant genomic elements in my case study.

This case is not just about AIDS, you know? So let's talk about what it really is about: the general public's hatred, our loathing, our fear of homosexuals. And how that climate of hatred and fear translated into the firing of this particular homosexual: my client.

— Denzel Washington, *Philadelphia, 1993*

5

HIV-1 Genetic Recombination Analysis

Contents

5.1 Introduction 156
5.2 HIV-1 Microbiological Considerations 159
5.3 Equipment 163
5.4 Methodology Block Diagram 163
5.5 HIV-1 Complete Genomic Sequences Compendium . . 164
5.6 HIV-1 Dataset 169
5.7 Spectrogram Generation 169
5.8 Pre-trained CNN Selection Criteria 170
5.9 Best Performance Based Decision Criteria 171
5.10 Robustness Criteria 174
5.11 Explainability Tools' Result Generation 175
5.12 Phylogenetic Tree and Comparison with Explainability Tools' Patterns 176
5.13 Raw Results 179
5.13.1 Concatenated Spectrograms 180
5.13.2 Superposed Spectrograms 183
5.14 Results Processing 186
5.14.1 CNN Selection 186
5.14.2 Robustness 189
5.14.3 Hot Zones Determination 190
5.15 Biological Consistency Analysis 194
5.15.1 Analysis of Successful Classifications 195
5.15.2 Similarity between Sequences of Different Subtypes . . . 225
5.15.3 Analysis of Failed Classifications 237
5.16 Conclusions 246
5.17 Future Studies 249

5.1 Introduction

The HIV-1 pandemic had a profound societal impact during the 1980s. According to data from United Nations Programme on HIV/AIDS (UNAIDS) [2410], from the onset of the pandemic until 2023, a total of 40.4 million individuals (with a range between 32.9 million and 51.3 million) have succumbed to diseases stemming from Acquired Immunodeficiency Syndrome (AIDS). Consequently, substantial funding was allocated for the advancement of research and development efforts aimed at finding a cure for AIDS.

Compared to other viruses, HIV-1 displays significantly high rates of variation. The vast genetic heterogeneity in HIV-1 can be attributed to its rapid replication, the error rate of its reverse transcriptase during genetic copying, and the high probability of genetic recombination during the virus replication process [2411]. At present, a definitive cure for HIV-1 remains elusive [2412], but effective treatments have been developed to manage the infection. These treatments are based on antiretroviral therapy (ART) and typically involve a combination of antiretroviral drugs acting on different stages of the HIV-1 life cycle [2413–2415]. One of the primary objectives of ART is to reduce the viral load of HIV-1. When the HIV-1 viral load is reduced to undetectable levels, it means that the amount of virus in the blood is so low that it cannot be detected by conventional tests [2416]. This does not necessarily imply that the virus has been completely eradicated, but rather that its replication is under control. Achieving and maintaining an undetectable viral load through ART, contributes to the preservation of the patient's immunological function, as HIV-1 targets and undermines the immune system [2417]. Moreover, sustaining low levels of the virus in the body significantly diminishes the risk of progression to AIDS, the advanced stage of HIV infection characterized by severe immunological dysfunction and the emergence of potentially life-threatening opportunistic diseases [2418].

Given that genetic recombination is one of the key factors influencing variability, it is imperative to comprehend the genetic recombination process and identify all potential mechanisms that regulate its incidence and its consequences. The significance of genetic recombination in viral diversification [2419] underscores the importance of devising novel approaches for detecting recombinant viruses, deepening our understanding of them and their recombination processes to anticipate viral evolution, and consequently, providing more information to pre-design effective treatments and vaccines.

As mentioned in Section 4.1, the exploration of mathematical patterns within the genome (intrinsic mathematical regularities and structures in the DNA sequence) can significantly contribute to our understanding of various genome-related aspects, including genetic evolution and gene expression, among others. In my specific context, this exploration provides valuable insights into the processes underlying genetic recombination in viruses.

Utilizing emerging technologies, such as Deep Learning, has the potential to enhance the current depth of microbiological understanding through innovative methodologies. Deep Learning, as a subset of Machine Learning, employs artificial neural networks with multiple layers to automatically discover intricate patterns and relationships within microbiological and/or genomic datasets. This advanced computational approach can unlock previously hidden insights in microbial biology, enabling more comprehensive and accurate analyses transcending the capabilities of traditional techniques.

In this Chapter, my objective was to decrypt the genomic information, specifically pertaining to genetic recombination processes in viruses, employing classification models. To achieve this aim, a significant case study is required, characterized by a substantial dataset, allowing the neural network to develop the capacity to discern pertinent patterns.

In this context, HIV-1 stands as a highly valuable research model due to its comprehensive catalog of extensively documented and studied recombinant sequences, rendering it an ideal choice for my research and analysis purposes.

Here, I exposed a novel methodology that leveraged a genomic representation independent of sequence length, coupled with the integration of Deep Learning algorithms [46]. This methodology was designed to model the classification of recombinant sequences, while concurrently offering explainability tools that pinpoint the specific genomic regions related to genetic recombination.

The identification of genetic recombination processes is immersed in inherent complexity, further compounded by the simultaneous presence of mutations in the genome. The challenge lies in the necessity to distinguish between genetic recombination events and multiple accumulated point mutations, as both types of alterations can introduce variability in genetic sequences. The presence of multiple mutations complicates the task of accurately identifying and characterizing recombination regions, requiring highly sophisticated analytical approaches and tools in genomic research.

This methodology provides some additional advantages. The use of genomic spectrograms with Deep Learning represents an alignment-free system, resulting in computational savings, shorter execution times, and the avoidance of introducing additional noise (Chapter 2, Section 2.4), as I will discuss later in this Chapter.

In accordance with what was described in Chapter 4, Section 4.5, in the context of viral genomes, it was possible to address the analysis of their complete genomes, thereby avoiding the loss of information resulting from the specific selection of genomic regions.

My objective was to conduct a comprehensive research into the identification of fundamental mathematical patterns that contribute to the characterization of organisms. I initiated this pursuit by focusing on a particular case study: discerning whether an HIV-1 virus is recombinant or non-recombinant, utilizing a thorough analysis of its entire genomic composition.

This Chapter consists of an initial explanatory section about some microbiological considerations regarding HIV-1 to facilitate the understanding of the experimental results (Section 5.2). Sections 5.3 and 5.4 describe the equipment used and the methodology designed, respectively. Sections 5.5, 5.6, 5.7, 5.8, 5.9, 5.10, 5.11,

and 5.12 describe in detail each of the stages of the methodology used. Section 5.13 presents the results of the performance metrics obtained in the conducted trainings. Section 5.14 identifies the selected pre-trained CNN, as well as the results of the robustness tests and the determination of the hot zones. Section 5.15 provides a detailed biological consistency analysis with a selection of representative cases. The Chapter concludes with the corresponding Conclusions (Section 5.16) and Future Studies (Section 5.17).

5.2 HIV-1 Microbiological Considerations

HIV-1 is classified as a single-stranded RNA retrotranscribed lentivirus within the retrovirus subfamily, whose genome size is around 10,000 nucleotides (nts).

The scaled schematic gene map of the HIV-1 gene map is shown in Figure 5.1

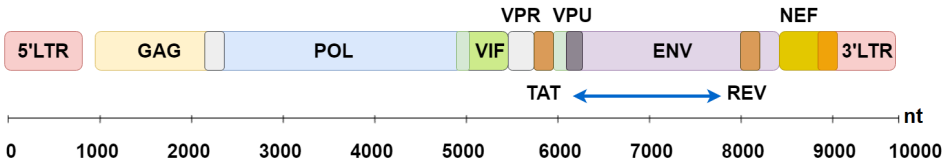


Figure 5.1: HIV-1 Gene Map. Based on the gene map of Los Álamos National Laboratory [41]. This diagram displays the main Open Reading Frames (ORFs) identified in the HIV-1 genome.

To aid in the visual comprehension of the structure of key proteins within the HIV-1 virion (the complete and infectious viral particle in its extracellular state), Figure 5.2 provides a schematic representation of the HIV-1 virus.

The functions of the proteins codified by HIV-1 genes are as follows:

- The *gag* gene is responsible for encoding the primary structural proteins of HIV-1. This includes the budding protein p6, p17, a membrane-anchored protein, and p24 (capsid), which assembles through polymerization into a conical nuclear structure. Inside this conical structure, there is a complex of proteins and nucleic acid composed of two copies of the HIV-1 genomic RNA and the nucleoprotein p7 (nucleocapsid), among other elements [2421, 2422].

- The *pol* gene encodes the Protease (PR), Reverse Transcriptase (RT) p66, and Integrase (IN), three enzymes that are essential for the infectious cycle of the

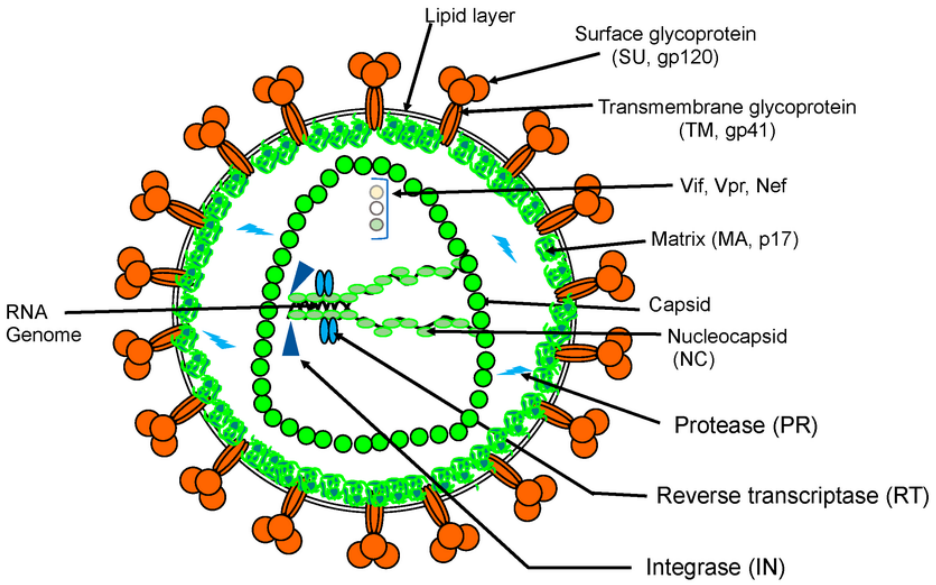


Figure 5.2: Schematic Representation of HIV-1 Virion. Image extracted from the research article by Eid et al. entitled "Advances in Continuous Microfluidics-Based Technologies for the Study of HIV Infection" [2420].

virus. They are generated as a Gag-Pol precursor polyprotein, which is processed by the viral protease. The Gag-Pol polyprotein is produced through a ribosome frameshifting event near the 3' end of *gag* [2423].

- The *env* gene codifies for an envelope glycoprotein that contains not only the CD4 binding site, but also seven transmembrane domain chemokine receptors that serve as co-receptors for HIV-1 [2424].

HIV-1 contains six other genes referred to as accessory genes that codify for proteins with varied functions:

- The *tat* gene's product is a transactivator protein of HIV gene expression [2425].

- The *rev*, along with *tat* codifies for a second necessary regulatory factor for HIV expression [2426].

- The *vif* gene codifies for the Viral Infectivity Factor, whose primary function is to interact with an antiviral defense protein in human cells called APOBEC3G (A3G), neutralizing its antiviral effect and promoting HIV replication [2427].

- The *vpr* gene, which codifies for a protein with some known functions that include directing preintegration complexes to the nucleus of the infected cell, inducing cell cycle arrest in the G2/M phase, transactivation of cellular genes (affecting

their regulation in the host cell), and inducing changes in the differentiation of hematopoietic cells and monocytes [2428].

- The *vpu* gene's product, among its functions, include promoting the degradation of the CD4 receptor in the cell membrane, facilitating the release of mature viral particles from infected cells, and counteracting the action of the antiviral protein Tetherin [2429, 2430].

- The *nef* gene codifies for a protein that induces negative regulation of CD4 and class I HLA molecules on the surface of infected cells, which may represent a significant escape mechanism by evading attack mediated by cytotoxic CD8+ lymphocytes. This protein also appears to interfere with T-cell activation by binding to various proteins involved in intracellular signal transduction pathways. It also inhibits apoptosis in HIV-infected cells [2431, 2432].

Regarding the replication process in host cells, retroviruses have a unique feature. They employ a process known as reverse transcription or retro-transcription. This involves the sequential conversion from viral RNA to DNA and subsequently to messenger RNA (mRNA) that will be translated into de viral proteins [2433].

Reverse transcription occurs via a series of steps, including two switches of nascent DNA, from one end to the opposite end of the RNA template [2434]. As a result, two identical sequences called Long Terminal Repeats (LTRs) are formed at each end of DNA. After reverse transcription, the viral DNA is integrated into the genome of the host cell, which is called provirus.

Within the HIV Genomic Structural Elements, the so-called Long Terminal Repeats (LTRs) are located at each end of the provirus and play an essential role in the initiation and regulation of viral transcription [2435]. Despite both LTRs display identical sequences, each one has different functions. 5'LTR contains the signals for the initiation and regulation of viral transcription while 3'LTR performs in post-transcriptional processes. 3'LTR has an efficient polyadenylation site (adding a polyA tail to messenger RNA), and is fundamental in the editing of viral transcripts [2436, 2437].

When HIV is ready to replicate and produce new copies of itself, 5'LTR serves as a promoter/activator region. This region contains signals that initiate and regulate the transcription process of the viral genome, allowing the synthesis of the messenger RNAs needed for the formation of new viral particles. On the other hand, 3'LTR fulfills its role in processes following transcription. It contains an efficient polyadenylation site, which is responsible for adding a polyadenine tail (a chain of adenines) to the viral messenger RNAs. This polyadenylation is a crucial step for the maturation of viral transcripts and affects their stability and processing [2436–2439].

Retroviruses possess another unique characteristic wherein each viral particle, or virion, contains two complete viral genomes. This particular feature is of considerable significance and may be intimately connected with the elevated occurrence of genetic recombination during the course of an HIV infection [2440]. This phenomenon can be attributed to the behavior of reverse transcriptase, an enzyme crucial in the replication of the retroviral genome, which demonstrates the intriguing ability to utilize portions of these two RNA molecules enclosed within a single virion as templates.

This characteristic amplifies the propensity for genetic recombination, when reverse transcriptase incorporates genetic material from two different viral genomes into a single DNA molecule. Consequently, the high error rate exhibited by reverse transcriptase during replication of the viral genome becomes a contributing factor to the heightened occurrence of genetic recombination [2441]. These errors result in the introduction of mutations, which, when combined with the dual-genome nature of retroviruses, further enhance the genetic diversity of the viral population.

The two copies of contained RNA, as well as the switches occurring during reverse transcription, may be related to the high level of genetic recombination that occurs in an HIV infectious process [2440, 2441].

In summary, the presence of two complete viral genomes within each retroviral virion, coupled with the indistinct use of these genomes by reverse transcriptase, creates a favorable environment for genetic recombination to occur. The high error rate during replication and the incorporation of genetic material from different viral

genomes into a single DNA molecule result in an elevated level of genetic recombination, contributing significantly to the genetic diversity observed in HIV populations.

The lack of a vaccine and highly effective treatments for HIV is partly due to the significant genetic recombination that takes place in HIV-infected individuals [2419]. To address this challenge, it is crucial to gain a comprehensive understanding of the genetic recombination process and identify all potential factors influencing it. This entails integrating various areas of expertise and cutting-edge technologies like Deep Learning to enhance our current understanding of microbiology from diverse angles.

HIV-1 is classified into four groups named M, N, O and P. Group M, which is the most widespread in the world, is divided into nine subtypes: A, B, C, D, F, G, H, J, K and L. Multiple recombinants between subtypes are currently circulating; they are called Circulating Recombinant Forms (CRF) [2442].

5.3 Equipment

All experiments were run in this equipment:

- Processing Unit: Intel(R) Xeon(R) Gold 5220R CPU @ 2.20 GHz.
- Installed RAM: 256 GB usable.
- Operative System: Windows 10 Education. Version: 21H1.
- GPU: NVIDIA RTX A6000. Total memory: 179.451 GB. VRAM: 48.571 GB.

5.4 Methodology Block Diagram

This genomics analysis process comprised three distinct phases.

In the first phase, the raw genomic sequences underwent a transformation into genomic spectrograms, which are effectively coherent biological images. This conversion into image data is a fundamental step as it facilitates the application of image analysis techniques to genomic information.

The second phase was centered around leveraging the power of Transfer Learning. Here, a pre-trained CNN was employed, using the dataset of these newly created

frequency-domain genomic images. Transfer learning allows the CNN to adapt its feature recognition capabilities to the specifics of this dataset. The CNN becomes adept at identifying intricate patterns and features within the transformed genomic data.

Upon the successful training of the CNN, the third and final phase came into play. In this stage, explainability tools were brought into action to dissect the inner workings of the CNN's decision-making process. These tools enabled me to pinpoint the specific areas of focus within the genomic spectrograms that were utilized by the CNN for classifying each image. My chosen explainability tools were GradCAM, LIME, and Gradient Attribution.

In essence, this three-phase approach encompasses the transformation of genetic data into interpretable images, the utilization of advanced neural networks to recognize complex genetic patterns, and the application of explainability tools to elucidate the decision-making process, ultimately leading to the identification of crucial viral features.

Figure 5.3 graphically summarizes the methodology utilized in the experimentation described in this Chapter.

5.5 HIV-1 Complete Genomic Sequences Compendium

I downloaded all complete HIV-1 sequences from all subtypes from the Los Alamos National Laboratory HIV Database repository in October 2021 [41].

Table 5.1 displays the number of complete HIV-1 sequences by non-recombinant subtype up to 2020.

These are subtypes which have not undergone genetic recombination with other subtypes. They are also referred to as pure subtypes. There are four primary groups: M, N, O, and P. Group M is the most prevalent globally and is further subdivided into ten subtypes: A, B, C, D, F, G, H, J, K, and L.

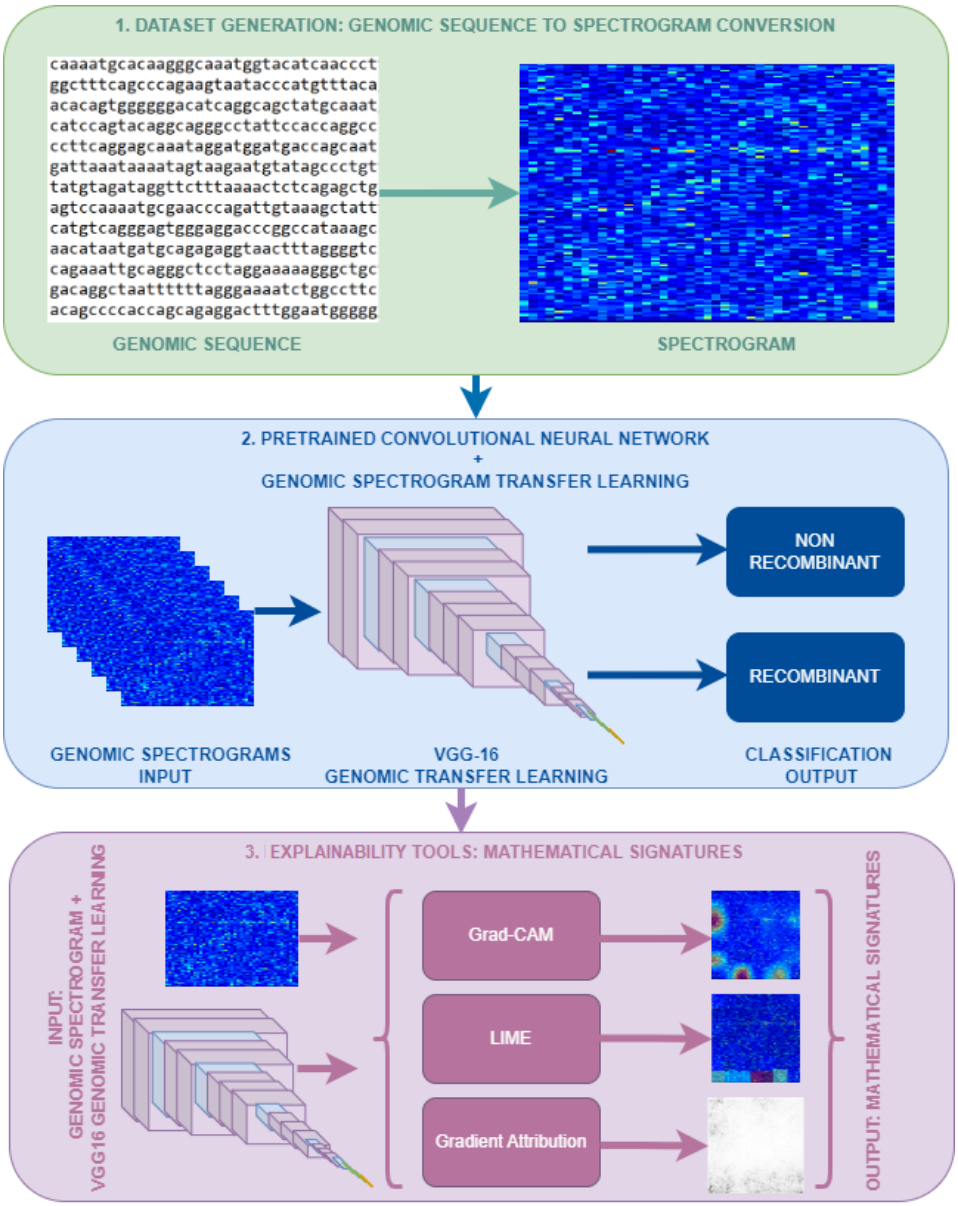


Figure 5.3: Methodology Block Diagram. The first phase involved converting genomic sequences into genomic spectrograms, which are biologically coherent images. The second phase entailed the application of Transfer Learning to a pre-trained Convolutional Neural Network using this dataset of frequency-domain images. This allowed the CNN to identify features within this image type. After the CNN was trained, in the third and final phase, I employed explainability tools to pinpoint the areas of focus used by the CNN for classifying each image. The outcome of this final phase was the identification of mathematical signatures associated with a viral feature's presence.

Table 5.1: Non-Recombinant HIV-1 Sequences up to 2020. Subtypes A-L are part of the primary group referred to as M. U stands for unclassified or unclassifiable sequences.

Subtype	Number of Sequences	Subtype	Number of Sequences
A	3	F2	14
A1	709	G	99
A2	4	H	10
A3	3	J	7
A4	3	K	2
A6	186	L	3
A7	3	N	11
A8	2	O	57
B	7,730	P	5
C	2,105	U	23
D	102	**TOTAL**	**11,163**
F1	82		

As can be seen in Table 5.1, the majority of non-recombinant sequences correspond to subtype B. It is one of the earliest subtypes identified and is currently more prevalent in the American continent and Europe [2443]. The non-recombinant subtype with the second highest number of complete sequences is subtype C, containing 2,105 sequences. This subtype is more prevalent in sub-Saharan Africa and has a significant presence in India and Nepal [2444]. The remaining 20 non-recombinant subtypes consist of fewer than 1,000 complete sequences.

Table 5.2 shows the number of complete HIV-1 sequences by recombinant subtype up to 2020.

These subtypes are also known as Circulating Recombinant Forms (CRFs). Their nomenclature carries the implicit composition of detected genetic recombination events. For example, the subtype CRF02_AG contains fragments from subtypes A and G. If the sequence is a mosaic of three or more subtypes, the subtype name includes the reference _cpx.

Table 5.2: Recombinant HIV-1 Sequences up to 2020. Total number of complete sequences for each CRF.

Subtype	Number of Sequences	Subtype	Number of Sequences
CRF01_AE	1,490	CRF60_BC	5
CRF02_AG	179	CRF61_BC	4
CRF03_A6B	5	CRF62_BC	3
CRF04_cpx	6	CRF63_02A6	13
CRF05_DF	4	CRF64_BC	9
CRF06_cpx	14	CRF65_cpx	6
CRF07_BC	43	CRF66_BF1	3
CRF08_BC	35	CRF67_01B	2
CRF09_cpx	5	CRF68_01B	3
CRF10_CD	3	CRF69_01B	7
CRF11_cpx	25	CRF70_BF1	4
CRF12_BF	15	CRF71_BF1	14
CRF13_cpx	10	CRF72_BF1	5
CRF14_BG	13	CRF73_BG	2
CRF15_01B	8	CRF74_01B	3
CRF16_A2D	4	CRF75_BF1	3
CRF17_BF1	7	CRF76_01B	3
CRF18_cpx	7	CRF77_cpx	4
CRF19_cpx	5	CRF78_cpx	3
CRF20_BG	23	CRF79_0107	3
CRF21_A2D	3	CRF80_0107	2
CRF22_01A1	21	CRF81_cpx	2
CRF23_BG	2	CRF82_cpx	6
CRF24_BG	6	CRF83_cpx	11
CRF25_cpx	5	CRF84_A1D	3
CRF26_A5U	4	CRF85_BC	11
CRF27_cpx	4	CRF86_BC	3
CRF28_BF1	5	CRF87_cpx	3
CRF29_BF1	6	CRF88_BC	3
CRF30_0206	4	CRF89_BF1	8
CRF31_BC	3	CRF90_BF1	6
CRF32_0206	3	CRF92_C2U	7
CRF33_01B	8	CRF93_cpx	3
CRF34_01B	3	CRF94_cpx	6
CRF35_A1D	22	CRF95_02B	3
CRF36_cpx	3	CRF96_cpx	3
CRF37_cpx	4	CRF97_01B	5
CRF38_BF1	5	CRF98_06B	1
CRF39_BF1	3	CRF99_BF1	2
CRF40_BF1	4	CRF100_01C	3
CRF41_CD	3	CRF101_01B	3

Table 5.2 – Continuation from the previous page

Subtype	Number of Sequences	Subtype	Number of Sequences
CRF42_BF1	17	CRF102_0107	2
CRF43_02G	5	CRF103_01B	4
CRF44_BF1	3	CRF104_0107	3
CRF45_cpx	5	CRF105_0108	5
CRF46_BF1	8	CRF106_cpx	6
CRF47_BF1	5	CRF107_01B	4
CRF48_01B	3	CRF108_BC	5
CRF49_cpx	4	CRF109_0107	2
CRF50_A1D	5	CRF110_BC	3
CRF51_01B	7	CRF111_01C	8
CRF52_01B	3	CRF112_01B	4
CRF53_01B	4	CRF113_0107	3
CRF54_01B	3	CRF114_0155	7
CRF55_01B	9	CRF115_01C	2
CRF56_cpx	4	CRF116_0108	3
CRF57_BC	7	CRF117_0107	7
CRF58_01B	6	CRF118_BC	1
CRF59_01B	9	**TOTAL**	**2,391**

Referring to Table 5.2, the number of identified recombinant subtypes is significantly higher than that of non-recombinant subtypes, which is indicative of the frequency of genetic recombination events in HIV-1. Until 2020, there were 118 different recombinant subtypes identified with complete genomic sequences. The only recombinant subtype with more than 1,000 sequences is CRF01_AE, commonly found in Southeast and East Asia [2445]. The second most numerous recombinant subtype is CRF02_AG, containing 179 sequences. This subtype has been most frequently observed in West and Central Africa but has also been detected in different geographic areas [2446]. The remaining 116 subtypes each comprise fewer than 50 complete sequences. Overall, their sequence count per subtype is lower compared to non-recombinant subtypes. However, the identification and description of 118 recombinant subtypes underscore the prevalence of genetic recombination.

5.6 HIV-1 Dataset

To build the dataset, I labeled all CRF sequences as recombinant and pure subtypes as non-recombinant. Subsequently, I randomly allocated these sequences to training, validation, and test folders in accordance with the guidelines outlined in Table 5.3. Out of a total of 13,554 complete HIV-1 sequences available until 2020, 82.36% of these sequences are classified as Non-Recombinants (11,267 sequences), while the remaining 17.64% are classified as Recombinants (2,391 sequences).

Table 5.3: Dataset Structure

Dataset	Recombinant	Non-Recombinant	TOTAL
Training	1,903 (16.89%)	9,364 (83.11%)	11,267
Validation	338 (17.01%)	1,649 (82.99%)	1,987
Test	150 (50.00%)	150 (50.00%)	300
TOTAL	**2,391**	**11,163**	**13,554**

I maintained this scenario in the training and validation folders. However, to mitigate biases stemming from this imbalance in the available biological evidence and to detect under-fitting biases during the training phases, I intentionally altered the structure of the Test Set, with a distribution of 50% / 50%, which differed from that of the Training and Validation Sets (17% / 83%) [38].

5.7 Spectrogram Generation

The utilization of concatenated and superposed spectrograms (see Chapter 4, Section 4.5) provided a means to access complementary information, thereby enhancing our ability to discern potential patterns that may elucidate the determinants of HIV-1 genome recombination. Each individual spectrogram was linked to the reference of its corresponding sequence, enabling precise identification and association with genomic data.

I performed all the experiments by randomly splitting the entire dataset into training, validation and test folders. Each sequence identified by its ID was present in the Training, Validation, and Test Sets for both concatenated and superposed

spectrograms. In essence, if a sequence was included in the Training Set for concatenated spectrograms, it was also included in the Training Set for superposed spectrograms. I applied the same principle to the Validation and Test Sets. This approach enhanced result comparability by eliminating discrepancies stemming from variations in the training data.

5.8 Pre-trained CNN Selection Criteria

Following the generation of both spectrogram datasets, the subsequent stage entailed the execution of a comprehensive test bench. The overarching goal was not solely limited to the selection of the CNN model with the highest performance, but it also encompassed the meticulous exploration and optimization of hyperparameter configurations, striving to attain the highest accuracy levels, particularly in achieving a well-balanced performance across non-recombinant and recombinant viral genomes.

CNNs of choice were:

- VGG-16 [2447].
- ResNET-101 [2448].
- Inception V3 [2449].

The experiments were conducted utilizing the MATLAB 2021b App Deep Learning Designer. The Convolutional Neural Networks (CNNs) employed in these experiments are pre-trained on a vast dataset containing millions of images sourced from the ImageNet database, enabling them to classify images into 1,000 distinct categories. Further details regarding the pre-trained CNNs are available in MATLAB [2450].

Through Transfer Learning, the CNN learned the specific features of both datasets. As explained in Section 3.5, freezing the weights of all intermediate layers intact, I replaced the last fully connected layer, specifying the new number of categories to analyze. I also replaced the classification layer of the pre-trained model. In this way, I extracted all previous knowledge from the other datasets and adapted it to my new problem [48].

As evidenced through initial testing, the application of Data Augmentation had a detrimental impact on performance, presumably owing to its potential distortion of the biological context inherent in the sequence. The augmentation process, especially when it includes rotations, may introduce inconsistencies that compromise the accuracy of the representation. It is important to note that the biological significance of the image is intricately linked to its precise orientation. As a result, I abstained from the utilization of this technique, as it would have counterproductively altered the fundamental biological context essential for accurate analysis.

The following test bench was conducted with the aim of ensuring uniformity in the evaluation of concatenated and superposed spectrogram datasets, allowing for the most direct and consistent comparisons among various CNN models (as summarized in Table 5.4).

Table 5.4: Test Bench. The same set of conditions was uniformly applied to all the chosen CNN models. These conditions encompassed various combinations of Learning Rates (0.0001 and 0.0100), Batch Sizes (52 and 128), and the number of training Epochs (10 and 30).

Test	Learning Rate	Batch Size	Epochs
1)	0.0001	52	10
2)	0.0100	52	10
3)	0.0001	52	30
4)	0.0100	52	30
5)	0.0001	128	10
6)	0.0100	128	10
7)	0.0001	128	30
8)	0.0100	128	30

5.9 Best Performance Based Decision Criteria

The assessment of optimal results should not rely exclusively on the aggregate count of successful classifications. The evaluation of CNN performance was guided by four specific and selected performance metrics, which are as follows:

- Validation Accuracy. The resulting percentage obtained by dividing the number of correct classifications by the total number of samples in the Validation Set [2451].

- Training Time. This is the time required to develop a model that effectively learned discernible patterns from the dataset [2452].

- Confusion Matrix Results. It summarizes the predictions of the model on the Test Set data, comparing them to the true class labels [2453]. The more balanced Confusion Matrix is, the better the performance of the CNN. The graphical representation obtained is similar to the one shown in Figure 5.4.

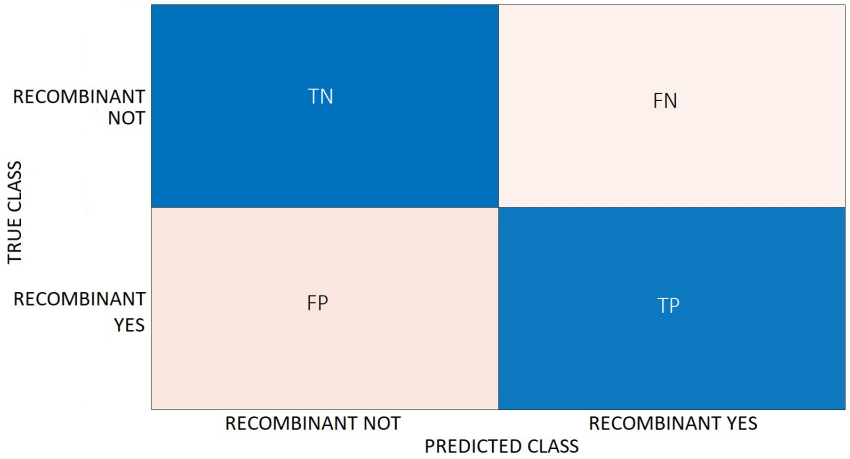


Figure 5.4: Confusion Matrix. Blue sections represent hits and orange sections represent errors. TP: true positive; TN: true negative; FP: false positive; FN: false negative.

Being:

TP: True Positive. The classification model made a positive prediction, and this prediction is correct in the real data.

TN: True Negative. The classification model made a negative prediction, and this prediction is correct in the real data.

FP: False Positive. The classification model made a positive prediction, and this prediction is incorrect in the real data.

FN: False Negative. The classification model made a negative prediction, and this prediction is incorrect in the real data.

- AUC (Area Under de Curve). The Receiver Operating Characteristic curve (ROC Curve) is the graphical representation of the performance of a classification model, plotting True Positive Rate (TPR = sensitivity) vs. False Positive Rate (FPR = 1 - specificity) [2454], which are defined by:

$$TPR = \frac{TP}{TP + FN} \tag{5.1}$$

$$FPR = \frac{FP}{FP + TN} = 1 - \text{specificity} \tag{5.2}$$

The greater the AUC value, the higher the model's discriminatory power in distinguishing between the two classes. A value of 0.5 signifies a model with no discriminatory ability, essentially making random decisions. Theoretically, an AUC value of 1 indicates a model with perfect discriminative capacity, capable of flawless class separation. However, in the context of Deep Learning, an AUC value of 1 is often a strong indicator of overfitting, signifying that the model has essentially memorized the training data without effectively learning to detect underlying patterns and generalize appropriately.

Figure 5.5 provides an illustration of a Receiver Operating Characteristic (ROC) curve for a Deep Learning model exhibiting satisfactory performance. In this case, the AUC approaches approximately 0.99, but it never reaches 1.

Training Time acts as a measure of the required computational resources. When similar values of Validation Accuracy and AUC were achieved, the decisive criterion became the maximum Test Accuracy in both recombinant and non-recombinant categories (Confusion Matrix Results). The best performance was attained when the hit rates in both categories were balanced and optimized.

From this standpoint, the CNN model that demonstrated superior performance across concatenated and superposed datasets, was the one chosen. This selection was based not only on achieving equilibrium within both categories but also ensuring a harmonious balance across both datasets.

As a measure of the balance between concatenated and superposed spectrograms, I determined that the optimal pre-trained CNN was the one with the lowest total distance value in both datasets. That is, the one with the smallest sum of differences in Test Accuracy in the two categories.

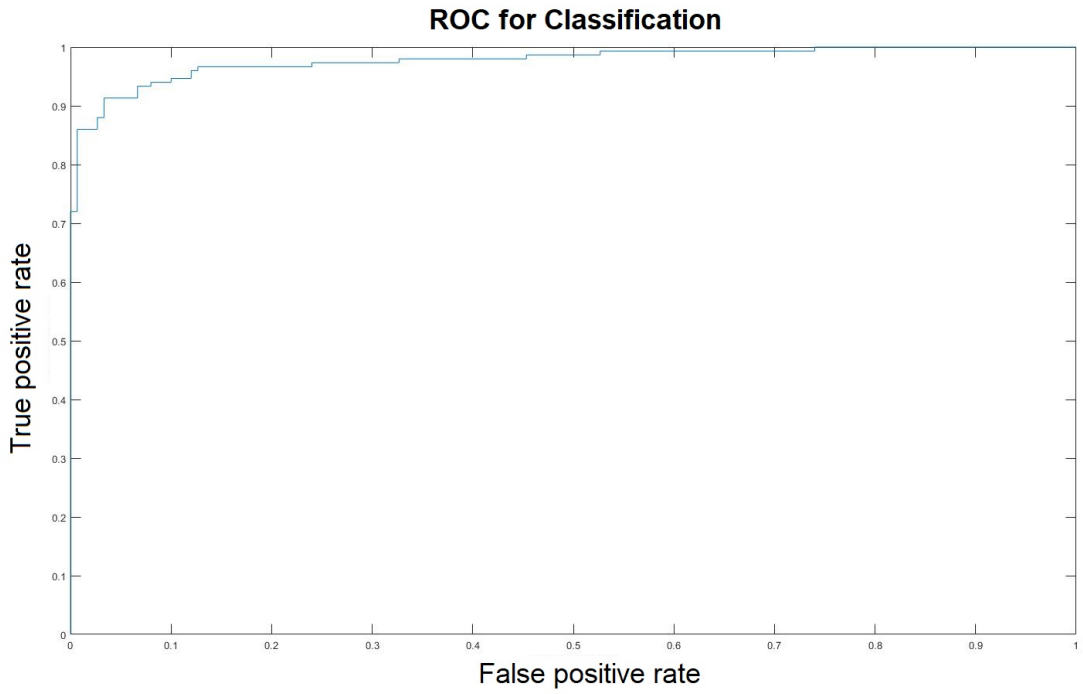


Figure 5.5: ROC Curve. On the x-axis, I represented the false positives. On the y-axis, I depicted the true positives. This graphical representation provides insight into the trade-off between these two rates and helps assess the model's performance in binary classification tasks.

$$\text{Total Distance} = \sum_{i=1}^{n} (\text{Test Accuracy Non-Rec}_i - \text{Test Accuracy Rec}_i) \tag{5.3}$$

Where:

- i: Number of datasets involved in the calculation. In my case, i=2 (concatenated and superposed).

- Test Accuracy Non-Rec: Value of Test Accuracy in the non-recombinant category.

- Test Accuracy Rec: Value of Test Accuracy in the recombinant category.

5.10 Robustness Criteria

In order to assess the robustness of the outcomes, the process of optimized training, utilizing an identical dataset, was iterated three times. If the results of these three iterations exhibited a high degree of consistency, the system could be deemed robust.

5.11 Explainability Tools' Result Generation

The application of explainability tools can provide information on which areas are determinant to the subsequent classification [223, 1897]. The results of the explainability analysis made possible to draw conclusions about the biological determination of a given feature. This is, which factors determined whether an HIV-1 sequence is recombinant or not.

The use of different explainability tools allowed to compare the areas of interest marked by each one of them and to complement their results in order to obtain the maximum possible information to draw conclusions. In this researching, three tools were applied:

1) Grad-CAM (Gradient-weighted Class Activation Mapping) is a technique for detecting the most important areas of an image so that the Deep Learning tool can classify it into a certain category. It uses gradients of the classification score with respect to the final convolutional feature map [225, 2455].

2) LIME (Local Interpretable Model-agnostic Explanations). This explainability tool computes feature importance by segmenting an image into several features and generating "artificial" observations by randomly including or excluding these features. If a concrete feature is excluded, the pixels corresponding to this feature are replaced with the value of the average image pixel. The network classifies these synthetic observations, and uses both the absence or presence of a feature and its corresponding scores for the predicted class to train a regression problem with a simpler model [1943, 2456, 2457].

3) Gradient Attribution points out with high accuracy the specific pixels of the image which are critical for the decision making of the Deep Learning tool. Returning a pixel map that highlights the most affecting pixels of the class score when changed [1945, 2458].

Grad-CAM is a very intuitive tool but with a low resolution, and LIME has variability in accuracy with a good visual explainability. Gradient Attribution indicates with a very high level of accuracy the determining pixels for classification, although its results can be difficult for the human eye to interpret (visually very noisy).

In terms of accuracy and ease of use for the human eye, the advantages and disadvantages of these three tools made it particularly interesting to work with all of them simultaneously to compare results and obtain complementary data on features and patterns.

The MATLAB functions used for each of these explainability tools were gradCAM, imageLIME, and gradientMap.

The output of these three functions is a scoremap, a structured numerical matrix aimed at accentuating and delineating pertinent regions within an input data or image that actively influence the outcome of a classification task. The scoremap serves as a critical component in the explainability and visualization of the model's decision-making process, allowing for the identification and assessment of salient features and patterns that inform the final classification. The score map is presented as an image, that is accomplished by superimposing a color map onto the original image, which highlights the areas of interest with prominent, visually discernible "hot" regions.

In the case of Grad-CAM and LIME, a 256-color Jet Color Map is applied, as depicted in Figure 4.3, ranging from 0 (deep blue) to 1 (deep red). The Gradient Attribution technique yields a pixel-level mapping, pinpointing the most pertinent pixels in the network's classification process, and is represented across a 255-color spectrum, varying from 0 (white) to 1 (black).

5.12 Phylogenetic Tree and Comparison with Explainability Tools' Patterns

A comparison was made between sequences with similar explainability patterns and the phylogenetic tree of the 13,554 HIV-1 complete sequences [2459, 2460].

First, I downloaded the HIV-1 Complete Consensus Multiple Sequence Alignment from the Los Alamos National Laboratory HIV-1 Database (until 2020). These alignments contain a single sequence per patient [2461]. These predefined alignments offer a significant advantage in the quality and utility of available genetic data. In this context, curation involves a meticulous selection and review process by the

global community of HIV-1 experts to ensure the accuracy and representativeness of genetic sequences. Furthermore, by conducting codon-level alignment, it ensures that the sequences correspond correctly at the coding level, which is essential for understanding genetic and protein structures. This results in reliable multiple sequence alignments, which can be considered optimal.

These alignments are comprehensive, encompassing all the complete sequences present in Los Alamos HIV-1 Database, with only a few exceptions:

- Each patient is represented by a single sequence.
- Highly similar sequences were removed. The decision to delete such sequences was made based on an examination of the distance distribution, and the cut-off point varies depending on the gene.
- Problematic sequences were excluded.

Problematic sequences are defined as those that exhibit a high presence of non-ACTG characters, meeting one of the following criteria:

- Contain more than 100 consecutive non-ACTG characters.
- Exhibit over 7% non-ACTG characters for sequences with a length of less than 1,000 nts.
- Display more than 5% non-ACTG characters for sequences with a length between 1,000 and 2,999 nts.
- Show more than 3% non-ACTG characters for sequences with a length of 3,000 nts or above.

Additionally, sequences are classified as problematic when there is a likelihood of contamination with a laboratory strain. And if they are hypermutant (G–>A hypermutation). Hypermutated sequences can carry substitutions that are not present in viable viruses, altering the branch lengths of the phylogenetic tree and complicating the determination of appropriate evolutionary models.

Synthetic samples are also excluded, as well as those containing an artifactual deletion of >100 nts, sequences shorter than 50 base pairs, or those deposited as the reverse complement (opposite) DNA strand.

These alignments were produced through an iterative procedure involving automated alignment using the software suite HMMER [2462], followed by manual editing with various software tools such as BioEdit [2463], Se-Al [2464], AliView [2465], and the Multiple Alignment Sequence Editor (MASE) [2466]. Most gaps were inserted in multiples of three bases to preserve ORFs. The alignment process represents a balance between achieving the best possible alignment, ensuring readability, and striving to maintain intact codons.

The protein alignments presented for each gene were created by utilizing both nucleotide sequences and their corresponding translated amino acid sequences. It is important to note that these translations, which are alignment-based, may vary from a direct, non-aligned translation. For example, an aligned translation takes into account frameshift compensation.

Sequences that are recognized as recombinants are typically marked as such, even if they do not exhibit recombination in the specific region being examined.

Hence, I prevent biases resulting from an over-representation of data from several sources and aim to create a more user-friendly phylogenetic tree.

Figure 5.6 displays the visual representation of the complete phylogenetic tree.

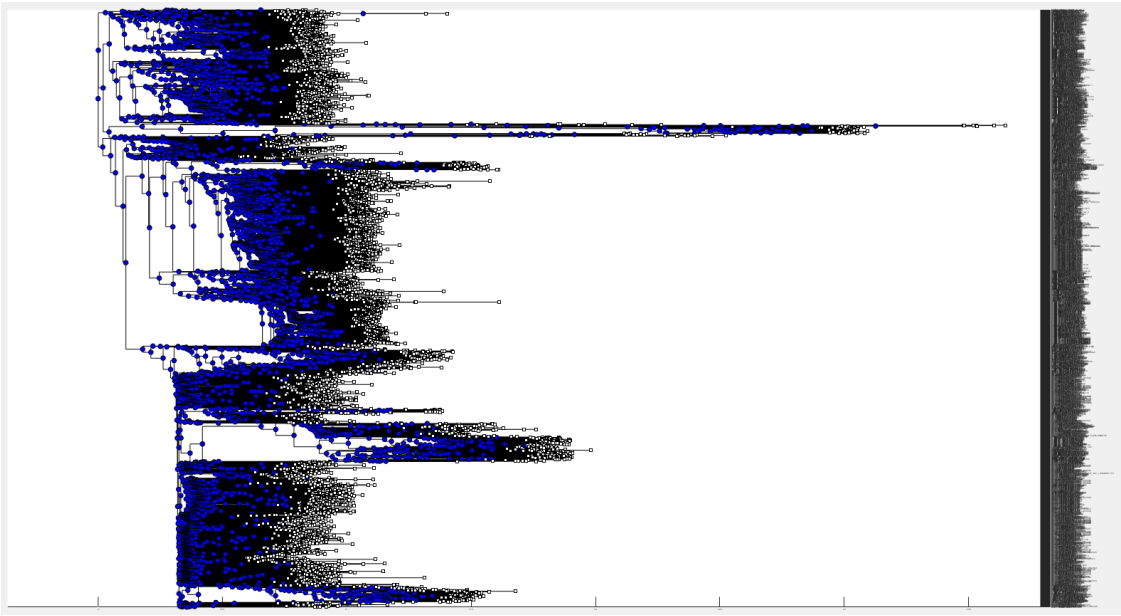


Figure 5.6: Phylogenetic Tree of HIV-1. This image is the visual representation of the phylogenetic tree generated from the Consensus Multiple Alignment of all complete HIV-1 sequences in the Los Alamos HIV-1 Database.

Subsequently, I conducted an analysis to assess the level of similarity among the patterns identified by explainability tools and their proximity within the phylogenetic tree. The construction of the phylogenetic tree based on these multiple sequence alignments and the calculation of phylogenetic distances between two sequences (Path Length) were carried out using MATLAB 2021b - Computational Biology Apps [2450].

In cases where the analyzed sequence was not present in the HIV-1 consensus multialignment, the representative sequence from the same patient was used as a point of reference. At times, it proved challenging to determine the placement of certain sequences within the phylogenetic tree due to the absence of either the sequence itself or the patient's reference sequence in the alignments. Additionally, in some instances, the patient's identity was not discernible from the sequences.

The traceability of intra-patient sequences (virus sequences produced within the same patient's organism) was established through the Patient Code (ID) field.

I imported the full HIV-1 Consensus Multiple Sequence Alignment into the Sequence Alignment Tool. The phylogenetic tree was automatically generated by utilizing the "View Tree" function. To view the phylogenetic path and distance between two sequences, you can select them directly on the tree.

5.13 Raw Results

In this Section, I provide the test bench results for each of the three selected representative pre-trained CNNs. At this stage, the optimal result should not only excel in terms of accuracy rates but should also demonstrate an optimal balance between both categories. In line with the criteria detailed in Section 5.8, I highlighted the best results in green.

To enhance the comprehension of Confusion Matrix Results in the upcoming sections, I exclusively presented the counts of correct classifications and errors, as depicted in Figure 5.7.

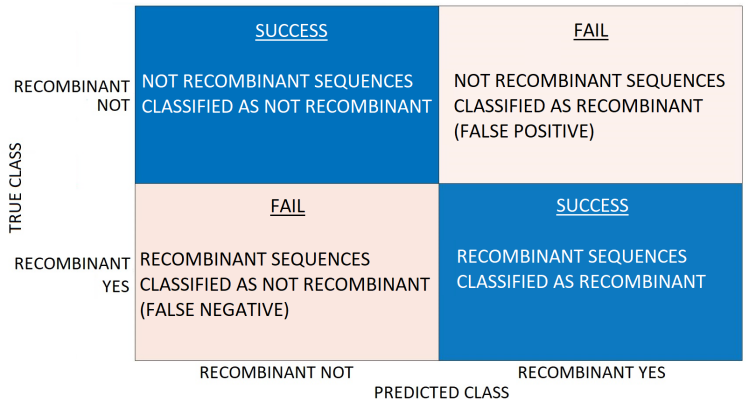


Figure 5.7: Confusion Matrix Scheme. The non-recombinants are located in the upper row, with hits located on the left and misses on the right. On the other hand, the lower row pertains to recombinants, with failures positioned on the left and hits on the right.

5.13.1 Concatenated Spectrograms

Table 5.5 shows the results obtained for the concatenated spectrograms dataset using VGG-16. For a Learning Rate of 0.0100, the network exhibited negligible performance in all scenarios. Notably, when employing a Batch Size of 128, there was a slight decrease in performance compared to using a Batch Size of 52. Moreover, extending the training to 30 Epochs, despite significantly increasing computational time and yielding slightly improved AUC values, resulted in a noticeable decline in accuracy.

Table 5.5: VGG-16 Concatenated Spectrograms Results. The best and most balanced levels of Confusion Matrix results and Validation Accuracy were highlighted in green.

Concatenated Spectrograms - VGG-16				
Description	**Validation Accuracy**	**Training Time**	**Confusion Matrix Results**	**AUC**
Learning Rate: 0.0001 Batch Size: 52 Epochs: 10	94.92%	17 min 36 sec	145 - 5 13 - 137	0.9772
Learning Rate: 0.0100 Batch Size: 52 Epochs: 10	82.99%	14 min 57 sec	150 0 150 0	0.5
Learning Rate: 0.0001 Batch Size: 52 Epochs: 30	96.18%	46 min 55 sec	147 3 22 128	0.9872
Learning Rate: 0.0100 Batch Size: 52 Epochs: 30	82.99%	18 sec	0 0 0 0	NaN
Learning Rate: 0.0001 Batch Size: 128 Epochs: 10	91.44%	16 min 39 sec	145 5 37 113	0.9610
Learning Rate: 0.0100 Batch Size: 128 Epochs: 10	82.99%	14 min 57 sec	150 0 150 0	0.5
Learning Rate: 0.0001 Batch Size: 128 Epochs: 30	94.01%	50 min 16 sec	144 6 26 124	0.9754
Learning Rate: 0.0100 Batch Size: 128 Epochs: 30	82.99%	49 min 17 sec	150 0 150 0	0.5

Table 5.6 shows the results obtained for the concatenated spectrograms dataset using ResNET-101. For a Learning Rate of 0.0001, the failure rate in the case of recombinants was disproportionately high compared to non-recombinants in all configurations. With a Learning Rate of 0.0100, the imbalance was reduced but still persisted. Results slightly deteriorated with a Batch Size of 128, as well as when increasing the number of Epochs to 30.

Table 5.6: ResNET-101 Results Concatenated Spectrograms. The configuration highlighted in green achieved the best Accuracy, Confusion Matrix results and AUC values, with noticeably reduced Training Time.

Concatenated Spectrograms - ResNET-101				
Description	**Validation Accuracy**	**Training Time**	**Confusion Matrix Results**	**AUC**
Learning Rate: 0.0001 Batch Size: 52 Epochs: 10	86.91%	44 min 56 sec	147 3 51 99	0.9571
Learning Rate: 0.01 Batch Size: 52 Epochs: 10	94.82%	41 min 7 sec	149 1 16 134	0.9928
Learning Rate: 0.0001 Batch Size: 52 Epochs: 30	88.42%	125 min 11 sec	149 1 57 93	0.9526
Learning Rate: 0.0100 Batch Size: 52 Epochs: 30	94.26%	131 min 48 sec	148 2 23 127	0.9853
Learning Rate: 0.0001 Batch Size: 128 Epochs: 10	84.20%	28 min 40 sec	148 2 75 75	0.8961
Learning Rate: 0.0100 Batch Size: 128 Epochs: 10	92.70%	27 min 9 sec	148 2 17 133	0.9928
Learning Rate: 0.0001 Batch Size: 128 Epochs: 30	86.76%	83 min 31 sec	149 1 65 85	0.9266
Learning Rate: 0.01 Batch Size: 128 Epochs: 30	92.10%	81 min 54 sec	145 5 18 132	0.9833

The results obtained using Inception V3 with concatenated spectrograms are detailed in Table 5.7. In all configurations where the Learning Rate was set at 0.0001, the failure rate for recombinants was significantly high and unbalanced compared to non-recombinants. Across all configurations with a Batch Size of 128, the hit rate was slightly worse. Training with 30 Epochs yielded optimal results compared to 10 Epochs.

Table 5.7: Inception V3 Results Concatenated Spectrograms. The configuration that achieved the best results was highlighted in green.

Concatenated Spectrograms - Inception V3				
Description	**Validation Accuracy**	**Training Time**	**Confusion Matrix Results**	**AUC**
Learning Rate: 0.0001 Batch Size: 52 Epochs: 10	89.38%	48 min 28 sec	143 7 43 107	0.9226
Learning Rate: 0.0100 Batch Size: 52 Epochs: 10	92.60%	49 min 7 sec	149 1 13 137	0.9863
Learning Rate: 0.0001 Batch Size: 52 Epochs: 30	89.08%	128 min 18 sec	145 5 38 112	0.9499
Learning Rate: 0.01 Batch Size: 52 Epochs: 30	94.82%	148 min 1 sec	146 4 7 143	0.9943
Learning Rate: 0.0001 Batch Size: 128 Epochs: 10	88.68%	33 min 54 sec	145 5 54 96	0.8729
Learning Rate: 0.0100 Batch Size: 128 Epochs: 10	93.21%	34 min 54 sec	147 3 21 129	0.9850
Learning Rate: 0.0001 Batch Size: 128 Epochs: 30	87.82%	102 min 6 sec	141 9 48 102	0.9075
Learning Rate: 0.0100 Batch Size: 128 Epochs: 30	93.51%	109 min 35 sec	147 3 14 136	0.9867

5.13.2 Superposed Spectrograms

In the case of superposed spectrograms, the results obtained with VGG-16 are shown in Table 5.8. For a Learning Rate of 0.0100, the network demonstrated poor performance across all scenarios. It is worth noting that when a Batch Size of 128 was employed, there was a slight decline in performance compared to a batch size of 52. Furthermore, extending the training duration to 30 Epochs, while significantly increasing computational time, led to a very slight deterioration in the AUC values and Confusion Matrix Results.

Table 5.8: VGG-16 Results Superposed Spectrograms. The optimal and most balanced levels of total accuracy were indicated in green.

Superposed Spectrograms - VGG-16				
Description	**Validation Accuracy**	**Training Time**	**Confusion Matrix Results**	**AUC**
Learning Rate: 0.0001 Batch Size: 52 Epochs: 10	92.90%	15 min 59 sec	143 7 11 139	0.9888
Learning Rate: 0.0100 Batch Size: 52 Epochs: 10	82.99%	21 sec	0 0 0 0	NaN
Learning Rate: 0.0001 Batch Size: 52 Epochs: 30	95.42%	49 min 15 sec	146 4 22 128	0.9861
Learning Rate: 0.0100 Batch Size: 52 Epochs: 30	82.99%	21 sec	0 0 0 0	NaN
Learning Rate: 0.0001 Batch Size: 128 Epochs: 10	94.87%	17 min 21 sec	146 4 18 132	0.9847
Learning Rate: 0.0100 Batch Size: 128 Epochs: 10	82.99%	23 sec	0 0 0 0	NaN
Learning Rate: 0.0001 Batch Size: 128 Epochs: 30	95.82%	51 min 17 sec	148 2 24 126	0.9876
Learning Rate: 0.0100 Batch Size: 128 Epochs: 30	82.99%	23 sec	0 0 0 0	NaN

Table 5.9 shows the results obtained in the case of ResNET-101 applied to superposed spectrograms. With a Learning Rate of 0.0001, there was a significantly higher failure rate for recombinants compared to non-recombinants across all configurations. While a Learning Rate of 0.0100 mitigated this imbalance to some extent, it still lingered. There was a slight decline in results when using a Batch Size of 128 and extending the training duration to 30 Epochs.

Table 5.9: ResNET-101 V3 Results Superposed Spectrograms. The configuration marked in green attained the highest accuracy and AUC values while also considerably reducing the Training Time.

Superposed Spectrograms - ResNET-101				
Description	**Validation Accuracy**	**Training Time**	**Confusion Matrix Results**	**AUC**
Learning Rate: 0.0001 Batch Size: 52 Epochs: 10	90.99%	47 min 11 sec	146 4 55 95	0.9202
Learning Rate: 0.0100 Batch Size: 52 Epochs: 10	96.68%	42 min 55 sec	148 2 19 131	0.9902
Learning Rate: 0.0001 Batch Size: 52 Epochs: 30	91.39%	101 min 44 sec	143 7 51 99	0.9238
Learning Rate: 0.01 Batch Size: 52 Epochs: 30	96.83%	136 min 24 sec	147 3 15 135	0.9899
Learning Rate: 0.0001 Batch Size: 128 Epochs: 10	89.13%	24 min 19 sec	145 5 68 82	0.9094
Learning Rate: 0.0100 Batch Size: 128 Epochs: 10	94.51%	30 min 31 sec	148 2 34 116	0.9832
Learning Rate: 0.0001 Batch Size: 128 Epochs: 30	90.29%	73 min 15 sec	148 2 56 94	0.9127
Learning Rate: 0.0100 Batch Size: 128 Epochs: 30	95.47%	91 min 49 sec	148 2 32 118	0.9860

Table 5.10 details the set of results obtained using Inception V3 on the superposed spectrograms dataset. In all setups employing a Learning Rate of 0.0001, the misclassification rate for recombinant sequences was markedly elevated and unevenly distributed when compared to non-recombinant sequences. Across all scenarios with a Batch Size of 128, the performance showed a slight decline. Extending the training to 30 Epochs yielded superior outcomes compared to training for only 10 Epochs.

Table 5.10: Inception V3 Results Superposed Spectrograms. The configuration that attained the most favorable results was indicated in green.

Superposed Spectrograms - Inception V3				
Description	**Validation Accuracy**	**Training Time**	**Confusion Matrix Results**	**AUC**
Learning Rate: 0.0001 Batch Size: 52 Epochs: 10	89.73%	49 min 33 sec	147 3 49 101	0.9480
Learning Rate: 0.0100 Batch Size: 52 Epochs: 10	96.38%	48 min 49 sec	148 2 20 130	0.9924
Learning Rate: 0.0001 Batch Size: 52 Epochs: 30	91.29%	134 min 2 sec	147 3 45 105	0.9599
Learning Rate: 0.01 Batch Size: 52 Epochs: 30	96.48%	149 min 6 sec	149 1 14 136	0.9879
Learning Rate: 0.0001 Batch Size: 128 Epochs: 10	86.86%	35 min 43 sec	147 3 84 66	0.9002
Learning Rate: 0.0100 Batch Size: 128 Epochs: 10	95.37%	35 min 7 sec	148 2 23 127	0.9928
Learning Rate: 0.0001 Batch Size: 128 Epochs: 30	89.53%	101 min 45 sec	144 6 48 102	0.9288
Learning Rate: 0.0100 Batch Size: 128 Epochs: 30	95.62%	104 min 41 sec	149 1 22 128	0.9944

5.14 Results Processing

5.14.1 CNN Selection

As illustrated in Table 5.11, for both concatenated and superposed spectrograms, it is noteworthy that:

- VGG-16 demonstrated superior performance when trained with a Learning Rate of 0.0001.

- ResNET-101 and Inception V3 achieved their optimal performance with a Learning Rate of 0.0100.

Table 5.11: Summary of Configurations that Achieved the Best Results

Concatenated Spectrogram Dataset							
CNN	**Hyperparameters**	**Validation Accuracy**	**Training Time**	**Training Time per Epoch**	**Test Accuracy**	**Confusion Matrix Results**	**AUC**
VGG-16	Learning Rate: 0.0001 Batch Size: 52 Epochs: 10	94.92%	17 min 36 sec	1 min 45.60 sec	94.00%	145 5 13 137	0.9772
Resnet-101	Learning Rate: 0.0100 Batch Size: 52 Epochs: 10	94.82%	41 min 7 sec	4 min 6.70 sec	94.33%	149 1 16 134	0.9928
Inception-V3	Learning Rate: 0.0100 Batch Size: 52 Epochs: 30	94.82%	148 min 1 sec	4 min 56.03 sec	96.33%	146 4 7 143	0.9943
Superposed Spectrogram Dataset							
CNN	**Hyperparameters**	**Validation Accuracy**	**Training Time**	**Training Time per Epoch**	**Test Accuracy**	**Confusion Matrix Results**	**AUC**
VGG-16	Learning Rate: 0.0001 Batch Size: 52 Epochs: 10	92.90%	15 min 59 sec	1 min 35.90 sec	94.00%	143 7 11 139	0.9888
Resnet-101	Learning Rate: 0.0100 Batch Size: 52 Epochs: 30	96.83%	136 min 24 sec	4 min 32.80 sec	92.67%	143 7 15 135	0.9899
Inception-V3	Learning Rate: 0.0100 Batch Size: 52 Epochs: 30	96.48%	149 min 6 sec	4 min 58.20 sec	95.00%	149 1 14 136	0.9879

In all three cases, a reduction in the Batch Size from 128 to 52 led to performance enhancements for the respective CNNs.

These three CNNs demonstrated consistent behavior as the number of epochs increased. The accuracy rate exhibited a linear, direct relationship with the number of epochs when classifying non-recombinant sequences, achieving a 100% accuracy rate. Conversely, for recombinant sequences, the accuracy rate increased until it reached its maximum, after which it started to decline.

VGG-16 achieved superior outcomes with an epoch count of 10 compared to 30. In the case of ResNET-101, better performance was observed with an epoch count of 10 when working with concatenated spectrograms. However, when dealing with superposed spectrograms, an epoch count of 30 proved to be more effective. Inception V3, for both datasets, exhibited a slight performance improvement when using an epoch count of 30.

The Training Time for VGG-16 was the shortest among the three models. On the other hand, Inception V3 exhibited a slower learning process. Inception V3's slower learning process may be due to its more intricate architecture and may require longer Training Time to reach optimal performance.

Due to Res-NET 101 consistently yielding the lowest accuracy for both datasets, it was excluded from further consideration. Inception V3, on the other hand, demonstrated superior performance when processing concatenated spectrograms, albeit with a slightly diminished accuracy when dealing with superposed recombinant spectrograms.

Therefore, in accordance with the performance measurement parameters defined in Section 5.9, VGG-16 was the CNN selected for this research.

As seen in Table 5.12, for similar ranges of accuracy, VGG-16 obtained the lowest Total Distance value among the three models. So, it yielded more balanced hit rates on the test set between non-recombinant and recombinant categories, and its training times were substantially lower. The best success rate and the most balanced results were achieved with 10 Epochs.

Table 5.12: Total Distances. The column "Cat." refers to the two classification categories, where "NR" stands for non-recombinant and "R" for recombinant. The column "Test Acc." stands for Test Accuracy. The column "Diff." refers to the difference between the test accuracy for the recombinant category and the test accuracy for the non-recombinant category. "Total Distance" is the sum of the "Diff." values in concatenated and superposed.

	Cat.	**Concatenated**		**Superposed**		
		Test Acc.	**Diff.**	**Test Acc.**	**Diff.**	**Total Distance**
VGG-16	NR	96.67%	5.33%	95.33%	2.67%	8.00%
	R	91.33%		92.67%		
ResNET-101	NR	99.33%	10.00%	98.00%	8.00%	18.00%
	R	89.33%		90.00%		
Inception V3	NR	97.33%	2.00%	99.33%	8.67%	10.67%
	R	95.33%		90.67%		

The optimal configuration for VGG-16, suitable for both concatenated and superposed spectrograms, featured the following settings:

- Learning Rate = 0.0001
- Batch Size = 52
- Epochs = 10

5.14.2 Robustness

I replicated the experiment three times with concatenated and superposed spectrograms using the selected configuration of VGG-16, following the criteria outlined in Section 5.10.

Table 5.13: Test Results of Three Replicates - Concatenated Spectrograms. Performed with VGG-16 and its best resulting configuration: Learning Rate = 0.0001, Batch Size = 52 and Epochs = 10.

Concatenated Spectrogram Dataset					
VGG-16	**Validation Accuracy**	**Training Time**	**Test Accuracy**	**Confusion Matrix Results**	**AUC**
Test 1	94.92%	17 min 36 sec	94.00%	145 5 13 137	0.9772
Test 2	95.82%	15 min 51 sec	92.67%	144 6 16 134	0.9828
Test 3	95.02%	17 min 36 sec	90.33%	142 8 21 129	0.9714

Table 5.14: Test Results of Three Replicates - Superposed Spectrograms. Performed with VGG-16 and its best resulting configuration, which ass the same for concatenated spectrograms: Learning Rate = 0.0001, Batch Size = 52, and Epochs = 10.

Superposed Spectrogram Dataset					
VGG-16	**Validation Accuracy**	**Training Time**	**Test Accuracy**	**Confusion Matrix Results**	**AUC**
Test 1	92.90%	15 min 59 sec	94.00%	143 7 11 139	0.9888
Test 2	92.25%	16 min 9 sec	93.33%	144 6 14 136	0.9904
Test 3	94.46%	16 min 39 sec	93.33%	144 6 14 136	0.9872

The outcomes exhibited a high degree of consistency across all three instances for both datasets, as depicted in Tables 5.13 and 5.14, affirming the robustness of the performance. The selection of the most effective training regimen adhered to the criteria delineated in Section 5.9. Notably, in both datasets, Test 1 emerged as the optimal choice, achieving superior hit rates and a remarkable balance between recombinants and non-recombinants.

5.14.3 Hot Zones Determination

Explainability tools facilitated the localization of regions within the CNN that contribute to the determination of a sequence's recombinant features. These specific regions within the spectrogram are commonly referred to as "hot zones", representing the focal points that the CNN scrutinizes during the classification process.

Figures 5.8 and 5.9 depict the prevalent regions of interest identified through explainability analysis. To discern and extract the most frequently occurring hot zones with the utmost precision, an individual scrutiny of the explainability results was conducted for each sequence within the test set, encompassing a total of 300 sequences. For the Test Set sequences that were accurately classified by the CNN, I isolated the most commonly recurring primary hot zones as well as secondary ones. In the case of those sequences where the network's predictions were incorrect, I identified the most frequently recurring erroneous primary hot zones as well as secondary ones.

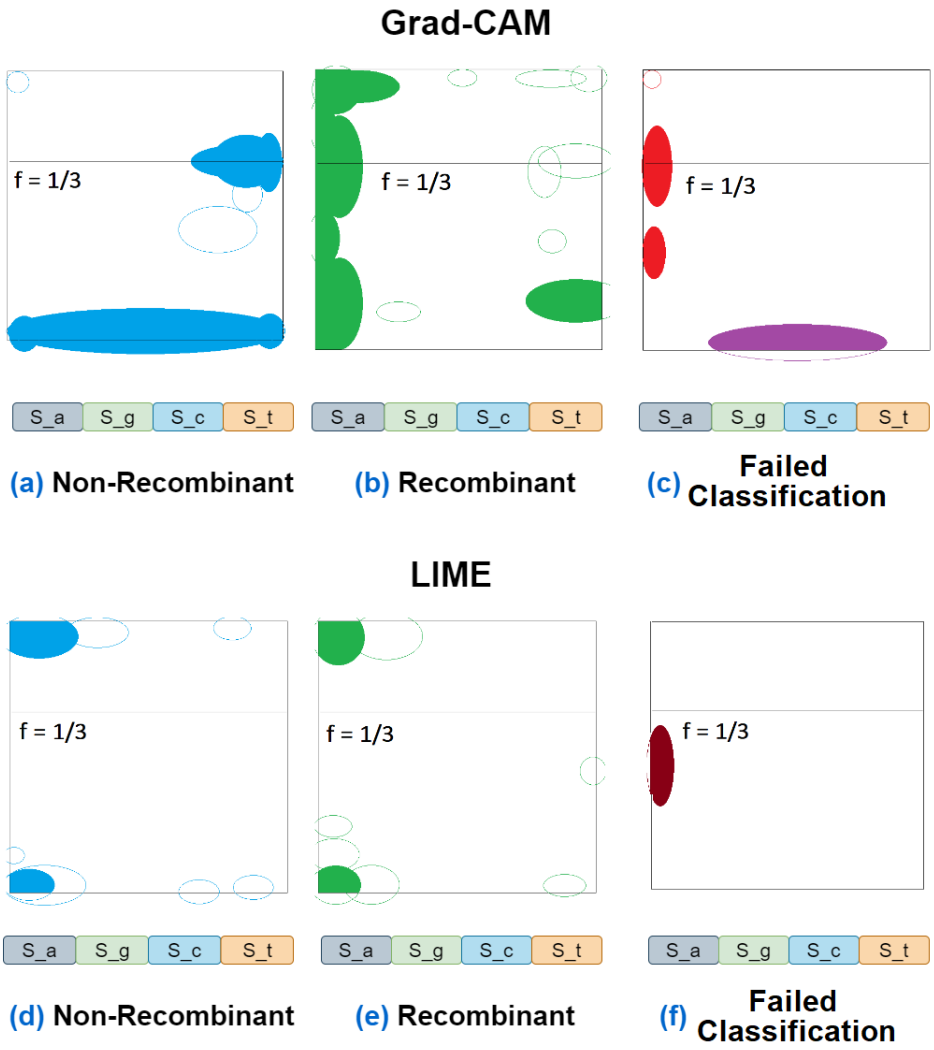


Figure 5.8: Hot Zones Interpreted by Grad-CAM and LIME on Concatenated Spectrograms Classified as Non-Recombinants (a), Recombinants (b), or Misclassified (c). The prevalent hot areas are colored, signifying a higher level of consistency in these regions. Conversely, the unfilled areas represent hot zones with lower reoccurrence. In the context of Grad-CAM, erroneously classified spectrograms (c), hot zones of "Recombinant classified as non-recombinant" are denoted in purple, whereas those corresponding to "Non-recombinant classified as recombinant" are represented in red. As for LIME, when the CNN's classification is erroneous (f), hot zones are highlighted in brown.

The Grad-CAM and LIME results, while exhibiting somewhat lower accuracy, offer an advantage in that they readily present clearly identifiable hot zones that are discernible to the human eye. Consequently, they are particularly well-suited for the initial assessment and identification of regions of interest. In contrast, Gradient Attribution, while a valuable technique, often presents challenges in terms of visual interpretation. It tends to generate visual representations that require a higher level of expertise and interpretation to identify the relevant areas, making it a better choice for advanced analysis but potentially less intuitive for an initial approximation.

The analysis of Grad-CAM hot zones revealed distinct patterns in the concatenated spectrograms (Figure 5.8). Notably, within non-recombinant sequences (Figure 5.8a), the low-frequency range encompassing all four nucleotides, along with frequencies proximate to one-third of the T nucleotide, emerged as critical factors influencing the classification as non-recombinant. In the case of recombinant concatenated spectrograms (Figure 5.8b), the primary Grad-CAM hot zone predominantly encompassed the left border, corresponding to the A nucleotide region, and extended to the lower third of the right border, associated with the T nucleotide area. Additionally, a shared zone of interest was observed for both non-recombinant and recombinant sequences, positioned on the right border at the one-third mark of the T-nucleotide area. In instances of misclassification, the Convolutional Neural Network (CNN) directed its attention to areas corresponding to the opposite category. The most commonly recurring regions associated with classification errors are illustrated in Figure 5.8c.

In the context of LIME applied to concatenated spectrograms (Figures 5.8d, e, and f), LIME consistently emphasized the left corners, characterized by high and low-frequency regions associated with the A nucleotide, for both non-recombinant (Figure 5.8d) and recombinant sequences (Figure 5.8e). However, in several cases of classification failure, the CNN focused on locations that did not align with the established hot zones. This region of erroneous focus was situated around the midpoint of the left axis, corresponding to the average A nucleotide frequencies (Figure 5.8f).

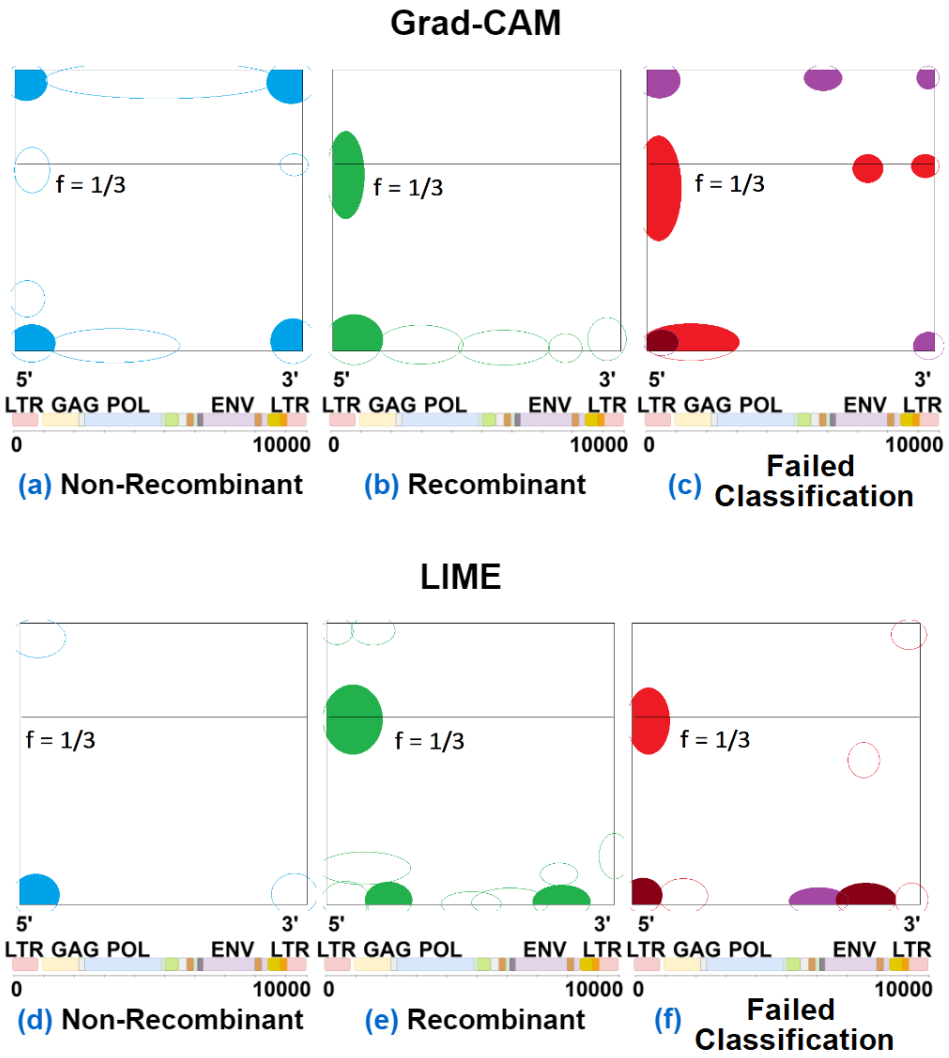


Figure 5.9: Hot Zones Interpreted by Grad-CAM and LIME on Superposed Spectrograms Classified as Non-Recombinants (a), Recombinants (b), or Misclassified (c). The hot areas that recurrently appear are represented in filled colors, indicating a higher degree of consistency in these regions. Conversely, unfilled areas denote hot zones with lower repeatability. Figures (c) and (f) indicate the areas where the CNN's focus was directed in cases of misclassification. In these instances, purple-colored hot zones signify "Recombinant classified as non-recombinant", while red-colored zones represent "Non-recombinant classified as recombinant". Notably, the brown-colored area indicates a common error zone shared between the two types of misclassification.

The Grad-CAM hot zones of superposed spectrograms are illustrated in Figure 5.9. The predominant hot zone for non-recombinant superposed spectrograms pertained to the high and low-frequency regions, primarily corresponding to the 5'LTR and 3'LTR of the HIV-1 genome, and are located at the four corners of the spectrogram. Secondary hot zones were concentrated along the upper and lower edges as well as in regions with high and low frequencies. As shown in Figure 5.9b, the primary influential regions of recombinant genomes identified by Grad-CAM were situated within the low-frequency range and in the proximity to $f = 1/3$ at the 5'LTR region.

In non-recombinant superposed spectrograms, LIME exhibited a pronounced focus in the lower-left corner, primarily associated with the low-frequency region at the sequence's outset, specifically the 5'LTR (Figure 5.9d). To a lesser extent, LIME also displayed activity in the upper-left corner, corresponding to the high-frequency area at the sequence's beginning, as well as in the lower-right corner, linked to the low-frequency region at the sequence's end, which corresponds to the 3'LTR. In the superposed spectrograms of recombinant sequences (Figure 5.9e), the prominent regions were located around the one-third mark of the frequency range at the sequence's onset and in regions closely associated with the *gag* and *env* genes.

In superposed spectrograms, both Grad-CAM and LIME indicated that the CNN misclassified by focusing on the hot zones of the opposite category (Figures 5.9c and 5.9f, respectively).

The performance of this two explainability tools exhibited consistency in both concatenated and superposed spectrograms. Both approaches yielded reliable results and demonstrate concordant areas of interest within the neighboring sequences.

5.15 Biological Consistency Analysis

As described in Section 5.12, once I established that the CNN could effectively distinguish between recombinant and non-recombinant sequences through discernible decision patterns, my subsequent objective was to delve deeper into the biological implications of this phenomenon. This involved constructing a phylogenetic

tree including all HIV-1 complete sequences up to 2020. After this, I assessed the correlation between the regions of heightened activity (hot zones) and their phylogenetic divergence. Additionally, I cross-referenced these hot zones with expert microbiological knowledge to corroborate the biological consistency of my findings.

To enhance the biological interpretation of the hot regions, I present multiple instances of specific sequences and their corresponding hot areas in Sections 5.15.1, 5.15.2, and 5.15.3. These examples are accompanied by a scaled reference for visual verification of the results obtained, along with detailed discussions.

5.15.1 Analysis of Successful Classifications

The hot areas of sequences belonging to the same subtype were usually similar. I addressed this observation through an analysis divided into four distinct study groups, each designed to explore specific aspects of my dataset and gain a better understanding of the relationship between genomic spectrograms and non-recombinant and recombinant sequences. These groups were divided as follows:

- Concatenated spectrograms, non-recombinant sequences (Table 5.15).
- Concatenated spectrograms, recombinant sequences (Table 5.16).
- Superposed spectrograms, non-recombinant sequences (Table 5.17).
- Superposed spectrograms, recombinant sequences (Table 5.18).

First, I present a selection of distinct instances of non-recombinant sequences belonging to the same subtype, showcasing analogous hot zones within the context of concatenated spectrograms.

Table 5.15 graphically illustrates the first example of the detected proportional relationship between phylogenetic distance (measure of evolutionary divergence between two sequences) and the degree of similarity of the hot zones.

Table 5.15: Comparison of Explainability of Non-Recombinant Concatenated Spectrograms. Each example includes two non-recombinant sequences with their IDs, subtypes, and hot zones from Grad-CAM, LIME, and Gradient Attribution, referenced to a scaled concatenated spectrogram image.

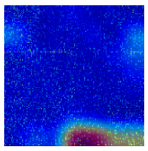
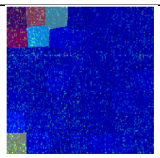
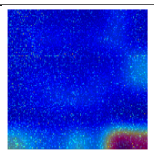
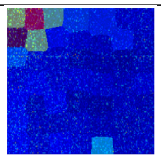
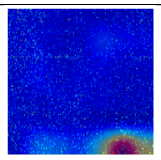
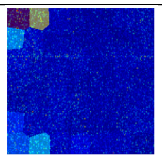
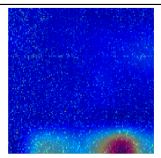
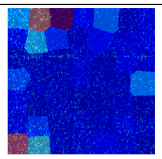
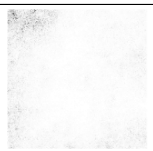
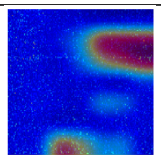
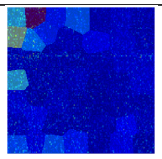
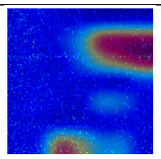
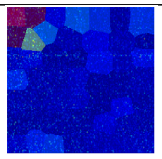
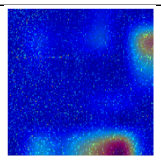
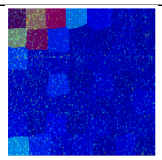
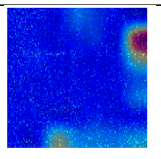
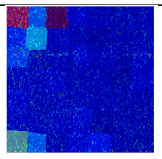
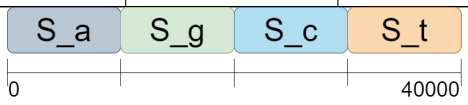
Subtype	Sequence ID	Grad-CAM	LIME	Gradient Attribution
A1	AF457081			
A1	KP223847			
	Reference			
A6	EF589043			
A6	EU861977			
	Reference			
B	MW061642			
B	MW061672			
	Reference			
O	KU168283			
O	MH705149			
	Reference			

Table 5.15 – Continuation from the previous page

Subtype	Sequence ID	Grad-CAM	LIME	Gradient Attribution
				

Analyzing the example corresponding to the sequences subtype A1 AF457081 and KP223847, in Grad-CAM, the main hot zone was centered on the low frequencies around the G and T nucleotides. However, LIME results focused on the high frequencies of A. Gradient Attribution agreed with this result, although it also slightly emphasized the low frequencies of A.

In this case, sequence KP223847 is not found in the phylogenetic tree as it lacks a complete intra-patient reference sequence. Therefore, it was not possible to calculate the phylogenetic distance between them. However, their membership in the same subgroup indicates a close phylogenetic relationship. This suggests that, despite the lack of precise phylogenetic distance data, these sequences share significant evolutionary proximity due to their grouping within the same subgroup.

The classification of sequences into subtypes is based on notable genetic similarities among the members of each subtype. Therefore, the membership of two sequences in the same subtype implies the presence of common genetic features and a shared ancestor in their evolutionary history. Although it is not possible to quantify the exact phylogenetic distance in this particular scenario, the shared membership in the same subtype strengthens the hypothesis that these sequences are related from an evolutionary perspective.

The next representative example corresponded to A6 subtype EF589043 and EU861977 sequences. Grad-CAM clearly pointed to low frequencies in the T nucleotide. LIME focused on high frequencies of A, especially in the case of EF589043, and G in the case of EU861977. Gradient Attribution primarily targeted the range corresponding to A, especially in the high-frequency region. In the case of EF589043, the lower half of the frequency range also held significance.

EF589043 (Patient ID: 300480) is not present in the phylogenetic tree and lacks a complete intra-patient reference sequence. Despite the absence of a complete reference and precise phylogenetic distance data, these sequences share a close evolutionary relationship and are classified in the same subtype.

The example of subtype B MW061642 and MW061672 sequences was very significant. I emphasize the high level of similarity in the hot zones of both frequencies. Grad-CAM showed a strong hot zone around $f = 1/3$ in T and, to a lesser extent, in G. There was a second significant hot zone in the low frequencies of G. LIME focused on the high frequencies of A, and Gradient Attribution also pointed to $f = 1/3$ in T and G, as well as the high and low frequencies of A.

Both sequences are intra-patient (Patient ID: 113317), so their phylogenetic distance is minimal. Indeed, the similarity level between the hot zones of both sequences is extremely high.

The final representative example in the case of non-recombinant sequences in concatenated spectrograms was subtype O KU168283 and MH705149. In both sequences, there was a main hot zone perfectly located at $f = 1/3$ in T. KU168283 exhibited another hot zone clearly identified in the low frequencies around C. LIME once again pointed to the high frequencies of A in both sequences, as well as G. Gradient Attribution highlighted the high frequencies of A as the most significant hot zone.

The sequence KU168283 and the sequence MH705149 (within subtype O both of them), presented similar explainability hot areas. Indeed, as seen in Figure 5.10, the proximity of both sequences (highlighted in red to facilitate their identification) can be seen in the phylogenetic tree extract.

As I exposed, the similarity level is the highest in intra-patient sequences, those virus sequences produced in the same patients organism. Since the phylogenetic tree only includes one representative intra-patient sequence, it is not possible to include an excerpt from it. Obviously, the phylogenetic distance is minimal in the case of these intra-patient sequences. So it is to be expected that explainability

tools will yield even more similar results. Such is the case of subtype B intra-patient sequences MW061642 and MW061672 (Figure 5.15).

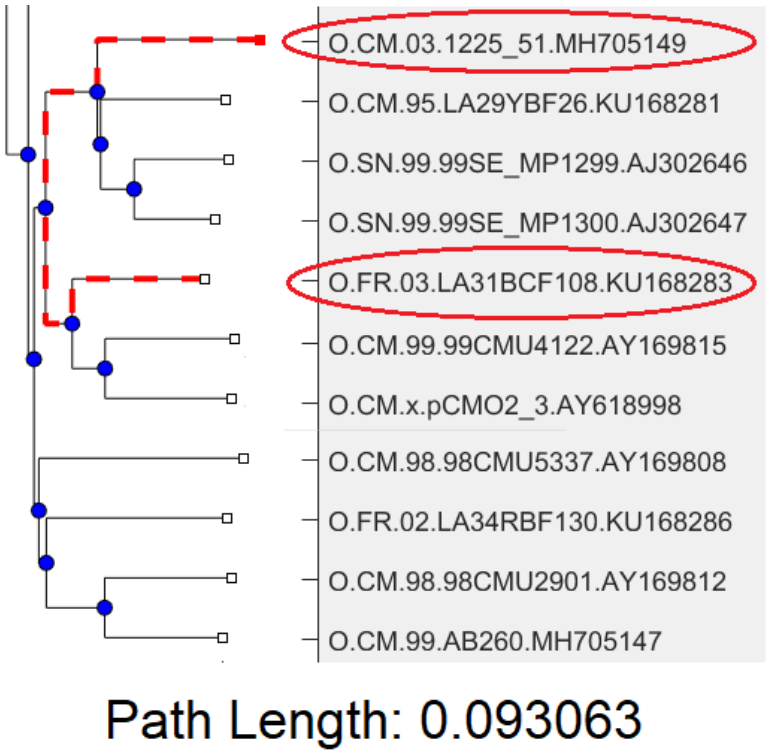


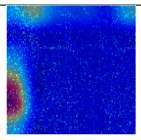
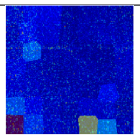
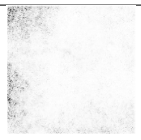
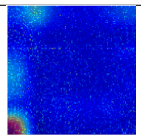
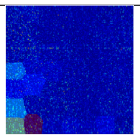
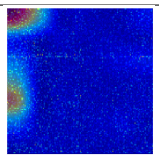
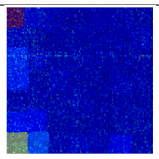
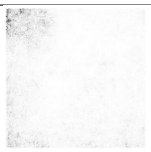
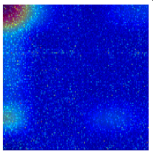
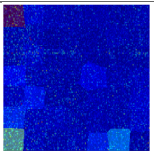
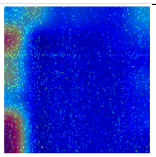
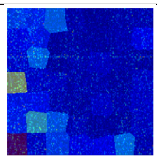
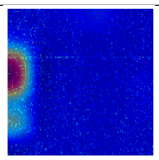
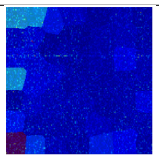
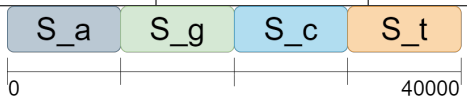
Figure 5.10: Partial Phylogenetic Tree Including Subtype O MH705149 and KU168283 Sequences. Whose non-recombinant subtype O sequences' path distance is 0.093063. Their clade (seven sequences in total) shares a common node with a subtype O clade of four sequences.

Secondly, I present several examples of the same phenomenon in recombinant sequences (concatenated spectrograms).

A representative example was subtype CRF22_01A1 AY371165 and KF716462 sequences, the latter being considered as its reference sequence JN864058 (Table 5.16). The phylogenetic closeness between the two sequences is shown in Figure 5.11, and their path is highlighted in red.

Subtype CRF22_01A1 AY371165 and KF716462 sequences have close phylogenetic positions and their hot zones are similar. Grad-CAM pointed to low and one-sixth frequencies of A. Another hot zone was at high frequencies of A and the beginning of G. LIME basically pointed to low frequencies of A. Gradient Attribution was noisy but visually centered in A nucleotide area and T (low frequencies).

Table 5.16: Comparison of Explainability Recombinant Concatenated Spectrograms. Each example includes two recombinant sequences identified by their ID and subtype, along with their respective hot zones identified by Grad-CAM, LIME, and Gradient Attribution, mapped to a scaled image of the concatenated spectrogram composition.

Subtype	Sequence ID	Grad-CAM	LIME	Gradient Attribution
CRF22_01A1	AY371165			
CRF22_01A1	KF716462			
	Reference			
CRF01_AE	MH078560			
CRF01_AE	MH327763			
	Reference			
CRF106_cpx	MT276998			
CRF106_cpx	MT277002			
	Reference			
				

From a phylogenetic perspective, both sequences are very similar as they are closely related and cluster within a region of the phylogenetic tree composed of sequences of the same subtype, known as CRF22_01A1. This subtype is the result of recombination between the recombinant subtype CRF01_AE and the non-recombinant subtype A1. Therefore, it is understandable that these two sequences exhibit a high degree of relatedness, given their common origin in this recombination event.

The similarity of the hot zones in both sequences is also consistent with the microbiological evidence presented.

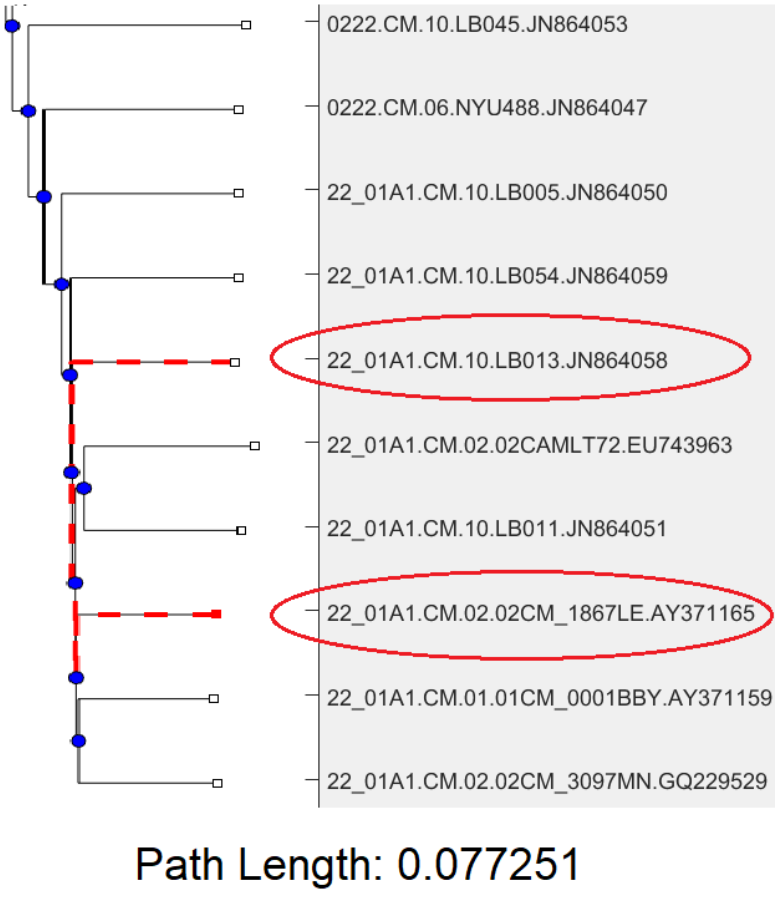


Figure 5.11: Partial Phylogenetic Tree Including Subtype CRF22_01A1 AY371165 and JN864058 Sequences. These sequences are included in the recombinant subtype CRF22_01A1, which contains 21 sequences. As shown, both are phylogenetically close sequences, with a path distance of 0.077251. Their last common node results in a clade of seven sequences.

The two sequences affiliated with the CRF01_AE subtype (MH078560 and

MH327763) exhibited a notably high level of concordance of their hot zones (Table 5.16). In both sequences, Grad-CAM pointed to the high-frequency region in A, with a second significant hot zone around $f = 1/6$ approximately for the MH078560 case. LIME similarly indicated the high frequencies of A in both cases, and, as a secondary area, the low frequencies of A as well. And Gradient Attribution clearly highlighted the high frequencies of A in both cases. Figure 5.12 shows the excerpt of the phylogenetic tree where both sequences are located.

The CRF01_AE subtype contains the highest number of sequences. Its prevalence in certain regions of Asia, coupled with its geographic diversity in that area, can lead to the independent evolution of different lineages of this subtype due to various evolutionary pressures [2467].

The proximity in the divergence from a recent common ancestor results in both sequences sharing many genetic and morphological similarities. Therefore, the coherence between the high similarity level in the hot zones calculated using explainability tools and the remarkably small phylogenetic distance between both sequences (0.061158) is striking.

In the case of the CRF106_cpx MT276998 and MT277002 sequences, despite a greater variability in Grad-CAM results, in both sequences, the importance was focused on A. In the case of MT276998, it was more focused on the range of low frequencies and around $f = 1/3$, and in the case of MT277002, it was highly concentrated around $f = 1/3$. The results from LIME were more consistent, centered on the low frequencies of A. In the case of Gradient Attribution, there was a common main hot zone located in the high frequencies of A, although in the case of MT277002, there were more notable areas across the entire frequency range in A.

Subtype CRF106_cpx MT276998 and MT277002 sequences, though next to each other in the phylogenetic tree (Figure 5.13), had divergent hot zones in their concatenated spectrograms (Table 5.16).

In Figure 5.13, I present the case with the closest phylogenetic proximity in this Chapter, with a path length of only 0.054178. Both sequences form a clade with a

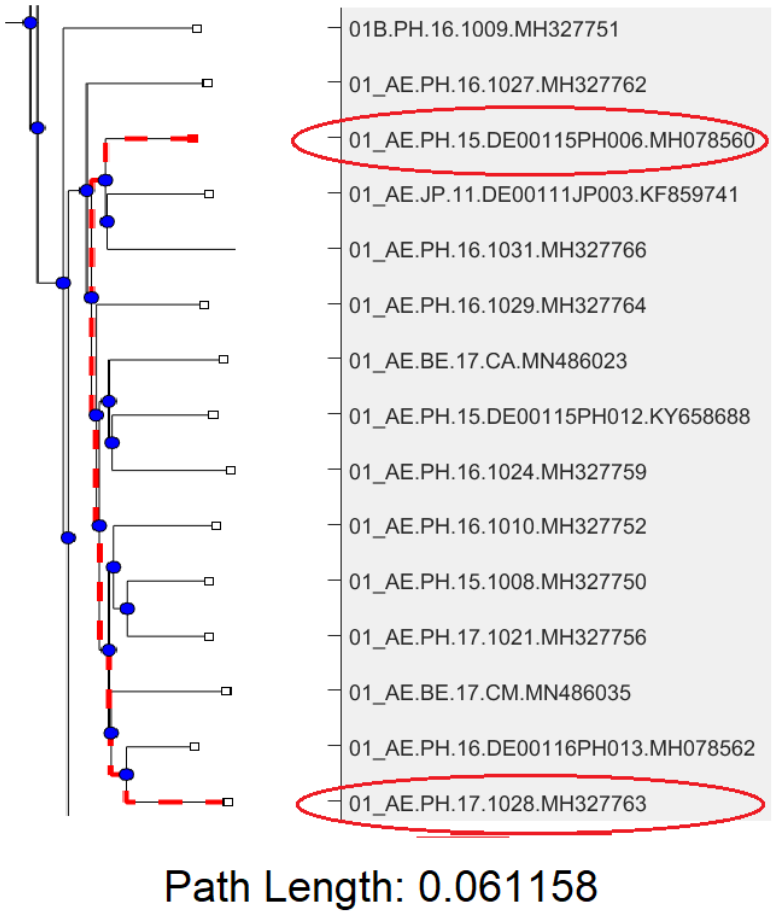


Figure 5.12: Partial Phylogenetic Tree Including Subtype CRF01_AE MH078560 and MH327763 Sequences. CRF01_AE is a recombinant subtype between pure subtypes A and E. In both sequences, their high degree of similarity between the hot zones was aligned with the narrow phylogenetic distance (0.061158 only). Both sequences belong to a clade composed of only 14 sequences.

single common ancestor. Although this fact was in line with the concordance level in the hot zones of both sequences, more resemblance in their hot zones was expected.

In light of these representative examples in the second study group (concatenated spectrograms of recombinant sequences), I observed that there was a directly proportional relationship between phylogenetic closeness and the degree of agreement of hot zones, though it was occasionally more diffuse in the case of recombinants.

Similarly, in the case of superposed spectrograms, the degree of similarity among hot zones was also higher among sequences of the same subtype. Table 5.17 serves as further confirmation and displays various representative examples for non-recombinant sequences with superposed spectrograms.

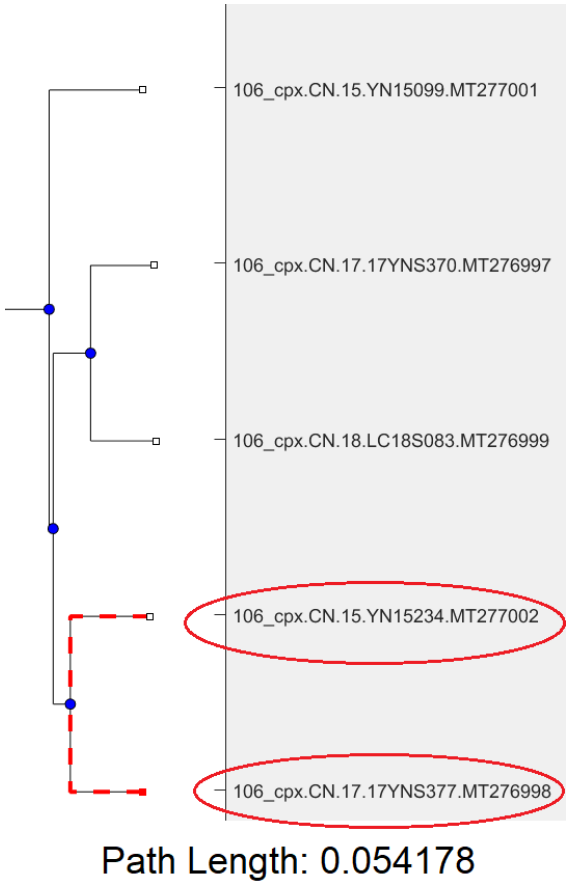


Figure 5.13: Partial Phylogenetic Tree Including Subtype CRF106_cpx MT276998 and MT277002 Sequences. The recombinant subtype CRF106_cpx MT276998 and MT277002, are phylogenetically "sisters" with a shared common ancestor and a very small phylogenetic distance (0.0554178).

The first example shows two sequences belonging to subtype A1 (MN791186 and MN791188) with high levels of resemblance between their hot zones. Grad-CAM clearly targeted high frequencies in both sequences within the *gag* gene and part of *pol*. Additionally, it also specifically pointed to both high and low frequencies of the 3'LTR. In the case of LIME, the most relevant hot zone was perfectly located in the high frequencies of the 5'LTR. A secondary hot zone was found in the low frequencies of the 5'LTR as well. The results of Gradient Attribution were more diffuse, although remarkably similar in both sequences. There appeared to be a higher concentration of hot zone points in the low-frequency range of both the 5'LTR and 3'LTR.

These sequences are indeed intra-patient (ID Patient = 116937), hence their phylogenetic distance is minimal. This case was highly illustrative of the pro-

portional relationship detected between the degree of hot zone similarity and phylogenetic proximity.

Table 5.17: Comparison of Explainability of Non-Recombinant Superposed Spectrograms. Each example contains two non-recombinant sequences with their ID and subtype, along with their respective hot zones identified by Grad-CAM, LIME, and Gradient Attribution, referenced to a scaled image of the superposed spectrogram composition.

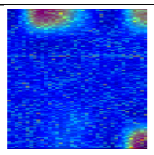
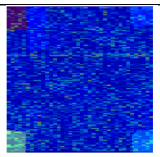
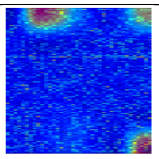
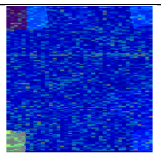
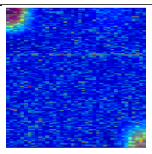
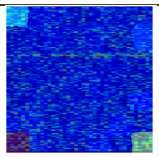
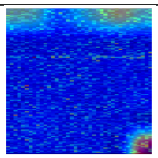
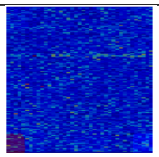
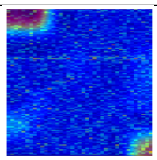
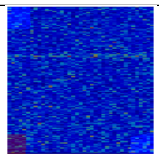
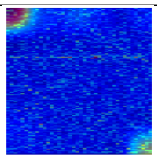
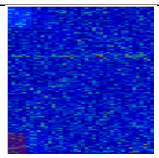
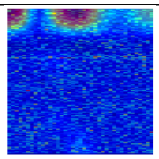
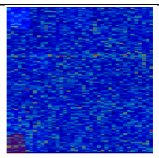
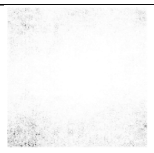
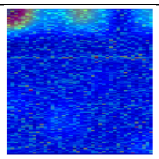
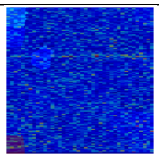
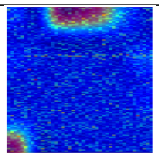
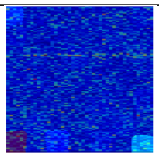
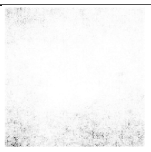
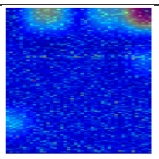
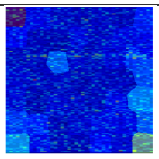
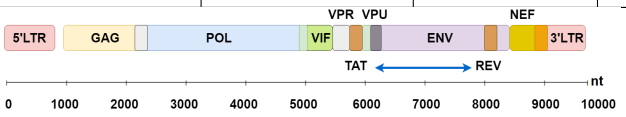
Subtype	Sequence ID	Grad-CAM	LIME	Gradient Attribution
A1	MN791186			
A1	MN791188			
	Reference			
A6	MH330346			
A6	MT369977			
	Reference			
B	FJ469740			
B	FJ496164			
	Reference			
C	DQ207941			

Table 5.17 – Continuation from the previous page

Subtype	Sequence ID	Grad-CAM	LIME	Gradient Attribution
C	DQ351235			
	Reference			
O	AB485666			
O	MH705149			
	Reference			
				

The next example corresponds to the subtype A6 MH330346 and MT369977 sequences. In Grad-CAM, the main hot zone common to both sequences was located in the low frequencies of the 3'LTR. In the case of MH330346, there was an additional main hot zone in the high frequencies of the 5'LTR. MT369977 exhibited multiple secondary zones within the high-frequency range spanning almost the entire genome, except for a segment between the *gag* and *pol* genes. With LIME, the primary hot zone was precisely situated in the low frequencies of the 5'LTR. The results of Gradient Attribution, in this instance, were diffuse and slightly divergent between the two sequences, although there were faint indications of certain common areas in the high and low frequencies around the 5'LTR and *gag*.

Regarding the two sequences of non-recombinant subtype A6, MT369977 is not present in the phylogenetic tree and lacks a complete intra-patient reference sequence, it was not possible to determine the exact phylogenetic distance between both sequences. Nevertheless, both sequences belong to the same subtype indicating sufficient phylogenetic proximity for the purposes of this example.

The following case pertains to subtype B and FJ496164 sequences. Grad-CAM and LIME exhibited remarkable concurrence in their hot zones. In both sequences, Grad-CAM distinctly highlighted the high frequencies of the 5'LTR and the low frequencies of the 3'LTR. In the case of LIME, the main hot zone corresponded to the low frequencies of the 5'LTR. Gradient Attribution produced divergent results. However, in both sequences, the range of high and low frequencies in the 5'LTR and *gag* stood out.

The absence of a complete reference intra-patient sequence for FJ496164 limited my ability to assess phylogenetic proximity in the same manner as the previous example. In the absence of this reference sequence, the primary basis for inferring phylogenetic proximity is the shared subtype designation, which implies that they share a common ancestral origin.

It is important to note that while subtype classification is a valuable indicator of relatedness, it does not offer the same level of specificity and granularity as phylogenetic distance. However, the shared subtype is still a strong indicator of a common ancestry, which is why it is often used as a proxy for phylogenetic relatedness when complete intra-patient data is unavailable.

The case of subtype C DQ207941 and DQ351235 provides an additional illustrative instance featuring non-recombinant sequences in superposed spectrograms. The hot zones in Grad-CAM were situated in the high frequencies of 5'LTR and *pol*. A less prominent hot zone was observed in the high frequencies of 3'LTR. LIME yielded nearly identical results, with all the emphasis on the low frequencies of 5'LTR. The hot zones in Gradient Attribution clustered in the low frequencies around 5'LTR and *gag*, as well as around 3'LTR and *nef*. In this case, the degree of dispersion within the hot zones was lower than in the two previous examples.

The subtype C ranks second in the number of complete sequences, surpassed only by subtype B. The significant quantity of C subtype sequences is associated with an expected increase in phylogenetic distance compared to previous examples. This is especially applicable to HIV-1, where a greater presence of diverse sequences

implies higher genetic variability, often resulting in an increased phylogenetic distance between sequences.

In Figure 5.14, I observed that the phylogenetic distance was greater than 0.1, with multiple points of evolutionary divergence along the phylogenetic path between both sequences. Although, given the degree of similarity in the hot zones, a closer phylogenetic proximity was expected, the diversification within subtype C can account for this slight divergence.

This evolutionary variability is logical and consistent with the high expansion and genetic diversity of subtype C [2468].

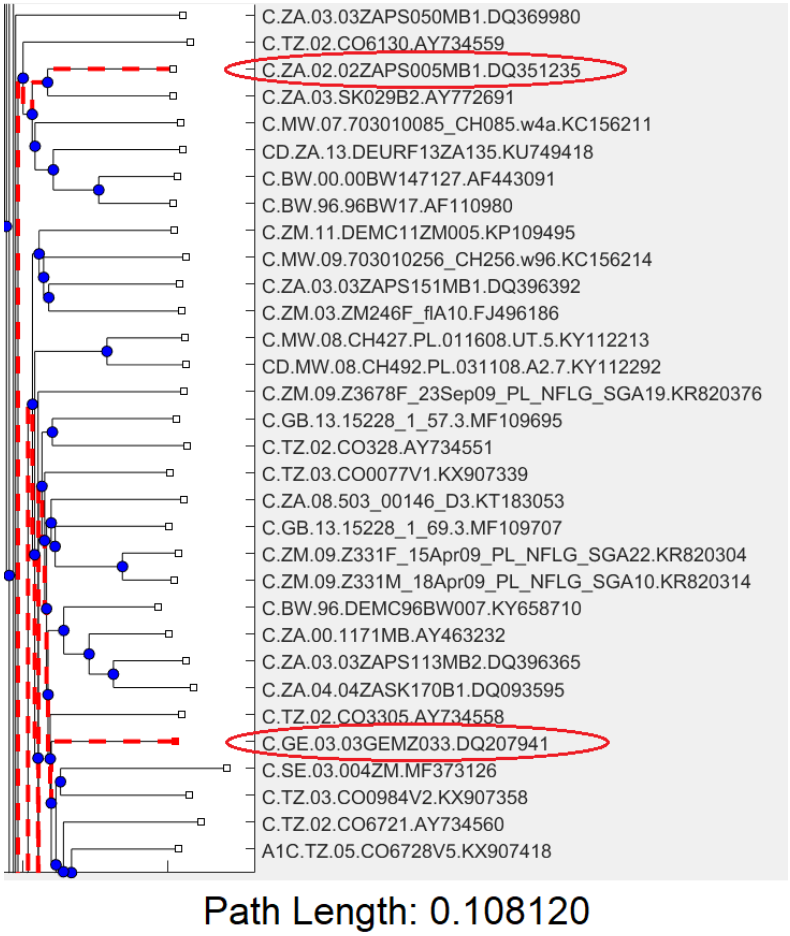


Figure 5.14: Partial Phylogenetic Tree Including Subtype C DQ207941 and DQ351235 Sequences. The non-recombinant subtype C comprises 2,105 complete sequences. The path length between DQ207941 and DQ351235 is 0.108120.

In the study case of subtype O AB485666 and MH705149 sequences, according to the Grad-CAM results, AB485666 exhibited a clear main hot zone in the high-

frequency range within the *pol* gene, as well as in the *vif*, *vpr*, *tat*, *vpu* genes, and a portion of *env*. Another main hot zone was clearly situated in the low frequencies of the 5'LTR. In the case of MH705149, the main hot zone was in the high frequencies of 3'LTR and *nef*. Secondary zones, particularly in the high frequencies of *gag* and *pol*, were also clearly visible. LIME results were more precise, focusing on the low and high frequencies of 5'LTR in AB485666 and MH705149, respectively. Other secondary zones were located in the low frequencies of 3'LTR. However, the results of Gradient Attribution were more similar in both sequences. They were prominently centered across the entire genome in the low-frequency range, with slightly greater intensity in the 5'LTR and *gag* regions, as well as in the 3'LTR and *nef* regions. Their relative degree of phylogenetic relatedness (Figure 5.15) aligns with the relative degree of similarity in their respective hot zones.

When considering non-recombinant sequences within the context of superposed spectrograms, the proportionality between phylogenetic proximity and the degree of similarity in hot zones was more volatile than in the case of concatenated ones, from these representative examples, I could observe that there was a direct relationship between both factors.

As for the fourth study group (superposed spectrograms of recombinant sequences), Table 5.18 shows three representative examples of the relationship between phylogenetic proximity and the degree of similarity in hot zones.

The first representative example consist of two CRF01_AE subtype sequences (AY358044 and AY358061) with a high degree of similarity in their hot zones.

The results of the three explainability tools were very similar in the case of these two sequences. In Grad-CAM, they were highly focused on low frequencies and around $f = 1/3$. In the case of AY358061, there was an additional hot zone in the high frequencies of 3'LTR. In LIME, the main hot zone was clearly defined in the low frequencies of *gag*. In Gradient Attribution, the hot zones were concentrated around *gag*, from *pol* to *vif* and 3'LTR, all within the low-frequency range. There was also a noticeable concentration around $f = 1/3$ in 5'LTR.

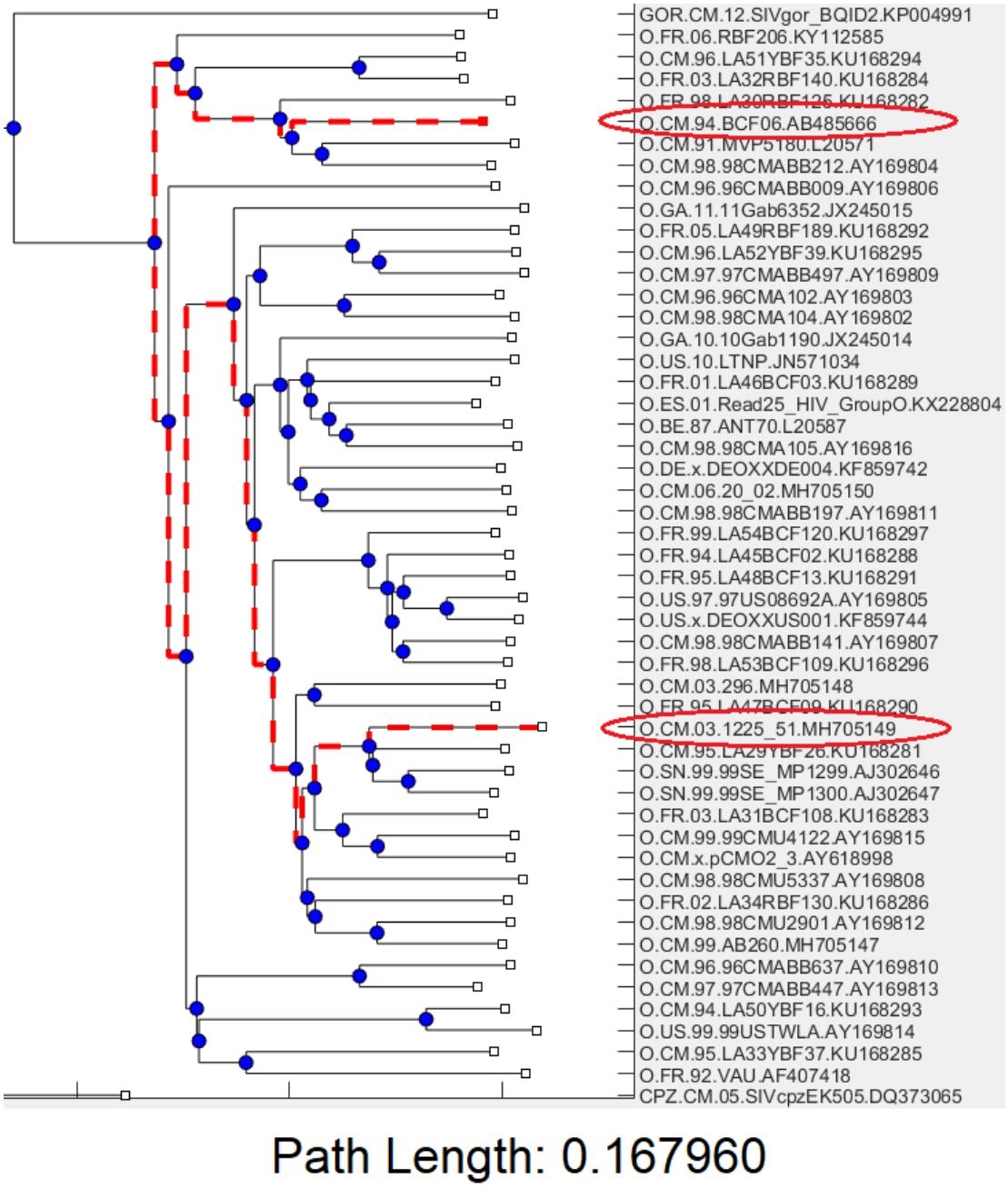
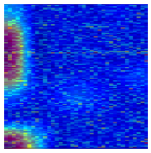
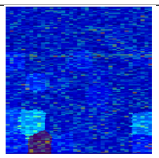
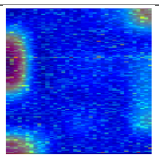
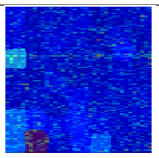
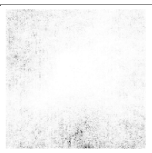
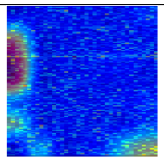
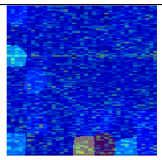
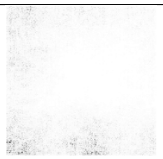
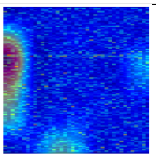
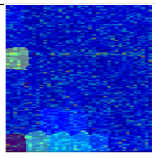
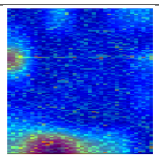
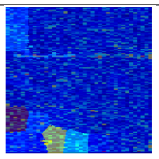
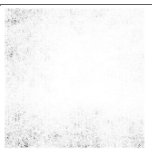
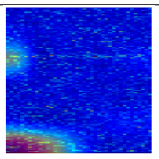
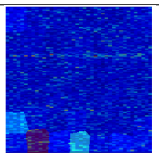
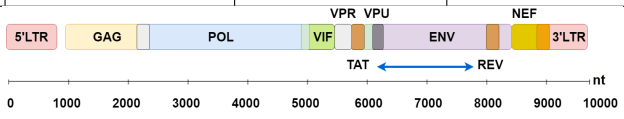


Figure 5.15: Partial Phylogenetic Tree Including Subtype O AB485666 and MH705149 Sequences. Up until 2020, there are 57 sequences classified as belonging to subtype O. The path length between AB485666 and MH705149 is 0.167960.

Table 5.18: Comparison of Explainability of Non-Recombinant Superposed Spectrograms. Each example comprises two recombinant sequences with their ID and subtype, along with their respective hot zones identified by Grad-CAM, LIME, and Gradient Attribution, mapped to a scaled image of the superposed spectrogram composition.

Subtype	Sequence ID	Grad-CAM	LIME	Gradient Attribution
CRF01_AE	AY358044			
CRF01_AE	AY358061			
	Reference			
CRF22_01A1	AY371165			
CRF22_01A1	KP109499			
Reference				
CRF20_BG	KT276270			
CRF20_BG	MW062985			
Reference				
				

Despite both sequences appear to be quite distant in the phylogenetic tree (Figure 5.16), their actual distance is very limited. In this segment of the tree, we can observe clades containing a significant number of sequences, which can affect the visual representation of the phylogenetic tree, making them seem more distant than their true phylogenetic relationship indicates.

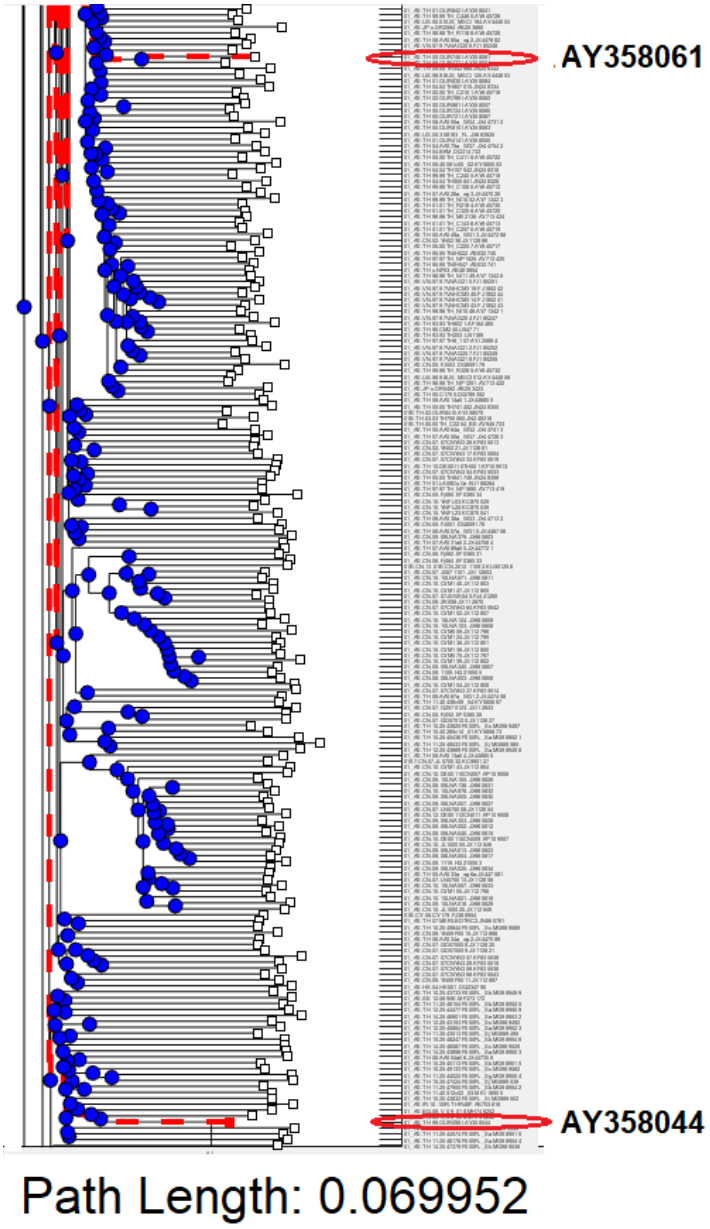


Figure 5.16: Partial Phylogenetic Tree Including Subtype CRF01_AE AY358044 and AY358061 Sequences. The recombinant subtype CRF01_AE comprises 1,490 complete sequences, being the most abundant recombinant subtype. The path length between AY358044 and AY358061 is 0.069952.

The pair of subtype CRF22_01A1 sequences also exhibited a similar pattern of hot zones. The main hot zones in Grad-CAM highly coincided in the vicinity of $f = 1/3$. In LIME, there was a significant agreement in a secondary hot zone located at $f = 1/3$ in 5'LTR. The main hot zone for AY371165 was in the low frequencies of*env*. For KP109499, it was in the low frequencies of 5'LTR. In Gradient Attribution, concentrations of hot points were observed in the low-frequency range in the areas of 5'LTR, *gag*, *vif*, *vpr*, *tat*, *vpu*, certain regions of *env*, and 3'LTR.

In this case, it is not possible to display the excerpt of the phylogenetic tree since KP109499 lacks a complete intra-patient reference sequence. Nevertheless, the fact that both sequences belong to the same subtype indicates a close phylogenetic relationship.

In the case of the two sequences subtype CRF20_BG KT276270 and MW062985, Grad-CAM focused on the low-frequency range, in the case of KT276270, it pointed to the vicinity of *gag* and *pol*. In the case of MW062985, the hot zone was similar but perhaps more centered around the 5'LTR and *gag*. LIME also centered on the low frequencies. There was a common secondary zone in the second half of *pol*. The main zone in the case of KT276270 was around $f = 1/9$ in 5'LTR. In MW062985, the primary hot zone was in the low frequencies of *gag*. The results of Gradient Attribution ware variable, although more intensity was detected in 5'LTR at $f = 1/3$ and the low frequencies of 3'LTR.

While sequences KT276270 and MW062981, which is the complete intra-patient reference sequence of MW062985 (Patient ID: 113475), are categorized under the same recombinant subtype, they display a substantial phylogenetic divergence, indicated by a separation length of 0.28 (Figure 5.17).

In this particular case, the observed genetic distance raised intriguing questions that warrant further investigation, to understand the factors responsible for this discrepancy.

Once the four study groups were analyzed, in most cases, the closer the phylogenetic distance, the greater the similarity of the hot zones, and ever more for intra-patient sequences [2469].

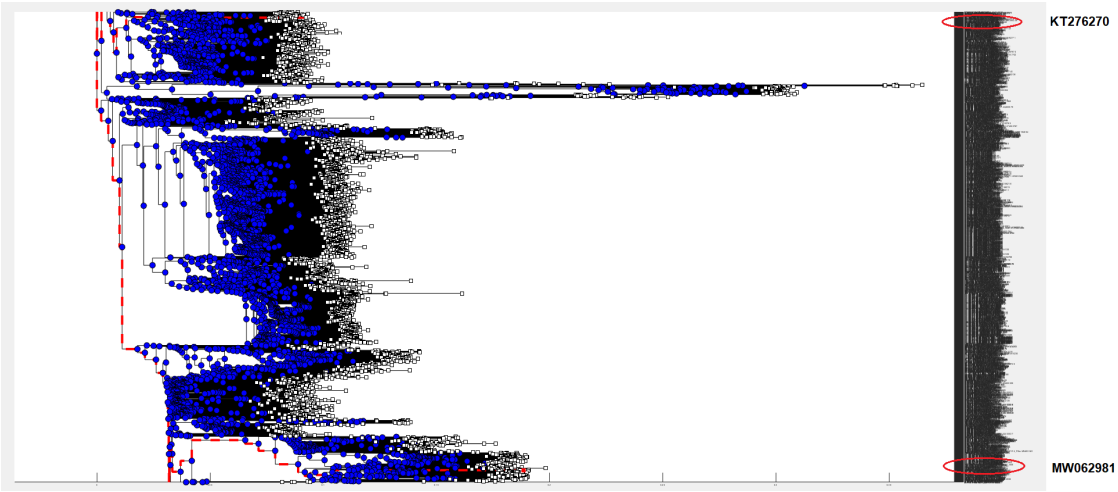


Figure 5.17: Partial Phylogenetic Tree Including Subtype CRF20_BG KT276270 and MW062981 Sequences. Up until 2020, there are 23 sequences classified as belonging to subtype CRF20_BG. The path length between them is around 0.28.

Furthermore, this correlation was observed in both types of spectrograms. If the hot zones of two sequences exhibited a high similarity level in concatenated spectrograms, the same degree of concordance was found in superposed, and vice versa. There was a consistency in the behavior of the CNN in both dataset types. This fact, combined with the correlation between the neural network's performance in both datasets and the degree of relatedness of the sequences, indicated a coherent and robust behavior. I provide some representative cases of this observation by way of endorsement.

For instance, Table 5.19 demonstrated a significant resemblance between the hot zones in concatenated and superposed spectrograms for sequences KR820306 and KR820312, which are both of subtype C and from the same patient (Patient ID: ZM331F).

As expected, their decision areas were absolutely similar. Both sequences shared a clear determinant hot zone in Grad-CAM, located at $f = 1/3$ and the low frequencies of T. KR820306 had another hot zone at $f = 1/3$ in C. LIME clearly targeted the high frequencies of A. Gradient Attribution pointed to high frequencies in both sequences and around $f = 1/6$ of A, and to a lesser extent, the low frequencies and $f = 1/3$ of T. In superposed spectrograms, Grad-CAM fixed in high frequencies of 5'LTR in both cases. KR820312 had a secondary area at

low frequencies of 3'LTR. A very peripheral area could be perceived at *pol* gene. LIME highlighted low frequencies of 5'LTR, followed by high frequencies of this area. Gradient Attribution seemed to emphasize high and low frequencies of 5'LTR and *gag*, and low frequencies of some area near *vif*, *env* and, specially, 3'LTR.

The second case corresponds to the sequences MW061429 and MW061455, which are intra-patient sequences of subtype B (Table 5.20). In the case of concatenated spectrograms, there was a primary hot zone at $f = 1/3$ in the region corresponding to nucleotides T and G. Another significant area was found at low frequencies in C. LIME results were less convergent but focus on the vicinity of high frequencies of A and C. Gradient Attribution results were highly similar, with a greater concentration of hot points at high and low frequencies of A and at $f = 1/3$ of T. Regarding superposed spectrograms, the results were nearly identical for all three explainability tools. In the case of Grad-CAM, they were centered on the low frequencies of 3'LTR. LIME pointed to the low frequencies of 5'LTR, while Gradient Attribution mainly focused on the low-frequency range of 5'LTR and 3'LTR.

Moreover, the same behavior can be registered for recombinant sequences (Table 5.20) with the two intra-patient subtype B sequences MW061429 and MW061455 (Patient ID: 9252).

As an illustration of the identical phenomenon within the context of recombinant sequences (Table 5.21), KC522031 and KC522033 are subtype CRF58_01B sequences. In concatenated spectrograms, Grad-CAM exhibited a significantly high level of similarity, with the primary hot zones centered at $f = 1/6$. In the case of LIME, although the results showed some variability, they consistently emphasized nucleotide A. For KC522031, LIME emphasized high frequencies, while for KC522033, it focused on low frequencies. Gradient Attribution results, on the other hand, were remarkably consistent, highlighting high frequencies in nucleotide A as the primary hot zones. In superposed spectrograms, the similarity between Grad-CAM's hot zones in both sequences was striking. The patterns of the hot zones were distinctive for this subtype. There was a very clear area at $f = 1/3$ in the 5'LTR. The other main hot zone was a quarter-circle that started at $f = 1/9$

Table 5.19: Comparison of Explainability of HIV-1 Phylogenetically Close Non-Recombinant Sequences in Concatenated and Superposed Spectrograms Case 1. The sequences are identified by their ID and subtype. The hot zones identified by Grad-CAM, LIME, and Gradient Attribution are referenced with a scaled representation of the concatenated and superposed spectrograms.

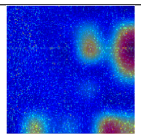
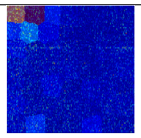
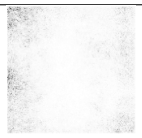
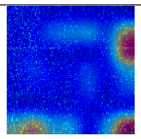
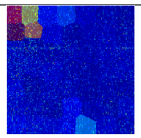
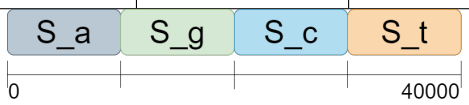
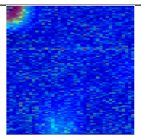
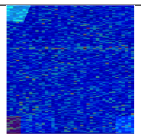
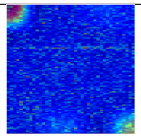
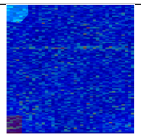
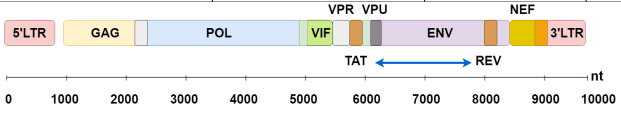
Similar hot zones in phylogenetically close sequences				
Non-Recombinant				
Concatenated Spectrograms				
Subtype	**Sequence ID**	**Grad-CAM**	**LIME**	**Gradient Attribution**
C	KR820306			
C	KR820312			
	Reference			
				
Superposed Spectrograms				
Subtype	**Sequence ID**	**Grad-CAM**	**LIME**	**Gradient Attribution**
C	KR820306			
C	KR820312			
	Reference			
				

Table 5.20: Comparison of Explainability of HIV-1 Phylogenetically Close Non-Recombinant Sequences in Concatenated and Superposed Spectrograms Case 2. The sequences are identified by their subtype and ID. The hot zones are referenced with scaled schematics of the concatenated and superposed spectrogram composition.

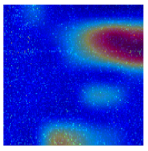
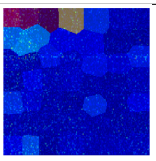
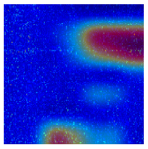
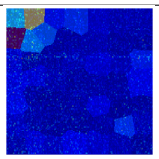
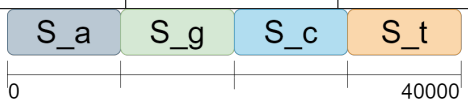
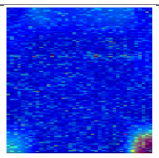
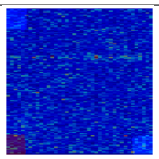
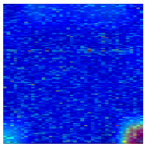
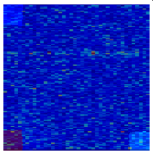
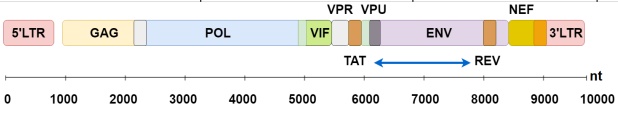
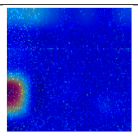
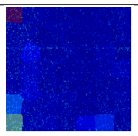
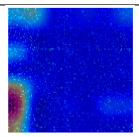
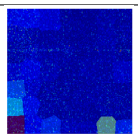
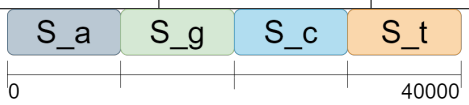
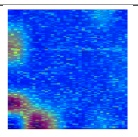
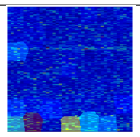
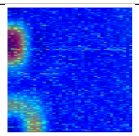
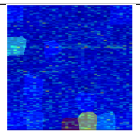
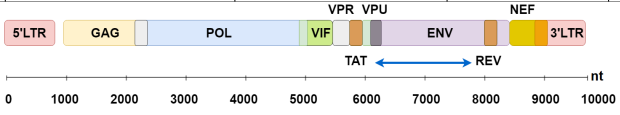
Similar hot zones in phylogenetically close sequences				
Non-Recombinant				
Concatenated Spectrograms				
Subtype	**Sequence ID**	**Grad-CAM**	**LIME**	**Gradient Attribution**
B	MW061429			
B	MW061455			
	Reference			
				
Superposed Spectrograms				
Subtype	**Sequence ID**	**Grad-CAM**	**LIME**	**Gradient Attribution**
B	MW061429			
B	MW061455			
	Reference			
				

Table 5.21: Comparison of Explainability of HIV-1 Phylogenetically Close Recombinant Sequences in Concatenated and Superposed Spectrograms Case 1. The recombinant sequences are identified by their ID and subtype. The hot zones from Grad-CAM, LIME, and Gradient Attribution are referenced with scaled schematics of the concatenated and superposed spectrogram composition.

Similar hot zones in phylogenetically close sequences				
Recombinant				
Concatenated Spectrograms				
Subtype	**Sequence ID**	**Grad-CAM**	**LIME**	**Gradient Attribution**
CRF58_01B	KC522031			
CRF58_01B	KC522033			
	Reference			
				
Superposed Spectrograms				
Subtype	**Sequence ID**	**Grad-CAM**	**LIME**	**Gradient Attribution**
CRF58_01B	KC522031			
CRF58_01B	KC522033			
	Reference			
				

in the 5'LTR and ended in the low frequencies of *gag*. The results from LIME were variable. In the case of KC522031, the main hot zone was located in the low frequencies of *gag*. In the case of KC522033, however, it was in the vicinity of *pol* and *vif*. Gradient Attribution pointed to $f = 1/3$ in the 5'LTR and *gag* area, as well as the low-frequency spectrum, especially in the 5'LTR and *gag* region, the area around *vif*, *vpr*, *tat*, *vpu*, *nef*, and 3'LTR.

The subtype CRF58_01B sequences KC522031 and KC522033 exhibit an extremely minimal phylogenetic distance (Figure 5.18). Likewise, the superposed spectrograms' hot zones demonstrated a remarkable similarity level.

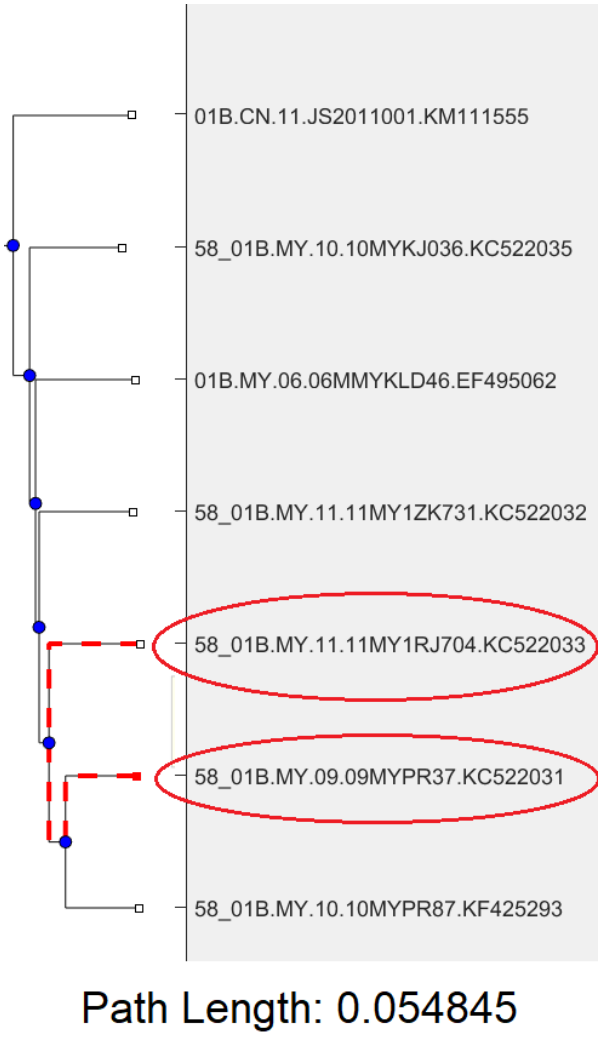
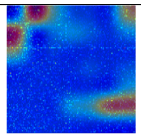
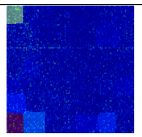
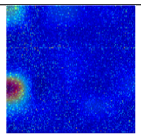
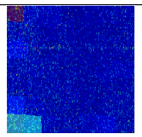
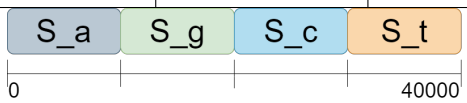
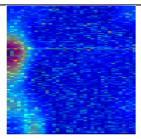
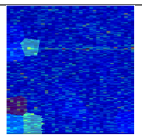
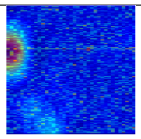
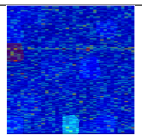
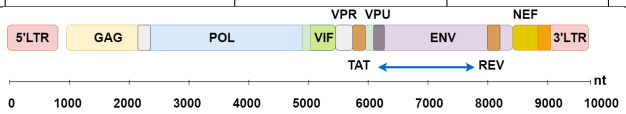


Figure 5.18: Partial Phylogenetic Tree Including Subtype CRF58_01B KC522031 and KC522033 Sequences. Subtype CRF58_01B includes six complete genomes cataloged as of 2020. The high degree of phylogenetic relatedness is evident, with a path length of only 0.054845.

Table 5.22: Comparison of Explainability of HIV-1 Phylogenetically Close Recombinant Sequences in Concatenated and Superposed Spectrograms Case 2. Sequences are labeled with their ID and subtype. The regions of interest identified by Grad-CAM, LIME, and Gradient Attribution are depicted on scaled diagrams of the concatenated and superposed spectrogram composition.

Concatenated Spectrograms				
Subtype	**Sequence ID**	**Grad-CAM**	**LIME**	**Gradient Attribution**
CRF02_AG	AY371129			
CRF02_AG	AY371130			
	Reference			
				
Superposed Spectrograms				
Subtype	**Sequence ID**	**Grad-CAM**	**LIME**	**Gradient Attribution**
CRF02_AG	AY371129			
CRF02_AG	AY371130			
	Reference			
				

In some cases of recombinant spectrograms, hot areas were slightly more volatile, particularly in the concatenated graphical representation. In the case of AY371129 and AY371130, both from subtype CRF02_AG, their concatenated spectrograms (Table 5.22) showed divergences between the explainability hot zones despite being phylogenetically close sequences, with a path length of 0.085774 (Figure 5.19). However, in the case of superposed spectrograms, their hot zones were similar.

Given the phylogenetic closeness between both sequences, a higher level of similarity between hot zones in concatenated spectrograms would be expected. This fact requires further research.

Grad-CAM hot zones were clearly different in concatenated spectrograms. Hot zones in AY371129 were around high frequencies of A, $f = 1/9$ and high frequencies in T. However, AY371130 most relevant zones were one-sixth and high frequencies of A. LIME had no deviations. The most determinant areas were high and low frequencies in A. In Gradient Attribution, the most important region seemed to be the high frequency range of A. In superposed, Grad-CAM relevant zone was located around one-third frequency of 5'LTR. Secondary areas were surrounding low frequencies of 5'LTR. LIME results were more dispersed. In AY371129, a zone close to one-ninth frequency at 5'LTR outlined. Less relevant sites were identified at low frequencies and one-third frequency at *gag*. In AY371130, the most determinant zone was one-third of 5'LTR. There was a secondary area in the low frequency range of the central part of the genome. Gradient Attribution point clouds were visually different, more noisy in the case of AY371129. In AY371129, the most relevant areas were low frequencies in 5'LTR and the first half of *gag*, *vif* and surroundings and 3'LTR. Another remarkable zone was between one-third and high frequencies of 5'LTR and *gag*. In AY371130, Gradient Attribution showed low activity, highlighting low frequencies of *gag* and maybe 5'LTR, around *vif* and, very weak, 3'LTR.

Despite the differences observed in the highlighted areas by Grad-CAM for concatenated spectrograms (not the case with LIME and Gradient Attribution), in the case of superposed spectrograms, a significantly high level of similarity was particularly remarkable with Grad-CAM.

Tables 5.23 and 5.24 illustrate two examples involving recombinant sequences belonging to subtype CRF01_AE.

As seen in Table 5.23, intra-patient recombinant sequences MW443219 and MW443225 (Patient Code 40503) developed the same high concordance in concatenated and superposed spectrograms. Both sequences belong to the same patient,

Table 5.23: Comparison of Explainability of HIV-1 Phylogenetically Close Recombinant Sequences in Concatenated and Superposed Spectrograms Case 3. Sequences are identified by their ID and subtype. The hot areas pointed by Grad-CAM, LIME, and Gradient Attribution are shown on scaled representations of the concatenated and superposed spectrogram composition.

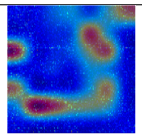
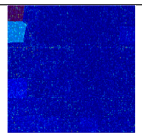
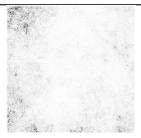
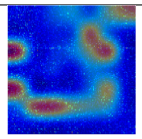
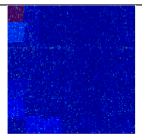
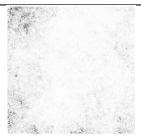
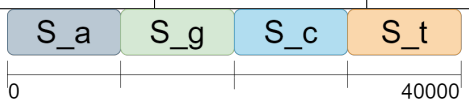
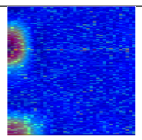
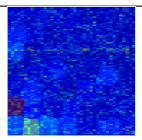
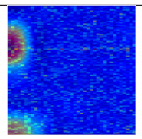
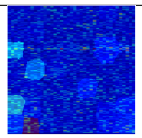
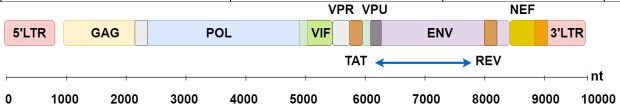
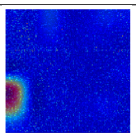
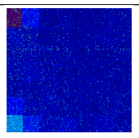
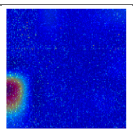
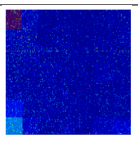
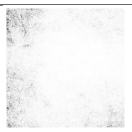
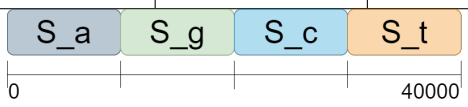
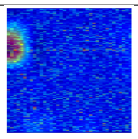
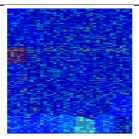
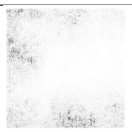
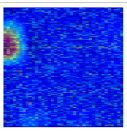
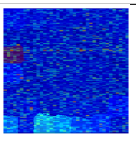
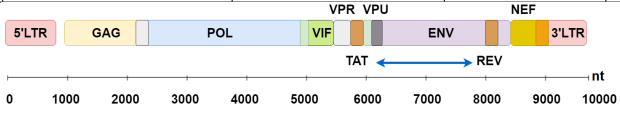
Similar hot zones in phylogenetically close sequences				
Recombinant				
Concatenated Spectrograms				
Subtype	**Sequence ID**	**Grad-CAM**	**LIME**	**Gradient Attribution**
CRF01_AE	MW443219			
CRF01_AE	MW443225			
	Reference			
				
Superposed Spectrograms				
Subtype	**Sequence ID**	**Grad-CAM**	**LIME**	**Gradient Attribution**
CRF01_AE	MW443219			
CRF01_AE	MW443225			
	Reference			
				

Table 5.24: Comparison of Explainability of HIV-1 Phylogenetically Close Recombinant Sequences in Concatenated and Superposed Spectrograms Case 4. Sequences are labeled with their ID and subtype. The hot zones highlighted by Grad-CAM, LIME, and Gradient Attribution are displayed on scaled representations of concatenated and superposed spectrograms.

Similar hot zones in phylogenetically close sequences				
Recombinant				
Concatenated Spectrograms				
Subtype	**Sequence ID**	**Grad-CAM**	**LIME**	**Gradient Attribution**
CRF01_AE	MK867566			
CRF01_AE	MK867567			
	Reference			
				
Superposed Spectrograms				
Subtype	**Sequence ID**	**Grad-CAM**	**LIME**	**Gradient Attribution**
CRF01_AE	MK867566			
CRF01_AE	MK867567			
	Reference			
				

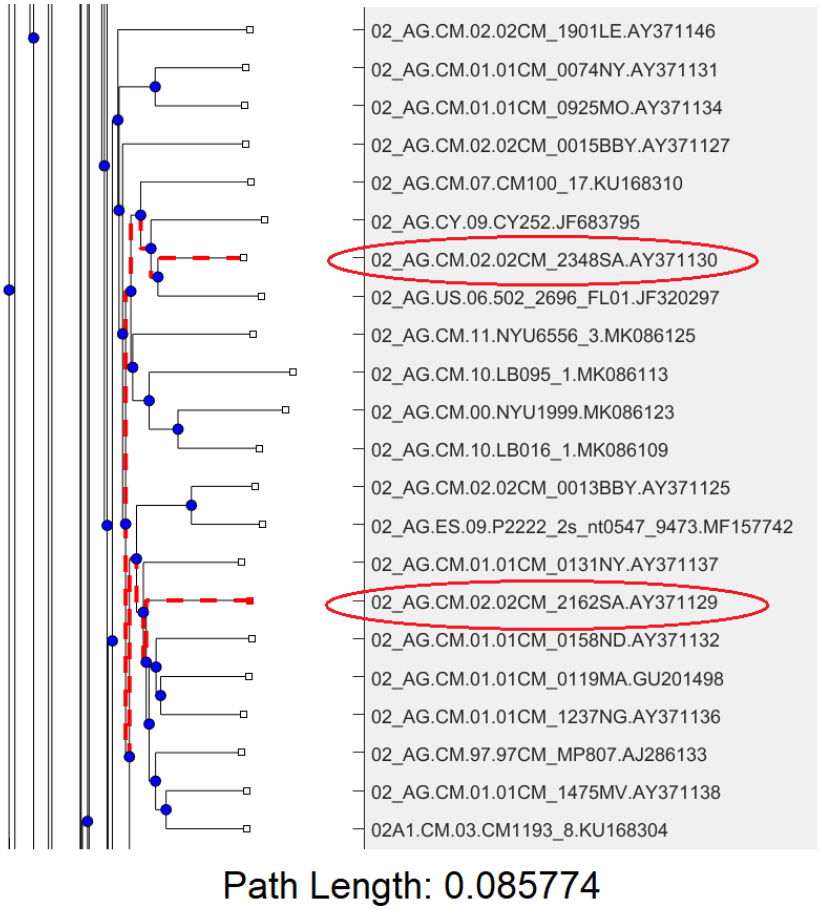


Figure 5.19: Partial Phylogenetic Tree Including Subtype CRF02_AG AY371129 and AY371130 Sequences. Subtype CRF02_AG has 179 complete genomes cataloged as of 2020. Their strong phylogenetic similarity is clearly demonstrated by a path length of 0.085774. These sequences belong to a clade that contains 18 sequences.

which implies a minimal phylogenetic distance, and entirely correlates with the high level of similarity in the hot areas.

MW443219 and MW443225 are intra-patient recombinant sequences of subtype CRF01_AE. The most remarkable feature of Grad-CAM was the high similarity level of the hot zones in concatenated spectrograms. The main key areas were one-third, one-sixth and one-ninth frequencies in A, one-ninth frequency in G, some points near one-third, one-sixth and one-ninth in C, and one-sixth, one third and high frequencies of T. However, LIME fixed in high frequencies of A. Noisy results of Gradient Attribution seemed to enhance low and high frequencies of A, some areas of low frequency range in G and C and low and one-third frequencies of T. In

superposed spectrograms, Grad-CAM pointed to one-third and low frequencies of 5'LTR. LIME focused at frequencies near one-ninth of 5'LTR and, low frequencies of *gag*. One minor zone was one-third in 5'LTR. Hot points in Gradient Attribution were grouped in one-third and high frequencies of 5'LTR, low frequencies in the central area of the genome and low frequencies around 3'LTR.

In the case of CRF01_AE MK867566 and MK867567 (Table 5.24), all explainability results were practically identical. In concatenated spectrograms, Grad-CAM clearly focused on frequencies around $f = 1/6$ in A. LIME also targeted the A region but within the high-frequency range. Gradient primarily centered on the high and low frequencies of A and the low frequencies of T.

These are intra-patient sequences (Patient ID: 6323), which implies a minimal phylogenetic distance. It was worth highlighting the extremely high resemblance level in their hot zones.

Thus, based on the findings of the explainability tools:

- The CNN exhibited consistent behavior for both datasets.

- This was particularly true for non-recombinant sequences, as hot zones tended to be more similar the phylogenetically closer the sequences were.

- In the case of recombinants there was greater volatility. Sometimes, a greater phylogenetic closeness did not correlate with a greater similarity in explainability results.

- Compared to concatenated spectrograms, superposed spectrograms reflected the relationship between phylogenetic closeness and the level of similarity in explainability tool results more accurately.

5.15.2 Similarity between Sequences of Different Subtypes

I identified an additional phenomenon that necessitates further research. It is noteworthy that distinct sequences from different recombinant subtypes exhibited remarkably analogous hotspots in both concatenated and superposed spectrograms.

To determine the reasons for these similarities, it is necessary to seek underlying contributing factors, such as convergent evolution, the result of specific genetic motifs or structural elements, or perhaps a response to common selective pressures.

Table 5.25 shows some representative examples of this phenomenon.

Table 5.25: Similarity between Sequences of Different Subtypes. Each example of similarity between two sequences is determined by the subtype and ID of the sequences, as well as the hot zones referenced to a scaled representation of the compared genomic spectrogram type.

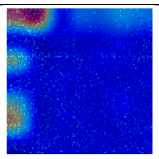
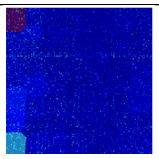
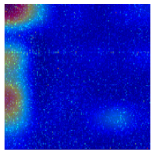
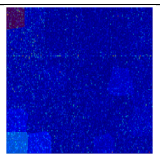
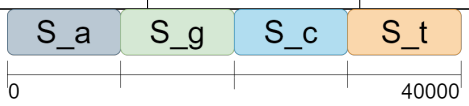
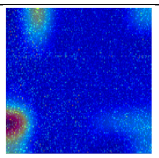
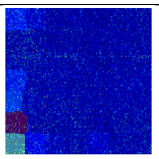
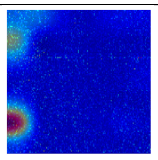
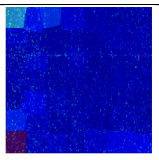
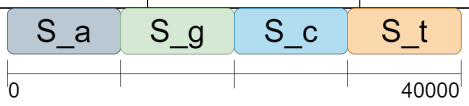
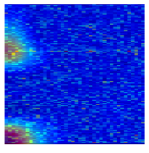
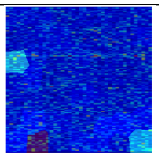
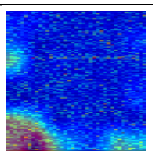
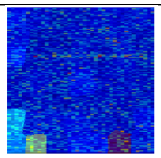
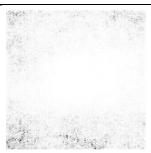
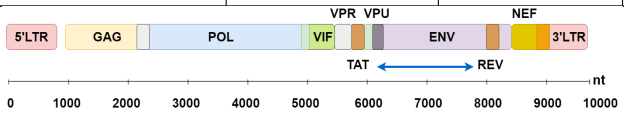
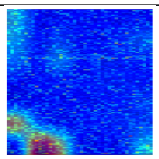
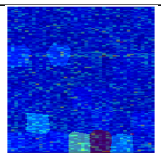
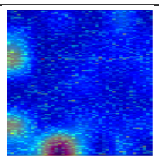
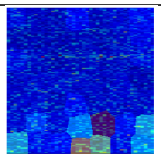
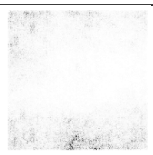
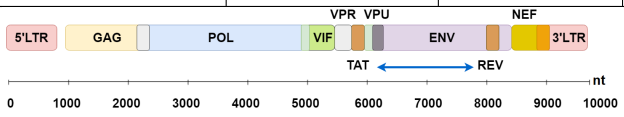
Subtype	Sequence ID	Grad-CAM	LIME	Gradient Attribution
CRF01_AE	JX960631			
CRF02_AG	KF561435			
	Reference			
				
CRF01_AE	AB253724			
CRF69_01B	AB845349			
	Reference			
				
CRF01_AE	JX448067			

Table 5.25 – Continuation from the previous page

Subtype	Sequence ID	Grad-CAM	LIME	Gradient Attribution
CRF55_01B	JX574661			
	Reference			
				
CRF01_AE	AB253424			
CRF69_01B	AB845349			
	Reference			
				

In the case of the recombinant sequences JX960631 (subtype CRF01_AE) and KF561435 (subtype CRF02_AG), there was a high similarity between the hot zones of concatenated sequences. In the case of Grad-CAM for both sequences, the primary hot zones were identified in the region corresponding to A, at high frequencies, and in the vicinity of $f = 1/6$. There was a secondary zone in A at the frequency $f = 1/3$. In LIME, the primary zone was clearly located in the high frequencies of A, as it was the case with Gradient Attribution.

Figure 5.20 is a visual representation of the high phylogenetic distance between both sequences, where this two sequences under study are located at opposite ends, with a path length of 0.19-0.20, which indicates that they are genetically distant sequences that have followed different evolutionary paths.

This is evident given the fact that they belong to different recombinant subtypes. Although both subtypes contain fragments identified as subtype A, from a virological

and phylogenetic standpoint, their lineages are distinct.

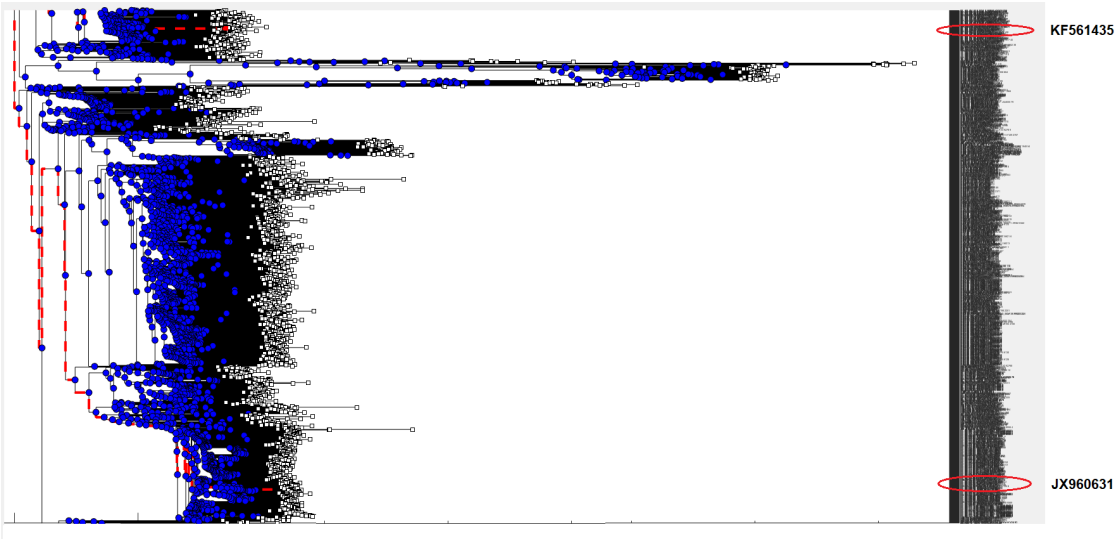


Figure 5.20: Partial Phylogenetic Tree Including Subtype CRF01_AE JX960631 and Subtype CRF02_AG KF561435 Sequences. The path length between both sequences is 0.19-0.20, indicating that they are phylogenetically distant sequences.

Figure 5.21 depicts the genome map of CRF01_AE. The uncolored zones could not be classified with any subtype.

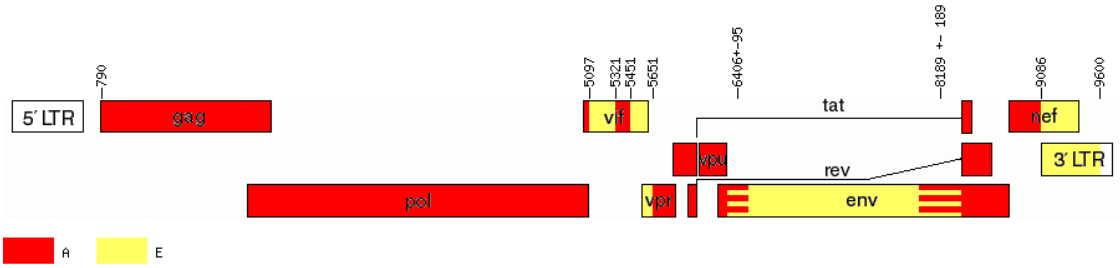


Figure 5.21: Genome Map of Recombinant Subtype CRF01_AE. This subtype is a recombinant between non-recombinant subtypes A and E. Image sourced from the Los Alamos National Laboratory HIV-1 Database HIV Circulating Recombinant Forms (CRFs) [41].

The genome map of subtype CRF02_AG is represented in Figure 5.22.

In broad terms, both subtypes coincide in that the *gag* gene belongs to subtype A, as well as a central region in *pol*, the final segment of *vpu*, and the start of *env*.

A starting point to elucidate the possible reasons that explain the similarity between the hot zones of JX960631 (subtype CRF01_AE) and KF561435 (subtype CRF02_AG) could be to delve deeper into the common genomic elements between

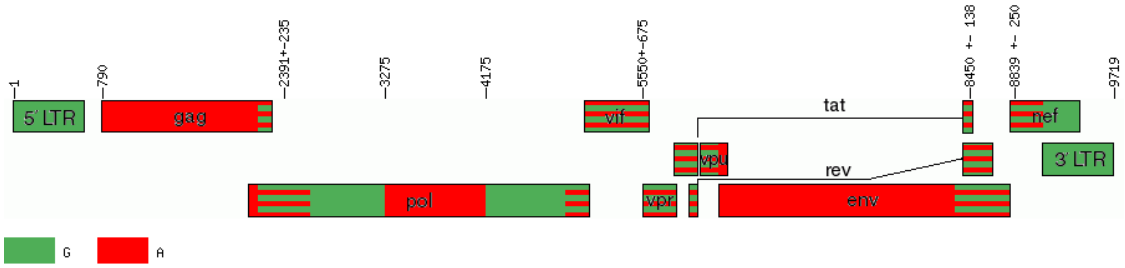


Figure 5.22: Genome Map of Recombinant Subtype CRF02_AG. This subtype is a recombinant between non-recombinant subtypes A and G. Image sourced from the Los Alamos National Laboratory HIV-1 Database HIV Circulating Recombinant Forms (CRFs) [41].

both subtypes and both sequences. Perhaps the coincidences in certain genomic regions and their relationship to phylogeny may help explain their coinciding hot zones despite belonging to different subtypes.

The next illustrative instance of this event corresponds to the hot zones corresponding to concatenated spectrograms of recombinant sequences AB253724 (subtype CRF01_AE) and AB845349 (subtype CRF69_01B). In both sequences, Grad-CAM pointed to $f = 1/9$ in A. In the case of LIME, for AB253724, it pointed exactly to the same area. However, for AB845349, it pointed to the low-frequency region of A. Gradient Attribution pointed to the high-frequency region of A.

As seen in Figure 5.23, these are two relatively distant sequences in the phylogenetic tree, with a path length of 0.20-0.21.

AB253423 is the intra-patient reference sequence of AB253724 (Patient ID: NH19-203).

The genome map of CRF01_AE can be found in Figure 5.21, while that of subtype CRF69_01B is presented in Figure 5.24. The uncolored zones could not be assigned to any subtype.

This large phylogenetic distance is expected in genomes of different subtypes, even though in this case the subtype CRF01_AE is one of the parental subtypes of CRF69_01B, meaning it provides genetic material during recombination to generate this new recombinant subtype.

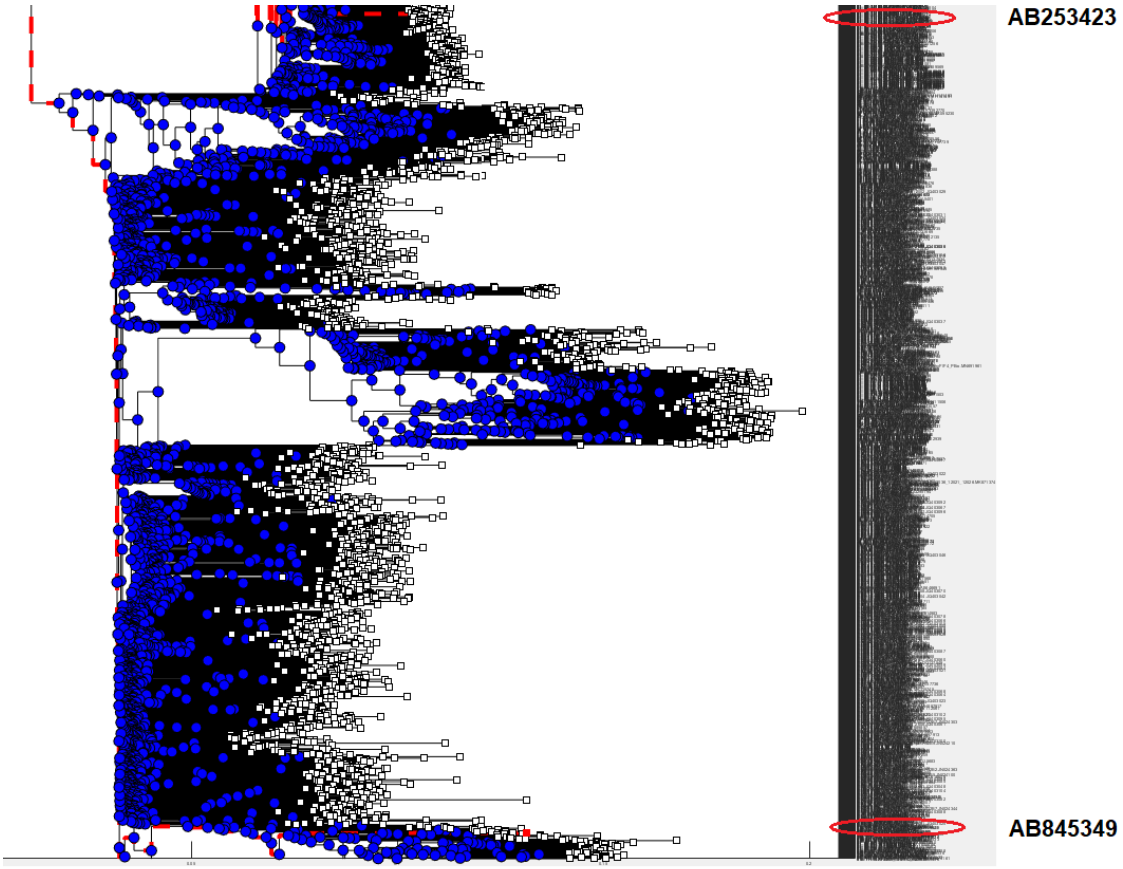


Figure 5.23: Partial Phylogenetic Tree Subtype CRF01_AE AB253423 and Subtype CRF69_01B AB845349 Sequences. The phylogenetic distance between these two sequences falls in the range of 0.20-0.21, underscoring their considerable genetic divergence.

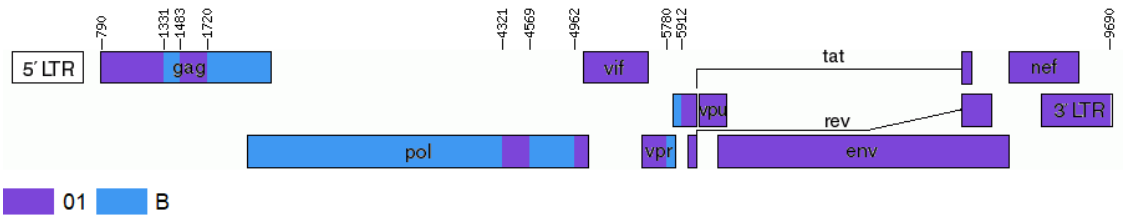


Figure 5.24: Genome Map of Recombinant Subtype CRF69_01B. This subtype is a recombinant between the recombinant subtype CRF01_AE and the non-recombinant subtype B. Image sourced from the Los Alamos National Laboratory HIV-1 Database HIV Circulating Recombinant Forms (CRFs) [41].

The parts of the CRF69_01B genome that correspond to subtype B are located in a small portion in *tat*, the last third of *vpr*, in *gag*, representing nearly half of it, and most of *pol*. The rest of the CRF69_01B genome corresponds to CRF01_AE.

Nevertheless, the high similarity of their hot zones prompted me to investigate whether this parent-child relationship plays any intrinsic role in the CNN's detection of the featuring recombinant.

This same phenomenon of similarity among hot areas in sequences from different subtypes was also consistently detected in superposed spectrograms. The first example presented corresponds to the recombinant sequences JX448067 (subtype CRF01_AE) and JX574661 (subtype CRF55_01B).

The main hot zones in Grad-CAM were located in the low-frequency region and at $f = 1/3$ in 5'LTR and *gag*. The hot zones in LIME were more divergent but concentrated in the low-frequency range around *gag*. In Gradient Attribution, the divergence was even greater, although it was centered in the low-frequency range in the 5'LTR, *gag*, and *env*. A certain significance of $f = 1/3$ in 5'LTR was visually apparent.

To delimit the visualization of phylogenetic distance, I considered sequence JX448059, which is the intra-patient reference sequence of JX448067 (Patient ID: AA111).

In view of Figure 5.26 and observing a path length of 0.23-0.24, there was no clear correlation between the degree of similarity of the hot zones and their relative position within the tree.

The subtype CRF55_01B (Figure 5.25), like in the previous case, descends from the same CRF01_AE subtype (Figure 5.21), which explains a certain distance in their relationship based on the ancestor-descendant connection. The uncolored zones could not be categorized into any subtype.

Almost the entire genome of CRF55_01B corresponds to CRF01_AE, except for two fragments located in the central region of *pol*, which correspond to the parental subtype B.

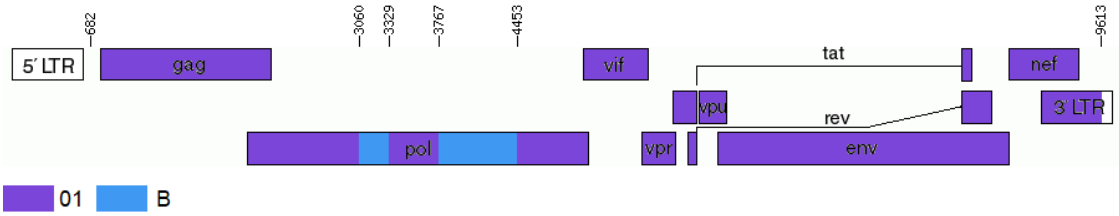


Figure 5.25: Genome Map of Recombinant Subtype CRF55_01B. This subtype results from a recombination event involving CRF01_AE, a recombinant subtype, and subtype B, which is non-recombinant. Image sourced from the Los Alamos National Laboratory HIV-1 Database HIV Circulating Recombinant Forms (CRFs) [41].

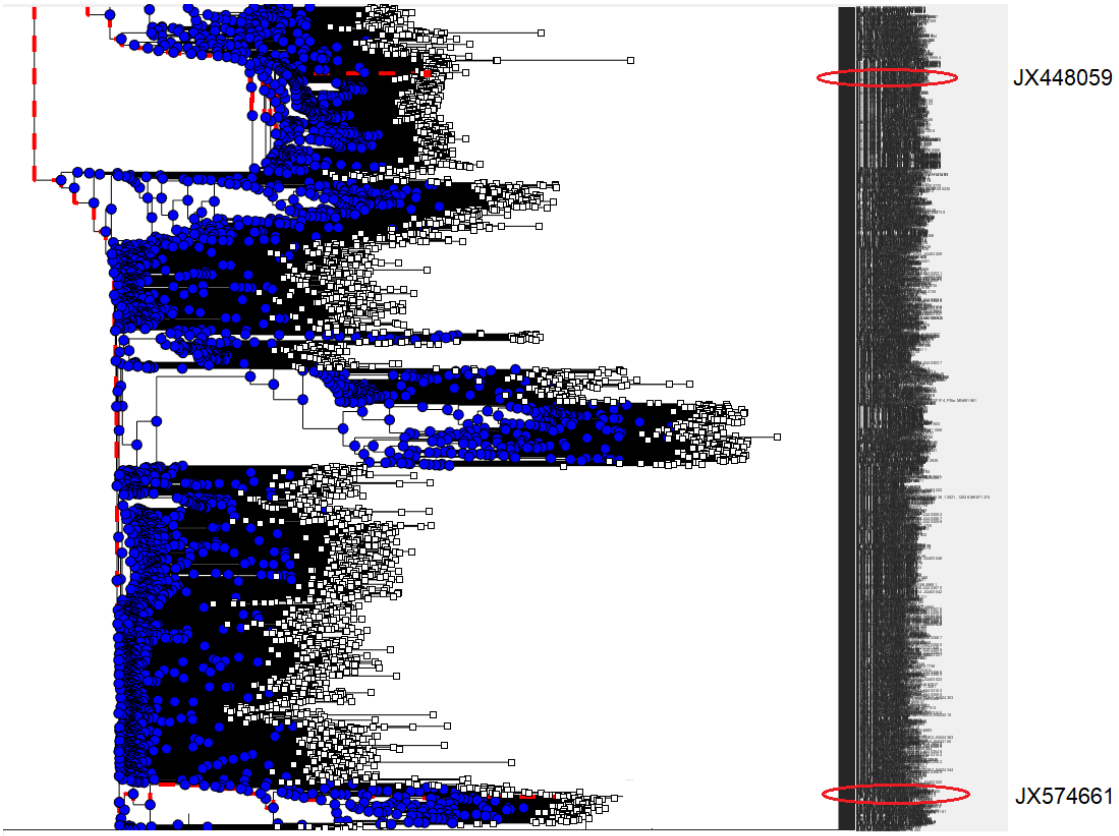


Figure 5.26: Partial Phylogenetic Tree Including Subtype CRF01_AE JX448059 and Subtype CRF55_01B JX574661 Sequences. The phylogenetic distance between these two sequences falls within the range of 0.23-0.24, highlighting a significant genetic divergence, which is visually depicted in the image.

It is necessary to delve into the exact analysis of the lineages involved in this case, as well as into microbiological and genomic aspects that may influence the similarity of the relevant regions for the detection of the recombinant feature.

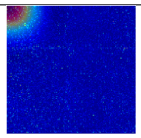
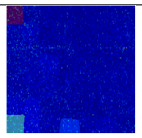
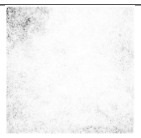
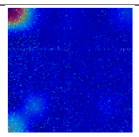
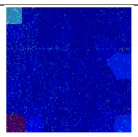
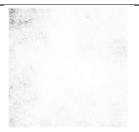
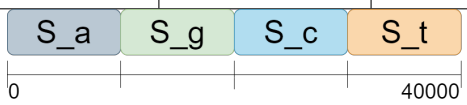
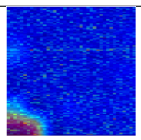
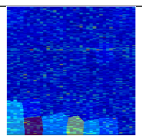
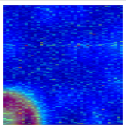
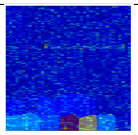
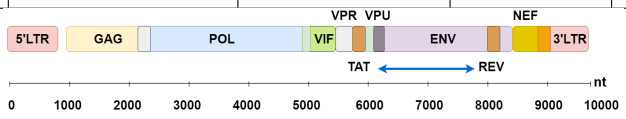
The next example in the case of superposed spectrograms corresponds to the recombinant sequences AB253424 (subtype CRF01_AE) and AB845349 (subtype CRF69_01B). This is the second study case involving sequences of these subtypes. In both sequences, the main areas of Grad-CAM pointed to the low-frequency region in *pol*, around $f = 1/9$ in the 5'LTR, and $f = 1/3$ in the 5'LTR. LIME pointed to the low-frequency range in the central genome region. Gradient Attribution was more volatile but centered on the low-frequency range. In AB253424, the highest concentration occurred in the 5'LTR and *gag* regions, while in AB845349, it occurred in the final region of *pol* and the following genes.

In this case, it was not possible to extract the phylogenetic tree as AB253424 does not have a complete intra-patient reference sequence. Patient ID: TH057. An analysis of the lineages involved in this case is necessary, as well as a thorough investigation of the subtype CRF69_01B and its relationship with its parental subtypes.

The persistence of this finding across multiple instances suggests the existence of hidden factors influencing this degree of similarity that cannot be discerned solely through phylogenetic analysis. It is imperative to conduct an in-depth analysis of the genomes of both subtypes involved.

In the example shown in Table 5.26, we can observe a case where this phenomenon occurs in both types of spectrogram. In concatenated spectrograms, Grad-CAM pointed to high frequencies in A. Likewise, LIME pointed to both high and low frequencies in A, and Gradient Attribution had the highest concentration of hot spots in high frequencies of A. In the case of superposed spectrograms, Grad-CAM clearly pointed to low frequencies in 5'LTR, *gag*, and the initial fragment of *pol*. LIME was more volatile but covered the entire range of low frequencies except for the *nef* gene and 3'LTR areas. Gradient Attribution also pointed to the entire range of low frequencies and, to a lesser extent, to high frequencies in the 5'LTR and 3'LTR regions

Table 5.26: Similarity between CRF53_01B DQ366663 and CRF52_01B DQ366664. Both sequences are identified by their subtype and ID. The hot zones are attached, referenced to scaled representations of concatenated and superposed spectrograms, respectively.

Concatenated Spectrograms				
Subtype	**Sequence ID**	**Grad-CAM**	**LIME**	**Gradient Attribution**
CRF53_01B	DQ366663			
CRF52_01B	DQ366664			
	Reference			
				
Superposed Spectrograms				
Subtype	**Sequence ID**	**Grad-CAM**	**LIME**	**Gradient Attribution**
CRF53_01B	DQ366663			
CRF52_01B	DQ366664			
	Reference			
				

Although they are different types of recombinants, both CRF53_01B and CRF52_01B are recombinants of subtypes CRF_01AE and B. Therefore, it is coherent that, even though they are from different subtypes, there are factors that influence their degree of similarity.

Then, I analyzed the composition of subtype CRF53_01B based on the illustration of its genome map (Figure 5.27).

The segments corresponding to subtype B include from nucleotide 2058 $(+/-5$

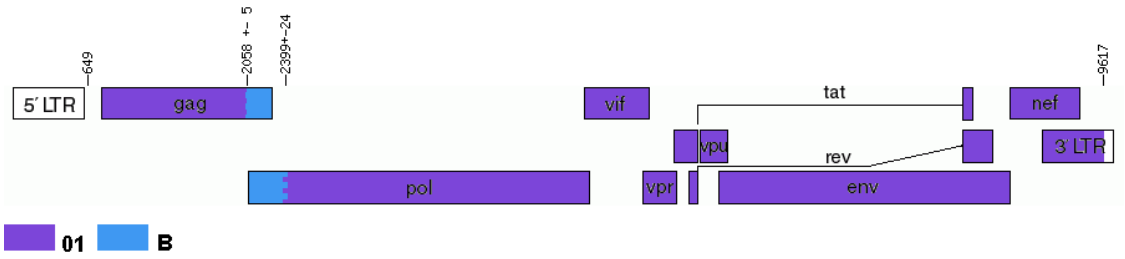


Figure 5.27: Genome Map of Recombinant Subtype CRF53_01B. This subtype results from genetic recombination between subtype CRF01_AE and subtype B. Image sourced from the Los Alamos National Laboratory HIV-1 Database HIV Circulating Recombinant Forms (CRFs) [41].

nts) to nucleotide 2399 ($+/-$ 24 nts), encompassing the final fragment of *gag* and the initial fragment of *pol*.

Regarding subtype CRF52_01B, the genome map is shown in Figure 5.28.

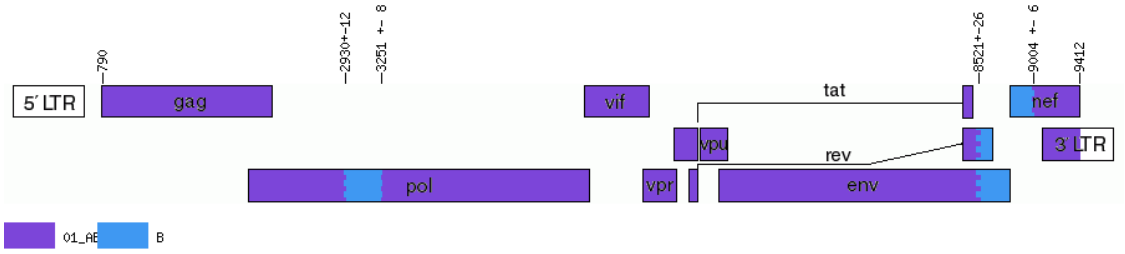


Figure 5.28: Genome Map of Recombinant Subtype CRF52_01B. Recombinant between recombinant subtype CRF01_AE and subtype B. Image sourced from the Los Alamos National Laboratory HIV-1 Database HIV Circulating Recombinant Forms (CRFs) [41].

The uncolored zones could not be identified with any subtype.

In this subtype, nearly the entire genome is classified as belonging to subtype CRF01_AE, with the exception of a fragment of the *pol* gene located between nucleotide 2930 ($+/-$ 12 nts) and nucleotide 3251 ($+/-$ 8 nts), which is classified as belonging to subtype B. Another fragment identified as subtype B is located between position 8521 ($+/-$ 26 nts) and position 9004 ($+/-$ 6 nts). This region involves the *env*, *rev*, and *nef* genes.

Although both recombinants are a mixture of the same subtypes (CRF01_AE and B), the result of recombination is notably different, as can be seen from both

genome maps. It would be interesting to analyze if there is any relationship between the predominance of CRF01_AE in both subtypes and the similarities between the hot zones of DQ366663 and DQ366664.

Figure 5.29 displays the excerpt of the phylogenetic tree corresponding to these two sequences.

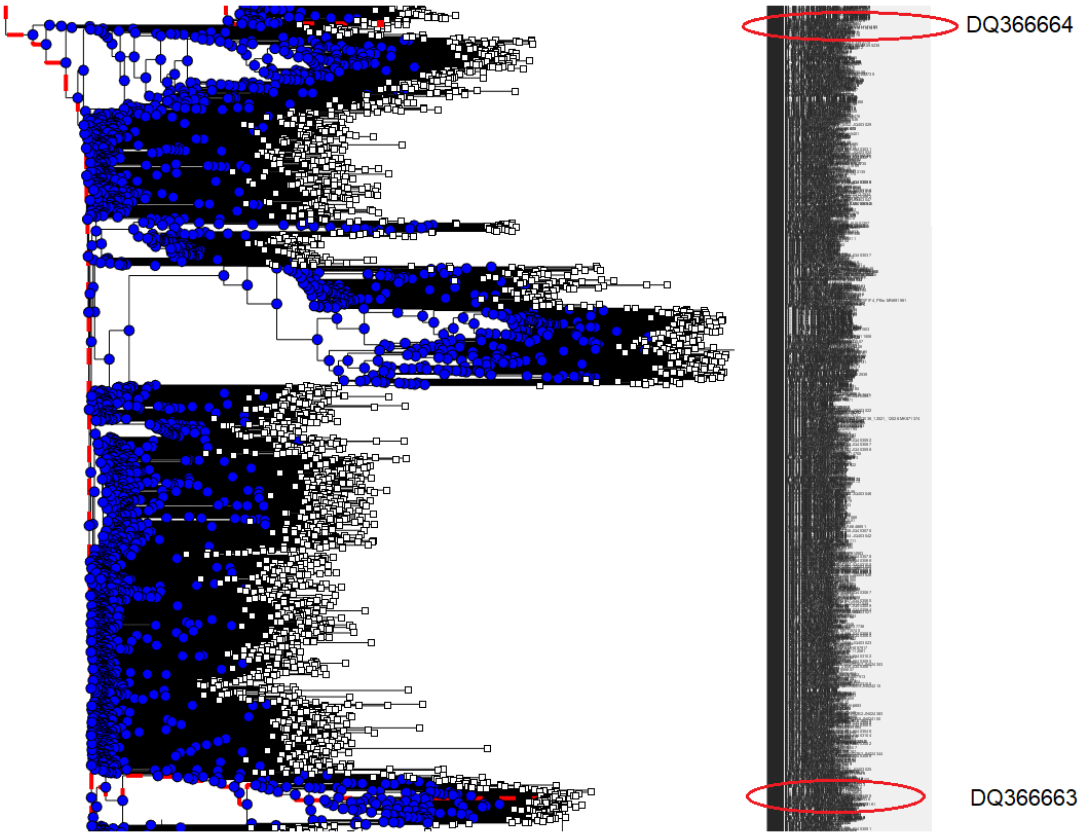


Figure 5.29: Phylogenetic Tree Including Subtype CRF53_01B DQ366663 and Subtype CRF52_01B DQ366664 Sequences. The phylogenetic distance between these two sequences falls in the range of 0.22-0.23, emphasizing a notable genetic divergence that is visually evident in the image.

The phylogenetic distance between both sequences is 0.22-0.23. It is interesting to note that despite this considerable distance, the hot zones were remarkably similar in concatenated and superposed spectrograms. This raises the possibility that genomic analysis using genomic spectrograms may reveal features that would not be apparent when working solely with phylogenetic distances. A deeper research of this phenomenon is needed.

5.15.3 Analysis of Failed Classifications

A failed sequence is one that the CNN misclassifies in the Test Set, meaning that a non-recombinant sequence is incorrectly categorized as recombinant, and vice versa.

The failed sequences of Non-Recombinant Concatenated Spectrograms, are shown in Table 5.27.

Out of a total of 150 non-recombinant sequences in the Test Set, the CNN misclassified only five sequences. Consequently, the Test Accuracy for concatenated non-recombinant sequences was 96.67%.

Table 5.27: Non-Recombinant Concatenated Failed Results. Those sequences misclassified also in superposed spectrograms are highlighted in orange.

Sequence ID	Subtype
AY371158	F2
FJ389363	G
KU724104	B
KX389648	G
MH330346	A6

The failed sequences of Recombinant Concatenated Spectrograms in the Test Set are listed in Table 5.28.

There was a total of 13 errors in the classification of recombinant sequences in the concatenated spectrograms. Therefore, the accuracy rate in the Test Set was 91.33%.

Table 5.28: Recombinant Concatenated Failed Results. Those sequences misclassified also in superposed spectrograms are highlighted in orange.

Sequence ID	Subtype
FM877780	CRF26_A5U
GU230137	CRF14_BG
HQ385479	CRF49_cpx
JN864049	CRF22_01A1
JN864058	CRF22_01A1
JX390977	CRF54_01B
JX679207	CRF57_BC
KC870043	CRF64_BC
KC898979	CRF88_BC
KM248765	CRF73_BG
KU168263	CRF92_C2U
MN172223	CRF108_BC

Table 5.28 – Continuation from the previous page

Sequence ID	Subtype
MT559132	CRF108_BC

The Test Accuracy of concatenated spectrograms was higher for non-recombinants than recombinants, though both test accuracy rates exceeded 90.00%.

In the case of Non-Recombinant Superposed Spectrograms, failed sequences are listed in Table 5.29.

Out of the total 150 non-recombinant sequences that make up the Test Set, the CNN misclassified only seven sequences, resulting in a Test Accuracy of 95.33%.

Table 5.29: Non-Recombinant Superposed Failed Results. Those sequences misclassified also in concatenated spectrograms are highlighted in orange.

Sequence ID	Subtype
AF110977	C
DQ168576	G
EU861977	A6
FJ389363	G
JX140676	G
KU677994	B
MK643588	C

Although the number of misclassified sequences in superposed non-recombinants was slightly higher than in concatenated non-recombinants, the accuracy rates were remarkably high (Tables 5.11, 5.13, and 5.14).

In the case of Recombinant Superposed Spectrograms, the list of failed sequences is shown in the Table 5.30.

The Test Accuracy was 92.67% since the CNN made errors in 11 out of the 150 instances in the Test Set.

Table 5.30: Recombinant Superposed Failed Results. Those sequences misclassified also in concatenated spectrograms are highlighted in orange.

Sequence ID	Subtype
AY586541	CRF18_cpx
GU230137	CRF14_BG
HQ385479	CRF49_cpx
JN864058	CRF22_01A1

Table 5.30 – Continuation from the previous page

Sequence ID	Subtype
JX390977	CRF54_01B
KC473840	CRF47_BF1
KU749419	CRF02_AG
KX907417	CRF41_CD
MN172223	CRF108_BC
MT276998	CRF106_cpx
MT559132	CRF108_BC

In the case of superposed spectrograms for recombinants, the error rate was slightly lower than in the concatenated case, although this difference was not significant.

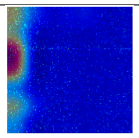
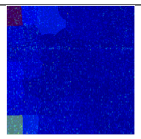
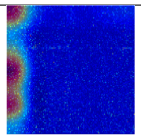
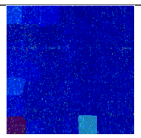
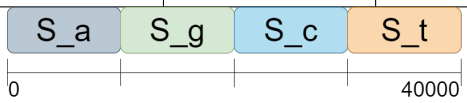
Thus, both in concatenated and superposed spectrograms, the Test Accuracy was higher for non-recombinants than for recombinants, although high values of hit rate were achieved in all cases.

I need to highlight that in multiple cases, the CNN misclassified the same sequence in both datasets. Specifically, the non-recombinant sequence FJ389363, and the recombinant sequences GU230127, HQ385479, JN864058, JX390977, MN172223, and MT559132. These findings could imply that these sequences exhibit distinct, special biological features that set them apart. As a result, it is advisable to conduct an extensive analysis regarding their genome integrity and accurate classification within their designated subtypes. These dual misclassifications may not be random and could point to underlying challenges in categorizing these sequences within their respective subtypes.

Below, I provide examples of incorrect classification and their comparative analysis between the hot zones and the position of the involved sequences in the phylogenetic tree.

In Table 5.31, we can observe an instance where the CNN made an incorrect classification by labeling the concatenated spectrogram of the non-recombinant sequence FJ389363 (subtype G) as a recombinant. Grad-CAM centered on the A spectrogram, LIME on the lower frequencies of A, and Gradient Attribution on the higher frequencies of A, as well as the lower frequencies of T and C.

Table 5.31: Non-Recombinant FJ389363 (subtype G) Erroneously Classified as Recombinant (Concatenated Spectrograms). Comparison with the hot zones of correctly classified recombinant sequence AJ508596 (subtype CRF30_0206).

Decision Patterns Comparison				
Non-Recombinant: failed classification as Recombinant				
Concatenated Spectrograms				
Subtype	**Sequence ID**	**Grad-CAM**	**LIME**	**Gradient Attribution**
G	FJ389363			
CRF30_0206	AJ508596			
	Reference			
				

Upon analyzing the Grad-CAM hot zones, I noticed a strong resemblance to the hot zones of AJ508596 (a recombinant sequence of subtype CRF30_0206). To a lesser extent, Gradient Attribution heatmaps also exhibited similarities, while LIME, although distinct, still shared some resemblances.

If we extract the position of both sequences in the phylogenetic tree (see Figure 5.30), we observe that the phylogenetic distance is not excessively high (0.142280). Upon a brief examination of the phylogenetic tree excerpt, we observe that AJ508596 (subtype CRF30_0206) is situated within a clade consisting of various types of recombinant sequences. The majority of these sequences belong to subtype CRF06_cpx, which is one of the parental subtypes of CRF30_0206. FJ389363 is grouped with sequences of subtype G.

A comprehensive analysis of this region of the phylogenetic tree, including the specific lineages to which these sequences belong and their evolutionary dynamics, is needed.

In the following example (Table 5.32), I encountered an instance in which a recombinant sequence was inaccurately categorized as non-recombinant. JN864049

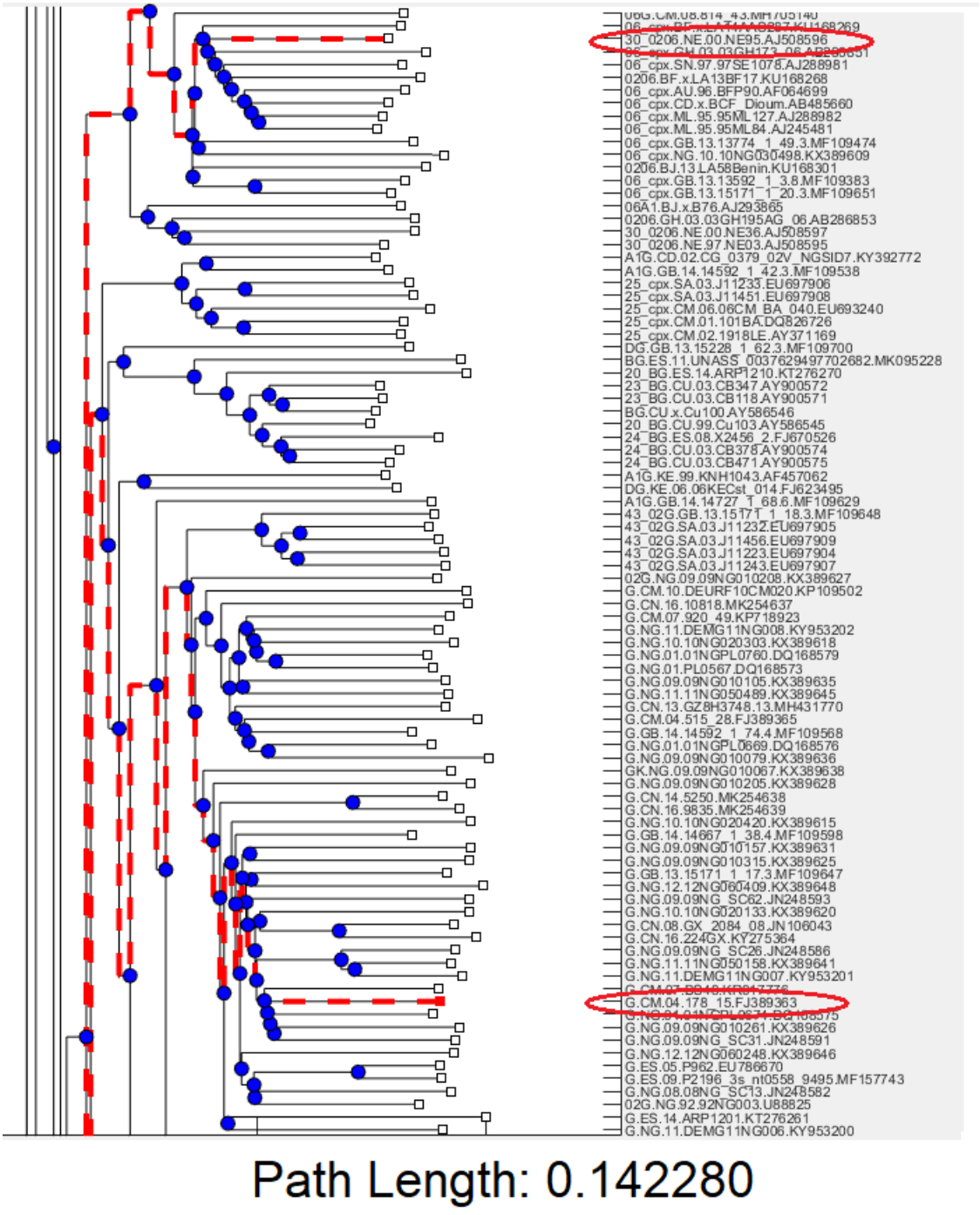
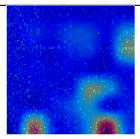
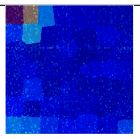
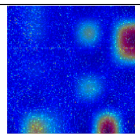
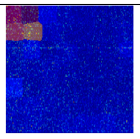
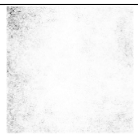
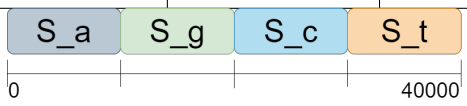


Figure 5.30: Partial Phylogenetic Tree Including Subtype G FJ389363 and Subtype CRF30_0206 AJ508596 Sequences. Their path length is 0.142280.

(subtype CRF22_01A1) was erroneously classified as non-recombinant. A comparison of its hot zones with those of DQ164119 (a non-recombinant sequence of subtype C) once again revealed a remarkable degree of agreement between them, even with Gradient Attribution.

Table 5.32: Recombinant JN864049 (subtype CRF22_01A1) Erroneously Classified as Non-Recombinant (Concatenated Spectrograms). Comparison with the hot zones of correctly classified non-recombinant sequence DQ164119 (subtype C).

Decision Patterns Comparison				
Recombinant: failed classification as Non-Recombinant				
Concatenated Spectrograms				
Subtype	**Sequence ID**	**Grad-CAM**	**LIME**	**Gradient Attribution**
CRF22_01A1	JN864049			
C	DQ164119			
	Reference			
				

Grad-CAM highlighted low frequencies between A and G, along with some areas near $f = 1/6$, $f = 1/3$, and low frequencies in C and T. LIME focused on the high frequencies of A, and Gradient Attribution was most active across the entire frequency spectrum of A (especially high frequencies), followed by T (primarily in $f = 1/3$ and low frequencies)

In this case, it is not possible to determine the phylogenetic distance between both sequences, as JN864049 does not have a complete intra-patient reference sequence (Patient ID: LPH27MF).

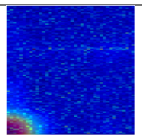
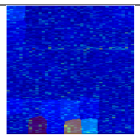
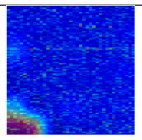
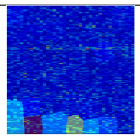
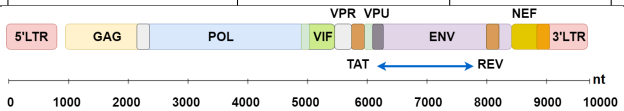
Below, I describe two examples of misclassification in the case of superposed spectrograms.

In the example corresponding to Table 5.33, FJ389363 (subtype G) was misclassified as recombinant. When comparing its hot areas with those of DQ366663

(recombinant sequence of subtype CRF53_01B), a significant level of agreement was observed between them, even in Gradient Attribution. The incorrect patterns resembled those of the recombinant determinant areas.

Likewise, the CNN mislabeled the superposed spectrogram of the FJ389363 sequence. In this scenario, the Grad-CAM and LIME heatmaps in both sequences under scrutiny exhibited striking similarities. Conversely, the Gradient Attribution heatmaps displayed noticeable distinctions that were readily discernible. In any case, there was a direct correlation between the misclassification and the presence of hot zones corresponding to the incorrect category.

Table 5.33: Non-Recombinant FJ389363 (subtype G) Erroneously Classified as Recombinant (Superposed Spectrograms). Comparison with the hot zones of correctly classified recombinant sequence DQ366663 (subtype CRF53_01B).

Decision Patterns Comparison				
Non-Recombinant: failed classification as Recombinant				
Superposed Spectrograms				
Subtype	**Sequence ID**	**Grad-CAM**	**LIME**	**Gradient Attribution**
G	FJ389363			
CRF53_01B	DQ366663			
	Reference			
				

Grad-CAM highlighted the low frequencies of 5'LTR. LIME, on the other hand, exhibited a more scattered distribution across the low-frequency range, with prominent areas located between 5'LTR and *gag*, and in some instances near *vif*. Gradient Attribution hot zones appeared to concentrate on both low and high frequencies.

In view of the phylogenetic position of both sequences (see Figure 5.31), I observed that these sequences are phylogenetically distant. Their path length is approximately 0.23-0.24. Therefore, I could not draw any clear conclusions from this data.

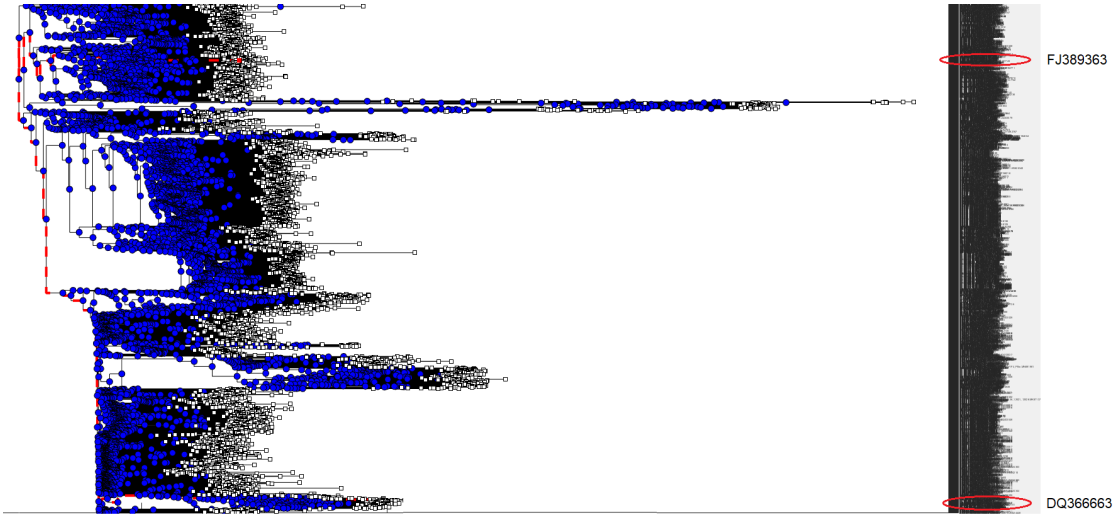


Figure 5.31: Partial Phylogenetic Tree Including Subtype G FJ389363 and Subtype CRF53_01B DQ366663 Sequences. The phylogenetic distance between these two sequences falls within the range of 0.23-0.24.

A more in-depth analysis is required to ascertain if there is any underlying reason why the CNN classified FJ389363 as recombinant. For instance, whether hidden genetic recombination phenomena may exist within the genome of this sequence or within subtype G, or if there could be other biological factors that are more discernible in the frequency spectrum than with traditional genomic analysis methodologies.

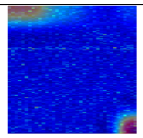
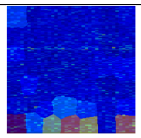
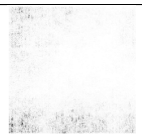
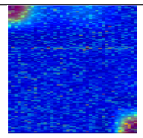
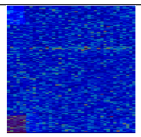
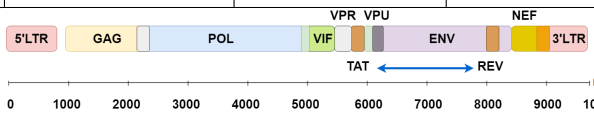
The data extracted from the phylogenetic tree is inconclusive since the phylogenetic distance does not correspond to the degree of similarity between the hot zones. However, this disparity has also been observed in cases of accuracy, as seen in Table 5.26 and Figure 5.29.

In the example shown in Table 5.34, in superposed spectrograms, the recombinant sequence KX907417 (subtype CRF41_CD) was erroneously classified as non-recombinant. Its hot zones exhibited significant similarity to those of the non-recombinant sequence MN090373 (subtype B). High similarity was notably observed

in the Grad-CAM hot zones. In the case of LIME, the KX907417 hot zones were less specific than those in the case of MN090373. Nevertheless, similarities in the Gradient Attribution hot zones could be easily discerned.

The Grad-CAM hot regions corresponded to the high frequencies of 5'LTR and the low frequencies of 3'LTR. In contrast, the LIME results displayed a more scattered pattern, with its primary determinant zone located in the low-frequency range of 5'LTR. Additional areas of interest were detected within the low-frequency range, particularly in the central region of the genome (around the end of *pol* and *vif*), as well as in the 3'LTR. As for Gradient Attribution, its determinant points spanned across both the low-frequency and high-frequency ranges of 5'LTR.

Table 5.34: Recombinant KX907417 (subtype CRF41_CD) Erroneously Classified as Non-Recombinant (Superposed Spectrograms). Comparison with the hot zones of correctly classified non-recombinant sequence MN090373 (subtype B).

Decision Patterns Comparison				
Recombinant: failed classification as Non-Recombinant				
Superposed Spectrograms				
Subtype	**Sequence ID**	**Grad-CAM**	**LIME**	**Gradient Attribution**
CRF41_CD	KX907417			
B	MN090373			
	Reference			
				

In this case, it is not possible to determine the phylogenetic distance because KX907417 does not have an assigned complete intra-patient reference sequence (Patient ID: 6650).

Upon analyzing non-recombinant sequences that were inaccurately categorized as recombinant, it appeared that the CNN's decision-making relies on the recombinant

hot areas. The same phenomenon occurred in the reverse direction, when the CNN failed to classify a recombinant sequence as non-recombinant, the decision zones looked like those associated with non-recombinants. It is necessary to analyze these phenomena considering other perspectives beyond phylogenetic closeness.

5.16 Conclusions

The methodology implemented in this Chapter, as discussed in Section 5.4, yielded outstanding and resilient outcomes, taking into consideration the inherent challenge of mutations occurring in both recombinant and non-recombinant sequences. Consequently, the complexity of the task was heightened. To gain insights into the CNN's decision-making process for sample classification, I harnessed a range of explainability tools.

This methodology not only facilitated the detection of the recombinant feature in each sample but also enabled the identification of the specific genomic regions where this characterization was located. This approach enhanced our understanding of genetic recombination processes in HIV-1 by providing detailed information about the precise locations of recombinant features. This additional knowledge could contribute a deeper insight into the evolution and genetic diversity of HIV-1.

Furthermore, when utilizing genomic spectrograms, I could identify mathematical patterns related to the frequency spectrum and mathematical signatures embedded in the genome that characterize an HIV-1 sequence as recombinant. My methodology enabled the swift, robust, and systematic identification of specific genomic regions crucial for characterizing genetic recombination in HIV-1. This targeted approach could significantly expedite genomic research, allowing microbiological experts to concentrate their efforts on these precise genomic areas. This not only accelerates the research process but also enhances the efficiency and accuracy of HIV-1 genetic analysis.

It is important to emphasize that this type of analysis using traditional manual genomics approaches would be exceedingly complex, further underscoring the utility of deep learning technology in this field.

My methodology allows for a complementary representation of genomic sequences, in addition to the standard representation of genomic sequences utilizing letters that indicate DNA bases. The conversion of genomic sequences into biologically consistent images, such as spectrograms, enables the application of powerful Deep Learning and Machine Learning tools in the field of genomics, thus harnessing their full potential. This implies that advanced image processing and Machine Learning algorithms can be utilized to analyze and gain a better understanding of genomic structure and features, consequently opening new research and discovery opportunities in genomics.

Despite genomic spectrograms having notably different characteristics compared to traditional images such as those in ImageNet, this successful application demonstrates the adaptability and effectiveness of these tools in new domains. The capability to efficiently identify regions of interest within genomic sequences through extensive information comparison is a crucial aspect of this technology.

The results obtained highlight the versatility and applicability of Deep Learning and image processing techniques in unconventional contexts, such as genomics.

The CNN achieved a test accuracy of 94% in classifying viral sequences as either recombinants or non-recombinants, and this level of accuracy was maintained consistently for both concatenated and superposed spectrograms. Moreover, the decision boundaries for each category were well-defined and distinguishable.

A correlation was observed between the degree of similarity in the highlighted areas of sequences of the same subtype and their phylogenetic proximity. This suggests that the similarities in the regions of interest of the sequences are linked to their evolutionary relatedness.

When the CNN made classification errors, both Grad-CAM and LIME focused on the hot zones of the incorrect category. This finding suggests consistency in the behavior of the CNN regarding the identification of areas of interest, even in situations where misclassification occurs.

As this research is entirely multidisciplinary, my goal should not be limited to analyzing the methodology and results solely from a computational perspective. It is essential to evaluate the results from microbiological viewpoints. The validation

of the obtained results should primarily rely on biological evidence. Therefore, I conducted a comprehensive biological analysis.

The importance of the 5' and 3' LTR zones and their U3, R, and U5 regions as critical areas for classifying a sequence as recombinant may be linked to the roles played by these terminal genome sections during viral transcription [2470–2472]. These regions are crucial in regulating gene expression and virus replication. As a result, their relevance to CNN-based classification may be associated with specific genomic elements within the frequency spectrum in these determinant zones, which are essential for detecting the recombinant feature. The adjacent *gag*, *nef*, and *env* genes could also play a role in HIV-1 genetic recombination.

The significance of A and T nucleotides in the recombinant feature is associated with the elevated presence of polyA and polyT stretches in the HIV-1 sequence. These specific regions are prone to polymerase skipping, thereby facilitating genetic recombination. The prominent role of the A nucleotide might be linked to the abundant Adenine content within lentiviruses, particularly in the case of HIV-1 [2473].

As postulated by Olabode et al. (2022) [2474], the similarity in hot zones between sequences of different subtypes could stem from the presence of more instances of genetic recombination than currently acknowledged. It is plausible that certain viruses traditionally categorized as pure subtypes may harbor hidden recombination events.

Furthermore, the identified recombinant sequences are the ones that prevailed due to the competitive advantages conferred by the resulting viruses. While there may be various recombination events taking place, those events will not be favored if the resulting virus is not viable or exhibits lower fitness compared to the parental viruses. This research provides us with the opportunity to gain a deeper understanding of genomic regions that are particularly prone to recombination and could potentially offer an evolutionary advantage to the virus [2475].

Certain sequences exhibited misclassification in both concatenated and superposed spectrograms. It is essential to confirm the integrity and consistency of these

published genomes. Furthermore, the criteria for categorizing these sequences into their respective subtypes should be reevaluated.

This phenomenon could potentially reinforce the CNN's proper functioning. In cases where these sequences are incorrectly classified by subtype or contain genomic errors, it is biologically reasonable for the CNN to exhibit failures in both datasets.

In summary, the primary hot zones for distinguishing between recombinant and non-recombinant sequences are the 5'LTR and 3'LTR regions, and nucleotide A and, to a lesser extent, T. These findings align with the biological attributes of genome structure and retrotranscription functions. This further supports the notion that the CNN identifies frequency-domain mathematical patterns that define the nature of a genomic sequence as recombinant or non-recombinant, along with their precise genomic locations.

5.17 Future Studies

Utilizing techniques grounded in multiple sequence alignment and manual adjustments did not allow us to reveal the underlying genetic recombination patterns in viruses. Despite the loss of information resulting from converting sequences into spectrograms, it becomes feasible to unveil previously undiscovered insights encoded within the genome. These insights expose patterns that remain hidden within the nucleotide sequence itself.

The potential of Deep Learning algorithms to uncover concealed patterns has the potential to significantly advance research in this field. To accomplish this, it is imperative to delve into comprehending the functioning of these networks and their pattern detection mechanisms.

Detailed researchs of established pre-trained architectures and the creation of explainability tools are indispensable at this point. With the advancement of more precise neural networks applied to genomics, scientists will gain the capability to pinpoint the exact genomic regions that significantly contribute to sequence characterization with heightened accuracy and confidence. As a result, the time needed for sequence analysis and comparison will undergo a substantial

reduction. This has particular implications for larger organisms with genomes spanning millions of nucleotides. Traditional analysis methods for such sequences involve intensive computational processes and are exceedingly time-consuming, often requiring manual adjustments at the level of individual letters. In these situations, it becomes paramount to have a tool capable of identifying the specific genomic regions relevant to the study's focus. My methodology can address this requirement by providing accurate identification, thus significantly improving the efficiency of microbiological analyses.

Moreover, it is crucial to provide these algorithms with meticulously curated datasets to assist in pattern recognition. To train and test the neural network effectively, I had to utilize the entire dataset of available HIV-1 complete sequences, which demands a substantial volume of data samples for robust performance. To address this requirement, there is a growing need to expand sequencing efforts across various organisms to bolster the availability of diverse and comprehensive datasets.

This study marked the pioneering application of a pre-trained CNN to spectrogram datasets for the detection and classification of HIV-1 sequences as recombinant. The results were highly successful and found strong biological support. Additionally, the utilization of explainability tools enabled the identification of specific genomic regions associated with this classification. While my methodology currently lacks direct clinical applications, it contributes significantly to advancing our understanding of genetic recombination in HIV-1. This understanding has potential implications for the future, such as aiding in the search for new antiretroviral drugs targeting precise genomic regions critical for the development of effective treatments.

This methodology offers a significantly expedited process with reduced computational expenses compared to traditional techniques. It eliminates the necessity for labor-intensive multiple sequence alignments and manual adjustments. The findings strongly suggest the presence of mathematical patterns within the genome that define distinct viral features. This prompts further exploration to identify and define these mathematical formulations. Among the explainability methods, Gradient Attribution stands out as the most suitable for precise signature localization due to

its high accuracy. Although it is the least intuitive tool for the human eye due to its noisy appearance, it may be the one that allows extracting the most information from its hot maps with the application of image preprocessing technique.

I have successfully pinpointed the locations where the CNN identifies the recombinant feature. My subsequent objective involves the quest for a signature or distinctive genetic marker that can facilitate the identification of analogous variants or sequences with significantly reduced computational resources.

Projects like the one outlined here have the potential to effectively lead and streamline the challenging process of characterizing critical regions. This approach may find broader applications in the study of more intricate life forms and their defining traits.

Many exciting questions and fields emerge from this research.

It's bad enough people are dying of AIDS, but no one should die of ignorance.

— Elizabeth Taylor, *One of the people who has done the most for the well-being of AIDS patients. She also said, "I don't think President Bush (senior) is doing anything at all about AIDS. In fact, I'm not sure he can even spell the word AIDS".*

6 Analysis of Total Hot Zones in HIV-1 by Subtype

Contents

6.1 Introduction 253
6.2 Equipment 254
6.3 Methods 255
6.4 HIV-1 Classification Error Rates per Subtype 257
6.5 Total Hot Zones per Category 259
6.6 HIV-1 Total Hot Zones per Subtype 261
6.6.1 Non-recombinant Subtypes 262
6.6.2 Recombinant Subtypes 274
6.7 Conclusions 290
6.8 Future Studies 293

6.1 Introduction

In Chapter 5, I applied a pre-trained CNN (VGG-16) to a dataset containing genomic spectrograms of all existing complete HIV-1 sequences until 2020, and I classified HIV-1 sequences as recombinant or non-recombinant. Thanks to explainability tools, I determined sequence by sequence the relevant areas where the network looks to detect the recombinant feature and subsequently classify the genomic sequence.

I also identified distinct hot zone patterns in each of the two categories in the

Test Set. In the case of concatenated spectrograms, the hot zones provided us with information on relevant nucleotides in categorizing a sequence as recombinant or non-recombinant. In the case of superposed spectrograms, I discriminated the genomic areas involved.

In light of the results obtained in Chapter 5, I wanted to delve deeper into the understanding of the mathematical patterns inherent in classifying sequences as recombinants by subtype. Furthermore, aligned with the aforementioned, I aimed to deepen our intrinsic understanding of the functioning of the CNN by classifying these genomic sequences as recombinant and non-recombinant. For the latter, I aimed to establish a comparison between the classification error rates associated with each subtype and the resulting total hot zones, not only from the sequences comprising the Test Set but also from the 13,554 complete HIV-1 sequences identified until 2020.

This Chapter consists of a description of the equipment (Section 6.2) and the methods used (Section 6.3). Section 6.4 presents the error rates obtained for non-recombinant and recombinant subtypes. Additionally, Section 6.5 shows the total hot zones for both categories, as well as their description from a microbiological perspective. Subsequently, I conducted a comparison between the total hot zones by subtype and those corresponding to their category in Section 6.6. This Chapter concludes with the findings derived from this part of the research (Section 6.7), as well as future research directions based on the analysis performed (Section 6.8).

6.2 Equipment

All experiments were run in this equipment:

- Processing Unit: Intel(R) Xeon(R) Gold 5220R CPU @ 2.20 GHz.
- Installed RAM: 256 GB usable.
- Operative System: Windows 10 Education. Version: 21H1.
- GPU: NVIDIA RTX A6000. Total memory: 179.451 GB. VRAM: 48.571 GB.

6.3 Methods

As outlined in Chapter 5, Section 5.11, Grad-CAM is the most intuitive and visually comprehensible explainability tool among the three utilized. Therefore, I consider it to be the tool of choice for the study undertaken in this Chapter.

To establish the total Grad-CAM hot zones, I performed a scalar summation of the scoremaps for each of the sequences using MATLAB 2021b.

$$TotalScoreMap = \sum_{i=1}^{n} ScoreMap_i \tag{6.1}$$

Once the summative scoremap of all the scoremaps for each sequence was obtained, an image was generated by applying a jetmap color map to visualize this total scoremap.

This technique highlights areas where there is a higher accumulation of activation in the scoremaps in the dataset, optimizing computation. Conceptually, it can be considered similar to an aggregation of relevant information. The scalar summation represents a weighted combination of activations for each pixel in the different feature maps. By applying the jetmap, I assigned a color coding to visualize the intensity of the combined activations.

In this way, I determined the common hot zones within the study group and graphically represent them for subsequent interpretation and analysis.

I calculated the summative scoremap for all non-recombinant sequences and, separately, for all recombinant sequences. This allowed me to obtain the total hot zones for non-recombinant sequences and the total hot zones when working with the complete dataset. Furthermore, I performed the same procedure by subtype, summing all sequences classified within each of the subtypes under study. The calculation and visualization of the total hot zones were only carried out for those subtypes that include more than 15 sequences. I referred to them as representative subtypes. This approach allowed me to reduce the variability inherent in handling few samples and obtain a representative image of the true central tendency of

the data. I performed this calculation and this analysis for both concatenated and superposed spectrograms.

As an additional component for conducting the analysis, I also computed the CNN error rates for each representative subtype. Additionally, as supplementary data, I calculated the hit rate for each representative subtype.

The next step was to perform a comparison between the total hot zones of each representative subtype and the total hot zones of the corresponding category. If the subtype was non-recombinant, I compared it with the total hot zones corresponding to the non-recombinant subtypes, and similarly if the representative subtype was recombinant.

In this comparison, I included the error rate per subtype as an additional element in the comparative analysis of the hot zones by subtype versus the total hot zone corresponding to its category. The error rate is calculated as:

$$\text{Error rate} = \frac{\text{Number of incorrect classifications}}{\text{Total number of classifications}} \tag{6.2}$$

In a similar manner, I calculated the hit rate as:

$$\text{Hit rate} = \frac{\text{Number of correct classifications}}{\text{Total number of classifications}} \tag{6.3}$$

By including error rates in the analysis comparing the similarity between the hot zones by subtype with the total ones corresponding to their category, I obtained additional information regarding the divergences that may occur between both images. Therefore, I indirectly obtained a measure of the consistency in the CNN's operation.

I also consulted expert literature by subtype to gather additional data that could be pertinent in this study.

6.4 HIV-1 Classification Error Rates per Subtype

All representative subtypes with a classification error rate greater than 0.00%, along with the error rates for each subtype are displayed in In Tables 6.1 and 6.2. Table 6.1 corresponds to non-recombinant subtypes, and Table 6.2 to recombinant subtypes.

Representative errors, defined as those above 15%, are highlighted in orange color. The error of any subtype with 15 sequences or less is not significant due to the lack of data. All the subtypes not included in these tables had an error rate of 0.00%.

I calculated the error rates by subtype for the case of superposed spectrograms, given the greater balance in terms of errors in the Test Set between the recombinant and non-recombinant categories compared to concatenated ones. As shown in Table 5.11, for VGG-16 (the selected pre-trained CNN), there were five errors in the recombinant category versus 13 in the non-recombinant category for concatenated spectrograms. For superposed spectrograms, there were seven errors in the recombinant category and 11 in the non-recombinant category.

As highlighted in Table 6.1, subtype U had the most striking error rate (39.13%), followed by subtype G (24.24%) and F1 (19.51%).

Table 6.1: HIV-1 Non-recombinant Error Rates per Subtype. Subtypes consisting of more than 15 sequences and whose error rate exceeds 15% are highlighted in orange.

Subtype	No. Errors	No. Sequences	Error (%)
A	2	3	66.67%
A1	19	709	2.68%
A6	6	186	3.23%
B	13	7730	0.17%
C	37	2105	1.76%
D	2	102	1.96%
F1	16	82	19.51%
F2	2	14	14.29%
G	25	99	25.25%
H	1	10	10.00%
J	1	7	14.29%
L	1	3	33.33%
O	1	57	1.75%
P	1	5	20.00%
U	9	23	39,13%

Regarding recombinant subtypes CRF, there are over 100 CRF subtypes identified. Except CRF01_AE (1,490 complete sequences) and CRF02_AG (179 complete sequences), CRF subtypes include less than 100 complete sequences. And in many cases this number rounds 5-15 sequences, so their error rates were not considered. As highlighted in Table 6.2, the only significant error rate corresponded to CRF22_01A1 (19.05%).

Table 6.2: HIV-1 Recombinant Error Rates per Subtype. Subtypes consisting of more than 15 sequences and whose error rate exceeds 15% are highlighted in orange.

Subtype	No. Errors	No. Sequences	Error (%)
CRF02_AG	2	179	1.12%
CRF03_A6B	1	5	20.00%
CRF05_DF	2	4	50.00%
CRF07_BC	1	43	2.33%
CRF08_BC	1	35	2.86%
CRF11_cpx	3	25	12.00%
CRF12_BF	1	15	6.67%
CRF13_cpx	1	13	7.69%
CRF14_BG	7	13	53.85%
CRF17_BF1	1	7	14.29%
CRF18_cpx	1	7	14.29%
CRF19_cpx	3	5	60.00%
CRF21_A2D	1	3	33.33%
CRF22_01A1	4	21	19.05%
CRF28_BF1	2	5	40.00%
CRF31_BC	2	3	66.67%
CRF38_BF1	1	5	20.00%
CRF41_CD	1	3	33.33%
CRF47_BF1	2	5	40.00%
CRF49_cpx	1	4	25.00%
CRF50_A1D	2	5	40.00%
CRF51_01B	1	7	14.29%
CRF54_01B	1	3	33.33%
CRF57_BC	3	7	42.86%
CRF60_BC	1	5	20.00%
CRF61_BC	1	4	25.00%
CRF69_01B	1	7	14.29%
CRF70_BF1	1	4	25.00%
CRF71_BF1	3	14	21.43%
CRF72_BF1	1	5	20.00%
CRF87_cpx	3	3	100.00%

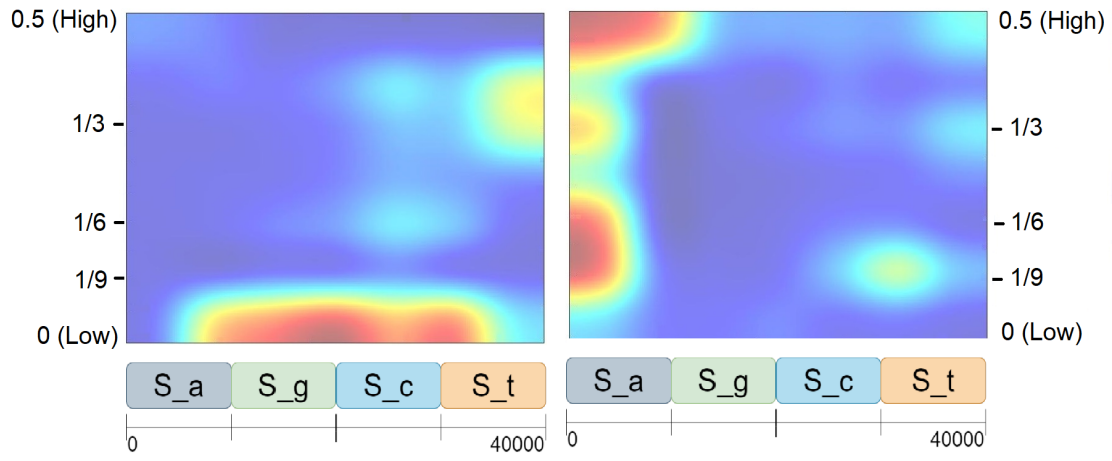
Table 6.2 – Continuation from the previous page

Subtype	No. Errors	No. Sequences	Error (%)
CRF90_BF1	3	6	50.00%
CRF105_0108	1	5	20.00%
CRF106_cpx	1	6	16.67%
CRF107_01B	1	4	25.00%
CRF108_BC	3	5	60.00%

In view of the error rates of all non-recombinant and recombinant subtypes, non-recombinant subtypes U and G exhibited the highest classification error rates, at 39.13% and 25.25% respectively. This phenomenon requires further research.

6.5 Total Hot Zones per Category

As seen in Figure 6.1, in concatenated spectrograms, the total hot zones across the entire dataset per category were clearly distinguishable between non-recombinant (see Figure 6.1a) and recombinant (see Figure 6.1b). That is, each category exhibited unequivocal patterns in its total hot areas.



(a) Non-recombinant **(b)** Recombinant

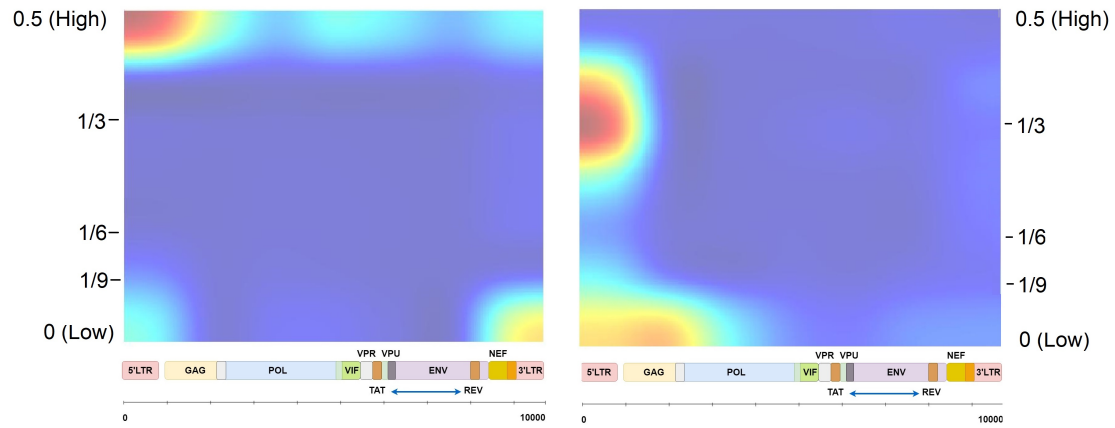
Figure 6.1: Total Hot Zones per Category (Concatenated Spectrograms)

In concatenated spectrograms, the main hot zones (in red + orange) of the non-recombinant category were located in the low-frequency range from the final third or quarter of the spectrogram corresponding to A, the spectrograms of G and C, and the first third of the T spectrogram. As for the secondary zones (from

yellow to lighter blue), the main total hot zone was located at $f = 1/3$ in the T spectrogram. Even more secondary zones were visually identified at $f = 1/3$ around the C spectrogram, in the frequency range from $f = 1/6$ to $f = 1/3$.

The main total hot zones (in red + orange) in the recombinant category were primarily located around the range between $f = 1/9$ and $f = 1/6$ in the A spectrogram and in the high-frequency range around A. A lesser-power total hot zone was detected at $f = 1/3$, also in the A spectrogram. Frequencies outside these total hot zones encompassed secondary total hot zones, likely due to radiation from the more prominent hot zones, all within the A spectrogram. Other secondary zones were detected at $f = 1/9$ in the transition zone between the C and T spectrograms, as well as at $f = 1/3$ and high frequencies in the second half of the T spectrogram. There were very minor zones, practically imperceptible, across the high-frequency range from the second half of the G spectrogram to the second half of the T spectrogram, as well as at $f = 1/3$ from the C spectrogram to the second half of the T spectrogram.

In the case of superposed spectrograms, I found the same phenomenon. The total hot zones for each category were perfectly distinguishable. See Figure 6.2. Figure 6.2a shows the total hot zones corresponding to the non-recombinant category in superposed spectrograms, and Figure 6.2b shows the total hot zones for the recombinant category in superposed spectrograms.



(a) Non-recombinant **(b)** Recombinant

Figure 6.2: Total Hot Zones per Category (Superposed Spectrograms)

Operating with superposed spectrograms, in the case of the non-recombinant category, the main total hot zone (in red + orange) was clearly located in the high-frequency range centered on the 5'LTR, extending up to the first half of *gag*. This zone irradiated up to the first third of *pol* as a more secondary zone. The next most significant total hot zone was located in the low-frequency range centered on the 3'LTR, extending to the vicinity of the end of *nef*. More secondary zones, shown in light blue, were visible in the low-frequency range from 5'LTR to the first half of *gag*, and in the high-frequency range from the second half of *pol* to the beginning of the second half of *env*, as well as from the final area of *env* to the end of the 3'LTR.

Regarding the recombinant category in superposed spectrograms, the main total hot zone (in red + orange) was clearly located in the vicinity of $f = 1/3$, with a distinct epicenter at 5'LTR, radiating up to the first half of *gag*. The second most significant total hot zone was in the low-frequency range in *gag*, irradiating on the right up to the first half of *pol*, and on the left, to the initial stretch of 5'LTR, connecting with another total hot zone of lesser power at the beginning of 5'LTR. Very secondary zones were perceived, barely visible, in the low-frequency range from the second half of *env* to the end of 3'LTR, as well as at $f = 1/3$ in 3'LTR.

The results in both categories in both types of spectrograms, showing a clear and significant distinction between the two categories in the explainability results reinforce the effectiveness of the CNN. It is indicative that the model learned relevant discriminative features for each category and can generalize them appropriately. This result suggests that the CNN was effectively capturing the distinctive characteristics associated with each class, supporting its solid performance in the classification task.

6.6 HIV-1 Total Hot Zones per Subtype

In this section, I present the resulting images from the calculation of the total Grad-CAM hot maps applicable to each representative subtype that includes more than 15 sequences, in both concatenated and superposed spectrograms, for non-recombinant and as recombinant categories.

I also included the relevant hot zones identified in Chapter 5 to compare the resulting patterns by subtype with the general hot zones.

Alongside the total hot zones of each representative subtype, I included total error rates and hit rates to establish a direct comparison between the obtained total hot zones and the performance of the CNN.

6.6.1 Non-recombinant Subtypes

Table 6.3 is a compilation of the total hot zones for the non-recombinant category and for each of the representative non-recombinant subtypes, facilitating comparability among all of them.

Table 6.3: Total Hot Zones per Non-Recombinant Category and Representative Subtypes. The columns refer to the two types of spectrograms utilized, and the rows to the non-recombinant category and each representative subtype, including scaled representations of the schematics for each type of spectrogram to facilitate the direct location of the hot zones.

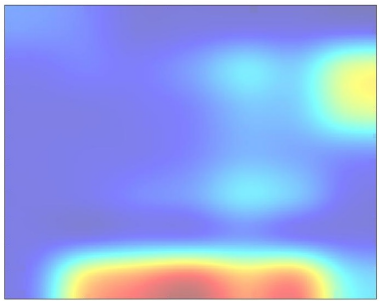
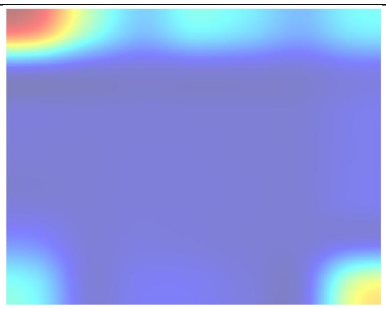
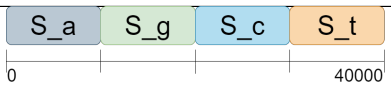
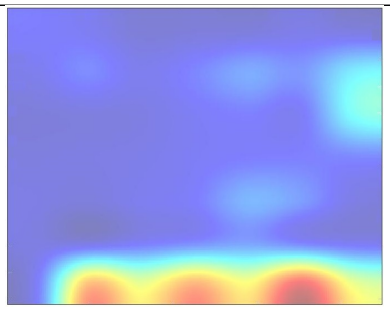
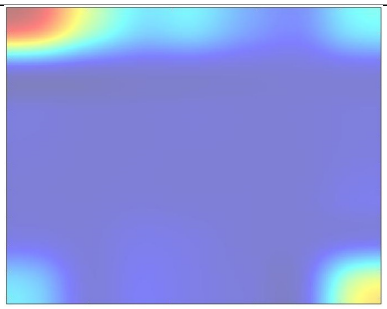
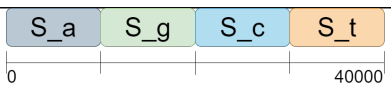
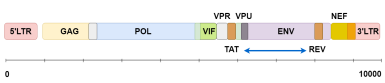
	Concatenated Spectrograms	**Superposed Spectrograms**
Non-Recombinant Category		
		
Subtype A1		
		

Table 6.3 – Continuation from the previous page

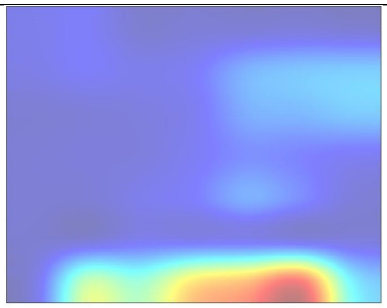
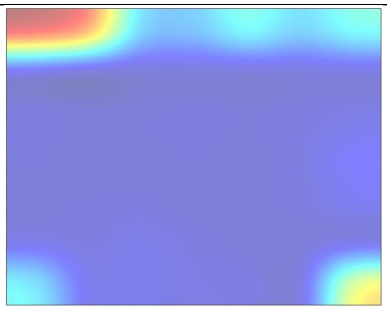
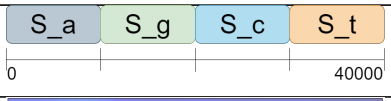
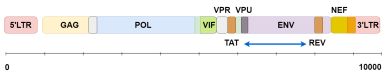
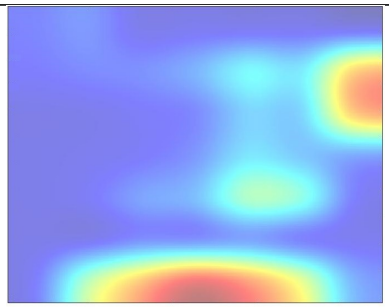
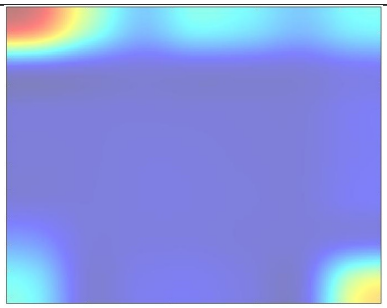
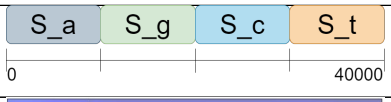
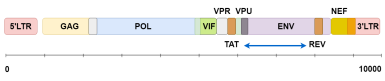
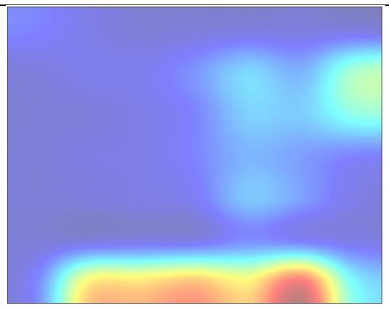
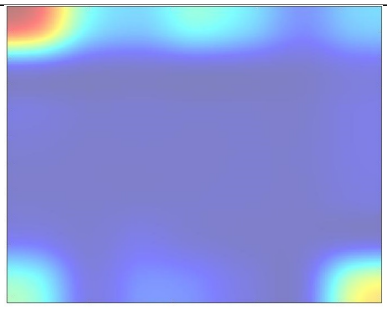
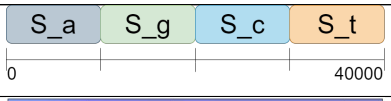
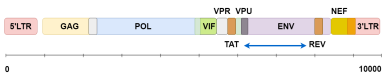
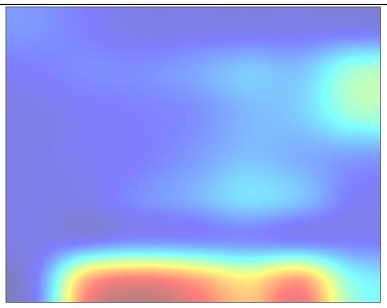
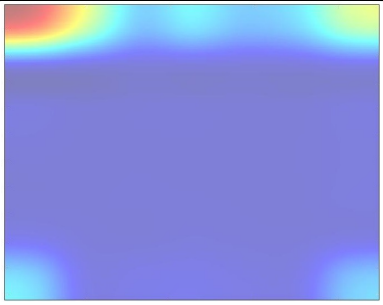
	Concatenated Spectrograms	Superposed Spectrograms
Subtype A6		
		
Subtype B		
		
Subtype C		
		
Subtype D		

Table 6.3 – Continuation from the previous page

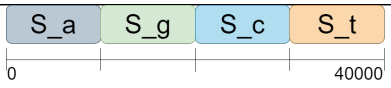
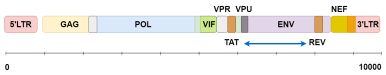
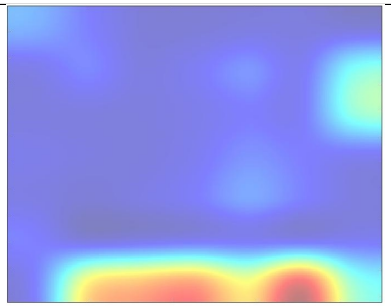
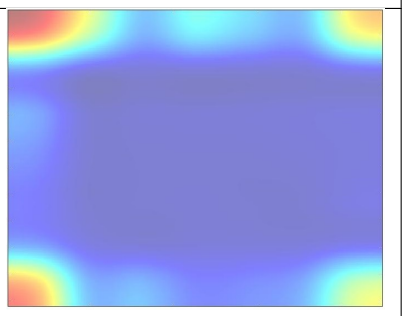
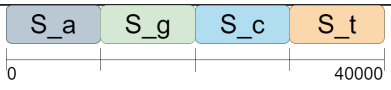
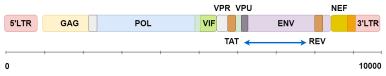
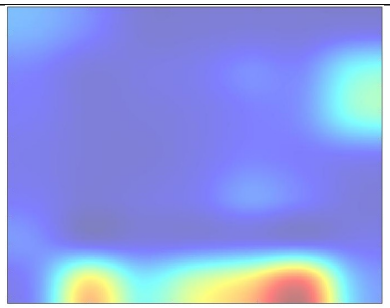
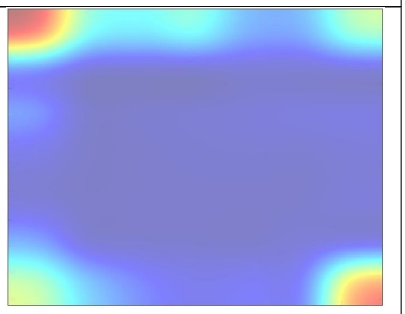
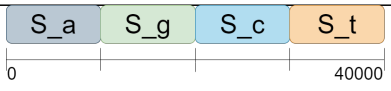
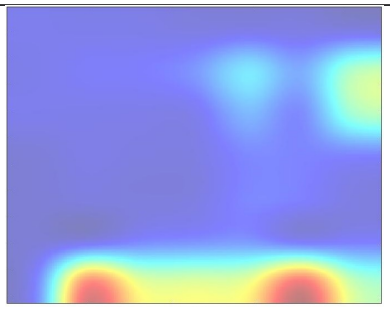
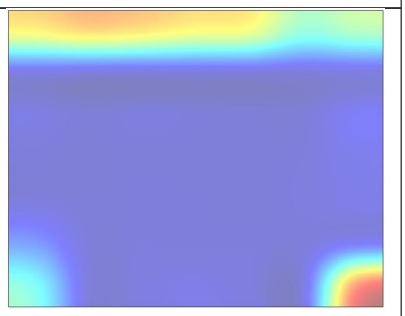
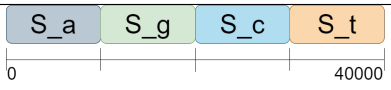
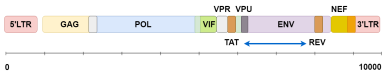
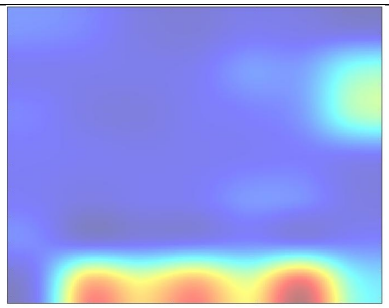
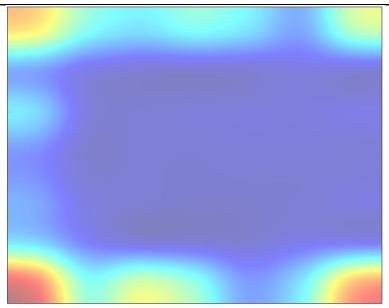
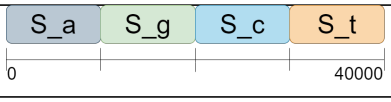
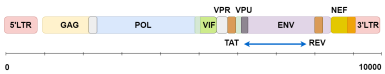
	Concatenated Spectrograms	Superposed Spectrograms
		
Subtype F1		
		
Subtype G		
		
Subtype O		
		

Table 6.3 – Continuation from the previous page

	Concatenated Spectrograms	Superposed Spectrograms
Subtype U		
		

I could easily appreciate the high degree of similarity between the hot zones corresponding to subtype A1 and the hot zones corresponding to the non-recombinant category in concatenated spectrograms. Regarding the differences between the total hot zones of A1 compared to the total non-recombinant hot zones, there was a loss of weight in all hot zones above the low-frequency range, as well as a greater segmentation in the hot zones within the low-frequency range in A1, compared to the total non-recombinant ones, which were more spread throughout this range.

The error rate corresponding to this subtype was only 2.68% (19 errors out of a total of 709 complete sequences). There was a clear correlation between the high hit rate corresponding to this subtype (97.32%) and the high degree of similarity between both images. Additionally, this subtype includes 709 sequences, making it the third most abundant subtype among the non-recombinants. Therefore, the specific weight of these positive results in the total of the category is notable. Ultimately, the degree of coherence in the CNN's operation was consistent.

Similarly to the case of concatenated spectrograms, in superposed spectrograms, the degree of similarity between the hot zones of subtype A1 and those corresponding to the non-recombinant category was outstanding. It exceeded even what is observable in concatenated spectrograms.

The only appreciable difference with the total non-recombinant hot zones was a certain loss of power in the secondary hot zones located in the high-frequency range.

Therefore, the hot zones of this subtype in both types of spectrograms were absolutely consistent with those corresponding to the non-recombinant category.

The A1 subtype is reliably considered by the expert community as a pure subtype [2476], meaning it shows no virological evidence of having recombined with other subtypes. Hence, these levels of accuracy and similarities between the total hot zones were aligned with this fact. The results shown by the CNN denoted a clearly non-recombinant character in this subtype.

Regarding subtype A6, the results in concatenated spectrograms were not as clear as in the case of A1. Although the profile of the hot zones corresponding to A6 aligned with the total non-recombinant hot zones, the degree of divergence between both images was greater.

Comparing these hot zones with the total in the non-recombinant category, there was a loss of significance in the low-frequency range, mainly in the spectrogram corresponding to C. All zones located above the low-frequency range also weakened.

However, the error rate in this subtype was similar to the case of A1. The CNN misclassified six sequences out of the 186 comprising this subtype, resulting in an error rate of only 3.23%. Despite these slight variations, the pattern visible in its hot zones correlated to a high degree with the non-recombinant category.

Interestingly, in the case of superposed spectrograms, these slight variations were not so evident, and the degree of similarity between both images was very high.

The main difference between the total hot zones of A6 and those corresponding to the non-recombinant category was that in the high-frequency range, the main hot zone of subtype A6 was broader, extending beyond *gag* to approximately the first quarter of *pol*. The very secondary hot zones in the high-frequency range around *pol* had practically disappeared.

Despite the slight divergences detectable in the case of concatenated spectrograms, it is evident that the hot zones corresponding to subtype A6 fitted perfectly within the non-recombinant category. Therefore, as in the previous case, the behavior of the CNN was consistent.

In the scientific literature, subtype A6 continues to be considered as non-recombinant, although multiple studies regard it as a subvariant of subtype A or A1 [2477, 2478]. It is interesting to delve into the relationship between these divergences in the hot zones in concatenated spectrograms and this phylogenetic peculiarity detected in subtype A6.

Is it possible that this evolutionary trait is being reflected in the behavior of the CNN itself?

Why is it more visualizable in concatenated spectrograms, not so much in superposed ones?

These facts require further research.

Regarding subtype B, this subtype is the most abundant among non-recombinants, by far. It consists of 7,730 sequences, followed by subtype C, with 2,105 sequences. The error rate in subtype B was minimal, only 0.17%. That is, the CNN only failed in 13 sequences out of the 7,730 classified as B. Being the subtype with the highest specific weight among non-recombinants and with such a high success rate, it is logical that its hot zones were very similar to the hot zones of its category. Indeed, the degree of homology was very high, both in its main zones and in the secondary ones.

Comparing their total hot zones with those corresponding to the non-recombinant category in concatenated spectrograms, I observed greater homogeneity in the main hot zone in the low-frequency range and greater delineation in the spectrograms of G and C. The hot zone at $f = 1/3$ in T gained importance compared to its equivalent in the non-recombinant category, as well as the rest of the secondary zones.

In the case of superposed spectrograms, the hot zones were almost identical to those corresponding to the total of the non-recombinant category. No evident difference was observed.

The subtype B is the most globally spread [2479] and, therefore, one of the most epidemiologically investigated [2443]. Consequently, its consideration as a pure subtype by the expert scientific community is well established. The high degree of convergence between the hot zones of this subtype and the total non-recombinant hot zones was not only consistent with the proper functioning of the CNN and

with the fact that it carries the most weight due to its abundance, but also with this fact from the perspective of microbiology and virology.

Subtype C is the second most abundantly sequenced, with 2,105 complete sequences up to 2020. Its error rate was 1.76%, with only 37 errors in classification by the CNN. Consequently, the hit rate in this subtype reached 98.24%. Its error rates were slightly higher than those of subtype B, but in any case, the success rate in the functioning of the CNN in this subtype was very high.

As a result, the similarity between the total hot zones of this subtype and those corresponding to non-recombinant was also very high, as can be verified by the high similarity level in concatenated spectrograms.

In concatenated spectrograms, comparing with the total hot zones in the non-recombinant category, there was a loss of importance in the low-frequency zone in A, G, and the initial segment of C, as well as an increase in power between C and T. In the rest of the hot zones, there was a general loss of weight.

In superposed spectrograms, both images appeared to be practically identical.

Comparing the total hot zones of C with those corresponding to the non-recombinant category, the main hot zone in the high frequencies was more centered on 5'LTR, with less expansion towards *gag*. There was a very slight gain in weight in the secondary zone in the high frequency range around *vif*, while the secondary zone in the high frequency range of 3'LTR was milder.

The abundance of sequenced subtype C genomes, like subtype B, makes it one of the most studied subtypes [2468, 2480]. Its classification as non-recombinant by HIV-1 scientific community experts remains valid [2481]. Therefore, the CNN's accuracy rates and subsequent explainability patterns were consistent, even with the current expert knowledge in virology.

The next representative subtype is D, with two classification errors out of the 102 sequences that compose it. Therefore, the error rate was small, only 1.96%, which resulted in a hit rate of 98.04%.

Considering the results shown in Table 6.3, the degree of similarity in concatenated spectrograms between both hot areas is surprising. It was even higher than

that detected in other subtypes such as B or C, both having much lower error rates and a much larger specific weight on the total of non-recombinant sequences.

Subtype D, despite not having the lowest error rate or being a subtype with many sequences, presented an extremely high degree of similarity in concatenated spectrograms with respect to its category. This is an interesting phenomenon that will require further research.

In concatenated spectrograms, compared to the total hot zones in the non-recombinant category, a displacement of importance was observed in the main hot zone in the low-frequency range towards the first half of G, as well as a lighter presence in all zones identified above the low-frequency range.

Interestingly, in the case of superposed spectrograms, slight divergences between the hot zones were observed, although those of subtype D were very clearly framed within the non-recombinant category.

In superposed spectrograms, the main hot zone was located in the high-frequency range of 5'LTR and *gag*. A less important hot zone was in the high frequencies of 3'LTR and *nef*. More secondary zones were in the high frequencies in *pol* and in the low frequencies in both LTRs. Regarding the total hot zones in the non-recombinant category, there was a clear loss of power in the hot zone located in the low frequencies of 3'LTR and, to a lesser extent, in the low frequencies of 5'LTR.

This loss of importance was also observed in the secondary zones located in the high frequencies around *vpr*, *tat*, *vpu*, and approximately the first half of *env*. In D, these secondary hot zones were practically nonexistent. However, concerning the total non-recombinant hot zones, the zone located in the high frequencies of 3'LTR and *nef* gained power.

In view of the results in both types of spectrograms, it is evident that subtype D, classified as non-recombinant by the scientific community [2482], showed accuracy levels and hot zones consistent with this fact.

In the case of F1, this subtype obtained an error rate that we can classify as representative, reaching 19.51%, with 16 classification errors out of a total of

82 sequences in this subtype. It was the third highest error rate among non-recombinant subtypes and the fourth highest if we consider both recombinant and non-recombinant ones. Therefore, the CNN's accuracy rate with this subtype only reached 81.71%.

Given such an error rate and analyzing Table 6.3, when I examined the degree of homology between its hot zones and those corresponding to the non-recombinant subtypes in concatenated spectrograms, there were not many differences in the most relevant hot zones (in red). The noticeable differences focused on the secondary zones, which lost specific weight in the case of F1 vs. the non-recombinant category.

Similarly, there was a slight loss of significance of the main hot zone in the low-frequency range in G and C, while the zone between C and T experienced a slight increase in power.

In the case of superposed spectrograms, certain differences were also observed at first glance, centered around the low frequencies of 5'LTR and the high frequencies of 3'LTR. In the case of F1, these zones carried more weight than those in the non-recombinant category.

In superposed spectrograms, when compared with the total non-recombinant hot zones, the main one in the high frequencies in 5'LTR was slightly more extended towards *gag*, and the hot zone in the high frequencies around 3'LTR also gained power. This same phenomenon, even more intensified, occurred in the low-frequency range of 5'LTR. In contrast, the low frequencies of 3'LTR experienced a very slight loss of importance, as well as the secondary zone located in the high frequencies around *vif* to *env*.

These divergences between the hot zones of subtype F1 and those corresponding to non-recombinants were consistent with its high error classification rate. Despite being a representative error (over 15%), the accuracy rate was relatively high, over 80%, so certain differences in the hot areas are expected, but without losing the clear non-recombinant character in this subtype.

Currently, subtype F1 continues to be considered non-recombinant [2483], so the observed error rate of this subtype has no underlying biological explanation. It

would be interesting to delve into the reasons that may cause this high error rate in the CNN.

Subtype G had the second highest error rate, at 25.25%. This means that the CNN misclassified a total of 25 out of the 99 existing complete G sequences up to 2020. The overall hit rate applicable to this subtype was only 74.75%.

For concatenated spectrograms, there were notable differences between the hot zones of subtype G and those of the non-recombinant category, especially in the most relevant zones (in red), located in the low-frequency range.

Compared to the total hot zones of the non-recombinant category, the most notable aspect was the loss of importance in the low frequencies of G and the first half of C. The zone of low frequencies between C and T gained power. In the rest of the secondary zones, there was a general loss of weight compared to the total hot zones of the non-recombinant category.

In the superposed spectrograms, the same phenomenon of divergence occurred, although not as noticeable at first glance as in the case of concatenated spectrograms.

The first notable difference compared to the total hot zones of the non-recombinant category was the increased importance of the hot zone in the low frequencies of 3'LTR and *nef*, as well as, to a lesser extent, in the low frequencies of 5'LTR and the high frequencies of 3'LTR and *nef*. The main hot zone in the high frequencies around 5'LTR was less expanded towards *gag* than in the case of the total hot zones of the non-recombinant category.

In addition, the secondary zones that were observed in the range of high frequencies in *pol*, *vif*, and *env* of the non-recombinant category were displaced to the left and centered in the *pol* and *vif* region.

The degree of divergence between the hot zones of subtype G and the non-recombinant hot zones was consistent with the high error rate of this subtype. It is extremely interesting that this subtype is classified by the scientific community as problematic due to a possible recombinant origin that has not yet been unequivocally defined. Despite this, expert committees decided to maintain this subtype within the non-recombinant subtypes, albeit with that annotation of being problematic [120].

This fact is simply fascinating and made me think that the operation of the CNN was sometimes even more coherent than expected. Coincidentally (or not), the second highest registered error rate was precisely in a subtype classified as problematic due to a possible recombinant origin. The high interest in delving into this subtype from the perspective of CNN results and explainability led me to extend the research. The obtained results are shown in Chapter 9.

Regarding the next representative subtype, subtype O, the CNN only failed in one sequence out of 57, resulting in a small error rate of 1.75% (hit rate of 98.25%).

Surprisingly, the degree of divergence in the hot zones in concatenated spectrograms was notable compared to those corresponding to its category. This degree of divergence was even greater than in subtype G, despite having a much lower error rate.

Compared to the non-recombinant total hot zones, in the case of concatenated spectrograms, the main hot zones of subtype O were located in the low frequencies between A and G and between C and T. Secondary zones were present in the low-frequency range in the remaining areas of C and G, as well as at $f = 1/3$ in T and to a lesser extent in C at $f = 1/6$.

This high degree of divergence also occurred in superposed spectrograms. Compared to the total non-recombinant spectrograms, the main hot zone with greater power was clearly located in the low frequencies in 3'LTR, extending to part of *nef*. A secondary zone was in the high-frequency range throughout the genome, with greater power in *gag*, and *pol*. Conversely, the secondary zone located in the low frequencies of 5'LTR was quite similar to the non-recombinant reference.

The most notable difference compared to the total non-recombinant hot zones was in the high-frequency range, with a less delimited zone and more extended throughout the genome. I also highlighted the greater importance of the hot zone in the low-frequency range in 3'LTR and part of *nef*.

The O subtype was discovered in the 1990s, and was named "O" to denote "Outlier" because initially it was considered an atypical strain of HIV-1 compared to subtypes A, B, C, etc.[2484, 2485].

This inconsistency between the CNN results and the divergence in explainability might be related to this fact. This subtype is an interesting case study that will require further research to determine if there is a relationship between computational dysfunction in accuracy rate and the degree of convergence of total hot zones, and the fact that this subtype was considered atypical by the HIV-1 expert community.

In any case, at this moment I can only note that the accuracy rate of the CNN did not correspond to the fact that the hot zones of this subtype, both in concatenated and superposed spectrograms, were so different from those corresponding to the non-recombinant category.

The last non-recombinant subtype under study was subtype U. It reached the highest error rate among all subtypes, both non-recombinant and recombinant. The CNN failed to classify nine U sequences out of the 23 classified in this subtype. Therefore, the error rate reached 39.13%, while the hit rate was limited to 60.87%.

At first glance in the case of concatenated spectrograms, there were differences between the hot zones of subtype U and those corresponding to the non-recombinant category, though a greater divergence would have been expected for such a high error rate.

The major difference focused on the main hot zone in the range of low frequencies, being more centered in the transition zones between A and G, between G and C, and between C and T. The rest of the hot zones lost relevance in U compared to their equivalents in the non-recombinant category.

In the case of superposed spectrograms, the difference was evident and coherent with the high error rate. Even the total hot zone corresponding to recombinants was glimpsed in the area corresponding to high frequency 5'LTR. This phenomenon was consistent with the high error rate of this subtype.

In the case of superposed spectrograms, there were multiple and notable differences between the total hot zones of U and those of the non-recombinant category. The most significant ones occurred in the low-frequency range. The hot zones in the low frequencies of 5'LTR and the first half of *gag*, as well as the low frequencies in 3'LTR and *nef*, gained relevant importance. There was also greater

significance in the high frequencies of 3'LTR and *nef*. The secondary zone located in the low frequencies in the second half of *gag*, *pol*, and *vif* did not exist in the total hot zones of the non-recombinant category.

Subtype U is not actually a subtype but rather a category that includes all sequences containing unclassified regions, as stated in the HIV Database of Los Alamos National Laboratory in "How the HIV Database Classifies Subtypes" [41]. This fact is fascinating because it is perfectly coherent with both the error rates found in this "pseudo-subtype" and the high degree of divergence in the hot zones of both types of spectrograms compared to the total non-recombinant hot zones.

Furthermore, the fact that in superposed spectrograms the total hot zones applicable to the recombinant category were glimpsed (this did not happen in any other subtype classified as non-recombinant) led me to think about the possible existence of hidden recombination phenomena in these unclassifiable sequences, requiring further research. Therefore, I decided to delve deeper into this subtype, and the results of this additional research seeking hidden recombination phenomena were presented in Chapter 8.

In light of the results obtained in subtype U and considering the true meaning of the nomenclature of this subtype, the high error rate of the CNN and the variability in the total hot zones of U compared to the non-recombinant category were consistent with virological reality.

6.6.2 Recombinant Subtypes

Table 6.4 presents a compilation of the total hot zones for the recombinant category and for each of the representative non-recombinant subtypes, enabling easy comparison among them.

Table 6.4: Total Hot Zones per Recombinant Subtype. The columns correspond to the two types of spectrograms used, while the rows represent the non-recombinant category and each representative subtype. Scaled schematics for each type of spectrogram are included to facilitate the direct identification of the hot zones.

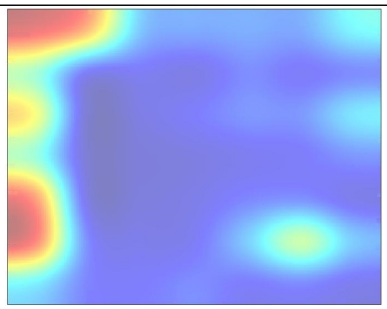
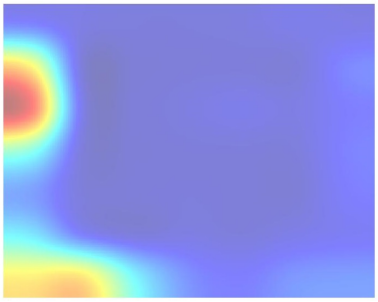
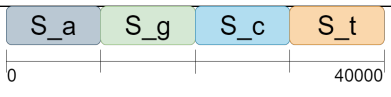
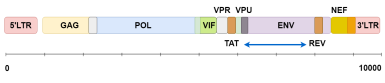
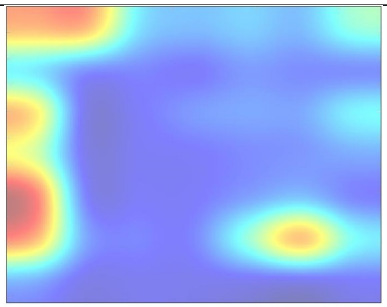
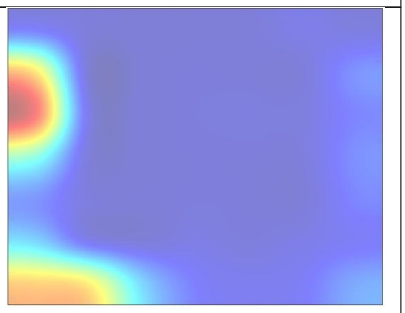
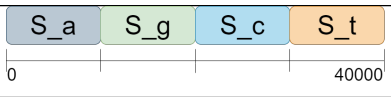
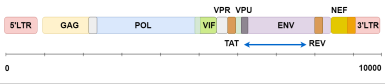
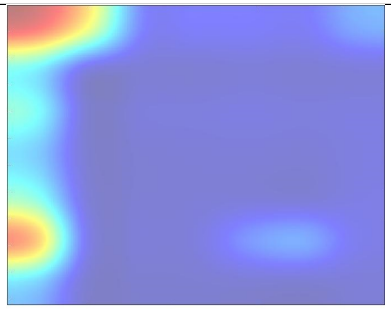
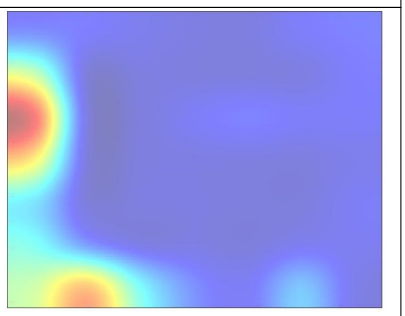
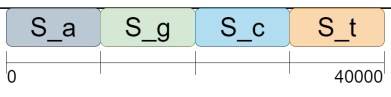
	Concatenated Spectrograms	Superposed Spectrograms
Recombinant Category		
		
Subtype CRF01_AE		
		
Subtype CRF02_AG		
		

Table 6.4 – Continuation from the previous page

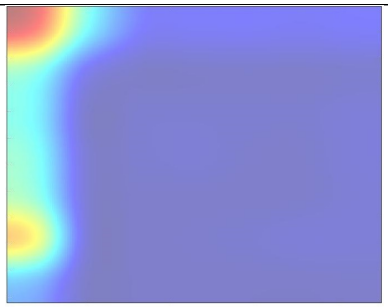
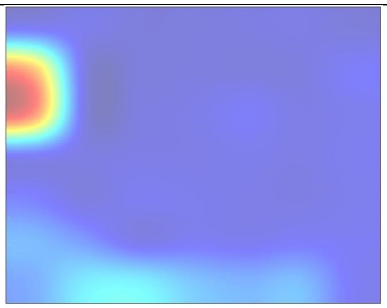
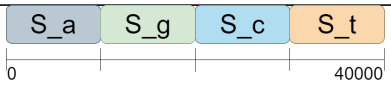
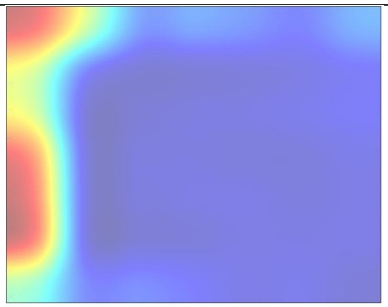
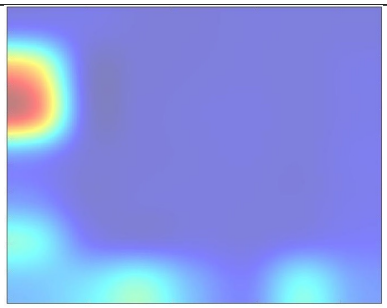
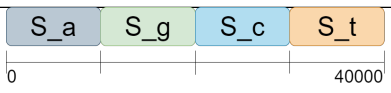
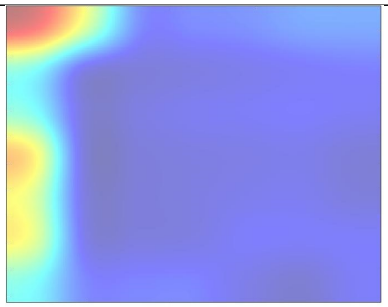
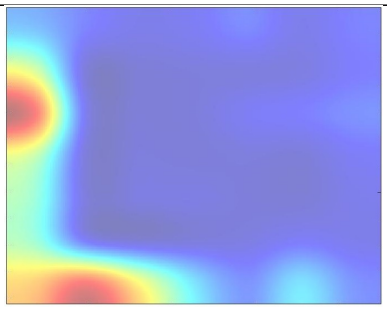
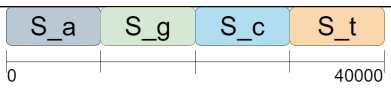
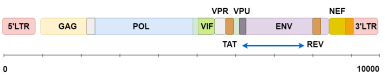
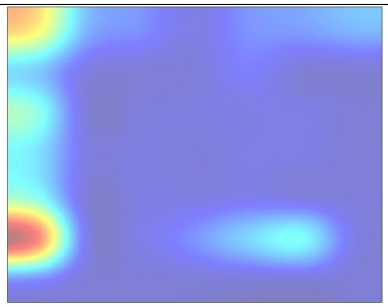
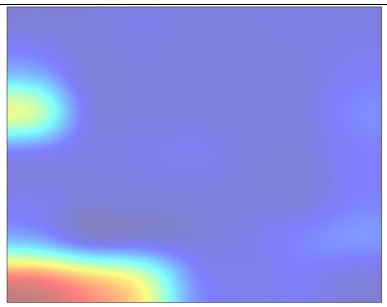
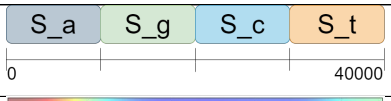
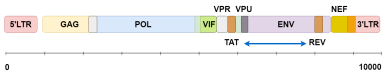
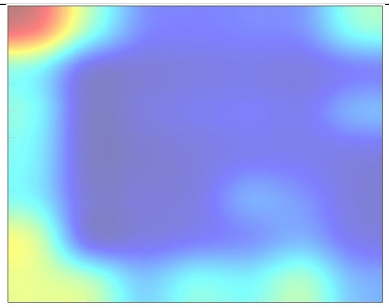
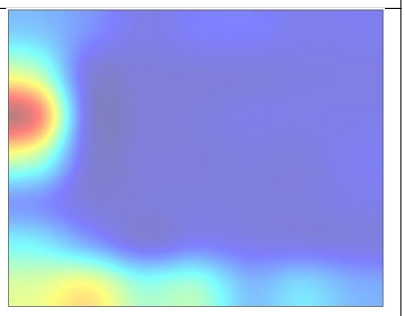
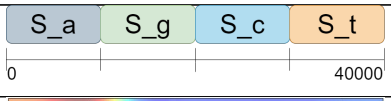
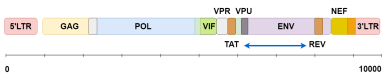
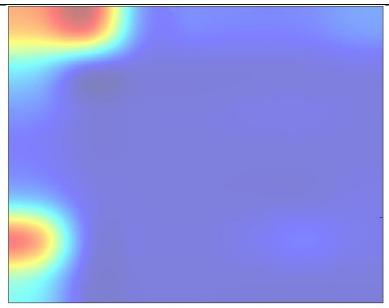
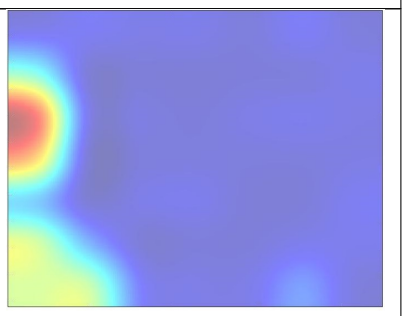
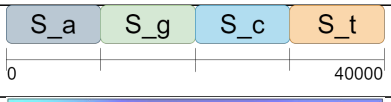
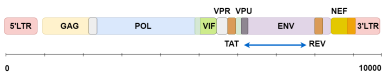
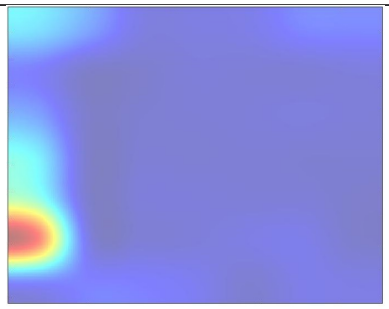
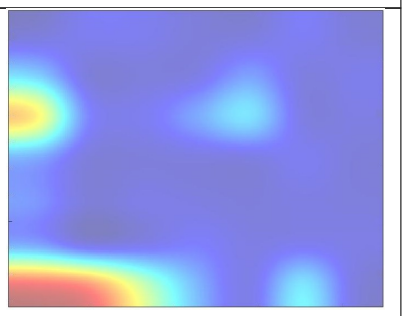
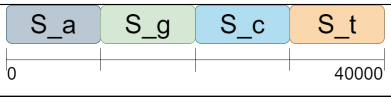
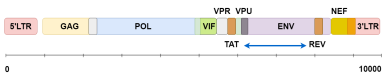
	Concatenated Spectrograms	Superposed Spectrograms
Subtype CRF07_BC		
		
Subtype CRF08_BC		
		
Subtype CRF11_cpx		
		
Subtype CRF20_BG		

Table 6.4 – Continuation from the previous page

	Concatenated Spectrograms	Superposed Spectrograms
		
Subtype CRF22_01A1		
		
Subtype CRF35_A1D		
		
Subtype CRF42_BF1		
		

Regarding recombinant subtypes, the number of representative subtypes (more than five identified sequences) is substantially higher compared to non-recombinant ones. In total, there are over 100 recombinant subtypes, many of them with a few sequences.

The first case study within the representative recombinant subtypes was Subtype CRF01_AE. It is the most abundant recombinant subtype by far, with 1,490 complete sequences identified until 2020. The CNN correctly classified all 1,490 as recombinant, resulting in an error rate of 0.00% for this subtype. Therefore, it is not listed in Table 6.2.

As seen in Table 6.4, the degree of similarity between the hot zones of CRF01_AE and the total hot zones corresponding to the recombinant category was very high. This is logical given the high specific weight that CRF01_AE has in the total of recombinant complete sequences. CRF01_AE represents 62.32% of the total recombinant sequences (1,490 out of a total of 2,391). Nevertheless, we can distinguish slight differences between both images at first glance. This may be due to the greater genomic diversity in the set of sequences and recombinant subtypes compared to non-recombinant ones.

Regarding concatenated spectrograms, the main hot zones were located in the A nucleotide region, at high frequencies, $f = 1/3$, and in the range around from $f = 1/9$ to $f = 1/6$. There was a main zone around $f = 1/9$ in the transition zone between C and T. Secondary zones were present at high frequencies and $f = 1/3$ in T. Comparing the main hot zones of CRF01_AE with the total hot zones corresponding to the recombinant category, I detected a slight loss of power in the high-frequency zone in A, as well as a slight increase in importance at $f = 1/3$, also in A. The other noticeable increase in importance with respect to the totals of its category was located around $f = 1/9$ in the transition zone between C and T. Besides CRF01_AE, only CRF20_AG showed greater intensity in that region, being a secondary zone, with a greenish shadow visible in it.

However, in the case of superposed spectrograms, the hot zones were practically identical, especially in the zone corresponding to $f = 1/3$ in 5'LTR.

The most influential hot zone was clearly located at $f = 1/3$ in 5'LTR, as well as at low frequencies in 5'LTR and *gag*, radiating with secondary power up to the first half of *pol*. Compared to the total hot zones corresponding to the entire

recombinant category, the only notable difference was a greater homogenization of power in the hot zone in the low-frequency range in 5'LTR and *gag*.

In the case of CRF01_AE, the CNN achieved a hit rate of 100% in the most numerous subtype within the recombinant subtypes. The high homology between the total hot zones of this subtype and those corresponding to the recombinant category was entirely coherent. After all, the genomes of this subtype set the predominant pattern for the recombinant category.

Regardless of the controversy surrounding the initially labeled subtype E, CRF01_AE is clearly classified by the HIV-1 expert community as recombinant [120]. Therefore, the CNN's behavior towards this subtype is clearly aligned with virological evidence.

Subtype CRF02_AG is the second most abundant after CRF01_AE, with 179 complete sequences identified until 2020. Of these, the CNN failed to classify only two cases as recombinants, resulting in an error rate of 1.12%. Thus, the hit rate achieved in this subtype was 98.88%.

In concatenated spectrograms, the hot zones of CRF02_AG were very similar to the recombinant total hot zones, with slight noticeable differences. All of this unfolded in a manner very akin to the case of subtype CRF01_AE.

Comparing the total hot zones with the total hot zones of the recombinant category, the most significant difference corresponded to the substantial loss of power at the interval between $f = 1/9$ and $f = 1/6$ in A, with the highest centrality observed at $f = 1/9$. There was also a less relevant loss of power at $f = 1/3$ in A.

The hot zone located in the range of high frequencies in A was similar to the total in recombinants, albeit with slight modifications in the contours, which were more rounded. There was a loss of importance in the remaining secondary zones of the total recombinants. In CRF02_AG, these remaining secondary zones were practically nonexistent.

Analyzing the superposed spectrograms, small differences were observed in the hot zones located in the range of low frequencies. The zone corresponding to

$f = 1/3$ in 5'LTR was practically identical in CRF02_AG and in the recombinant total hot zones.

Comparing the total hot zones in superposed spectrograms with the total recombinant ones, the main total hot zone at $f = 1/3$ in 5'LTR had a slightly higher frequency amplitude than the total of the category. Very secondary zones emerged in the 5'LTR area from the range of low frequencies to the vicinity of $f = 1/3$. The hot zone located in the low frequencies in *gag* was sharper and very slightly more powerful than the equivalent in the recombinant total hot zone. The secondary zone detectable in the second half of *env* was practically undetectable in the total hot zones of the recombinant category.

The very high homology in the hot zones of CRF02_AG compared to the total hot zones of its recombinant category was absolutely aligned with the behavior of the CNN and with the compilation of scientific evidence determining that this subtype is clearly recombinant [2486].

In the case of the next representative subtype, CRF07_BC, the CNN failed to classify in one sequence out of 43, which resulted in an error rate of 2.33% (hit rate of 97.67%).

This good performance of the network with this subtype translated into a notable similarity between the hot zones of CRF07_BC and those of the corresponding recombinant category.

However, slight differences noticeable at first glance are shown in concatenated spectrograms. These differences do not prevent understanding that this subtype has patterns clearly interpretable as recombinants.

In the case of concatenated spectrograms, the main hot zones were clearly visualized in the high frequencies in A and at $f = 1/6$. Secondary zones covered the rest of the frequency range in A.

Comparing the total hot zones of CRF07_BC with those corresponding to the recombinant category, the most notable difference was the significant loss of importance of the hot zone located at the range between $f = 1/9$ and $f = 1/6$ in A. In this recombinant subtype CRF07_BC, this hot zone was more secondary

and with a notable centrality at $f = 1/9$. The importance in $f = 1/3$ in A also diminished and faded away.

In CRF07_BC, the zone located at high frequencies in A was more framed in the spectrogram of A, with a smaller and more circular contour than the same zone in the complete recombinant category. The rest of the secondary hot zones detected in the recombinants disappeared in this subtype.

With regard of superposed spectrograms, the differences between both images were clearly appreciable in the low-frequency range, with hot zones being much more moderate in the case of CRF07_BC compared to the total hot zones corresponding to all recombinant subtypes. The hot zone at $f = 1/3$ in 5'LTR remained once again unchanged.

The most relevant difference compared to the total hot zones of the recombinant category was the significant loss of power in the low-frequency spectrum in *gag* and 5'LTR. In the case of CRF07_BC hot zones, only a secondary hot zone was observed, more oriented towards *pol* than towards *gag*. Secondary zones were observed in the low frequencies in *gag*, *pol*, in the area between *vif* and *vpu*, as well as in *env*.

Despite the slight variabilities in the hot zones in both concatenated and superposed spectrograms, their character was clearly identifiable as recombinant. This was consistent with the high accuracy of the CNN, as well as with the recombinant nature of the subtype itself, according to scientific literature [2487].

In the CRF08_BC subtype, the CNN failed on one out of 35, resulting in an error rate of 2.86% (hit rate of 97.14%). The total hot zones corresponding to CRF08_BC in concatenated spectrograms were manifestly similar to those corresponding to recombinant, once again with slight discrepancies.

In concatenated spectrograms, the main hot zones were located in the range between $f = 1/9$ to $f = 1/3$ in A, as well as in the high frequencies also of A.

In the comparison between the total hot zones of CRF08_BC and those of the recombinant category, the most notable difference was the increase in power and expansion in the range from $f = 1/6$ to $f = 1/3$ approximately in the A spectrogram. In the high-frequency range, the total hot zone in A radiated less

towards the first fragment of the G spectrogram, compared to the recombinant total hot zones. There was a general loss of power in the rest of the secondary zones, with the overall total hot zone in the recombinants at $f = 1/9$ in the transition between C and T being nonexistent in the total hot zones of CRF08_BC.

In the case of superposed spectrograms, similar to previous recombinant subtypes, the region at $f = 1/3$ in 5'LTR remained very stable, unlike the rest of the zones.

Comparing with the total hot zones corresponding to the recombinant category, the most relevant difference was that in CRF08_BC, the hot zone in the low-frequency range affecting 5'LTR and *gag* practically disappeared. Secondary hot zones emerged in the low frequencies in *pol* and in the vicinity of the second half of *env*.

The CNN's accuracy rates and the homology of the hot zones in the CRF08_BC subtype with those corresponding to the recombinant category demonstrated that the CNN's operation was correct and consistent. On the other hand, the recombinant nature of the CRF08_BC subtype is clearly studied and endorsed in the scientific literature [2488], which further supports the CNN's correct behavior.

The next representative recombinant subtype is CRF11_cpx. The error rate of this subtype was moderately high, reaching 12.00% due to three misclassifications out of the 25 sequences that compose CRF11_cpx. Therefore, the hit rate was limited to 88%. It was the second highest error rate, only behind the CRF22_01A1 subtype, which reached a total error of 19.05%.

Aligned with this, Table 6.4 (concatenated spectrograms) shows slight divergences between the hot zones corresponding to CRF11_cpx and those corresponding to the recombinant category, although the recombinant nature of CRF11_cpx's hot zones fits perfectly within the recombinant character. I did not detect a significant increase in such divergences compared to other subtypes with lower error rates.

In the case of concatenated spectrograms, the most relevant main hot zone was located in the high-frequency range of A. Lesser zones were detected in the vicinity between $f = 1/9$ and $f = 1/3$ in A as well. Comparing the total hot zones of this subtype with those corresponding to the recombinant category, the magnitude of

the power loss in the vicinity of $f = 1/9 - f = 1/6$ stood out. This zone in the case of CRF11_cpx was merely secondary. In the high-frequency range, the total hot zone of CRF11_cpx had more rounded contours and was less extended into the G spectrogram. The rest of the detectable secondary zones in the recombinant category practically disappeared in the case of CRF11_cpx.

In the case of superposed spectrograms, Table 6.4 shows the same phenomenon as in the concatenated ones; slight variations were observed, although the main hot zone at $f = 1/3$ in 5'LTR remained stable. Similar to the concatenated spectrograms, there was no significant increase in the divergence between both images, despite the error rate being notably higher than that of other recombinant subtypes.

The main hot zone was clearly located around $f = 1/3$ in 5'LTR. Another main hot zone was situated in the low-frequency range between *gag* and the first half of *pol*. There were secondary hot zones in the low-frequency range around 5'LTR, as well as in the interval from approximately $f = 1/9$ to $f = 1/3$ also in 5'LTR, and a very secondary zone located approximately in the low-frequency range around the second half of *env*. Compared to the hot zones of the recombinant category, an increase in power was easily detected in the low-frequency range affecting *gag* and *pol*. In the area corresponding to 5'LTR, a slight increase in importance was also detected. There was also a slight increase in importance in the frequency range from low frequencies to $f = 1/3$. In CRF11_cpx, the emergence of a secondary hot zone in the low-frequency range located around the second half of *env* could be glimpsed. In the total hot zones of the recombinant category, this zone was practically indistinguishable.

CRF11_cpx is the only representative recombinant subtype classified as cpx. This notation indicates that it is a subtype composed of multiple genetic recombination events involving at least three different subtypes [2489]. Despite the evident recombinant nature of this subtype, the CNN curiously achieved an error rate that, while not representative, was considerable. It is unknown whether there may be any relationship between the complexity of the recombination events in this subtype and the error rate reached by the CNN.

It is also noteworthy that the increase in the error rate did not apparently translate into an increase in the divergences between the hot zones of CRF11_cpx and the total hot zones corresponding to the recombinant category.

The recombinant subtype CRF20_BG consists of 23 complete sequences up to 2020. The CNN's hit rate for this subtype was 100%. For this reason, this subtype is not listed in Table 6.2.

In concatenated spectrograms, it is striking that, despite an error rate of 0.00%, the total hot zones corresponding to CRF20_BG showed observable divergences with respect to the total hot zones of the recombinant category, as observed in concatenated spectrograms.

Along with the recombinant subtype CRF01_AE, this subtype is the only one where a hot zone around $f = 1/9$ in the transition zone between C and T is clearly visible.

Comparing the total hot zones of CRF20_BG with those of the recombinant category, the loss of power in the total hot zone at high frequencies of A was noteworthy, with a consequent loss of radiation towards G. The hot zone at $f = 1/9$ and, to a lesser extent, at $f = 1/3$, both in A, also lost importance and amplitude. The remaining secondary zones of the recombinant category lose power in the case of CRF20_BG.

In the case of superposed spectrograms, the phenomenon repeated itself. Even the highly stable hot zone in the previous subtypes, $f = 1/3$ in 5'LTR, lost power in this subtype. Although the main hot zones of CRF20_BG were the same as those of the recombinant category, their intensities were reversed.

Comparing with the total hot zones of the recombinant category, on the one hand, the hot zone at $f = 1/3$ in 5'LTR lost importance and amplitude, being relegated to a secondary role in the case of the total hot zones of CRF20_BG. On the contrary, the hot zone gained power and extension in the low frequencies from 5'LTR to *pol*.

In the CRF20_BG subtype, there was a discrepancy between the error rate and the degree of convergence between the total hot zone of this subtype and that of the recombinant category. Despite its low error rate of only 1.12% (2 misclassifications

out of a total of 179 sequences), the total hot zones of this subtype show visually perceptible differences compared to the recombinant total hot zones.

In the specialized literature, I did not find any apparent reason that could explain these differences detectable in a subtype with a 100% accuracy rate. Some studies highlight the high degree of mutations present in this subtype, but it is a subtype with a clearly recombinant nature. A deeper study of this recombinant subtype CRF20_BG is interesting to try to elucidate if there are underlying reasons behind this certain inconsistency between the CNN's performance and the differences between the hot zones.

The CRF22_01A1 subtype had the highest error rate among recombinant subtypes and the third-highest error rate when considering recombinant and non-recombinant types together. The CNN misclassified four out of the 21 complete sequences that make up this subtype, resulting in an error rate of 19.05%. The corresponding accuracy rate was 80.95%.

Upon comparative analysis of the results shown in Table 6.4 for concatenated spectrograms, there was a clear divergence between the total hot zones corresponding to CRF22_01A1 and the total hot zones corresponding to the complete recombinant feature.

Furthermore, the secondary hot zones observed in the low-frequency range for CRF22_01A1 bore some resemblance to the hot zones corresponding to the non-recombinant category. This is consistent with the high error rate in this subtype.

The main hot zone was clearly located in the high-frequency range in A. There was a secondary zone easily visible in the interval from the low-frequency range to approximately $f = 1/6$, also in A. From $f = 1/6$ to the high-frequency range in A, there was another very slight secondary zone. Another secondary zone was in the high-frequency range in T, as well as in the low-frequency range in C and the first half of the corresponding T spectrogram. Another secondary zone was detected in the low frequencies from the second half of G to the first two-thirds of C. Comparing with the total hot zones of the recombinant category, there were multiple and striking differences. The most notable was the loss of importance at $f = 1/6$ in A. In the

case of CRF22_01A1, this zone was relegated to a secondary role and extended to the low frequencies radiating up to G. Likewise, in CRF22_01A1, the hot zone at $f = 1/3$ in A disappeared, being replaced by a strip of secondary importance along the entire frequency range in A, while the high-frequency zone remained the main one, although with a smaller extension towards the G spectrogram. In the case of CRF22_01A1, the secondary hot zone at $f = 1/6$ in the transition between C and T practically disappeared, rather moving to the low-frequency range. The secondary hot zone from G to C was practically non-existent in the case of the total recombinant hot zones. Regarding the rest of the recombinant secondary zones, there was a general loss of power in the case of CRF22_01A1.

However, in the case of superposed spectrograms, these differences were more subtle. There was a greater fuzziness in the low-frequency zone in CRF22_01A1 compared to the corresponding hot zones in the recombinant category. However, despite the high error rate, the hot zone located at $f = 1/3$ in 5'LTR was clearly discerned as a sharp main hot zone, with a high similarity to the equivalent pattern in the recombinant category.

The main hot zone was clearly established at $f = 1/3$ in 5'LTR. The others were secondary zones located in the low-frequency range, mainly affecting *gag* and, to a lesser extent, 5'LTR, *pol*, *vif*, and the second half of *env*, very secondarily. In superposed spectrograms, comparing the total hot zones of this subtype with those of the recombinant category, I detected a slight loss of power in the hot zone in *gag* and 5'LTR. The secondary zones appearing in the low-frequency range in the second half of *pol*, *vif*, and the second half of *env* were practically nonexistent in the total hot zones of the category. The main hot zone ($f = 1/3$ in 5'LTR) was very similar in both images, slightly more centered and rounded at $f = 1/3$ in the case of CRF22_01A1.

I could not find any data in the scientific literature that could apparently justify this high error rate. The genetic recombination events present in the CRF22_01A1 subtype are perfectly identified, so there is currently no doubt that it is a recombinant subtype [2490]. Some scientific studies highlight its role in the emergence of new recombinant subtypes [2491]. However, I did not find any reason

in the microbiological and virological field that could justify the high error rate committed by the CNN. Further research is needed to delve into the existence of certain factors peculiar to this subtype that may be related to the CNN's behavior.

In any case, there was a correlation between the high error rate and the difference between the hot zones corresponding to this subtype and the hot zones corresponding to the recombinant category.

Regarding the next representative recombinant subtype, CRF35_A1D, it consists of 22 complete sequences up to 2020, and the CNN correctly classified all 22 sequences as recombinants, achieving a total accuracy rate of 100%.

Similarly to the rest of the subtypes, despite a total accuracy rate in the subtype, in concatenated spectrograms clear differences could be observed between the total hot zones corresponding to CRF35_A1D and the total hot zones of the recombinant feature. However, the pattern of the hot zones of this subtype was aligned with the pattern corresponding to the recombinant feature.

In concatenated spectrograms, the main hot zone primarily affected the range of high frequencies in A, especially in its second half, radiating up to the first quarter of G. Another main hot zone was located around $f = 1/9$ in the first half of A. Comparing with the total hot zones of all recombinant subtypes, the disappearance of the hot zone at $f = 1/3$ in A stood out in the case of CRF35_A1D. The other main hot zone was more focused at $f = 1/9$, with lower power and amplitude than its equivalent considering all recombinant subtypes. In the high frequencies, there was a shift in power towards the second half of A. The other secondary zones of the recombinant category practically disappeared in CRF35_A1D.

Differences in the hot zones located at low frequencies were also observed in superposed spectrograms of CRF35_A1D and the entire recombinant category, except for the $f = 1/3$ in 5'LTR, which appeared very similar.

In the case of superposed spectrograms, there was a clear main hot zone whose epicenter seems to be slightly below $f = 1/3$ in 5'LTR. There was a secondary zone that extends from the low-frequency range to approximately $f = 1/6$, affecting 5'LTR, *gag*, and the first quarter of *pol*. When comparing the total hot zones of

CRF35_A1D with those of the recombinant category in superposed spectrograms, the first notable fact was the loss of importance in the low frequencies around *gag* and the first half of *pol*. This zone played a more secondary role in CRF35_A1D, also extending beyond $f = 1/6$ in the area from 5'LTR to the first half of *gag*. Additionally, the amplitude of the main hot zone with an epicenter at $f = 1/3$ in 5'LTR expanded, especially towards lower frequencies.

Like in the rest of the analyzed recombinant subtypes, despite the consistency between the CNN's behavior and the calculated explainability patterns with the recombinant feature, there was some variability in the hot zones of CRF35_A1D with reference to he pattern corresponding to all recombinant subtypes. This variability did not seem as aligned with the degree of classification success in this subtype, which stood at 100%. This occurred in both concatenated and superposed spectrograms. In the latter, the main hot zone at $f = 1/3$ in 5'LTR remained very stable and highly similar to the one corresponding to the recombinant category. CRF35_A1D was classified as a recombinant subtype at the microbiological and virological level, according to the literature [2492], so the CNN's behavior was consistent with this circumstance.

The last representative subtype is CRF42_BF1, consisting of 17 complete sequences until 2020. The CNN correctly classified all of them, resulting in a hit rate of 100% for this subtype. Striking differences between the hot zones corresponding to this subtype and the total hot zones in the recombinant category became apparent. While the pattern of the hot zones of CRF42_BF1 fell within the recombinant pattern, possibly this is the subtype with the greatest deviation from the recombinant total hot zones in concatenated spectrograms.

In the case of concatenated spectrograms, the main hot zone was located at $f = 1/9$ in the spectrogram corresponding to A, while secondary zones were located in the range between $f = 1/6$ and $f = 1/3$ approximately and also in the high-frequency range. When comparing the total hot zones of Subtype CRF42_BF1 with the total hot zones of the recombinant category in concatenated spectrograms, a general loss of importance in all hot zones was observed except at $f = 1/9$ in A,

though the latter was much smaller and centered at $f = 1/9$. All secondary hot zones in C and T in the recombinant category disappeared in CRF42_BF1.

In superposed spectrograms, the most notable peculiarity was the loss of power in the $f = 1/3$ zone in 5'LTR hot zones of CRF42_BF1 with reference to the total ones in the recombinant category. This is a phenomenon similar to that detected in the case of the subtype CRF20_BG. In fact, the total hot zone of CRF42_BF1 and CRF20_BG were remarkably similar.

In any case, similar to the case of concatenated spectrograms, there was a certain variance between the CNN's accuracy rate (100%) in this subtype and these noticeable differences between the hot zones of CRF42_BF1 and the total recombinant ones.

In superposed spectrograms, the main primary hot zone corresponded to the low-frequency range of 5'LTR, *gag*, and the first half of *pol*. Another lighter main zone was identified at $f = 1/3$ in 5'LTR. Secondary zones were perceived at $f = 1/3$ in the vicinity of *vpr*, *tat*, and *vpu*, as well as in the low-frequency range in the second half of *env*. The increase in importance and amplitude of the total recombinant hot zone in the low frequencies from 5'LTR to the first half of *pol* was notable when compared to all recombinant subtypes pattern. Secondary zones at $f = 1/3$ in the vicinity of *vpr*, *tat*, and *vpu*, as well as in the low-frequency range in the second half of *env*, were practically nonexistent in the recombinant category.

Therefore, as with the other representative recombinant subtypes, a direct proportionality between the CNN's accuracy rate and homology was not defined.

I could not find any data in the expert literature that clearly justifies the high degree of similarity between the hot zones of CRF20_BG and CRF42_BF1 [2493]. It will be interesting to delve deeper into the study of both subtypes in order to determine possible phylogenetic relationships that justify the high degree of similarity between their total hot zones.

The consideration of subtype CRF42_BF1 as recombinant is robust, so the behavior of the CNN and the alignment of the total hot zones of this subtype

with those corresponding to the recombinant category were coherent factors with virological evidence.

6.7 Conclusions

Firstly, the observation of a clear and significant distinction between the total hot zones of the entire dataset between the non-recombinant and recombinant categories suggests a highly discriminative representation by the CNN. This phenomenon indicates that the CNN detected intrinsic and specific characteristics of each category both in concatenated and superposed spectrograms. The repetition of the phenomenon using two different representations reinforces the possible existence of mathematical signatures embedded in the genome that determine, in this case, whether a complete HIV-1 sequence is recombinant or not.

Furthermore, there is coherence between the homology of the hot zones in the individual sequences, the total hot zones per representative subtype, and between the total hot zones per category. From any Grad-CAM result image, it is possible to identify at first glance whether it belongs to the non-recombinant category or the recombinant category, whether it is an individual sequence or a subtype.

The results achieved in this Chapter reinforce the robustness and generalization capability of the model, and therefore are aligned with what has already been discussed in Chapter 5.

Analyzing the total hot zones per subtype, I detected greater stability and a clearer correlation between the degree of homology among the hot zones per subtype and the total non-recombinant hot zones. There was also a correlation with the CNN's accuracy rate in each subtype under study.

In general terms, the differences between the hot zones per subtype and the total hot zones corresponding to their category were more easily appreciable in concatenated spectrograms than in superposed ones. A clear example of this was the case of subtype A6.

In the case of non-recombinant spectrograms, the hot zones using superposed spectrograms were more stable than those observed in non-recombinant concatenated

spectrograms, as well as in the results obtained for recombinant subtypes (both in concatenated and superposed spectrograms).

Within the category of non-recombinants, several subtypes exhibited peculiar results,specifically F1, G, O, and U.

Subtype G is considered problematic by the HIV-1 expert community due to a potential recombinant origin, while subtype U refers to all sequences that are simply unclassified or unclassifiable. The high error rates in these two cases and the divergences in the hot zones, far from indicating poor performance, clearly point to an even more coherent behavior than expected. In the case of subtype G, certain patterns corresponding to the recombinant category were discerned. This also occurs in the case of subtype U, where the hot zones were also noisy.

The behavior of the CNN regarding these two subtypes was fascinating, therefore I decided to study in-depth the presence of possible hidden genetic recombination events in sequences classified as U (Chapter 8) and to explore the recombinant origin in subtype G through the analysis of the hot zones of the sequences that compose it (Chapter 9).

Regarding subtype F1, there was no apparent reason for the abnormal error rate observed in this subtype. As for subtype O, labeled as an "Outlier" among experts, further investigation will be necessary to determine if there is a connection between its outlier character and the limited correlation between the high accuracy rate and the degree of convergence between its total hot zones and those of its total category.

In Chapter 7, I performed a preliminary analysis of subtypes F1 and O, as well as the recombinant subtype CRF22_01A1 (the only recombinant subtype with a significant classification error rate), in an initial effort to elucidate the underlying reasons for the observed peculiarities in the results for these subtypes.

The hot zones of each of the analyzed recombinant subtypes were more variable and noticeable at a first glance with respect to the total hot zones in the recombinant category. This variability was again more perceptible in concatenated spectrograms than in superposed ones. This may be due to the greater genomic variability in the set of sequences and recombinant subtypes compared to non-recombinants.

It is worth noting that there are more than 100 types of CRF recombinants, of which only nine can be considered representative (composed of more than 15 sequences). Therefore, the number of the considered non-representative recombinant subtypes is very high. This limited information, combined with the greater genetic variability inherent in CRF recombinant subtypes, logically may affect the patterns in the total hot zones per subtype. Consequently, it will also affect the degree of homology between the total hot zones per subtype with respect to the total recombinant ones, which do include the genetic variability associated with all CRFs.

Therefore, once again, the results yielded by the CNN and by explainability tools are aligned with the biological reality affecting these subtypes.

The total number of sequences in the set of analyzed representative recombinant subtypes was 1,855, representing 77.58% of the total of 2,391 complete recombinant sequences until 2020. It is possible that the remaining 22.42%, although a minority in relation to the total, may have associated genetic variability that sufficiently interferes in the correlation between the total hot zones of the recombinant category and the total hot zones per subtype, especially considering the genetic variability among the analyzed representative recombinant subtypes.

This fact requires further research, expanding this study beyond representative subtypes, including all recombinant and non-recombinant subtypes, and delving into comparisons of their hot zones.

It would also be interesting to study the similarity relationships between the total hot zones of different subtypes and establish a comparison with the phylogenetic relationships between them, in order to deepen the virological plausibility of the Grad-CAM results.

In the case of recombinants, I did not see a direct relationship between the CNN's error rate and the degree of divergence between the hot zones and the total of the recombinants. Regardless of this error rate, there were slight differences, but a clear direct proportionality was not perceived. For example, in the case of CRF11_cpx, the error rate reached 12%, but its total hot zones were not particularly different. On the contrary, in the case of CRF20_BG the error rate was 0%, and yet the

divergences between the hot zones were perfectly noticeable. Although all of them can be clearly considered as recombinants, the differences existing were mild.

In recombinants, the stability of the hot zone in superpositions at $f = 1/3$ in 5'LTR is highly striking. This main hot zone remained virtually unchanged in all representative recombinant subtypes, except in the case of CRF20_BG. Considering the genetic variability associated with recombinant subtypes, as well as the variability observed in the total hot zones per subtype, this stability at $f = 1/3$ in 5'LTR became even more remarkable.

There are subtypes such as CRF20_BG and CRF42_BF1 whose hot zones per subtype were notably similar. It would be interesting to dedicate possible future research to the search for possible relationships between these similarities in the patterns of hot zones of these subtypes and the existence of perhaps unknown phylogenetic relationships among them.

Regarding the error rates detected in the recombinant subtypes, subtype CRF22_01A1 was the only representative recombinant subtype with an error rate exceeding 15%. It would be interesting to delve into the possible microbiological reasons that could explain this error rate.

Overall, this investigation of total hot zones per subtype provided additional information not only about the degree of coherence in the behavior of the CNN but also interesting clues to deepen the virological knowledge of the different subtypes of HIV-1.

6.8 Future Studies

I consider a line of research of high interest to extrapolate the analysis conducted in this chapter to all HIV-1 subtypes, not just limiting ourselves to the representative subtypes. It is very interesting to deepen the phylogenetic knowledge among all subtypes, as well as to exhaustively analyze the relationship between phylogenetic distances and the degree of similarity between the total hot zones per subtype.

Another line of research of high interest should be the in-depth study of the mathematical pattern identified at $f = 1/3$ in 5'LTR in recombinant sequences.

It is a very powerful total hot zone and, furthermore, very stable, so I consider it of high interest for the mathematical modeling of patterns existing in that region of the HIV-1 genome. This is the most significant contribution of the research presented in this Chapter.

The variability among recombinant subtypes is to be expected due to the myriad of possible combinations. However, it would also be interesting to verify whether the hot zones highlighted by the CNN correspond to regions of greater (or possibly lesser) genetic stability.

Other lines of research associated with the results obtained include a comprehensive study of the F1, CRF22_01A1 and, especially, O subtypes, given their particularities. As the beginning of this future research, Chapter 7 is a brief analysis of the results obtained in this research.

Chapters 8 and 9 address the U and G subtypes, respectively.

I have noticed that even those who assert that everything is predestined and that we can change nothing about it still look both ways before they cross the street.

— Stephen Hawking, *he never received a Nobel because his theories could never be proven. He didn't need it anyway.*

7

A Brief Analysis of HIV-1 Subtypes F1, CRF22_01A1, and O

Contents

7.1 Introduction 295
7.2 Equipment 296
7.3 Methods 296
7.4 Subtype F1 297
7.5 Subtype CRF22_01A1 326
7.6 Subtype O 341
7.7 Conclusions 360
7.8 Future Studies 362

7.1 Introduction

In view of the results of the comparison conducted in Chapter 6, Section 6.6, it was interesting to delve into the relationship between the divergences detected among the total hot zones by category and the total hot zones of certain peculiar subtypes. Besides those subtypes whose error rates were significant (such as subtypes U, G, F1, and CRF22_01A1), the phylogenetic particularities of subtype O ("outlier") make this subtype an interesting case study.

The detailed analysis of subtype U is developed in Chapter 8 and that of subtype

G in Chapter 9. In the case of subtypes F1, O, and CRF22_01A1, the analysis was more brief and is developed in the following sections.

In these sections, based on the results of the CNN classification and the Grad-CAM hot zones of each sequence belonging to these subtypes, I conducted an analysis of the divergences between the total hot zones by category and each of these three subtypes, extracting data that may be of interest for further study of the phylogeny and evolution of these subtypes.

This Chapter consists of Sections 7.2 and 7.3, where I present the equipment used and the methodology followed. In Sections 7.4, 7.5, and 7.6 I show the results of the analysis conducted on the classification and explainability results for all sequences of subtypes F1, CRF22_01A1, and O, respectively. This Chapter concludes with Sections 7.7 (Conclusions) and 7.8 (Future Studies).

7.2 Equipment

All experiments were run in this equipment:

- Processing Unit: Intel(R) Xeon(R) Gold 5220R CPU @ 2.20 GHz.
- Installed RAM: 256 GB usable.
- Operative System: Windows 10 Education. Version: 21H1.
- GPU: NVIDIA RTX A6000. Total memory: 179.451 GB. VRAM: 48.571 GB.

7.3 Methods

The experimental methodology followed the same scheme explained in Chapter 5, Section 5.4. I generated genomic spectrograms for all complete HIV-1 sequences (concatenated and superposed). Next, I fine-tuned a pre-trained CNN using Transfer Learning with this dataset of genomic spectrograms. After confirming the proper functioning of the pre-trained CNN, I applied three different explainability tools (Grad-CAM, LIME, and Gradient Attribution) to obtain relevant hot zones in the recombinant feature. Finally, I constructed the phylogenetic tree using the HIV-1 complete consensus multiple sequence alignment from the Los Alamos National

Laboratory HIV-1 Database (up to 2020) and compared the phylogenetic closeness to the degree of similarity of the hot zones.

The complete description about this process is detailed in Chapter 5, Sections 5.5, 5.6, 5.7, 5.8, 5.9, 5.10, and 5.11.

In this case study, I focused solely on the Grad-CAM hot zones, given the visual clarity of the results obtained with this explainability tool.

I studied the explainability patterns of the sequences of each of the three subtypes under study.

I analyzed also the CNN classification and the degree of convergence of the hot zones with the total hot zones corresponding to their category. In the case of subtypes F1 and CRF22_01A1, I also analyzed the impact of the misclassified sequences on the total hot zones by subtype and their relationship with the observed divergences with respect to the total hot zones of their respective categories. In the case of subtype O, I analyzed the impact of the individual hot zones on the divergences between the total hot zones of O and the non-recombinant category, despite the low error rate of this subtype.

7.4 Subtype F1

Appendix A shows the classification results as well as the hot zones of the sequences of this subtype in concatenated and superposed spectrograms.

Subtype F1 consists of 82 complete sequences up to 2020.

Using concatenated spectrograms, the CNN misclassified eight out of the 82 sequences. This resulted in an error rate of 9.76%, below the representative error determined at 15%. Regarding the accuracy, the CNN classified 46 sequences as non-recombinant with a confidence score above 90%, which represents 56.10% of the total. Thirteen sequences obtained a confidence score between 70% and 90% (15.85% of the total), and 15 were correctly classified as non-recombinant but with a more moderate confidence score, between 50% and 70%, accounting for 18.29% of the total.

In the case of the superposed spectrograms, the CNN's failure rate reached 19.51%, making it the third non-recombinant subtype, behind subtypes U and G

respectively (Table 6.1). The CNN failed to classify a total of 16 sequences. The imbalance between the failure rates of concatenated and superposed spectrograms was noteworthy.

Regarding the successes, the CNN classified 38 sequences with a confidence score of 90% or higher, which represented the 46.34% of the total. This ratio for these confidence score levels was equivalent to that observed in concatenated spectrograms. In fact, for 29 sequences, the confidence score exceeded 90% in both concatenated and superposed spectrograms.

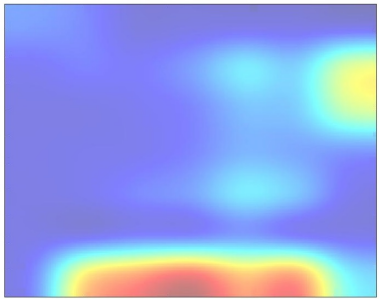
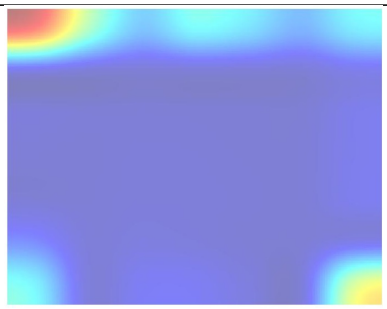
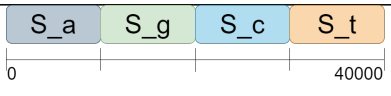
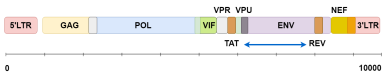
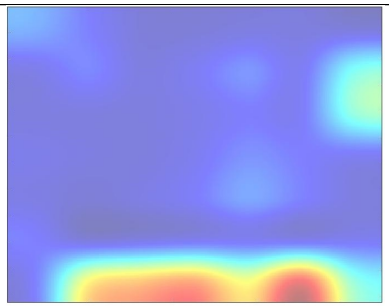
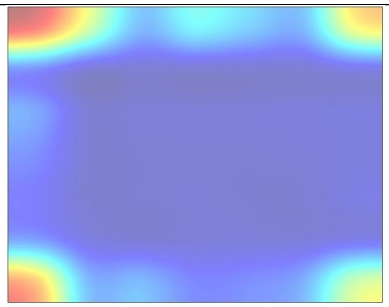
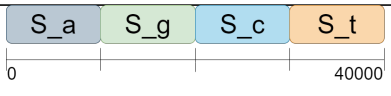
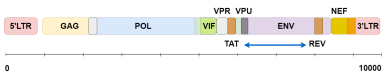
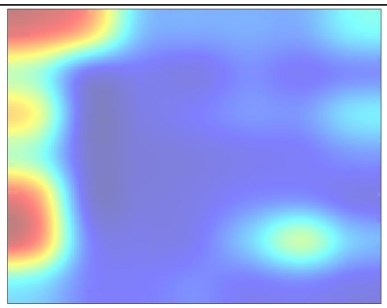
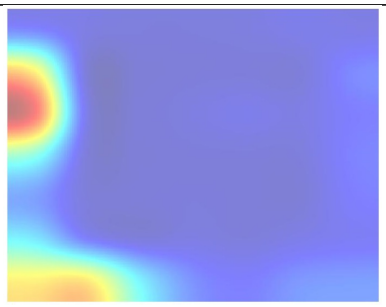
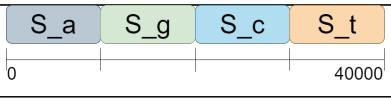
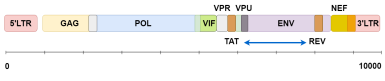
The total number of sequences classified as non-recombinant with a confidence score between 70% and 90% was 16 (19.51%), and with a confidence score between 50% and 69%, 12 sequences (14.63%).

There were no significant differences between concatenated and superposed spectrograms at none of the accuracy levels, which means that the increase in sequences incorrectly classified as recombinant was spread across all levels of confidence score.

As seen in Appendix A, in multiple instances, sequences correctly classified with notable confidence levels in one type of spectrogram were misclassified with low confidence scores in the other. This is the case for sequences AJ249238, KJ849782, MF109526, MH746263, MT417748, MT417763, MT417765, and MT417769. It would be interesting to explore the reasons for this phenomenon.

The analysis of variations in the total hot zones between the non-recombinant category and subtype F1 is shown in Table 7.1. Additionally, to facilitate the comparison of the hot zones of misclassified sequences, the total hot zones for the recombinant category are also included in the Table.

Table 7.1: Total Hot Zones per Category and Subtype F1. They include scaled representations of the schematics for each type of spectrogram to facilitate the direct identification of the hot zones.

	Concatenated Spectrograms	Superposed Spectrograms
Non-Recombinant Category		
		
Subtype F1		
		
Recombinant Category		
		

Despite the high error rate of subtype F1, I did not detect a significant impact on the main total hot zones in superposed spectrograms. The impact was more evident in secondary zones.

By analyzing the hot zones of each sequence grouped by confidence score levels,

I was able to obtain additional data to further investigate the phenomenon.

Starting with concatenated spectrograms, Table 7.2 consolidates all the correct classifications by the CNN with 90% confidence score or higher.

Table 7.2: Subtype F1 Sequences Classified by the CNN as Non-Recombinant, Confidence Score $\geq$ 90% in Concatenated Spectrograms. Sequences are alphabetically ordered.

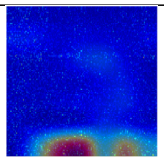
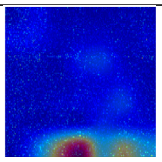
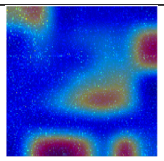
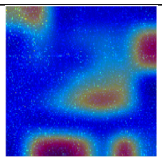
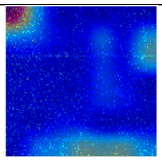
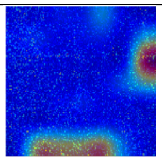
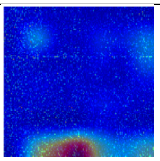
Concatenated Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
AB485656	0.97126138	
AB485657	0.97564381	
AB485658	0.99716210	
AB485659	0.99716210	
AF005494	0.98673069	
AJ249238	0.93734807	
AY173957	0.98636252	

Table 7.2 – Continuation from the previous page

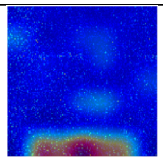
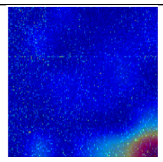
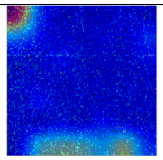
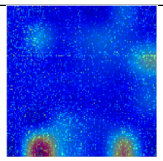
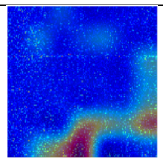
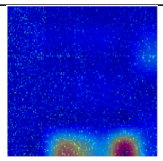
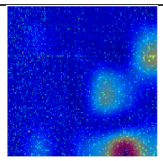
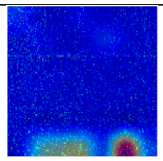
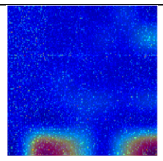
Concatenated Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
DQ189088	0.91134757	
DQ979024	0.94219297	
FJ771008	0.92909813	
FJ900268	0.95488888	
JF683771	0.98674184	
JX140671	0.99903536	
KC113009	0.99790549	
KJ849782	0.99814367	
KJ849791	0.99428540	

Table 7.2 – Continuation from the previous page

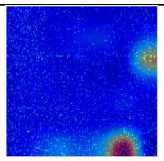
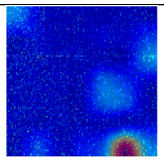
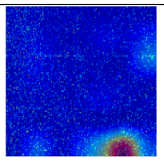
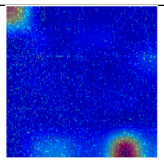
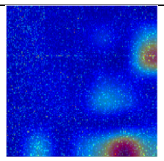
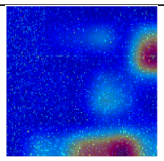
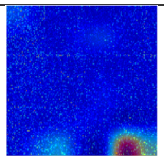
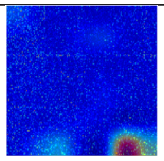
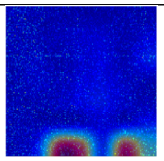
Concatenated Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
KJ883137	0.99396616	
KJ883139	0.99662578	
KJ883140	0.96537137	
KJ883141	0.97730917	
KJ883142	0.98620671	
KJ883143	0.99202955	
KJ883144	0.99928987	
KJ883145	0.96053547	
KJ883146	0.97278959	

Table 7.2 – Continuation from the previous page

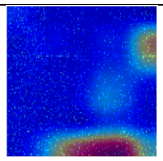
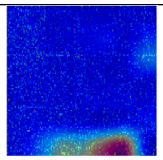
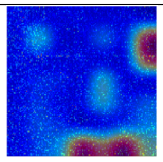
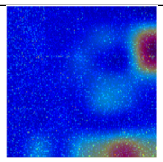
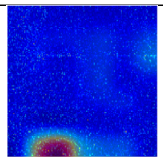
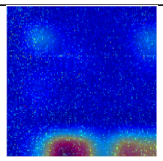
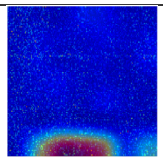
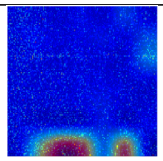
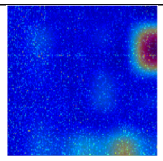
Concatenated Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
KJ883148	0.99589890	
KJ883149	0.98480254	
KJ883151	0.98516744	
KJ883152	0.96697491	
KT276257	0.98942530	
KU168274	0.94078940	
KU168275	0.97266847	
KU749396	0.98291439	
KY514084	0.99781597	

Table 7.2 – Continuation from the previous page

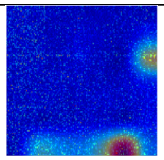
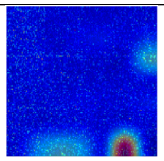
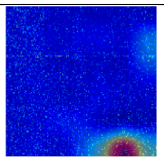
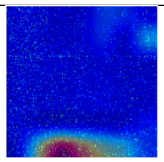
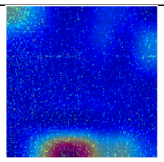
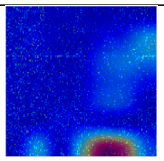
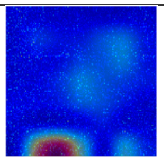
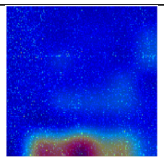
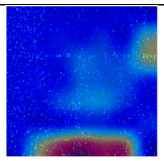
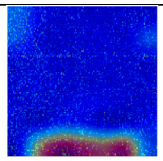
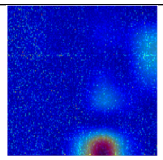
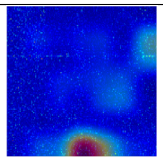
Concatenated Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
KY514085	0.99957353	
KY989955	0.99525923	
KY989957	0.99448121	
MG365764	0.99214339	
MK041565	0.98091447	
MK177827	0.95731449	
MT417748	0.99254322	
MT417762	0.99634862	
MT417763	0.98270667	

Table 7.2 – Continuation from the previous page

Concatenated Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
MT417764	0.94145989	
MT417765	0.98940897	
MT417769	0.91214973	

It was very striking that the sequences AB485658 and AB485659 were classified with identical confidence scores, and their hot zones are indistinguishable. Both sequences are intra-patient (Patient ID: BCI7). Suspecting a coincidence in the results, I compared both sequences using Python and confirmed that there is no difference between their genomes. It is necessary to review these sequences and, if appropriate, eliminate one of the two IDs.

In general, the hot zones were varied in their patterns, with no single pattern being overwhelmingly predominant. Despite the general predominance of the low-frequency range in the main hot zones, I was able to identify certain sub-patterns within subsets of sequences. For example, the pattern similar to the hot zones in sequence AB485656, the one detected in sequence AJ249238, the pattern in sequence KJ883137, or the one in sequence KJ883139.

In the case of sequences correctly classified with a 70% - 90% confidence score in concatenated spectrograms (Table 7.3), the degree of variability in sub-patterns was very similar to those in Table 7.2.

Table 7.3: Subtype F1 Sequences Classified by the CNN as Non-Recombinant, 70% ≤ Confidence Score < 90% in Concatenated Spectrograms. Sequences are alphabetically ordered.

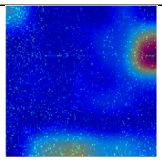
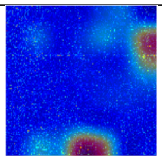
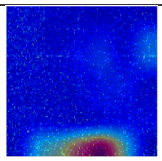
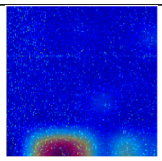
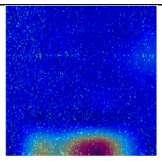
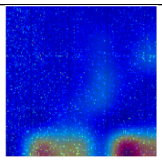
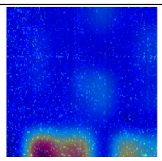
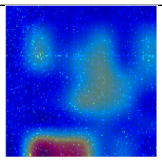
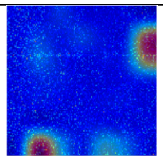
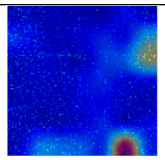
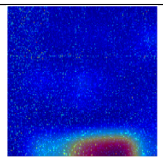
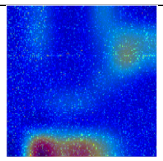
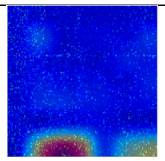
Concatenated Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
DQ979023	0.84386384	
FJ900266	0.88960779	
KJ883147	0.75896615	
KT427774	0.77205658	
KU749395	0.72366655	
KY639271	0.88291824	
KY639272	0.89017004	
KY639273	0.81883091	

Table 7.3 – Continuation from the previous page

Concatenated Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
MF109526	0.85129678	
MF109640	0.80146289	
MG365763	0.83651102	
MT222953	0.80201364	
MT417747	0.76550716	

In the case of correctly classified sequences with a more uncertain confidence score, the variability of the hot zones increased, although they remained predominantly centered in the low-frequency range (Table 7.4).

Table 7.4: Subtype F1 Sequences Classified by the CNN as Non-Recombinant, $70\% \leq$ Confidence Score $< 90\%$ in Concatenated Spectrograms. Sequences are alphabetically ordered.

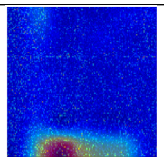
Concatenated Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
AF077336	0.62676984	

Table 7.4 – Continuation from the previous page

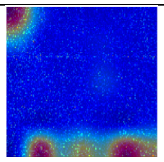
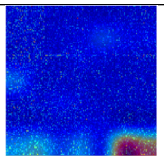
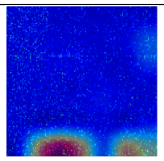
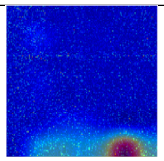
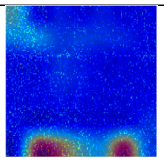
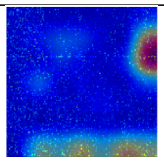
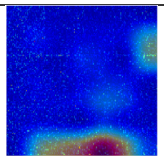
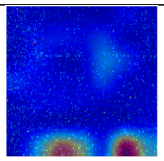
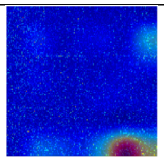
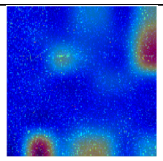
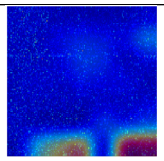
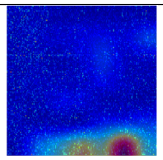
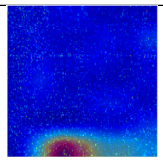
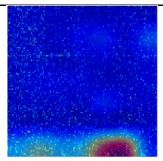
Concatenated Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
AY173958	0.66884786	
FJ670516	0.69824451	
FJ771006	0.53531951	
FJ771009	0.59062725	
FJ771010	0.60734957	
FJ900267	0.52454239	
KJ883138	0.63806075	
KJ883150	0.65480053	
KT276260	0.67440534	

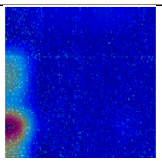
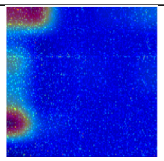
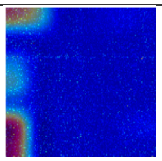
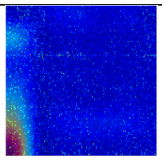
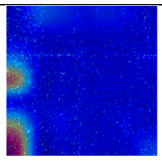
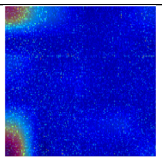
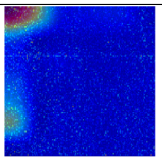
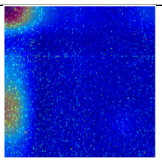
Table 7.4 – Continuation from the previous page

Concatenated Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
KU168276	0.65043950	
MG365762	0.64324737	
MG365766	0.61434537	
MG365768	0.54682845	
MN485989	0.67678934	

Indeed, the hot zones of the correctly classified sequences were clearly aligned with the non-recombinant total hot zones, with a strong predominance of the main hot zone located in the low-frequency range from the final segment of the spectrogram corresponding to A, extending to the initial segment of the spectrogram of T; followed by the hot zone located at $f = 1/3$ in the spectrogram of T.

In the case of the eight misclassified subtype F1 sequences (Table 7.5), their hot zones clearly corresponded to the total hot zones of the recombinant category.

Table 7.5: Subtype F1 Sequences Misclassified by the CNN as Recombinant in Concatenated Spectrograms (Confidence Score < 50%). Sequences are alphabetically ordered.

Concatenated Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
AF075703	0.47632891	
DQ979025	0.34399492	
EU446022	0.35458723	
KT427814	0.39004901	
MF109434	0.48077449	
MG365767	0.14730887	
MH746263	0.053940132	
MN486030	0.43544948	

Therefore, there was coherence between the hot zones of each sequence and those corresponding to each category, both in cases where the CNN succeeded

and where it failed. This indicates that the CNN identified robust and distinct patterns in each category.

There was some variability in the Grad-CAM patterns for correct classifications, which invites further investigation into the possible relationships between each identified sub-pattern and the phylogenetic relationships of the sequences that compose them. Cases like those of sequences AB485658 and AB485659 demonstrate the proportionality between phylogenetic proximity and the degree of similarity of the sequences. This shows a robust behavior of the CNN in identifying these patterns.

A curious phenomenon occurred in superposed spectrograms. The CNN misclassified 16 out of the 82 sequences belonging to this subtype. However, the total hot zones were not excessively discordant compared to the total hot zones of the recombinant category (Table 6.3).

The hot zones of the 38 sequences correctly classified by the CNN with 90% confidence score or higher are shown in Table 7.6.

Table 7.6: Subtype F1 Sequences Classified by the CNN as Non-Recombinant, Confidence Score $\geq$ 90% in Superposed Spectrograms. Sequences are alphabetically ordered.

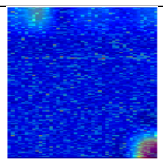
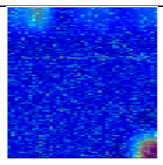
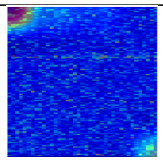
Superposed Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
AB485658	0.97130698	
AB485659	0.97130698	
AF005494	0.99218082	

Table 7.6 – Continuation from the previous page

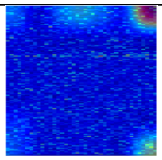
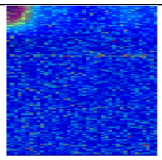
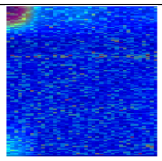
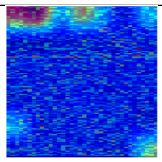
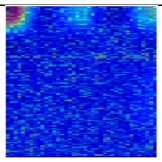
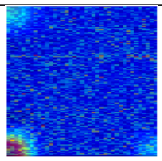
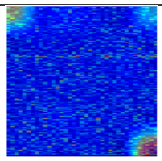
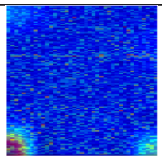
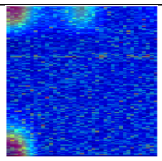
Superposed Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
AF075703	0.92978340	
AY173957	0.99943846	
AY173958	0.99181050	
FJ900268	0.93622065	
JX140671	0.99953818	
KC113009	0.97341305	
KJ849791	0.98963320	
KJ883137	0.99370015	
KJ883138	0.96232021	

Table 7.6 – Continuation from the previous page

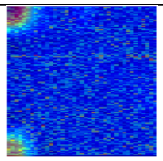
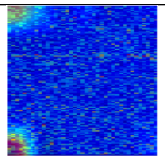
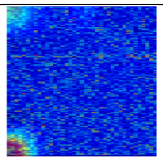
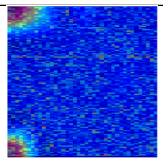
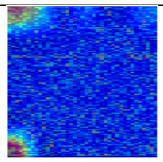
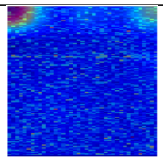
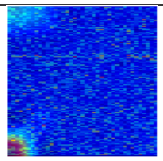
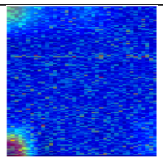
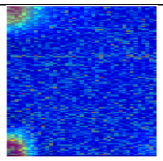
Superposed Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
KJ883139	0.99578285	
KJ883141	0.99214596	
KJ883142	0.93870032	
KJ883143	0.99570090	
KJ883144	0.98756963	
KJ883146	0.99960250	
KJ883147	0.95498258	
KJ883148	0.92667770	
KJ883149	0.99303508	

Table 7.6 – Continuation from the previous page

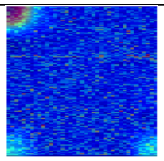
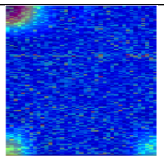
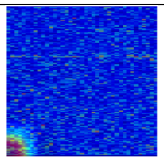
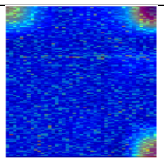
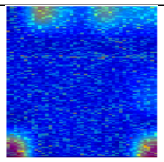
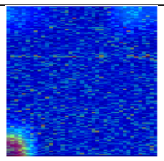
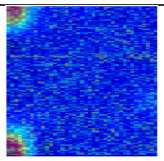
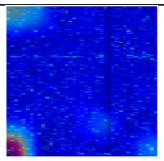
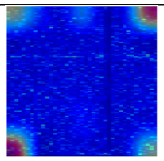
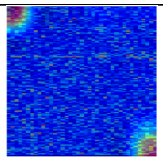
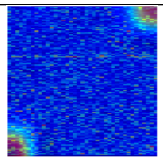
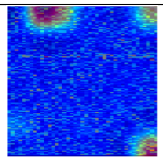
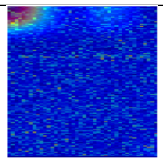
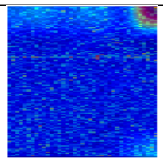
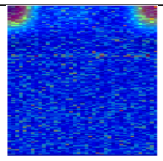
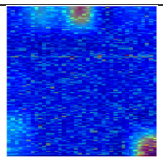
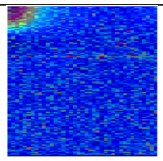
Superposed Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
KJ883151	0.99717110	
KJ883152	0.99047381	
KT276257	0.97901309	
KU168274	0.98875976	
KU168275	0.98760313	
KY514084	0.99184263	
KY514085	0.99274033	
KY639272	0.97767538	
KY639273	0.93598765	

Table 7.6 – Continuation from the previous page

Superposed Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
KY989955	0.99510574	
KY989957	0.98087388	
MG365762	0.97378516	
MG365764	0.98222381	
MH746263	0.97540295	
MK041565	0.93857902	
MN486030	0.99667943	
MT417764	0.95977890	

Just like in the case of concatenated spectrograms, sequences AB485658 and AB485659 yielded identical results, as they are indeed a single sequence with duplicated IDs.

The hot zones were consistent with those corresponding to the non-recombinant category. Although certain sub-patterns were also detectable, the hot zones of all the sequences were highly focused on the high and low frequencies of 5' and 3' LTR.

In the case of the subset of sequences correctly classified with a confidence score between 70% and 90%, a clear sub-pattern was detected, although not as abundant as in the previous case, with hot zones in the high-frequency range in the central area of the superposed spectrogram. This is exemplified by the sequence FJ771006 as a reference.

Table 7.7: Subtype F1 Sequences Classified by the CNN as Non-Recombinant, 70% $\leq$ Confidence Score $<$ 90% in Superposed Spectrograms. Sequences are alphabetically ordered.

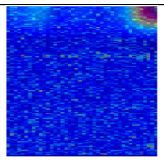
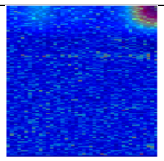
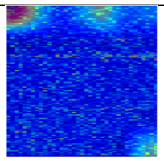
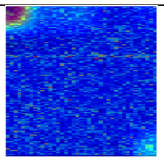
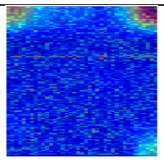
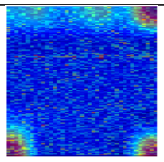
Superposed Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
AB485656	0.86337656	
AB485657	0.71860975	
DQ189088	0.80327213	
DQ979023	0.83786082	
DQ979024	0.87136734	
FJ670516	0.87515742	

Table 7.7 – Continuation from the previous page

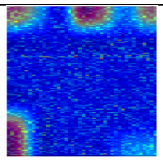
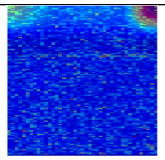
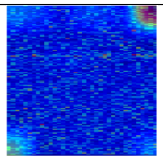
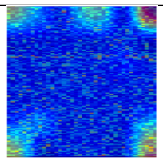
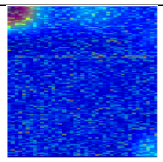
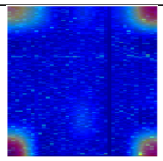
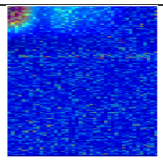
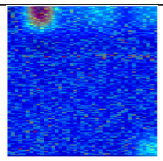
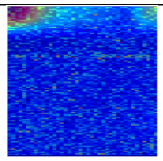
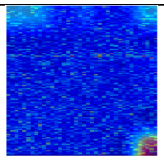
Superposed Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
FJ771006	0.84371614	
FJ771010	0.88615108	
JF683771	0.80479991	
KJ883140	0.89492697	
KU749395	0.79233700	
KY639271	0.79526740	
MG365763	0.77941793	
MG365766	0.80349880	
MT222953	0.84729683	

Table 7.7 – Continuation from the previous page

Superposed Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
MT417747	0.88939387	

This variability was repeated in those sequences correctly classified as non-recombinant with a moderate/low confidence score (Table 7.8).

Table 7.8: Subtype F1 Sequences Classified by the CNN as Non-Recombinant, $50\% \leq$ Confidence Score $< 70\%$ in Superposed Spectrograms. Sequences are alphabetically ordered.

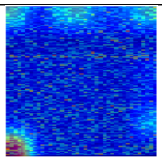
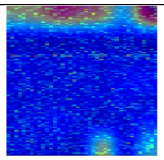
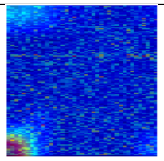
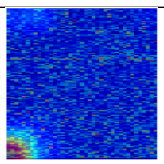
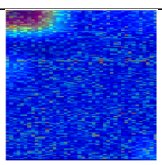
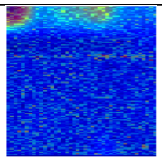
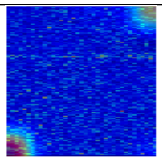
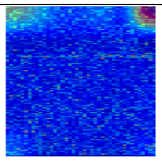
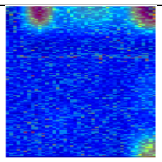
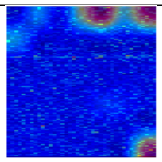
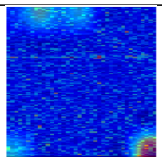
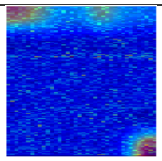
Superposed Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
AF077336	0.53128004	
EU446022	0.56270295	
KJ883145	0.56175154	
KT276260	0.69034636	
KU168276	0.66593117	

Table 7.8 – Continuation from the previous page

Superposed Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
KU749396	0.61182326	
MF109640	0.65761119	
MG365767	0.61304283	
MG365768	0.61507100	
MK177827	0.53774112	
MN485989	0.62142700	
MT417762	0.68152523	

Therefore, in the case of superposed spectrograms, I could identify a sub-pattern more focused on high and low frequencies in the LTRs, which is more prevalent in cases where the sequence is clearly non-recombinant for the CNN, with a confidence score equal to or higher than 90%. A second, more dispersed sub-pattern throughout the high-frequency range appeared more frequently in moderately correct classifications. It would be interesting to further investigate this

phenomenon and its relationship with the phylogeny of subtype F1.

The hot zones corresponding to the 16 misclassified sequences (Table 7.9) were consistent with the total hot zones of the recombinant category (Table 7.1), with a particular predominance of the hot zone at $f = 1/3$ in 5'LTR, which was clearly visible in nearly all cases, except for MT417769, where it was not present, and KJ849782, KJ883150, KT427774, and MT417748, where it appeared to be shifted toward $f = 1/6$.

Table 7.9: Subtype F1 Sequences Misclassified by the CNN as Recombinant in Superposed Spectrograms (Confidence Score < 50%). Sequences are alphabetically ordered.

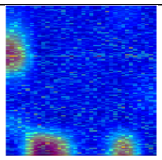
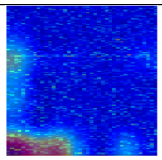
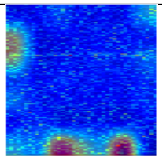
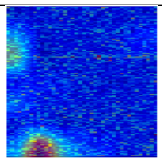
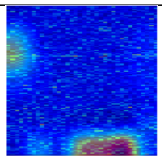
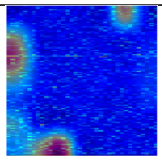
Superposed Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
AJ249238	0.21464163	
DQ979025	0.20889668	
FJ771008	0.40365908	
FJ771009	0.39664441	
FJ900266	0.38207427	
FJ900267	0.21767899	

Table 7.9 – Continuation from the previous page

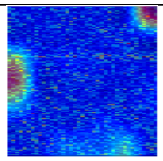
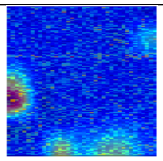
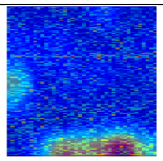
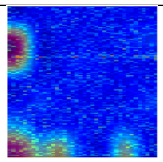
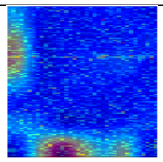
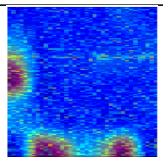
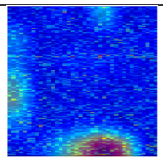
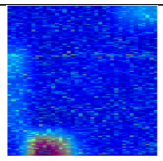
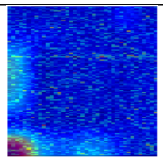
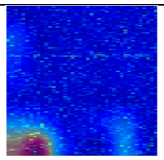
Superposed Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
KJ849782	0.34296542	
KJ883150	0.49418384	
KT427774	0.31397808	
KT427814	0.012349190	
MF109434	0.36552623	
MF109526	0.29212782	
MT417748	0.14745644	
MT417763	0.29784909	
MT417765	0.18301243	

Table 7.9 – Continuation from the previous page

Superposed Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
MT417769	0.015416335	

Therefore, in general terms, the hot zones in correct classifications were consistent with those corresponding to the non-recombinant category, while in incorrect classifications, they were consistent with the recombinant category. I found variability in the sub-patterns of the hot zones, but interestingly, in the case of superposed spectrograms correctly classified with Confidence Score $\geq 90\%$, the hot zones were more focused on the LTRs in the high and low frequencies. Perhaps the high variability in the individual hot zones of the subtype F1 sequences was related to the limited impact of the high error rate on the total hot zones of subtype F1 compared with the non-recombinant category.

In several cases, the CNN succeeded with one type of spectrogram but failed with the other. Such was the case for the sequences listed in Table 7.10.

Table 7.10: Subtype F1 Misclassified Sequences as Recombinant in One Type of Spectrogram. Sequences are alphabetically ordered.

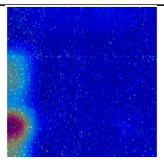
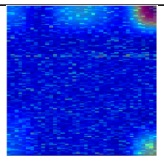
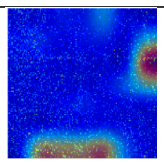
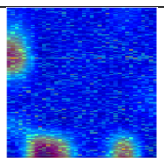
Non-Recombinant Category				
	Concatenated		**Superposed**	
Sequence ID	**Confidence Score**	**Grad-CAM**	**Confidence Score**	**Grad-CAM**
AF075703	0.47632891		0.92978340	
AJ249238	0.93734807		0.21464163	

Table 7.10 – Continuation from the previous page

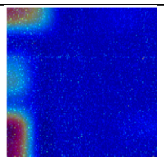
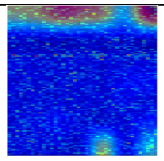
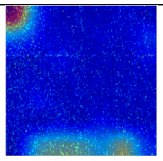
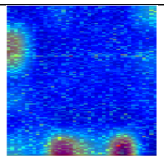
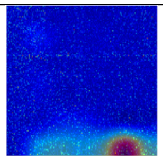
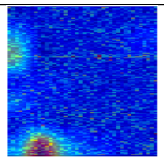
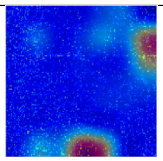
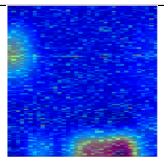
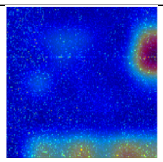
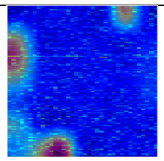
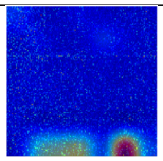
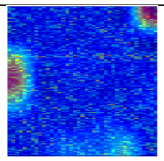
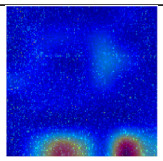
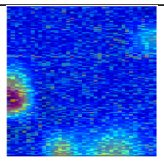
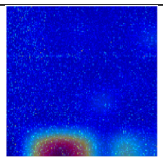
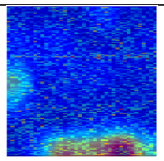
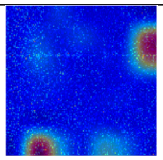
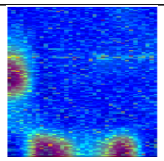
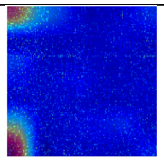
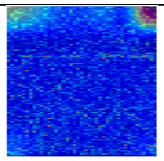
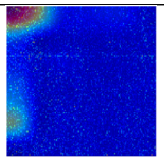
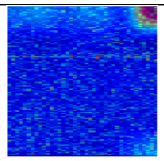
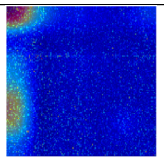
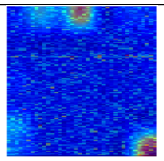
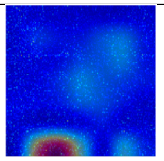
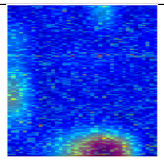
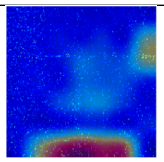
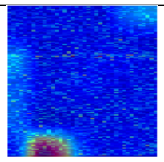
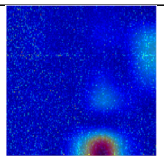
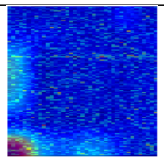
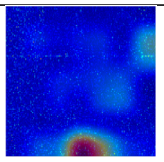
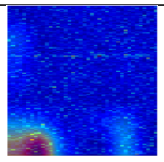
Non-Recombinant Category				
	Concatenated		**Superposed**	
Sequence ID	**Confidence Score**	**Grad-CAM**	**Confidence Score**	**Grad-CAM**
EU446022	0.35458723		0.56270295	
FJ771008	0.92909813		0.40365908	
FJ771009	0.59062725		0.39664441	
FJ900266	0.88960779		0.38207427	
FJ900267	0.52454239		0.21767899	
KJ849782	0.99814367		0.34296542	
KJ883150	0.65480053		0.49418384	
KT427774	0.77205658		0.31397808	
MF109526	0.85129678		0.29212782	

Table 7.10 – Continuation from the previous page

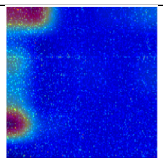
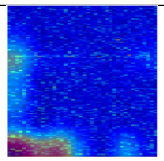
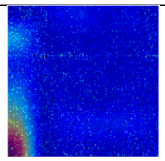
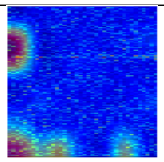
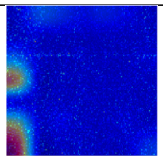
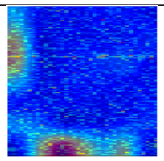
Non-Recombinant Category				
	Concatenated		**Superposed**	
Sequence ID	**Confidence Score**	**Grad-CAM**	**Confidence Score**	**Grad-CAM**
MG365767	0.14730887		0.61304283	
MH746263	0.053940132		0.97540295	
MN486030	0.43544948		0.99667943	
MT417748	0.99254322		0.14745644	
MT417763	0.98270667		0.29784909	
MT417765	0.98940897		0.18301243	
MT417769	0.91214973		0.015416335	

In some cases, such as EU446022, FJ771009, or FJ900267, the success in one type of spectrogram was achieved with a relatively low confidence score. These sequences are good candidates for studying genomic factors that may explain the classification results and the explainability, such as indicators of hidden recombination phenomena. In other cases, there was a notable discrepancy between the results in concatenated and superposed spectrograms. This was the case for KJ849782, MH746263,

MT417748, MT417763, MT417765, and MT417769. These cases are both curious and interesting and warrant further investigation into this phenomenon.

The sequences DQ979025, KT427814, and MF109434 were incorrectly classified in both concatenated and superposed spectrograms (Table 7.11). Indeed, their hot zones were consistent with the total hot zones of the recombinant category in both concatenated and superposed spectrograms.

Table 7.11: Subtype F1 Misclassified Sequences as Recombinant in Both Types of Spectrogram. Sequences are alphabetically ordered.

Non-Recombinant Category				
	Concatenated		**Superposed**	
Sequence ID	**Confidence Score**	**Grad-CAM**	**Confidence Score**	**Grad-CAM**
DQ979025	0.34399492		0.20889668	
KT427814	0.39004901		0.012349190	
MF109434	0.48077449		0.36552623	

This dual misclassification may indicate the presence of hidden genetic recombination phenomena that the CNN detected, to classify them as recombinant in both datasets. These three cases are particularly interesting to study due to the controversial nature of this dual classification as recombinant.

In light of the preliminary analysis conducted, subtype F1 exhibited certain peculiarities that make it extremely interesting for understanding aspects such as the genetic evolution of HIV-1, the impact of genetic recombination, as well as for deepening knowledge about phylogeny and the relationships among subtypes and their sequences. Specifically, future research should explore the relationships between the obtained results and phylogeny to identify factors influencing the

detected phenomena, as well as their potential implications for revising the current cataloging of sequences and subtypes.

7.5 Subtype CRF22_01A1

Appendix B presents the classification results and the hot zones for the 21 sequences of this subtype in both concatenated and superposed spectrograms.

The CRF22_01A1 subtype comprises 21 complete sequences up to 2020.

In the case of concatenated spectrograms, the CNN misclassified seven out of the 21 sequences, resulting in an error rate of 33.33%, which is significantly higher than the representative value of 15%. Regarding accuracy, the CNN correctly classified 12 sequences as recombinant with a confidence score above 90%, accounting for the 57.14% of the total. For more moderate confidence scores, the CNN correctly classified one sequence with a confidence level in the range from 70% to below 90%, and one sequence in the range from 50% to below 70%, representing error rates of 4.76%, respectively.

In the case of superposed spectrograms, the CNN's error rate was 19.05%, with four classification errors, making it the only recombinant subtype with error rates exceeding 15% (Table 6.2). In both types of spectrograms, the error rates in classification were significant. It would be interesting to further investigate this subtype to explore whether there is an underlying cause for the non-recombinant characteristics detected by the CNN. Regarding the successes, the CNN classified 12 out of 21 sequences with a confidence score of 90% or higher, which constitutes the 57.14% of the total, the same amount as in the case of concatenated spectrograms. The total number of sequences classified as recombinant with a confidence score between 70% and 90% was three (14.29%), and between 50% and 69%, it was two sequences (9.52%).

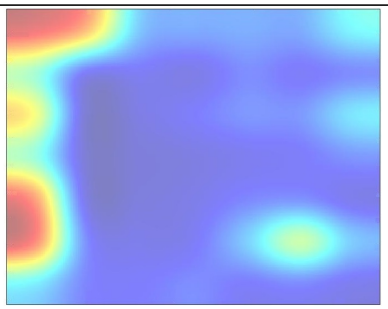
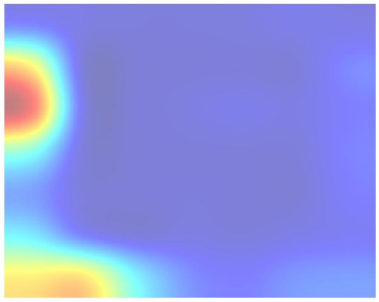
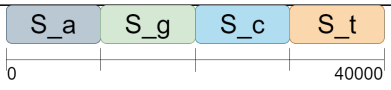
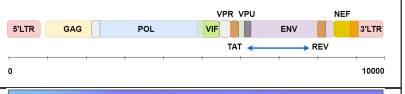
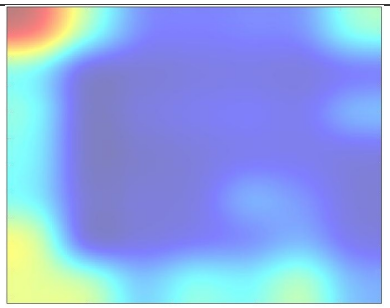
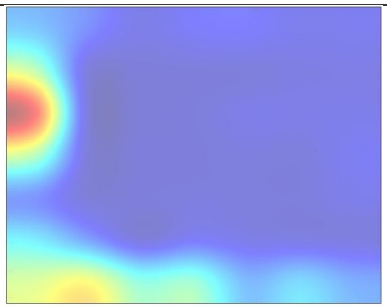
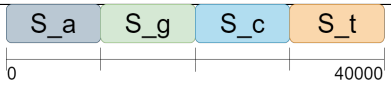
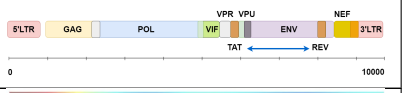
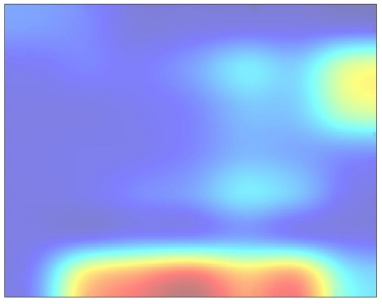
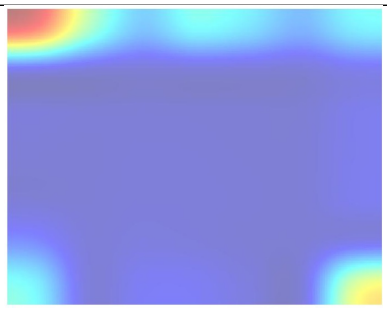
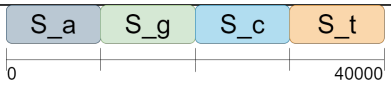
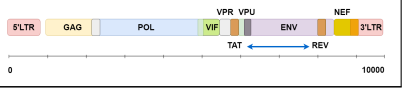
As shown in Appendix B, there was generally a correlation between the classification results in concatenated and superposed spectrograms. For sequences EU743963 and JN864051, the confidence scores were highly discrepant, with values significantly above 90% in one type of spectrogram and misclassifications in the other. In the case

of JN864049, there was also a discrepancy, but to a lesser extent (approximately 20% misclassification in concatenated spectrograms and a moderately reliable classification in superposed spectrograms, with a confidence score 63%).

In four cases, there was a misclassification in both types of spectrograms: sequences JN864051, JN864058, JN864059, and KX228817. Notably, the case of JN864058 stood out, with very low confidence scores close to 0%. According to the CNN, this sequence exhibited a very strong non-recombinant character.

The analysis of variations in the total hot zones between the recombinant category and the CRF22_01A1 subtype is presented in Table 7.12. Additionally, to facilitate the comparison of hot zones for misclassified sequences, the total hot zones for the non-recombinant category are also included.

Table 7.12: Total Hot Zones per Category and Subtype CRF22_01A1. They include scaled representations of the schematics for each type of spectrogram to facilitate the direct identification of the hot zones.

	Concatenated Spectrograms	Superposed Spectrograms
Recombinant Category		
		
Subtype CRF22_01A1		
		
Non-Recombinant Category		
		

As shown in Table 7.12, in concatenated spectrograms, there was a clear divergence between the total hot zones corresponding to CRF22_01A1 and those corresponding to the recombinant feature. This is consistent with the high error rate of this subtype in concatenated spectrograms (33.33%), as well as the fact

that the secondary zones in the low-frequency range throughout the concatenated spectrogram closely resemble the main total hot zones of the non-recombinant category. The primary hot zone was distinctly located in the high-frequency range in A. Additionally, a secondary zone was clearly visible in the interval from the low-frequency range up to approximately $f = 1/6$, also in A. Another very slight secondary zone was present from $f = 1/6$ to the high-frequency range in A. There was another secondary zone in the high-frequency range in T, as well as in the low-frequency range in C and the first half of the corresponding T spectrogram. Additionally, a secondary zone was observed in the low frequencies from the second half of G to the first two-thirds of C. When compared to the total hot zones of the recombinant category, there were numerous and notable differences. The most notable one was the reduced significance at $f = 1/6$ in A. For CRF22_01A1, this zone became secondary and extended into the low frequencies, radiating up to G. Similarly, the hot zone at $f = 1/3$ in A vanished, being replaced by a secondary strip spanning the entire frequency range in A, while the high-frequency zone remained the primary one, though with a lesser extension towards the G spectrogram. The secondary hot zone at $f = 1/6$ in the transition between C and T nearly vanished, shifting instead to the low-frequency range. The secondary hot zone from G to C was almost non-existent for the total recombinant hot zones. In general, the remaining recombinant secondary zones showed a loss of power in the case of CRF22_01A1.

In the case of superposed spectrograms, these differences were more nuanced. The low-frequency zone in CRF22_01A1 showed greater fuzziness compared to the corresponding hot zones in the recombinant category. Despite the high error rate, the hot zone at $f = 1/3$ in 5'LTR was distinctly identified as a prominent main hot zone, closely resembling the equivalent pattern in the recombinant category. The primary hot zone was clearly identified at $f = 1/3$ in 5'LTR. Other zones were secondary and located in the low-frequency range, mainly affecting *gag*, and to a lesser extent, 5'LTR, *pol*, *vif*, and the latter half of *env*, though very secondarily. In superposed spectrograms, comparing the total hot zones of this subtype with those of the recombinant category, there was a slight reduction in power in the hot

zones of *gag* and 5'LTR. The secondary zones in the low-frequency range of the second half of *pol*, *vif*, and the latter half of *env* were nearly absent in the total hot zones of the category. The primary hot zone ($f = 1/3$ in 5'LTR) appeared very similar in both images, being slightly more centered and rounded at $f = 1/3$ in the case of CRF22_01A1. In any case, there was a correlation between the high error rate and the difference between the hot zones corresponding to this subtype and the hot zones corresponding to the recombinant category.

As seen in Table 7.13, the 12 sequences that were correctly classified as recombinant with a trust level above 90% exhibited varied patterns in terms of interpretability.

Table 7.13: Subtype CRF22_01A1 Sequences Classified by the CNN as Recombinant, Confidence Score $\geq$ 90% in Concatenated Spectrograms. Sequences are alphabetically ordered.

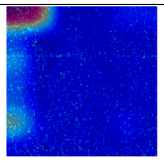
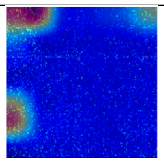
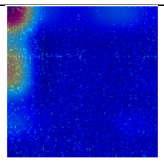
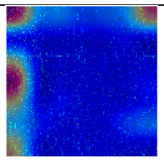
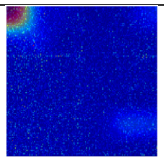
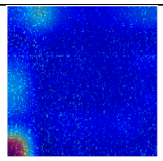
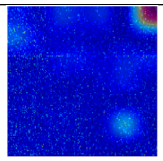
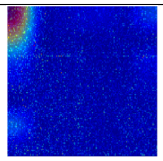
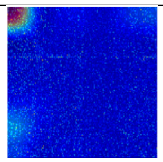
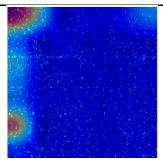
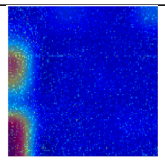
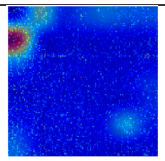
Concatenated Spectrograms Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
AY371159	0.96656334	
GQ229529	0.99928892	
JN864050	0.94469112	
KF716460	0.99614245	
KF716461	0.99901760	

Table 7.13 – Continuation from the previous page

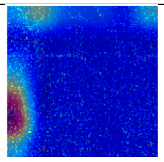
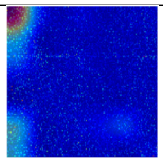
Concatenated Spectrograms Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
KF716462	0.99236321	
KF716463	0.99983811	
KP109498	0.99646217	
KP109500	0.99889731	
MN153487	0.99250811	
MT349410	0.99835867	
MT349413	0.99722987	

Despite this variety, there was a predominance of the hot zone located in the high frequencies in the A spectrogram, which was clearly evident in eight out of the 12 sequences. Except for KF716462, KF716463, MT349410, and MT349413, where this area was not identified as a hot zone, it was a very stable and prominent region in the remaining sequences. Consistent with the total hot zones of this subtype, the main hot zones of the recombinant category in the range from $f = 1/9$ to $f = 1/6$ or

at $f = 1/3$ did not have as significant a presence in this set of sequences, appearing in red or orange in six of the 12 sequences and mostly with less intensity.

Only two sequences were correctly classified by the CNN with moderate/low trust levels (Table 7.14).

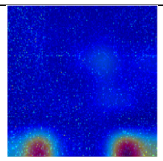
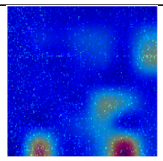
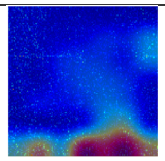
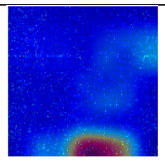
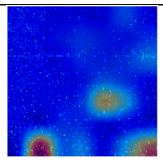
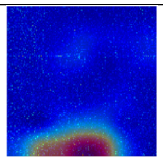
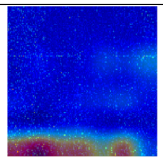
Table 7.14: Subtype CRF22_01A1 Sequences Classified by the CNN as Recombinant, 50% $\leq$ Confidence Score $<$ 90% in Concatenated Spectrograms. Sequences are alphabetically ordered.

Concatenated Spectrograms Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
AY371165	0.70579261	
KP109499	0.55124038	

The hot zones of these two sequences aligned with the total hot zones of the subtype and were similar to those obtained in sequences with high confidence scores. I did not detect any differences that would establish a direct correlation between confidence scores and hot zones. The hot zones of AY371165 exhibited a more unique pattern, while those of KP109499 are similar to those of other sequences, such as KP109500.

Table 7.15 summarizes the results for the seven sequences misclassified as non-recombinant.

Table 7.15: Subtype CRF22_01A1 Sequences Misclassified by the CNN as Non-Recombinant in Concatenated Spectrograms (Confidence Score $< 50\%$). Sequences are alphabetically ordered.

Concatenated Spectrograms Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
EU743963	0.095893860	
JN864049	0.19887774	
JN864051	0.099829063	
JN864058	0.071222097	
JN864059	0.26577246	
KX228816	0.43857804	
KX228817	0.17623627	

In concatenated spectrograms, a very high error rate was observed (33.33%). Notably, some sequences, such as JN864051 and JN864058, had very low confidence scores, close to 0, which indicates a clear classification by the CNN of these sequences as non-recombinant. The only sequence where the non-recombinant classification was less definitive was KX228816, with a confidence score of approximately 44%.

Regarding the hot zones, all of them were clearly aligned with the total hot zones of the non-recombinant category (Table 7.12). Notably, the similarity between EU743963 and JN864049, and to a lesser extent with JN864059, stood out. These patterns bore some resemblance to those of the subtype F1 sequences KJ883150 (Table 7.4), and in the case of JN864059, with KJ849791 (Table 7.2). It would be interesting to study the phylogenetic relationships between these and other sequences with similar patterns, from the same or different subtypes, to explore the existence of potential unrecognized parental relationships, as well as factors related to the non-recombinant nature that might be present in these sequences and challenge their recombinant classification.

In the case of concatenated spectrograms, it was striking that the CNN was clear both in correct and incorrect classifications, with significant confidence scores, high in correct classifications and very low in incorrect ones. Ten of the correctly classified sequences surpassed 99% confidence score. There were hardly any sequences where the CNN showed hesitation. The most doubtful cases were KP109499, classified as recombinant with 55% confidence score, and KX228816, classified as non-recombinant with 44% confidence score.

Given the results, it seems that the CNN clearly identified elements corresponding to both categories in the CRF22_01A1 sequences. The high error rate and the hot zones of the misclassified sequences justify the prominence of the low-frequency range in the total hot zones of this subtype.

In the case of superposed spectrograms, the error rate 19.05% was lower than in concatenated spectrograms, though still significant. For correct classifications, the variability in confidence scores was greater, increasing in the more moderate ranges and remaining stable (12 sequences) at high confidence score values.

The hot zones corresponding to the 12 sequences clearly classified by the CNN as recombinant (Table 7.16) were extremely stable and related to the recombinant category, particularly with the hot zone at $f = 1/3$ in the 5'LTR. This hot zone was visible in all 12 sequences, while the low-frequency range patterns were more variable. This explains the presence of secondary total hot zones across

the low-frequency range in this subtype, which is less evident in the total hot zones of the recombinant category.

Table 7.16: Subtype CRF22_01A1 Sequences Classified by the CNN as Recombinant, Confidence Score $\geq$ 90% in Superposed Spectrograms. Sequences are alphabetically ordered.

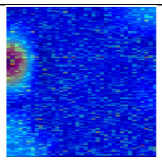
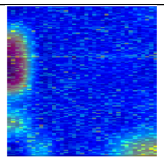
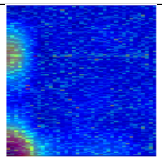
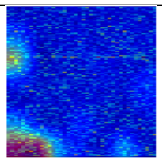
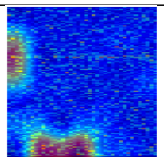
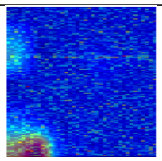
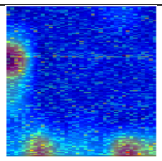
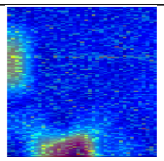
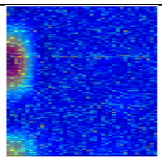
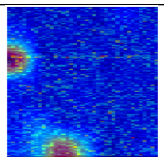
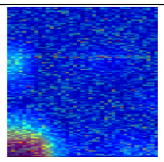
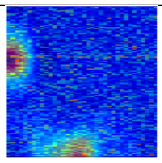
Superposed Spectrograms Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
AY371159	0.99985778	
AY371165	0.97759801	
EU743963	0.95634651	
KF716460	0.99985349	
KF716461	0.99980301	
KF716462	0.99981803	
KF716463	0.99392575	
KP109498	0.99925226	

Table 7.16 – Continuation from the previous page

Superposed Spectrograms Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
KP109500	0.99980980	
MN153487	0.98742801	
MT349410	0.99716872	
MT349413	0.93800962	

In three cases, moderate confidence scores were obtained, ranging from 70% to less than 90% (Table 7.17). Regarding the hot zones, the same phenomenon of the previous case was repeated in these sequences, as well as in the two most doubtful cases (Table 7.18).

Table 7.17: Subtype CRF22_01A1 Sequences Classified by the CNN as Recombinant, 70% ≤ Confidence Score < 90% in Superposed Spectrograms. Sequences are alphabetically ordered.

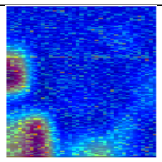
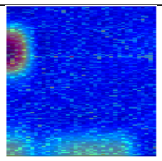
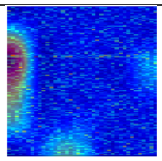
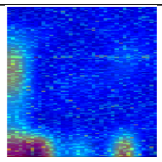
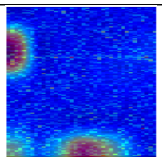
Superposed Spectrograms Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
GQ229529	0.86511111	
JN864050	0.82171077	
KP109499	0.86139542	

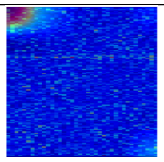
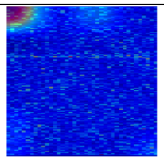
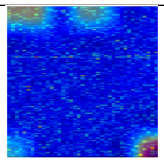
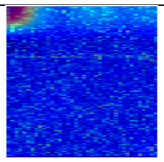
Table 7.18: Subtype CRF22_01A1 Sequences Classified by the CNN as Recombinant, 50% ≤ Confidence Score < 70% in Superposed Spectrograms. Sequences are alphabetically ordered.

Superposed Spectrograms Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
JN864049	0.63258851	
KX228816	0.67217767	

Therefore, the results obtained from the hot zones of correctly classified sequences in superposed spectrograms justified the variations in the total hot zones of the CRF22_01A1 subtype, particularly in the low-frequency range. It would be interesting to delve deeper into the patterns in this frequency range and their relationship to the recombinant nature of this subtype.

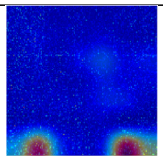
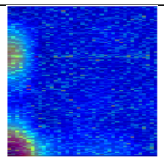
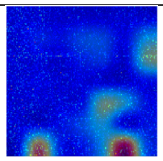
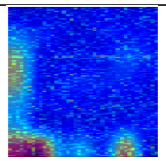
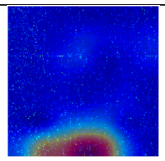
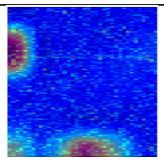
As shown in Table 7.19, the four failures observed in superposed spectrograms exhibited hot zones consistent with the non-recombinant category (Table 7.12), with a predominance of a main hot zone in high frequencies in 5'LTR. The impact of these hot zones on the total hot zones of the CRF22_01A1 subtype was mainly visible in the detection of very secondary hot zones in high frequencies from 5'LTR to *gag*. However, this impact, while present and observable, was subtle.

Table 7.19: Subtype CRF22_01A1 Sequences Misclassified by the CNN as Non-Recombinant in Superposed Spectrograms (Confidence Score $< 50\%$). Sequences are alphabetically ordered.

Superposed Spectrograms Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
JN864051	0.49274105	
JN864058	0.081556499	
JN864059	0.39441884	
KX228817	0.38101846	

In several cases, the CNN succeeded with one type of spectrogram but failed with the other. This is the case for the sequences shown in Table 7.20, including EU743963, JN864049, and KX228816.

Table 7.20: Subtype CRF22_01A1 Misclassified Sequences as Non-Recombinant in One Type of Spectrogram. Sequences are alphabetically ordered.

Non-Recombinant Category				
	Concatenated		**Superposed**	
Sequence ID	**Confidence Score**	**Grad-CAM**	**Confidence Score**	**Grad-CAM**
EU743963	0.095893860		0.95634651	
JN864049	0.19887774		0.63258851	
KX228816	0.43857804		0.67217767	

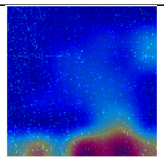
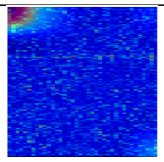
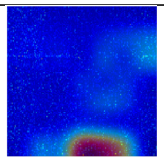
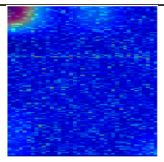
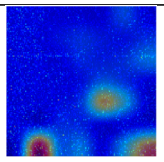
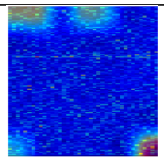
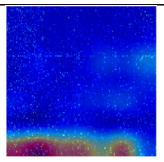
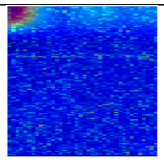
The CNN classified the sequences JN864049 and KX228816 as non-recombinant in the case of superposed spectrograms with low confidence, making it interesting to determine if the recombinant nature of these sequences is weak and the reasons behind it.

EU743963 is an interesting case study because the results in concatenated and superposed spectrograms were highly discrepant. In superposed spectrograms, it was correctly classified as non-recombinant with a high confidence score, while in concatenated spectrograms, it was incorrectly classified as recombinant with high confidence from the CNN.

I also cannot rule out the possibility that factors associated with the non-recombinant feature in these sequences were somehow diluted when performing the arithmetic sum of the FFT to generate the superposed spectrograms.

Another curious phenomenon for this subtype was the high overlap of misclassified sequences in both concatenated and superposed spectrograms. This was the case for the four sequences JN864051, JN864058, JN864059, and KX228817 (Table 7.21).

Table 7.21: Subtype CRF22_01A1 Misclassified Sequences as Non-Recombinant in Both Types of Spectrogram. Sequences are alphabetically ordered.

Non-Recombinant Category				
	Concatenated		**Superposed**	
Sequence ID	**Confidence Score**	**Grad-CAM**	**Confidence Score**	**Grad-CAM**
JN864051	0.099829063		0.49274105	
JN864058	0.071222097		0.081556499	
JN864059	0.26577246		0.39441884	
KX228817	0.17623627		0.38101846	

All sequences that were misclassified by the CNN in superposed spectrograms were also misclassified in concatenated spectrograms. These double misclassifications may indicate that these sequences are controversial from a purely microbiological perspective and may require a review of the subtype they were categorized under. Based on the CNN results, it seems clear that they have a strong non-recombinant component that sets them apart from other sequences of their subtype and category. Further research is needed.

In conclusion, the CRF22_01A1 subtype has proven fascinating in its variety and behavior during experimentation. Given the consistency of the results obtained and the impact of the hot zones on the calculation of the total hot zones for this subtype, it is evident that it is a highly interesting subtype for deepening the understanding of the phylogenetic relationships between different HIV-1 strains from the perspective of genetic recombination.

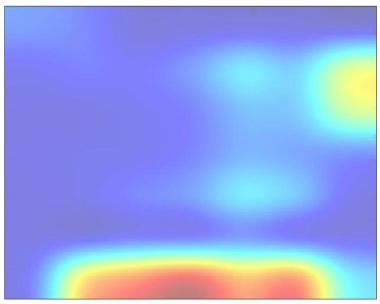
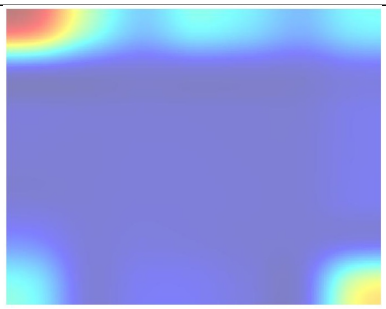
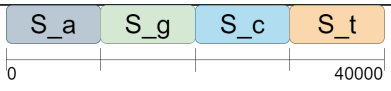
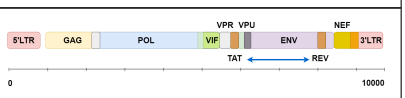
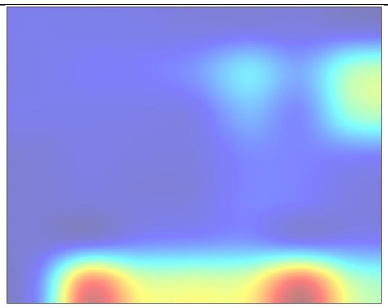
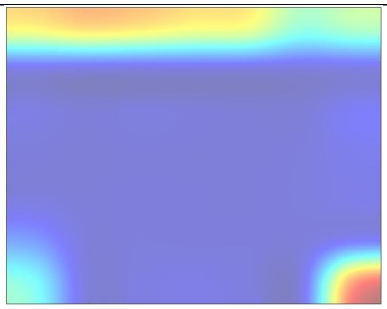
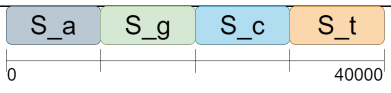
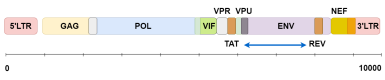
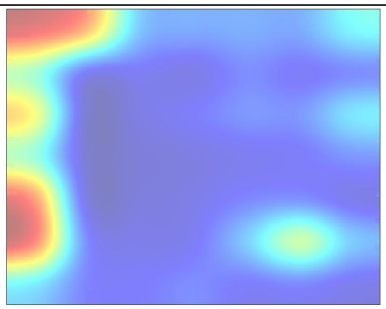
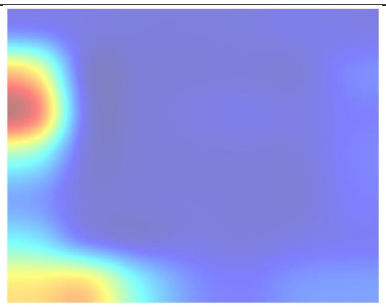
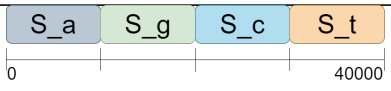
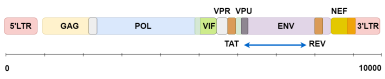
7.6 Subtype O

Appendix C shows the classification results as well as the hot zones for the 57 sequences of subtype O (up to 2020) in concatenated and superposed spectrograms.

The O subtype case differs from the previous ones. In this case, the results in both concatenated and superposed spectrograms were distinctly positive, with correct classifications and high and very high confidence scores. However, the total hot zones for this subtype were visibly divergent from the total hot zones of the non-recombinant category, particularly in the case of superposed spectrograms. These divergences did not correspond to the total hot zones of the recombinant category, necessitating a different approach compared to the previous subtypes.

Table 7.22 displays the total hot zones for the non-recombinant category, the O subtype, and the recombinant category, where we can observe that the error rate did not align with the divergence between the hot zones of the O subtype and those of the non-recombinant category.

Table 7.22: Total Hot Zones per Category and O. The columns represent the two types of spectrograms used, while the rows correspond to each category and the O subtype. Scaled schematics for each type of spectrogram are included to aid in the direct identification of the hot zones.

	Concatenated Spectrograms	Superposed Spectrograms
Non-Recombinant Category		
		
Subtype O		
		
Recombinant Category		
		

In the case of concatenated spectrograms, there were no errors in classifying the 57 sequences as non-recombinant. Moreover, for 52 of these sequences, the confidence scores were above 90%. Exceptions include AF407418 with a confidence

score of 79%, AY169806 with 80%, KF859742 with 84%, KU168293 with 88%, and KU168295 with 78%. Therefore, it is clear that the non-recombinant feature detected by the CNN was highly reliable in its classifications.

For superposed spectrograms, only one misclassification occurred with AY169803, which was classified as recombinant with a confidence score so low it could be considered nearly random (49%). This same sequence had a trust level of 99% in concatenated spectrograms. For MH705149, the classification as non-recombinant had a trust level of 85%. In the remaining 55 sequences, the CNNs confidence level was well above 90%.

Therefore, the non-recombinant nature of this subtype is clear both in terms of correct classification and in terms of high and very high confidence levels.

Surprisingly, the divergence in the hot zones of the concatenated spectrograms was significant compared to those of its category. This divergence was even more pronounced than in subtype G, despite the latter having a much lower error rate.

In the case of concatenated spectrograms, compared to the non-recombinant total hot zones, the main hot zones of subtype O were located in the low frequencies between A and G and between C and T. Secondary zones appeared in the low-frequency range in the remaining areas of C and G, as well as at $f = 1/3$ in T, and to a lesser extent, at $f = 1/6$ in C.

This significant divergence was also evident in the superposed spectrograms. Compared to the total non-recombinant spectrograms, the primary hot zone with the highest power was clearly positioned in the low frequencies of 3'LTR, extending into part of *nef*. A secondary zone was present across the genome in the high-frequency range, with increased intensity in *gag* and *pol*. In contrast, the secondary zone in the low frequencies of 5'LTR closely resembled the non-recombinant reference. The most significant difference from the total non-recombinant hot zones was observed in the high-frequency range, where the zone was less well-defined and more widespread across the genome. Additionally, the importance of the hot zone in the low-frequency range within 3'LTR and part of *nef* was notably greater.

Analyzing the Grad-CAM heat maps of the 57 sequences in concatenated spectrograms, I detected certain common sub-patterns. The main difference between the total hot zones of subtype O and the non-recombinant total hot zones was the loss of intensity in the hot zone located in the low-frequency range between the G and C spectrograms. In subtype O, this zone was more centered on the transition from A to G and from C to T. That is, in most subtype O sequences, the Grad-CAM results showed two prominent main hot zones separated in the low-frequency range. The 49 sequences corresponding to this pattern are shown in Table 7.23.

Table 7.23: Subtype O Sequences with a Grad-CAM Hot Zone in the Low Frequency Range of the Medium Area of Concatenated Spectrograms

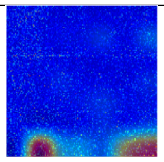
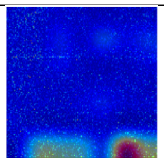
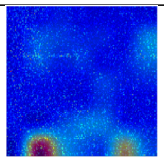
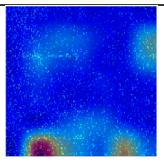
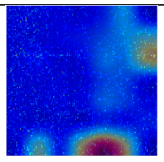
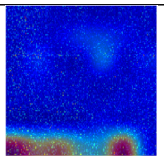
Concatenated Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
AB485666	0.99854183	
AB485667	0.99626046	
AB485668	0.94492567	
AB485669	0.96329194	
AJ302646	0.99786025	
AJ302647	0.99987161	

Table 7.23 – Continuation from the previous page

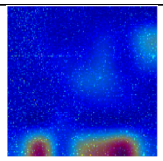
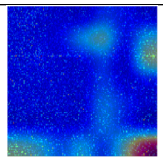
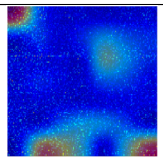
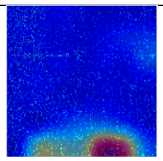
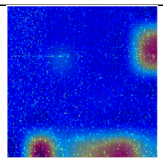
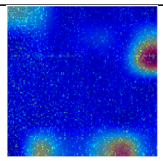
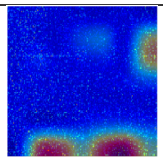
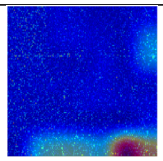
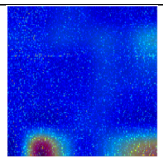
Concatenated Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
AY169802	0.99038130	
AY169803	0.98940021	
AY169804	0.98750579	
AY169805	0.96536005	
AY169808	0.99990940	
AY169810	0.99721354	
AY169811	0.95042008	
AY169812	0.98030341	
AY169813	0.96824318	

Table 7.23 – Continuation from the previous page

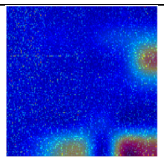
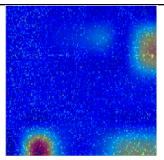
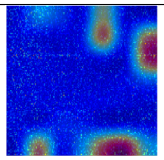
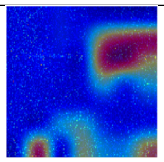
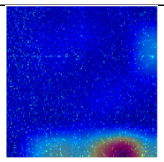
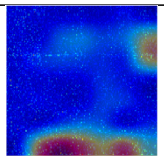
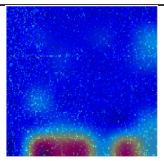
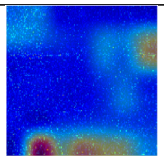
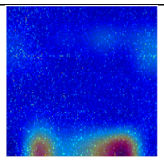
Concatenated Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
AY169816	0.99049020	
AY618998	0.97496450	
AY623602	0.99964368	
JN571034	0.99498367	
JX245014	0.91018569	
JX245015	0.98681462	
KF859742	0.84006888	
KF859744	0.94214135	
KU168281	0.91542292	

Table 7.23 – Continuation from the previous page

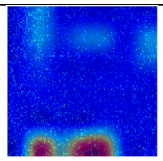
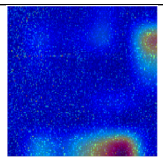
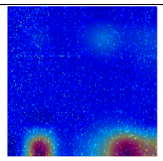
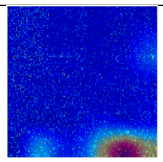
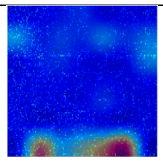
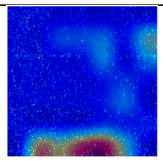
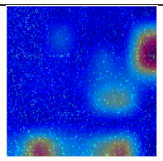
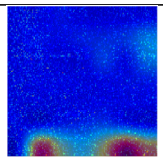
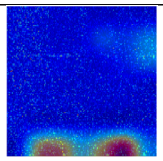
Concatenated Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
KU168282	0.99890673	
KU168283	0.99699128	
KU168284	0.91554320	
KU168285	0.98266107	
KU168286	0.95827639	
KU168289	0.97666639	
KU168290	0.99255329	
KU168291	0.97698820	
KU168292	0.96091431	

Table 7.23 – Continuation from the previous page

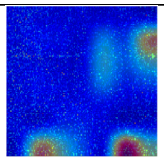
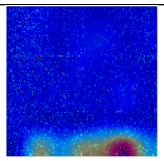
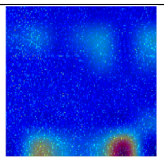
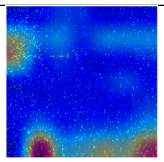
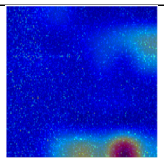
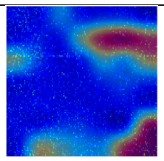
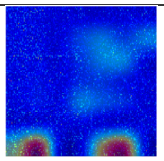
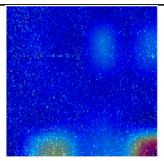
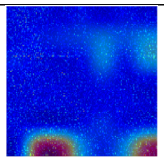
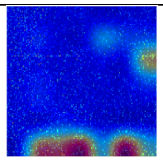
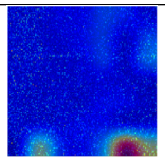
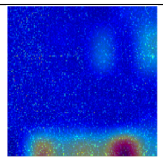
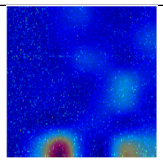
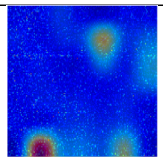
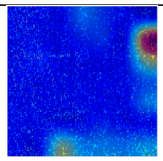
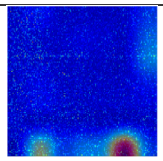
Concatenated Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
KU168293	0.88265246	
KU168294	0.99952984	
KU168295	0.78010982	
KU168296	0.95705414	
KU168297	0.99640274	
KU168298	0.96031851	
KX228804	0.98936754	
KY112585	0.99319106	
KY953205	0.99129784	

Table 7.23 – Continuation from the previous page

Concatenated Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
KY953206	0.95345253	
L20571	0.98114705	
L20587	0.98487145	
MH705147	0.98381597	
MH705148	0.95619524	
MH705149	0.99949920	
MH705150	0.98717475	

The clear predominance of this pattern, consisting of two distinctly separated focal points in the low-frequency range of the concatenated spectrograms, as well as their relatively variable positioning, was absolutely aligned with the divergences between the total hot zones of subtype O and those corresponding to the non-recombinant category.

This pattern was detected, albeit rarely, in some sequences of the CRF22_01A1

subtype, such as EU743963, classified as non-recombinant with a confidence score of 10%, and JN864049, also classified as non-recombinant with a confidence score of 20% (Appendix B). There were also similarities with several F1 sequences, such as AB485656, FJ771006, or JX140671 (Appendix A).

It would be extremely interesting to analyze the potential relationships between sequences and subtypes to determine phylogenetic reasons that could explain the similarity in hot zones.

The eight subtype O sequences that did not follow the previous pattern are shown in Table 7.24

Table 7.24: Subtype O Sequences with a Grad-CAM Hot Zone Different of the One in the Low Frequency Range of the Medium Area of Concatenated Spectrograms

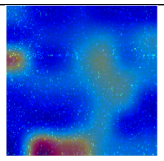
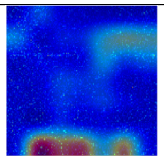
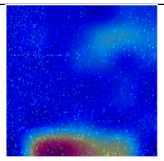
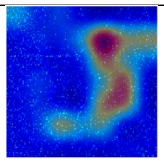
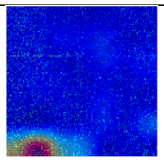
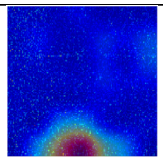
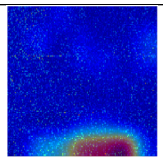
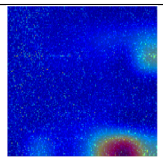
Concatenated Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
AF407418	0.78692532	
AY169806	0.80012530	
AY169807	0.98317564	
AY169809	0.99912959	
AY169814	0.99618173	

Table 7.24 – Continuation from the previous page

Concatenated Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
AY169815	0.99258476	
KF859743	0.99150473	
KU168288	0.99835664	

These eight sequences exhibited patterns more aligned with the total hot zones of the non-recombinant category, particularly AF407418, AY169806, AY169807, AY169815, and KF859743, which showed a central and homogeneous main hot zone across the low-frequency range of the concatenated spectrogram.

It would be interesting to delve into the phylogeny of this subtype to determine if there is a relationship between these patterns and the phylogenetic distance among the sequences of this subtype.

In the case of superposed spectrograms, the main difference between the total hot zones of subtype O and those of the non-recombinant category lied in the high-frequency range. For subtype O, there was greater power distributed across the high-frequency range throughout the genome, not centered on 5'LTR as seen in the non-recombinant total. Additionally, subtype O had more power in the low frequencies of 3'LTR.

Based on the superposed hot zones shown in Appendix C, I detected certain sub-patterns within the O sequences. The first sub-pattern was the presence of main hot zones in the high-frequency range that were not located in 5'LTR

(in colors ranging from yellow to red). The 32 sequences that exhibited this O pattern are shown in Table 7.25.

Table 7.25: Subtype O Sequences with a Grad-CAM Hot Zone in the High Frequency Range Not in 5'LTR of Superposed Spectrograms

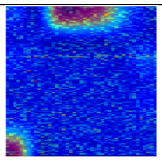
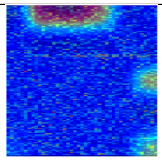
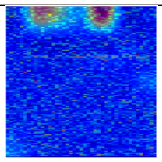
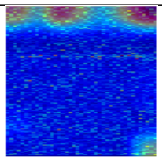
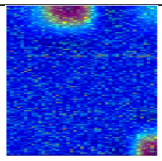
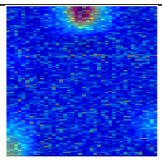
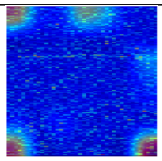
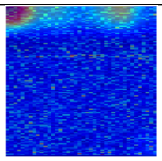
Superposed Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
AB485666	0.99952483	
AB485668	0.95757324	
AB485669	0.95260042	
AF407418	0.98849803	
AJ302646	0.99822253	
AY169805	0.99978799	
AY169806	0.99805754	
AY169807	0.999897480	

Table 7.25 – Continuation from the previous page

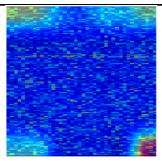
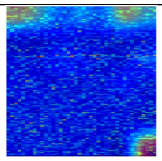
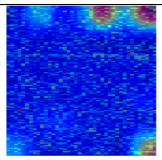
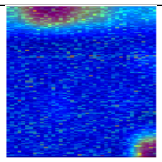
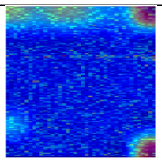
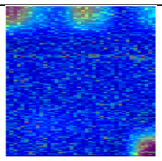
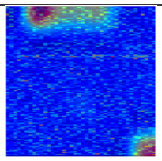
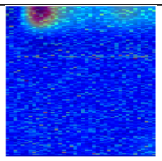
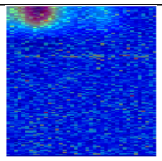
Superposed Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
AY169808	0.94435722	
AY169809	0.99481046	
AY169811	0.98550040	
AY169815	0.99446338	
JN571034	0.99845481	
JX245014	0.99405020	
KF859742	0.98930490	
KF859743	0.97085339	
KF859744	0.98728174	

Table 7.25 – Continuation from the previous page

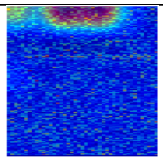
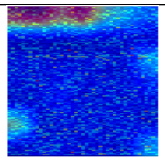
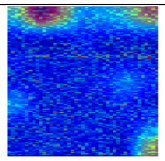
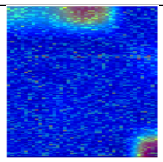
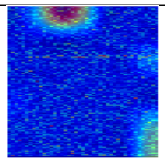
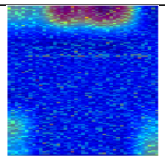
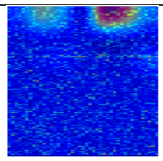
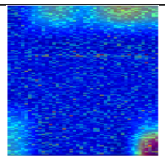
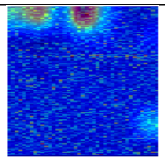
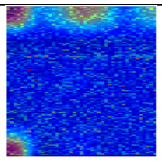
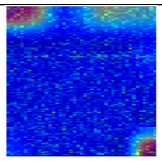
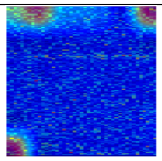
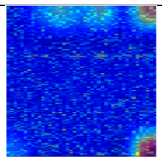
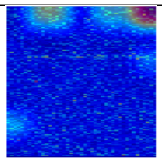
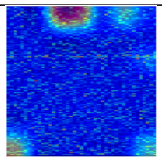
Superposed Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
KU168282	0.97326827	
KU168285	0.99845552	
KU168288	0.99952126	
KU168289	0.99635351	
KU168290	0.99797899	
KU168293	0.98878479	
KU168294	0.99887353	
KU168296	0.98576689	
KY953206	0.99682850	

Table 7.25 – Continuation from the previous page

Superposed Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
L20571	0.95797539	
L20587	0.97713041	
MH705147	0.95675308	
MH705148	0.94489014	
MH705149	0.85092968	
MH705150	0.95876294	

The predominance of hot zones with some power in the high-frequency spectrum in regions other than 5'LTR justifies the characteristic pattern of subtype O, which consisted of a hot zone with significantly greater power across the entire genome in the high-frequency spectrum. The positional variability of the main hot zones in each sequence accounted for the even distribution of power in the high frequencies of subtype O, similar to what was observed in concatenated spectrograms.

Similarly, the increase in power in the low frequencies at 3'LTR was explained by a sub-pattern found in several subtype O sequences where the only significant main hot zone was precisely located at that point (Table 7.26).

Table 7.26: Subtype O Sequences with a Grad-CAM Hot Zone in the Low Frequency Range at 3'LTR of Superposed Spectrograms

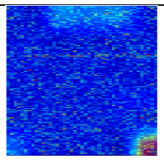
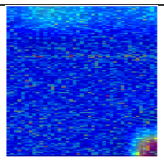
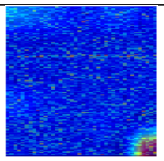
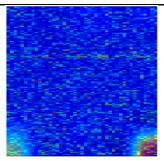
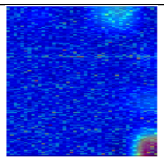
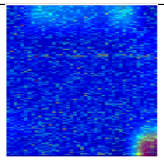
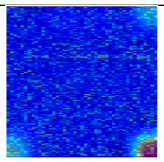
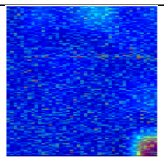
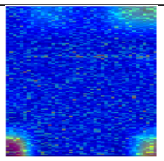
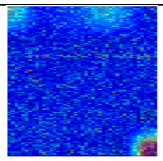
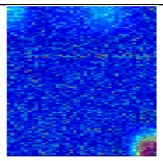
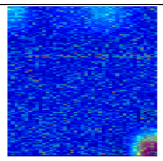
Superposed Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
AB485667	0.99785477	
AY169810	0.99311793	
AY169816	0.99820375	
AY623602	0.99855655	
KU168284	0.99761105	
KU168286	0.99852991	
KU168291	0.99969530	
KU168295	0.99981850	
KU168297	0.99945492	

Table 7.26 – Continuation from the previous page

Superposed Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
KU168298	0.99612576	
KY112585	0.99622190	
KY953205	0.99622190	

It would be interesting to delve into the phylogeny of this group of 12 sequences to research the correlation between their high degree of similarity in hot zones and their exact parental relationship, as well as their phylogenetic relationship with other subsets.

The following subset of 12 sequences of subtype O consists of those whose hot zones did not align with any of the two previous sub-patterns (Table 7.27). Some of these, such as AJ302647, AY169804, or KX228804, were more aligned with the total hot zones in superposed spectrograms for the non-recombinant category.

Table 7.27: Subtype O Sequences with a Grad-CAM Hot Zone Different from the Previous Ones of Superposed Spectrograms

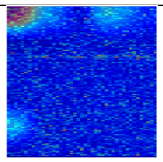
Superposed Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
AJ302647	0.99909902	

Table 7.27 – Continuation from the previous page

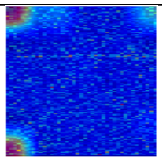
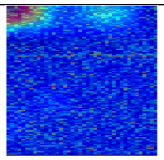
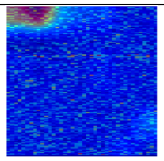
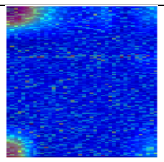
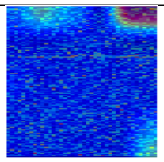
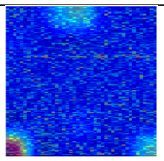
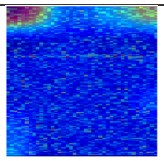
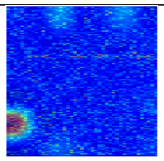
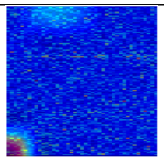
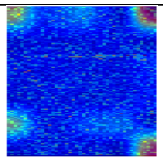
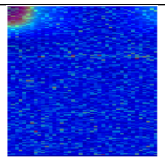
Superposed Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
AY169802	0.99958569	
AY169804	0.90259385	
AY169812	0.98134768	
AY169813	0.99572456	
AY169814	0.99442297	
AY618998	0.99994183	
JX245015	0.99542987	
KU168281	0.99951732	
KU168283	0.96098906	

Table 7.27 – Continuation from the previous page

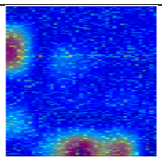
Superposed Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
KU168292	0.99970585	
KX228804	0.94765824	

It is worth noting the high degree of similarity between the hot zones of AY169804 and AJ302647, so a phylogenetic analysis of their relationship would be necessary. The pattern of KU168281 was peculiar and did not align with the others, featuring a single main hot zone clearly located at $f = 1/9$ in the 5'LTR region. This hot zone was very secondary in subtype O sequences such as AJ302647, or those In table 7.25 JN571034, KU168285, KU168293, and MH705149. It would be interesting to determine if this hot zone is related to the degree of relatedness among these sequences.

Finally, Table 7.28 shows the sequence that was incorrectly classified as recombinant by the CNN. This recombinant classification was essentially random, as the confidence score is 49%. However, the hot zones were relatively consistent with the total hot zones of the recombinant category regarding $f = 1/3$ in 5'LTR, though not as much with the hot zone located in the central genomic region in the low-frequency range. The hot zone located at $f = 1/9$ in 5'LTR was also observed, albeit secondarily.

The combination of hot zones in this sequence, with patterns clearly oriented towards the recombinant feature and patterns observed in other sequences correctly classified as non-recombinant, may be related to the CNN's low confidence in its classification as recombinant.

Table 7.28: Subtype O Misclassified Sequence in Both Types of Spectrogram

Superposed Spectrograms Non-Recombinant Category		
Sequence ID	**Confidence Score**	**Grad-CAM**
AY169803	0.49227378	

The peculiarities of this sequence make it an interesting case study for uncovering new information about the origin of subtype O and its true position in HIV-1 phylogeny.

In the case of subtype O, its peculiar Grad-CAM results were highly consistent with the distinctive profile of this subtype compared to other HIV-1 subtypes. A preliminary analysis of the explainability results allowed me to identify various subpatterns that may indicate not only phylogenetic relationships among the analyzed sequences, such as different evolutionary branches, but also possible connections with sequences from other subtypes.

Understanding this "Outlier", its origin, and its evolution could enhance and clarify our understanding of HIV-1 evolution, its genetic diversity, and the current classification of subtypes and sequences.

7.7 Conclusions

In this Chapter, I conducted a brief analysis of subtypes F1, CRF22_01A1, and O, as they presented peculiar results in the analysis of the total hot zones developed in Chapter 6.

In the cases of F1 and CRF22_01A1, I detected significant error rates that led me to delve deeper into the analysis of the hot zones of their sequences and to perform a comparison with the total hot zones of both subtypes and with those corresponding to their respective categories. In both subtypes, the differences between their total hot zones and those corresponding to their categories matched the patterns found

in the hot zones of the incorrectly classified sequences. This alignment underscores the consistent and robust performance of the CNN, with a visibly notable impact on subtype CRF22_01A1 and a less pronounced effect on subtype F1.

The high error rates of subtype F1 and the peculiarities in terms of variability and sub-patterns in Grad-CAM hot zones detected in the analysis of the hot zones of its sequences, as well as certain similarities with sequences of other subtypes, make this subtype extremely interesting for understanding aspects such as the genetic evolution of HIV-1, the impact of genetic recombination on this evolution, and for expanding knowledge about phylogeny and the relationships between subtypes and their corresponding sequences. In future research, it will be necessary to delve deeper into the connections between the obtained results and the phylogeny to identify factors that may influence the observed phenomena, as well as to evaluate their potential impact on the revision of the current classification of sequences and subtypes.

Regarding subtype CRF22_01A1, its high error rates and high confidence scores in both classifying each sequence as non-recombinant and as recombinant seem to indicate that the CNN identified clear patterns in multiple sequences, making this subtype interesting to review. Its variability and the sub-patterns detected, as well as the fact that all misclassified sequences in superposed were also misclassified in concatenated, make this subtype fascinating due to its diversity and behavior in the experiments. After reviewing the consistency of the results and the impact of the hot zones on the calculation of the total hot zones for this subtype, it is clear that this subtype is extremely interesting for deepening the understanding of the relationships between different HIV-1 strains from the perspective of genetic recombination.

The case of subtype O (referred to as "Outlier") was even more peculiar. Despite its very low error rates and high confidence scores in correctly classifying most of its sequences, notable differences between the hot zones of this subtype and those corresponding to the recombinant category were clearly visible. Upon analyzing the hot zones of the subtype O sequences, I detected certain peculiar sub-patterns that justified the divergences in the total hot zones of subtype O. Therefore, the peculiar results of Grad-CAM showed high consistency with the distinctive profile

of this subtype compared to other HIV-1 subtypes. The identified sub-patterns can indicate not only phylogenetic relationships among the analyzed sequences, such as different evolutionary branches, but also connections with sequences of other subtypes. Understanding this "Outlier," its origin, and evolution, can complete and clarify the evolution of HIV-1 itself, its genetic diversity, and the current classification of subtypes and sequences.

7.8 Future Studies

Evidently, it is necessary to expand this preliminary analysis with a thorough review of the phylogeny of these three subtypes and their relationships with other subtypes and sequences to broaden and complement the initial conclusions obtained in this brief analysis.

Constant reviews conducted by experts in the field demonstrate that much remains to be discovered regarding the phylogeny of HIV-1. Approaches from different perspectives, such as the research presented, can provide new insights and guide new information about the evolutionary history of this fascinating virus, HIV-1.

What beautifies the desert is that it hides a well of water somewhere.

— Antoine de Saint-Exupéry, *distracted flyer. It is said that sometimes he refused to land until he finished the novel he was reading at the time.*

8

Analysis of HIV-1 Subtype U

Contents

8.1 Introduction 364
8.2 Equipment 365
8.3 Methods 366
8.4 Global Results of the Experiment 368
8.5 Analysis of Phylogeny and Explainability Patterns . . 370
8.5.1 AB485664, AB485665 and KY392773 370
8.5.2 AY046058, KY392780, MH078559 and MK095236 379
8.5.3 FJ388921, EF029066, EF029067, EF029068, EF02906, and KY392775 386
8.5.4 HM215249 and HM215251 391
8.5.5 JF683772 399
8.5.6 JX140679 405
8.5.7 KP718920, CRF11_cpx, CRF13_cpx, and CRF27_cpx 413
8.5.8 MN153489, A6 and A1 422
8.5.9 MN153492, F1, CRF12_BF, and F2 427
8.5.10 MN736701 and MN736702 431
8.5.11 Preliminary Conclusions 440
8.6 Conclusive Biological Coherence 441
8.7 Non-Conclusive Biological Coherence 457
8.8 Another Similarities 469
8.9 Conclusions 470

8.1 Introduction

HIV-1 exhibits remarkable evolutionary complexity [2494, 2495]. This virus is highly mutable due to its error-prone reverse transcriptase, leading to the generation of numerous genetic variants within a single host [2496]. Furthermore, the high replication rate of HIV-1 [2497] and the lack of proofreading of its genome contribute to its genetic diversity [2498].

Genetic recombination is another crucial factor in the evolutionary complexity of HIV-1 [2499]. This virus can exchange segments of its genetic material when two different strains infect the same host cell, leading to the emergence of new viral variants. This leads to the rapid emergence of recombinant strains that can acquire drug resistance features and evade immune responses [2500]. Genetic recombination in HIV-1 adds an additional layer of diversity and adaptability to this virus, making it even more challenging to combat [2501].

Natural selection and immunological pressure eliminate less fit strains and promote the evolution of resistant strains [2502], which complicates the development of an effective vaccine and long-lasting antiretroviral therapies [2503]. This evolutionary complexity has consistently challenged efforts to control and eradicate HIV-1 infection.

Some fragments and complete sequences do not match with any of the already described subtypes and are included in the U subtype (unclassified). Moreover, if a recombinant contains unclassified (or unclassifiable) regions, U is included in the name. There are two general circumstances for the assignment of subtype U. One is when the sequence is so divergent from previously-classified sequences that a subtype cannot be assigned based on current reference sequences. The second situation is when the sequence is of such poor quality that the subtype cannot be determined with any reasonable certainty [120].

There are 23 U subtype complete sequences until 2020. And about 10 CRF subtypes contain U fragments [41].

Despite extensive efforts devoted to HIV/AIDS research and control, a significant level of uncertainty persists regarding the diversity of viral families and subfamilies

inherent in the HIV-1 genome [2504, 2505]. This knowledge gap is exemplified by the inclusion of unclassified sequences designated as "U subtype". The U classification can be seen as an indication of our limited understanding of HIV-1 variability, and it is hypothesized that, with time and improved characterization of missing genomic regions, this category of U sequences might be reevaluated and potentially assigned to previously unknown subtypes or subfamilies.

Further knowledge of these sequences and fragments may provide insight into the phylogeny of HIV-1 and its viral evolution.

This Chapter is composed of Sections 8.2 and 8.3, where I describe the equipment and methodology used, respectively. Section 8.4 presents the CNN classification results in terms of accuracy and total hot zones by category. In Section 8.5, the results of the comparative analysis between the phylogeny and the hot zones of the subtype U sequences with their phylogenetically related sequences are shown. I then conducted a final analysis with the existing literature on a case-by-case basis, finding conclusive cases of coherence among my results, the phylogeny, and scientific knowledge (Section 8.6) and non-conclusive cases (Section 8.7). In Section 8.8, additional similarities identified in the hot zones of subtype U sequences with those from other subtypes, which are of particular interest, are also discussed. This Chapter ends with a section on Conclusions (Section 8.9).

8.2 Equipment

All experiments were run in this equipment:

- Processing Unit: Intel(R) Xeon(R) Gold 5220R CPU @ 2.20 GHz.
- Installed RAM: 256 GB usable.
- Operative System: Windows 10 Education. Version: 21H1.
- GPU: NVIDIA RTX A6000. Total memory: 179.451 GB. VRAM: 48.571 GB.

8.3 Methods

The experimental methodology followed the same scheme explained in Chapter 5, Section 5.4. I generated genomic spectrograms for all complete HIV-1 sequences (concatenated and superposed). Next, I fine-tuned a pre-trained CNN using Transfer Learning with this dataset of genomic spectrograms. After confirming the proper functioning of the pre-trained CNN, I applied three different explainability tools (Grad-CAM, LIME, and Gradient Attribution) to obtain relevant hot zones in the recombinant feature. Finally, I constructed the phylogenetic tree using the HIV-1 complete consensus multiple sequence alignment from the Los Alamos National Laboratory HIV-1 Database (up to 2020) and compared the phylogenetic closeness to the degree of similarity of the hot zones.

The complete description about this process is detailed in Chapter 5, Sections 5.5, 5.6, 5.7, 5.8, 5.9, 5.10, 5.11, and 5.12.

In this case study, I focused solely on the Grad-CAM hot zones, given the visual clarity of the results obtained with this explainability tool. Likewise, I worked exclusively with superposed spectrograms, due to their enhanced visibility to the human eye compared to concatenated spectrograms. The x-axis of a concatenated spectrogram measures four times the length of the x-axis of a superposed spectrogram, resulting in more precise localization of the hot zones due to the smaller amplitude of the x-axis.

On the basis of the phylogenetic position of all U subtype sequences in the HIV-1 complete sequences tree, I studied their closeness relationships and their explainability patterns. This analysis was complemented with data from samples registered in the Los Alamos Sequence Database [41].

In this Chapter, I describe the more comprehensive phylogenetic analysis I conducted on all sequences closely related to each of the complete sequences of the U subtype.

Additionally, I generated all the Grad-CAM heat map images for the 13,554 complete HIV-1 sequences until 2020.

I analyzed each branch of the phylogenetic tree where the U sequences are located, both from a phylogenetic perspective and in terms of their similarity to the hot zones of the phylogenetically close sequences.

I analyzed also the CNN classification and the degree of convergence of the hot zones with the total hot zones corresponding to their category. I also examined in detail the level of homology of the hot zones in both U subtype sequences and phylogenetically close sequences of other subtypes. See Section 8.5.

Although I worked with superposed spectrograms, it was not a priority at this point to discriminate the genomic areas involved in classifying a complete HIV-1 sequence as recombinant or non-recombinant in this case study. Introducing this variable could excessively complicate my analysis.

I attempted to determine the presence of hidden genetic recombination events in sequences classified as U, as well as identify potential ancestral relationships between the U sequences and their phylogenetically close sequences. Additionally, I analyzed possible concealed relationships between non-U subtypes present in the regions of the phylogenetic tree under research.

And finally, I conducted a thorough investigation of the existing microbiological evidence to attempt to determine the level of alignment between the results produced by the CNN and the expert knowledge in microbiology and virology. See Sections 8.6, 8.7, 8.8, and 8.9.

By analyzing the results obtained within the U subtype, I was able to extract additional information from the perspective of recombinant featuring. Understanding HIV-1 genetic recombination can contribute to understanding HIV-1 viral evolution itself.

A better knowledge of both complete U sequences and U fragments detected in multiple CRFs will allow us to establish more precisely how the viral evolution of HIV-1 has been and how it is. The use of Deep Learning tools for this purpose may provide new approaches to elucidate knowledge.

Appendix D is a compilation of the classification results for all the sequences involved in this chapter, along with their corresponding Grad-CAM hot zones

(superposed spectrograms). All sequences were grouped by subtypes and in alphanumeric order.

8.4 Global Results of the Experiment

More information about the experiment results can be found in Chapter 5, Sections 5.13, 5.14, and 5.15.

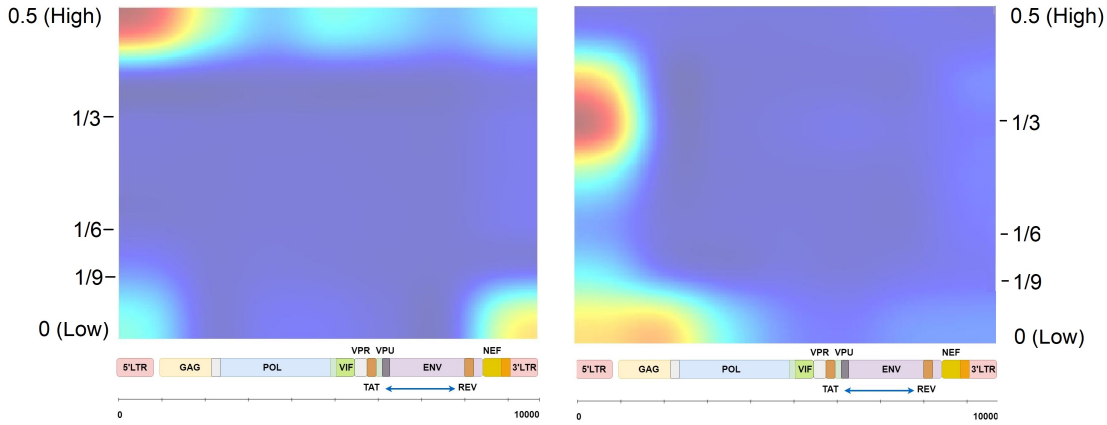
As seen in Table 8.1, CNN hit rate detecting recombinant feature reached 98.27% on the total of the 13,554 complete HIV-1 sequences. The CNN successfully classified 10,993 HIV-1 complete sequences of the total of 11,163 non-recombinant sequences. This meant a success rate of 98.48%. Of the total of 2,391 complete recombinant sequences, the CNN correctly identified 2,326 as recombinants. So the recombinant hit rate was 97.28%. We must take into account that the CNN previously learned with sequences of Train and Validation Sets, so these had higher hit rate than those of the Test Set.

Table 8.1: CNN Error Rates of Classification. Both the errors and the successes are expressed in terms of the number of sequences as well as in %.

Non-recombinant Error Rate		
Errors	170	1.52%
Successes	10,993	98.48%
TOTAL	**11,163**	**100.00%**
Recombinant Error Rate		
Errors	65	2.72%
Successes	2,326	97.28%
TOTAL	**2,391**	**100.00%**

Figure 8.1 shows explainability patterns for the two categories. Figure 8.1a is the result of the mathematical superposition of the 11,163 non-recombinant patterns and Figure 8.1b is the recombinant global pattern of the 2,391 recombinant sequences.

As seen in Figure 8.1a, high frequencies of 5'LTR and low frequencies of 3'LTR were the main hot areas where the CNN looks at to classify a sequence as non-recombinant. Secondary zones were high frequencies of the central part of the genome (around *pol* and *vif* genes) and low frequencies of 5'LTR. In the case of recombinant



(a) Non-recombinant Total Hot Zones **(b)** Recombinant Total Hot Zones

Figure 8.1: Total Explainability Patterns by Category (Superposed Spectrograms)

sequences, hot areas were one-third frequency at 5'LTR (the most determinant point) and low frequency range in 5'LTR and *gag*. Other much more minor points were low frequencies of 3'LTR and the end of *env* and *nef* genes, and one third of 3'LTR.

Chapter 6, Section 6.4 detailed error rates per subtype. Among the subtypes (non-recombinant and recombinant) with more than 15 complete sequences, U subtype had the highest error rate (39.13%), significantly surpassing the following subtypes. The next subtype with the highest error rate was G (25.25%). The U subtype includes 23 complete genomic sequences, of which nine are incorrectly classified by the CNN as recombinant. This was a remarkable error rate, especially when compared to the error rates of the other subtypes.

Compared to the rest of the non-recombinant subtypes, the high error rate of subtype U suggests that there may be recombination phenomena that can help, leading not only to propose a review of this subtype sequences, but also to extract similarities through their explainability patterns by low-computing algorithms.

I analyzed in depth explainability patterns related with recombinant feature of each U subtype sequence.

My results were biologically consistent and robust. I obtained not only valuable information about subtype U, but I acquired knowledge of inter-subtype parentage and their common and distinct genomic components. All this with a fast, efficient and low computing methodology.

Following sections describe each level of homology between explainability patterns and biological evidences.

8.5 Analysis of Phylogeny and Explainability Patterns

In the phylogenetic tree corresponding to the complete HIV-1 sequences, each of the nodes (blue points in the phylogenetic tree) represents a common ancestor, which is unknown in this case [2506].

Unfortunately, in virology, tracking evolutionary relationships is extremely complex, and therefore, unequivocally determining common ancestors in the tree is challenging [2507]. However, despite not knowing the exact common ancestor, we can extract information related to evolutionary trends and branches.

In addition, genetic recombination in viruses like HIV-1 represents a significant challenge for understanding the evolutionary and phylogenetic relationships between different strains. As HIV-1 replicates and spreads within the human body, interactions between multiple genetic variants create opportunities for recombination. This results in a complex network of genetic relationships rather than a clear and linear phylogenetic tree. Recombination can occur in different regions of the viral genome, further complicating the inference of common ancestors and the traceability of viral evolution [2508].

In this section, I analyzed all sequences that are phylogenetically close to the sequences of subtype U, so I extracted their Grad-CAM images in an attempt to discern in detail the existence of common patterns and their individual and collective relationship with their phylogenetic interconnections.

8.5.1 AB485664, AB485665 and KY392773

The CNN classified AB485664 and AB485665 as non-recombinant with a confidence level of over 90%, and KY392773 as non-recombinant as well, but with a relatively lower confidence level (around 69%).

AB485664 and AB485665 are intra-patient (Patient Code: KITA), the latter being the complete reference sequence for AB485664, and therefore, only AB485665 appears in the tree.

In the partial phylogenetic tree shown in Figure 8.2, I observed a certain proximity between AB485665 (along with AB485664) and KY392773.

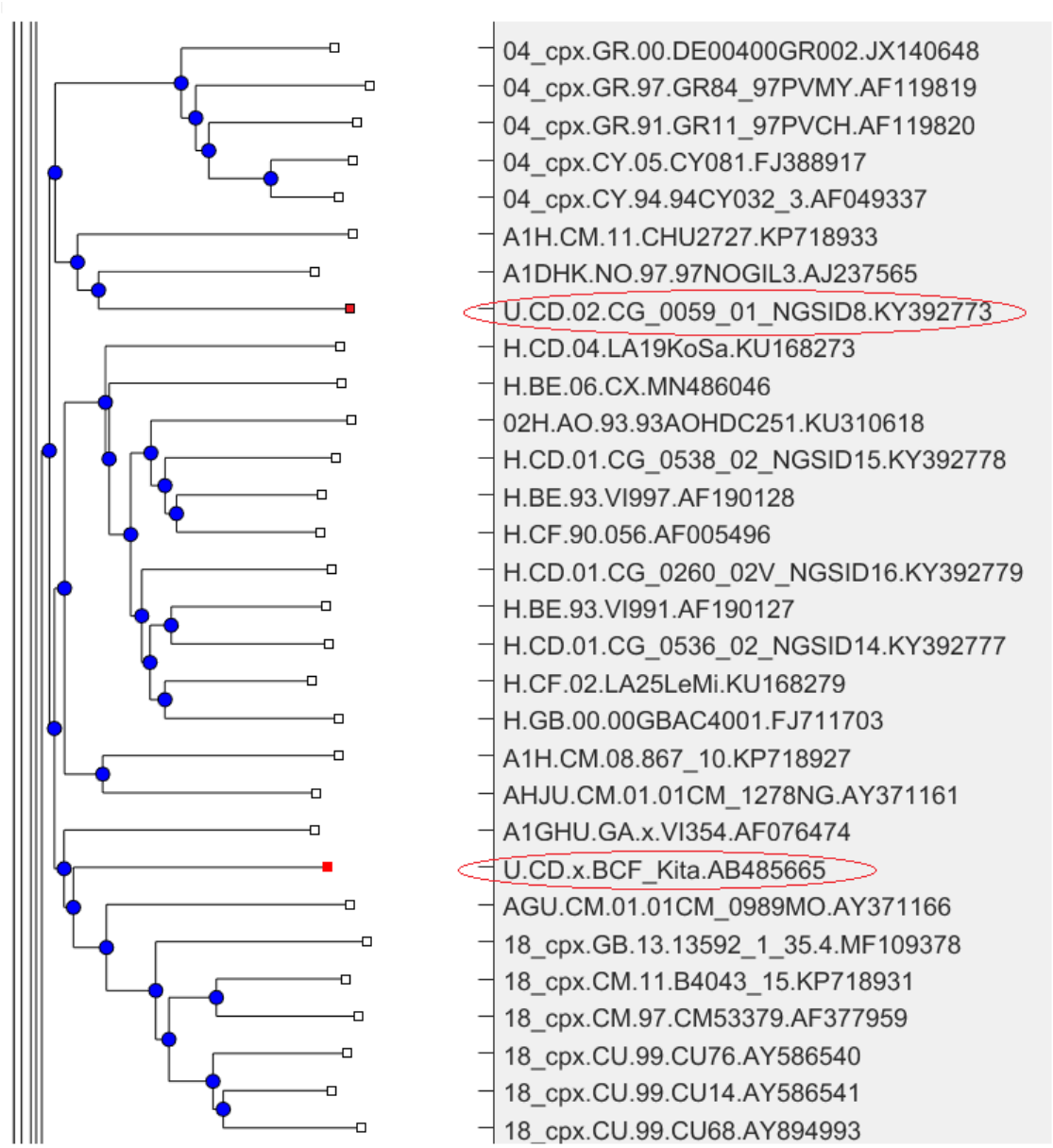


Figure 8.2: Partial Phylogenetic Tree Including Subtype U AB485665 and KY392773 Sequences

AB485665 shares a common ancestor with subtype A1GHU AF076474, which is the product of the recombination among A1, G, H and U. There is an evolutionary

bifurcation between AB485665 and a clade composed of subtype AGU AY371166 and six CRF18_cpx sequences (mosaic of subtypes A, G, H, K, and unclassified regions). Phylogenetically nearest to KY392773 are its sister sequences AJ237565 (subtype A1DHK) and KP718933 (subtype A1H), which branched off from a common ancestor. Their closest clade, which includes these three sequences, also contains five CRF04_cpx sequences (mosaic of subtypes A, G, H, K, and unclassified regions).

Table 8.2 displays the hot zones (Grad-CAM) of KY392773 and the sequences belonging to the CRF04_cpx subtype. Although subtype U KY392773 was classified as non-recombinant, the CNN's confidence level was relatively low (less than 70%). All CRF04_cpx sequences were correctly classified as recombinant by the CNN, with high confidence levels. The hot zones exhibited significant variability across all sequences, with only a slight similarity observed in AF119819 and AF119820.

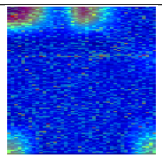
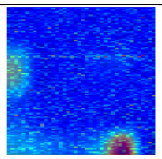
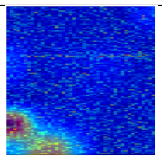
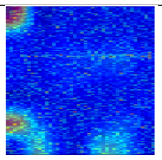
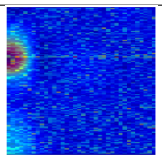
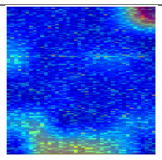
KY292773 hot zones were substantially different from CRF04_cpx. There was also a high variability among CRF04_cpx explainability patterns.

The high variability of KY392773 with the other recombinant sequences, as well as the limited homogeneity among the recombinant sequences themselves, could be related to the complex mosaic of subtypes involved in this CRF. In some cases, the similarity of explainability patterns was higher between closely related sequences (see AF119819 and AF119820).

The high divergence among the hot zones of CRF04_cpx sequences may be attributed to the complexity of recombination events in this subtype, as it is a mosaic of four different subtypes (A, G, K, and H). The genome map of CRF04_cpx subtype is shown in Figure 8.3.

In this subtype, certain fragments (uncolored) have not been cataloged within the identified subtypes. These fragments include the 5'LTR, approximately half of *pol*, substantial portions of *vif* and *vpr*, one-third of *env* and *nef*, as well as the complete 3'LTR. The existence of this relevant, unclassifiable portion of the genome may explain, at least partially, the diversity of its explainability patterns. Perhaps there are common elements in the KY392773 sequence genome that are correlated with those unclassifiable parts in the CRF04_cpx subtype. This could

Table 8.2: KY392773 (Subtype U) and CRF04_cpx Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
KY392773	U	Non-recombinant	69.660%	
JX140648	CRF04_cpx	Recombinant	97.716%	
AF119819	CRF04_cpx	Recombinant	96.615%	
AF119820	CRF04_cpx	Recombinant	92.580%	
FJ388917	CRF04_cpx	Recombinant	96.297%	
AF049337	CRF04_cpx	Recombinant	88.671%	

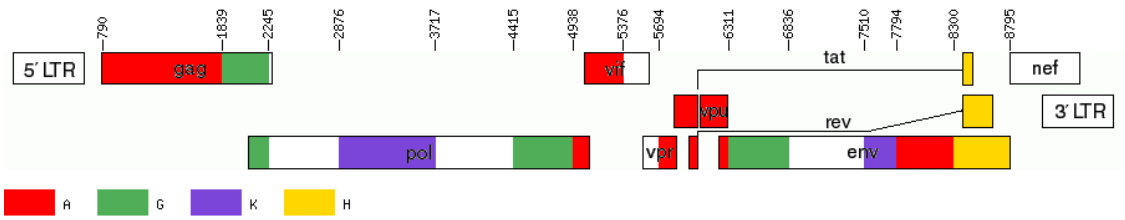


Figure 8.3: Genome Map of Recombinant Subtype CRF04_cpx. This subtype is a mosaic of the non-recombinant subtypes A, G, K, and H. Image sourced from the Los Alamos National Laboratory HIV-1 Database HIV Circulating Recombinant Forms (CRFs) [41].

account for the phylogenetic proximity between the KY392773 clade and the clade formed by these CRF04_cpx sequences.

Although a clear interrelation between phylogenetic position and similarity in hot areas was not expressed in this case, there were certain indications that can explain this phenomenon and shed light on certain genomic and evolutionary aspects of the unclassifiable fragments of CRF04_cpx, as well as their origin.

The next step was to analyze the clade to which AB485664 and AB485665 belong. Table 8.3 compiles their hot zones together with the phylogenetically close CRF18_cpx sequences. All sequences were correctly classified with high confidence levels, except in the case of AY586541. This sequence was classified by the CNN as non-recombinant with a relative confidence level, although slightly below 75%. Generally, the degree of similarity between the hot zones of AB485664 and AB485665 with the recombinant sequences MF109378 was very high. To a lesser extent, there was some similarity with AY894993 and AY586540.

The sequences AB485664 and AB485665 are intra-patients, so their phylogenetic closeness is maximum. This fact was consistent with the high degree of similarity in their hot zones.

An analysis of the genome map of the CRF18_cpx subtype can provide additional information about this case study.

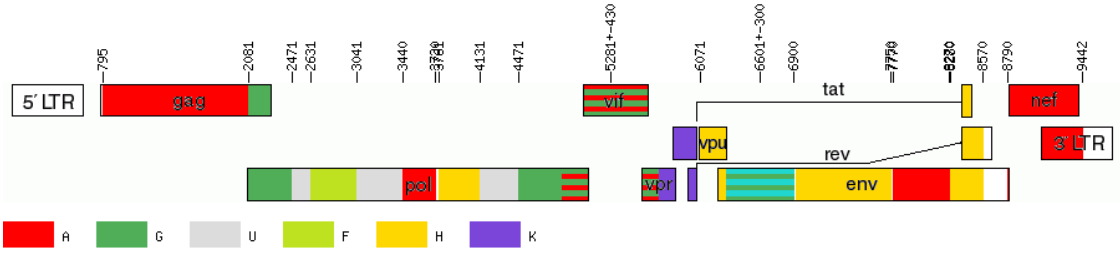
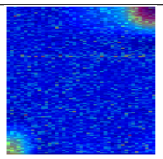
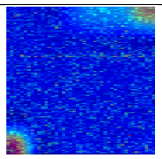
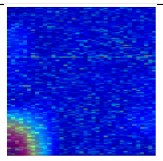
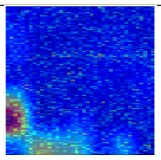
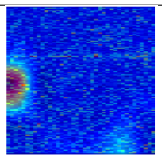
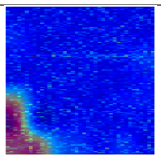
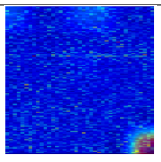
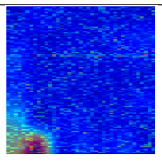


Figure 8.4: Genome Map of Recombinant Subtype CRF18_cpx. This subtype is a complex mosaic of subtypes A, G, U, F, H, and K. Image sourced from the Los Alamos National Laboratory HIV-1 Database HIV Circulating Recombinant Forms (CRFs) [41].

As seen in Figure 8.4, the CRF18_cpx subtype exhibits multiple genetic recombination events, which are extremely complex, as well as fragments that are not identified (uncolored). This is evident, for example, in the *vif* gene, where there are

Table 8.3: AB585664 and AB585665 (Subtype U) and CRF18_cpx Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
AB485664	U	Non-recombinant	90.331%	
AB485665	U	Non-recombinant	92.995%	
MF109378	CRF18_cpx	Recombinant	98.499%	
KP718931	CRF18_cpx	Recombinant	99.569%	
AF377959	CRF18_cpx	Recombinant	96.008%	
AY586540	CRF18_cpx	Recombinant	96.447%	
AY586541	CRF18_cpx	Non-recombinant	74.161%	
AY894993	CRF18_cpx	Recombinant	98.893%	

no genomic fragments belonging to a single subtype; rather, it has been determined to be a complete mixture of subtypes A and G. Clearly, this complexity and the absence of a fully traceable and unequivocal explanation for the generation and development of this subtype significantly hinder the establishment of unequivocal evolutionary relationships with phylogenetically close sequences, including this case study.

Interestingly, the degree of similarity in the hot zones was consistent with the phylogeny.

There was a very high similarity between AB585665 (and AB585664) and its phylogenetically closest CRF sequence, MF109378. It would be necessary to analyze the genomic U zones of this sequence to determine its degree of similarity with AB585664 and AB585665. The phylogenetic proximity of both sequences and their high degree of similarity in the hot zones (despite the CNN correctly classifying them into their respective categories) suggests the possibility that the U fragments of MF109378 specifically originate from a common ancestor with AB585664 and AB585665.

Another phenomenon consistent with phylogeny is that sequences phylogenetically close to MF109378 displayed similar explainability patterns, especially KP718931, AY586540, and AY894993.

In the case of AY586541, the divergence was related to the CNN classification of this sequence as non-recombinant, with a moderate confidence level of 74.16%. Interestingly, in all three sequences, hot zones were present in the 3'LTR region: in the high-frequency range for AB585664 and AB585665, and in the low-frequency range for AY586541. However, regardless of this incorrect classification, the possible relationship between AB585664, AB585665, and AY586541 should be explored.

Regardless of this point, the similarity level of explainability patterns was in agreement with phylogeny.

With the aim of extracting additional information from phylogenetically close sequences, I also analyzed the hot zones of the non-recombinant sequences of subtype H that are included in the excerpt from the phylogenetic tree.

Both CRF04_cpx and CRF18_cpx contain genomic fragments belonging to subtype H, so a certain phylogenetic closeness between the three subtypes seems logical, as these CRFs result from evolutionary drifts from H, due to genetic recombination events. Table 8.4 compiles the hot zones of the H sequences present in the excerpt from the phylogenetic tree.

Table 8.4: Subtype H Explainability Patterns. Each sequence is identified by its ID and subtype. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

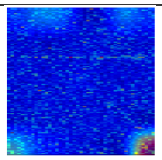
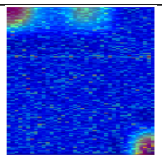
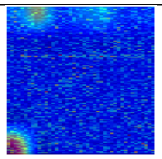
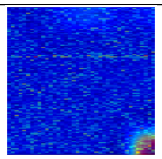
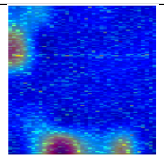
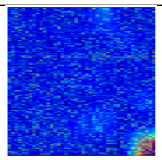
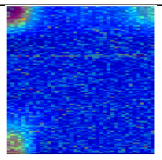
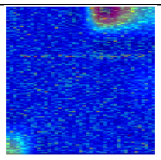
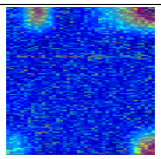
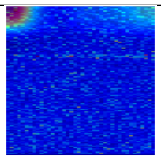
Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
KU168273	H	Non-recombinant	89.265%	
MN486046	H	Non-recombinant	89.830%	
KY392778	H	Non-recombinant	84.282%	
AF190128	H	Non-recombinant	98.441%	
AF005496	H	Recombinant	77.936%	
KY392779	H	Non-recombinant	99.995%	
AF190127	H	Non-recombinant	92.135%	

Table 8.4 – Continuation from the previous page

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
KY392777	H	Non-recombinant	72.514%	
KU168279	H	Non-recombinant	93.215%	
FJ711703	H	Non-recombinant	98.114%	

In general terms, the CNN correctly classified all sequences with a high degree of confidence. The exceptions were AF005496, classified as recombinant with a relative confidence level slightly below 78%, and KY392777. This sequence was correctly classified, but its confidence level is moderate, around 72%. The hot zones were similar and robust in the majority of sequences.

Excluding AF005496, the majority of the hot zones in these H sequences were similar to each other and consistent with the non-recombinant category.

There was a clear correlation between the degree of similarity in the hot zones and phylogenetic proximity. See the case of KU168273 and MN486046, and AF190127 with FJ711703.

All hot zones were coincident with global non-recombinant explainability areas, unless AF005496, whose patterns clearly defined this sequence as recombinant (confidence score of 77.94%).

KY392778, AF190127 and KY392777 hot zones had certain similarity level with AB485664 and AB485665 ones. Similarly, a certain degree of resemblance was detected between the hot zones of KY392773 and the H subtype sequence MN486046. Given the phylogenetic proximity of these H sequences, it would be necessary to analyze the genomes of these sequences in detail to attempt to establish levels of homology, clarify the parental relationship between them,

and obtain more information about the origin and evolution of these sequences currently classified as U.

In this case, the combined analysis of phylogeny and the degree of similarity of the hot zones could provide additional information that may guide relationships of kinship and evolutionary branching affecting these sequences and related subtypes. Additionally, a detailed analysis of the underlying genomic factors determining the explainability patterns of subtype U AB485664 and AB485665 (classified as non-recombinant), and those of sequences such as subtype CRF18_cpx MF109378 (classified as recombinant), could shed light on the presence of mathematical patterns involved in the recombinant feature, as well as potential hidden phylogenetic relationships among these sequences.

While the analysis of the case involving KY392773 and CRF04_cpx sequences was not as conclusive as that of AB485664 and AB485665 with CRF18_cpx sequences, in general, there was a correlation between the phylogenetic position and the degree of similarity of the hot zones.

8.5.2 AY046058, KY392780, MH078559 and MK095236

In the tree region shown in Figure 8.5, we can find four subtype U sequences that are phylogenetically related and close to each other. MK095236 and the other two clades share a common ancestor. the first of these clades is subdivided in two subclades. The first subdivision contains subtype A7 and A1U sequences. The second subdivision includes KY392780 and their parent sequences, consisting in one A1U sequence and five subtype CRF45_cpx sequences (mosaic of subtypes A, K y U). The second main clade is structured with AY046058 and its parent sequences belonging to subtype CRF02_AG (recombination between A and G), subtype 0209 (mixture of CRF02_AG and CRF09_cpx) and subtype CRF09_cpx (mosaic of A, G and U). Subtype U MH078559 and subtype AKU DQ886038 are the unique sisters of its clade.

First, I proceeded with the analysis of the hot zones of subtype U sequences in this specific section of the phylogenetic tree (Table 8.5).

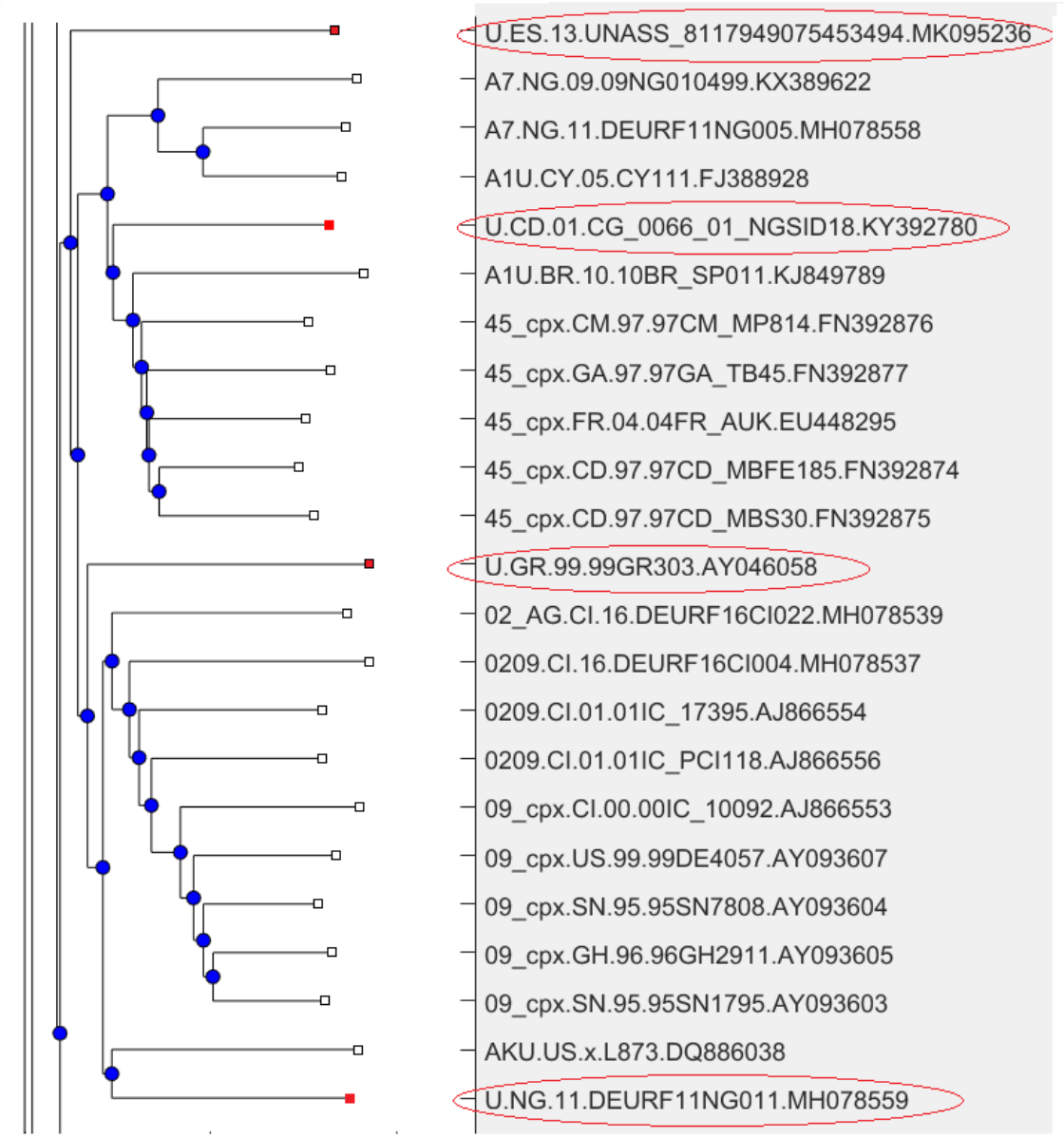
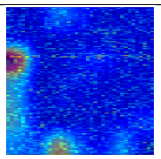
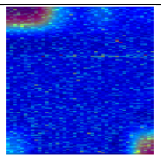
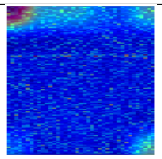
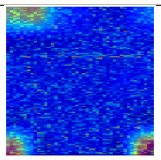


Figure 8.5: Partial Phylogenetic Tree Including Subtype U AY046058, KY392780, MH078559 and MK095236 Sequences

The CNN classified MK095236 as recombinant. Its hot zones were consistent with this classification. The remaining U sequences were classified as non-recombinant, and their hot zones are similar, although KY392780 was classified as non-recombinant with a low confidence level, around 62%.

KY392780, AY046058 and MH078559 shared similar patters aligned with non-recombinant global hot zones. In addition, hot areas of these three sequences were deeply related to their phylogenetic closeness. Conversely, the CNN classified

Table 8.5: AY046058, KY392780, MH078559 and MK095236 (Subtype U) Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
MK095236	U	Recombinant	82.815%	
KY392780	U	Non-recombinant	62.076%	
AY046058	U	Non-recombinant	90.426%	
MH078559	U	Non-recombinant	91.760%	

MK095236 as recombinant in accordance with its explainability pattern.

In light of the data in Table 8.5, several hypotheses arise. There may be a possible common ancestor between MK095236 and the rest of the sequences in the tree excerpt, where genetic diversity led to genetic recombination phenomena that the CNN somehow elucidates, hence classifying MK095236 as recombinant.

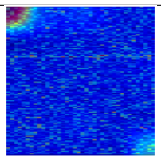
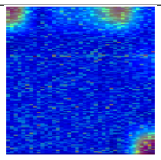
KY392780, AY046058, and MH078559, although clearly classified as non-recombinant, belong to clades with an overwhelmingly predominant presence of recombinant sequences (except for two sequences belonging to subtype A7). It is possible that these sequences may contain genome elements related to the recombinant feature, despite being classified as non-recombinant by the CNN. Delving deeper into understanding their genomes, as well as those of the adjacent recombinant sequences, is of interest to identify the reasons behind this phenomenon and the exact genomic relationship between these non-recombinant subtype U

sequences and their related sequences. It is also interesting to research the relationship among those sequences belonging to subtype A7, the subtype U sequences, and the compilation of recombinant sequences that make up this zone of the phylogenetic tree.

Table 8.6 summarizes the results of the CNN classification, as well as the corresponding hot zones for the A7 sequences in the studied phylogenetic region.

The CNN classified both sequences as non-recombinant with a high degree of confidence, though slightly below 90%, around 85%. The hot zones of both sequences were relatively similar.

Table 8.6: Subtype A7 Sequences Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
KX389622	A7	Non-recombinant	86.314%	
MH078558	A7	Non-recombinant	84.715%	

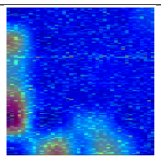
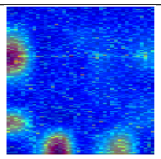
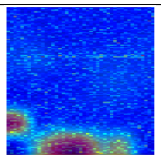
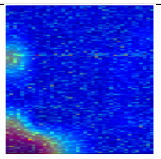
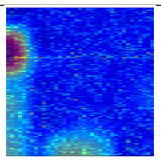
Their hot zones were relatively similar and consistent with their non-recombinant classification. There was a high resemblance with the explainability patterns of subtype U AY046058, MH078559, and especially with those of KY392780. The degree of similarity was particularly striking between KY392780 (subtype U) and MH078558 (subtype A7). This was entirely accordant with the phylogeny, as these two sequences are precisely the closest in the tree.

KY392780 is part of a clade that includes several CRF45_cpx sequences.

Table 8.7 presents the compilation of the CRF45_cpx sequences involved in this case study. All sequences were correctly classified by the CNN with high confidence

levels. The degree of similarity in their hot zones was noticeable, especially between FN392876, FN392877, and FN392875, as well as between EU448295 and FN392874.

Table 8.7: Subtype CRF45_cpx Sequences Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
FN392876	CRF45_cpx	Recombinant	95.330%	
FN392877	CRF45_cpx	Recombinant	98.648%	
EU448295	CRF45_cpx	Recombinant	83.951%	
FN392874	CRF45_cpx	Recombinant	99.806%	
FN392875	CRF45_cpx	Recombinant	98.699%	

Subtype CRF45_cpx is a mosaic of subtypes A, K and U. Figure 8.6 clarifies the genome map of this subtype, with all its genomic regions identified in relation to its parental subtypes.

Several fragments of *pol*, parts of *env*, *tat*, and *ref*, along with the second half of *nef* and nearly the entire 3'LTR, originate from sequences classified as subtype U. This may rationalize the phylogenetic proximity to the U sequences in this case study.

Although this CRF is a complex mixture of subtypes, these sequences are moderately similar. As offspring from a common ancestor is produced, explainability

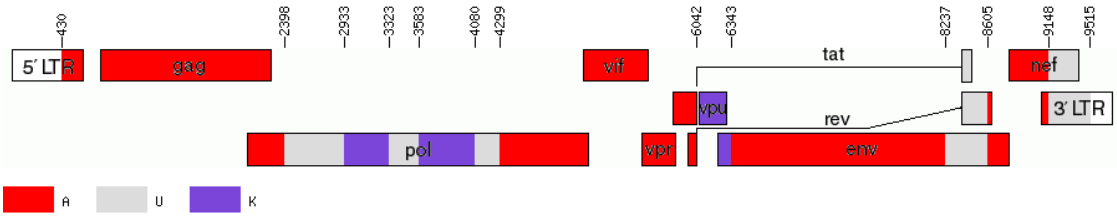


Figure 8.6: Genome Map of Recombinant Subtype CRF45_cpx. This subtype is a mosaic of subtypes A, U, and K. Image sourced from the Los Alamos National Laboratory HIV-1 Database HIV Circulating Recombinant Forms (CRFs) [41].

patterns evolve. FN392877 patterns are the closest with FN392876. EU448295 differs very slightly from FN392877. FN392874 explainability patterns are similar to EU448295 and FN392875. FN392876, FN392875, and FN392877 hot zones are similar to subtype U MK095236. In fact, these sequences derive from a close common ancestor.

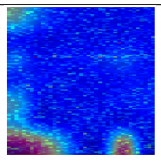
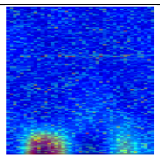
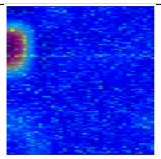
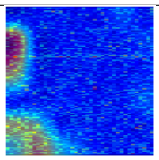
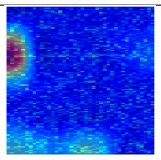
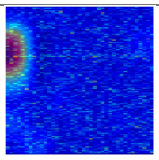
The second subcluster is composed, in addition to subtype U AY0046058 and MH078559, of sequences belonging to various recombinant subtypes. Table 8.8 is a compilation of the hot zones of the sequences belonging to this subcluster.

All sequences are correctly classified by the CNN as recombinant with very high confidence levels. Subtype U AY046058 (classified as non-recombinant) is the philogenetically closest sequence of all of them. Explainability patterns between recombinant MH078539 and non-recombinant AY046058 were determined by their opposite classifications. Their phylogenetic proximity did not result in a greater degree of similarity. However, MH078539 and CRF09_cpx sequence AJ866553 shared some similar hot areas. AJ866553 sisters had deeply similar patterns, especially AY093607 and AY093603. Recombinant feature, phylogenetic closeness and explainability patterns homology were highly consistent.

The genome map of subtype CRF09_cpx does not have an available image. This subtype is composed of fragments of subtypes A, F, G, and U, and it possibly shares ancestors with CRF02_AG [2509].

In conclusion, there were multiple similarities between the phylogenetic relationship and the degree of similarity in the hot zones of the U sequences with

Table 8.8: Subtype CRF02_AG and Subtype CRF09_cpx Sequences Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
MH078539	CRF02_AG	Recombinant	98.766%	
AJ866553	CRF09_cpx	Recombinant	98.759%	
AY093607	CRF09_cpx	Recombinant	99.688%	
AY093604	CRF09_cpx	Recombinant	99.894%	
AY093605	CRF09_cpx	Recombinant	99.189%	
AY093603	CRF09_cpx	Recombinant	99.972%	

the sequences in their phylogenetic vicinity. It is striking how closely the hot zones of subtype U MK095236 (classified as recombinant by the CNN) matched with CRF45_cpx sequences FN392876 and FN392877, as well as with FN392875. There was also a similarity with CRF09_cpx sequences MH078539, AY093607, AY093604, AY093605, and AY093603. Regarding KY392780, AY046058, and MH078559 (classified as non-recombinant), their hot zones indeed showed more similarity with A7 sequences KX389622 and MH078558.

This case study is an example of the phylogenetic complexity added by multiple instances of genetic recombination in HIV-1. The current lack of knowledge regarding the circumstances under which these recombinations occurred and the absence of extensive sequencing that would clearly establish the paths of genetic diversification significantly hinder the phylogenetic analysis of this virus.

It would be necessary to thoroughly analyze the genomes of the sequences involved in this case study to try to correlate the identified subtype U regions in CRF45_cpx with these U sequences. Similarly, the same approach applies to CRF09_cpx. Additionally, it is interesting to explore the relationship of these sequences with subtype A7 to more clearly establish their possible evolutionary relationships.

8.5.3 FJ388921, EF029066, EF029067, EF029068, EF02906, and KY392775

Figure 8.7 displays the excerpt from the phylogenetic tree where all subtype U sequences from this case study are located. In this order, FJ388921, EF029066 and KY392775 are immediate relatives. KY392775 share a common ancestor with subtype FK MH705156 and two subtype K sequences (AJ249235 and AJ249239). This cluster has a complex sister clade containing 11 subtype F2 sequences and multiple F2-related recombinant sequences.

The sequences EF029066, EF029067, EF029068, and EF029069 are intra-patient (Patient Code: H10986). Therefore, only the intra-patient reference sequence EF029066 is included in the phylogenetic tree, as the phylogenetic distance between these four sequences is minimal.

The three U sequences in the tree form a clade along with a recombinant sequence FK (between subtypes F and K) and two sequences identified as subtype K. The classification results of all sequences in this clade, as well as their hot zones, are shown in Table 8.9.

With the exception of the intra-patient sequences from EF029066 (EF029066, EF029067, EF029068, and EF029069), all sequences were classified by the CNN

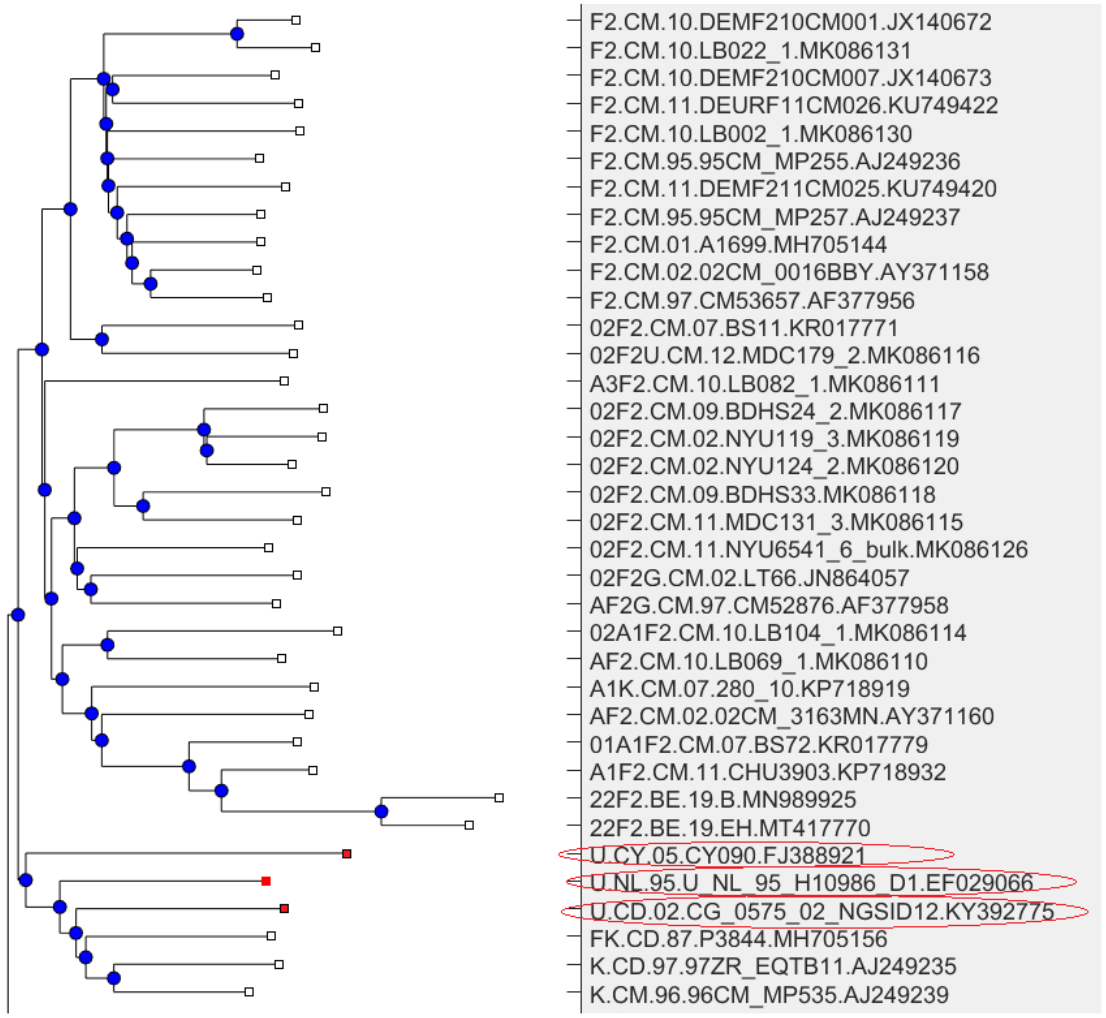
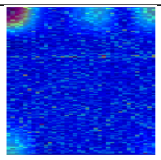
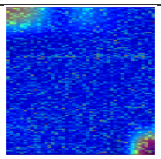
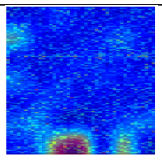
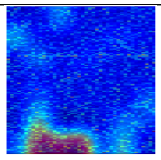
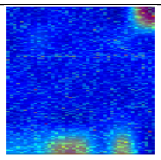
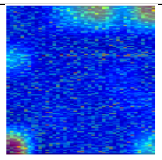
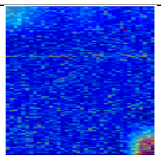
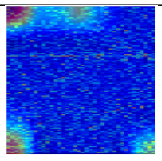


Figure 8.7: Partial Phylogenetic Tree Including Subtype U FJ388921, EF029066, EF029067, EF029068, EF029069 and KY392775 Sequences

as non-recombinant with high confidence levels. Their hot zones were noticeably similar. In the case of EF029067, EF029068, and EF029069, the CNN classified them as recombinant. Their hot zones were extremely similar and directly proportional to the phylogenetic proximity.

Table 8.9: FJ388921, EF029066, F029067, EF029068, EF029069, and KY392775 (Subtype U), and Subtype K Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
FJ388921	U	Non-recombinant	91.853%	
EF029066	U	Non-recombinant	51.424%	
EF029067	U	Recombinant	76.708%	
EF029068	U	Recombinant	69.344%	
EF029069	U	Recombinant	73.163%	
KY392775	U	Non-recombinant	89.344%	
AJ249235	K	Non-recombinant	97.356%	
AJ249239	K	Non-recombinant	99.312%	

FJ388921, EF029066, and KY392775 are very closely related phylogenetically. Their high similarity in the hot zones was consistent with this fact.

The CNN classified EF029066 as non-recombinant but its confidence score

was extremely low (51.42%). Its Grad-CAM hot zones were highly similar to FJ388921 ones. EF029067, EF029068 and EF029069 (intra-patient with EF029066) were classified as recombinant with moderate reliability rates. Their recombinant explainability patterns were notably similar, which is in accordance with intra-patient feature.

Despite the low confidence in classifying EF029066 as non-recombinant, its hot zones were very similar to its closest subtype U sequences, as well as to the K sequences with which it shares a clade, such as AJ249239. The similarity between the patterns of subtype K sequences and those of FJ388921 and EF029066 was also noteworthy.

The fact that the intra-patient sequences of EF029066 were classified as recombinant with a high similarity in their hot zones could indicate the presence of a hidden recombination phenomenon in these intra-patient sequences that the CNN is detecting. This observation is fascinating and requires further research. The set formed by EF029066 and its recombinant intra-patient sequences EF029067, EF029068, EF029069 is a good candidate to identify recombination frequency-related phenomena.

My next step was to also analyze the sister clade to the clade of these U sequences to extract additional information. This clade is composed of a subcluster almost entirely made up of F2 sequences. The other subcluster is formed mostly by non-CRF recombinants derived from subtype F2.

Table 8.10 displays the hot zones of phylogenetically related sequences belonging to F2. Except for KU749422, which was incorrectly classified as recombinant with a low degree of confidence, the rest of the sequences were classified as non-recombinant with high confidence levels. In all cases, the hot zones were consistent with their respective categories. Except for KU749422, the degree of similarity among the remaining sequences was very high.

Table 8.10: Subtype F2 Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

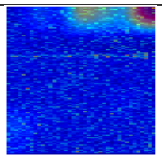
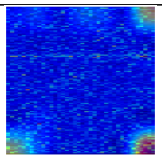
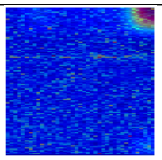
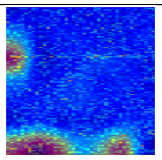
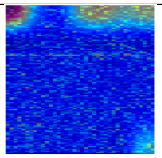
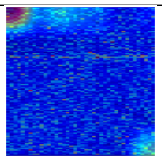
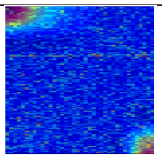
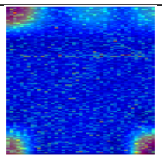
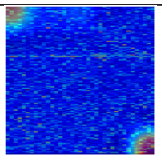
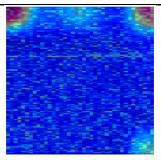
Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
JX140672	F2	Non-recombinant	96.758%	
MK086131	F2	Non-recombinant	90.423%	
JX140673	F2	Non-recombinant	99.456%	
KU749422	F2	Recombinant	60.523%	
AJ249236	F2	Non-recombinant	97.015%	
KU749420	F2	Non-recombinant	91.191%	
AJ249237	F2	Non-recombinant	98.966%	
MH705144	F2	Non-recombinant	91.671%	
AY371158	F2	Non-recombinant	87.008%	

Table 8.10 – Continuation from the previous page

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
AF377956	F2	Non-recombinant	99.293%	

Their explainability patterns were robust and coherent with their non-recombinant category. Except in the case of KU749422, which was classified as recombinant with a low trust level (60.523%). Indeed, its hot zones were consistent with the recombinant category and very different from the rest of the sequences classified as non-recombinant, shown in Table 8.10.

The hot zones of subtype U FJ388921 and EF029066 were very similar to those of the F2 sequences KU749420, MH705144, and AF377956. Those of KY392775 were similar to those of F2 JX140672. EF029066 also resembled the hot zones of F2 AJ249237 and AY371158.

The hot zones of EF029067, EF029068, and EF029069 might show some similarity to those of KU749422 in the low-frequency range, although I considered this similarity inconsistent for drawing conclusions.

In conclusion, there was coherence in the degree of similarity in the hot zones of subtype U sequences present in the tree with phylogenetically close non-recombinant sequences.

I consider the existence of a hidden intra-patient genetic recombination phenomenon that affects sequences EF029066, EF029067, EF029068, and EF029069 as a possibility.

This phenomenon requires further research and can be extremely interesting to unravel information about the genetic recombination mechanisms themselves.

8.5.4 HM215249 and HM215251

Figure 8.8 shows the region on the phylogenetic tree where subtype U sequences HM215249 and HM215251 are located. These two sequences are unique sisters.

There is an evolutionary divergence phenomenon between HM215249 and HM215251 and a clade formed by various CRF and non-CRF recombinant sequences. This sister clade is composed of subtype 26CU FM877784, subtype 26C sequences FM877779, FM877783, and FM877778 (the most distant in relation to the other two 26C sequences), and subtype CRF26_A5U sequences FM877781, FM877782, FM877777, and FM877780. The lower-located clades cluster multiple recombinant subtypes.

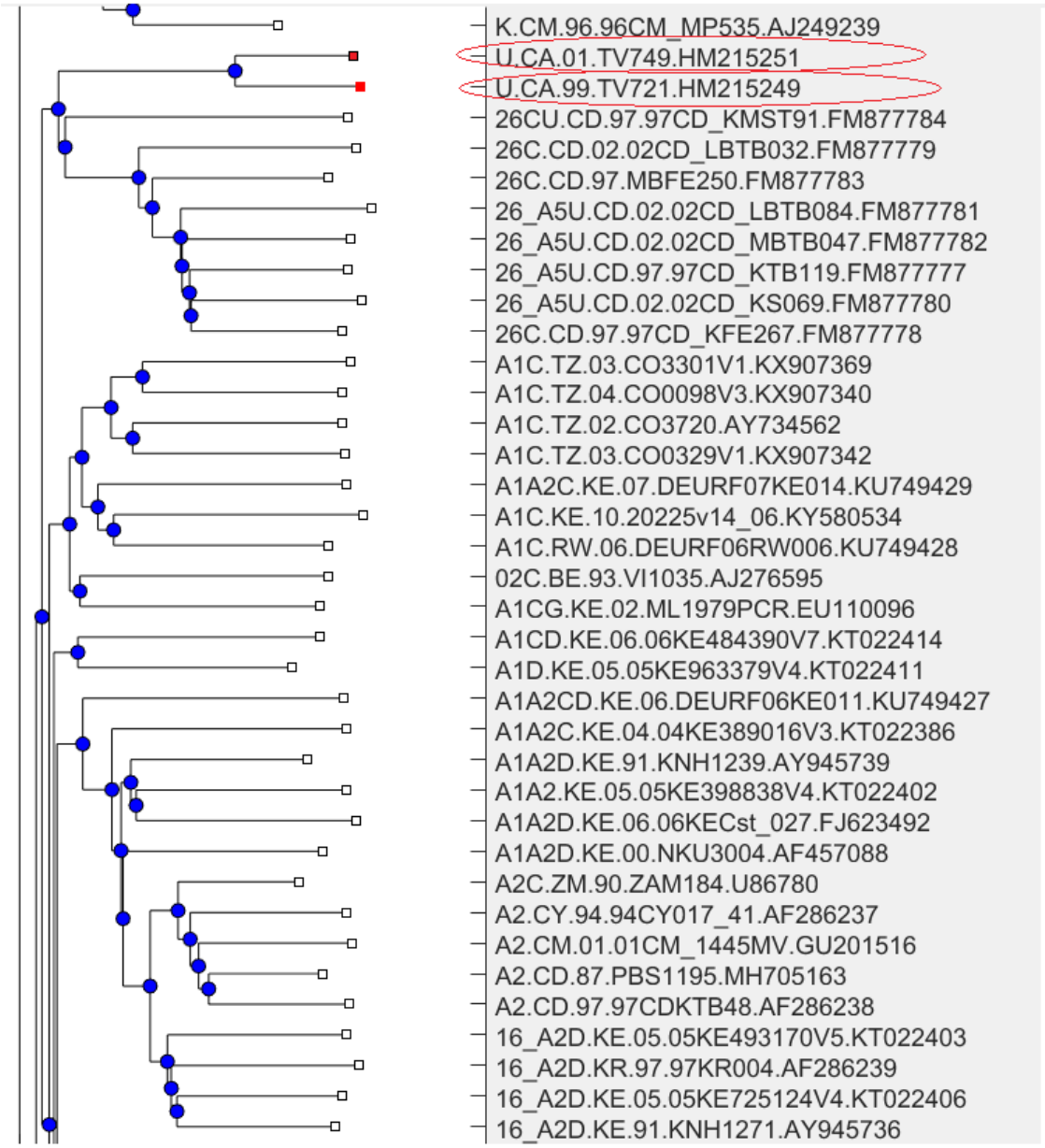
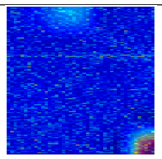
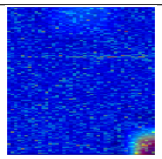


Figure 8.8: Partial Phylogenetic Tree Including Subtype U HM215249 and HM215251 Sequences

The CNN classified HM215249 and HM215251 sequences as non-recombinant with a confidence level exceeding 99%. Their hot zones were consistent with their classification and virtually identical (Table 8.11).

Table 8.11: HM215249 and HM215251 (Subtype U) Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
HM215249	U	Non-recombinant	99.553%	
HM215251	U	Non-recombinant	99.289%	

Their closest phylogenetic distance was accordant with their almost identical explainability patterns. Their phylogenetically closest sequences are recombinants belonging to the non-CRF subtypes 26CU and 26C, as well as CRF26_A5U sequences.

Non-CRF subtypes are those recombinant subtypes that do not comply with the requierements for a CRF to be defined[120]:

- Two or more genome sequences. If only two are provided, then additional partial genomes with the same mosaic structure must also be provided.

- The genome sequences must contain full coding sequence information, minimally from the beginning of *gag* through the end of *nef*. Full-length genomes, including LTRs, are preferred.

- All sequences must be from at least three epidemiologically unrelated infections.

- All sequences must have the same mosaic structure.

- All sequences (and segments of sequences) should be phylogenetically related, i.e., they will cluster together on a tree that includes other sequences of the same subtype(s).

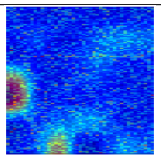
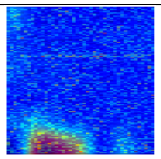
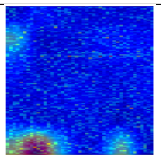
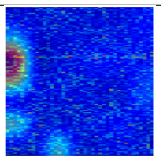
Broadly speaking, non-CRF subtypes (i.e., Unique Recombinant Forms, URFs) are recombinants that have not spread significantly in the population. Each URF is specific to an individual or a very small group.

The reiterated presence of subtype U in the recombination events of sequences close to HM215249 and HM215251 is striking, both within CRF26_A5U itself and in subtype 26CU. This is phylogenetically consistent. The clear genetic branching between HM215249 and HM215251 (classified as non-recombinant by the CNN) and the rest of the sequences in their clade (all of which are recombinant) may indicate genetic recombination events causing this separation between the two subclades. A detailed study of the genomes of the sequences comprising this clade can provide information about hidden genetic recombination phenomena, as well as the exact relationship between these U sequences and their adjacent sequences. It is also interesting to conduct further research to detail the relationship between subtype U HM215249 and HM215251 and subtypes A5, A1, and A2. Additionally, it is important to identify, if present, potential determining elements in the non-recombinant feature that are common among these subtypes.

Table 8.12 displays the CNN classification results as well as the hot zones of the CRF26 sequences. The CNN classified all CRF26_A5U sequences correctly as recombinant, with high confidence levels exceeding 99% for FM877781, FM877777, and FM877780. In the case of FM877782, the confidence level surpassed 90%. The hot zones were consistent with the recombinant category, although a clear relationship between phylogenetic proximity and the degree of similarity of the hot zones was not perceived.

Their hot zones were notably different from HM215249 and HM215251, as expecteed due to the non-recombinant classification of these two sequences. Explainability patterns of these four subtype CRF26_A5U sequences were correlated with phylogenetic distance. FM877782 and FM877777 were the most representative examples of this parallelism given their high degree of similarity in their hot zones and their phylogenetic proximity.

Table 8.12: Subtype CRF26_A5U Sequences Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
FM877781	CRF26_A5U	Recombinant	99.733%	
FM877782	CRF26_A5U	Recombinant	90.827%	
FM877777	CRF26_A5U	Recombinant	99.886%	
FM877780	CRF26_A5U	Recombinant	99.931%	

The observed resemblance between the hot zones of FM877777 and those corresponding to the U subtype sequences EF029067 (Table 8.9) was intriguing. A comprehensive genomic analysis of these sequences would be valuable to uncover potential latent phylogenetic relationships.

Similarly, the similarity level in the hot zones of the sequence CRF26_A5U FM877782 (Table 8.12) with the explainability patterns of subtype U EF029067 and EF029068 (Table 8.9) is striking. This similarity provides additional information to unravel the degree of relatedness or possible hidden relationships among these three sequences, as well as among the other sequences belonging to their respective subtypes.

Figure 8.9 is a graphical representation of the genome map of subtype CRF26_A5U.

Given that this subtype exhibits segments from subtype U, it is of particular interest to analyze whether these segments originate specifically from these U sequences.

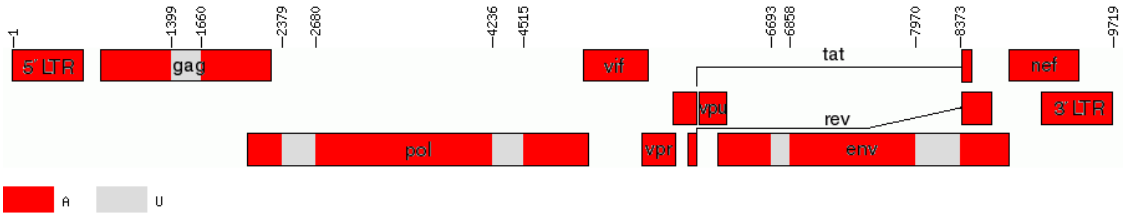


Figure 8.9: Genome Map of Recombinant Subtype CRF26_A5U. This subtype is a mixture of subtypes A5 and U. Image sourced from the Los Alamos National Laboratory HIV-1 Database HIV Circulating Recombinant Forms (CRFs) [41].

The segments labeled as U are located in the central region of *gag*, one segment in the first half of *pol*, and another in its second half, one segment in the first half of *env*, and another in its second half.

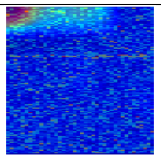
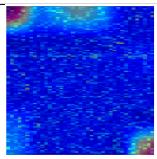
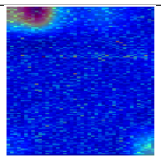
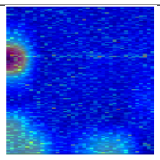
I must note the variability in the hot zones of these four CRF26_A5U sequences despite their phylogenetic proximity. It would be interesting to delve into the understanding of the genomic diversity of this subtype and verify the specific contribution of U fragments to this diversity.

One of the clades located in the lower size of the partial phylogenetic tree, apart from the one containing subtype U HM215249 and HM215251, is composed, among others, of several sequences of the non-recombinant subtype A2. I analyzed their hot zones (Table 8.13) to attempt to establish a connection with the non-recombinant nature of these U sequences, as indicated by the CNN.

The CNN classified subtype A2 sequences AF286237 and GU201516 clearly as non-recombinant. In the case of sequence MH705163, the confidence level of the classification was very low, at only 53.498%. The hot zones of these three sequences were consistent with the non-recombinant category and highly similar to each other. Regarding sequence AF286238, it was clearly classified as recombinant with a confidence level of 99.801%. Its hot zones were consistent with the recombinant feature and, consequently, diverged from the other three sequences.

AF286238 hot zones were divergent because of its recombinant classification (confidence score of 99.80%), in fact its explainability patterns were similar to

Table 8.13: Subtype A2 Sequences Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
AF286237	A2	Non-recombinant	98.597%	
GU201516	A2	Non-recombinant	92.866%	
MH705163	A2	Non-recombinant	53.498%	
AF286238	A2	Recombinant	99.801%	

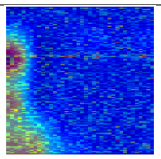
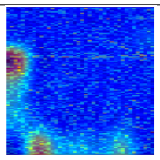
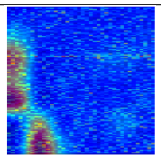
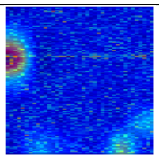
subtype CRF16_A2D and subtype CRF26_A5U FM877780. Its categorical classification as recombinant suggested the existence of hidden recombination phenomena in AF286238. MH705163 was classified as non-recombinant but its trust level is notably low (53.50%). Explainability pattern similarity level and the progressive increase of recombinant feature in MH705163 and AF286238 were correlated with phylogenetic distance.

Interestingly, the hot zones of U subtype HM215249 and HM215251 (Table 8.11) did not show excessive similarity with the hot zones of subtype A2 AF286237, GU201516, and MH705163. It would be interesting to explore into the factors behind these distinct hot zones (within the recombinant category) between these subtype U sequences and the A2 sequences.

Table 8.14 displays the CNN classification results and hot zones of the sequences of subtype CRF16_A2D. The CNN classified all sequences as recombinant with very high confidence levels, ranging between 94.97% and 99.65%. The hot zones

of all sequences were consistent with the ones corresponding to the recombinant category and exhibited very high degrees of homology among themselves.

Table 8.14: Subtype CRF16_A2D Sequences Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
KT022403	CRF16_A2D	Recombinant	99.389%	
AF286239	CRF16_A2D	Recombinant	95.870%	
KT022406	CRF16_A2D	Recombinant	94.974%	
AY945736	CRF16_A2D	Recombinant	99.652%	

Their explainability patterns were notably aligned with global recombinant hot zones. These four sequences were closely related. This was coherent with their high similar explainability patterns.

Figure 8.10 displays the genomic map of subtype CRF16_A2D.

Clear conclusions were not drawn from the analysis of the lower-located clades, neither regarding phylogeny nor the similarity of the hot zones. These clades do not have a close kinship with the clade to which HM215249 and HM215251 belong. This could be associated with the limited similarity of the hot zones. In such a scenario, there continued to be a proportional relationship between phylogenetic proximity and the degree of similarity of the hot zones.

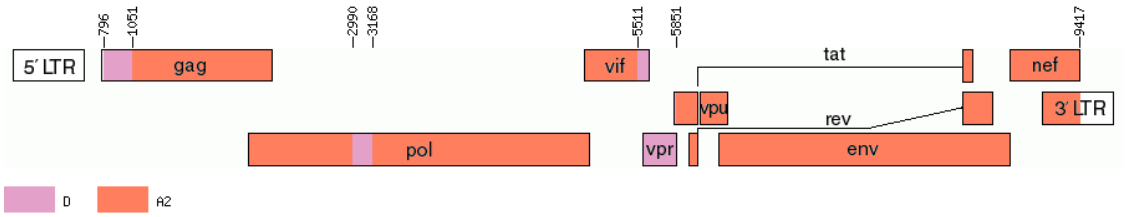


Figure 8.10: Genome Map of Recombinant Subtype CRF16_A2D. This subtype is a mixture of subtypes A2 and D. Image sourced from the Los Alamos National Laboratory HIV-1 Database HIV Circulating Recombinant Forms (CRFs) [41].

At the genomic level, however, interesting conclusions were drawn by working on the clade of these two sequences. There was coherence between the hot zones of both and a close degree of relatedness. The CNN identified a clear non-recombinant feature in both sequences. Phylogenetically, the existence of common ancestors with sequences of the recombinant subtype in the clade to which they belong is evident. The evolutionary divergence between these two U subtype sequences compared to the rest of the sequences, classified as recombinants, is consistent and possibly indicative that the divergence is precisely motivated by the recombinant feature. It is extremely interesting to delve into the genomic knowledge of this set of sequences in an attempt to elucidate the elements that drive this branching and its underlying causes.

8.5.5 JF683772

Figure 8.11 shows the excerpt from the phylogenetic tree where the subtype U sequence JF683772 is located.

Subtype 01G MN486045 is the phylogenetically closest sequence to JF683772 (unique sisters). Subtype 01G means a recombination between CRF01_AE and G subtypes. Related clades are abundant and phylogenetically diverse. One of these clades contains multiple CRF02_AG sequences. Another clade is composed by multiple recombinant subtypes, including three CRF63_02A6 sequences.

Table 8.15 displays the CNN classification results and the hot zones of the CRF63_02A6 sequences in this case study. All non-CRF sequences were excluded from the experimentation due to their spurious nature in terms of epidemiological

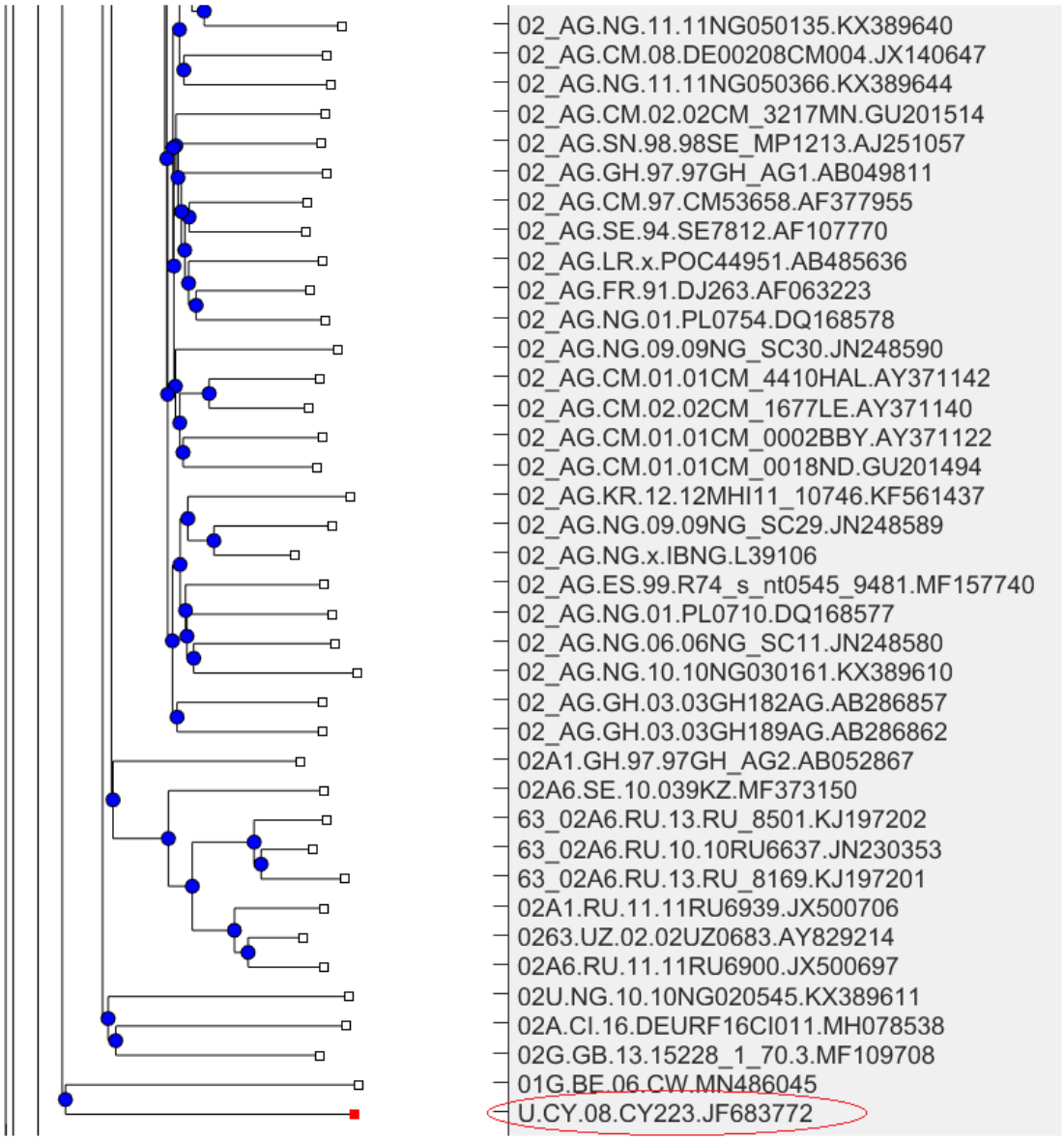


Figure 8.11: Partial Phylogenetic Tree Including Subtype U JF683772 Sequence

presence. The experimentation was conducted considering only sequences belonging to pure subtypes and CRFs, leaving out marginal sequences, so I did not include data on the classification and explainability results of the sister sequence subtype 01G MN486045 or other closely related non-CRF sequences.

Table 8.15: JF683772 (Subtype U), CRF63_02A6 and CRF02_AG Sequences Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

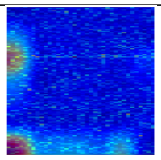
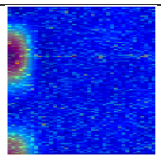
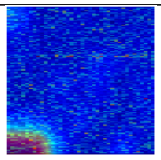
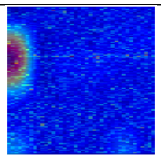
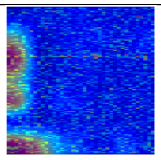
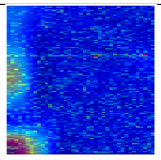
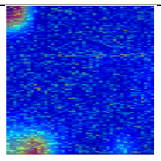
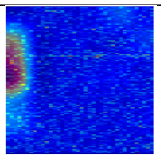
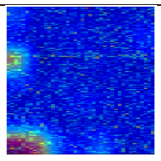
Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
JF683772	U	Recombinant	75.854%	
KJ197202	CRF63_02A6	Recombinant	99.967%	
JN230353	CRF63_02A6	Recombinant	99.935%	
KJ197201	CRF63_02A6	Recombinant	99.830%	
GU201514	CRF02_AG	Recombinant	99.978%	
AJ251057	CRF02_AG	Recombinant	99.999%	
AB049811	CRF02_AG	Recombinant	99.952%	
AF377955	CRF02_AG	Recombinant	99.984%	
AF107770	CRF02_AG	Recombinant	99.934%	

Table 8.15 – Continuation from the previous page

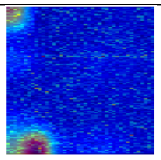
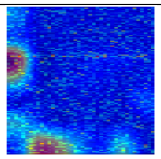
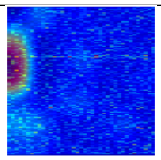
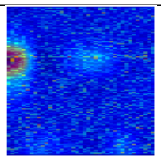
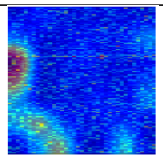
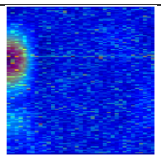
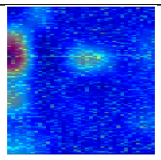
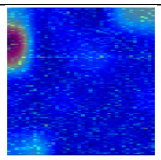
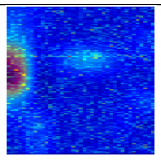
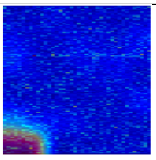
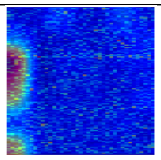
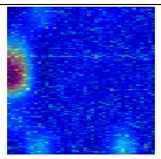
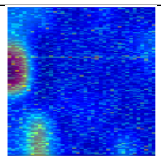
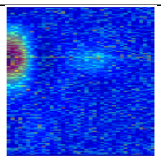
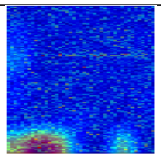
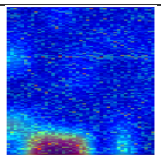
Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
AB485636	CRF02_AG	Recombinant	99.579%	
AF063223	CRF02_AG	Recombinant	99.830%	
DQ168578	CRF02_AG	Recombinant	99.996%	
JN248590	CRF02_AG	Recombinant	99.968%	
AY371140	CRF02_AG	Recombinant	99.737%	
AY371122	CRF02_AG	Recombinant	99.966%	
GU201494	CRF02_AG	Recombinant	99.984%	
KF561437	CRF02_AG	Recombinant	99.599%	
JN248589	CRF02_AG	Recombinant	99.997%	
L39106	CRF02_AG	Recombinant	98.921%	

Table 8.15 – Continuation from the previous page

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
MF157740	CRF02_AG	Recombinant	99.960%	
DQ168577	CRF02_AG	Recombinant	96.880%	
JN248580	CRF02_AG	Recombinant	99.992%	
KX389610	CRF02_AG	Recombinant	99.845%	
AB286857	CRF02_AG	Recombinant	99.732%	
AB286862	CRF02_AG	Recombinant	99.905%	

The CNN classified subtype U JF683772 as recombinant with a confidence level slightly above 75%. In the case of recombinant sequences from subtypes CRF02_AG and CRF63_02A6, all sequences were clearly classified as recombinant, with very high confidence levels (98%-99%). The hot zones corresponded clearly in all cases to those of recombinant sequences. Overall, the degree of similarity of the hot zones was mostly proportional to phylogenetic proximity.

In this segment of the tree, there were sequences where the level of similarity among hot zones was closely correlated with the degree of closeness in their parentage. This was evident in the case of the sister sequences CRF02_AG AB286857 and AB286862, as well as the sequences CRF63_02A6. However, it is important to

note that, in the case of CRF63_02A6 JN230353, the patterns of hot zones differed from its related CRF63_02A6 sequences.

The CNN classified the sequence JF683772 as recombinant with a confidence level exceeding 75%. The hot zones were consistent with the recombinant feature and similar to their recombinant relatives. The similarity level was coherent with the phylogenetic relationship. The explainability patterns of subtype CRF02_AG AF107770 were the most similar to JF683772 and to subtype CRF63_02A6 KJ197202, with visible hot zones at $f = 1/3$ and in the low frequency range of 5'LTR and possibly *gag*.

At this point, it is interesting to visualize the composition of the genome map of both CRFs. Figure 8.12 shows the genome map of subtype CRF02_AG.

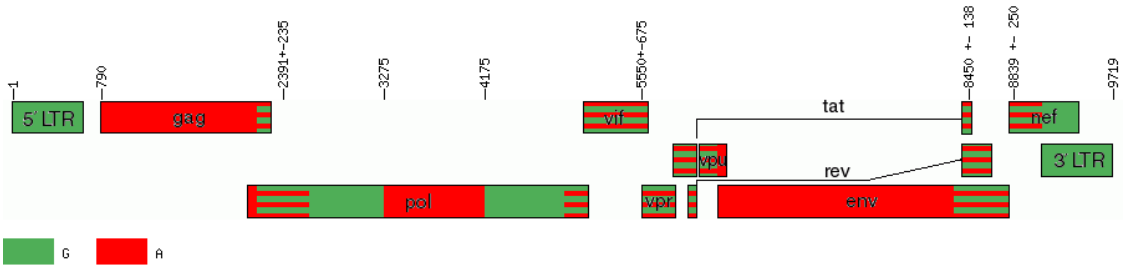


Figure 8.12: Genome Map of Recombinant Subtype CRF02_AG. This subtype is a mixture of subtypes A and G. Image sourced from the Los Alamos National Laboratory HIV-1 Database HIV Circulating Recombinant Forms (CRFs) [41].

The genome map of subtype CRF63_02A6 is shown in Figure 8.13.

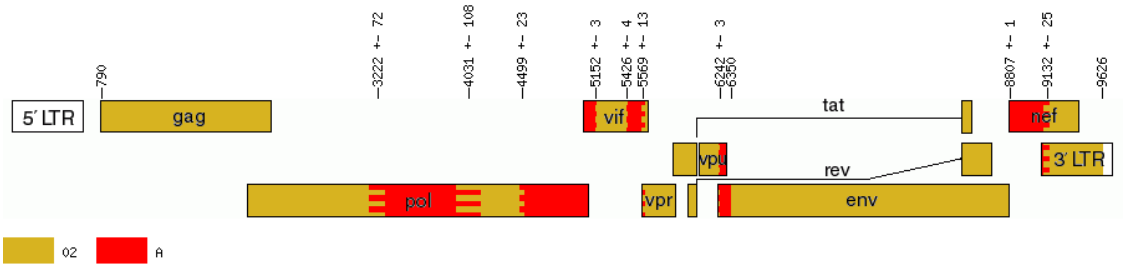


Figure 8.13: Genome Map of Recombinant Subtype CRF63_02A6. This subtype is a mixture of subtypes CRF02_AG and A (more specifically A6). Image sourced from the Los Alamos National Laboratory HIV-1 Database HIV Circulating Recombinant Forms (CRFs) [41].

The complexity of this subtype lies, among other factors, in that one of its parental subtypes is, in turn, another recombinant CRF of subtypes A and G.

The phylogenetic connection between CRF02_AG and CRF63_02A6 is consistent with the existence of closely related clades in the tree.

The subtype U sequence JF683772 shares its entire phylogenetic environment with recombinant sequences. This was consistent with the classification made by the CNN. Furthermore, the high degrees of similarity in the hot zones of the involved sequences reinforced this classification.

The presence of hidden recombination events in this sequence cannot be excluded. It would be necessary to analyze the genomes of these sequences in detail to elucidate the origin of the genomic regions of JF683772 and their degree of relationship with phylogenetically close sequences.

8.5.6 JX140679

In Figure 8.14, we can see the phylogenetic region where subtype U JX140679 is located. JX140679 shares clade with subtype 02GK AJ866557 and subtype 0206 KU168299, recombination between CRF02_AG (mixture of A and G) and CRF06_cpx (mosaic of A, G, J and K). From the point of view of recombinant feature, the parentage of the JX140679 clade is relatively close to recombinant sequences of multiple subtypes, both CRF and non-CRF, as well as sequences of subtype A1. Among these CRFs, subtype CRF02_AG stand out. And among the non-CRF recombinants, those related to CRF02_AG and pure subtypes A, A1 and G highlight, such as 02G, 02A1, 02A. Subtype CRF37_cpx itself is a recombinant mosaic among subtypes CRF01_AE, CRF02_AG, A, G and U.

Phylogenetically, this sequence is part of a relatively isolated clade, sharing a common ancestor with two non-CRF recombinant sequences whose non-recombinant subtypes include A, G, J, and K. This is a complex case study from a phylogenetic perspective. There is likely a lack of evolutionary information to delve deeper into this case study.

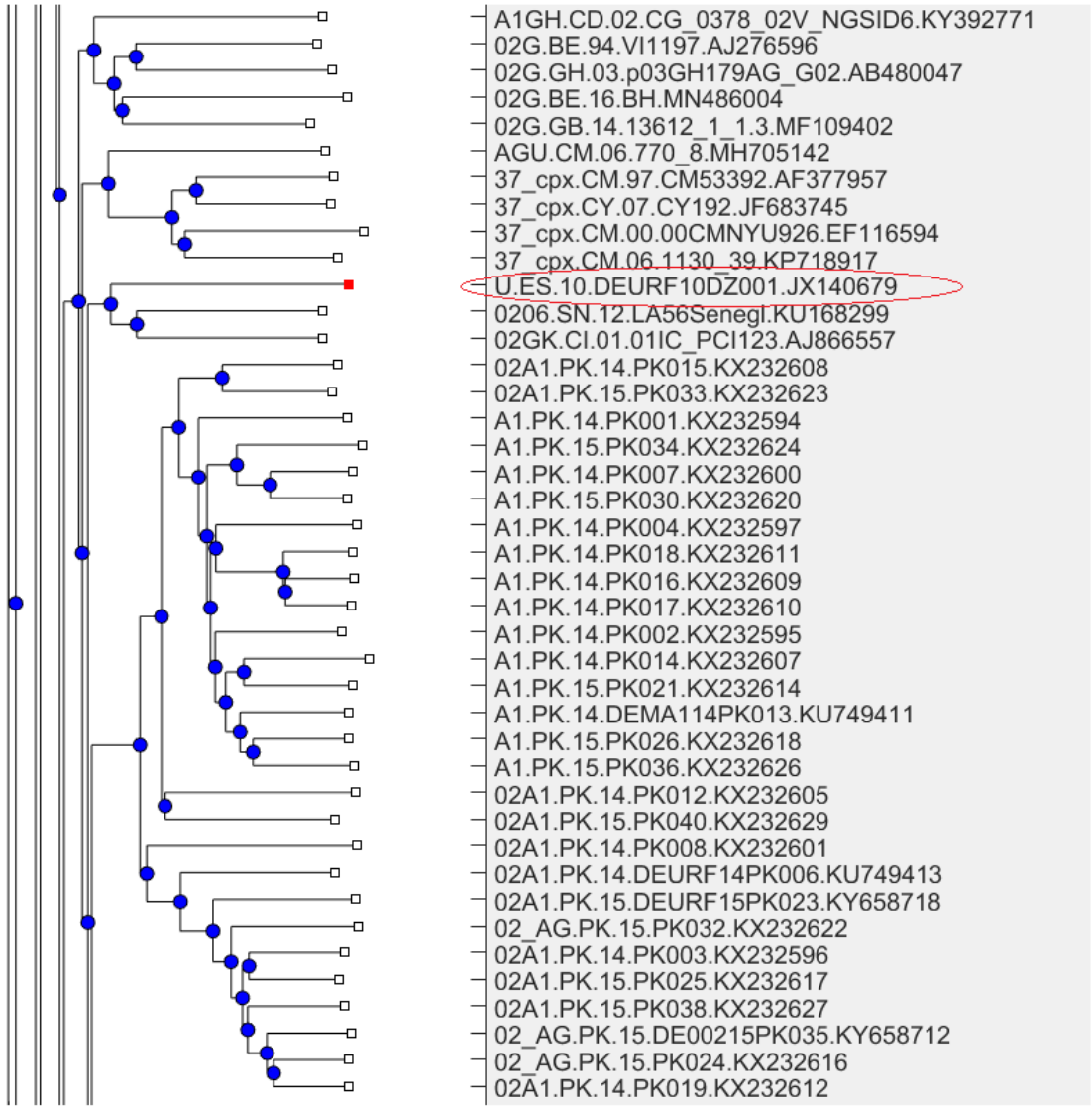


Figure 8.14: Partial Phylogenetic Tree Including Subtype U JX140679 Sequence

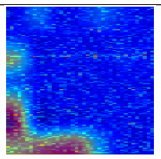
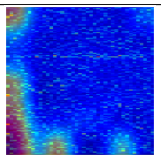
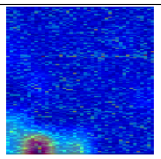
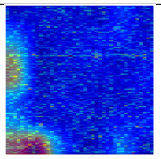
One of the clades closest to the JX140679 clade is composed of non-recombinant subtype A1 sequences and, to a lesser extent, recombinants of A1. Another nearby clade is entirely composed of recombinant sequences, mostly CRF37_cpx and one non-CRF AGU recombinant. There is also a nearby clade consisting entirely of recombinant sequences, CRF02_AG and non-CRF recombinants between CRF02_AG and A1. I approached this case study by analyzing the hot zones of the sequences in these clades to extract information.

The genome map of the CRF02_AG subtype is available in Figure 8.12.

Table 8.16 summarizes the classification results and hot zones of JX140679

(subtype U) and adjacent CRF02_AG subtype sequences. The CNN classified JX140679 as recombinant with a confidence level close to 78%. All CRF02_AG sequences were clearly classified as recombinant with an accuracy rate exceeding 99% in all cases. The explainability patterns were consistent with the recombinant featuring and, overall, were related to phylogenetic distance.

Table 8.16: JX140679 (Subtype U) and Subtype CRF02_AG Sequences Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
JX140679	U	Recombinant	77.945%	
KX232622	CRF02_AG	Recombinant	99.974%	
KY658712	CRF02_AG	Recombinant	99.527%	
KX232616	CRF02_AG	Recombinant	99.957%	

The classification of recombinant JX140679 was robust and exhibited interpretable patterns, closely resembling those found in the CRF02_AG sequences. This was also consistent with the two sequences that are "sisters" to JX140679, which are recombinants derived from subtype CRF02_AG.

The hot zones of KY658712 were similar to those of the CRF26_A5U sequence FM877777 (see table 8.12). The resemblance of hot zones to recombinants where one of their parental viruses is U, raises the need to investigate possible parental

relationships between these recombinant subtypes, beyond their evident connection through subtype A.

Table 8.17 shows the compilation of hot zones and classification results of A1 sequences that form a clade close to that of JX140679. All sequences were classified as non-recombinant with a high degree of confidence, except for KX232611, which was clearly classified as recombinant with a confidence level exceeding 96%. The other exception was KX232618, classified as recombinant but with a confidence level below 70%. In all sequences, the hot zones were consistent with those corresponding to their category. The high degree of similarity between the hot zones of sequences classified as non-recombinant was noteworthy.

Table 8.17: Subtype A1 Sequences Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

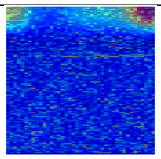
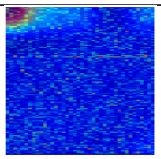
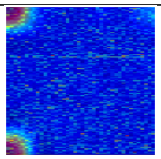
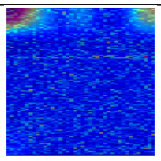
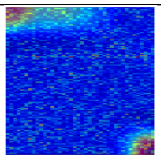
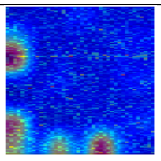
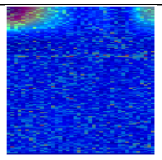
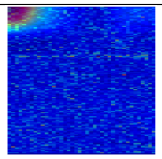
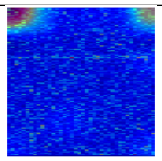
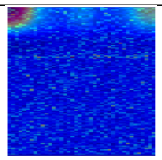
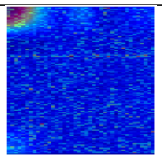
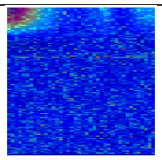
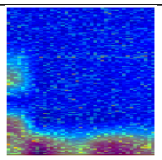
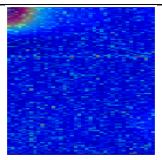
Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
KX232594	A1	Non-recombinant	88.062%	
KX232624	A1	Non-recombinant	99.239%	
KX232600	A1	Non-recombinant	98.494%	
KX232620	A1	Non-recombinant	99.969%	
KX232597	A1	Non-recombinant	99.701%	
KX232611	A1	Recombinant	96.440%	

Table 8.17 – Continuation from the previous page

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
KX232609	A1	Non-recombinant	90.666%	
KX232610	A1	Non-recombinant	97.200%	
KX232595	A1	Non-recombinant	82.135%	
KX232607	A1	Non-recombinant	93.285%	
KX232614	A1	Non-recombinant	84.851%	
KU749411	A1	Non-recombinant	87.714%	
KX232618	A1	Recombinant	69.297%	
KX232626	A1	Non-recombinant	93.556%	

The hot zones of subtype U JX140679 evidently differed from the hot zones of sequences correctly classified as non-recombinant. However, the degree of similarity, specially in the hot zones related to 5'LTR, with sequence KX232611 is striking, which was also classified by the CNN as recombinant with a very high degree of confidence. It is necessary to investigate this sequence currently classified as

A1 to elucidate whether hidden recombination phenomena exist that warrant a reconsideration of its current subtype, as well as the relationship between this specific sequence and recombinant subtypes with similar hot zones, such as CRF02_AG (see hot zones of recombinant sequences in Table 8.16).

The similarity with the other A1 sequence classified as recombinant (KX232618) was lower. Likewise, it would be interesting to determine if there are also hidden recombination events in this sequence.

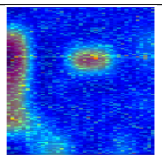
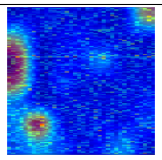
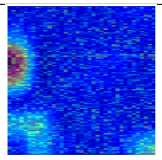
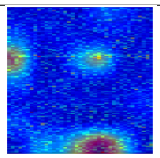
KX232611 explainability patterns differed from the rest of the subtype A1 sequences. This was because this sequence is classified as recombinant with a confidence level of 96.44%. Its patterns were similar to those of the CRF02_AG sequence KX232622. In the case of recombinant-classified KX232618, divergent hot zones could be related to its low confidence score (69.30%). Its patterns had a slight similarity with KX232622 (subtype CRF02_AG). Further search for hidden recombination phenomena in these sequences is needed.

Finally, I analyzed the hot zones of the CRF37_cpx sequences present in a phylogenetically close clade to that of JX140679 (Table 8.18). The CNN effectively classified all sequences as recombinant, mostly with a high degree of confidence, except for KP718917, which, although correctly classified, had a low trust level, slightly above 58%. The degree of similarity in the hot zones was very high, especially in sequences with a high confidence level. The hot zones of KP718917 were slightly divergent. All hot zones were consistent with those corresponding to their category, and their similarity levels corresponded to their phylogenetic proximity.

Explainability patterns were consistent with phylogenetic closeness and related with global recombinant hot zones. However, comparing with another recombinant subtypes, CRF37_cpx explainability tools had certain peculiarities. There was a hot point at $f = 1/3$ located around *vif* gene. This point did not appear in the rest of the recombinant subtypes included in this research. Interestingly, its location corresponded to U genomic areas.

The genome map of subtype CRF37_cpx is shown in Figure 8.15.

Table 8.18: Subtype CRF37_cpx Sequences Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
AF377957	CRF37_cpx	Recombinant	99.943%	
JF683745	CRF37_cpx	Recombinant	94.574%	
EF116594	CRF37_cpx	Recombinant	99.397%	
KP718917	CRF37_cpx	Recombinant	58.757%	

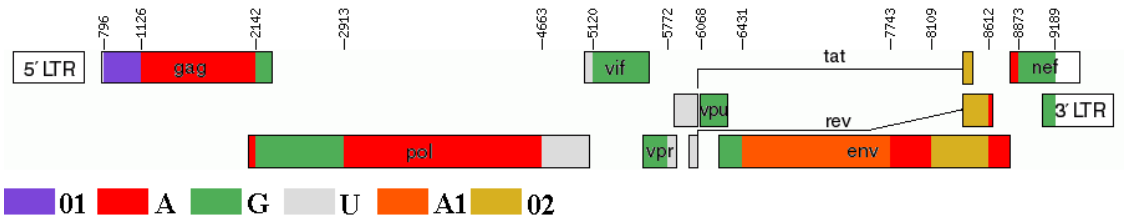


Figure 8.15: Genome Map of Recombinant Subtype CRF37_cpx. This subtype is a complex mosaic of subtypes CRF01_AE, A, G, U, A1, and CRF02_AG. Image sourced from the Los Alamos National Laboratory HIV-1 Database HIV Circulating Recombinant Forms (CRFs) [41].

The regions identified as U are located in the final region of *pol*, the beginning of *vif*, the final region of *vpr*, and the initial regions of *tat* and *rev*.

Exploring the relationship between CRF37_cpx and sequences classified as U, and precisely determining which U sequences share a higher degree of homology, would be interesting. This would help delve into the viral evolutionary mechanisms that affect these subtypes.

Since subtype CRF37_cpx is also a recombinant of CRF01_AE, I attach its genome map again to facilitate the analysis tracking.

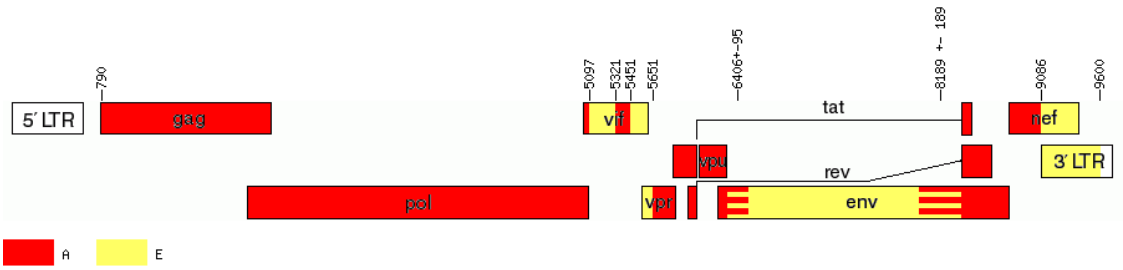


Figure 8.16: Genome Map of Recombinant Subtype CRF01_AE. This subtype is a mixture of subtypes A and E. Image sourced from the Los Alamos National Laboratory HIV-1 Database HIV Circulating Recombinant Forms (CRFs) [41].

This case study opens interesting avenues of research, not only regarding subtype U JX140679 itself but also concerning its phylogenetically close sequences, both recombinant and non-recombinant. The classification of JX140679 as recombinant was consistent with the phylogenetic relationships, both with the sequences it shares a clade with and those in adjacent clades. The coherence of the hot zones with those corresponding to the recombinant feature category was evident, and there was homology between the hot zones of JX140679 and phylogenetically close recombinant sequences.

This case study opens interesting avenues of research, not only regarding subtype U JX140679 but also concerning its phylogenetically close sequences, both recombinant and non-recombinant. The degrees of similarity in the complementary hot zones between recombinant subtypes and A1 sequences classified as recombinant, along with the composition of the genome map of the analyzed CRFs, raise the possibility of yet-to-be-determined familial relationships among them.

8.5.7 KP718920, CRF11_cpx, CRF13_cpx, and CRF27_cpx

Figure 8.17 shows the excerpt from the phylogenetic tree where the subtype U sequence KP718920 is located. KP718920 has a common ancestor with a clade including multiple subtype CRF11_cpx sequences (mosaic of subtypes A, E, G, J and U) and subtype 11A1 KR017778 sequence. The sister clade of this cluster contains two subclades. The first subclade clusters two subtype CRF27_cpx sequences (complex recombination of A, E, G, H, J, K and U). The second subclade includes 10 CRF13_cpx sequencesand one subtype 01A1G sequence.

There is a common ancestor among the sequence KP718920 and numerous CRF11_cpx sequences. The sister clade is also entirely composed of recombinant sequences.

Analyzing the hot zones and classification results in Table 8.19 allows to obtain more information about this case study.

The CNN classified KP718920 as non-recombinant with a confidence level close to 85%. Regarding the CRF11_cpx sequences, the majority were classified as recombinant with high confidence levels. An exception was AJ291718, classified as recombinant with an extremely low confidence level (around 52%), MF373206 with a confidence level just below 70%. Additionally, in the case of the CRF11_cpx sequence KP718929, the CNN classified it as non-recombinant, although with a low confidence level (around 59%). In the case of JF683802, the CNN clearly determined that this sequence was non-recombinant, with a confidence level of almost 99%. The hot zones of all involved sequences were consistent with the hot zones of their respective categories. In the case of CRF11_cpx sequences, there was a relationship between phylogenetic proximity and the degree of similarity in the hot zones, although the hot zones of sequences in this subtype exhibited notable variability compared to other analyzed CRFs.

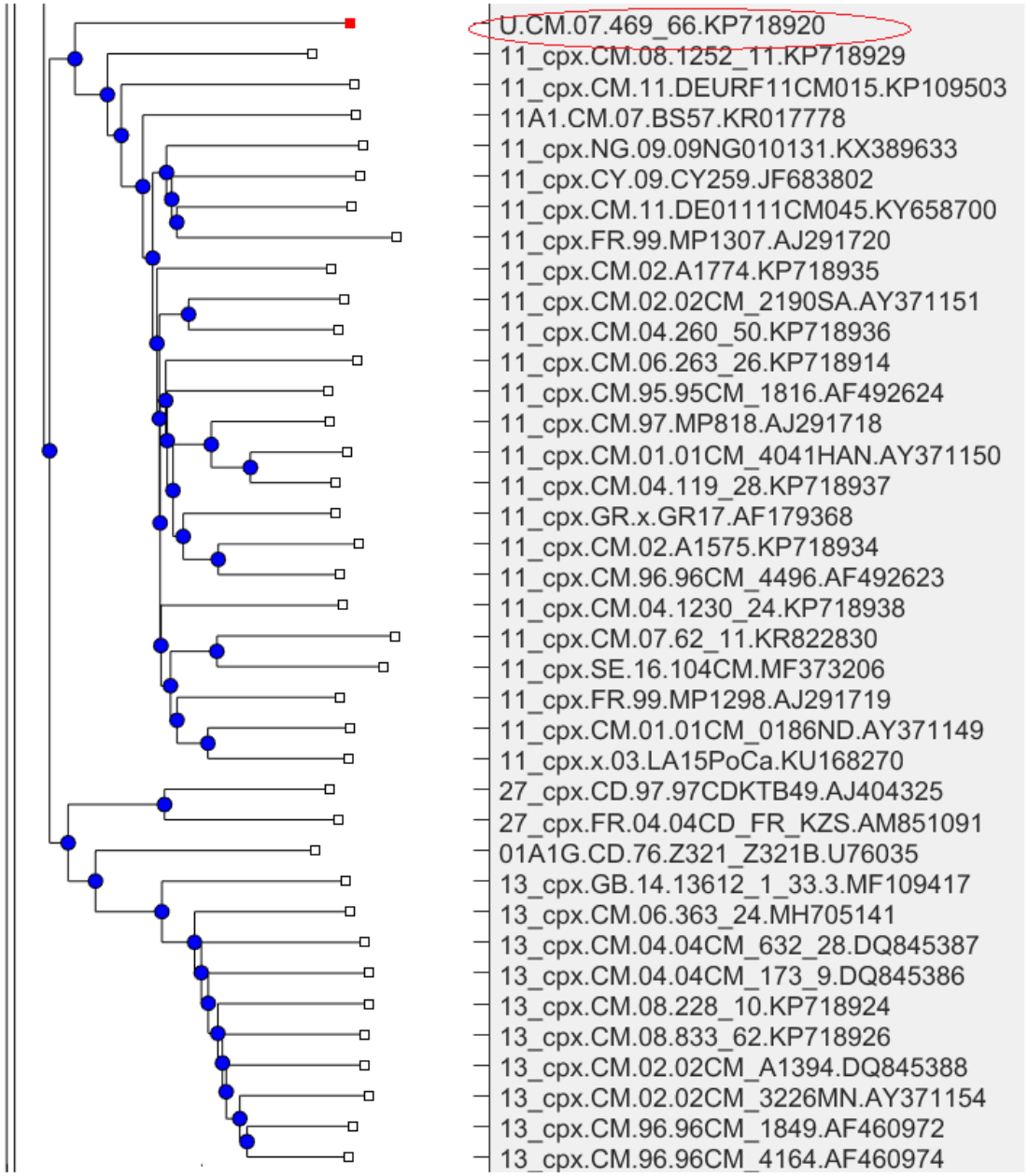


Figure 8.17: Partial Phylogenetic Tree Including Subtype U KP718920 Sequence

Table 8.19: KP718920 (Subtype U) and Subtype CRF11_cpx Sequences Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

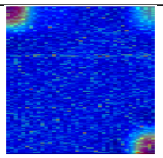
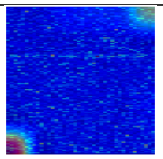
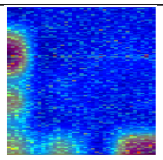
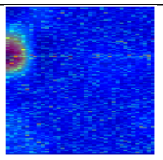
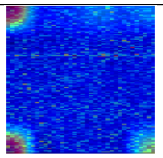
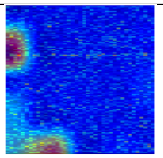
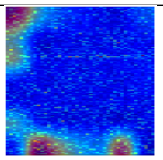
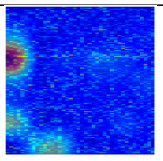
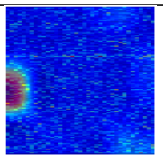
Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
KP718920	U	Non-recombinant	84.895%	
KP718929	CRF11_cpx	Non-recombinant	59.627%	
KP109503	CRF11_cpx	Recombinant	96.425%	
KX389633	CRF11_cpx	Recombinant	99.706%	
JF683802	CRF11_cpx	Non-recombinant	98.782%	
KY658700	CRF11_cpx	Recombinant	99.995%	
AJ291720	CRF11_cpx	Recombinant	93.096%	
KP718935	CRF11_cpx	Recombinant	95.160%	
AY371151	CRF11_cpx	Recombinant	99.686%	

Table 8.19 – Continuation from the previous page

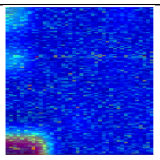
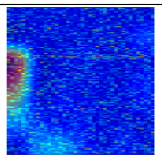
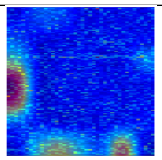
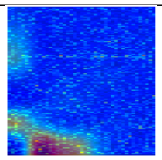
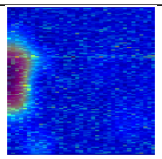
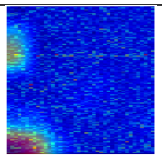
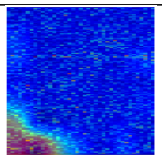
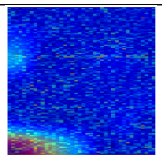
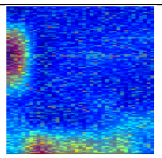
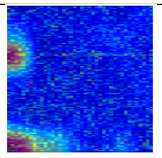
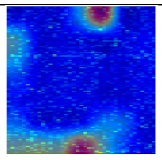
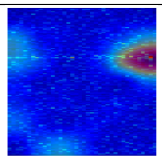
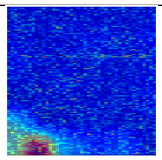
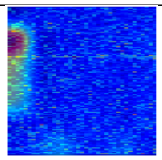
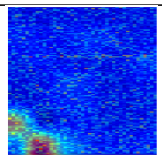
Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
KP718936	CRF11_cpx	Recombinant	98.496%	
KP718914	CRF11_cpx	Recombinant	99.799%	
AF492624	CRF11_cpx	Recombinant	99.352%	
AJ291718	CRF11_cpx	Recombinant	52.328%	
AY371150	CRF11_cpx	Recombinant	99.859%	
KP718937	CRF11_cpx	Recombinant	99.862%	
AF179368	CRF11_cpx	Recombinant	98.542%	
KP718934	CRF11_cpx	Recombinant	99.956%	
AF492623	CRF11_cpx	Recombinant	85.659%	
KP718938	CRF11_cpx	Recombinant	96.437%	

Table 8.19 – Continuation from the previous page

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
KR822830	CRF11_cpx	Recombinant	84.367%	
MF373206	CRF11_cpx	Recombinant	69.889%	
AJ291719	CRF11_cpx	Recombinant	98.856%	
AY371149	CRF11_cpx	Recombinant	92.801%	
KU168270	CRF11_cpx	Recombinant	91.791%	

The pairs of sister sequences whose explainability data are shown in Table 8.19 are:

- KY658700 and AJ291720.
- AY371151 and KP718936.
- AY371150 and KP718937.
- KP718934 and AF492623.
- KR822830 and MF373206.
- AY371149 and KU168270.

The CNN classified KP718920 as non-recombinant with a confidence level of 84.90%. Interestingly, the CNN also classified KP718929 as non-recombinant, with a low trust level (59.63%). This CRF11_cpx sequence is very close phylogenetically to subtype U KP718920 and, moreover, shows a phylogenetic disruption with the

rest of the CRF11_cpx sequences in its clade. These characterizations as non-recombinant were consistent with the phylogeny. All hot zones of all sequences in Table 8.19 were consistent with those of their respective categories, and there was a certain relationship between phylogenetic proximity and the degree of similarity in those relevant areas. However, in cases of sister sequences such as AY371151 and KP718936, AY371150 and KP718937, KR822830 and MF373206, or AY371149 and KU168270, more similarity was expected.

Consistency between the divergence of hot zones in CRF11_cpx KP718929 from those associated with the recombinant category supports the classification of this sequence as non-recombinant. JF683802 was similarly categorized as non-recombinant with a high confidence score (98.782%). Its hot zones exhibited resemblance to subtype U KP718920. These sequences might possess common non-recombinant attributes, notably more pronounced in JF683802 compared to KP718929.

The degree of similarity in the hot zones of subtype U KP718920 with the CRF11_cpx sequences KP718929 and JF683802 (classified as non-recombinant) may indicate the existence of common genomic elements diverging for some reason from the recombinant feature.

The variability of hot zones in the analyzed CRF11_cpx sequences is noteworthy. Perhaps this fact is related to the complex mosaic of genetic recombination events affecting this subtype (Figure 8.18).

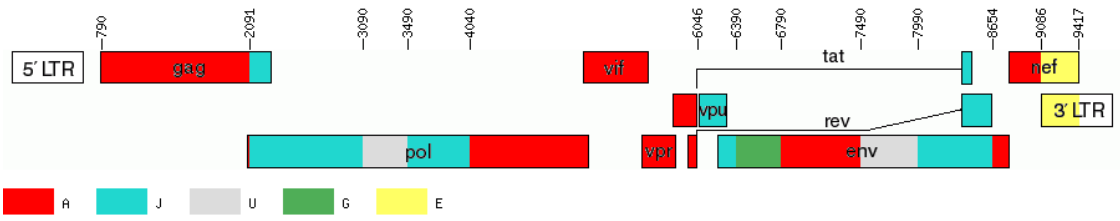


Figure 8.18: Genome Map of Recombinant Subtype CRF11_cpx. This subtype is a mosaic of subtypes A, J, U, G, and E. Image sourced from the Los Alamos National Laboratory HIV-1 Database HIV Circulating Recombinant Forms (CRFs) [41].

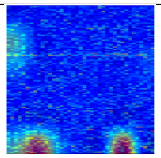
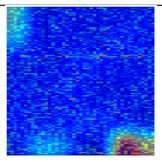
The zones identified as arising from U subtype are located in the central region of *pol* and *env*. It would be interesting to delve into the relationship between

subtype U and these U fragments identified in CRF11_cpx. It would also be interesting to determine the degree of homology of the genomes of the CRF11_cpx sequences classified as non-recombinant with that of the sequence KP718920 itself to elucidate differences between their genomes and the rest of the recombinant sequences in their clade. This way, we can gather more information about hidden familial relationships that may exist with subtype U.

The phylogenetically closest clade has an evolutionary divergence resulting in a subclade formed by two CRF27_cpx sequences and another subclade formed by several CRF13_cpx sequences. Table 8.20 shows the hot zones and classification results of these sister sequences CRF27_cpx.

Both sequences were classified as recombinant with a confidence level exceeding 99%. Their hot zones were notably similar, although perhaps to a lesser extent than expected, given their phylogenetic proximity. Similarly, while their hot zones were consistent with the recombinant category, they were slightly different from the more common hot zones in recombinant sequences.

Table 8.20: Subtype CRF27_cpx Sequences Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
AJ404325	CRF27_cpx	Recombinant	99.275%	
AM851091	CRF27_cpx	Recombinant	99.769%	

AJ404325 and AM851091 are sister sequences clearly classified as recombinant, with a confidence level exceeding 99%. Their closest phylogenetic distance was consistent with their highly similar explainability patterns, although a higher degree

of resemblance was expected. The hot zones, especially for AJ404325, resembled the hot zones of the CRF11_cpx sequences AJ291720 and, to a lesser extent), AF492624. Their hot zones were also similar to some of their CRF13_cpx-related sequences in the clade, such as MF109417, DQ845387, or DQ845386 (Table 8.21).

Table 8.21: Subtype CRF13_cpx Sequences Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

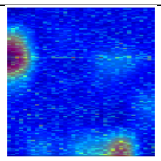
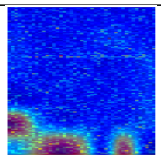
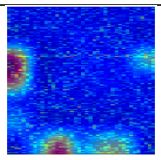
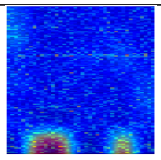
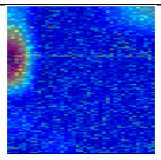
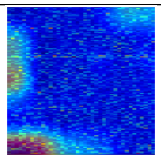
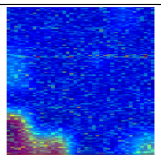
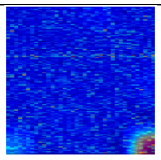
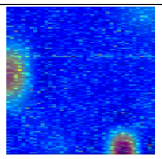
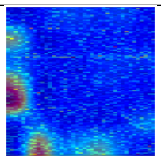
Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
MF109417	CRF13_cpx	Recombinant	51.066%	
MH705141	CRF13_cpx	Recombinant	99.350%	
DQ845387	CRF13_cpx	Recombinant	70.371%	
DQ845386	CRF13_cpx	Recombinant	98.412%	
KP718924	CRF13_cpx	Recombinant	99.760%	
KP718926	CRF13_cpx	Recombinant	96.912%	
DQ845388	CRF13_cpx	Recombinant	93.773%	

Table 8.21 – Continuation from the previous page

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
AY371154	CRF13_cpx	Non-recombinant	54.231%	
AF460972	CRF13_cpx	Recombinant	76.672%	
AF460974	CRF13_cpx	Recombinant	63.107%	

The CNN classified almost all sequences as recombinant with high confidence levels. The first exception was AY371154, which the CNN classified as non-recombinant with a very low confidence level (around 54%). The other exception was AF460974. In this case, the CNN classifies it as recombinant but with a trust level of approximately 63%. The hot zones were consistent with their corresponding category according to the classification by the CNN, and the degree of similarity aligned with the relatedness of the sequences.

AY371154 explainability divergence was due to its non-recombinant classification (low confidence score of 54.23%). AY371154 hot zones were similar to KP718920 (also a non-recombinant classified sequence).

Furthermore, similarities were detected among the hot zones of the different CRFs in this excerpt from the tree. See the example of CRF13_cpx KP718926 with the CRF11_cpx sequence AJ291718.

These high similarity levels among phylogenetically close sequences but belonging to different CRFs might be due to the existence of common genomic elements that might be detected in some manner by the CNN.

The genome map of subtype CRF13_cpx is shown in Figure 8.19.

The subtype CRF13_cpx is a subtype with complex genetic recombination phenomena among various subtypes. Presence of its U subtype is detected in *vpu*

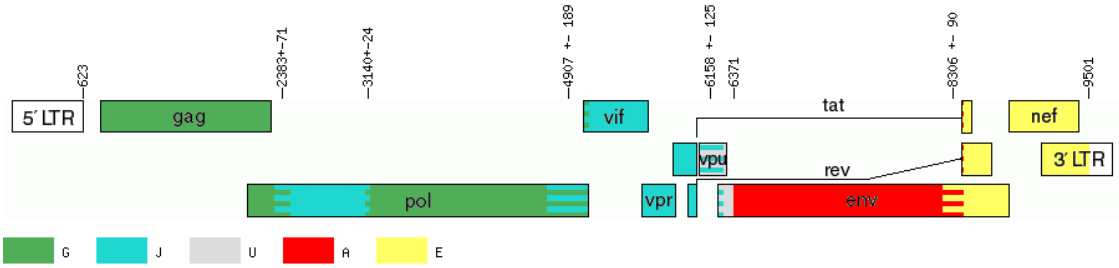


Figure 8.19: Genome Map of Recombinant Subtype CRF13_cpx. This subtype is a mosaic of non-recombinant subtypes G, J, U, A, and E. Image sourced from the Los Alamos National Laboratory HIV-1 Database HIV Circulating Recombinant Forms (CRFs) [41].

and the start of *env*. It would be interesting to delve into the relationship or degree of homology between the subtype U sequence KP718920 and these genomic regions in phylogenetically close CRF13_cpx sequences to deepen our understanding of these genetic recombination phenomena.

In this case study, curious phenomena occurred regarding the CNN's behavior with the recombinant sequences that are phylogenetically surrounding KP718920. The CNN's misclassifications in these CRF sequences led to hot zones similar to those of the sequence KP718920 itself, which deserve further study.

The exploration of genomic elements related to the non-recombinant feature detected by the CNN is meaningful.

8.5.8 MN153489, A6 and A1

The fragment of the phylogenetic tree containing MN153489 (Subtype U) is shown in Figure 8.20.

Its phylogenetically closest sequence is subtype AU MN736709 (unique sister sequence). A clade positioned close to this one (see upper clade) is in turn divided into two subclades. Its first subclade contains mainly subtype A1 sequences. Its second subclade is mainly composed of subtype A6 sequences. Both blocks include certain recombinant sequences. Another phylogenetically distant clade (see lower clade to that containing MN153489) is primarily composed of non-CRF recombinants. The majority of recombinant sequences in this clade are A1D.

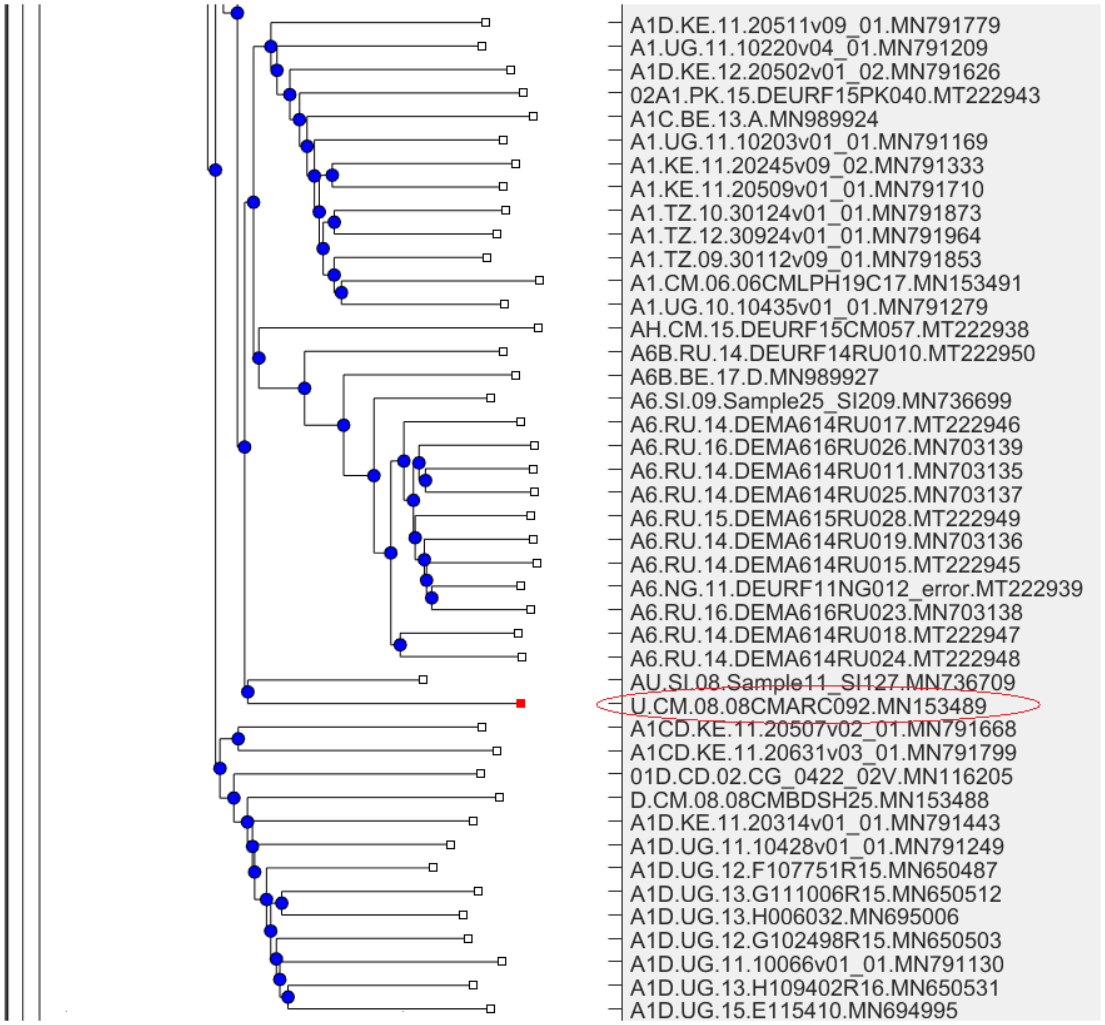


Figure 8.20: Partial Phylogenetic Tree Including Subtype U MN153489 Sequence

Table 8.22 shows the hot zones and classification results of MN153489.

The CNN classified MN153489 as non-recombinant. Its explainability patterns were coincident with global non-recombinant hot zones. MN153489 determinant areas were notably similar with subtype A1 MN791169 sequence ones. This sequence was probably the most similar descendant in terms of explainability.

Table 8.23 shows the classification results of the A1 sequences from the clade close to MN153489. The CNN classified all sequences as non-recombinant with confidence levels exceeding 99%. The hot zones were consistent with those corresponding to their category and showed a high degree of similarity among them.

Table 8.22: MN153489 (Subtype U) Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

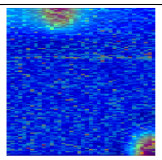
Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
MN153489	U	Non-recombinant	93.586%	

Table 8.23: Subtype A1 Sequences Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

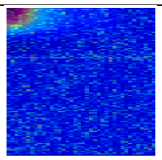
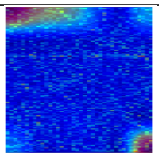
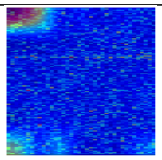
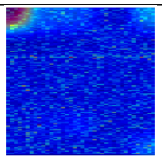
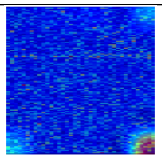
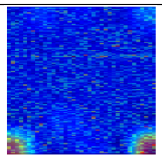
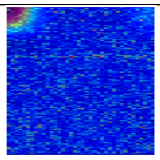
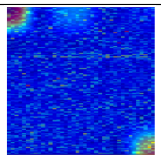
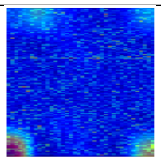
Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
MN791209	A1	Non-recombinant	99.703%	
MN791169	A1	Non-recombinant	99.987%	
MN791333	A1	Non-recombinant	99.864%	
MN791710	A1	Non-recombinant	99.580%	
MN791873	A1	Non-recombinant	99.975%	
MN791964	A1	Non-recombinant	99.852%	

Table 8.23 – Continuation from the previous page

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
MN791853	A1	Non-recombinant	99.926%	
MN153491	A1	Non-recombinant	95.920%	
MN791279	A1	Non-recombinant	99.817%	

MN791169 classification as non-recombinant had a confidence score extremely high (99.987%). Its hot zones were quite similar to MN153489 ones so these two sequences might have common non-recombinant featuring components. In general terms, A1 explainability patterns assimilated to MN153489 hot zones.

As for the A6 sequences, the results are compiled in Table 8.24. All sequences were classified by the CNN as non-recombinant with confidence levels exceeding 90%. There were two exceptions. The first one was MN703138, classified as recombinant with a low confidence level (around 56.50%). The second one was MN703137, which, although classified as non-recombinant, had a slightly lower confidence level than the majority of cases, around 81.50%. The hot zones were generally consistent with those corresponding to their categories.

Table 8.24: Subtype A6 Sequences Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

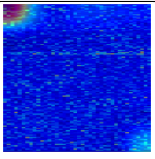
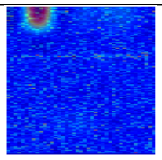
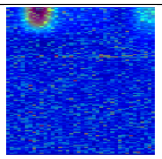
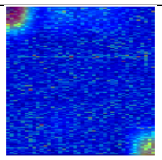
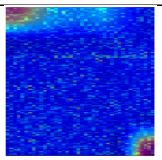
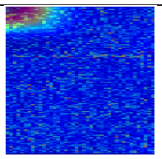
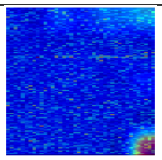
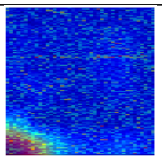
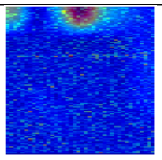
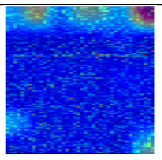
Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
MN736699	A6	Non-recombinant	94.463%	

Table 8.24 – Continuation from the previous page

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
MT222946	A6	Non-recombinant	99.926%	
MN703139	A6	Non-recombinant	94.302%	
MN703135	A6	Non-recombinant	99.075%	
MN703137	A6	Non-recombinant	81.689%	
MT222949	A6	Non-recombinant	94.863%	
MN703136	A6	Non-recombinant	91.241%	
MN703138	A6	Recombinant	56.614%	
MT222947	A6	Non-recombinant	99.856%	
MT222948	A6	Non-recombinant	99.002%	

Explainability patterns were homogeneous and consistent with those of non-recombinant sequences.

The degree of similarity of the hot zones was consistent with phylogenetic proximity. An example of this was the case of the A6 sequences MT222946 and MN703139. Likewise, MN703135 and MN703137, and to a lesser extent, MN736699.

MN703138 hot zones were the most dissenting patterns. This sequence was classified as recombinant (low confidence score of 56.61%). This questionable sequence requires a detailed analysis of its genome. Sequences such as MN703139 showed very similar patterns with the subtype U sequence MN153489.

The classification of subtype U MN153489 as non-recombinant was apparently consistent with the close phylogenetic relationships visualized in the tree. Likewise, the degree of similarity of the hot zones between this sequence and its close A1 and A6 sequences was high, as well as between the A1 and A6 sequences themselves. This homology in the hot zones of subtypes A1 and A6 was correlated with their inherent phylogenetic closeness, as both are grouped within subtype A.

It is interesting to research the possible relationships of kinship between MN153489 and both subtypes.

8.5.9 MN153492, F1, CRF12_BF, and F2

The excerpt from the phylogenetic tree where subtype U MN153492 is located is shown in Figure 8.21. MN153492 closest sequence is MN153476 (unique sister). MN153476 is identified as subtype 01F2, recombinant between CRF01_AE (mixture between subtypes A and E) and F2. The cluster composed of these two sister sequences has a common ancestor with a clade containing two subtype F2 sequences. The sister clade of this cluster contains subtype F1 sequences and F1-related recombinants such as BF1 (a mixture of B and F1), CRF12_BF (recombinant between B and F), 01F1G (a mixture of CRF01_AE, F1, and G). This tree branch is strongly influenced by subtype F and its sub-subtypes.

The classifications, confidence levels, and hot zones of MN153492 and its phylogenetically closest sequences are grouped in Table 8.25. Subtype U MN153492 was classified by CNN as non-recombinant with a confidence level exceeding 90%, and its hot zones were consistent with this category. As for the F2 sequences, both were

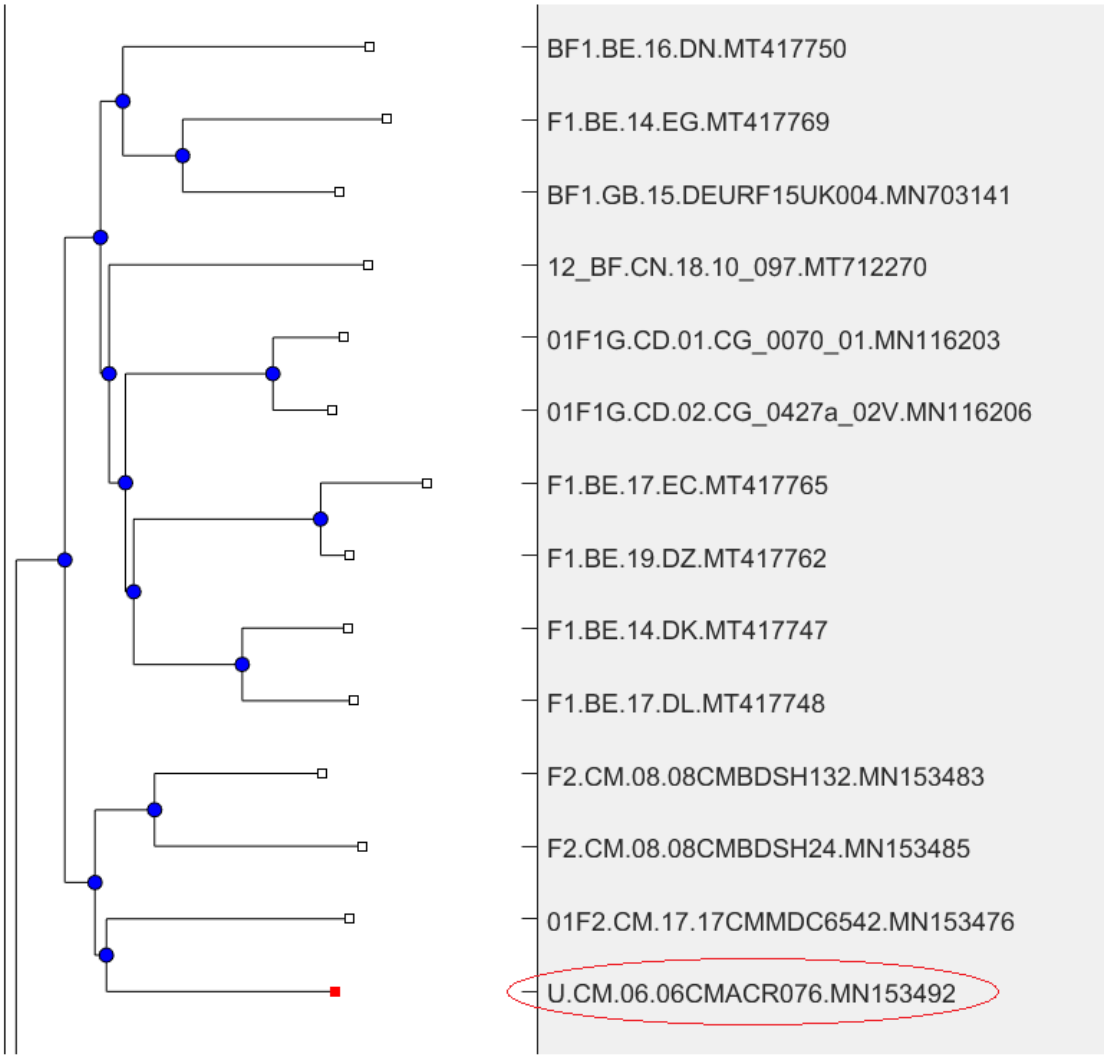
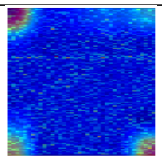
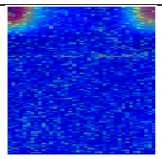
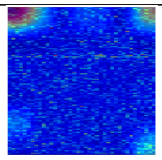
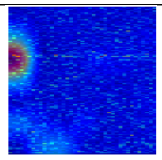
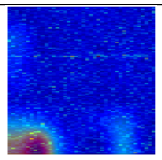
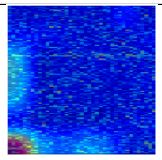
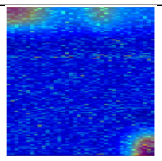
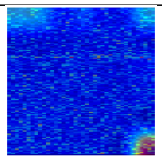
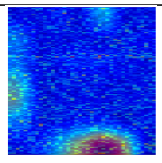


Figure 8.21: Partial Phylogenetic Tree Including Subtype U MN153492 Sequence

classified as non-recombinant, with slightly different confidence levels, almost 96% in the case of MN153483 and slightly above 84% in the case of MN153485. Nevertheless, the hot zones were similar and consistent with their non-recombinant category. The CRF12_BF sequence was classified as recombinant with a confidence level exceeding 99.50% and hot zones in line with the recombinant feature. Regarding the F2 sequences, interestingly, the majority were incorrectly classified as recombinant. Furthermore, their hot zones were diverse.

Table 8.25: MN153492 (Subtype U) and Subtypes F2, F1 and CRF12_BF Sequences Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
MN153492	U	Non-recombinant	91.835%	
MN153483	F2	Non-recombinant	95.832%	
MN153485	F2	Non-recombinant	84.152%	
MT712270	CRF12_BF	Recombinant	99.571%	
MT417769	F1	Recombinant	98.458%	
MT417765	F1	Recombinant	81.698%	
MT417762	F1	Non-recombinant	68.153%	
MT417747	F1	Non-recombinant	88.939%	
MT417748	F1	Recombinant	85.254%	

Regarding the F2 sequences, the disparity between phylogeny and the degree of similarity in hot zones, as well as the decisions made by the CNN, was noteworthy.Sister sequences, as in the case of F2 MT417747 and MT417748, are expected to have similar classifications and hot zones. However, MT417747 was classified as non-recombinant, while MT417748 was incorrectly classified as recombinant. Their hot zones also differed, although they were consistent with their respective categories. The same phenomenon occurred in the case of MT417765 with MT417762.

In the case of MT417769, its sister sequence is a non-CRF BF1 recombinant, meaning a recombinant between subtype B and F1. Both have a common ancestor with another BF1 sequence. Perhaps the CNN's classification of MT417769 as recombinant indicates hidden recombination events in this sequence that make it phylogenetically close to these BF1 sequences. This point requires further investigation.

The high error rate in the classification of this group of F1 sequences could indicate hidden genetic recombination phenomena in this subtype.

Subtype F2 MN153485 explainability patterns were the most similar to MN153492 ones. MT712270 hot zones were coherent with global recombinant ones. Subtype F1 MT417769, MT417765 and MT417748 were classified as recombinant. Consequently, their explainability patterns were diffuse and could be assimilated to the recombinant category. There were some diffuse non-recombinant sequences whose trust levels were moderate. These sequences and their confidence rates were subtype F1 MT417762 (68.15%) and MT417747 (88.94%), and subtype F2 MN153485 (84.15%). MN153483 non-recombinant feature was stronger and their hot zones were more coincident with global non-recombinant explainability patterns.

In view of the genome map of subtype CRF12_BF (Figure 8.22), it is indeed a recombinant between subtypes B and F. Therefore, there is a direct relationship between this CRF and F1 and F2 sequences.

Considering the phylogeny and CNN results, there were certain ambiguities in terms of classification and the degree of similarity between hot zones. Several striking phenomena was detected that, while not providing conclusive results, offer some insights into phylogenetic relationships and possible related evolutionary divergences.

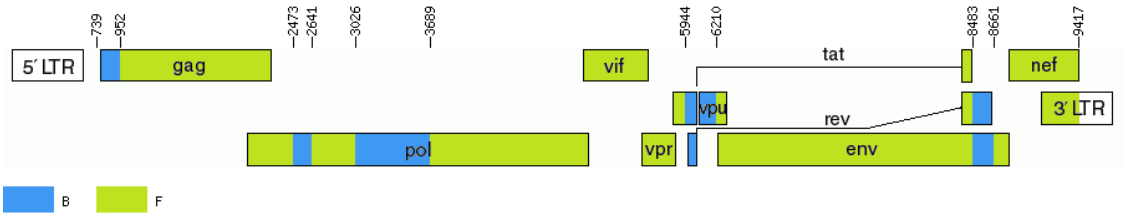


Figure 8.22: Genome Map of Recombinant Subtype CRF12_BF. This subtype is a mixture of subtype B and F. Image sourced from the Los Alamos National Laboratory HIV-1 Database HIV Circulating Recombinant Forms (CRFs) [41].

All of this makes it an especially interesting case study that requires a deeper exploration of the possible relationships between MN153492 and subtypes F1 and F2. It is also interesting to search for potential hidden genetic recombination phenomena in the F1 sequences classified by the CNN as recombinant.

8.5.10 MN736701 and MN736702

The excerpt from the phylogenetic tree where the subtype U sequences MN736701 and MN736702 are located is shown in Figure 8.23. Most of the sequences in this area are recombinants of multiple types. There is a common ancestor between MN736702 and a complex clade that includes several recombinant subtypes: 02G, CRF20_BG, A1BG, 0102A1 (recombinant among CRF01_AE, CRF02_AG and A1), A1BG, 0102A1, CRF2201A1, 0102, CRF02_AG, 02A, 02A1. This clade also includes a non-recombinant sequence of subtype G: MN153484. MN736701 belongs to a clade of heterogeneous recombinant sequences. MN736701 has a unique sister sequence, subtype CRF27_cpx MN116204. The four remaining sequences of this clade belong to subtypes 06U, CRF06_cpx, 0218 and CRF18_cpx. MN736701 and MN736702 belong to sister clades.

The results of the CNN classification and the hot zones for subtype U sequences MN736701 and MN736702 are shown in Table 8.26. The CNN classified both sequences as recombinant. In the case of MN736701, the confidence level was higher, close to 74.50%. However, for the sequence MN736702, the confidence level was

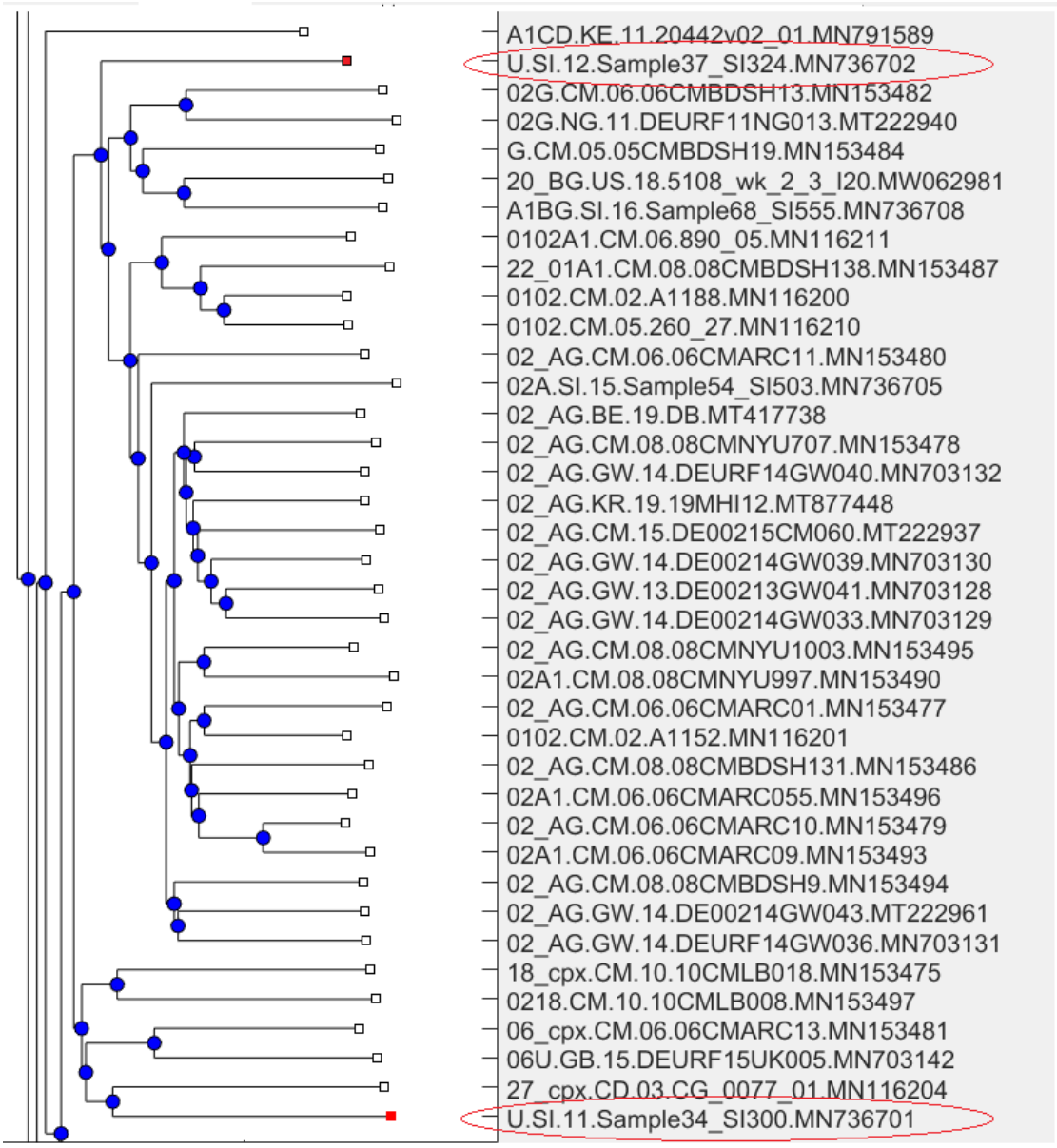
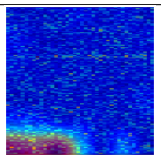
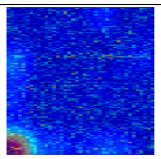


Figure 8.23: Partial Phylogenetic Tree Including Subtype U MN736701 and MN736702 Sequences

very low, only around 53%. The hot zones were similar and consistent with the overall hot zones of the recombinant category.

MN736701 and MN736702 were classified as recombinant and belong to sister clades. Their explainability patterns were similar even though the low confidence score of MN736702 recombinant classification (52.86%). Subtype CRF02_AG sequences MN703132 and MT222961 (Table 8.29), and subtype CRF20_BG MW062981 (Table 8.27) had the most similar explainability patterns with MN736701 and

Table 8.26: MN736701 and MN736702 (Subtype U) Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
MN736701	U	Recombinant	74.420%	
MN736702	U	Recombinant	52.857%	

MN736702.

Table 8.27 shows the classification results of sequences phylogenetically closest to MN736702. The CNN classified subtype G MN153484 as non-recombinant with a high confidence level, exceeding 98.50%. Its hot zones were consistent with the non-recombinant feature. Subtype CRF20_BG MW062981 and and subtype CRF22_01A1 MN153487 were also correctly classified as recombinant with high trust levels. Their hot zones were in line with the recombinant feature and showed a high degree of similarity.

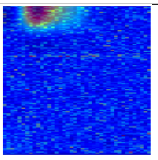
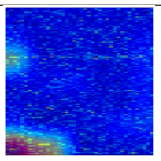
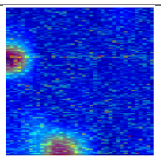
Despite belonging to different subtypes, the level of similarity between the hot zones of the sequence CRF20_BG MW062981 and the sequence CRF22_01A1 MN153487 correlated with their phylogenetic proximity.

Remarkably, there was a high resemblance between the sequence CRF20_BG MW069281 and that of subtype U MN736701, even though the phylogenetic distance is greater compared to MN736702. It would be necessary to delve deeper into understanding the relationships of kinship between these two sequences.

Figure 8.24 displays the genome map of subtype CRF20_BG.

This genome map justifies the composition of the clade where MN736702 is located, consisting of a single G sequence and recombinants derived from this subtype. The absence of more G sequences in this tree excerpt is also noteworthy.

Table 8.27: Subtypes G, CRF20_BG, CRF22_01A1 Sequences Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
MN153484	G	Non-recombinant	98.570%	
MW062981	CRF20_BG	Recombinant	99.958%	
MN153487	CRF22_01A1	Recombinant	98.743%	

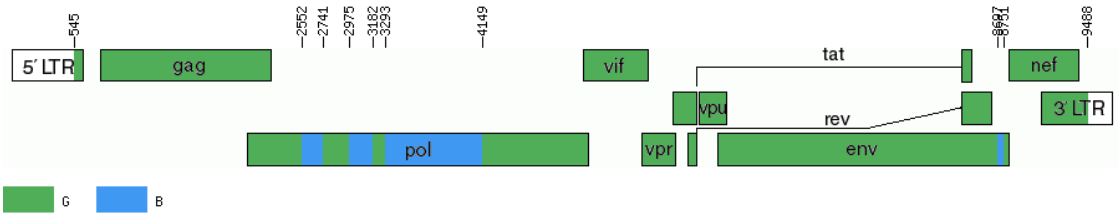


Figure 8.24: Genome Map of Recombinant Subtype CRF20_BG. This subtype is a mixture of subtypes B and G. Image sourced from the Los Alamos National Laboratory HIV-1 Database HIV Circulating Recombinant Forms (CRFs) [41].

A deep genomic analysis of this G sequence MN153484 is interesting at this point to determine unknown phylogenetic relationships between this sequence and the U sequences MN736701 and MN736702. Based on the hot zones, no similarities are observed that provide additional information between them.

The genome map of subtype CRF22_01A1 is shown in Figure 8.25.

In this case, the presence in the tree extract of the sequence CRF22_01A1 is consistent with the fact that in that part of the clade, several sequences related to subtypes CRF01_AE and subtype A1 are clustered. It would be interesting

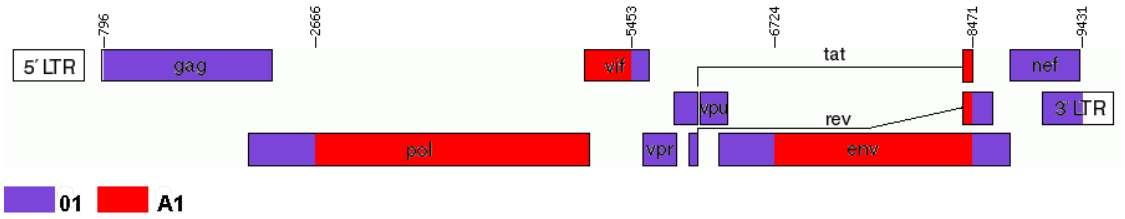


Figure 8.25: Genome Map of Recombinant Subtype CRF22_01A1. This subtype is a mixture of subtypes CRF01_AE and A1. Image sourced from the Los Alamos National Laboratory HIV-1 Database HIV Circulating Recombinant Forms (CRFs) [41].

to delve into possible hidden parental relationships and the evolutionary divergences that caused them between these A1 and CRF01_AE subtypes with the G subtypes and derivatives present in the other clade region. As well as with the U sequence MN736702.

Regarding the similarity degrees of the hot zones, the high concordance between the hot zones of CRF20_BG MW062981 and CRF22_01A1 MN153487 was striking. This is noteworthy, especially considering that they are two different subtypes with parental viruses belonging to subtypes that are not initially closely related. Further research is needed.

Regarding the other U subtype sequence involved in this case study, Table 8.28 shows the classification results of sequences phylogenetically closest to MN736701. The CNN correctly classified all three sequences as recombinant with confidence levels exceeding 99% in all cases. Their hot zones were similar and fit perfectly into those corresponding to the recombinant category.

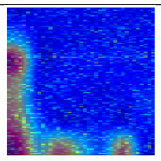
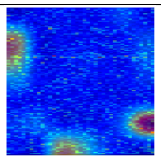
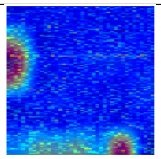
The resemblance level in their hot zones was high and directly proportional to the phylogenetic closeness between these sequences.

Similarly, there was a certain degree of similarity with the hot zones of the CRF22_01A1 sequence MN153487.

MW062981 (subtype CRF20_BG) had the most similar patterns with explainability hot zones of MN736701 and MN736702.

Although MN116204 (subtype CRF27_cpx) is the phylogenetically closest sequence to MN736701, their similarity patterns were substantially different. MN116204

Table 8.28: Subtypes CRF18_cpx, CRF06_cpx, CRF27_cpx Sequences Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
MN153475	CRF18_cpx	Recombinant	99.120%	
MN153481	CRF06_cpx	Recombinant	99.772%	
MN116204	CRF27_cpx	Recombinant	99.750%	

trust level was 99.75% and MN736701 was only 74.42%. The sequence phylogenetically closest to MN736702 is MW062981 (subtype CRF20_BG). In this case there was more coherence.

Figure 8.26 illustrates the genome map of subtype CRF06_cpx.

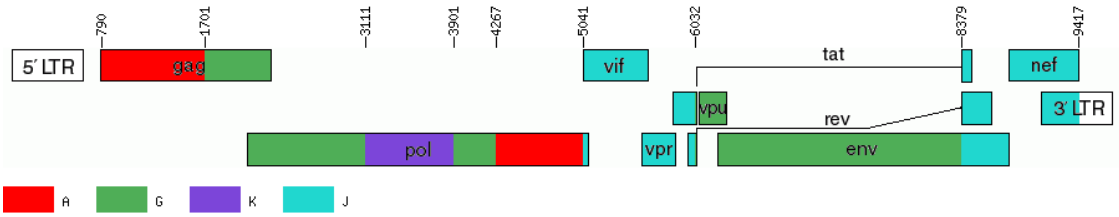


Figure 8.26: Genome Map of Recombinant Subtype CRF06_cpx. This subtype is a mosaic of subtypes A, G, K, and J. Image sourced from the Los Alamos National Laboratory HIV-1 Database HIV Circulating Recombinant Forms (CRFs) [41].

Subtypes A and G are parental to this subtype, which partly explains both the presence of this sequence in this area of the tree and the high similarity among the hot zones of the involved sequences, regardless of their belonging to different CRFs.

I provide the genome map of subtype CRF27_cpx in Figure 8.27.

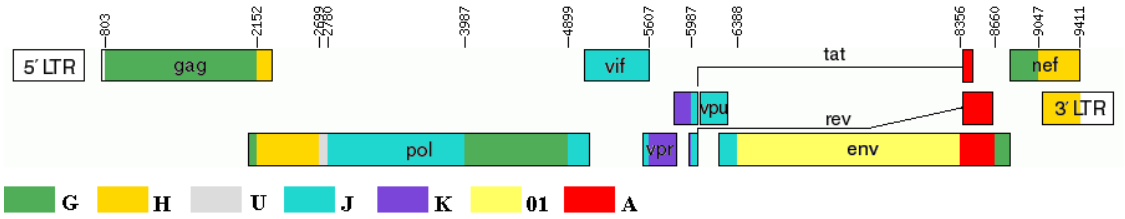


Figure 8.27: Genome Map of Recombinant Subtype CRF27_cpx. This subtype is a mosaic of G, H, U, J, K, CRF01_AE, and A. Image sourced from the Los Alamos National Laboratory HIV-1 Database HIV Circulating Recombinant Forms (CRFs) [41].

This recombinant subtype is composed of multiple genetic recombination events from various subtypes, including once again subtypes A and G. Additionally, there is a small fragment in the first half of the *pol* gene identified as subtype U. A detailed analysis of the genomes of CRF27_cpx MN116204 and subtype U MN736701 might provide additional insights into the phylogenetic relationship between both sequences and unknown evolutionary divergences among sequences in this area of the phylogenetic tree.

The genome map of CRF18_cpx, available in Figure 8.4, is a mosaic of subtypes A, G, U, F, H, and K. The genomic regions identified as U are located in three different fragments along the *pol* gene.

To delve deeper into the parental relationships between these mosaic CRF subtypes and the U sequences MN736701 and MN736702, it is interesting to analyze the degrees of homology between the U genomic regions of these recombinant subtypes and their counterparts in the U sequences.

In the clade where subtype U MN736702 is located, there is a predominant presence of CRF02_AG sequences, as well as non-CRF recombinants between CRF02_AG and A1. Table 8.29 compiles the CNN classification results and the hot zones of each intervening sequence. The CNN correctly classified all sequences as recombinants with a very high degree of confidence. The hot zones were consistent with the recombinant feature of this subtype, and their level of similarity aligned with the phylogenetic distance.

Table 8.29: Subtype CRF02_AG Sequences Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

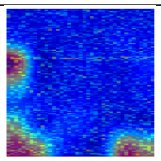
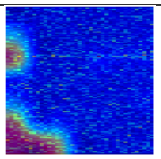
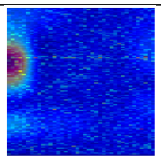
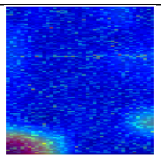
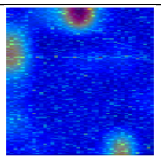
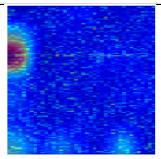
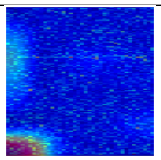
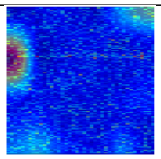
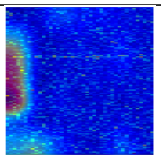
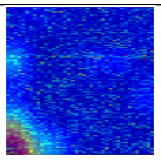
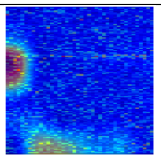
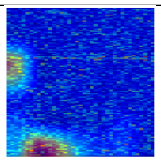
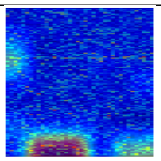
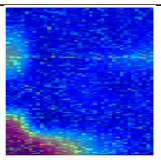
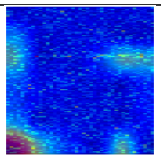
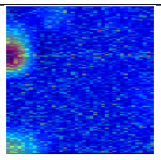
Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
MN153480	CRF02_AG	Recombinant	99.960%	
MT417738	CRF02_AG	Recombinant	98.258%	
MN153478	CRF02_AG	Recombinant	99.934%	
MN703132	CRF02_AG	Recombinant	99.799%	
MT877448	CRF02_AG	Recombinant	99.914%	
MT222937	CRF02_AG	Recombinant	99.623%	
MN703130	CRF02_AG	Recombinant	99.573%	
MN703128	CRF02_AG	Recombinant	99.644%	
MN703129	CRF02_AG	Recombinant	99.931%	

Table 8.29 – Continuation from the previous page

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
MN153495	CRF02_AG	Recombinant	99.134%	
MN153477	CRF02_AG	Recombinant	99.926%	
MN153486	CRF02_AG	Recombinant	97.607%	
MN153479	CRF02_AG	Recombinant	99.590%	
MN153494	CRF02_AG	Recombinant	97.899%	
MT222961	CRF02_AG	Recombinant	99.834%	
MN703131	CRF02_AG	Recombinant	99.991%	

Their hot zones were aligned with their recombinant feature. And there was a high correlation between phylogenetic closeness and the similarity level of their explainability patterns. An example of this fact was the degree of similarity between the sister sequences MN703128 and MN703129.

MN703132 and MT222961 were the most similar to MN736701 and MN736702. The explainability patterns of MN153480, MT417738 and MN153494 were remarkably similar.

Given the joint analysis of phylogeny, the CNN classification results, and the degrees of similarity in the hot zones of all phylogenetically close sequences, it is plausible to consider the potential existence of hidden genetic recombination phenomena in subtype U sequences MN736701 and MN736702. Additionally, I obtained certain data that prompt further investigation into hidden parental relationships among the recombinant subtypes in this area of the phylogenetic tree.

8.5.11 Preliminary Conclusions

The sequences designated as subtype U are moderately dispersed throughout the phylogenetic tree. While some grouping exists in certain cases, there is no clear phylogenetic relationship linking them all. This is consistent with the fact that, from a microbiological standpoint, we are not actually referring to a specific subtype, but rather a collection of unclassifiable sequences that have been grouped under the same name. Therefore, we are not discussing a family, subfamily, or lineage whose elements share a certain degree of common ancestry, nor do their genomes exhibit a specific degree of homology or share particular viral characteristics, etc. [2510–2512].

Due to this peculiarity, the high degree of "errors" in the CNN classification of these sequences seems logical, as well as the subsequent divergence in their hot zones.

Taking into consideration this inherent peculiarity of subtype U, the detailed analysis of the explainability patterns along with the phylogeny yields interesting results.

Some case studies resulted in more conclusive and better-related outcomes than others, at least superficially. The most important aspect is that in all cases, research directions have been suggested that can complement and shed light on the sequences classified as U themselves, their parental relationships with other subtypes, and among different non-U subtypes.

In the following Sections 8.6, 8.7, and 8.8, based on the evidence gathered at this point, I will compare the results obtained with the biological validation against the existing literature for each of the sequences of type U.

Once the comparison and joint analysis of phylogeny and similarity degree of each U sequence with its phylogenetically close sequences are completed, in this

section, I will complement the previous results with scientific literature, grouping the outcomes based on the level of biological coherence.

8.6 Conclusive Biological Coherence

The first case of high biological homology is the group of intra-patient sequences EF029066, EF029067, EF029068 and EF029069 and their related sequences FJ388921 and KY392775 (all of them subtype U).

FJ388921 (Patient Code: CY090) [2513], EF029066 (Patient Code: H10986) and KY392775 (Patient Code: NGSID12, intra-patient with partial sequences KY365200 and KY365214) [2514] are three subtype U direct relatives (Figure 8.9). Table 8.30 shows classification results.

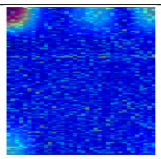
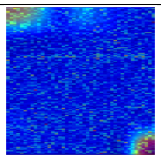
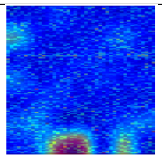
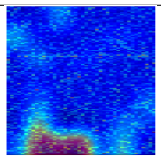
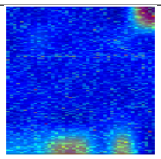
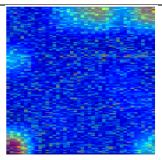
According to the CNN, FJ388921 had a strong non-recombinant component, with standard explainability patterns. EF029066 (classified as non-recombinant) had the lowest confidence score of all U subtype (under 52%). And the CNN classified its intra-patient sequences (EF029067, EF029068, EF029069) as recombinant, with medium confidence levels, around 70-75%. EF029066 explainability patterns were coherent with global non-recombinant hot zones. EF029067, EF029068 and EF029069 had close determinant areas, fairly diffuse in EF029067 and EF029068. Their patterns did not particularly match global recombinant hot zones. KY392775 was classified as non-recombinant just under 90%. However, explainability patterns were slightly controversial with non-recombinant ones.

FJ388921, EF029066 and KY392775 explainability patterns were similar. This correlated with their phylogenetic positions.

These three sequences share clade with recombinant (non-CRF) subtype FK MH705156 and the non-recombinant subtype K AJ249235 and AJ249239 sequences.

EF029066, EF029067, EF029068 and EF029069 are intra-patient sequences (Patient Code H10986). The CNN classified EF029066 as non-recombinant, but its CNN reliability level was minimum (near 51.50%). On the contrary, EF029067, EF029068 and EF029069 were classified by the CNN as recombinant and their

Table 8.30: FJ388921, EF029066, EF029067, EF029068, EF029069 and KY392775 Results. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
FJ388921	U	Non-recombinant	91.853%	
EF029066	U	Non-recombinant	51.424%	
EF029067	U	Recombinant	76.708%	
EF029068	U	Recombinant	69.344%	
EF029069	U	Recombinant	73.163%	
KY392775	U	Non-recombinant	89.344%	

explainability patterns were highly similar and divergent from EF029066. The intra-patient feature of these four sequences unables to analyze deeply their phylogenetic relationship, but according to the results obtained, there is a clear recombination phenomenon. EF029067, EF029068 and EF029069 decision areas were similar to several sequences of different recombinant subtypes, and this was coherent with CNN classification (Table 8.31). The striking similarity level between these different subtypes suggests the existence of genomic components common to several different

recombinant lineages, related with these three U subtype sequences.

Most similar hot areas corresponded to those of CRF26_A5U FM877782, and to a lesser extent to CRF02_AG AB286862, CRF11_cpx AJ291718 and CRF13_cpx DQ845386. Their confidence scores were higher than 90%, except for CRF11_cpx AJ291718 (around 52%).

Another point that requires further research is the possible relationship between low confidence scores and similarity between explainability patterns (case of AJ291718).

EF029066 and its intra-patient subtype U sequences EF029067, EF029068 and EF029069 are currently cataloged as a new subtype, or a new branch of subtype K [2515]. This supports EF029066 classification as non-recombinant. EF029066 low reliability level and the similar recombinant hot zones of EF029067, EF029068 and EF029069 make it necessary to search for recombination events in these sequences.

A sister clade including subtype F2 sequences and the resemblance between EF029066 and these F2 sequences explainability patterns suggest certain common features between them. As seen in Table 8.32, Subtype F2 sequences KU749420, and especially AJ249237 and AY371158 have strongly similar hot zones to EF029066. This endorses EF029066 non-recombinant classification. However, its low confidence-score and recombinant sister sequences could make me consider the presence of recombinant factors embedded.

Subtype F2 sequences KU749420, and especially AJ249237 and AY371158 had strongly similar hot zones to EF029066. All these sequences were clearly classified as non-recombinant by the CNN. More information about involved subtype F2 sequences is available at Table 8.10.

Another highly coherent case is related with subtype U HM215249 and HM215251. HM215249 (Patient Code: TV721 together with partial sequence DQ059736) and HM215251 (Patient Code: TV749 with partial sequence DQ059737) were classified as non-recombinant by the CNN (Table 8.33). Their confidence scores above 99% were the maximum of all U subtype sequences and their Grad-CAM patterns were almost identical. HM215249 and HM215251 are the unique components of

Table 8.31: EF029067, EF029068 and EF029069 Most Analogous Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms. Data of non-U-subtype sequences extracted from Tables 8.12, 8.15, 8.19 and 8.21

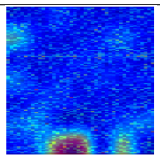
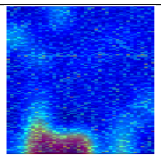
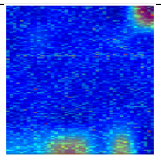
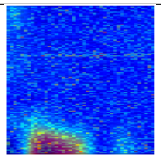
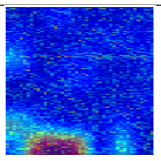
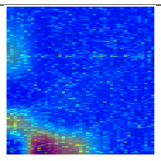
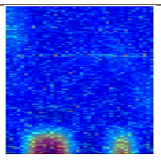
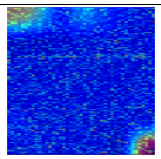
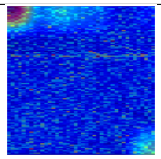
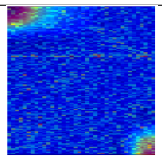
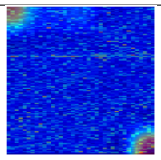
Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
EF029067	U	Recombinant	76.708%	
EF029068	U	Recombinant	69.344%	
EF029069	U	Recombinant	73.163%	
FM877782	CRF26_A5U	Recombinant	90.827%	
AB286862	CRF02_AG	Recombinant	99.905%	
AJ291718	CRF11_cpx	Recombinant	52.328%	
DQ845386	CRF13_cpx	Recombinant	98.412%	

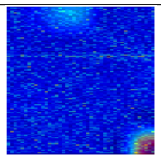
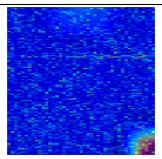
Table 8.32: EF029066 Most Analogous Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
EF029066	U	Non-recombinant	51.424%	
KU749420	F2	Non-recombinant	91.191%	
AJ249237	F2	Non-recombinant	98.966%	
AY371158	F2	Non-recombinant	87.008%	

their clade (Figure 8.8). Quesnel-Vallières et al. considered that HIV-1 isolates infecting patients TV721, TV725, TV749, and TV919 must be classified as URF (Unique Recombinant Form) and therefore they cannot be considered CRFs [2516]. Although there was a discrepancy between CNN's classification as non-recombinant and the work of Quesnel-Vallières et al., we should consider that the CNN was not trained with URF sequences to avoid introducing spurious data into the experiment. Therefore, my results were coherent with the virological evidence. These sequence had highly similar non-recombinant explainability patterns and extremely high confidence scores (>99%). The CNN clearly classified these two sequences as non-recombinant with the highest reliability.

Their nearest relatives are subtype 26CU FM877784, two subtype 26C sequences (FM877779 and FM877783), and a subset of CRF26_A5U sequences (mixture of A5 and U [2517]).

Table 8.33: HM215249 and HM215251 Results. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
HM215249	U	Non-recombinant	99.553%	
HM215251	U	Non-recombinant	99.289%	

Tables 8.11 and 8.12 (Section 8.5.4) show the coherence between non-recombinant explainability patterns and recombinant ones. In the case of subtype U HM215249 and HM215251, both sequences were classified by the CNN as non-recombinant. Their hot zones were correlated with their category and showed a very high degree of similarity, consistent with their close phylogenetic proximity. Similarly, the recombinant sequences CRF26_A5U contained in Table 8.12 were classified by the CNN as recombinant with high trust levels, and their hot areas were also analogous and phylogenetic-distance dependent.

The third case of high biological coherence is subtype U JF683772 (Patient Code: CY223 - unique sequence) whose only sister is subtype 01G MN486045. JF683772 was classified as recombinant (trust level over 75%) and its patterns were similar to Table 8.15 subtypes CRF02_AG and CRF63_02A6 (recombination between CRF02_AG and A6 [2518]), endorsing the JF683772 recombinant feature. This sequence may have recombination phenomena which approach JF683772 to already-defined CRF subtypes.

Wei Shau Hu and Howard M. Temin point the recombinant feature of this sequence [2519], and Mary A. Rodgers et al. suggests a recombination between A1, G, and H [2514]. Explainability patterns were consistent with this.

JF683772 mathematical signature was similar to some recombinant sequences of different subtypes as well as to subtype F1 MT417765, which was classified

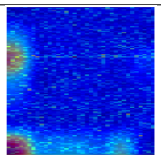
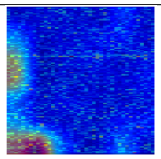
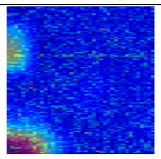
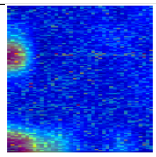
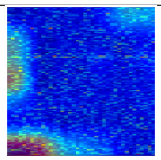
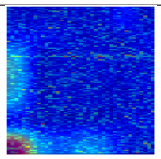
as recombinant with a relatively high trust level (Table 8.34). Coherence with global recombinant hot areas suggested common genomic components among their subtypes that should be analyzed. And all these similarities endorse the JF683772 classification as recombinant. In addition, according to the obtained results, MT417765 could feature recombinant factors that should be researched.

JF683772 was classified as recombinant with a moderate confidence level. Its explainability patterns were coherent with recombinant feature. Its hot zones were similar to CRF02_AG KX232616, CRF11_CPX KP718937 and KP718938, and CRF_13 KP718926. Subtype F1 MT417765 explainability patterns were also similar. This sequence was classified as recombinant with a confidence score upper than 80%.

The fourth case is JX140679. This is intra-patient with sequences JX422198, KC113002 and KC113003 (Patient Code: P2468_ARP1158). JX140679 together with KC113002 and KC113003 can be framed as CRF02_AG / CRF06_cpx, while JX422198 as CRF01_AE, A1, G [2520]. This means that, although these sequences can not be cataloged in any subtype, they are assimilated to recombinant subtypes, with a pronounced tendency towards CRF02_AG and CRF06_cpx (mosaic of A, G, J and K) [2521]. In fact, it is clear from Table 8.16 that the most similar sequence was subtype CRF02_AG KX232622. As seen in the phylogenetic tree (Figure 8.14), the immediate related sequences of JX140679 are 0206 and 02GK subtypes. So, I perceived a solid relationship between explainability patterns data and microbiological evidence, detecting the recombinant feature of JX140679. The resemblance of the hot zones of KY392774 and JX140679 (Table 8.45) made me intuit that these sequences could be phylogenetically close.

A detailed analysis of surrounding clades in conjunction with hot zones comparison, revealed exciting phenomena (Table 8.35). JX140679 explainability patterns were similar to CRF02_AG KX232622 followed by KX232616. The clade aside the cluster of JX140679 contains several A1 sequences. JX140679 was moderately similar to KX232611, which theorically is A1 but the CNN classified it as recombinant with a trust level of 96.440%.

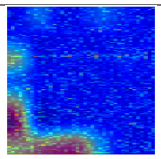
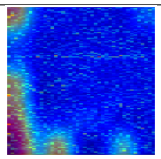
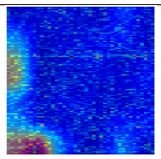
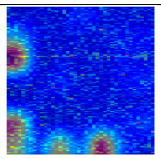
Table 8.34: JF683772 Results and Most Analogous Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms. More info about involved non-U sequences is available in Tables 8.16, 8.19 and 8.21.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
JF683772	U	Recombinant	75.854%	
KX232616	CRF02_AG	Recombinant	99.957%	
KP718937	CRF11_cpx	Recombinant	99.862%	
KP718938	CRF11_cpx	Recombinant	96.436%	
KP718926	CRF13_cpx	Recombinant	96.912%	
MT417765	F1	Recombinant	81.698%	

Another neighboring clade consists of a subtype AGU common ancestor whose offspring belong to CRF37_cpx subtype (mosaic of CRF01_AE, CRF02_AG, A, G, U) [2522]. Subtype CRF37_cpx and JX140679 are certain similar (Table 8.18). All explainability patterns were coincident with global recombinant hot zones. However, CRF37_cpx determinant areas were quite peculiar, compared with other recombinant subtypes. CRF37_cpx sequences had a particular hot point at one-third frequency located around *vif* gene. Interestingly, its location corresponded

to U genomic areas. Further research is needed.

Table 8.35: JX140679 Most Analogous Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms. The complete explainability patterns of all involved sequences is available at Section 8.5.6.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
JX140679	U	Recombinant	77.945%	
KX232622	CRF02_AG	Recombinant	99.974%	
KX232616	CRF02_AG	Recombinant	99.957%	
KX232611	A1	Recombinant	96.440%	

JX140679 hot zones were significantly similar to those of CRF02_AG KX232622 and KX232616, clearly classified as recombinants (confidence score around 99%). Another remarkable resemblance was with A1 KX232611, classified as recombinant with a confidence score over 96%.

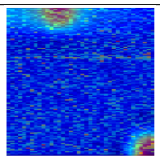
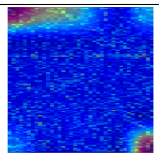
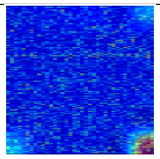
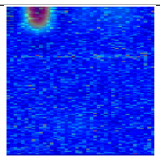
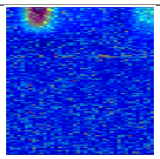
This particular likeness with obviously recombinant classified sequences supports my results.

Next highly consistent U sequence was MN153489 (isolate 08CMARC092 - unique sequence) that was classified as non-recombinant. However, it is described as a complex mosaic pattern CRF02_AG, D, and U recombination [2523]. Its high non-recombinant confidence score (93.586%) was related with its phylogenetic position. Its phylogenetic correlation with mainly subtypes A1 and A6 sequences supports

this point. MN153489 looked similar to sequences with a strong non-recombinant component. This factor endorses my results (Table 8.36).

The CNN classified this sequence as non-recombinant with a high confidence level (upper than 90%). MN153489 explainability patterns were closest to subtype A1 MN791169 ones, followed by subtype A1 MN791873 and subtype A6 MT222946 and MN703139. All these non-U sequences were classified as non-recombinants with a confidence score higher than 94%.

Table 8.36: MN153489 Results and Most Analogous Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms. All explainability patterns are listed in Section 8.5.8.

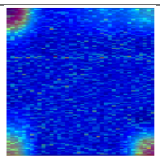
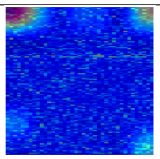
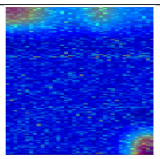
Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
MN153489	U	Non-recombinant	93.586%	
MN791169	A1	Non-recombinant	99.987%	
MN791873	A1	Non-recombinant	99.975%	
MT222946	A6	Non-recombinant	99.926%	
MN703139	A6	Non-recombinant	94.302%	

The next sequence was MN153492 (Patient Code: ARC076 - isolate 06CMACR076, unique sequence). This is intra-patient with partial sequences FJ014630, FJ014689 and GQ344961. The CNN classified MN153492 as non-recombinant (trust level of

91.835%). As seen in Figure 8.21, its nearest sequence was MN153476 (recombinant, subtype 01F2). Both share clade with another two sister subtype F2 sequences. MN153492 explainability patterns were remarkably similar to subtypes F2 and F1 ones, especially to sequences MN153485 (subtype F2) and MT417762 (subtype F1).

MN13492 had a high reliable classification as non-recombinant, above 90%. Explainability patterns matched with its category. Its patterns matched with subtype F2 MN153485 ones. And, to a lesser extent, with subtype F1 MT417762 (Table 8.37). These F2 and F1 sequences were classified as non-recombinant but their confidence scores are medium, specially in the case of MT417762 (under 69%).

Table 8.37: MN153492 Results and Most Analogous Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
MN153492	U	Non-recombinant	91.835%	
MN153485	F2	Non-recombinant	84.152%	
MT417762	F1	Non-recombinant	68.152%	

MN153492 probably features some genomic similarity to a recombinant sequence that makes its classification as F2 questionable. However, explainability patterns denoted clearly that it is non-recombinant.

My results might be inconsistent with the assumption that MN153492 was the result of genetic recombination between F1, CRF02_AG and U [2523]. Further research is needed.

MN153485 and MN153492 reliance levels were fairly low. This was solid and coherent: same patterns, quite low reliances and phylogenetically closeness. Exploring the cause of their low confidence score could shed light about possible recombinant components embedded in their genomes, even more if we consider that its closest sequence to MN153492 was a recombinant one (subtype 01F2 MN153476). In addition, CRF12_BF MT712270 is the unique recombinant sequence in this extract of the phylogenetic tree (Figure 8.21), mixture of B and F1 [2524].

In addition to subtype F2 MN153485, low confidence levels were also measured in F1 sequences MT417762 (about 68%) and MT417747 (89%). This phenomenon, together with the classifications as recombinant of F1 sequences MT417769, MT417765 and MT417748 (Table 8.38), led me to consider the possibility that there were genomic components in these F1 sequences that dissolve their non-recombinant nature.

In the case of MT417769, trust level was near 98.5%. MT417765 and MT417748 confidence scores were moderate, over 80%. Their explainability patterns were rather similar and coherent with recombinant featuring.

The same discussion arises as in the previous case of MN153489. The similarity level of their patterns might indicate the presence of similar factors producing this phenomenon.

The last case corresponded to subtype U MN736701 and MN736702. The CNN classified MN736701 (Patient Code: SI300 - sample 34, unique sequence) and MN736702 (Patient Code: SI324 - sample 37, unique sequence) as recombinants (Table 8.39). MN736701 confidence score was under 75%. MN736702 decision was very controversial, with a reliability of only 52.857%. Explainability patterns of both sequences were similar.

In the case of MN736702, such a low confidence suggested that the CNN had serious trouble in determining its recombinant feature.

The phylogenetic tree section contains only recombinant sequences (Figure 8.23). MN736701 closest sequence is CRF27_cpx MN116204. The other relative sequences are recombinants of various subtypes: 06U, CRF06_cpx, 0218 and

Table 8.38: Phylogenetically-Close-to-MN153492 F1 Sequences, classified as Recombinant. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms. Table 8.25 is the complete compendium of explainability patterns in this case.

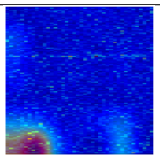
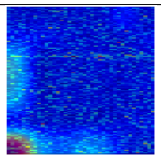
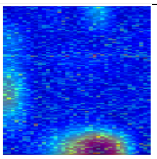
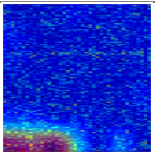
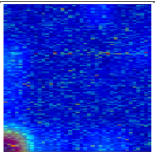
Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
MT417769	F1	Recombinant	98.458%	
MT417765	F1	Recombinant	81.698%	
MT417748	F1	Recombinant	85.253%	

Table 8.39: MN736701 and MN736702 Results. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
MN736701	U	Recombinant	74.420%	
MN736702	U	Recombinant	52.857%	

CRF18_cpx. All phylogenetic relationships are with recombinant sequences, so its recombinant classification was coherent.

MN736701 is a complex recombinant of subtypes A, G, J, CRF01_AE, and U [2525]. My results supported its recombinant feature. Interestingly, MN736701 is intra-patient with partial sequence KF753731 (Patient Code: SI300), included in G subtype. Further research about the possible recombinant origin of subtype G would help to clarify, not only this phylogenetic relationship, but also the phylogeny of HIV-1.

Table 8.40 summarizes MN736701 most similar explainability patterns. All of them were almost identical. All sequences had a strong recombinant feature with confidence scores surrounding 99%. This phenomena endorsed the classification of MN736701 as recombinant. The sequences whose interpretability patterns were most similar to MN736701 are CRF20_BG MW062981, CRF02_AG MN703130 and MN703132, followed by CRF11_cpx KP718936, CRF02_AG AB286857 and CRF63_0206 JN230353. .

MN736702 is intra-patient (Patient Code: SI324) with partial sequence KF753704, assessed as NA. MN736702 is composed of CRF02_AG, CRF56_cpx (mosaic of CRF02_AG, B and G) [2526], and subtype B fragments [2525].

Figure 8.28 displays the genome map of subtype CRF56_cpx.

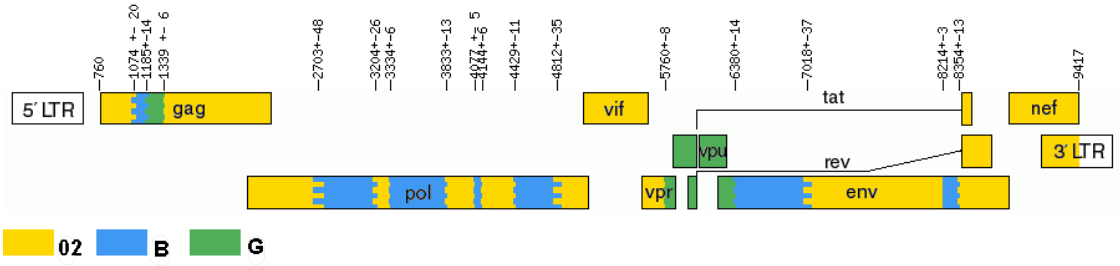
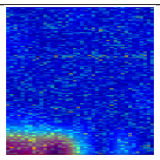
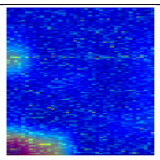
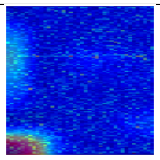
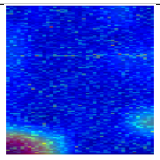
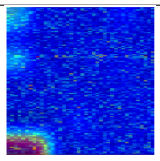
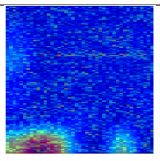
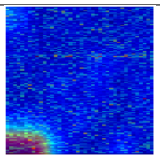


Figure 8.28: Genome Map of Recombinant Subtype CRF56_cpx. This subtype is a mosaic of subtypes CRF02_AG, B, and G. Image sourced from the Los Alamos National Laboratory HIV-1 Database HIV Circulating Recombinant Forms (CRFs) [41].

As in the case of MN736701, all its sharing-clade sequences are recombinants of diverse subtypes, including CRF22_01A1 MN153487 (mixture of CRF01_AE

Table 8.40: MN736701 Most Analogous Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms. Complete information about explainability patterns of these sequences is available at Tables 8.27, 8.29, 8.19 and 8.15

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
MN736701	U	Recombinant	74.420%	
MW062981	CRF20_BG	Recombinant	99.958%	
MN703130	CRF02_AG	Recombinant	99.573%	
MN703132	CRF02_AG	Recombinant	99.798%	
KP718936	CRF11_cpx	Recombinant	98.496%	
AB286857	CRF02_AG	Recombinant	99.732%	
JN230353	CRF63_02A6	Recombinant	99.935%	

and A1) [2527]. MN736702 is prevalently related to subtype CRF02_AG sequences. My results were supported by these evidences.

CRF02_AG MN703130, MN703132 (same case as MN736701), MN153495 and MT222961 were the most similar sequences to MN736702. To a lesser extent, CRF18_cpx MF109378, CRF02_AG L39106 and CRF11_cpx KP718929 had also similar explainability patterns. MN736702 differed from MN736701 in its resemblance with non-recombinant hot zones, for example, subtypes A1 MN791279, A6 MN703138 and F1 MT417765. These recombinant and non-recombinant subtype hot zones were virtually identical. This phenomenon is especially exciting. By locating their frequency-domain peculiarities, we can advance towards the formulation of these mathematical signatures. It may be easier to deduce them by direct comparison (Table 8.41).

Recombinant sequences CRF02_AG MN703130, MN703132, MN153495 and MT222961, CRF18_cpx MF109378 were similar. Idem non-recombinant sequences A1 MN791279, A6 MN703138 and F1 MT417765. Although CRF02_AG L39106 and CRF11_cpx KP718929 were classified as non-recombinants, their explainability patterns were similar to the rest of recombinant classified sequences.

Table 8.41: MN736702 Most Analogous Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms. More information is available at Tables 8.29, 8.3, 8.15, 8.23, 8.24, 8.25 and 8.19.

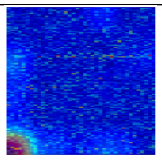
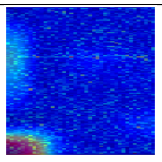
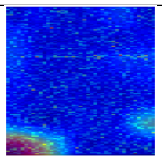
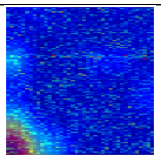
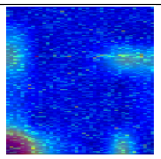
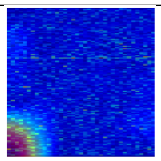
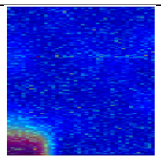
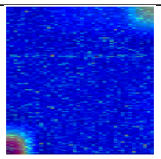
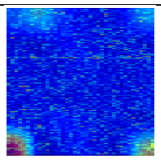
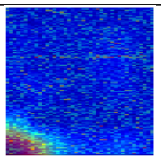
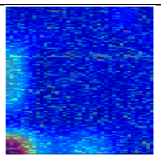
Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
MN736702	U	Recombinant	52.857%	
MN703130	CRF02_AG	Recombinant	99.573%	
MN703132	CRF02_AG	Recombinant	99.798%	

Table 8.41 – Continuation from the previous page

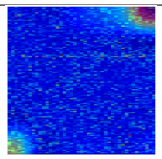
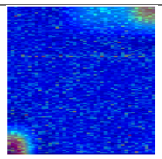
Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
MN153495	CRF02_AG	Recombinant	99.134%	
MT222961	CRF02_AG	Recombinant	99.834%	
MF109378	CRF18_cpx	Recombinant	98.499%	
L39106	CRF02_AG	Recombinant	98.921%	
KP718929	CRF11_cpx	Non-recombinant	59.627%	
MN791279	A1	Non-recombinant	99.817%	
MN703138	A6	Recombinant	56.613%	
MT417765	F1	Recombinant	81.698%	

8.7 Non-Conclusive Biological Coherence

The first non-conclusive case involves AB485664 and AB485665. The CNN classified both sequences as non-recombinant with high confidence scores (>90%). Their explainability patterns were strongly similar and consistent with non-recombinant

category (Table 8.42).

Table 8.42: AB485664 and AB485665 Results. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
AB485664	U	Non-recombinant	90.331%	
AB485665	U	Non-recombinant	92.995%	

AB485664 and AB485665 are intra-patient sequences (ID Patient KITA) along with partial sequences AF447852 and FJ183726 [2528], being AB485665 the reference sequence.

AB485664 and AB485665 explainability patterns were deeply similar. This was coherent with their highest phylogenetic proximity (intra-patient sequences).

AB485665 is phylogenetically close to AF076474 (recombinant subtype A1GHU) and AY371166 (recombinant subtype AGU). While CNN clearly classified AB485665 as non-recombinant, some researching propose that it could be recombinant [2529].

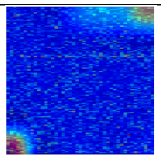
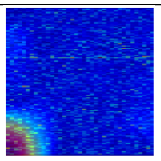
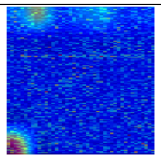
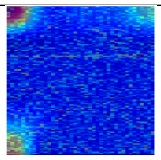
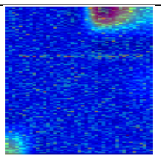
As explained in Section 8.5.1, according to the CNN criteria, AB485665 was the only non-recombinant sequence, surrounded by recombinants in the phylogenetic tree (Figure 8.2). The CNN classified this sequence and AB485664 with a high confidence score and explainability patterns were clear and coherent. Same phenomena happened with subtype U sequences AY046058 and KY392780 (Section 8.5.2), HM215249 and HM215251 (Section 8.5.4) and KP718920 (Section 8.5.7). Future researching should deeply analyze genomic differences of these non-recombinant sequences and their recombinant relatives.

The sister clade of AB485665 consists mostly of non-recombinant subtype H sequences. These subtype H sequences share explainability patterns with AB485665, specially KY392778, AF190127 and KY392777, whose hot zones clearly belonged

to non-recombinant category. Curiously, the most similar explainability patterns to AB485665 were those of MF109378 (CRF18_cpx), clearly classified as recombinant and being the closest CRF18_cpx offspring of AB485665. There was a correlation between pattern similarity and phylogenetic closeness (Table 8.43).

CRF18_cpx MF109378 and subtype H KY392778, AF190127 and KY392777 had the most similar hot zones. MF109378 was clearly classified as recombinant, with a confidence score of almost 93.0%. Subtype H sequences had unequal trust levels as non-recombinants. From 72.5% of KY392777 to 92.1% of AF190127. MF109378 and KY392778 explainability patterns were almost identical to AB485665 ones.

Table 8.43: AB485665 Most Analogous Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms. The complete information is shown at Tables 8.3 and 8.4.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
AB485665	U	Non-recombinant	92.995%	
MF109378	CRF18_cpx	Recombinant	98.499%	
KY392778	H	Non-recombinant	84.282%	
AF190127	H	Non-recombinant	92.134%	
KY392777	H	Non-recombinant	72.514%	

Recombinant MF109378 had the same pattern as the involved non-recombinant sequences. This is, low frequencies of 5'LTR. So, this area may contain critical factors related with recombinant featuring.

One interesting study case is the identification of frequency-domain differences between subtype CRF18_cpx MF109378 and subtype H KY392778 in order to unravel one possible mathematical signature that characterizes HIV-1 recombinant feature. The Grad-CAM hot zones of both sequences showed high power in the low frequency range in the 5'LTR. Therefore, in a more in-depth frequency analysis, the differences we might find between the frequency patterns of this two sequences in this area could provide information about the recombinant feature.

There is an additional interesting fact. According to Figure 8.2, almost all AB485665 clade-mates are CRF18_cpx (recombination between A1, F, G, H, K, U) [2530], except MN153475. This CRF18_cpx sequence is phylogenetically close to subtype U MN736701 (Figure 8.23). The role played by U subtype on CRF18_cpx should be analyzed in depth.

The clade-sisterhood with subtype H sequences invites me to investigate the commonalities between this subtype and AB485665 and its intra-patient sequence AB485664.

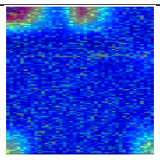
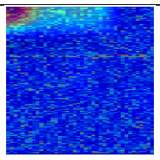
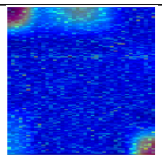
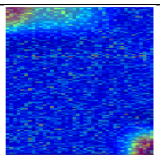
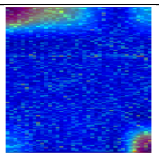
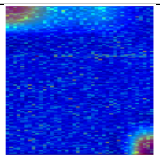
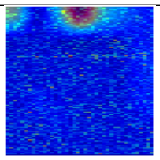
The second non-conclusive case was KY392773. This is intra-patient (Patient Code: NGSID8) with KY365209 (partial sequence, genotype U) [2514]. KY392773 was classified by the CNN as non-recombinant but its trust level was quite low (69.660%).

As seen in Table 8.44, the most similar explainability patterns corresponded to clearly non-recombinant sequences. This phenomenon remarked the non-recombinant feature detected by the CNN at KY392773.

The CNN classified this sequence as non-recombinant but with a low confidence score (under 70%). Explainability patterns corresponded with non-recombinant ones, with a prominent role of high frequencies in the central area of the sequence (secondary non-recombinant pattern). The involved sequences whose explainability patterns were most similar with KY392773 ones were subtype A2 AF286237 and GU201516, A1 KX232597 and MN791169, and A6 MN703137 and MT222947.

Except for MN703137 (trust level of 81.7%), all of them were classified as non-recombinant with confidence scores over 92%. The most similar sequence was subtype A2 GU201516.

Table 8.44: KY392773 Results and Most Analogous Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms. Complete information is available at Tables 8.13, 8.17, 8.23 and 8.24.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
KY392773	U	Non-recombinant	69.660%	
AF286237	A2	Non-recombinant	98.597%	
GU201516	A2	Non-recombinant	92.866%	
KX232597	A1	Non-recombinant	99.701%	
MN791169	A1	Non-recombinant	99.987%	
MN703137	A6	Non-recombinant	81.688%	
MT222947	A6	Non-recombinant	99.856%	

As seen in Section 8.5.1, KY392773 shares clade with recombinant sequences. Its sister sequence is AJ237565, subtype A1DHK. Both sequences share clade with subtype A1H KP718933. This three-sequence clade has a sister clade which consists

on five CRF04_cpx sequences. In accordance with Table 8.2, the hot zones of KY392773 did not show an obvious similarity to the hot zones of the CRF04_cpx sequences of its clade. Among these recombinant sequences, the most similar explainability patterns with KY392773 could those from CRF04_cpx AF119820 in the high-frequency range of 5'LTR. The CNN clearly identified AF119820 as recombinant. This fact might denote certain recombinant components that make KY392773 phylogenetically close to recombinant CRF sequences.

KY392774 sequence is not mapped in HIV-1 phylogenetic tree. KY392774 is intra-patient (Patient Code NGSID10) with KY365188 (incomplete sequence genotype = A1) [2514]. The CNN classified KY392774 as recombinant but its coincidence score was low. Explainability patterns were coherent with recombinant feature. And, as seen in Table 8.45, KY392774 hot zones were highly coincident with those of JX140679 (Table 8.45). Despite the similarities between their hot zones, they are not particularly close sequences in the phylogenetic tree. However, the CNN classified both as recombinant with highly similar hot zones. Further research is needed to determine the phylogenetic closeness between these two sequences.

Their hot zones were remarkably similar and rather divergent with global recombinant patterns. The main difference was a KY392774 primary hot area at the second half of the genome. JX140679 had a medium confidence score (about 78%). KY392774 classification was unclear (less than 60%).

The third non-conclusive case included subtype U AY046058, KY392780, MH078559 and MK095236. There was a phylogenetic correlation among them, as explained in Section 8.5.2.

MK095236 classification was recombinant and their patterns corresponded with the recombinant ones, but there was a certain hesitation level (under 90%). There was a very secondary pattern along the frequency range of 5'LTR. AY046058 and MH078559 were clearly classified as non-recombinants (confidence score over 90%). The CNN had a low confidence in KY392780 (around 62%). Non-recombinant patterns of these three sequences were similar, especially between KY392780 and AY046058. Table 8.46 provides an abstract of classification results of these sequences.

Table 8.45: Comparison between KY392774 and JX140679 Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

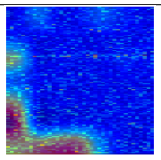
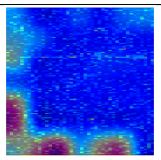
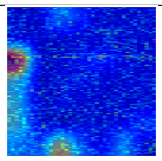
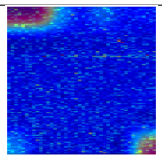
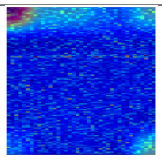
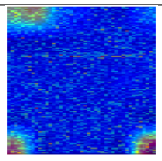
Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
JX140679	U	Recombinant	77.945%	
KY392774	U	Recombinant	58.879%	

Table 8.46: MK095236, KY392780, AY046058 and MH078559 Results. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

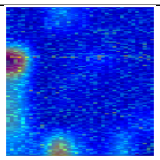
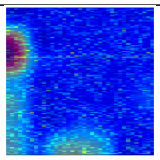
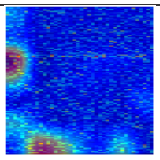
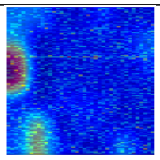
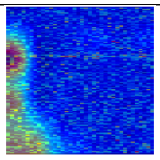
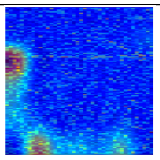
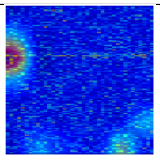
Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
MK095236	U	Recombinant	82.815%	
KY392780	U	Non-recombinant	62.076%	
AY046058	U	Non-recombinant	90.426%	
MH078559	U	Non-recombinant	91.760%	

MK095236 (classified as recombinant) is separated from all other sequences by a common ancestor (Figure 8.5). Curiously, subtype U KY392780, AY046058 and MH078559 were classified as non-recombinant, as evidenced by their explainability patterns (Table 8.46). Besides two subtype A7 sequences, the only non-recombinant ones were specifically these subtype U sequences. Given the coherence of their explainability patterns, there could be a common element in their genome that identifies these sequences as non-recombinant despite their resemblance to recombinant sequences. Finding this common element could clarify this phenomenon.

A group of CRF09_cpx sequences (mixture of A, G, U) stands out among the rest of the CRF sequences in this area of the phylogenetic tree. It is possible that CRF09_cpx and CRF02_AG have a common ancestor [2509]. Precisely, sequence MH078539 (CRF02_AG) has a common ancestor with some 0209 sequences and this set of CRF09_cpx sequences. This fact is tangential to the scope of the present research, but I consider it is worth mentioning as an interesting observation.

The other relevant cluster consists of several CRF45_cpx sequences (mosaic of A, K and U) [2531]. The most determinant explainability pattern was one-third frequency of 5'LTR. This frequency was one of the main recombinant hot zones, so this was coherent with CNN classification. MK095236 relevant areas were similar to those of CRF45_cpx FN392875, CRF02_AG AF063223 and JN248580, CRF16_A2D KT022403, AF286239 and AY945736 Explainability Patterns. All involved sequences were classified as recombinant with high confidence scores.

Table 8.47: MK095236 Most Analogous Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms. Complete Explainability Patterns of all involved sequences are available at Tables 8.7, 8.15 and 8.14.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
MK095236	U	Recombinant	82.815%	
FN392875	CRF45_cpx	Recombinant	98.699%	
AF063223	CRF02_AG	Recombinant	99.829%	
JN248580	CRF02_AG	Recombinant	99.992%	
KT022403	CRF16_A2D	Recombinant	99.389%	
AF286239	CRF16_A2D	Recombinant	95.870%	
AY945736	CRF16_A2D	Recombinant	99.652%	

As summarized in Table 8.47, most similar explainability patterns with MK095236 were those of subtype CRF45_cpx FN392875 (complete information in Table 8.7) and their hot zones were similar to recombinant global patterns. MK095236 hot zones were also similar to subtype 02_AG sequences AF063223 and JN248580 and subtype 16_A2D sequences KT022403, AF286239 and AY945736. Subtype

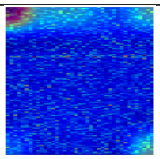
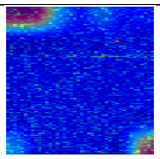
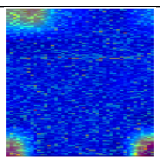
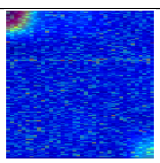
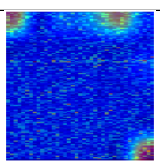
16_A2D is a recombinant between A2 and D [2532, 2533].

No Patient Code is associated with AY046058 while MH078559 Patient Code is 316688 (unique sequence), and KY392780 is intra-patient sequence with partial sequence KY365183 (Patient Code: NGSID18) [2514].

AY046058, KY392780 and MH078559 explainability patterns exhibited similarities with subtype A7 KX389622 and MH078558 (Table 8.48). It is noteworthy that these three subtype U sequences are the only non-recombinant ones in their phylogenetic tree extract, and their hot areas were similar. They may share a common factor, foreign to the rest of the sequences, which determines their non-recombinant character. Further research is needed.

Their explainability patterns were coherent with global non-recombinant ones and remarkably similar to those of subtype A7 KX389622 and MH078558.

Table 8.48: AY046058, KY392780 and MH078559 Most Analogous Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms. More information is available at Tables 8.5 and 8.6.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
AY046058	U	Non-recombinant	90.426%	
KY392780	U	Non-recombinant	62.076%	
MH078559	U	Non-recombinant	91.760%	
KX389622	A7	Non-recombinant	86.313%	
MH078558	A7	Non-recombinant	84.715%	

AY046058 non-recombinant featuring is controversial with the conclusion that AY046058 is a mosaic of A, G, J, K and regions of indeterminate classification [2534].

The third non-conclusive case was KP718920 (Patient Code: 469_66 - unique sequence) that was classified as non-recombinant. Their explainability patterns were coincident with global non-recombinant ones, and the most homologous hot zones corresponded to subtype F2 and A1 several sequences (Table 8.49). All this factors validated the KP718920 non-recombinant feature.

Classified as non-recombinant, its explainability patterns agreed with global non-recombinant patterns, but confidence score was not completely evident, under 85%. These patterns were similar to subtype F2 AF377956, AY371158, AJ249237 and AF377956. They also were moderately alike to subtype A1 KX232595, KX232607 and subtype F2 MN153485 and MN153483.

Table 8.49: KP718920 Results and Most Analogous Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms. More information about non-U subtypes is shown at Tables 8.10, 8.17, 8.23, 8.24 and 8.25.

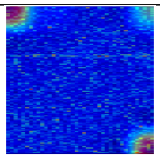
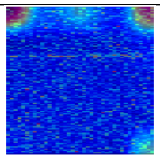
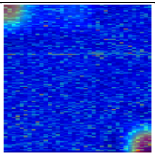
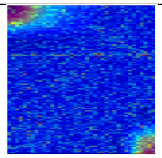
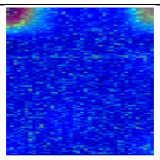
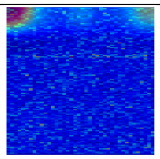
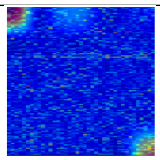
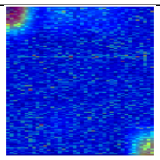
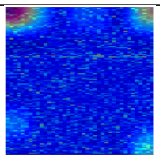
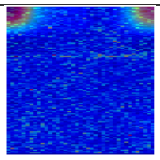
Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
KP718920	U	Non-recombinant	84.895%	
AF377956	F2	Non-recombinant	99.293%	
AY371158	F2	Non-recombinant	87.008%	
AJ249237	F2	Non-recombinant	98.966%	

Table 8.49 – Continuation from the previous page

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
KX232595	A1	Non-recombinant	82.135%	
KX232607	A1	Non-recombinant	93.285%	
MN153491	A1	Non-recombinant	95.919%	
MN703135	A6	Non-recombinant	99.075%	
MN153485	F2	Non-recombinant	84.152%	
MN153483	F2	Non-recombinant	95.832%	

With the exception of KR017778, which belongs to subtype 11A1, all KP718920 related sequences belong to subtype CRF11_cpx (Figure 8.17), mosaic of A, E, G, J and U [2521]. A deep analysis of its genome structure suggests that KP718920 is the result of a recombination among CRF36_cpx (CRF01_AE, CRF02_AG, A, G) [2535], CRF11_cpx, CRF02_AG and U [2536].

The genome map of subtype CRF36 is available in Figure 8.29.

The sister clade of KP718920 cluster contains mainly CRF27_cpx sequences (mosaic of A, E, G, H, J, K and U) [2537] and CRF13_cpx (multiple recombination of CRF01_AE, A, G, J and U) [2538].

Phylogenetic linkage with CRF11_cpx, followed by CRF13_cpx [2539] requires

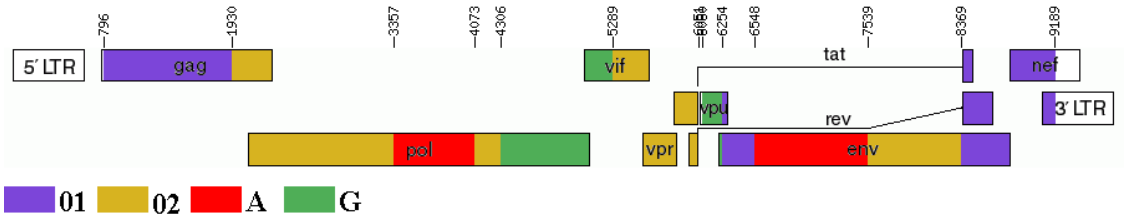


Figure 8.29: Genome Map of Recombinant Subtype CRF36_cpx. This subtype is a mosaic of CRF01_AE, CRF02_AG, A, and G. Image sourced from the Los Alamos National Laboratory HIV-1 Database HIV Circulating Recombinant Forms (CRFs) [41].

further research on KP718920 classification as non-recombinant.

Interestingly, the CNN classified subtype CRF11_cpx KP718929 and JF683802 as non-recombinants (Table 8.19). CRF11_cpx JF683802 and subtype U KP718920 were highly similar. Same case happened with CRF13_cpx AY371154 (Table 8.21). There may be hidden non-recombinant-like phenomena in their genomes. Further research is needed.

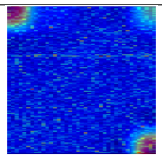
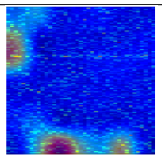
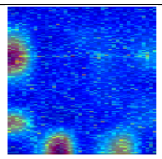
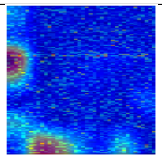
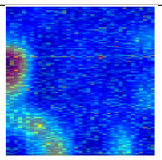
8.8 Another Similarities

As seen in Table 8.50, AF005496 was classified as recombinant (confidence score around 78%). Its hot zones were highly similar to those of subtypes CRF45_cpx FN392877 and CRF02_AG AF063223 and AY371140. Their main and secondary areas provided a distinctive signature which could point to the existence of common elements among these sequences of different subtypes.

Their main and secondary explainability patters were significantly homologous and coherent with global recombinant hot zones. All non-U sequences were clearly classified as recombinant, with confidence scores over 98.5%.

As in the previous case, non-recombinant classified AF190128 (subtype H) had almost identical patterns with subtype A1 MN791873 (Table 8.51). These resemblances demonstrated the interrelationships between the different subtypes and might help to identify new common features.

Table 8.50: Similarities among AF005496 (Subtype H), FN392877 (CRF45_cpx) and subtype CRF02_AG Sequences AF063223 and AY371140. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms. Complete information is provided in Tables 8.4, 8.7, and 8.15.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
KP718920	U	Non-recombinant	84.895%	
AF005496	H	Recombinant	77.935%	
FN392877	CRF45_cpx	Recombinant	98.648%	
AF063223	CRF02_AG	Recombinant	99.829%	
AY371140	CRF02_AG	Recombinant	99.737%	

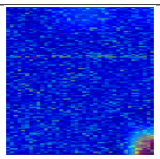
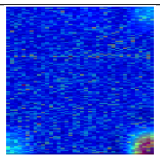
Both sequences had a strong non-recombinant component (confidence scores of 98-99%) and their hot areas are coherent with non-recombinant global explainability patterns.

8.9 Conclusions

CNN classifications as recombinant or non-recombinant and their corresponding patterns were phylogenetically coherent and consistent with sample data, and similarity level correlated with phylogenetic closeness.

In multiple cases, most CRF subtypes phylogenetically close to subtype U

Table 8.51: Similarities between the Explainability Patterns of AF190128 (Subtype H) and MN791873 (Subtype A1). The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms. Additional information of these and their neighboring sequences is shown in Tables 8.4 and 8.23.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
AF190128	H	Non-recombinant	98.441%	
MN791873	A1	Non-recombinant	99.975%	

sequences are recombinants that include fragments categorized as U. Although this appears logical, exploring the details of genomic similarities between the U fragments of these recombinants and complete subtype U sequences would be beneficial to determine the precise reasons behind this evolutionary proximity. Moreover, acquiring additional data could clarify the origin of these U sequences.

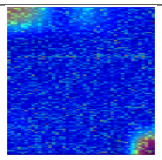
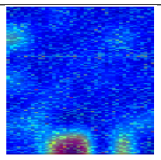
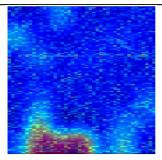
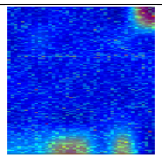
I detected a particularly interesting case where, according to the CNN results, there appeared to be a hidden recombination phenomenon in the intra-patient subtype U sequences EF029066, EF029067, EF029068, and EF029069 (Patient Code: H10986) (Table 8.9).

Table 8.52 compiles the results related to these four subtype U sequences to more easily visualize the results obtained.

EF029066 was classified as non-recombinant by the CNN, with a confidence level that is nearly random, at only 51.424%. However, the remaining three intra-patient sequences were classified by the CNN as recombinants, with moderate confidence levels ranging between 69% and 76%. The confidence levels of these three sequences were relatively consistent with those of the recombinant category. Their very high degrees of homology were in line with being intra-patient (maximum phylogenetic

closeness). Regarding the hot zones of EF029066, they aligned with the recombinant theory and consequently differed from its three remaining intra-patient sequences.

Table 8.52: EF029066, EF029067, EF029068, and EF029069 Explainability Patterns. The Grad-CAM result for each sequence is provided in the corresponding column for superposed spectrograms.

Sequence ID	Subtype	Classification	Confidence Score	Grad-CAM
EF029066	U	Non-recombinant	51.424%	
EF029067	U	Recombinant	76.708%	
EF029068	U	Recombinant	69.344%	
EF029069	U	Recombinant	73.163%	

A detailed study of the differences between the recombinant and non-recombinant sequences, as well as the identification of the genomic elements determining the recombinant feature, could provide information about the mechanisms that drive and produce genetic recombination in HIV-1. Therefore, it is an extremely interesting case study that deserves further research.

My results were coincident with multiple microbiological evidences, such is the case of EF029066, EF029067, EF029068, EF029069, FJ388921, HM215249, HM215251, JF683772, JX140679, KY392775, MN153489, MN153492, MN736701 and MN736702. Their explainability patterns were consistent with phylogeny and multiple biological evidences (60.870% of the total U sequences). The 34.783% was rather inconclusive (AB485664, AB485665, KP718920, KY392773, KY392774,

KY392780, MH078559 and MK095236). Sequence AY046058 was the most controversial case (4.348%).

I also was able to establish similarities between explainability patterns of different subtypes whose sequences are phylogenetically close. This can help me to search for common genomic elements in the hot areas pointed by the CNN. These results also reflected the leading role of genetic recombination in HIV-1 phylogeny.

Many of CRF_cpx involved in this research contain subtype G sections. Engagement between subtypes U and G may denote some parentage level between them. And the possible recombinant origin of subtype G could influence subtype U explainability patterns.

Now we know where the CNN looks at to decide, my next step is searching what exactly is seeing. MN736702 can significantly clarify this point, giving me the opportunity to make a direct comparison of both categories in a very specific genomic area (low frequencies of 5'LTR, located in the vertical strip positioned furthest to the left, see Figure 6.2). Perhaps we can unravel those mathematical signatures that define a sequence as recombinant and better understand HIV-1 genetic recombination.

A deeper understanding of both U complete sequences and U fragments detected in multiple CRFs will allow to establish more precisely how HIV-1 viral evolution is. The use of Deep Learning tools for this purpose may provide new approaches to elucidate knowledge.

This research is a demonstration of this point. This methodology is much faster and low computing, not requiring manual adjustments and multiple alignment. By applying an alternative methodology, I deepened in the knowledge of subtype U, whose sequences are more problematic for classification. And more unknown.

I am not young enough to know everything.

— Oscar Wilde, *The only novel he published during his lifetime was "The Picture of Dorian Gray", one of the greatest ethical debates ever written.*

9 Analysis of HIV-1 Subtype G

Contents

9.1 Introduction 475
9.2 Microbiological Considerations and Recombinant Origin of Subtype G 476
9.3 Equipment 477
9.4 Methods 478
9.5 Global Results of the Experiment 478
9.6 Analysis of Phylogeny and Explainability Patterns . . 480
9.6.1 Subtype G Philogenetic Group 1 480
9.6.2 Subtype G Philogenetic Group 2 499
9.6.3 Subtype G Philogenetic Group 3 509
9.7 Discussion 512
9.7.1 Comments on Subtype G Phylogeny 512
9.7.2 Analysis of Hot Zones in Intra-patient Sequences 516
9.7.3 Prevalence Analysis of Recombinant Subtype G Sequences 521
9.7.4 Analysis of Similarity Patterns among Subtype G Hot Zones 522
9.7.5 Influence of dual infections 540
9.8 Conclusions 542

9.1 Introduction

In this Chapter, I present the results of a comprehensive analysis of subtype G, examining the results of the CNN classification of all sequences that make up this

subtype and comparing them with the phylogeny to provide additional information on the correlation between my findings and the ongoing controversy among HIV-1 experts regarding the possible recombinant origin of subtype G.

This Chapter consists of an explanatory section on some microbiological considerations of subtype G, as well as data from the scientific literature on its possible recombinant origin (Section 9.2). Subsequently, Sections 9.3 and 9.4 describe the characteristics of the equipment used and the methodology employed, respectively. The overall experimental results are presented in Section 9.5. Next, I conducted a comparative analysis between the phylogeny and the patterns detected in the interpretability results (Section 9.6). Section 9.7 offers a detailed discussion. The chapter concludes with Section 9.8, where I present the conclusions of this study.

9.2 Microbiological Considerations and Recombinant Origin of Subtype G

The phylogenetic evolution of HIV-1 is extremely complex, as mentioned in Chapter 5, Section 5.1 and in Chapter 8 Section 8.1. The high genetic variability of HIV-1 poses a significant challenge in the precise identification of its variants and evolutionary divergences. As a result, there is limited knowledge about numerous events that have occurred thus far [2540].

Such circumstances directly affect the nomenclature of the different subtypes. This nomenclature necessitates periodic scrutiny by the expert scientific community involved in HIV-1 research.

Technological advancements such as improvements in long PCR technologies enable the detection and classification of new lineages or emerging subtypes [2541].

All of these updates in the discovery of new strains, lineages, or previously hidden phylogenetic relationships, as well as the detection of new recombination breakpoints between subtypes, can have a direct impact on the nomenclature of HIV-1 subtypes.

To date, the decision of HIV-1 expert community is to maintain non-recombinant subtypes A-D [2542], F-H [2543], J [2544], and K [2545].

Furthermore, the assignment of a subgenomic sequence (an incomplete genomic sequence) to a specific subtype based on genomic regions corresponding to accessory genes is notably intricate. Currently, subtype G is classified among those considered "problematic" [41]. All recognized variants of subtype G demonstrate a closer genetic resemblance to subtype A in the *vif/vpr* region. This observation could stem from an anomaly or a shared recombinant ancestor. Despite ongoing research, controversy persists in this domain, and a definitive conclusion remains elusive [2543, 2546].

Due to the absence of a representative "pure" strain of subtype G and the limited length of the region in question, the lineage denomination of subtype G is maintained rather than redefining it as recombinant.

Therefore, although the decision is to maintain the classification of subtype G as non-recombinant, the exception of the accessory genes of subtype G as non-recombinant remains controversial.

Regardless of the nomenclature decisions made, recent studies, such as the one conducted by Abecasis et al., assert that subtype G is of recombinant origin [2547]. In this research, they clearly point out the classification of subtype G as a CRF (Circulating Recombinant Form).Other studies suggest that the ability of subtype G to undergo recombination may stem from the formation of heterozygous viral particles or the creation of functional chimeras mixed with another HIV-1 strain, resulting in the emergence of a novel and complex CRF [2548].

9.3 Equipment

All experiments were run in this equipment:

- Processing Unit: Intel(R) Xeon(R) Gold 5220R CPU @ 2.20 GHz.
- Installed RAM: 256 GB usable.
- Operative System: Windows 10 Education. Version: 21H1.
- GPU: NVIDIA RTX A6000. Total memory: 179.451 GB. VRAM: 48.571 GB.

9.4 Methods

The research methodology followed the structure represented in Chapter 5, Section 5.4. Initially, I generated genomic spectrograms for all complete HIV-1 sequences up until 2020 (concatenated and superposed). Subsequently, I conducted training using Transfer Learning with a pre-trained CNN using these two datasets of genomic spectrograms. Once the correct functioning of the pre-trained CNN was verified, I employed three distinct explainability tools (Grad-CAM, LIME, and Gradient Attribution) to identify the relevant hot zones within the recombinant feature. Finally, I constructed a phylogenetic tree based on the Complete Consensus Multiple Sequence Alignment of HIV-1 sourced from the Los Alamos National Laboratory HIV-1 Database [41] (up until 2020) and compared the phylogenetic proximity to the similarity level of these identified hot zones.

The detailed information regarding this process is described in Chapter 5, Sections 5.5, 5.6, 5.7, 5.8, 5.9, 5.10, 5.11,and 5.12.

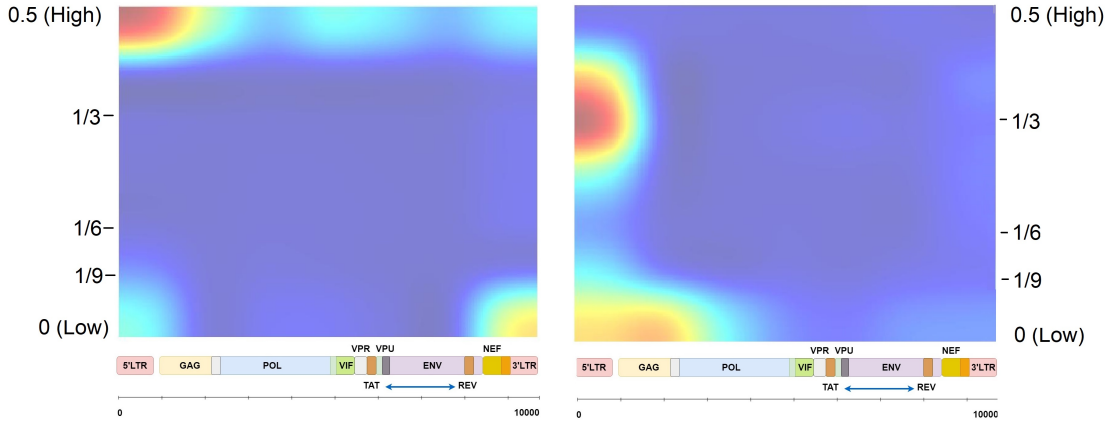
Appendix E is a compilation of the classification results for all subtype G sequences, accompanied by their respective Grad-CAM hot zones (superposed spectrograms). The sequences are arranged in alphanumeric order.

9.5 Global Results of the Experiment

As seen in Table 8.1, the CNN hit rate detecting recombinant feature reached 98.27% on the total of the 13,554 complete HIV-1 sequences.

In non-recombinant sequences, the CNN exhibited only 170 errors out of a total of 11,163 sequences (10,993 correct classifications), yielding an accuracy rate of 98.48%. Conversely, among the recombinant sequences, out of a total of 2,391 sequences, only 65 errors were identified (2,326 correct classifications), resulting in an accuracy rate of 97.28%. Despite being slightly lower in the recombinant sequences, both accuracy rates are considered excellent. However, we must take into account that the CNN previously learned with sequences of Train and Validation Sets, so these have higher hit rate than those of the Test Set.

Figure 9.1 shows explainability patterns for the two categories. Figure 9.1a is the result of the mathematical superposition of the 11,163 non-recombinant patterns and Figure 9.1b is the recombinant global pattern of the 2,391 recombinant sequences. As seen in Figure 9.1a, high frequencies of 5'LTR and low frequencies of 3'LTR were the main hot areas where the CNN looked at to classify a sequence as non-recombinant. Secondary zones were high frequencies of the central part of the genome (around *pol* and *vif* genes) and low frequencies of 5'LTR. In the case of recombinant sequences (see Figure 9.1b, hot areas were one-third frequency at 5'LTR (the most determinant point) and low frequency range in 5'LTR and *gag*. Other much more minor points were low frequencies of 3'LTR and the end of *env* and *nef* genes, and one third of 3'LTR.



(a) Non-recombinant Total Hot Zones **(b)** Recombinant Total Hot Zones

Figure 9.1: Total Explainability Patterns by Category (Superposed Spectrograms)

In an initial visual analysis of the complete phylogenetic tree, subtype G sequences form two major phylogenetic groups. Sequence MN153484 remains isolated from the rest. I analyzed the phylogeny of these two groups and their respective explainability patterns.

The subtype G sequences MK254638 and KY658701 are partial genomes present in the HIV-1 Complete Consensus Multiple Sequence Alignment sourced from the Los Alamos National Laboratory HIV-1 Database (until 2020), so they are not included in the dataset, and then, not depicted in the phylogenetic tree.

9.6 Analysis of Phylogeny and Explainability Patterns

9.6.1 Subtype G Philogenetic Group 1

Figure 9.2 contains the first group of subtype G sequences. All G sequences in this excerpt of the phylogenetic tree stem from an evolutionary divergence with the subtype A1G sequence (a recombinant between G and A1) KU168306. Furthermore, this sequence, in turn, originated from another evolutionary divergence with the subtype A2G sequence AF316544. The entire grouping of G sequences can be further subdivided into a series of clades. There exists an evolutionary branching that separates the upper clade (from KX389619 to JN248584) from the lower clades. Within this evolutionary branching, the sequence KJ948662 shares a common ancestor with the four lower clades.

As is evident, there is a significant clustering of G sequences in a specific area of the tree, indicating a clear phylogenetic relationship among them.

The finding that sequences A1G and A2G exhibit a specific combination of subtypes A and G is noteworthy, particularly considering the relationship identified between both subtypes in their accessory genes (see Section 9.2).

Considering the clustering of G sequences evident in this excerpt of the phylogenetic tree, I will subdivide it to conduct a detailed analysis by clades of this initial Group 1. Figure 9.3 presents in detail the first clade of the subdivision (Group 1 Clade 1). This clade, in turn, subdivides into two subclades. The first subclade encompasses nine sequences of subtype G. The other subclade includes two sister G sequences, namely KY658701 and JN248584.

Table 9.1 consolidates the CNN classification results, as well as the hot zones for each of the sequences within this Group 1 Clade 1.

The CNN classified nearly all sequences as non-recombinant with high confidence levels. There was a slight divergence in the case of KU168278, where the confidence level was slightly below 75%, although it was classified as non-recombinant. The only sequence classified as recombinant was FJ289266, with a confidence level

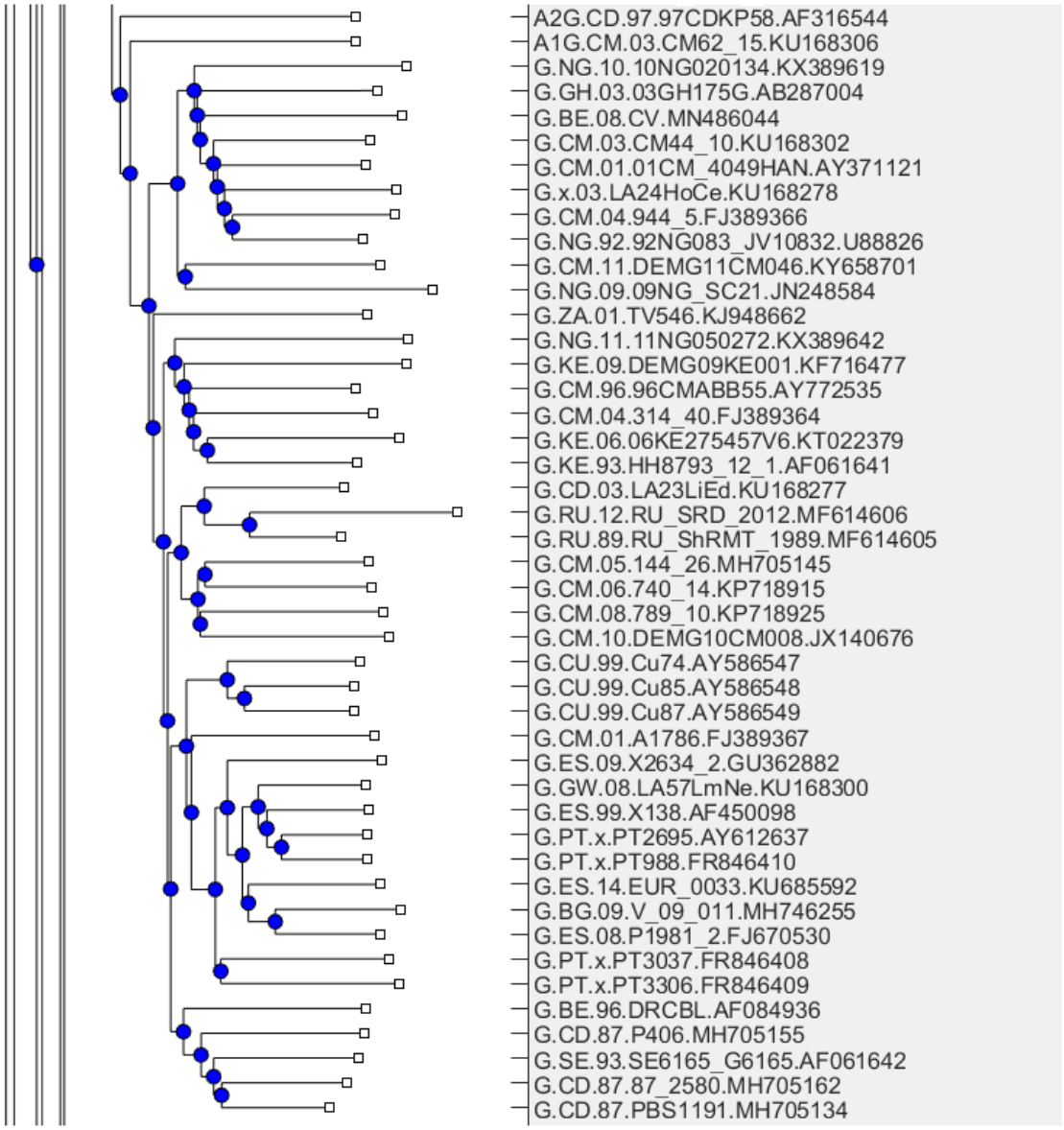


Figure 9.2: Subtype G Phylogenetic Group 1

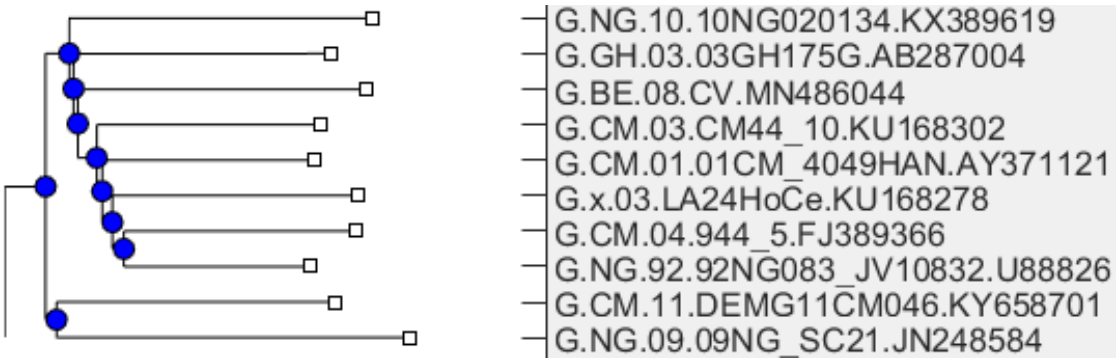


Figure 9.3: Subtype G Phylogenetic Group 1 Clade 1

close to 80%. The hot zones aligned consistently with their respective categories, and except for FJ289266, there was a clear correlation between the phylogenetic proximity and the similarity degree of explainability results.

Table 9.1: Subtype G Phylogenetic Group 1 Clade 1 Results

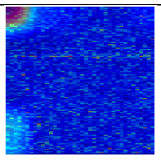
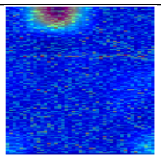
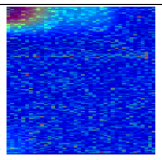
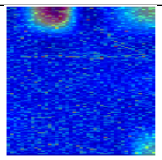
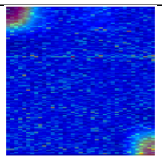
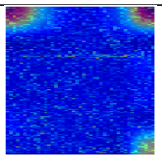
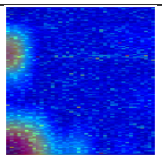
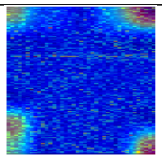
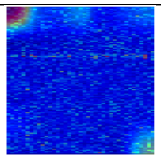
Sequence ID	Classification	Confidence Score	Grad-CAM
KX389619	Non-recombinant	87.054%	
AB287004	Non-recombinant	96.559%	
MN486044	Non-recombinant	95.916%	
KU168302	Non-recombinant	86.075%	
AY371121	Non-recombinant	89.034%	
KU168278	Non-recombinant	74.893%	
FJ389366	Recombinant	79.741%	
U88826	Non-recombinant	97.288%	

Table 9.1 – Continuation from the previous page

Sequence ID	Classification	Confidence Score	Grad-CAM
JN248584	Non-recombinant	94.672%	

Although included in the HIV-1 Complete Consensus Multiple Sequence Alignment from the Los Alamos National Laboratory HIV-1 Database (until 2020), KY658701 (HIV-1 isolate DEMG11CM046 from Cameroon, partial genome) is not listed in Table 9.1 as it is not a complete sequence and therefore not included in the dataset.

In the subtype G sequences classified as non-recombinant, there was coherence between the similarities of explainability patterns and phylogenetic proximity. I observed high degrees of similarity in the case of subtype G sequences MN486044 and KU168302. Additionally, notable is the high degree of similarity among KX389619, MN486044, AY371121, KU168278, U88826 (particularly with KU168278), and JN248584.

I observed a certain peculiarity in the hot zones of U88826, compared to the hot zones of the remaining sequences classified as non-recombinant. Despite the high confidence level that it is a non-recombinant sequence, its hot zones showed slight divergence from the rest. It is interesting to determine the existence of potential peculiarities in its genome that might be related to the recombinant feature of its sister sequence FJ389366 (clearly classified as recombinant with hot zones in line with that classification).

In Group 1 Clade 1, most of the G sequences were classified as non-recombinant. The sequence classified as recombinant, FJ389366, due to its high confidence level and the representativeness of its hot zones, is a good candidate for studying hidden genetic recombination phenomena or factors that may help determine the possible recombinant origin of this subtype.

I analyzed Group 1 Clade 2 of the phylogenetic tree depicted in Figure 9.4.

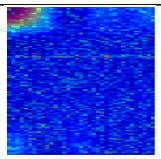
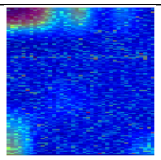
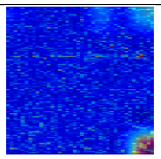
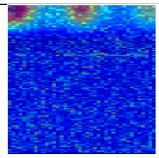
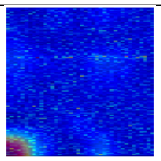
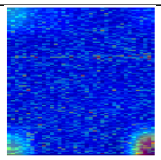
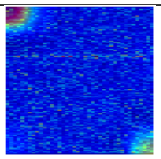
I must note that, although Sequence KJ948662 appears in Figure 9.4, it actually diverges not only with Group 1 Clade 2 but also with Clades 3, 4, and 5. This clade is entirely composed of G sequences. Within this clade, consecutive evolutionary branches are exhibited among the six sequences comprising the clade (excluding KJ948662).



Figure 9.4: Subtype G Phylogenetic Group 1 Clade 2

Table 9.2 shows the classification results from the CNN and their corresponding hot zones. The CNN classified almost all sequences as non-recombinant with varying degrees of confidence. Sequences like KX389642 had a confidence level of only around 61%, compared to AF061641 with a trust level close to 74.5%, and KF716477 with a degree of confidence exceeding 99%. The only classification as recombinant corresponded to sequence FJ389364, with a high level of confidence. In the case of sequences classified as non-recombinant, the hot zones aligned consistently with their category and demonstrate clear proportionality between the similarity of hot zones and phylogenetic proximity.

Table 9.2: Subtype G Phylogenetic Group 1 Clade 2 Results

Sequence ID	Classification	Confidence Score	Grad-CAM
KJ948662	Non-recombinant	99.914%	
KX389642	Non-recombinant	61.288%	
KF716477	Non-recombinant	99.266%	
AY772535	Non-recombinant	80.442%	
FJ389364	Recombinant	97.485%	
KT022379	Non-recombinant	97.569%	
AF061641	Non-recombinant	78.448%	

Their confidence levels were heterogeneous, ranging from 61.288% of KX389642 to 99.914% of KJ948662. Nevertheless, there was a correlation between phylogenetic proximity and the degree of similarity of the hot zones. Particularly noteworthy was the degree of similarity in the hot zones of sister sequences KT022379 and AF061641. Likewise, sequence KX389642 showed high similarity with KJ948662 and, to a lesser extent, with AY772535.

It is striking the high confidence level in classifying FJ389364 as recombinant

(97.485%). Its hot zones were consistent with the recombinant feature. Interestingly, it showed a certain degree of similarity with the hot zones of KT022379, a sequence clearly classified as non-recombinant and phylogenetically very close. It is interesting to analyze the homologies between the genomes of both sequences to try to determine common genetic elements that, in turn, explain their different classifications.

Similar to the scenario observed in Group 1 Clade 1, this phylogenetic area is characterized by the prevalent presence of non-recombinant sequences.

The next phylogenetic cluster to analyze is referred to as Group 1 Clade 3 in the phylogenetic tree depicted in Figure 9.5. At this juncture, I must note that the subtype G sequence KJ948662 exhibits evolutionary branches within Group 1 Clades 2, 3, 4, and 5.

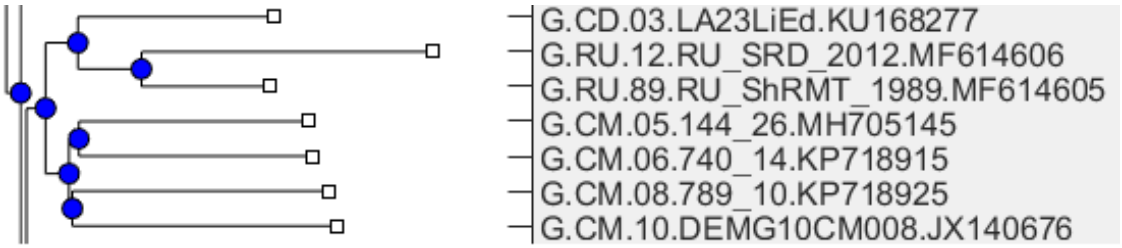
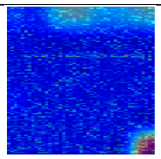
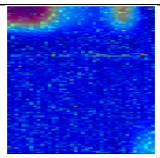
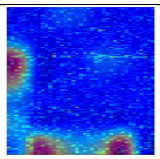
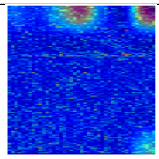
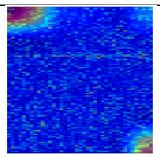
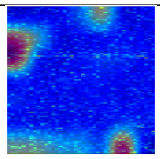
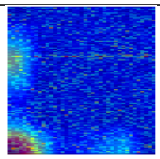


Figure 9.5: Subtype G Phylogenetic Group 1 Clade 3. This clade consists of seven sequences, subdivided into two subclades. All sequences belong to subtype G.

Table 9.3 displays the compilation of classification results and hot zones for the sequences that form this clade. The CNN classified four out of the seven sequences that form the clade as non-recombinant. Confidence levels varied from 55.22% in the case of MH705145 to 99.15% in the case of MF614606. The hot zones aligned with the non-recombinant category, although the relationship between phylogenetic distance and similarity degree was more ambiguous. On the other hand, the network classified 3 out of the seven sequences as recombinant. In the cases of KP718925 and JX140676, the confidence level of their classification was high. However, in the case of MF614605, the confidence level in its classification as recombinant was below 60%. The hot zones of these three sequences aligned with the recombinant category, and there was a correlation between the degree of similarity in the hot zones and their phylogenetic proximity.

Table 9.3: Subtype G Phylogenetic Group 1 Clade 3 Results

Sequence ID	Classification	Confidence Score	Grad-CAM
KU168277	Non-recombinant	71.045%	
MF614606	Non-recombinant	99.152%	
MF614605	Recombinant	59.337%	
MH705145	Non-recombinant	55.218%	
KP718915	Non-recombinant	99.821%	
KP718925	Recombinant	96.602%	
JX140676	Recombinant	98.180%	

In this case study, the CNN classified 43% of the sequences as recombinant. It is a noteworthy percentage that may indicate the presence of recombinant components within the seven sequences that form this clade.

Subdividing this clade into two subclades, the first of which comprises KU168277, diverges from the other two sister sequences within it (MF614606 and MF614605). KU168277 and MF614606 were classified by the CNN as non-recombinant, and their hot zones were similar. However, MF614605 was classified as recombinant,

although with a low confidence level, around 59%. Their hot zones aligned with this recombinant category. It could be of interest to conduct a detailed comparative study of the genomes of MF614606 and MF614605 to determine the location of the less homologous regions between both sequences and search for potential locations of hidden genetic recombination phenomena within MF614605.

In the second subclade, there is a notable concurrence between its phylogeny and the classifications made by the CNN. The first two sister sequences are classified as non-recombinant (MH705145 and KP718915), whereas the other two sister sequences are classified as recombinant (KP718925 and JX140676).

In the case of the non-recombinant sister sequences, the confidence level for classifying MH705145 as non-recombinant was low, close to 55%. However, for KP718915, the confidence level was 99.82%. Interestingly, the hot zones of these two sister sequences were divergent. The low confidence level of MH705145 might be influencing this discrepancy. There might be factors in its genome that render this sequence ambiguous in terms of a recombinant feature.

In the case of the recombinant sister sequences, their classification as recombinant exhibited a high confidence level in both cases, and there was a clear similarity between their hot zones.

Furthermore, their hot zones were similar to the other sequence classified as recombinant (MF614605). This phenomenon, in addition to being consistent with the phylogeny among these three sequences, denoted the existence of potential hidden recombination events in their genomes. These events could provide additional information regarding the origin of subtype G and its possible relationship with subtype A.

The hot zones of both recombinant sister sequences also exhibited similarities with the hot zones of FJ389366 (Table 9.1), although not as much with those of FJ389364 (Table 9.2).

The similarity between the hot zones of MF614606 (classified as non-recombinant with a confidence level of 99.15%) and KP718915 (classified as non-recombinant with a confidence level of 99.82%) with the hot zones of KJ948662 (classified as

non-recombinant with a confidence level of 99.91%, see Table 9.2) was notable. This sequence, KJ948662, was the result of an evolutionary branching with Clades 2, 3, 4, and 5.

The similarity in the hot zones of these three sequences was consistent with their strong non-recombinant characteristic and may denote the existence of common genomic elements among them.

In conclusion, it is remarkable that in this section of the tree, the presence of sequences classified as recombinant among G sequences reaches 43%, contrasting with previous cases where the majority consisted of G sequences classified as non-recombinant. In this scenario, the consistency between phylogeny and CNN classifications was particularly interesting. I must point out that phylogenetically close sequences were precisely classified as recombinant by the CNN. Therefore, it is plausible that there is some common characteristic among them that determines their recombinant feature.

A detailed analysis of these sequences classified as recombinant is necessary to attempt to elucidate hidden genetic recombination events within them, as well as to gather information that could reveal insights about the origin of subtype G.

Figure 9.6 shows the next clade of Group 1 (Group 1 Clade 4). This clade consists entirely of a total of fourteen subtype G sequences. There is a first subclade comprising three sequences and a second subclade consisting of eleven sequences grouped into several clusters. Subtype G sequence KJ948662 has evolutionary branches within Group Clades 2, 3, 4, and 5.

Table 9.4 displays the CNN classification results for these sequences, as well as their corresponding hot zones. Out of the fourteen sequences, the CNN classified only seven as non-recombinant. Among these seven sequences, most were classified with medium to low confidence levels. For four of these sequences, the trust degree range fell between 52% and 70%. In two cases, the confidence level range was slightly higher (86%-88%), and in one case, it reached 98%. 50% of the sequences were classified as recombinant (seven out of a total of fourteen). Two of these sequences reached a confidence level in the range of 69%-79%. For two others, the confidence

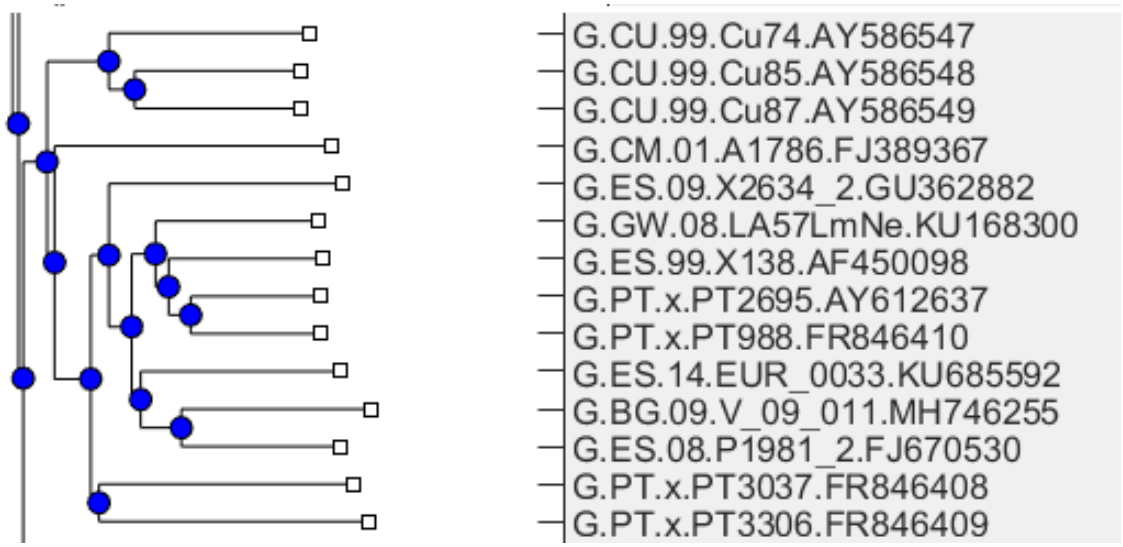


Figure 9.6: Subtype G Phylogenetic Group 1 Clade 4

level fell within the 60%-65% range. In three sequences, the confidence level ranged between 50% and 55%. The hot zones aligned consistently with their corresponding category. However, in this case, it is challenging to establish a direct correlation between phylogenetic proximity and the degree of similarity of the hot zones.

Table 9.4: Subtype G Phylogenetic Group 1 Clade 4 Results

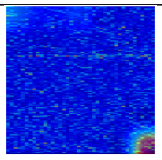
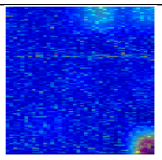
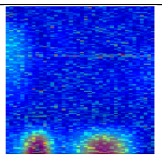
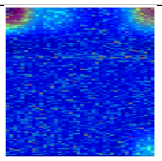
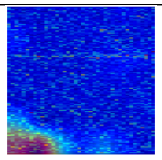
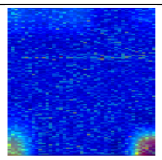
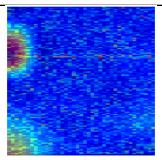
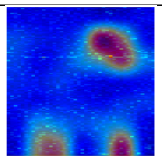
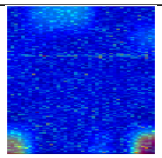
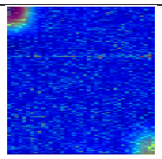
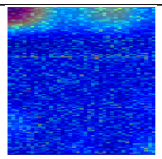
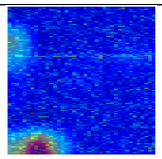
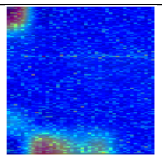
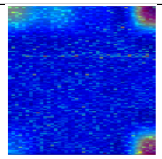
Sequence ID	Classification	Confidence Score	Grad-CAM
AY586547	Non-recombinant	52.077%	
AY586548	Non-recombinant	59.551%	
AY586549	Recombinant	69.026%	
FJ389367	Non-recombinant	88.919%	

Table 9.4 – Continuation from the previous page

Sequence ID	Classification	Confidence Score	Grad-CAM
GU362882	Recombinant	65.706%	
KU168300	Non-recombinant	86.844%	
AF450098	Recombinant	79.366%	
AY612637	Recombinant	51.362%	
FR846410	Recombinant	54.472%	
KU685592	Non-recombinant	60.288%	
MH746255	Non-recombinant	70.052%	
FJ670530	Recombinant	60.648%	
FR846408	Recombinant	53.208%	
FR846409	Non-recombinant	98.008%	

This clade represents a complex case where the majority of confidence levels returned by the CNN were unusually low. Out of the fourteen sequences, nine of them had a confidence level below 70%. In five of these sequences, the confidence level fell below 60%. Therefore, it is a clade composed of subtype G sequences that are ambiguous enough in terms of being recombinant or non-recombinant, causing issues for the CNN in categorizing them.

The phylogenetic alignment matching the cluster of problematic sequences for the CNN is highly intriguing. It is possible that the genomes of these sequences exhibit peculiarities related to the detected recombinant feature by the CNN, precisely resulting in those low confidence levels and the high percentage of classification as recombinant. The CNN classified 50% of the sequences within this clade as recombinant. This is an exceedingly high percentage, particularly when compared to the average error rates in the rest of the subtypes. This error rate is striking within a set comprising only fourteen sequences, which are also phylogenetically close.

Considering this, along with the low confidence level in some classifications as recombinant (and non-recombinant), I could possess sufficient data to suggest the possibility that in this case study, there might be information regarding the recombinant origin within subtype G, and these problematic sequences may contain relevant genomic fragments for this investigation.

There was a certain similarity between the hot zones of FJ389367 (classified as non-recombinant with a confidence level of 88.92%) and those of KU685592 (classified as non-recombinant with a confidence level of 60.65%) with the hot zones of KJ948662 (classified as non-recombinant with a confidence level of 99.91%, see Table 9.2). This sequence, KJ948662, results from an evolutionary branching with Group 1 Clades 2, 3, 4, and 5. Despite the different confidence levels in classifying these sequences as non-recombinant, the fact that the hot zones of these three sequences were similar might indicate the existence of common genomic elements among them regardless of the obtained confidence levels. This fact aligned with the phylogenetic relationships among them and is an interesting aspect to further explore in understanding the evolutionary history of this subtype G.

Upon detailed analysis of the upper subclade comprised of sequences AY586547, AY586548, and AY586549, I observed coherence among the hot zones and classification results concerning phylogeny. There is an evolutionary branching between AY586547 and the set formed by sister sequences AY586548 and AY586549. The CNN classified AY586547 as non-recombinant in an extremely uncertain manner, with a confidence level of 52.08%. Their hot zones were consistent with the non-recombinant feature. This same uncertainty in classification repeated in AY586548 and AY586549. AY586548 was classified as non-recombinant, but the confidence level was also low (59.55%). In contrast, sequence AY586549 was classified as recombinant with a moderate confidence level (69.03%). The degree of similarity between the hot zones of AY586547 and AY586548 was notable. In this case, there was a correlation between phylogenetic closeness and the degree of similarity of the hot zones. In the case of AY586548 and AY586549, despite being sister sequences, there was no evident similarity in their hot zones, probably due to their different classifications by the CNN.

It is possible that hidden genetic recombination events exist in AY586549. A detailed comparative study between AY586548 and AY586549 is necessary to attempt to identify non-homologous genomic regions where these phenomena might be located.

The lower subclade consists of eleven sequences. There is an evolutionary branching between FJ389367 and the other 10 sequences. FJ389367 was classified as non-recombinant with a confidence level of 88.92%.

Regarding the remaining 10 sequences, there is an evolutionary branching between the sister sequences FR846408 and FR846409 and the other eight sequences. Within these eight sequences, there is a further branching between GU362882 and the other seven sequences. GU362882 was classified as recombinant with a confidence level of 65.71%. Its hot zones were consistent with the corresponding recombinant feature and show a high degree of similarity with those of AF084936 (classified as recombinant with 98.74% confidence).

The sequence KU168300 was classified as non-recombinant with a confidence level of 86.84%. However, the sequences branching from KU168300 were classified as recombinant. In the case of AF450098, the CNN clearly classified it as recombinant, with a confidence level of 79.37%. Both sister sequences, AY612637 and FR846410, were classified as recombinant by the CNN, but their confidence levels were low (51.36% and 54.47%, respectively). It is striking that their hot zones did not resemble each other when a high similarity was expected.

It is also noteworthy the high degree of similarity between sequences FR846410 (recombinant with a confidence level of 54.47%) and KU168300 (non-recombinant with a confidence level of 86.84%). There was a certain similarity detected between the hot zones of AY612637 (classified as recombinant with a confidence level of 51.36%) and AY586549 (classified as recombinant with a confidence level of 69.03%).

Regarding the group formed by KU685592, MH746255, and FJ670530, the latter sequence was classified as recombinant by the CNN with a confidence level of 60.65%. KU685592 and MH746255 were classified as non-recombinant with confidence levels of 60.29% and 70.05%, respectively. Their hot zones showed a reasonable similarity in coherence with their phylogenetic closeness. However, the hot zones of MH746255 and its sister sequence FJ670530 were disparate due to their classifications in different categories. Noteworthy was the similarity between the hot zones of FJ670530 and those of AF450098.

Finally, the sister sequences FR846408 and FR846409 were classified in opposite categories by the CNN. In the case of FR846408, it was classified as recombinant with a low confidence level (53.21%). Its hot zones were consistent with the recombinant feature and, perhaps, with a slight resemblance to those of GU362882. On the contrary, CNN categorically classified FR846409 as non-recombinant, with a confidence level of 98.01%, and its hot zones were highly similar to those of FJ389367, which was also classified as recombinant with a confidence level of 88.92%. To a lesser extent, the hot zones of FR846409 were similar to those of other sequences also classified as non-recombinant, such as MH746255, AY586547, and AY586548.

Remarkably, in this clade, most of the phylogenetically sister sequences were classified in opposite categories by the CNN. Such is the case of FR846408 and FR846409, as well as MH746255 and FJ670530, and finally, the case of AY586548 and AY586549.

The case of AY586547, AY586548, and AY586549 is highly interesting. The samples for these three sequences were collected in Cuba on the same date (1999). The Patient Codes are CU74 for AY586547, CU85 for AY586548, and CU87 for AY586549. AY586548 and AY586549 were categorized by the CNN as non-recombinant with a degree of uncertainty, but AY586549 was already classified as recombinant. Their phylogenetic proximity aligns with the fact that samples were taken in a close temporal period and in the same geographic area. The low confidence levels in the classification as non-recombinant, coupled with the fact that AY586549 was classified as recombinant, might imply an evolutionary drift towards a recombinant feature.

In this case study, there were several particularities that could be highly significant regarding the "problematic" nature associated with subtype G. The remarkably high error rate in classification by the CNN was notable, with 50% of the sequences classified as recombinant. Their phylogenetic proximity suggests the presence of genomic components in these sequences with a recombinant load, which justifies both the high error rate and the inconclusive outcomes yielded by the CNN in most of these sequences.

The obtained results may support the potential recombinant origin of this subtype. I do believe that within this cluster of sequences, we can gather information about this recombinant origin. It is a promising candidate for study. Further research is needed.

Next, I present the analysis of the last clade corresponding to Group 1. Figure 9.7 displays Group 1 Clade 5. This clade comprises exclusively G sequences, totaling five sequences. Within these five sequences, a series of evolutionary branches occur, ultimately leading to the sister sequences MH705162 and MH705134.

I would like to highlight that the subtype G sequence KJ948662, exhibits evolutionary branching within Group 1 Clades 2, 3, 4, and 5.

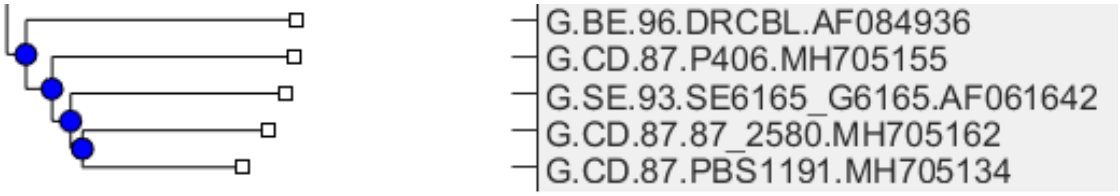
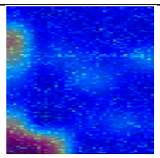
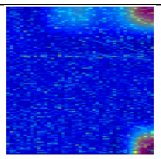
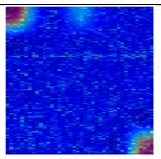
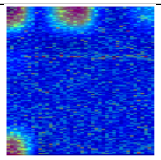
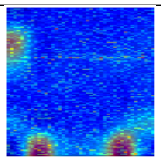


Figure 9.7: Subtype G Phylogenetic Group 1 Clade 5

Table 9.5 aggregates the classification results of all sequences composing Group 1 Clade 5, along with their respective hot zones.

The CNN clearly classified the sequence AF084936 as recombinant, with a confidence level close to 99%. Its hot zones aligned significantly with those corresponding to the recombinant category. The sequence MH705155 was classified as non-recombinant but with a confidence level of 59.60%. Thirdly, the sequence AF061642 was classified as non-recombinant with a high confidence level (over 91%). MH705162 was categorized as recombinant, yet the confidence level is slightly above 70%. In the case of MH705134, the CNN categorized it as recombinant with a confidence level of only 61.42%. The hot zones of the five sequences aligned significantly with those corresponding to their category. However, the degree of agreement between parental proximity and hot zone homology was highly constrained due to the high number of incorrect classifications. While in certain cases this agreement was more evident, it was not possible to assert this across the entire set of the five sequences.

Table 9.5: Subtype G Phylogenetic Group 1 Clade 5 Results

Sequence ID	Classification	Confidence Score	Grad-CAM
AF084936	Recombinant	98.735%	
MH705155	Non-recombinant	59.600%	
AF061642	Non-recombinant	91.067%	
MH705162	Non-recombinant	70.117%	
MH705134	Recombinant	61.419%	

Similar to the case of Group 1 Clade 4, the two sister sequences MH705162 and MH705134 were classified by the CNN into opposite categories. This accounted for the divergence between their hot zones. This recurring phenomenon of finding sister sequences with opposing classifications should be researched.

The hot zones of AF084936 (classified as recombinant with a confidence level of 98.74%) bore some resemblance to the hot zones of other sequences classified as recombinant, such as AF450098 (confidence level of 79.37%, see Table 9.4), JX140676 (classified as recombinant with a confidence level of 98.18%, see Table 9.3), or FJ389366 (confidence level of 79.74%, see Table 9.1).

Besides being consistent with the hot zones of the recombinant feature, this fact may indicate the existence of common elements related to a potential recombinant origin of subtype G.

MF614605 (Table 9.3) and MH705134 displayed highly analogous explainability patterns. Similarly, the hot zones of AY586549 (Table 9.4) also shared a certain degree of homology. All these sequences exhibited medium to low confidence scores in their CNN classifications. Upon observing KP718925 (Table 9.3), identified as recombinant with a confidence score of 96.602%, the patterns exhibited a certain degree of resemblance.

The CNN misclassified two out of five sequences in this clade. The hot zones of the sequences classified as recombinant (AF084936 and MH705134) were similar. Additionally, there was similarity among the hot zones of sequences classified as recombinant.

In this case, the comparison between phylogeny and the degree of similarity in the hot zones, while not yielding directly correlated results, could provide clues about the peculiarities that classify subtype G as "problematic".

Upon examining Group 1, I noticed a certain clustering of sequences classified as recombinant. This was evident in Group 1 Clade 3, where 43% of its sequences were classified as recombinant, or the complexity observed in Group 1 Clade 4 regarding CNN-provided results, with low confidence levels in multiple classifications as non-recombinant and 50% of sequences classified as recombinant. In Group 1 Clade 5, two out of the five sequences were classified as recombinant.

An interesting phenomenon arose from multiple cases where I observed two sister sequences (maximum phylogenetic closeness) where one was classified as non-recombinant while the other was labeled as recombinant. This disconnection between phylogenetic proximity and the recombinant status in several instances suggests the presence of highly homologous genetic factors that nonetheless determine the classification of a sequence as either recombinant or non-recombinant. A recombinant origin in subtype G could explain the compatibility between phylogenetic closeness and different classifications.

9.6.2 Subtype G Philogenetic Group 2

There is a predominant presence of subtype G sequences, with a minority presence of non-CRF recombinants 02G (recombinants between CRF02_AG and subtype G) and GK (non-CRF recombinant between subtypes G and K). There is an evolutionary divergence within the set of G sequences involving the 02G KX389627 sequence. Another evolutionary divergence among the G sequence set is observed with a clade consisting of three non-CRF recombinant 02G sequences. The grouping of G sequences can be subdivided into two visually distinguishable clades. The lower clade of G sequences includes the non-CRF recombinant sequence GK KX389638 and the 02G sequence U88825.

Figure 9.8 comprises the second set of subtype G sequences. Similar to the case of Group 1, most of the G sequences are clustered in this area of the phylogenetic tree. This is expected since all sequences belong to the same subtype.

I proceeded to divide Group 2 into two subclades with the purpose of conducting a visual and detailed analysis. Figure 9.9 details the phylogenetic structure of Group 2 Clade 1.

This subcluster is comprised entirely of fourteen subtype G sequences. There is an evolutionary divergence between KP109502 and subsequently with MK254637 from the rest of the G sequences. Following these two branches, the remaining G sequences split into two visually distinguishable subclusters.

Table 9.6 compiles the classification results performed by the CNN, as well as the hot zones of the G sequences present in Group 2 Clade 1.

Within this Group 2 Clade 1, most of the G sequences are clearly classified as non-recombinant, which is consistent with their phylogenetic proximity, as a high degree of genome similarity is expected.

The CNN practically classified all sequences as non-recombinant, with significant confidence levels. The exceptions were KP109502, with a relative confidence score (66.39%) despite being classified as non-recombinant. The sequences FJ389365 and DQ168576 were classified as recombinant with confidence scores of 75.94% and 95.41%, respectively. The hot zones of all sequences were coherent with their

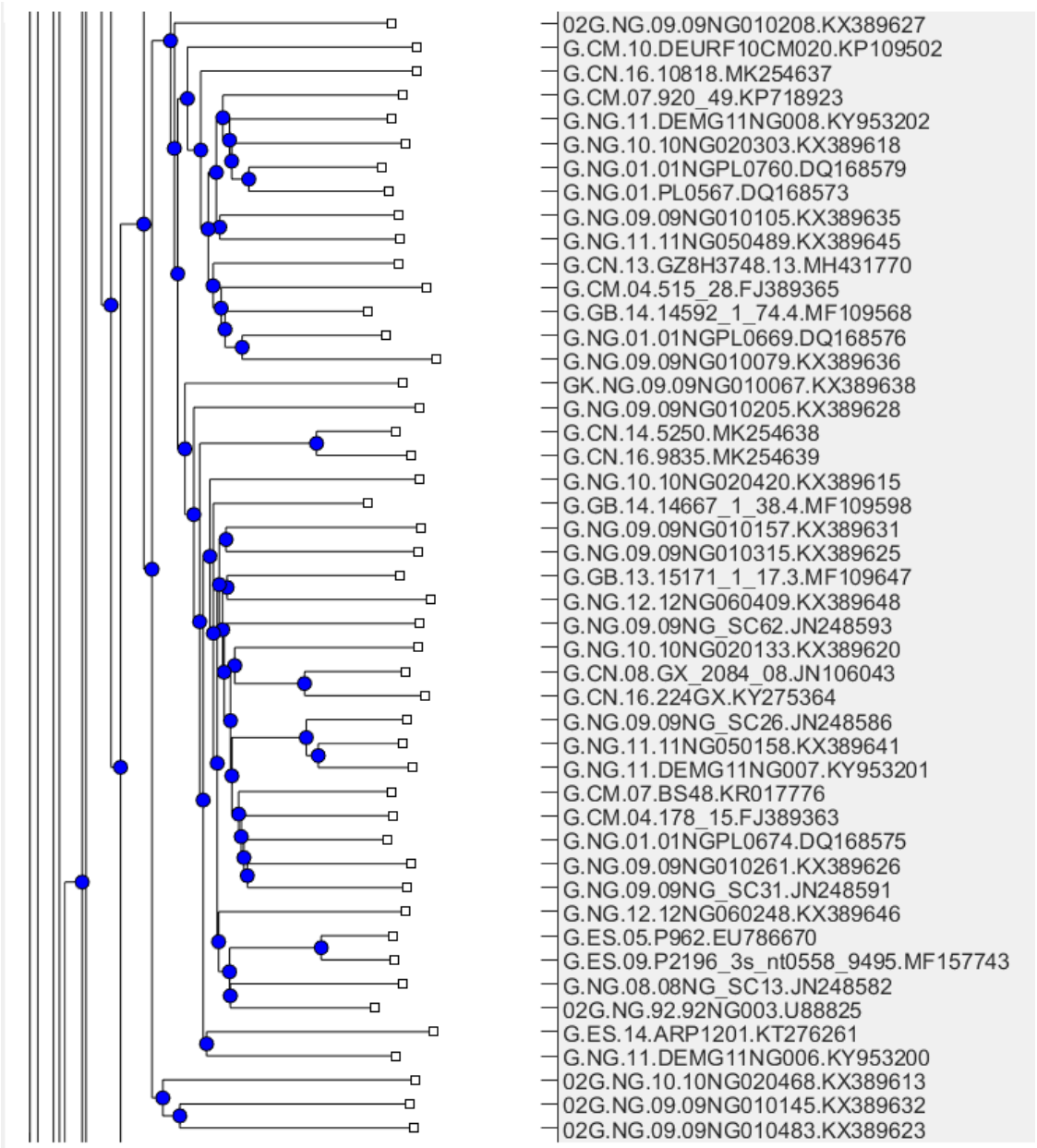


Figure 9.8: Subtype G Phylogenetic Group 2

respective categories, and there was a correlation between the phylogenetic distance and the degree of similarity among them.

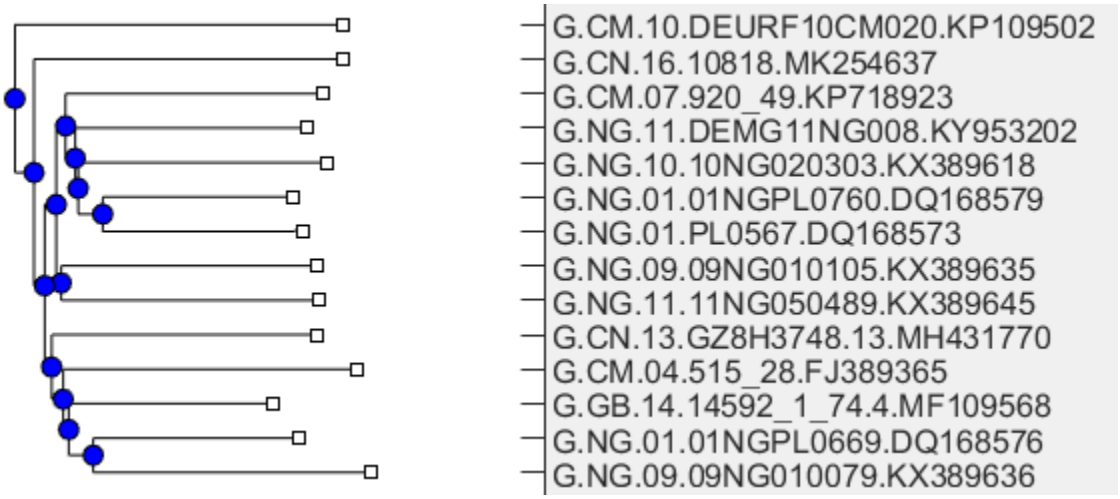


Figure 9.9: Subtype G Phylogenetic Group 2 Clade 1

Table 9.6: Subtype G Phylogenetic Group 2 Clade 1 Results

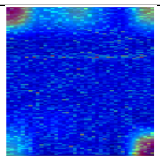
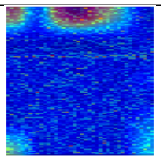
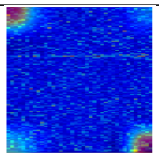
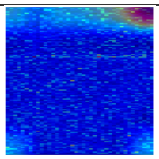
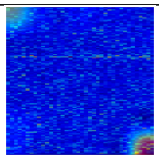
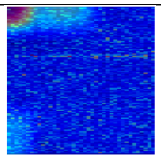
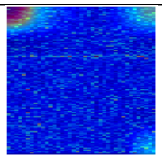
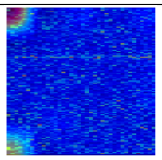
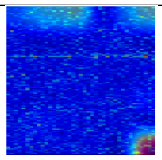
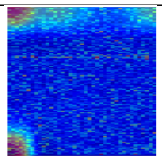
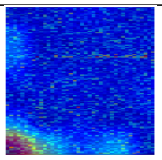
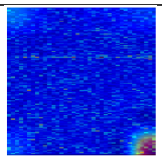
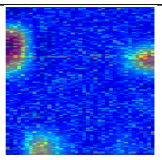
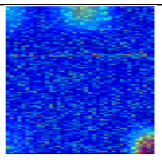
Sequence ID	Classification	Confidence Score	Grad-CAM
KP109502	Non-recombinant	66.358%	
MK254637	Non-recombinant	98.884%	
KP718923	Non-recombinant	99.697%	
KY953202	Non-recombinant	99.913%	
KX389618	Non-recombinant	91.092%	
DQ168579	Non-recombinant	98.203%	

Table 9.6 – Continuation from the previous page

Sequence ID	Classification	Confidence Score	Grad-CAM
DQ168573	Non-recombinant	98.712%	
KX389635	Non-recombinant	81.198%	
KX389645	Non-recombinant	98.702%	
MH431770	Non-recombinant	88.232%	
FJ389365	Recombinant	75.938%	
MF109568	Non-recombinant	95.335%	
DQ168576	Recombinant	95.409%	
KX389636	Non-recombinant	97.562%	

The fact that the G sequence farthest phylogenetically from the rest (KP109502) had a moderate confidence level in its classification, only 66.39%, is intriguing. The subsequent sequence branching evolutionarily from the rest is MK254637, which was also classified as non-recombinant but had a high confidence level (98.88%). The similarity degree in the hot zones of KP109502 and MK254637 was high and

consistent with their non-recombinant feature. Given their phylogenetic proximity, a clear correlation was observed between this closeness and the degree of similarity in their hot zones. It is possible that KP109502 contains genetic elements whose non-recombinant factor is weaker, more ambiguous, which might be the reason for its lower confidence level.

The remaining non-recombinant G sequences showed high similarity in their hot zones, both among themselves and with KP109502 and MK254637. There was generally a significant correlation between the degree of relatedness and the degree of similarity in the hot zones, as seen in the case of sister sequences DQ168579 and DQ168573. Another example was the similarity between KP718923 and KX389618. However, it is noteworthy that a higher degree of similarity in the hot zones was expected for sister sequences KX389635 and KX389645 given their high phylogenetic proximity. This discrepancy between the degree of similarity in the hot zones and their high phylogenetic proximity should be thoroughly studied.

As for sequences classified as recombinant, both confidence scores were noteworthy. In the case of FJ389365, the confidence level in its recombinant feature was 75.94%, while for DQ168576, it was 95.41%. Although their hot zones were coherent with their respective categories, a high degree of similarity between the hot zones of both sequences was not observed.

The hot zones of FJ389365 resembled those of sequences such as AF061640 (intra-patient from AF061641 and classified as recombinant with a confidence score of 52.58%, see Table 9.10), AF084936 (classified as recombinant with a confidence score of 98.74%, see Table 9.5), FJ389366 (classified as recombinant with a confidence score of 79.74%, see Table 9.1).

The hot zones of DQ168576 resembled those of sequences such as FJ670530 (classified as recombinant with a confidence score of 60.65%, see Table 9.4), JN248593 (classified as recombinant with a confidence score of 61.62%, see Table 9.7), or MH705134 (classified as recombinant with a confidence score of 61.42%, see Table 9.5). It is curious that, despite the high confidence level of DQ168576, sequences

with the greatest similarity showed relatively low confidence levels, around 60%. This phenomenon warrants further research.

The next case study is Group 2 Clade 2. The phylogeny of this group is visually depicted in Figure 9.10.

This clade is composed entirely of 29 subtype G sequences with multiple successive genetic branchings. There are approximately seven subclades observed, which represent groupings of phylogenetically closer sequences.

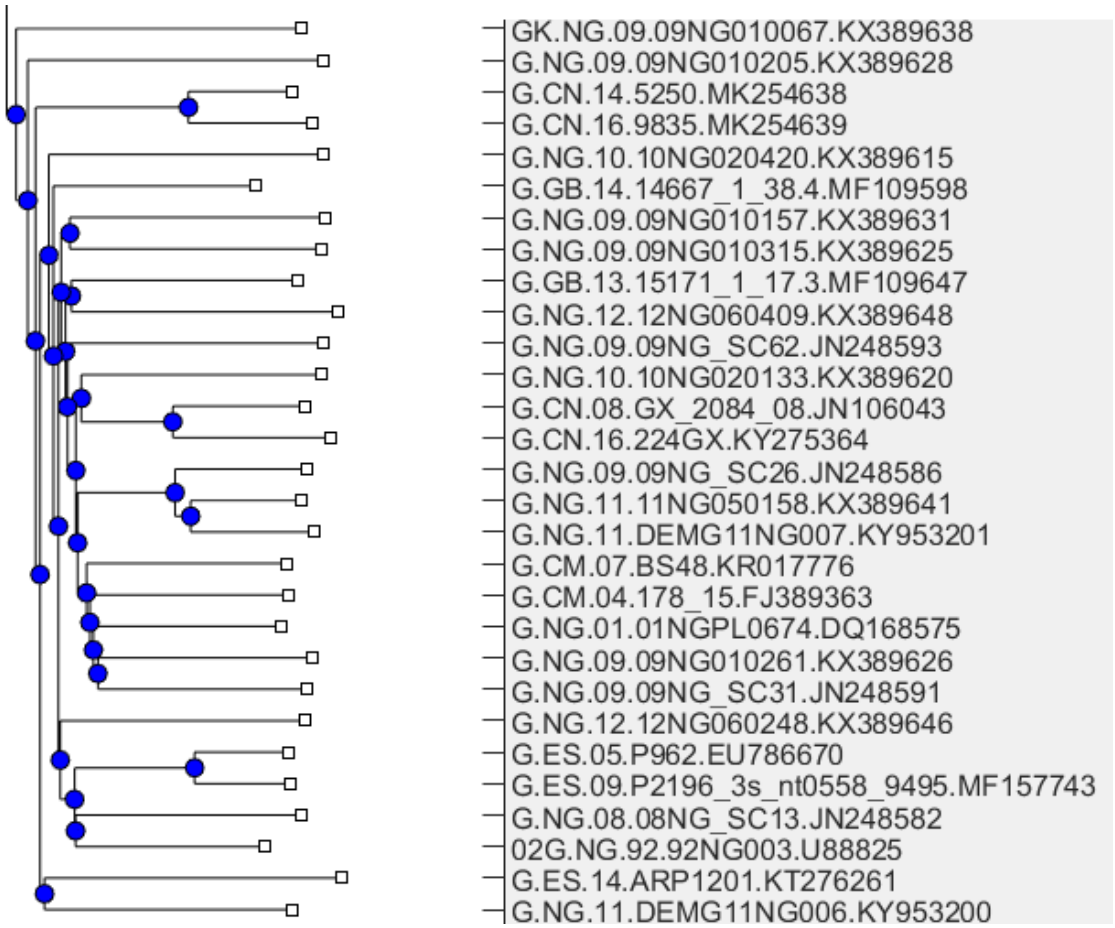


Figure 9.10: Subtype G Phylogenetic Group 2 Clade 2

Table 9.7 compiles the CNN classification results of the sequences belonging to Group 2 Clade 2, along with their respective hot zones. Most of the sequences were classified as non-recombinant. However, the confidence levels in the non-recombinant classification varied, ranging from 99.95% in the case of KX389615 to 56.53% in KT276261, including sequences such as KY953201, with a confidence

level of 60.07%. There were two sequences classified as recombinant, both with relatively low confidence levels. One of them was JN248593, with a confidence level of 61.62%, and the other sequence was JN248591, with a confidence level of only 51.69%. There was a third sequence classified as recombinant, FJ389363, with a confidence level of 98.29%. The hot zones were consistent with their respective category, and in general, there was a correlation between the phylogenetic proximity and the degree of similarity in the hot zones.

Table 9.7: Subtype G Phylogenetic Group 2 Clade 2 Results

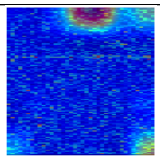
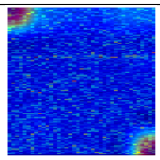
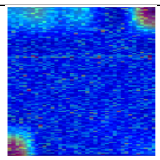
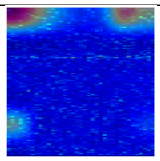
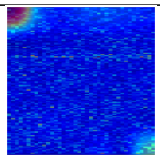
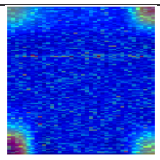
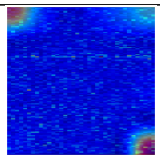
Sequence ID	Classification	Confidence Score	Grad-CAM
KX389628	Non-recombinant	96.122%	
MK254639	Non-recombinant	93.354%	
KX389615	Non-recombinant	99.948%	
MF109598	Non-recombinant	98.178%	
KX389631	Non-recombinant	89.406%	
KX389625	Non-recombinant	99.771%	
MF109647	Non-recombinant	69.453%	

Table 9.7 – Continuation from the previous page

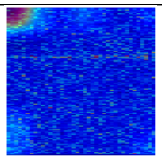
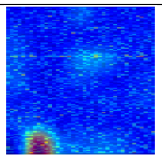
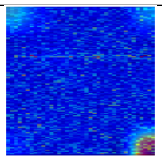
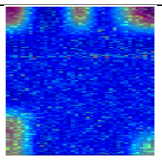
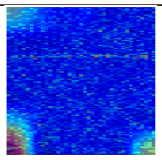
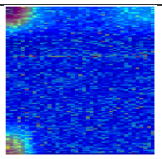
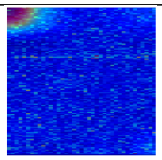
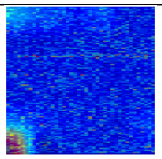
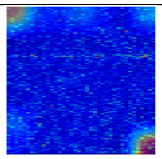
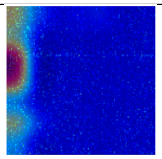
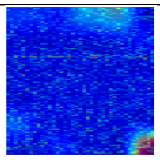
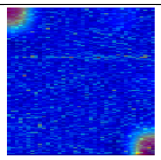
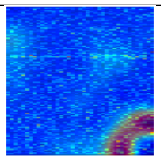
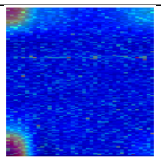
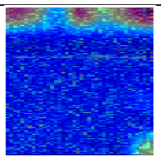
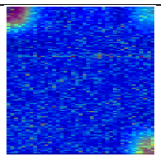
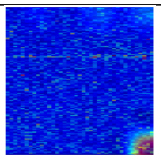
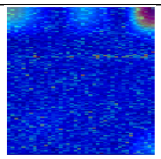
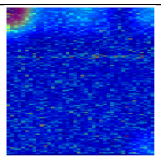
Sequence ID	Classification	Confidence Score	Grad-CAM
KX389648	Non-recombinant	88.867%	
JN248593	Recombinant	61.615%	
KX389620	Non-recombinant	99.147%	
JN106043	Non-recombinant	57.452%	
KY275364	Non-recombinant	99.843%	
JN248586	Non-recombinant	84.494%	
KX389641	Non-recombinant	57.324%	
KY953201	Non-recombinant	60.065%	
KR017776	Non-recombinant	99.521%	
FJ389363	Recombinant	98.285%	

Table 9.7 – Continuation from the previous page

Sequence ID	Classification	Confidence Score	Grad-CAM
DQ168575	Non-recombinant	99.994%	
KX389626	Non-recombinant	58.888%	
JN248591	Recombinant	51.692%	
KX389646	Non-recombinant	99.690%	
EU786670	Non-recombinant	59.745%	
MF157743	Non-recombinant	82.400%	
JN248582	Non-recombinant	99.939%	
KT276261	Non-recombinant	56.534%	
KY953200	Non-recombinant	75.915%	

Within this clade, most of the sequences were classified as non-recombinant. Additionally, a homology was observed between their hot zones and their degree

of relatedness. This is evident in sequences such as KX389631 and KX389648, KX389625 and MF109647, or KY953201 and FJ389363.

However, closer proximity in the hot zones of sister sequences such as KX389625 and KX389631, as well as MF109647 and KX389648, was expected.

Regarding sequence MK254638, it is worth noting that although it is included in the HIV-1 Complete Consensus Multiple Sequence Alignment from the Los Alamos National Laboratory HIV-1 Database (up until 2020), it is not listed in Table 9.1. This sequence is (quoted verbatim) *"HIV-1 isolate 5250 from China gag protein (gag) gene, complete cds; pol protein (pol) gene, partial cds; vif protein (vif) and vpr protein (vpr) genes, complete cds; tat protein (tat) and rev protein (rev) genes, partial cds; vpu protein (vpu) gene, complete cds; and envelope glycoprotein (env) gene, partial cds"* [41]. It is not a complete sequence and therefore is not listed in the dataset.

Additionally, the degree of concordance of hot zones between JN106043 (classified as non-recombinant with a confidence score of 57.45%) and EU786670 (classified as non-recombinant with a confidence score of 59.75%) was noteworthy. Their low confidence scores, combined with high homology in their hot zones, suggested the existence of common genomic factors that may indicate a low level of non-recombinant feature. A detailed analysis of their genomes is necessary to accurately discern these factors.

An interesting case involves JN248593 (classified as recombinant with a confidence score of 61.62%) and JN248591 (classified as recombinant with a confidence score of 51.69%). JN248591 (Patient Code: SC31) was sequenced from a sample taken in Nigeria in 2009. JN248593 (Patient Code: SC62) was also sampled from Nigeria around the same time. Therefore, there is geographical and temporal proximity in both samples collection, which aligns with their phylogeny, CNN results, and the degree of homology in their hot zones. Both sequences were classified as recombinant with low confidence scores, they are phylogenetically close, but their hot zones were divergent. Furthermore, they were also distinct from those of FJ389363.

In this case, there was a discrepancy among the hot zones of the three sequences classified as recombinants in this clade. It would be interesting to explore the

possibility that these three sequences have different genomic segments that somehow affect the recombinant feature. This might provide clues about a potential recombinant origin of this G subtype and potential evolutionary branches that, while maintaining a certain recombinant characteristic, result in different lineages.

This phylogenetic area called Group 2 has a predominant presence of non-recombinant sequences. In contrast to Group 1, there are no significant areas detected with the presence of sequences classified as recombinant. This fact is intriguing and apparently consistent with the phylogeny. The clustering of non-recombinant sequences in this area might provide us with new insights into the genetic evolution of this subtype G.

9.6.3 Subtype G Philogenetic Group 3

In Groups 1 and 2, I observed clusters of G sequences grouped in the same area. However, as seen in Figure 9.11, one of the G sequences is surprisingly isolated in the phylogenetic tree compared to the rest of the subtype G sequences.

This is the case of the subtype G sequence MN153484. Its phylogenetic relation to the subtype U sequence MN736702 prompted an analysis in Chapter 8 Section 8.5.10.

Phylogenetically, I encountered an exceptional case where this G sequence is isolated from the rest of its subtype sequences but is situated in a region of the tree predominantly occupied by recombinants related to the recombinant subtype CRF02_AG, whose parental subtypes are A and specifically G.

We should also take into account the phylogenetic proximity of a subtype U sequence. It would be interesting to meticulously determine the homologies between their respective genomes to try to identify homologous genomic regions that provide new evolutionary insights among the various subtypes involved in this case study, as well as additional information about the origin of subtype G.

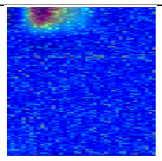
Based solely on the phylogenetic situation, this Group 3 is a very interesting case study that may provide insights into undiscovered details about the origin and evolution of subtype G, as well as its relationships with other subtypes.

The subtype U sequence MN736702 and a complex clade comprising several recombinant subtypes share a common unknown ancestor. The subtype G sequence MN153484 is part of a subcluster consisting of five sequences. Besides this G sequence, the rest includes two sister sequences of the non-CRF recombinant 02G subtype (a recombinant between the CRF02_AG and the G subtype itself) and two other sister sequences that are phylogenetically closer to MN153484. These latter two belong to the CRF20_BG and the non-CRF A1BG subtypes (a recombinant between the A1, B, and G subtypes). The remaining subclusters close to this subcluster are mainly composed of CRF02_AG sequences and non-CRF recombinants with CRF02_AG as their major parental subtype.

Table 9.8 displays the classification results and hot zones of the subtype G sequence MN153484. The CNN classified MN153484 as non-recombinant with a confidence score of 75.92%. Its hot zones were coherent with those corresponding to the non-recombinant category.

For readers interested in exploring the hot zones of adjacent sequences from other subtypes, I refer to the tables in Chapter 8 Section 8.5.10 Table 8.27.

Table 9.8: Subtype G Group 3 Results

Sequence ID	Classification	Confidence Score	Grad-CAM
MN153484	Non-recombinant	75.915%	

In this case study, there were several exciting factors to consider. Firstly, the phylogenetic isolation of this sequence from the rest of its subtype sequences. This could suggest that this particular sequence possesses a set of genomic characteristics that make it divergent from other sequences within its subtype. A comprehensive analysis is crucial to uncover the reasons for this divergence between this sequence and the other subtype G sequences. Additionally, delving into the exact relationship between this sequence and the subtype U sequence MN736702 is of particular interest.

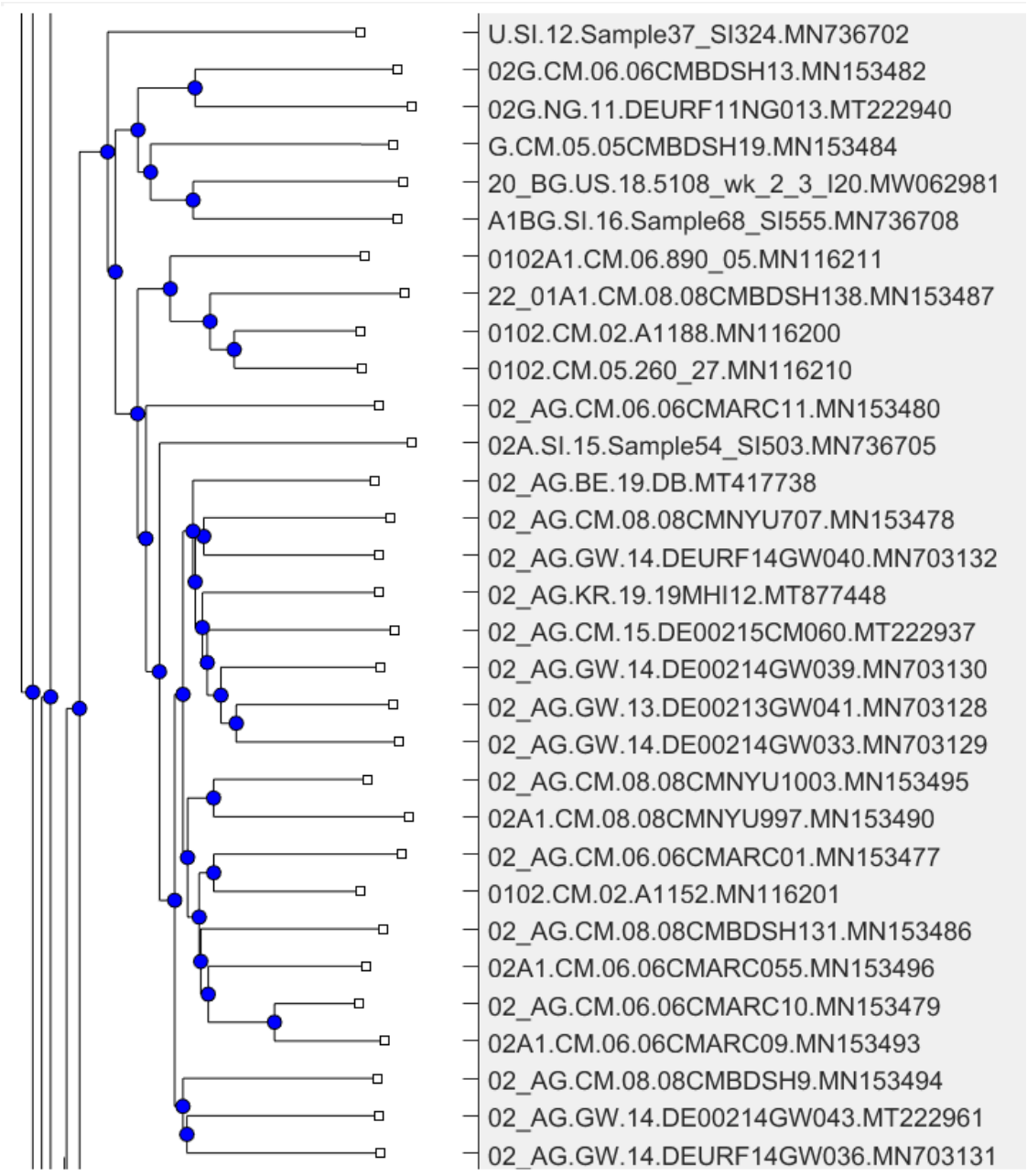


Figure 9.11: Subtype G Phylogenetic Group 3

On the other hand, it is surprising that the CNN classified this sequence as non-recombinant when its entire phylogenetic neighborhood comprises recombinant sequences. Even subtype U MN736702 is classified as recombinant, albeit with a confidence score of 52.86%. See Chapter 8 Section 8.5.10 Table 8.26.

A comparative analysis between subtype G MN153484 and adjacent sequences may shed light on the genomic elements divergent from those in other sequences that determine its classification as non-recombinant.

The hot zones of MN153484 closely resemble those of subtype G AB287003, AB287004, and, to a lesser extent, MN486044. These three sequences are located in the phylogenetic area corresponding to Group 1. It would be interesting to delve into the reasons behind these high similarities in their hot zones in order to uncover elements that may reveal unknown phylogenetic relationships.

Knowing in detail the factors that determine the isolation of this subtype G sequence from the rest may provide new insights into the origin and evolutionary development of the entire subtype G.

This sequence and its phylogenetic environment represent a fascinating case study.

9.7 Discussion

9.7.1 Comments on Subtype G Phylogeny

I must note that the difference between the sequences represented in the phylogenetic tree (approximately 83) and the 99 complete sequences available in the Los Alamos National Laboratory HIV Database until 2020 is due to the divergence in criteria for selection between the download of complete HIV-1 sequences from all subtypes from the Los Alamos National Laboratory HIV Database repository and the sequences included in the HIV-1 Complete Consensus Multiple Sequence Alignment from the Los Alamos National Laboratory HIV-1 Database (until 2020) [41].

As for the HIV-1 Complete Consensus Multiple Sequence Alignment from the Los Alamos National Laboratory HIV-1 Database (until 2020), the meticulous selection and review process by the global community of HIV-1 experts in selecting

representative sequences means that certain sequences are excluded from this multiple alignment, from which the phylogenetic tree is generated. For instance, since each patient is represented by a single sequence, subsequent intra-patient sequences are excluded from the alignment. Additionally, highly similar sequences and those considered problematic are removed. Consequently, out of the 99 identified complete sequences belonging to subtype G, only about 83 sequences are included in the consensus multiple alignment, and therefore, the remaining ones do not appear in the phylogenetic tree. The comprehensive definition of the inclusion criteria in this multiple alignment is detailed in Chapter 5, Section 5.12.

In this context, it is pertinent to offer several remarks:

- MK254638 (Patient Code: 5250) and KY658701 (Patient Code: MDC181) are not listed as complete genomes in the Los Alamos National Laboratory HIV Database repository. Despite being partial sequences, they are included in the HIV-1 Complete Consensus Multiple Sequence Alignment from the Los Alamos National Laboratory HIV-1 Database (until 2020). Hence, they are featured in the phylogenetic tree. The definition of MK254638 describes it as (quoted verbatim) *"HIV-1 isolate 5250 from China, encompassing the gag protein (gag) gene (complete cds), pol protein (pol) gene (partial cds), vif protein (vif) and vpr protein (vpr) genes (complete cds), tat protein (tat) and rev protein (rev) genes (partial cds), vpu protein (vpu) gene (complete cds), and envelope glycoprotein (env) gene (partial cds)"*. The definition of KY658701 is (quoted verbatim) *"HIV-1 isolate DEMG11CM046 from Cameroon, constituting a partial genome"*.

- In addition to these two sequences defined as partial, the phylogenetic tree includes 83 sequences out of the 99 complete subtype G sequences listed in the Los Alamos National Laboratory HIV Database repository. Therefore, a total of 85 subtype G sequences are positioned in the tree.

There are three main reasons that provide an explanation for this controversy. The first reason is the exclusion of intra-patient sequences in the HIV-1 Complete Consensus Multiple Sequence Alignment from the Los Alamos National Laboratory HIV-1 Database (until 2020), only one representative sequence per patient is

included. Second, certain sequences are considered partial in this consensus multiple alignment, whereas in the repository, they are categorized as complete genomes. Consequently, they are not included in the phylogenetic tree but are present in the dataset used. The third reason is the identification of certain sequences as problematic, leading to their exclusion from the consensus multiple alignment, although they remain in the repository.

The intra-patient sequences that were excluded from the HIV-1 Complete Consensus Multiple Sequence Alignment from the Los Alamos National Laboratory HIV-1 Database (until 2020) are as follows:

- AB231893 and AB287003. Those are intra-patient sequences of AB287004 (Patient Code: 175).

- AB485662, AB485663, and AF061640. Those are intra-patient sequences of AF061641 (Patient Code: HH8793).

- JX140674 and KX228812. Those are intra-patient sequences of EU786670 (Patient Code: P0962_ARP1077).

- JX140675. This is intra-patient sequence of KX228801 (Patient Code: X2634_ ARP1110).

- MF614607. This is intra-patient sequence of MF614606 (Patient Code: SRD).

The sequences considered partial in the HIV-1 Complete Consensus Multiple Sequence Alignment from the Los Alamos National Laboratory HIV-1 Database, which however are listed as complete in the Los Alamos National Laboratory HIV Database repository, are as follows:

- HM067749 (Patient Code not specified). The definition of this sequence is (quoted verbatim) *"HIV-1 isolate sh52 from China, partial genome"* [41]. I understand that it is considered problematic for this reason.

- KX228801 (Patient Code: X2634_ARP1110). The definition of this sequence is (quoted verbatim) *"HIV-1 isolate Read1_HIV_G from Spain, partial genome"* [41]. I comprehend that it is regarded as problematic for this particular reason.

- MT349406 (Patient Code: LB030). The definition of this sequence is (quoted verbatim) *"HIV-1 isolate 10CMLB030 from Cameroon, partial genome"* [41]. I understand that it is flagged as problematic due to this factor.

- MT349407 (Patient Code not specified). The definition of this sequence is (quoted verbatim) *"HIV-1 isolate 10CMLB040 from Cameroon, partial genome"* [41]. I understand it is marked as problematic because of these issues.

The problematic sequences that were consequently removed from the HIV-1 Complete Consensus Multiple Sequence Alignment from the Los Alamos National Laboratory HIV-1 Database are as follows:

- AF423760 (Patient Code: X558). Information available in the Los Alamos National Laboratory HIV-1 Database indicates that (quoted verbatim) *"Genome sequence lacks part of non-coding region. Sample from a female IDU taken in 2000 in Orense, Spain. This genome is considered to be representative of the parental strain of CRF14_BG"* [41]. I understand that it is considered problematic for these reasons.

- FJ670530 (Patient Code: 1981). Information available in the Los Alamos National Laboratory HIV-1 Database indicates that (quoted verbatim) *"Genome sequence lacks part of non-coding region. This sequence seems to be subtype G throughout the entire genome. It is much more similar to CRF14_BG than to other G, in the region where CRF14_BG is derived from subtype G. It thus seems to have shared a rather recent ancestor with the parental non-recombinant G, that recombined with B to create CRF14_BG"* [41]. I assume these are the reasons why it is considered problematic.

- MT334914 (Patient Code: 966). I did not find any data in the Los Alamos National Laboratory HIV-1 Database [41] that would give us information about this omission. I understand it is considered problematic.

In the context of the phylogenetic analysis, initially, 85 sequences were identified, with 16 sequences excluded from the set of 99 complete genome sequences from the Los Alamos National Laboratory HIV-1 Database repository. Therefore, the total number of subtype G sequences identified at this stage is 101.

Within the set of sequences included in the phylogenetic tree, I detected the presence of two sequences classified as partial, which despite being incomplete, are included in the HIV-1 Complete Consensus Multiple Sequence Alignment from the Los Alamos National Laboratory HIV-1 Database.

Out of these 101 sequences, 99 are classified and deposited as complete sequences in the Los Alamos National Laboratory HIV-1 Database repository.

Therefore, the control and inclusion of all relevant G sequences in the analysis were ensured.

9.7.2 Analysis of Hot Zones in Intra-patient Sequences

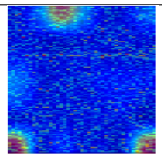
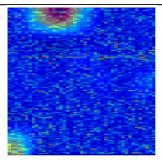
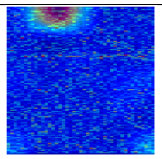
The intra-patient sequences refer to the genetic variability observed within an individual infected with HIV-1. In this context, it is expected that genetic variations among intra-patient sequences are minor, thus resulting in minimal phylogenetic distance [2549]. For this reason, the HIV-1 Complete Consensus Multiple Sequence Alignment from the Los Alamos National Laboratory HIV-1 Database contains only a single sequence per patient.

High similarity is expected among the hot zones of intra-patient sequences. To verify this, I analyzed all the intra-patient sequences categorized within subtype G.

The first case is that of AB287004 and its intra-patient sequences AB231893 and AB287003 (Patient Code: 175). Table 9.9 compiles the classification results from the CNN, as well as the resulting hot zones. This three intra-patient sequences were classified as non-recombinant, and their hot zones exhibited a high degree of similarity.

In this case, there was a high concordance between phylogenetic proximity, classification results, and the hot zones. The hot zones of AB287003 and AB287004 were nearly identical, while those of AB231893 showed a slight difference compared to the other two. This might be due to a greater evolutionary distance of AB231893 compared to the other two sequences. However, the results obtained by the CNN were highly robust and consistent with the close relationship among these three sequences.

Table 9.9: AB231893, AB287003, and AB287004 (Patient Code: 175)

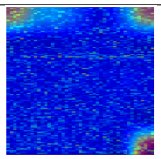
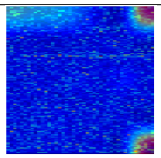
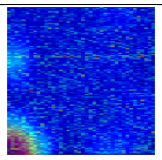
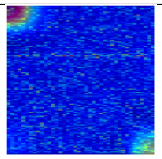
Sequence ID	Classification	Confidence Score	Grad-CAM
AB231893	Non-recombinant	98.911%	
AB287003	Non-recombinant	97.376%	
AB287004	Non-recombinant	96.559%	

The following case study involves AF061641 and its intra-patient sequences AB485662, AB485663, and AF061640 (Patient Code: HH8793). The classification results and hot zones are presented in Table 9.10.

The sequences AB485662, AB485663, and AF061641 were classified as non-recombinant. Confidence levels ranged from 67.11% for AB485663 to 87.52% for AB485662. The similarity degree of their hot zones was consistent with the non-recombinant feature and the close phylogenetic proximity of these sequences. Sequence AF061640 was classified as recombinant with a low confidence level of only 52.58%. The hot zones of this sequence diverged from the other three intra-patient sequences and were consistent with its recombinant nature.

The very high degree of similarity between the hot zones of AB485662 and AB485663 was noteworthy. This fact was consistent with their high phylogenetic proximity. The hot zones of AF061641 were slightly different, perhaps because this sequence harbors more genomic differences than those present between AB485662 and AB485663. Interestingly, sequence AF061640 was classified as recombinant, which was reflected in its hot zones. The divergence between the hot zones of this sequence and the rest of the intra-patient sequences may be driven by the CNN's detection of the recombinant feature. The presence of hidden genetic recombination

Table 9.10: AB485662, AB485663, AF061640, and AF061641 (Patient Code: HH8793)

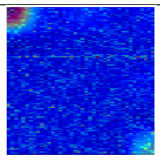
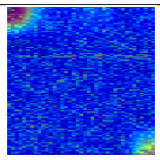
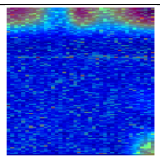
Sequence ID	Classification	Confidence Score	Grad-CAM
AB485662	Non-recombinant	87.516%	
AB485663	Non-recombinant	67.105%	
AF061640	Recombinant	52.583%	
AF061641	Non-recombinant	78.448%	

events in this sequence that do not manifest in the others cannot be ruled out. This phenomenon requires further research.

The following case study consists of EU786670 and its intra-patient sequences JX140674 and KX228812 (Patient Code: P0962_ARP1077). These three intra-patient sequences were classified as non-recombinant with varying confidence levels (Table 9.11). Ranging from a low level of 59.75% in the case of EU786670 to 88.91% in the case of KX228812. Their hot zones were consistent with their non-recombinant category and highly homologous between the hot zones of JX140674 and KX228812. The hot zone of EU786670 differed slightly from the rest of the intra-patient sequences.

In this case, all three sequences were classified as non-recombinant. It is interesting to note that the hot zones of EU786670 were slightly different from those of JX140674 and KX228812. One possible reason might be, within the presumed phylogenetic closeness of intra-patient sequences, a greater distance between EU786670 and the other two sequences. Furthermore, the confidence

Table 9.11: JX140674, KX228812, and EU786670 (Patient Code: P0962_ARP1077)

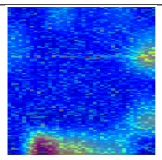
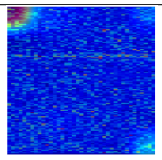
Sequence ID	Classification	Confidence Score	Grad-CAM
JX140674	Non-recombinant	67.745%	
KX228812	Non-recombinant	88.905%	
EU786670	Non-recombinant	59.745%	

level of EU786670 was less than 60%, while the other two intra-patient sequences had higher confidence levels. This might reflect the possible greater phylogenetic distance of EU786670, affecting the strength of the non-recombinant feature.

The next case study concerns KX228801 and its intra-patient sequence JX140675 (Patient Code: X2634_ARP1110). Table 9.12 summarizes the classification results by the CNN and the hot zones of both sequences. The sequence JX140675 was classified as recombinant with a confidence level above 70%. However, sequence KX228801 was classified as non-recombinant with a very high confidence level of 99.56%. Their hot zones were divergent and consistent with their respective classifications.

Interestingly, these two intra-patient sequences were classified differently by the CNN. According to the CNN, JX140675 was classified as recombinant with a substantial confidence level (70.17%). However, KX228801 was clearly classified as non-recombinant by the CNN, with a confidence level of 99.56%. The hot zones of both sequences were disparate, likely heavily influenced by the differing CNN classifications. This case study is highly significant in determining whether there are hidden genetic recombination phenomena in JX140675, which were detected by the CNN. It is also interesting to ascertain if there could be a relationship

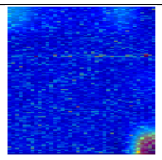
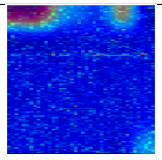
Table 9.12: JX140675 and KX228801 (Patient Code: X2634_ARP1110)

Sequence ID	Classification	Confidence Score	Grad-CAM
JX140675	Recombinant	70.169%	
KX228801	Non-recombinant	99.562%	

between the recombinant nature of this sequence and the possible recombinant origin of subtype G.

The last case study involves subtype G MF614606 and its intra-patient sequence MF614607 (Patient Code: SRD). Both sequences were classified as non-recombinant by the CNN with high confidence levels. Their hot zones were consistent with the non-recombinant category, yet they visually presented differences between each other (Table 9.13).

Table 9.13: MF614607 and MF614606 (Patient Code: SRD)

Sequence ID	Classification	Confidence Score	Grad-CAM
MF614607	Non-recombinant	97.817%	
MF614606	Non-recombinant	99.152%	

This case study is intriguing. Both sequences were clearly classified as non-recombinant and are intra-patient sequences, so their phylogenetic distance is presumably minimal. However, their hot zones were distinctly divergent, whereas a

high degree of similarity would be expected. This phenomenon is striking and requires further research.

The analysis of intra-patient sequences of subtype G proved puzzling due to the diversity of observed scenarios. Highly coherent cases were identified, where a direct correlation existed between phylogenetic proximity and the degree of similarity of the hot zones. However, surprisingly, cases with relative coherence were also documented. This variability in the results is highly significant and intriguing, potentially associated with factors that position subtype G as a "problematic" subtype within the specialized scientific community focused on HIV-1.

It is extremely interesting to delve deeper into the investigation of these intra-patient sequences, not only to determine additional information about subtype G evolutionary development and origin, but also to attempt to ascertain if there are specificities in subtype G replication and/or genetic recombination processes that could explain the variations observed even within their intra-patient sequences.

9.7.3 Prevalence Analysis of Recombinant Subtype G Sequences

Table 9.14 displays the prevalence analysis of recombinant sequences versus non-recombinant sequences for each analyzed phylogenetic zone and subzone.

I excluded Group 3 from Table 9.14 since in this fragment of the phylogenetic tree, only one subtype G sequence was present.

Upon examining the data presented in Table 9.14, I noticed a certain clustering of subtype G sequences classified as recombinant in Group 1, specifically in Clades 3, 4, and 5. In the remaining clades, the prevalence of recombinant sequences ranges around 11-15%. The most striking case is Group 1 Clade 4, where the recombinant prevalence reaches 50% in a relatively substantial clade comprising fourteen sequences.

The high rate of recombinant sequences in Group 1, Clades 3, 4, and 5 results in a total presence of recombinant sequences in Group 1 reaching 33.33%.

Table 9.14: Subtype G Recombinant Feature Prevalence

Phylogenetic Area	Total Number of Sequences	Number of Recombinant Sequences	Recombinant Prevalence (%)
Group 1 Clade 1	9	1	11.11%
Group 1 Clade 2	7	1	14.29%
Group 1 Clade 3	7	3	42.86%
Group 1 Clade 4	14	7	50.00%
Group 1 Clade 5	5	2	40.00%
Group 1 Total	**42**	**14**	**33.33%**
Group 2 Clade 1	14	2	14.29%
Group 2 Clade 2	26	3	11.54%
Group 2 Total	**40**	**5**	**12.50%**

Therefore, there is coherence between the phylogeny and the detected recombinant nature by the CNN. This implies that sequences classified as recombinant are peculiarly clustered nearby. This finding further supports the potential existence of a recombinant origin in this subtype, evident not only in the high error rates of the CNN but also in the phylogenetic causality of these problematic sequences.

In Group 2, the presence of 12.50% recombinant sequences is percentage-wise similar to that observed in Group 1 Clades 1 and 2.

Therefore, there is a phylogenetic relationship between sequences classified as recombinant and non-recombinant, showing a tendency for them to cluster together. It is necessary to delve deeper into this phenomenon to elucidate the origin and evolution of subtype G. The subset of sequences classified as recombinant is an appropriate subject for such study.

9.7.4 Analysis of Similarity Patterns among Subtype G Hot Zones

Considering the entirety of classification results and hot zones from the 99 complete subtype G sequences in the Los Alamos National Laboratory HIV Database repository, I grouped these sequences into clusters based on the degree of similarity in their hot zones.

This clustering may be insightful in attempting to elucidate information about the origin of this subtype and to uncover hidden or unknown phylogenetic relationships among the complete sequences that comprise subtype G.

For this purpose, I organized the sequences into visually distinctive clusters.

Given that the hot zones of the two categories classified by the CNN (recombinant and non-recombinant) were robust and distinguishable, it is anticipated that this clustering will be homogeneous in terms of the sequences comprising each group predominantly belonging to the same category.

The first cluster (referred to as Cluster 1) is comprised of sequences exhibiting visually detected hot zones in a band of high frequencies across the entire genome (Table 9.15), not just in the upper corners (areas corresponding to high frequencies in the LTRs, see Figure 9.1). These high-frequency hot areas may vary in significance, ranging from primary to secondary, yet are visually discernible at a glance.

This cluster consists of a total of 41 sequences. Nearly all sequences classifiable within this cluster are categorized as non-recombinant. Confidence levels in 26 sequences exceed 80%. This accounts for 63.41% of the total sequences in this cluster. Among the 41 sequences, seven have a classification confidence below 60%, representing 17.07% of the total. The only sequence in this cluster classified as recombinant is FR846410, with a confidence level of 54.47%. The hot zones align with the non-recombinant category. Interestingly, even the hot zones of FR846410 show a certain alignment with the total hot zones of the non-recombinant category.

Table 9.15: Subtype G Explainability Patterns Cluster 1. Subtype G sequences are listed in alphanumeric order.

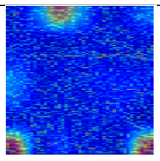
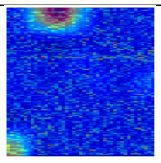
Sequence ID	Classification	Confidence Score	Grad-CAM
AB231893	Non-recombinant	98.911%	
AB287003	Non-recombinant	97.376%	

Table 9.15 – Continuation from the previous page

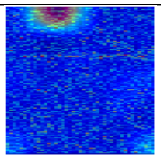
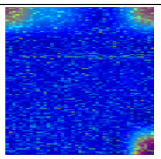
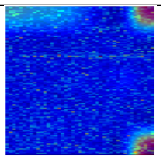
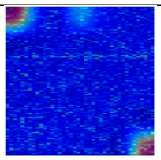
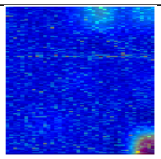
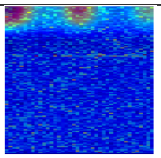
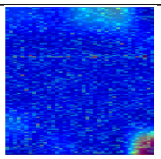
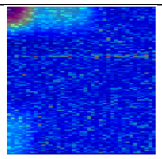
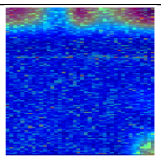
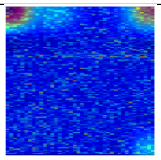
Sequence ID	Classification	Confidence Score	Grad-CAM
AB287004	Non-recombinant	96.559%	
AB485662	Non-recombinant	87.516%	
AB485663	Non-recombinant	67.105%	
AF061642	Non-recombinant	91.067%	
AY586548	Non-recombinant	59.551%	
AY772535	Non-recombinant	80.442%	
DQ168575	Non-recombinant	99.994%	
DQ168579	Non-recombinant	98.203%	
EU786670	Non-recombinant	59.745%	
FJ389367	Non-recombinant	88.919%	

Table 9.15 – Continuation from the previous page

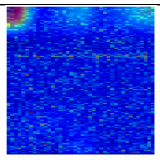
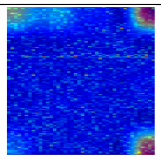
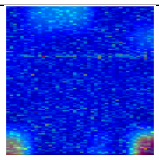
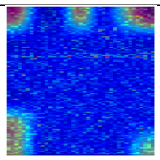
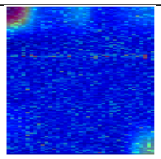
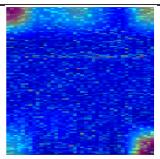
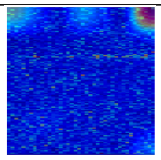
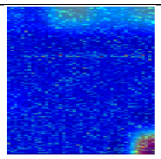
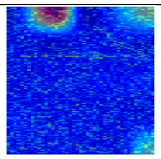
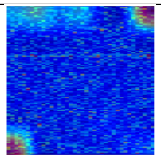
Sequence ID	Classification	Confidence Score	Grad-CAM
FJ670520	Non-recombinant	87.638%	
FR846409	Non-recombinant	98.008%	
FR846410	Recombinant	54.472%	
JN106043	Non-recombinant	57.452%	
JN248584	Non-recombinant	94.672%	
KP109502	Non-recombinant	66.358%	
KT276261	Non-recombinant	56.534%	
KU168277	Non-recombinant	71.045%	
KU168302	Non-recombinant	86.075%	
KX389615	Non-recombinant	99.948%	

Table 9.15 – Continuation from the previous page

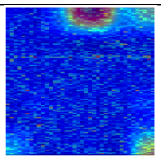
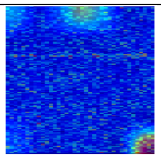
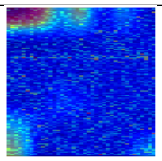
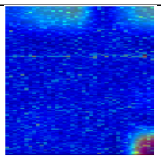
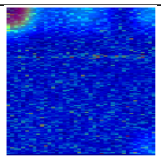
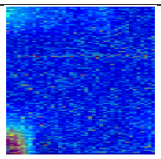
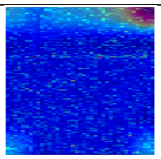
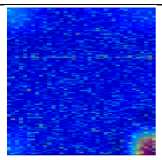
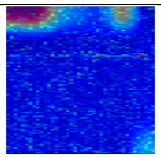
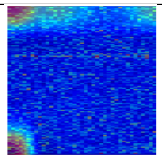
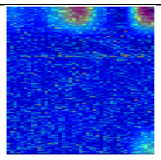
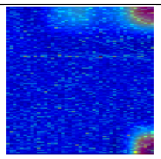
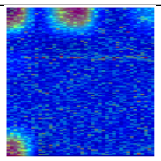
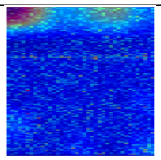
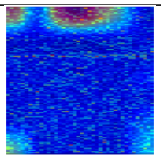
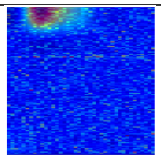
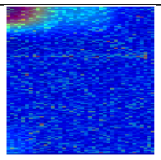
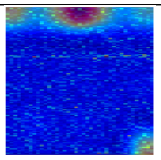
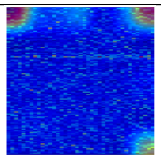
Sequence ID	Classification	Confidence Score	Grad-CAM
KX389628	Non-recombinant	96.122%	
KX389636	Non-recombinant	97.562%	
KX389642	Non-recombinant	61.288%	
KX389645	Non-recombinant	98.702%	
KY953200	Non-recombinant	75.915%	
KY953201	Non-recombinant	60.065%	
KY953202	Non-recombinant	99.913%	
MF109568	Non-recombinant	95.335%	
MF614606	Non-recombinant	99.152%	
MH431770	Non-recombinant	88.232%	

Table 9.15 – Continuation from the previous page

Sequence ID	Classification	Confidence Score	Grad-CAM
MH705145	Non-recombinant	55.218%	
MH705155	Non-recombinant	59.600%	
MH705162	Non-recombinant	70.117%	
MH746255	Non-recombinant	70.052%	
MK254637	Non-recombinant	98.884%	
MN153484	Non-recombinant	98.570%	
MN486044	Non-recombinant	95.916%	
MT334914	Non-recombinant	85.527%	
MT349406	Non-recombinant	86.871%	

This cluster is the largest among those established, accounting for 41.41% of the total 99 complete subtype G sequences. Although it is possible to further subdivide

this cluster into subgroups with a higher degree of similarity, for the purpose of this chapter's study, it is considered more appropriate to keep the 41 sequences in the same cluster in order to limit the complexity of the analysis conducted.

These sequences are distributed approximately equally between Group 1 and Group 2, so no conclusion can be drawn in this regard.

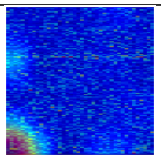
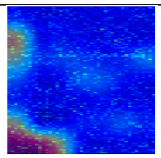
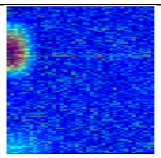
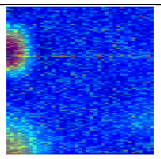
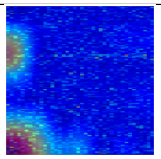
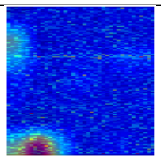
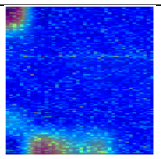
In any case, it seems that in a large part of the subtype G sequences in this cluster, the range of high frequencies across the entire genome plays a decisive role in detecting the non-recombinant feature.

The case of the sequence classified as recombinant (FR846410) is striking. Its confidence level was so low (54.47%) that it is plausible to consider that the CNN classified it erroneously. Despite this sequence being identified as recombinant, it is fascinating to observe the similarity of its hot zones with the rest of the sequences present in this cluster. This sequence is a good candidate for the search for genomic factors determining the recombinant feature.

I identified sequences belonging to Cluster 2 as those that have a main or secondary hot zone perfectly visible at $f = 1/3$ on the left-hand side of the genomic spectrogram (Table 9.16). This side corresponds to the 5'LTR. Additionally, the lower-left corner of the image must also appear as a main hot zone or be perfectly visible as a secondary hot zone. This area corresponds to the low-frequency zone of 5'LTR and/or *gag* and its surroundings. More information about the positioning of the hot zones concerning the genome map is available in Figure 9.1.

This cluster consists of seven sequences, all classified as recombinant. The confidence levels in the classification vary. Out of the seven sequences, three have a confidence level below 65%. Two sequences have confidence levels slightly below 80%. And two sequences have high confidence levels, exceeding 95%. The hot zones align with those corresponding to the recombinant category.

Table 9.16: Subtype G Explainability Patterns Cluster 2. Subtype G sequences are listed in alphanumeric order.

Sequence ID	Classification	Confidence Score	Grad-CAM
AF061640	Recombinant	52.583%	
AF084936	Recombinant	98.735%	
AF423760	Recombinant	95.373%	
AF450098	Recombinant	79.366%	
FJ389366	Recombinant	79.741%	
FJ670530	Recombinant	60.648%	
FR846408	Recombinant	53.208%	

The fact that this cluster consists entirely of recombinant sequences is logical and consistent with the hot zones corresponding to this category.

I must note that I included the sequence FR846408 in Cluster 2 because, although it does not exactly meet the standards that correspond to it, its high degree of similarity with the hot zones of the sequence FJ670530 led me to include it in this cluster.

These seven sequences are entirely located in the phylogenetic area corresponding to Group 1. It is an interesting case for study, as there is coherence among the phylogenetic positions, the recombinant feature detected by the CNN, and the degree of similarity in their hot zones.

I consider this an excellent case study to attempt to determine the genomic elements present in these sequences that cause the CNN to identify them as recombinant. Likewise, a detailed phylogenetic study is necessary to explore the existence of factors indicative of the recombinant origin of subtype G.

The third identified cluster includes sequences whose primary or secondary hot zones are perfectly visible in any of the upper or lower corners of the genomic spectrogram (Table 9.17). That is to say, in the high and low frequencies of the 5'LTR (left side) and 3'LTR (right side) and their respective surroundings. See Figure 9.1 for more information about the location of the hot zones on the genome map. There should be no hot zones in other areas unless they are practically imperceptible.

The CNN classified all sequences within this cluster as non-recombinant. Out of the 31 sequences comprising this cluster, three sequences exhibited confidence levels below 60%. This accounts for 9.68% of the total sequences in this cluster. Five sequences showed confidence levels within the range of 60-80%. This represents 16.13% of the total sequences. The remaining 23 sequences displayed confidence levels above 80%. This constitutes 74.19% of the total 31 sequences in this cluster. The hot zones were consistent with those corresponding to the non-recombinant category. The high degree of homology among the explainability patterns of these sequences correlates with their phylogenetic proximities.

Table 9.17: Subtype G Explainability Patterns Cluster 3. Subtype G sequences are listed in alphanumeric order.

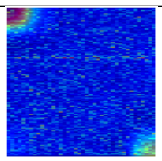
Sequence ID	Classification	Confidence Score	Grad-CAM
AF061641	Non-recombinant	78.448%	

Table 9.17 – Continuation from the previous page

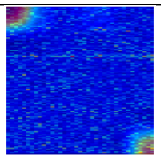
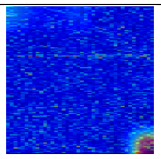
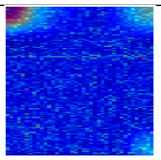
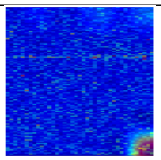
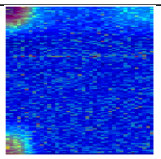
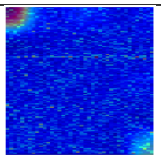
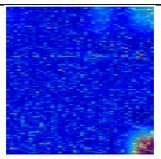
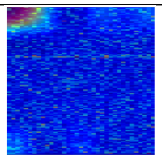
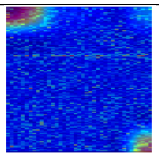
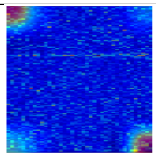
Sequence ID	Classification	Confidence Score	Grad-CAM
AY371121	Non-recombinant	89.034%	
AY586547	Non-recombinant	52.077%	
DQ168573	Non-recombinant	98.712%	
JN248582	Non-recombinant	99.939%	
JN248586	Non-recombinant	84.494%	
JX140674	Non-recombinant	67.745%	
KF716477	Non-recombinant	99.266%	
KJ948662	Non-recombinant	99.914%	
KP718915	Non-recombinant	99.821%	
KP718923	Non-recombinant	99.697%	

Table 9.17 – Continuation from the previous page

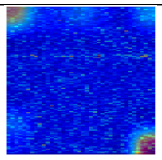
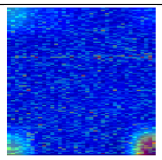
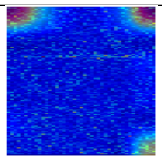
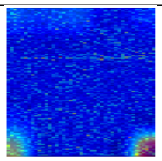
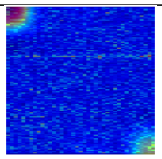
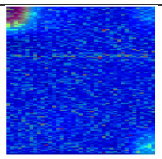
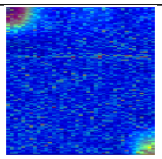
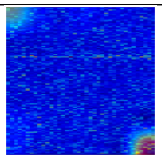
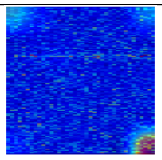
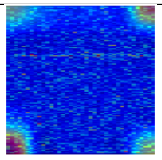
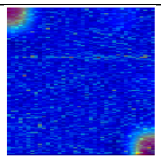
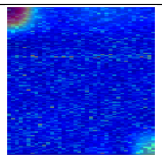
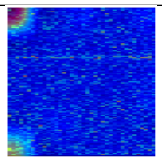
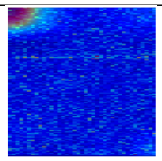
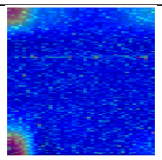
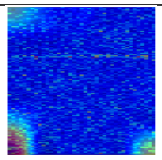
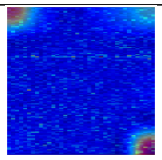
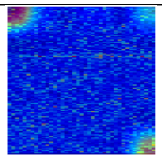
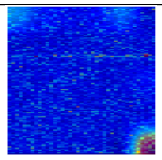
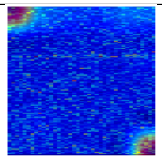
Sequence ID	Classification	Confidence Score	Grad-CAM
KR017776	Non-recombinant	99.521%	
KT022379	Non-recombinant	97.569%	
KU168278	Non-recombinant	74.893%	
KU168300	Non-recombinant	86.844%	
KU685592	Non-recombinant	60.288%	
KX228801	Non-recombinant	99.562%	
KX228812	Non-recombinant	88.905%	
KX389618	Non-recombinant	91.092%	
KX389620	Non-recombinant	99.147%	
KX389625	Non-recombinant	99.771%	

Table 9.17 – Continuation from the previous page

Sequence ID	Classification	Confidence Score	Grad-CAM
KX389626	Non-recombinant	58.888%	
KX389631	Non-recombinant	89.406%	
KX389635	Non-recombinant	81.198%	
KX389641	Non-recombinant	57.324%	
KX389646	Non-recombinant	99.690%	
KY275364	Non-recombinant	99.843%	
MF109647	Non-recombinant	69.453%	
MF157743	Non-recombinant	82.400%	
MF614607	Non-recombinant	97.817%	
MK254639	Non-recombinant	93.354%	

This cluster is the second largest after Cluster 1. Cluster 3 comprises 31 sequences, all of which were classified as non-recombinant by the CNN.

The homology degree of their hot zones was high despite the volatility in the confidence levels.

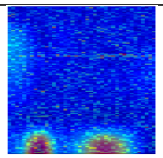
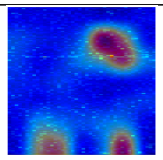
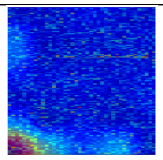
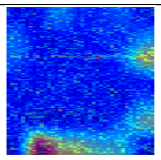
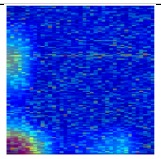
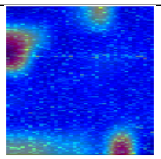
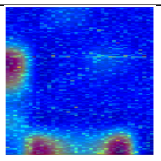
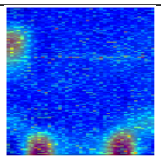
Phylogenetically, the 31 sequences are evenly distributed between the phylogenetic areas labeled as Group 1 and Group 2.

It is interesting that this clustering based on the degree of similarity of the hot zones affects only non-recombinant sequences. Therefore, this explainability pattern was exclusively linked to the non-recombinant feature. Identifying the genomic elements that cause this explainability pattern could provide information about the non-recombinant feature in these sequences.

The fourth cluster is composed of sequences whose main or secondary hot zones are perfectly visible in the vicinity of $f = 1/3$ in 5'LTR (or 3'LTR), especially in the low frequencies of the *gag* and *env* environments (Table 9.18).

Cluster 4 comprises eight sequences entirely classified as recombinant. 25% of the sequences had a confidence level below 60%, with one of them (AY612637) having a confidence level of only 51.36%. 50% had confidence levels ranging from 60-80%, and the remaining 25% had confidence levels exceeding 96%. The hot zones of the eight sequences were highly similar and consistent with those corresponding to the recombinant category.

Table 9.18: Subtype G Explainability Patterns Cluster 4. Subtype G sequences are listed in alphanumeric order.

Sequence ID	Classification	Confidence Score	Grad-CAM
AY586549	Recombinant	69.026%	
AY612637	Recombinant	51.362%	
FJ389365	Recombinant	75.938%	
JX140675	Recombinant	70.169%	
JX140676	Recombinant	98.180%	
KP718925	Recombinant	96.602%	
MF614605	Recombinant	59.337%	
MH705134	Recombinant	61.419%	

Like Cluster 2 (Table 9.16), all the sequences that comprise Cluster 4 are classified by the CNN as recombinant. Except for FJ389365, the rest of the sequences are located in the phylogenetic Group 1. We are unaware of the phylogenetic location of JX140675, as it is an intra-patient sequence of KX228801 (problematic

sequence, see Section 9.7.1).

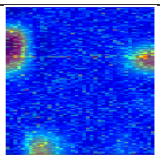
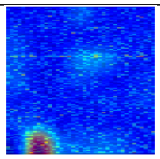
The hot zones of these sequences belonging to Cluster 4 exhibit an easily recognizable and distinct pattern compared to Cluster 2. It is an extremely interesting case study because the phylogenetic proximity of most sequences, coupled with this clear pattern of explainability, suggests the existence of distinctive genomic elements common in these sequences. These elements not only provide information about their origin but also determine their recombinant nature.

The peculiarity of their explainability patterns invites a thorough study of the genomes of the elements in this cluster in search of highly homologous fragments among them, distinguishable from the rest of the subtype G sequences, to extract unknown information about their common origin.

The sequences considered as belonging to Cluster 5 are those whose main and secondary hot zones are perfectly visible at $f = 1/3$ of 5'LTR and/or 3'LTR and in the low-frequency range in *gag* (Table 9.19). Consult Figure 9.1 to access further details regarding the positioning of the hot zones on the genomic map.

This cluster consists of two sequences classified as recombinant by the CNN with varying confidence levels. The confidence level for DQ168576 was 95.41%, whereas for JN248593, it was 61.62%. Their hot zones were consistent with the recombinant feature.

Table 9.19: Subtype G Explainability Patterns Cluster 5. Subtype G sequences are listed in alphanumeric order.

Sequence ID	Classification	Confidence Score	Grad-CAM
DQ168576	Recombinant	95.409%	
JN248593	Recombinant	61.615%	

Similar to Clusters 2 and 4, all sequences in this set were classified as recombinant by the CNN, and both are located in the phylogenetic zone referred to as Group

2. This case is coherent in terms of the correlation detected between phylogenetic proximity and the degree of similarity in their hot zones.

The explainability patterns were distinctive and differentiable from the rest of the clusters, even from those whose sequences are identified as recombinant. This may provide information about the common ancestors that may exist exclusively between these two sequences.

The Cluster 6 includes those G sequences whose primary and secondary hot zones are prominently located in the low-frequency range around 5'LTR and *gag* (Table 9.20). No visible hot zone is detected at $f = 1/3$ in the LTRs.

This Cluster consists of five sequences, all classified as recombinant. Three of them had confidence levels above 92%. The confidence levels of the other two sequences were within the range of 60-65%. The hot zones were consistent with their corresponding category and exhibit a high degree of similarity.

Table 9.20: Subtype G Explainability Patterns Cluster 6. Subtype G sequences are listed in alphanumeric order.

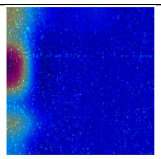
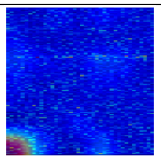
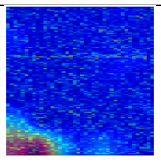
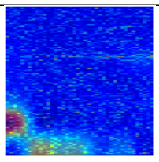
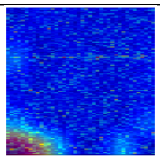
Sequence ID	Classification	Confidence Score	Grad-CAM
FJ389363	Recombinant	98.285%	
FJ389364	Recombinant	97.485%	
GU362882	Recombinant	65.706%	
HM067749	Recombinant	92.267%	

Table 9.20 – Continuation from the previous page

Sequence ID	Classification	Confidence Score	Grad-CAM
MT349407	Recombinant	61.462%	

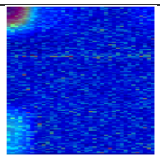
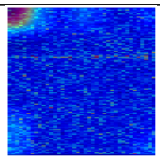
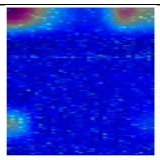
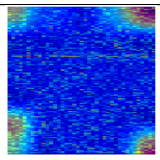
Despite FJ389363 being located in the phylogenetic zone named Group 2, while FJ389364 is in Group 1, the similarity between their hot zones is striking, as they were nearly identical. The resemblance between the hot zones of GU362882 and MT349407 was also interesting, considering these sequences have moderate confidence levels. This phenomenon might indicate the presence of common genomic fragments between them. We are unaware of the phylogenetic position of HM067749 and MT349407 since they are excluded from the HIV-1 Complete Consensus Multiple Sequence Alignment from the Los Alamos National Laboratory HIV-1 Database (until 2020) [41]. See Section 9.7.1.

Similar to Clades 2, 4, and 5, all sequences within this cluster were classified by the CNN as recombinant, displaying characteristic explainability patterns. It is interesting to further investigate the search for exclusive common genomic elements within this set of sequences to identify characteristics of a potential common ancestor among them.

The Cluster 7 is composed of sequences whose main or secondary hot zones were perfectly visible in the high frequencies of 5'LTR (upper left corner) and in the vicinity of $f = 1/9$ of 5'LTR (Table 9.21). Regardless of the existence of other main or secondary hot zones.

Cluster 7 consists of four sequences classified as non-recombinant, all with confidence levels above 87%. Their hot zones were consistent with the non-recombinant classification and exhibit high homology among themselves.

Table 9.21: Subtype G Explainability Patterns Cluster 7. Subtype G sequences are listed in alphanumeric order.

Sequence ID	Classification	Confidence Score	Grad-CAM
KX389619	Non-recombinant	87.054%	
KX389648	Non-recombinant	88.867%	
MF109598	Non-recombinant	98.178%	
U88826	Non-recombinant	97.288%	

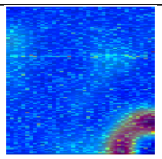
The explainability pattern of this cluster was clearly distinguishable from the rest of the clusters. KX389619 and U88826 are located in the phylogenetic zone referred to as Group 1, while KX389648 and MF109598 are situated in Group 2.

Especially interesting is the particular hot zone at $f = 1/9$ located in 5'LTR, as it provides a relevant clue to identify possible common ancestors among these sequences. The high confidence levels in their non-recombinant nature were also significant.

Finally, I left a single sequence into Cluster 8 (Table 9.22). Its explainability patterns bore no significant resemblance to the other 98 sequences.

JN248591 was classified as recombinant with a low confidence level, around 51.69%. Its hot zones were peculiar and unique within the subtype G, consisting of a ring-shaped section around the low frequencies in the 3'LTR.

Table 9.22: Subtype G Explainability Patterns Cluster 8

Sequence ID	Classification	Confidence Score	Grad-CAM
JN248591	Recombinant	51.692%	

It is intriguing that this sequence, classified as recombinant with a low confidence level, had such distinctive hot zones compared to the other 98 sequences. This sequence represents a unique case within subtype G. It is highly interesting to delve deeper into understanding this sequence at all levels to attempt to determine if it can provide clues about undetected evolutionary branches.

The clustering of subtype G sequences based on their explainability patterns was consistent in grouping sequences classified within a specific category. Consequently, except for JN248591 (the sole element of its cluster), non-recombinant sequences are predominantly grouped within the same clusters. Likewise, the same pattern was observed in the case of recombinant sequences.

Furthermore, I obtained additional information regarding the diverse explainability patterns within each of the clusters. This might provide us with insights into the existence of phylogenetic relationships among the sequences comprising each cluster.

The ability to accurately cluster the areas of interest, clearly identifying sets predominantly composed of homologous recombinant sequences in contrast to those consisting of non-recombinant sequences, is noteworthy. The diverse explainability patterns within these sets could yield significant insights into the underlying evolutionary divergences within this subtype. Specifically, the sets of recombinant sequences are of particular interest in this analysis.

Therefore, this analysis has proven to be extremely enriching.

9.7.5 Influence of dual infections

The multiple cases of sister sequences, including intra-patient sequences, with discordant classifications and/or whose hot zones did not exhibit the expected level

of similarity in sequences with very high phylogenetic proximity suggested to me the hypothesis that possible dual infections in the patients where these phenomena occurred might have had some degree of influence.

As is already known, dual infections lead to the emergence of recombinants and can be intra-subtype (different evolutionary branches of the same subtype) or inter-subtype (different subtypes). These dual infections are classified as co-infections or super-infections [2550]. A co-infection involves the simultaneous infection of two different strains or subtypes [2551], while superinfection refers to a second infection after a specific immune response to the first infecting strain has developed [2552].

While there are numerous documented cases of co-infection [2553–2562], the in-depth study of the phenomenon of superinfection came later and is still less well-known.

However, the current higher survival rate among HIV-1 infected individuals [2563], combined with lower awareness of taking protective measures during sexual intercourse, and the resulting increase in transmission rates of all types of STIs [2564–2566], will most likely lead to an increase in cases of HIV-1 superinfection.

Superinfection is associated with accelerated disease progression in certain patients [2567, 2568] (although not always [2569]) and may hinder the development of effective HIV-1 vaccines [2570].

There are multiple case studies of HIV-1 superinfection [2571–2576], and there is growing interest in this phenomenon [2577–2579]. In any case, multiple voices within the HIV-1 expert community indicate the existence of genetic recombination phenomena following a superinfection [2578, 2580–2582].

The hypothesis I propose at this point is whether past or ongoing dual infection phenomena in the patients might somehow influence the anomalous behaviors found in subtype G. Perhaps these effects manifest more clearly in one of the sister sequences, leading to the different classifications. Additionally, they might be reflected in greater divergence in the hot zones of intra-patient sequences, where each of the involved sequences could be closer to one of the original elements of the dual infection.

Verifying this hypothesis requires an in-depth microbiological and phylogenetic study focused on identifying the effects of these possible dual infections, and it may not be possible to obtain all the necessary information due to a lack of historical data.

Nevertheless, it is a hypothesis that I consider highly interesting, both for providing data that could explain these peculiar behaviors in sequences that are phylogenetically very close, and for shedding light on the origin of this mysterious subtype G.

9.8 Conclusions

Through a CNN, I identified singularities in a specifically problematic subtype with a potential recombinant origin. In this Chapter, I went beyond mere classification and identification of the areas of interest. I am substantiating the conclusions of multiple studies regarding the manifest peculiarities of subtype G from a wholly innovative perspective.

The error rate exhibited by the CNN in subtype G may be related to the subtype's recombinant origin, warranting an in-depth analysis of this matter. It is conceivable that this error rate goes beyond mere inaccuracies by the CNN; rather, this neural network is detecting genomic components related to the frequency spectrum, indicating the potential recombinant origin in specific sequences.

There are multiple factors drawing attention besides the high error rate. These include the presence of numerous cases where the non-recombinant feature is associated with low confidence levels, the similarity between hot zones of sequences identified in different categories, the phylogenetic clustering among sequences with similar CNN result peculiarities, and their hot zones, among others.

It is interesting to observe the significant variability in confidence levels, both in the classification as recombinant or non-recombinant. This subtype exhibits remarkable diversity across multiple aspects, not solely confined to classifications.

Although my approaches to uncover the underlying causes of this anomalous behavior of the CNN have not yielded conclusive and unambiguous results, I was able

to provide supplementary information to guide future investigations in unraveling the mysteries surrounding subtype G.

The phylogenetic analysis does not seem to reveal definitive results, although it demonstrates a clear tendency towards the proximity of both recombinant and non-recombinant sequences. This pattern aligns with previous observations regarding the relationship between the similarity of the areas of interest and phylogenetic proximity, as discussed in Chapter 5 Section 5.15.2.

In general, sequences classified as recombinant by the CNN tend to cluster within the same phylogenetic zones. This tendency may stem from a common, at least partial, origin among them. It would be interesting to delve deeper into the relationship between this phenomenon and scientific studies that propose a recombinant origin in this subtype.

In this subtype G, a curious and recurring phenomenon occurs in multiple cases. That is, phylogenetically related sequences are classified into disparate categories. One of these sequences is classified by the CNN as recombinant while its counterpart is labeled as non-recombinant. The repeatability of this phenomenon necessitates an exploration of the underlying reasons in these cases.

The analysis of intra-patient sequences has yielded surprisingly diverse outcomes. I identified highly coherent cases, where the hot zones are notably similar. However, I detected cases of non-recombinant sequences with markedly different hot zones. This variability in the hot zones requires further investigation to attempt to unravel the reasons behind such divergence.

Additionally, multiple study cases were identified where the CNN produced conflicting classifications. These latter cases are particularly intriguing, especially considering their recurrent detection. This phenomenon suggests a potential association with hidden genetic recombination events and/or the recombinant origin of this subtype.

Furthermore, I identified potential instances of latent genetic recombination even within intra-patient sequences via the pre-trained CNN. It would be of interest to pinpoint the exact location of the genomic regions that define the recombinant

trait in this context and investigate their potential linkage with the recombinant origin of subtype G.

The subtype G sequence MN153484 is an intriguing and highly significant case study. This sequence appears to be phylogenetically isolated from the rest of the subtype G sequences. There seems to be a certain degree of relatedness to the subtype U sequence MN736702. Identifying the genetic elements causing this evolutionary separation from the rest of its subtype sequences, as well as determining the exact relationship with subtype U MN736702, could provide valuable insights into the evolutionary ramifications affecting both subtype G and its relationship with sequences categorized under subtype U.

Another extremely intriguing case is that of subtype G JN248591, classified by the CNN as recombinant with a confidence level of only 51.692%, and whose hot zones notably differ from the explainability patterns of all other G sequences. A detailed genomic analysis would be highly interesting to delve into the reasons that make this sequence unique within its subtype concerning its hot zones.

The ability to accurately cluster areas of interest, identifying distinctly discernible sets predominantly composed of recombinant sequences on one hand and non-recombinant sequences on the other, is remarkable. The diverse and distinguishable explainability patterns within these clusters can provide valuable information about the underlying evolutionary ramifications in this subtype, with the patterns observed in the clusters of recombinant sequences being particularly interesting.

Leveraging a novel approach, such as Deep Learning technologies, I extracted multiple data points that could shed light on several current mysteries within this subtype.

Ultimately, subtype G is a fascinating subtype that can provide a wealth of information not only about the recombinant feature in HIV-1 but also about its phylogenetic relationships and evolutionary diversifications.

But I don't want to go among mad people," Alice remarked. "Oh, you can't help that," said the Cat: "we're all mad here. I'm mad. You're mad." "How do you know I'm mad?" said Alice. "You must be," said the Cat, "or you wouldn't have come here..

— "Alice in Wonderland" (1865), Lewis Carroll,*a microbiology enthusiast, he purchased one of the most modern microscopes of his time. With it, he devoted himself to observing microorganisms and insect larvae.*

10

Comprehensive Analysis of the Genomic Elements of SARS-CoV-2

Contents

10.1 Introduction 545
10.2 Genomic Protein Analysis in SARS-CoV-2 546
10.3 Equipment 561
10.4 Methods 562
10.4.1 Sequence Preprocessing 562
10.4.2 Sequence Filtering 565
10.4.3 Fragmentation of Filtered Sequences 566
10.4.4 Determination of Consistency Ratios 567
10.4.5 Calculation of Consistency Ratios 571
10.5 Results 572
10.5.1 Filtering Results 572
10.5.2 Consistency Level Results 579
10.6 Discussion 581
10.7 Conclusions 587

10.1 Introduction

At this point, the block of chapters corresponding to the analysis of genetic recombination in HIV-1 has concluded. The next block corresponds to a similar analysis in SARS-CoV-2, building on the knowledge acquired from HIV-1.

Prior to the study of genetic recombination in SARS-CoV-2 using a methodology based on genomic spectrograms, pre-trained CNNs, and explainability tools, I wanted to conduct a detailed analysis of the genomic elements of SARS-CoV-2. This analysis is based on statistical approaches to determine the degree of conservation of each element as a measure of its stability.

This detailed analysis is merely an approach to the genomic regions most prone to undergo modifications, whether mutations or genetic recombination events.

This pre-analysis may serve also as a basis for future research aimed at the effective implementation of algorithms focused on predicting the most likely viral evolutions in SARS-CoV-2, as well as in other viruses or microorganisms in future stages of investigation.

This Chapter consists of a section explaining the composition of the complete SARS-CoV-2 genome in terms of the proteins it encodes, to facilitate the understanding of the analysis performed (Section 10.2). Section 10.3 describes the equipment used, and Section 10.4 outlines the designed methodology. Section 10.5 presents the results regarding the selection of sequences with complete and correct EMBL (European Molecular Biology Laboratory) labeling, as well as the results of the consistency analysis for each sequence's genomic element in terms of length and nucleotide conservation degree. Section 10.6 discusses the analysis of the results obtained and their relation to microbiological evidence. Finally, Section 10.7 summarizes the main conclusions drawn from the exhaustive analysis of the SARS-CoV-2 genomic elements conducted in this Chapter.

10.2 Genomic Protein Analysis in SARS-CoV-2

The first step was to identify, analyze, and ultimately understand each and every one of the genomic elements of this virus. The complete genome of SARS-CoV-2 has a length that can range from 26,000 to 32,000 nucleotides (nts) [72]. It contains 11 Open Reading Frames (ORFs) which codify 27 proteins [2583]. In broad terms, the genome of SARS-CoV-2 contains:

1. An untranslated region at the 5' end (5'UTR).

2. An Open Reading Frame ORF1ab encoding 16 non-structural proteins (nsps) involved in viral replication.

3. Four genes encoding structural proteins: the Spike protein (S), the Envelope protein (E), the Membrane protein (M), and the Nucleocapsid protein (N).

4. Six accessory genes encoding accessory proteins: ORF3a, ORF6, ORF7a, ORF7b, ORF8, and ORF10.

5. An untranslated region at the 3' end (3'UTR).

To enhance the visual understanding of the structure of key proteins within the SARS-CoV-2 virion (the complete and infectious viral particle in its extracellular state), Figure 10.1 offers a schematic representation of the SARS-CoV-2 virus.

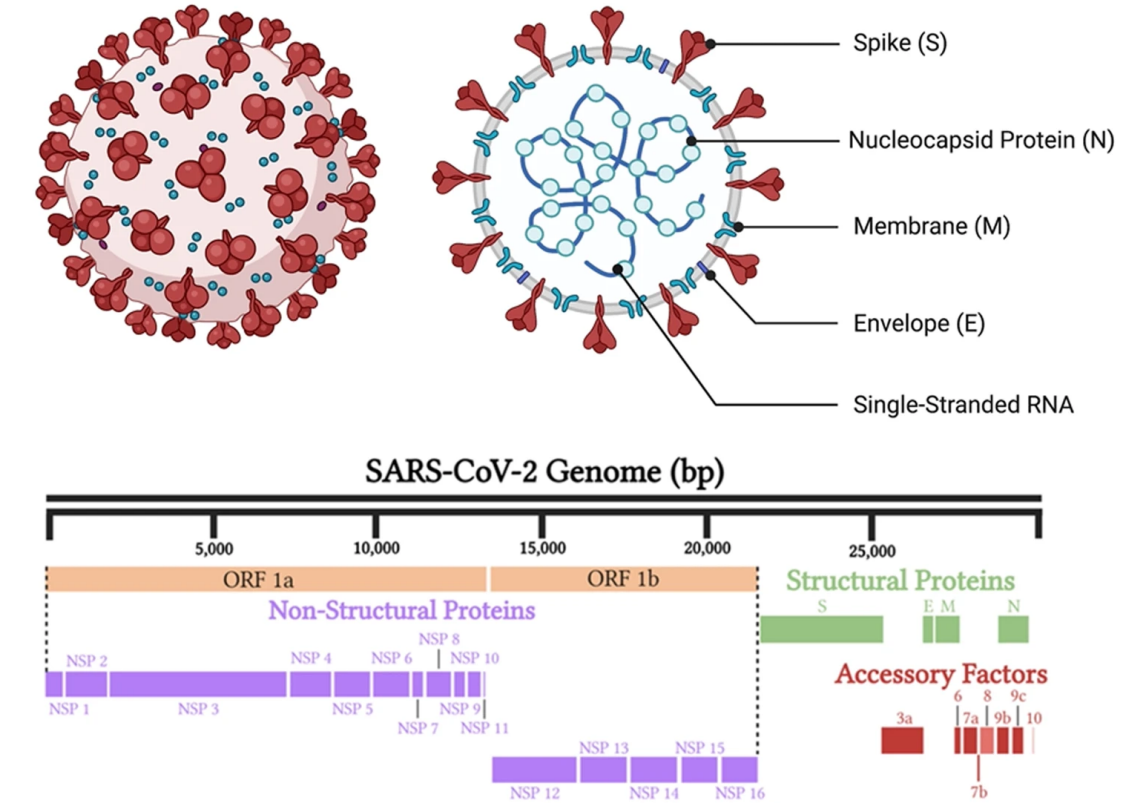


Figure 10.1: Schematic Representation of SARS-CoV-2 Virion. Image extracted from the research article by Jamison et al. [2584].

Various intergenic regions are located between the genes of SARS-CoV-2. These intergenic regions are segments of the viral genome found between individual genes

and do not encode proteins. The function of these regions is mostly unknown but they may potentially play a significant role in the virus lifecycle and its interaction with the host cell. Although they do not encode proteins, intergenic regions may contain crucial regulatory sequences for the replication, transcription, and translation of the virus, as well as for regulating viral gene expression and the host immune response [2585].

Some of the features and potential functions of intergenic regions in SARS-CoV-2 may include:

- Regulatory sequences. Intergenic regions may contain sequences of regulatory elements that control the activity of promoters and other elements necessary for gene expression [2586].

- RNA secondary structures. These regions can form RNA secondary structures that play critical roles in regulating viral replication, translation, and RNA stability [2587].

- Protein binding sites. Intergenic regions can serve as binding sites for viral and cellular proteins, facilitating the interaction between the virus and the host cell machinery [2588].

- Generation of subgenomic RNA. Some intergenic regions can serve as sites for initiation and termination of transcription, resulting in the generation of subgenomic RNA used for the synthesis of structural and accessory proteins of the virus [2589].

Table 10.1 compiles the genomic elements of SARS-CoV-2 Wuhan-Hu-1 Reference Genome Sequence NC_045512.2 (ORFs and intergenic regions) and their approximate consensus location [2590] :

Table 10.1: Breakdown of Genomic Elements in the SARS-CoV2 Wuhan-Hu-1 Reference Genome Sequence (NC_045512.2). The column "Nucleotide Position" refers to the nucleotide positions of the start and end of each genomic element in the SARS-CoV2 Wuhan-Hu-1 reference genome sequence (NC_045512.2) [2590].

Gene	Nucleotide Position
5'UTR	1-265
ORF1ab polyprotein	**266-21,555**
nsp1 (leader protein)	266-805

Table 10.1 – Continuation from the previous page

Gene	Nucleotide Position
nsp2	806-2719
nsp3 (papain-like protease)	2,720-8,554
nsp4	8,555-10,054
nsp5 (3C-like proteinase)	10,055-10,972
nsp6	10,973-11,842
nsp7	11,843-12,091
nsp8	12,092-12,685
nsp9	12,686-13,024
nsp10	13,025-13,441
nsp11	13,442-13,480
nsp12 (RNA-dependent RNA polymerase)	13,442-16,236
nsp13 (helicase)	16,237-18,039
nsp14 (3'-to-5' exonuclease)	18,040-19,620
nsp15 (endoRNAse)	19,621-20,658
nsp16 (2'-O-ribose methyltransferase)	20,659-21,552
ORF1ab/S intergenic region	21,553-21,562
S (surface glycoprotein)	**21,563-25,384**
S/ORF3a intergenic region	25,385-25,392
ORF3a	**25,393-26,220**
ORF3a/E intergenic region	26,221-26,244
E (envelope protein)	**26,245-26,472**
E/M intergenic region	26,473-26,522
M (membrane glycoprotein)	**26,523-27,191**
M/ORF6 intergenic region	27,192-27,201
ORF6	**27,202-27,387**
ORF6/ORF7a intergenic region	27,388-27,393
ORF7a	**27,394-27,759**
ORF7b	**27,756-27,887**
ORF7a/ORF8 intergenic region	27,888-27893
ORF8	**27,894-28,259**
ORF8/N intergenic region	28,260-28,273
N (nucleocapsid phosphoprotein)	**28,274-29,533**
N/ORF10 intergenic region	29,534-29,557
ORF10	**29,558-29,674**
3'UTR	**29,675-29,903**

The exact functions and operation of the 5'UTR and 3'UTR are still partly unknown.

However, 5'UTR is associated with the initiation of translation of viral proteins and regulates their synthesis by the ribosome [2591]. 5'UTR plays an important

role in controlling mRNA stability (the more stable, the greater the capacity to produce more protein copies from that mRNA). It can encode loops in the RNA, thus protecting it from the host's RNAases, ensuring the generation of more protein copies per mRNA molecule. Additionally, its variants can affect its binding efficiency to the ribosome, thus impacting protein synthesis [2592, 2593].

3'UTR regulates protein translation, promotes RNA stability, includes signals recognized by the cellular machinery for transcription termination, and adds a polyadenine tail (polyA) to the 3' end of the viral RNA [2594, 2595].

Furthermore, both the 5'UTR and 3'UTR contain elements to facilitate the dominance of the host cell environment [2595, 2596].

ORF1a produces a portion of the polyprotein 1ab (pp1ab), which includes the non-structural proteins nsp1-nsp11. ORF1b produces the remainder of polyprotein 1ab, which includes the non-structural proteins nsp12-nsp16, as well as part of nsp11, through a ribosomal -1 frame shift overlapping the STOP codon of ORF1a [2597]. This complex organization allows for efficient synthesis of multiple non-structural proteins from the same open reading frame.

The functions identified in the proteins encoded in ORF1ab are:

- Nsp1 (leader protein). The main function of nsp1 in SARS-CoV-2 is to suppress the host immune response and manipulate the cellular machinery to favor viral replication [2598].

- Nsp2. It is involved in viral replication processes [2599] and modulation of the immune response to favor such replication [2600].

- Nsp3 (papain-like protease). It promotes efficient viral RNA replication and is crucial for processing viral polyproteins, generating functional products necessary for viral replication. It is also involved in polyprotein processing, releasing individual proteins. It is also related to certain modulations in the host immune response, as well as to evasion from its effects [2601, 2602].

- Nsp4. It contributes to the formation of viral RNA replication complexes within the host cell, associates with other viral proteins to induce changes in cellular membranes, forming specific membranous structures that act as platforms

for viral RNA replication and transcription, facilitates viral RNA transcription, and contributes to efficient viral genome replication [2603, 2604]

- Nsp5 (3C-like proteinase). Its main function is the proteolysis of viral polyproteins, making specific cuts in the long chain of polyproteins produced during viral RNA translation. This generates individual proteins [2605]. Additionally, it may be related to mechanisms of immune response evasion [2606].

- Nsp6. It may be involved in inducing changes in cellular membranes, thus contributing to the formation of specific membranous structures. These structures, such as the remodeled endoplasmic reticulum, are essential for viral replication [2607], interfering with the cellular process of autophagy [2608]

- Nsp7. Its main function is co-stimulation of RNA polymerase activity. Nsp7 forms a complex with the protein nsp8, known as the RNA-dependent RNA replication-transcription complex. This nsp7/nsp8 complex acts as a co-activator that significantly enhances the activity of nsp12 (RNA-dependent RNA polymerase), where nsp7 contributes to the stability and function of the complex, facilitating efficient replication and transcription of viral RNA. The nsp7/nsp8 complex has also been implicated in viral RNA elongation during virus genome replication. Its presence is essential for the synthesis of long RNA chains, a critical process for successful viral replication [2609, 2610]. It also appears that nsp7 inhibits interferon production [2611].

- Nsp8. This protein associates with nsp7 to form a polymerase complex essential for viral replication. Both proteins act as a stimulating co-factor for the nsp12 polymerase, enhancing its RNA synthesis activity and thus facilitating viral replication [2612]. Additionally, nsp8 is implicated in the initiation process of viral RNA synthesis [2610], and may be involved in generating mutations during replication [2613]. It may also have immune response suppression functions [2614].

- Nsp9. It plays several key roles during viral replication. It participates in viral RNA synthesis, interacts with other proteins in the replicase complex, stabilizes this complex, is able to bind to nucleic acids, and has been suggested to influence the host immune response [2615]. It may also promote cytokine production [2616].

- Nsp10. This protein is the co-activator of nsp14 enzymatic activity, it stimulates the nsp13 helicase, and interacts with nsp14 and nsp16, thus nsp10 appears to play a role in regulating multiple enzymes essential for viral replication and RNA modification [2617, 2618].

- Nsp11. This protein may play a functional role in the affinity/interaction with the host cytosolic membrane [2619]. However, it is a highly disordered and poorly understood protein [2620].

- Nsp12 (RNA-dependent RNA polymerase, RdRp). It is the main subunit of the viral RNA polymerase, also known as RdRp. This enzyme is essential for viral genome replication and RNA synthesis during infection. Its main function is to catalyze the synthesis of complementary RNA strands using viral RNA as a template. Nsp12 associates with other viral proteins, such as nsp7 and nsp8, to form the replicase/transcriptase complex. This complex is involved in viral RNA transcription and replication [2612]. Given its importance in virus replication, it is a target in antiviral development. Nsp12 and its RdRp function are conserved in other coronaviruses [2621].

- Nsp13 (helicase). Its helicase activity can unwind or untwist DNA and RNA strands during viral replication. Nsp13 associates with other viral proteins, such as nsp7, nsp8, and nsp12, to form the replicase/transcriptase complex. This complex is involved in viral RNA replication and transcription. It has been suggested that nsp13 may stimulate nsp12 activity. In addition to its helicase activity, nsp13 has nucleotide repairing activity, meaning it can remove mispaired nucleotides incorporated during replication, thus contributing to the fidelity of the process [2622].

- Nsp14 (3'-to-5' exonuclease). It has 3'-5' exonuclease activity, meaning it can remove nucleotides from an RNA chain starting from the 3' end and moving towards the 5' end. This activity is involved in error correction during viral RNA synthesis, thus contributing to the fidelity of the viral genome. It also has a "Cap 0" methyltransferase function. That involves adding a methyl group to the 5' end of newly synthesized viral RNAs [2623]. Cap methylation is important for the stability and efficiency of viral mRNA translation. It also collaborates with the nsp10 protein,

which acts as a co-activator [2624]. Due to its crucial role in error correction during viral RNA replication, nsp14 is an important target for some antivirals [2625].

- Nsp15 (endoRNAse). Its endoribonuclease activity cleaves phosphodiester bonds in RNA. This activity is specific to certain RNA substrates and may be involved in RNA modification and processing during viral replication. It is also a component of the replicase/transcriptase complex, which is involved in viral RNA replication and transcription [2626]. It has been suggested that the endoribonuclease activity of nsp15 is involved in evasion of the host immune system. Nsp15 could influence cellular responses to stress and promote a favorable environment for viral replication [2627].

- Nsp16 (2'-O-ribose methyltransferase). The main function of nsp16 is to carry out cap methylation at the 5' end of viral messenger RNA (mRNA). This modification involves adding a methyl group to the guanosine at the mRNA's initial position. Cap methylation is essential for mRNA stability and efficiency of translation. Methylation of the mRNA cap is a strategy used by viruses to evade the host immune response. Cap modification allows the viral mRNA to mimic cellular mRNA more closely and escape detection by the immune system. It also interacts with the protein nsp10, which acts as a coactivator. Interaction with nsp10 is necessary for stimulating nsp16's methyltransferase activity. Nsp16 is a component of the replicase/transcriptase complex, which is involved in viral RNA replication and transcription. This complex is essential for the virus's lifecycle. Additionally, nsp16 may have nucleotide proofreading activity. This activity involves removing misincorporated nucleotides during replication, contributing to the fidelity of the viral genome [2628–2630].

As for the rest of the ORFs, in order, their functions are as follows:

- Surface glycoprotein (S). Structural protein located in the virus envelope that is responsible for facilitating virus entry into host cells. It consists of two main functional subunits: the S1 subunit contains the receptor-binding domain (RBD), and the S2 subunit mediates viral membrane fusion with the cell membrane. The receptor-binding domain (RBD) of the S protein interacts specifically with

the angiotensin-converting enzyme 2 (ACE2) receptor on the surface of host cells. This interaction is the initial and critical step for viral infection as it facilitates virus binding to the cell and subsequent entry. Once the RBD binds to ACE2, the S protein undergoes a series of conformational changes that lead to fusion of the viral membrane with the cell membrane. This process allows the release of the virus genetic material into the host cell, marking the beginning of viral replication and infection spread [2631].

Variability in the S protein of SARS-CoV-2 plays a crucial role in the virus evolution and the emergence of new variants. The interaction between the S protein and ACE2 on the surface of human cells is a critical step in the viral infection process, and any changes in the structure or composition of the Spike protein can influence its ability to bind to the receptor and enter the host cell.

Throughout the COVID-19 pandemic, numerous variants of SARS-CoV-2 have been identified that exhibit changes in the Spike protein. These changes can include point mutations, insertions, deletions, or genetic recombinations in the region encoding the Spike protein. The emergence of these variants is closely related to the evolutionary pressure to which the virus is subjected. For example, natural selection may favor the spread of variants that exhibit mutations in the Spike protein that increase their affinity for the ACE2 receptor or allow them to escape neutralization by antibodies generated from previous infections or vaccination. Additionally, variability in the Spike protein may be influenced by environmental factors such as exposure to different host populations, selective pressure from medical treatments, or interaction with other pathogens [2632, 2633].

- **ORF3a.** It codifies a viral protein from the coronavirus family that is expressed during SARS-CoV-2 replication. It is a multifaceted protein that has been associated with the following activities.

a) Induction of the immune response. ORF3a has been implicated in modulating the host's immune response. ORF3a expression has been observed to induce the release of proinflammatory cytokines and the activation of immune signaling pathways, contributing to the inflammatory response observed during

SARS-CoV-2 infection. This can trigger an exaggerated immune system response, which can be harmful to the host [2634].

b) Induction of cellular apoptosis. ORF3a has been shown to induce apoptosis (programmed cell death) in infected cells. This could contribute to the virus pathogenesis by promoting cell death and viral dissemination [2635].

c) Involvement in viral envelope formation. ORF3a has been associated with the formation of the viral envelope of SARS-CoV-2. It has been suggested to interact with other viral proteins and cellular components to facilitate the assembly of new virions and the release of viral particles [2636].

d) Modulation of the Endoplasmic Reticulum (ER) Stress Response. ORF3a can affect the Endoplasmic Reticulum (ER) homeostasis of the host cell, leading to the activation of the ER stress response. This may be important for viral replication and pathogenesis, as the ER is crucial for protein synthesis and proper folding of viral proteins [2637].

- Envelope protein (E). Structural protein, involved in several functions, as detailed below.

a) Viral envelope formation. The E protein is essential for the viral envelope formation of SARS-CoV-2. It is embedded in the virus membrane and plays a crucial role in viral morphogenesis, helping to shape and stabilize the virion structure [2638].

b) Involvement in viral release. It has been suggested that the E protein may be involved in the release of viral particles from infected cells. This can occur through its interaction with other viral proteins and cellular components that facilitate the release of complete virions [2639].

c) Modulation of host immune response. The E protein can influence the host immune response during SARS-CoV-2 infection. It has been observed to induce proinflammatory responses and may contribute to the inflammation observed in tissues affected by viral infection [2640].

d) Interaction with host proteins. The E protein can interact with various host proteins, potentially affecting multiple cellular processes. These

interactions may be important for viral replication, pathogenesis, and host immune response [2641].

- **Membrane glycoprotein (M).** This structural protein plays an essential role in the formation and morphology of the virion, as well as in its interaction with the cellular environment during infection.

  a) Morphogenesis and virion assembly. The M protein is crucial for the formation of the virion envelope and its proper morphology. It binds to the viral membrane and is responsible for shaping and stabilizing the virus structure. Additionally, the M protein interacts with other viral proteins and cellular components to facilitate the assembly of complete virions [2642].

  b) Virion shape determination. The M protein plays a significant role in determining the shape and size of the virion. Its interactions with other viral proteins and the cell membrane contribute to shaping the final viral particle [2643].

  c) Involvement in viral release. The M protein may be involved in the release of viral particles from infected cells. It has been suggested to interact with other viral proteins and cellular components to facilitate the release of complete virions from the host cell [2644].

  d) Interaction with the host immune system. The M protein can influence the host immune response during SARS-CoV-2 infection. It has been observed to induce specific immune responses and may be recognized by the host immune system as an antigen, triggering antibody production [2645].

- **ORF6.** This ORF encodes a viral protein with multiple functions during infection. Although its specific role is not yet fully defined, several biological activities associated with ORF6 have been identified.

  a) Inhibition of host immune response. ORF6 has been implicated in the inhibition of the host immune response. It has been observed that ORF6 interacts with key proteins of the interferon signaling pathway, such as the transcription factor STAT1 and the nuclear export complex CRM1, resulting in the inhibition of interferon-induced gene transcription and interference with the host's innate immune response [2646].

b) Retention of proteins in the endoplasmic reticulum (ER). ORF6 has been shown to localize to the endoplasmic reticulum (ER) of the host cell and retain proteins in this cellular structure, affecting multiple cellular processes, such as the immune response and protein synthesis [2647].

c) Modulation of transcriptional activity. ORF6 has been associated with the modulation of transcriptional activity in infected cells. It has been suggested that ORF6 may interact with transcriptional regulatory proteins and alter the expression of specific host genes, which could contribute to viral pathogenesis and evasion of the immune response [2648].

d) Regulation of viral gene expression. ORF6 may play a role in regulating viral gene expression by affecting the transcription and stability of viral RNAs. It has been proposed that ORF6 can interact with viral proteins and cellular components involved in the replication and transcription of the viral genome, influencing the production of viral proteins and virus replication [2649].

In summary, ORF6 is a multifunctional protein that plays a crucial role in evading the host immune system, manipulating cellular machinery, and regulating viral gene expression during SARS-CoV-2 infection. Its continued study is essential for fully understanding viral pathogenesis mechanisms and for developing therapeutic strategies targeted against COVID-19.

- ORF7a. The viral protein encoded by ORF7a performs several biological functions.

a) Modulation of Host Immune Response. ORF7a has been suggested to modulate the host immune response during SARS-CoV-2 infection. It has been observed that ORF7a interacts with proteins of the innate immune system, such as the 1 protein of the eIF3f initiation factor, which may influence the host immune response and viral pathogenesis [2650].

b) Induction of Cellular Apoptosis. ORF7a has been implicated in the induction of apoptosis (programmed cell death) in infected cells. It has been observed that the expression of ORF7a leads to characteristic cellular morphological changes associated with apoptosis, which may facilitate viral dissemination and SARS-CoV-2 pathogenesis [2651].

c) Interaction with Cellular and Viral Proteins. ORF7a can interact with various cellular and viral proteins, suggesting a multifaceted role during infection. It has been demonstrated that ORF7a interacts with the viral envelope protein E, as well as with proteins involved in intracellular transport machinery, which may influence viral replication and the release of viral particles [2652].

d) Subcellular Localization. ORF7a is predominantly localized to the endoplasmic reticulum (ER) membrane of infected cells. This subcellular localization suggests that ORF7a may be involved in manipulating the structure and function of the ER during SARS-CoV-2 infection, which can affect multiple cellular processes, including immune response and viral replication [2653].

- **ORF7b.** The viral protein identified with ORF7b is linked to the following biological activities:

a) Interaction with cellular and viral proteins. It has been demonstrated that ORF7b interacts with various cellular and viral proteins during SARS-CoV-2 infection. These interactions can impact multiple cellular and viral processes, including the immune response, viral replication, and release of viral particles [2654].

b) Regulation of the immune response. ORF7b has been implicated in modulating the host immune response during SARS-CoV-2 infection. It has been suggested that ORF7b may interfere with the signaling of certain immune pathways, resulting in the suppression or modulation of the inflammatory and antiviral response [2655].

c) Subcellular localization. ORF7b is primarily localized in the endoplasmic reticulum (ER) and Golgi apparatus of infected cells. This subcellular localization suggests that ORF7b may be involved in manipulating the structure and function of these cellular organelles during viral infection [2656].

d) Effects on viral pathogenesis. Although further research is needed, it has been proposed that ORF7b may play a role in viral pathogenesis by influencing virus replication, host immune response, and virus ability to evade the immune system [2657].

- **ORF8.** The viral protein encoded by ORF8 is relevant in terms of pathogenesis, thus it has been the subject of intensive research. Although its specific role is not fully defined, several biological activities associated with ORF8 have been identified.

a) Modulation of host immune response. ORF8 may modulate the host immune response during SARS-CoV-2 infection. Research has shown that ORF8 can affect the interferon signaling pathway and antigen presentation, which may compromise the immune system ability to combat viral infection [2658].

b) Interaction with cellular and viral proteins. ORF8 interacts with various cellular and viral proteins, suggesting a multifaceted role during infection that can influence viral replication, pathogenesis, and host immune response [2659].

c) Induction of cellular apoptosis. It has been suggested that ORF8 may induce apoptosis in infected cells. Apoptosis is a process of programmed cell death that can be beneficial for viral replication and virus dissemination [2660].

d) Genetic variants. ORF8 is known to have genetic variants, some of which have emerged during the COVID-19 pandemic. These variants can influence the virulence and ability of the virus to evade the host immune response [2661].

- **Nucleocapsid phosphoprotein (N).** The main function of the structural protein N is to bind and package viral RNA into the virion nucleus, forming the nucleocapsid, which is essential for viral genome replication and transcription. Some of its functions are listed below.

a) Binding to viral RNA. The N protein specifically binds to viral RNA through RNA-binding domains, facilitating its packaging into the nucleocapsid. This interaction is crucial for stabilizing and protecting the viral RNA from cellular nucleases and for maintaining the integrity of the viral genome during replication and transcription [2662].

b) Nucleocapsid formation. The N protein is the main structural component of the nucleocapsid, which is a helical structure formed by multiple units of the N protein and packaged viral RNA. This structure provides stability to the viral genome and facilitates its transport within the host cell during infection [2663].

c) Regulation of viral replication. It has been suggested that the N protein may play a role in regulating viral replication and genome transcription. Its interaction with other viral proteins and cellular components can influence the efficiency and speed of viral replication [2664].

d) Interaction with host proteins. The N protein can interact with cellular host proteins, which may affect the host immune response and virus pathogenesis. It has been observed that the N protein induces specific immune responses during infection, contributing to the elimination of the virus by the immune system [2665].

- **ORF10.** Its specific functions in SARS-CoV-2 are not yet fully understood, but several hypotheses have been proposed regarding its possible role in viral biology. Some of these hypothetical functions pending further research for confirmation are as follows.

a) Regulation of the immune response. ORF10 could be involved in regulating the host immune response during SARS-CoV-2 infection. ORF10 has been observed to interact with host proteins involved in the immune response, suggesting that it may have a role in immune system evasion or modulation of inflammation during viral infection [2666, 2667].

b) Interaction with cellular and viral proteins. ORF10 has shown interactions with other viral and cellular proteins, suggesting that it could be involved in regulating multiple cellular and viral processes. These interactions could influence viral replication, viral particle assembly, or the cellular response to the virus [2668].

c) Involvement in viral pathogenesis. Although further research is needed, it has been proposed that ORF10 could play a role in the pathogenesis of SARS-CoV-2 by affecting viral replication, the host immune response, or virus virulence. Changes in ORF10 expression have been observed in different infection contexts, suggesting that it could contribute to the virus adaptation to different environmental conditions or variability in the host immune response [2669].

Within the **3'UTR**, several specific structural elements of the genomic RNA of coronaviruses, including SARS-CoV-2, are located. The stem-loop and pseudoknot structures are mutually exclusive in their formation since their sequences overlap,

and it is postulated that their alternative formation plays some role in transcription regulation [2670]. A succinct description of both structures is following.

- **Coronavirus 3'UTR pseudoknot stem-loop 1 and 2.** These elements are pseudoknot structures present in a stem-loop within the viral RNA untranslated region 3' (3'UTR), and play an important role in the regulation of viral translation and replication [2671]. The stem-loop 1 and 2 elements are specific structures present in the genomic RNA of SARS-CoV-2 that are involved in the frameshift process during viral RNA translation. These elements play a crucial role in protein expression by inducing a -1 frameshift during viral RNA translation. This -1 frameshift is essential for the synthesis of the viral polyprotein, which is processed to produce the different proteins necessary for virus replication and structure [2672]. These RNA secondary structures are recognized by the cellular translation machinery and play a role in the efficiency and accuracy of frameshift during viral RNA translation [2673].
- **Coronavirus 3' stem-loop II-like motif (s2m).** This is a specific secondary structure present in the 3' untranslated region (3' UTR) of coronavirus RNA. This stem-loop structure has been identified as important for viral replication and viral RNA stability, and it has been suggested that may be involved in the interaction of viral RNA with cellular proteins and viral replication processes [2674, 2675].

These structural elements are critical to the virus biology and may be important targets for the development of antiviral therapies.

10.3 Equipment

All experiments were run in this equipment:

- Processing Unit: Intel(R) Core(TM) i7-4770K CPU. 3.5 GHz.
- Installed RAM: 32 GB usable.
- Operative System: Windows 10 Education. Version: 22H2.
- GPU: NVIDIA GeForce RTX 3090. Total memory: 40 GB.

10.4 Methods

All computer analyses were developed in a Python environment, using primarily the following libraries:

- Pandas, for data manipulation in the form of DataFrames. A DataFrame in Python is a labeled, mutable two-dimensional data structure that can contain data of different types [2676].

- Numpy, for implementing numerical operations [2677].

- Math, for applying certain mathematical functions [2678].

- Os, for operations related to the operating system [2679].

- Re, for regular expression operations [2680].

- Selenium, specifically the WebDriver function, for the automatic download of EMBL files [2681].

Figure 10.2 shows the main stages of the methodology presented in this Chapter.

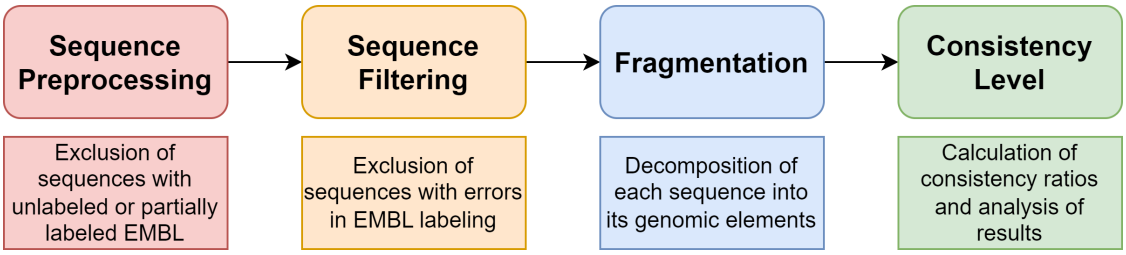


Figure 10.2: SARS-CoV-2 Genomic Element Analysis Block Diagram

10.4.1 Sequence Preprocessing

Starting with a dataset consisting of the FASTA files of 97,236 complete sequences of SARS-CoV-2, I implemented an automatic search and download system for the corresponding EMBL files for each of the identifiers of these sequences. Subsequently, I performed preprocessing of this dataset, determining the existence of labeling in the EMBL file for each genomic element.

The FASTA files provided the content of the genomic sequence, expressed in nucleotides. Although SARS-CoV-2 is an RNA virus, its genome is represented in FASTA format using the letters A, T, C, and G (not U). By convention,

most databases and bioinformatics tools transcribe RNA viral genomes using the traditional DNA nomenclature.

The first step was the calculation of certain parameters of interest, such as the total length of the complete sequence and the number of occurrences of each nucleotide (A, C, G, T) in the sequence, as well as the number of other characters that represent ambiguous nucleotides or codings for multiple nucleotides (W, S, M, K, R, Y, B, D, H, V, N, Z).

The next step was the automatic detection and download of the corresponding EMBL files for each of the sequences from the public repository of the European Nucleotide Archive (ENA) [39].

An EMBL file is a type of file used to store nucleic acid sequences and their associated information in the format of the European Molecular Biology Laboratory (EMBL) Sequence Database. This format is commonly used in bioinformatics to store and share genomic, transcriptomic, and proteomic sequence data.

EMBL files contain metadata about the sequence, such as accession numbers, descriptions, gene annotations, coding regions, DNA features (such as promoters, transcription factor binding sites, etc.), and any other relevant information. These files can be processed by a variety of bioinformatics programs for sequence analysis, annotations, and genomic comparisons [2682].

In these EMBL files, therefore, all genomic elements are labeled (or at least, they should be). This is the most detailed listing of labeled elements found in the EMBL files that correspond with those previously described in Section 10.2, and which I called "Protein Listing":

1) 5'UTR.

2) ORF1ab:

- Nsp1 (leader protein).
- Nsp2.
- Nsp3 (papain-like protease).
- Nsp4.
- Nsp5 (3C-like proteinase).

- Nsp6.
- Nsp7.
- Nsp8.
- Nsp9.
- Nsp10.
- Nsp11.
- Nsp12 (RNA-dependent RNA polymerase).
- Nsp13 (helicase).
- Nsp14 (3'-to-5' exonuclease).
- Nsp15 (endoRNAse).
- Nsp16 (2'-O-ribose methyltransferase).

3) Spike or S (Surface glycoprotein).

4) ORF3a.

5) E (Envelope protein).

6) M (Membrane glycoprotein).

7) ORF6.

8) ORF7a.

9) ORF7b.

10) ORF8

11) N (Nucleocapsid phosphoprotein).

12) ORF10

13) Coronavirus 3'UTR pseudoknot stem-loop 1.

14) Coronavirus 3'UTR pseudoknot stem-loop 2.

15) Coronavirus 3' stem-loop II-like motif (s2m).

Later, I added to this "Protein List" the nomenclatures of all intergenic regions detected between the ORFs. This total compilation of genomic elements was referred to as the "Genomic Elements Listing".

Once the EMBL files of all sequences were downloaded, I conducted a comprehensive analysis of each sequence based on their FASTA and EMBL files. Broadly speaking, the phases of the process were as follows:

1. Opening each file and reading its content.

2. Searching for the presence of each protein in "Genomic Elements Listing" within the content of the EMBL file.

3. Identifying the start and end positions of each protein within the EMBL file.

4. Verifying the presence of all proteins.

5. Listing files with the complete labeling of all proteins from the "Genomic Elements Listing".

In other words, the sequence preprocessing involved examining the EMBL files to find specific information about proteins and other genetic structures of interest, counting the occurrence of proteins, and determining which files contain all the proteins specified in a predefined list. Once examined, the next step was to conduct a comprehensive analysis of these EMBL files to identify the presence and location of specific proteins and other genetic structures of interest. The final phase involved saving this information in CSV files for further analysis.

10.4.2 Sequence Filtering

Due to errors in the labeling of the EMBL files or errors in the sequencing or transcription, sometimes the content of the same genomic element in different EMBL files does not match.

To avoid analysis with incorrect data, I conducted filtering in two phases.

First, I removed all those sequences in which the structural and non-structural proteins were not labeled in the EMBL, indicating start and end positions for each of them. That is, sequences in whose EMBL the following occurred:

- Absence of ORF labels.

Or:

- Partial labeling. That is, ORF1ab was labeled with its start and end positions, but not specified for each of the proteins included in this polyprotein.

In a second phase, I filtered out those sequences where, in addition to all their elements being fully labeled, the start and end positions were correct. To do

this, I worked by comparing the three starting nucleotides and the three ending nucleotides for each genomic element.

To work only with correct genomic elements, I needed to ensure that at least the first 3 initial and final nucleotides of each genomic element were the same. I opted for a rather pragmatic first step, which was to determine the frequency of occurrence of the first and last three letters of each protein sequence in the column and what types of nucleotides they were. This way, I obtained the most common patterns of start and end for each of the genomic elements in each original complete sequence. All sequences that had at least one genomic element not matching the most common pattern in the corresponding start and end were removed from the analysis.

I excluded from this criterion the genomic elements 5'UTR and ORF8/N intergenic region. The high variability of these two elements could distort the operability of the comparison. The ORF8/N intergenic region shared its first three nucleotides approximately 50%. The 5'UTR was an extremely variable zone, subjected to many possible modifications, including point nucleotide mutations (Single Nucleotide Polymorphisms, SNPs) and deletions of different fragments [2593]. The inherent distortion of both fragments discouraged their inclusion in the conservation analysis of genomic elements since it would imply the elimination of practically all sequences under analysis, resulting in the loss of relevant information from multiple sequences that I considered exploitable except in the case of these elements.

10.4.3 Fragmentation of Filtered Sequences

The next step involved splitting each complete genomic sequence of SARS-CoV-2 into each of its genomic fragments, including each of the proteins as well as the intergenic regions between ORFs and other non-structural elements.

Each of these fragments was stored independently in a CSV file with the following naming convention: "ID_Element.csv", where "ID" was the Identification Number of the complete genomic sequence and "Element" is the nomenclature of the genomic element (structural proteins, non-structural proteins, accessory elements, etc.).

10.4.4 Determination of Consistency Ratios

The next step involved establishing a set of metrics to determine the degree of consistency for each genomic element. My objective was to employ computationally efficient and straightforward ratios.

I determined the consistency of a genomic element based on two main factors. On one hand, the degree of conservation of genome length, which is related to the abundance of insertions, deletions, genetic recombination, or other mechanisms of genetic variability that can alter the sequence length of the element. On the other hand, the degree of nucleotide conservation, which can measure the extent of point mutations in that genomic element.

To determine consistency in both factors, I selected the following metrics:

- Average Length for each genomic element.

- Standard Deviation of the length of each genomic element. Low values of this ratio indicate lower variability.

And two more specific metrics tailored to the case study, based on the results of a previous simultaneous comparison, named the Length Consistency Index and Nucleotide Consistency Index.

For each genomic element, I simultaneously compared all genomic elements across all sequences, resulting in a comparative sequence called the Generated Final Sequence.

This sequence was composed of all common characters (nucleotides) at each position of the genomic element. If at least one nucleotide at that position differed across any sequence, a dash (-) was included at that position. Additionally, if a character was absent in any of the sequences, it was filled with zeros (0) to match the length of the longest sequence. Therefore, the number of zeros at the end of the Generated Final Sequence corresponds to the difference in nucleotide count between the longest and shortest sequences.

To facilitate the understanding of the analysis process and the generation of the Generated Final Sequence, I will now demonstrate this process using a series of fictional fragments as an example. Let us suppose the genomic fragments

named Sec1 Sec5 correspond to the same fictional genomic element in 5 different sequences, which have varying lengths and nucleotide variability. If we arrange them element by element (Table 10.2).

Table 10.2: Implementation Process of the Generated Final Sequence - Step 1

Name	**Nucleotide Positions**											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Seq1	A	C	T	A	A	G	C	T	C	A	T	T
Seq2	A	G	T	C	C	C	G	T				
Seq3	A	C	T	A	A	C	G	T	C	A	A	T
Seq4	A	T	T	C	C	G	C	T	G	C	T	
Seq5	A	C	T	A	A	G	C	T	G	T		

Starting from the length of the longest fragment, in this case, Seq1 and Seq3, I fill each of the remaining fragments with "0" at the end until they reach the longest length. In the Generated Final Sequence, the "0"s at the end will correspond to the difference between the shortest and longest sequences. In other words, the appearance of a "0" in any of the fragments means that the Generated Final Sequence will have a "0" in that position. Logically, the first "0" appears at the end of the shortest fragment until it reaches the length of the longest fragment. I demonstrate this padding with "0"s to equalize lengths in Table 10.3.

Table 10.3: Implementation Process of the Generated Final Sequence - Step 2

Name	**Nucleotide Positions**											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Seq1	A	C	T	A	A	G	C	T	C	A	T	T
Seq2	A	G	T	C	C	C	G	T	0	0	0	0
Seq3	A	C	T	A	A	C	G	T	C	A	A	T
Seq4	A	T	T	C	C	G	C	T	G	C	T	0
Seq5	A	C	T	A	A	G	C	T	G	T	0	0
Generated Final Sequence									**0**	**0**	**0**	**0**

At this point, all the fragments, as well as the Generated Final Sequence, have the same length (that of the longest fragment). Therefore, the number of "0"s provides a measure of the degree of variability between the length of the shortest and longest fragments.

The next step in completing the Generated Final Sequence involves comparing nucleotide matches in the remaining fragment (the part not filled with "0"), that is, from the first nucleotide to the end of the shortest fragment. Column by column, if all nucleotides are the same, the Generated Final Sequence will be filled with the value of that nucleotide in that position. If they do not match in all the analyzed fragments, a "-" is placed in that position. The number of "-" gives us a measure of the degree of nucleotide concordance in that genomic fragment. For example, in the illustrative case, at positions 1, 3, and 8, the nucleotide is the same in all the fragments, so the value in those positions in the Generated Final Sequence will be the value of those nucleotides: A at position 1, T at position 3, and T at position 8. In the remaining positions, there is at least one fragment where the nucleotide differs, so the corresponding value in the Generated Final Sequence at that position is "-", as shown in Table 10.4.

Table 10.4: Implementation Process of the Generated Final Sequence - Step 3

Name	**Nucleotide Positions**											
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Seq1	A	C	T	A	A	G	C	T	C	A	T	T
Seq2	A	G	T	C	C	C	G	T	0	0	0	0
Seq3	A	C	T	A	A	C	G	T	C	A	A	T
Seq4	A	T	T	C	C	G	C	T	G	C	T	0
Seq5	A	C	T	A	A	G	C	T	G	T	0	0
Generated Final Sequence	**A**	**-**	**T**	**-**	**-**	**-**	**-**	**T**	**0**	**0**	**0**	**0**

Based on this Generated Final Sequence for each genomic element, I calculated the specified ratios:

- Length Consistency Index. Expressed in %, high levels of this ratio indicate greater consistency in terms of the length of the genomic sequence.

The consistency index in terms of length (Length Consistency Index) was calculated according to the following formula:

$$\text{Length Consistency Index (\%)} = \left(\frac{\text{Total Length} - \text{No. of Zeros}}{\text{Total Length}} \right) \times 100 \quad (10.1)$$

Being:

Total Length. It is the length of the Generated Final Sequence. It returns the length of the final sequence generated for the protein in question. This sequence contains common characters among all protein sequences for that particular protein, as well as additional dashes that have been added so that all sequences have the same length as the longest sequence.

No. of Zeros. It counts the number of zeros (0) in the final sequence. These zeros were added to adjust the length of all sequences to the length of the longest one. Therefore, the difference between the total length and the number of zeros gives us the amount of actual characters in the final sequence.

Finally, I normalized by dividing the difference obtained in the previous step by the total length of the final sequence. This normalization gives us an index representing the proportion of actual characters in the final sequence in relation to its total length. The higher this index, the fewer zeros there are in the final sequence, indicating less variability in the length of the original sequences.

- Nucleotide Consistency Index. Similar to the previous ratio, high values denote greater consistency, this time in terms of nucleotide matching. It was also expressed in %.

The consistency index in terms of nucleotide coincidence (Nucleotide Consistency Index) was calculated according to the following formula:

$$\text{Nucleotide Consistency Index (\%)} = \left(\frac{\text{Total Length} - \text{No. of Zeros} - \text{No. of Dashes}}{\text{Total Length}}\right) \times 100 \tag{10.2}$$

Being:

1. Total Length. The same description as in the previous ratio.

2. No. of Zeros. The same description as in the previous ratio.

3. No. of Dashes. It counts the number of dashes (-) in the final sequence. These dashes represent the places where characters differ among the original sequences. Subtracting this count from the total length gives us the amount of common characters in the final sequence.

I normalized by dividing the difference obtained in the previous step by the total length of the final sequence. This gave us an index representing the proportion of actual characters in the final sequence relative to its total length. The higher this index, the fewer dashes there were in the final sequence, indicating less variability in the characters of the original sequences.

The Length Consistency Index and the Nucleotide Consistency Index measured slightly different aspects of protein sequences. While the Length Consistency Index focused on the uniformity of sequence lengths, the Nucleotide Consistency Index focused on the similarity of nucleotides present in those sequences. Both provided valuable information about variability and conservation in protein sequences, but from different perspectives.

The determination of these consistency ratios was based on similar applications of variability model metrics, particularly regarding commonality and variability degrees [2683], Language Sample Analysis (LSA) [2684, 2685], and statistical approaches to text categorization [2686]. In the case of genomic sequences, this method is more straightforward and intuitive than other character string comparison algorithms such as the Mongue-Elkan Method [2687], soft cardinality [2688], or hash-q exact string matching algorithm [2689].

10.4.5 Calculation of Consistency Ratios

Once all the genomic sequences under study were processed and segmented, and the consistency measurement metrics were determined, the final step involved calculating the consistency ratios and storing the results.

The process carried out for the calculation of these parameters included the following steps:

1. Data Loading. A CSV file containing filtered protein sequences was loaded into a Pandas DataFrame. I calculated the length of the DataFrame. Each column of this DataFrame represented one of the predefined genomic elements in the Genomic Elements Listing.

2. Analysis of Matching Nucleotides. I performed this analysis as follows:

- The program iterated over each column of the DataFrame, i.e., over each genomic element.

- For each column, a list was created with the content of each element.

- Starting from the shortest sequence, the character (nucleotide type) at each index was compared with the same index in all sequences (of greater length).

- A new sequence was constructed containing the common nucleotide among all sequences at that index, or a dash "-" if there were differences.

- This final sequence, adjusted to the length of the longest sequence, was saved in a list.

4. Calculation of statistics. This consisted of the following steps:

- A DataFrame was created to store the calculated statistics.

- For each element, the average length of the sequence, the standard deviation of the length, the consistency index in terms of length, and the consistency index in terms of nucleotide content were calculated.

5. Result saving. The summary DataFrame of statistics was stored in a CSV file.

- These statistics were added to a summary DataFrame of statistics per element.

10.5 Results

10.5.1 Filtering Results

The raw complete sequences were 97,236, out of which 97,229 were correctly processed beforehand, and 7 encountered an error during the download, primarily related to file format errors. That is, if the EMBL file had an unexpected format or was corrupted, the automated download tool was unable to process it properly, resulting in an error.

I removed all sequences whose EMBL is not labeled or partially labeled, for example, if there is no breakdown in nsps in ORF1ab. After this elimination, the resulting operational complete sequences were reduced to 39,138.

Finally, I eliminated sequences whose labels contain errors. For example, errors in the start or end position of any of their genomic elements. An error in at least

one genomic element results in the removal of the corresponding list. After this filtering, I was left with 38,047 sequences.

For safety, the detection of incorrect labels in the EMBL (inconsistent start and end positions, typographical errors, etc.) in any of its genomic elements resulted in the removal of the complete sequence in the final analysis even if the first and last triplets were consistent with the most common ones identified, to avoid the introduction of undetected errors despite such coincidence.

The presence of a labeling defect in a specific genomic element often meant in virtually all cases that this defect led to subsequent errors in other genomic elements. If the error is in the early proteins, in most cases, this error was carried over to the rest, especially in cases of inconsistent start and end positions. Therefore, the repeatability of incorrectly labeled sequences is noticeable. The presence in one element implied the repeated presence in the rest, especially in the elements following the first error.

The most common start and stop codons for each genomic element detected, as well as their occurrence percentage, are summarized in Table 10.5.

Table 10.5: Most Common First and Last Triplets of Genomic Elements of SARS-CoV-2 Sequences and their Number of Occurrences. Percentage was calculated with reference to the total of 39,138 sequences fully labeled in their EMBL.

Genomic Element	**First Triplet**	**Number of Occurrences**	**Last Triplet**	**Number of Occurrences**
nsp1 (leader protein)	ATG	39,118 (99.95%)	GGG	39,112 (99.93%)
nsp2	GCA	39,072 (99.83%)	GGT	39,089 (99.88%)
nsp3 (papain-like protease)	GCA	39,052 (99.78%)	GGT	39,054 (99.79%)
nsp4	AAA	39,045 (99.76%)	CAG	39,043 (99.76%)
nsp5 (3C-like proteinase)	AGT	39,030 (99.72%)	CAA	39,032 (99.73%)
nsp6	AGT	39,018 (99.69%)	CAG	39,008 (99.67%)
nsp7	TCT	39,000 (99.65%)	CAA	39,002 (99.65%)

Table 10.5 – Continuation from the previous page

Genomic Element	Fist Triplet	Number of Occurrences	Last Triplet	Number of Occurrences
nsp8	GCT	38,995 (99.64%)	CAG	38,985 (99.61%)
nsp9	AAT	38,979 (99.59%)	CAA	38,979 (99.59%)
nsp10	GCT	38,973 (99.59%)	CAG	38,968 (99.57%)
nsp11	TCA	38,963 (99.55%)	GTG	38,963 (99.55%)
nsp12 (RNA-dependent RNA polymerase)	TCA	38,964 (99.56%)	AAC	38,965 (99.56%)
nsp13 (helicase)	GCT	38,863 (99.30%)	CAA	38,873 (99.32%)
nsp14 (3'-to-5' exonuclease)	GCT	38,833 (99.22%)	CAG	38,837 (99.23%)
nsp15 (endoRNAse)	AGT	38,811 (99.17%)	CAA	38,807 (99.15%)
nsp16 (2'-O-ribose methyltransferase)	TCT	38,800 (99.14%)	AAC	38,747 (99.00%)
ORF1ab/S intergenic region	CTA	38,537 (98.46%)	CAA	38,528 (98.44%)
S	ATG	38,746 (99.00%)	TAA	38,744 (98.99%)
S/ORF3a intergenic region	AAC	38,523 (98.43%)	TTA	38,526 (98.44%)
ORF3a	ATG	38,742 (98.99%)	TAA	38,743 (99.00%)
ORF3a/E intergenic region	AGC	38,519 (98.42%)	TTA	38,520 (98.42%)
E	ATG	38,742 (98.99%)	TAA	38,742 (98.99%)
E/M intergenic region	AAC	38,514 (98.41%)	CCA	38,501 (98.37%)
M	ATG	38,739 (98.98%)	TAA	38,742 (98.99%)
M/ORF6 intergenic region	AGT	38,490 (98.34%)	AGA	38,491 (98.35%)
ORF6	ATG	38,738 (98.98%)	TAA	38,732 (98.96%)

Table 10.5 – Continuation from the previous page

Genomic Element	Fist Triplet	Number of Occurrences	Last Triplet	Number of Occurrences
ORF6/ORF7a intergenic region	AAC	38,226 (97.67%)	ACA	38,421 (98.17%)
ORF7a	ATG	38,731 (98.96%)	TGA	38,698 (98.88%)
ORF7b	ATG	38,698 (98.88%)	TAA	38,684 (98.84%)
ORF7b/ORF8 intergenic region	AAC	38,119 (97.40%)	ACA	38,156 (97.49%)
ORF8	ATG	38,678 (98.83%)	TAA	38,669 (98.80%)
ORF8/N intergenic region	AAC	20,221 (51.67%)	AAA	38,041 (97.20%)
N	ATG	38,662 (98.78%)	TAA	38,662 (98.78%)
N/ORF10 intergenic region	AAC	38,059 (97.24%)	AGA	38,100 (97.35%)
ORF10	ATG	38,660 (98.78%)	TAG	38,660 (98.78%)
3'UTR pseudoknot stemLoop 1	TTG	38,659 (98.78%)	CAA	38,652 (98.76%)
3'UTR pseudoknot stemLoop 2	TAA	38,649 (98.75%)	TTA	38,650 (98.75%)
3' stem-loop II-like motif	TTC	38,567 (98.54%)	GAA	38,614 (98.66%)

In general terms, for the coding regions of the 39,138 sequences analyzed, the first and last triplets were coincident between 98.7% and 99.7%. However, in the intergenic regions, this coincidence slightly decreased, ranging between 97.0% and 98.5% (except for the first triplet of ORF8/N Intergenic Region 51.67%).

For 3'UTR pseudoknot stemLoop 1 and 3'UTR pseudoknot stemLoop 2, their homogeneity degrees in their first and last triplets were very high and stable, within a narrow range around 98.5%. These genomic elements are indeed highly conserved, so my results were biologically coherent [2690, 2691].

The first exception was identified in the first triplet in the ORF8/N intergenic region, which only reached a majority coincidence rate of 51.67% in the case of AAC as the first triplet. The variability in the first triplet of this intergenic zone is

very striking. This fact may be related to the high number of insertions produced in this genomic element [2692], which may lead to high variability in this area. In any case, this fact requires further research.

However, the degree of coincidence of the last triplet in the ORF8/N intergenic region was aligned with the corresponding values for the rest of the intergenic regions. This degree of coincidence was quantified at 97.2%.

In any case, except for the first triplet in the ORF8/N intergenic region, the degree of coincidence between the first and last triplets of each genomic element was extremely high.

As a measure of the consistency of the results obtained, I observed that the first triplet of nsp1 is ATG (AUG in mRNA). Given that nsp1 is the first non-structural protein listed in ORF1ab, the first triplet of nsp1 is the initial triplet of ORF1ab. ATG is a start codon. Start codons are sequences of three nucleotides in messenger RNA (mRNA) that signal the beginning of translation of genetic information into proteins during the process of protein synthesis in cells [2693]. ATG is the most usual codon start [2694, 2695].

A start codon, such as ATG, initiates the process of translation, facilitating the binding of the ribosome and the incorporation of the first amino acid into the polypeptide chain, in this case ORF1ab. This codon signals the ribosome the starting point to begin translating the genetic message from mRNA into the amino acid sequence of a protein. That is, the main function of a start codon is to initiate the process of protein synthesis in the cell. Therefore, this match in my detection is consistent.

The last triplet of nsp16 was not a stop codon but rather is AAC (asparagine). This fact is correct and derives from the labeling standards in the EMBL files in the ORB1ab polyprotein. The stop codon of this polyprotein is after the last codon of nsp16 (the last non-structural protein encoded in ORF1ab). For example, in the case of the complete SARS-CoV-2 sequence OK322850, the labeling of ORF1ab is 212..21492. The labeling of nsp16 is 20596..21489. Therefore, logically, the stop codon refers solely to ORF1ab, which is the polyprotein encoding the

set of non-structural proteins nsp1-nsp16. Hence, the last codon of nsp16 does not coincide with a stop codon.

The first and last triplet of S corresponded to a start codon (ATG) and a stop codon (TAA), which was consistent with what is expected. The same case with the same triplets occurred in the case of ORF3a, E, M, ORF6, ORF7b, ORF8, N.

In the case of ORF7a and ORF10, the first triplet was the start codon ATG but the stop codon was TGA.

Therefore, the start codon in all ORFs was ATG and the stop codon was TAA, except in the case of ORF7a and ORF10, where it was TGA and TAG, respectively.

These coincidences in the first and last triplets with their respective translation as start and stop codons validated the results of my checks.

Once the most common first and last triplets in each genomic element were identified, the next step was to exclude from the analysis those sequences with any labeling errors in the EMBL file, in at least one genomic element. These labeling errors can be typos or other errors in the nucleotide positions of the start and end of each genomic element, among others.

Table 10.6: Number of Sequences Detected Without Labeling Errors. The "Total Number" column represents the total number of sequences correctly labeled in the EMBL for each genomic element. The "Percentage" column represents the % relative to the total of 39,138 fully labeled sequences in their EMBL.

Genomic Element	**Total Number**	**Percentage**
nsp1 (leader protein)	39,102	99.91%
nsp2	39,059	99.80%
nsp3 (papain-like protease)	39,047	99.77%
nsp4	39,041	99.75%
nsp5 (3C-like proteinase)	39,021	99.70%
nsp6	39,005	99.66%
nsp7	38,997	99.64%
nsp8	38,983	99.60%
nsp9	38,975	99.58%
nsp10	38,966	99.56%
nsp11	38,963	99.55%
nsp12 (RNA-dependent RNA polymerase)	38,963	99.55%
nsp13 (helicase)	38,853	99.27%
nsp14 (3'-to-5' exonuclease)	38,817	99.18%
nsp15 (endoRNAse)	38,801	99.14%

Table 10.6 – Continuation from the previous page

Genomic Element	Total Number	Percentage
nsp16 (2'-O-ribose methyltransferase)	38,746	99.00%
ORF1ab/S intergenic region	38,528	98.44%
S	38,744	98.99%
S/ORF3a intergenic region	38,521	98.42%
ORF3a	38,742	98.99%
ORF3a/E intergenic region	38,518	98.42%
E	38,742	98.99%
E/M intergenic region	38,497	98.36%
M	38,739	98.98%
M/ORF6 intergenic region	38,484	98.33%
ORF6	38,731	98.96%
ORF6/ORF7a intergenic region	38,163	97.51%
ORF7a	38,698	98.88%
ORF7b	38,684	98.84%
ORF7b/ORF8 intergenic region	38,112	97.38%
ORF8	38,663	98.79%
ORF8/N intergenic region	38,112	97.38%
N	38,661	98.78%
N/ORF10 intergenic region	38,047	97.21%
ORF10	38,660	98.78%
3'UTR pseudoknot stemLoop 1	38,651	98.76%
3'UTR pseudoknot stemLoop 2	38,648	98.75%
3' stem-loop II-like motif	38,537	98.46%

Therefore, out of the total of 39,138 sequences whose EMBL included labels for all genomic elements, 38,047 were correct. This means that 1,091 sequences with labeling errors in the EMBL were eliminated. This represents 2.79% of the analyzed sequences presenting errors in the EMBL file. It is necessary to demand greater attention to correct labeling in EMBL files prior to their availability in public repositories. I consider this error percentage excessive, indicating deficient verification procedures in the generation of EMBL files.

In most cases among these 1,091 sequences, an error in labeling a genomic element's start and/or end positions translated to subsequent genomic elements. For instance, a defect in the end position of nsp4 would affect the start position of nsp5 (end position of nsp4 + 1), and so forth. Thus, sequences incorrectly labeled in one genomic element would also be incorrect in others.

After excluding all sequences with labeling errors in at least one genomic element, I calculated variability results for 38,047 sequences. This was the lowest figure, corresponding to the N/ORF10 intergenic region. In this area, therefore, more labeling errors were detected (some inherited from other errors and specific to this region).

10.5.2 Consistency Level Results

The results of their ratios, considering these 38,047 sequences, are compiled in Table 10.7.

Table 10.7: Measures of Consistency Level by Genomic Element. "Average Length" represents the mean length of the genomic element in the analyzed sample set, and "Length Standard Deviation" indicates the SD in terms of length, both expressed in number of nucleotides. "Length Consistency Index" and "Nucleotide Consistency Index" respectively indicate the percentage of match in terms of length and nucleotide consistency.

Genomic Element	Average Length	Length Standard Deviation	Length Consistency Index	Nucleotide Consistency Index
nsp1 (leader protein)	539.91	0.97	94.51%	0.55%
nsp2	1,914.00	0.16	98.60%	2.02%
nsp3 (papain-like protease)	5,834.79	0.93	99.23%	0.39%
nsp4	1,450.00	0.15	98.41%	0.66%
nsp5 (3C-like proteinase)	918.0	0.0	100%	19.39%
nsp6	864.69	4.43	98.28%	2.06%
nsp7	249.00	0.00	100.00%	34.54%
nsp8	594.00	0.06	98.02%	31.35%
nsp9	339.00	0.00	100%	29.50%
nsp10	417.00	0.00	100%	28.30%
nsp11	39.00	0.03	84.62%	17.95%
FT stem-loop 1	28.00	0.00	100%	78.57%
FT stem-loop 2	55.00	0.00	100%	49.09%
nsp12 (RNA-dependent RNA polymerase)	27.00	0.03	77.78%	18.52%
nsp13 (helicase)	1,803.00	0.10	98.85%	15.05%

Table 10.7 – Continuation from the previous page

Genomic Element	Average Length	Length Standard Deviation	Length Consistency Index	Nucleotide Consistency Index
nsp14 (3'-to-5' exonuclease)	1,581.00	0.02	99.81%	15.34%
nsp15 (endoRNAse)	1,038.00	0.03	99.71%	3.66%
nsp16 (2'-O-ribose methyltransferase)	894.00	0.00	100%	11.30%
ORF1ab/S intergenic region	12.00	0.00	100%	75.00%
S	3,816.39	4.29	98.44%	0,23%
S/ORF3a intergenic region	10.00	0.00	100%	60.00%
ORF3a	827.99	0.44	96.40%	1.56%
ORF3a/E intergenic region	26.00	0.00	100%	53.85%
E	228.00	0.02	98.68%	23.25%
E/M intergenic region	52.00	0.04	88.68%	7.55%
M	669.00	0.00	100%	7.03%
M/ORF6 intergenic region	12.00	0.00	100%	75.00%
ORF6	185.99	0.66	38.71%	2.69%
ORF6/ORF7a intergenic region	8.00	0.44	8.51%	3.19%
ORF7a	365.96	1.26	55.29%	1.08%
ORF7b	132.00	0.19	88.64%	3.03%
ORF7b/ORF8 intergenic region	8.00	0.17	34.78%	13.04%
ORF8	232.21	141.73	14.96%	3.41%
ORF8/N intergenic region	148.70	141.79	0.62%	0.31%
N	1,259.18	2.60	97.64%	0.95%
N/ORF10 intergenic region	26.00	0.02	92.31%	19.23%
ORF10	117.00	0.01	98.32%	10.92%
3'UTR pseudoknot stemLoop 1	36.00	0.00	100%	38.89%

Table 10.7 – Continuation from the previous page

Genomic Element	Average Length	Length Standard Deviation	Length Consistency Index	Nucleotide Consistency Index
3'UTR pseudoknot stemLoop 2	29.00	0.00	100%	41.38%
3' stem-loop II-like motif	40.18	4.42	21.95%	7.32%

10.6 Discussion

The calculated ratios provide a measure of the consistency degree of each genomic element from multiple perspectives, all of which are complementary.

Consistency ratios are directly proportional measures of homogeneity among sequences, so high values of these ratios indicate high homogeneity.

Consistency and low variability in terms of length are measures of the presence of genomic rearrangements such as insertions or deletions. These phenomena significantly affect the length of the genomic element, so homogeneity in terms of length may imply a lower propensity for insertions and deletions in that genomic element.

Low values of standard deviation in terms of length indicate that the lengths of the sequences cluster in a narrow range around the mean length. Therefore, the lower the value, the more homogeneity in terms of length.

Regarding the proteins encoded by the SARS-CoV-2 genome, the results obtained varied depending on the genomic element.

The coding genomic elements better conserved in terms of nucleotide matching, whose degree of homology was over 30%, were nsp7, nsp8, and 3'UTR pseudoknot stemLoop 1 and 2. In all cases, homology in terms of length remained highly stable.

However, the genomic elements least conserved in terms of nucleotide matching were, firstly, the S protein, followed by nsp3, nsp1, nsp4, and N. In all cases, the degree of nucleotide conservation was less than 1%.

Notably, there was a high standard deviation in the case of ORF8, with its length in the studied set varying around 232.21 ± 141.73. This phenomenon indicates

high variability in terms of length.

The intergenic regions also exhibited heterogeneous behaviors, ranging from highly conserved regions such as those located between ORF1ab and S, S and ORF3a, IRF3a and E, and M and ORF6, to remarkably variable regions such as the ORF8/N intergenic region.

ORF6, ORF7a, 2'stem-loop II-like motif, and ORF8 had values below 60% in terms of Length Consistency Index, resulting in a greater difference between the longest and shortest sequence.

It was noteworthy the similarity in the patterns of these ratios in nsp1, nsp2, nsp3, nsp4, and nsp6. These proteins showed considerable regularity in their length; however, the amount of common nucleotides was conspicuously low, even falling to levels below 1% in the case of nsp1, nsp3, and nsp4.

Nsp5, nsp7, nsp9, and nsp10 also followed similar patterns in the measurements of these ratios. In these four proteins, the Length Consistency Index reached 100%, meaning that in all cases, the length of these non-structural proteins was the same. As for the Nucleotide Consistency Index (the amount of common nucleotides in all the sequences analyzed), it remained higher than in the case of proteins nsp1 to nsp4 and nsp6, ranging from 19.39% in the case of nsp5 to 34.54% in the case of nsp7.

The importance of the functions of these proteins (especially if they are enzymes) in the virus's reproductive cycle means that they cannot afford significant variations in their structure. Nsp5 is essential for the cleavage of the ORF1a and ORF1ab polyproteins [2605], and therefore, it requires specificity and cannot tolerate many modifications in its nucleotide sequence. Similarly, nsp7 and nsp8 are cofactors for the nsp12 polymerase [2612], and nsp8 is involved in the initiation of viral replication [2610]. Thus, it seems logical that these proteins should not undergo excessive variations in their RNA.

Nsp11 and nsp12 showed similar patterns. These are proteins with very short RNA. The genome of nsp11 consists of only 39 nucleotides, and that of nsp12 has only 27 nucleotides, making it the smallest genome among all the proteins of the SARS-CoV-2 genome. The significant value of the Length Consistency Index

in both cases was around 80%. The nucleotide homology was low, around 18%. We must keep in mind that, given the small size of their genomes, any variation has a notable impact in percentage terms.

The ratios of nsp13 and nsp14 showed similar patterns. These proteins turned out to be very stable in terms of length, with Length Consistency Index values ranging from 98.85% in the case of nsp13 to 99.81% in the case of nsp14. In both cases, the Nucleotide Consistency Index remained around 15%.

Nsp15 exhibited some similarity to the patterns of nsp13 and nsp14, but its Nucleotide Consistency Index was notably lower, at only 3.66%, making it a genomic region poorly conserved in terms of nucleotide coincidence.

Nsp16 exhibited a peculiar pattern. Despite the length being exactly the same across all 38,047 sequences (894 nucleotides), the degree of nucleotide agreement was reduced to a mere 11.30%. The low coincidence in this protein, which is otherwise stable in terms of length, is striking.

The length of the structural protein S was $3,816.39 \pm 4.29$. The low standard deviation value is indicative of high homology in the length of the genome of this protein, especially considering the large amplitude of said genome. The difference between the longest and shortest genome was only 1.56%. Therefore, it was a stable protein in terms of length. What was truly striking was the low value of the Nucleotide Consistency Index, at only 0.23%. It is the lowest value among all other genomic elements, including coding, non-coding, auxiliary, etc. This extremely high variability in nucleotide matches aligns with the fact that it is precisely in S where more viral evolution events have occurred, whether mutations and genetic recombination events [2696]. Given that S is responsible for facilitating virus entry into host cells, as it interacts with ACE2 to enter the host cell, it seems logical that a high rate of variations occurs precisely in this protein to find more effective methods to bind to ACE2 and evade the host's immune response [2697].

ORF3a had a length of 827.99 ± 0.44 and a Length Consistency Index reaching 96.40%. Therefore, its length was relatively stable. However, its nucleotide coincidence was only 1.56%, indicating high variability in its genome.

The next structural protein is E, with a length of 228.00 ± 0.02, making it a stable protein in terms of length. However, its common nucleotide content was not as stable, as it was only 23.25%.

In the case of the structural protein M, the contrast between its high homogeneity in terms of length, 669 nucleotides in all sequences, and its low number of coincident nucleotides, only 7.03%, was very striking.

M and E are structural elements and must fit together to form the virion, which explains their stability in terms of length. However, modifications in their sequence (nucleotide variability) may help the virus evade the host's immune system [2642, 2698–2700].

ORF6 exhibited variability in terms of length, 185.99 ± 0.66, with a considerable difference between the shortest and longest sequence, totaling 61.29%. The variability in terms of Nucleotide Consistency Index was even higher, with only 2.69% of common nucleotides.

ORF7a was also an unstable ORF, with a length of 365.96 ± 1.26, and a distance between sequences of extreme length reaching 44.71%. It was also notably variable in terms of common nucleotides, with common characters accounting for only 1.08%.

The length of ORF7b was short and more stable, with a length of 132.00 ± 0.19 and a Length Consistency Index of 88.64%. Common nucleotides accounted for only 3.03% of the total.

The case of ORF8 was remarkable for its high variability in all measured ratios. With an extremely high standard deviation, especially considering the total length of this ORF. The measured length was 232.21 ± 141.73. The Length Consistency Index was only 14.96%, resulting in a distance of 85.04% between the longest and shortest lengths. There was also a low nucleotide coincidence, only 3.41%. This variability is related to the evolution of ORF8, which underwent multiple variations aimed at evading the host immune response [2661]. The role of this protein in modulating the host's immune response and in interacting with cellular and viral proteins [2658, 2659] underscores the need to develop more effective variants to optimize these functions.

The next structural protein was N, more stable in terms of length ($1,259.18 \pm 2.60$), with a difference between extreme lengths of only 2.36%. This is consistent with the fact that it is a structural protein that must fit with the others for proper virion formation. It stood out for its extremely high variability in terms of common characters. Less than 1%, specifically 0.95%, of the nucleotides coincided in all sequences. This may be evolutionarily related to the need for new forms that optimize the role of N in its interaction with other viral proteins and cellular components, which can influence the efficiency and speed of viral replication [2664], as well as its capacity to mitigate the host immune response and virus pathogenesis [2665].

ORF10 had a length of 117.00 ± 0.01, therefore, its standard deviation was very low, and the Length Consistency Index is high, reaching 98.32%. All this denoted high stability in terms of genome length. However, this stability was low in terms of common content, with the Nucleotide Consistency Index being only 10.92%.

Up to this point, the analysis of the coding regions concludes. Upon analyzing the intergenic areas, several noteworthy phenomena were observed.

In general terms, the intergenic regions proved to be more stable in terms of genomic length and common elements than the coding regions, and they were short sequences (8-26 nts) with little variability in their length, except for E/M (52 nts) and ORF8/N (148.7 nts).

All intergenic regions followed the same pattern, except for E/M and ORF6/ORF7a, ORF7b/ORF8, and ORF8/N.

ORF6/ORF7a intergenic region consisted of only 8 nucleotides, subject to variability in length, 8.00 ± 0.44, so the standard deviation was considerable considering the small size of this element. The high difference between the longest and shortest sequences is striking, with this difference being 91.49%. The degree of nucleotide coincidence was only 3.19%. Probably, the very high variability in terms of length also affected the high degree of non-coincidence in terms of common nucleotides.It was a highly unstable region in any case. The same phenomenon of variability, though to a lesser extent, was observed in ORF7b/ORF8.

E/M and ORF8/N were much more variable than other intergenic regions, especially ORF8/N, which was the longest intergenic zone with an average length of 148.70 nucleotides. The standard deviation nearly reached this value, at 141.79%, which was the highest standard deviation observed in this analysis. Therefore, the length was subject to extremely high variability: 148.70 ± 141.79. Additionally, the Nucleotide Consistency Index was only 0.31%. In summary, ORF8/N was the most unstable intergenic region among all intergenic regions.

As for the following auxiliary elements located in the 3'UTR:

3'UTR pseudoknot stemLoop 1 was also absolutely stable in the length of its genome (36 nucleotides) but not as much in terms of common nucleotides. The collection of sequences only shared 38.89% of nucleotides, indicating considerable variability at that point.

A similar case was seen in 3'UTR pseudoknot stemLoop 2, with a completely stable length of 29 nucleotides but a Nucleotide Consistency Index of 41.38%.

Both elements exhibit a similar pattern in their ratios.

Regarding the 3' stem-loop II-like motif, the measured length was 40.18 ± 4.42, indicating variability in terms of length. The low value of the Length Consistency Index is striking, at only 21.95%, resulting in a difference of 78.05% between the shortest and longest genome. This high variability was aligned with the scarcity of common elements, only 7.32%. Therefore, this genomic element was highly variable.

The lowest value of the Length Consistency Index (excluding ORF8/N due to its extremely high Length Standard Deviation) was ORF8, with 14.96%. The second worst was ORF6 (38.71%).

The highest value of the Length Consistency Index was found in nsp5, nsp9, nsp10, nsp16, M, and the intergenic zones ORF1ab/S, S/ORF3a, ORF3a/E, M/ORF6, as well as in 3'UTR pseudoknot stemLoop 1 and 3'UTR pseudoknot stemLoop 2. In all these cases, it reached 100%.

The lowest value of the Nucleotide Consistency Index was in S (0.23%). This is an extremely low value, especially considering the relative stability regarding length-related ratios (length of 88.35 ± 32.21 versus a Length Consistency Index

of 98.44%). It is precisely this protein that has the highest mutation rate among the genes that make up SARS-CoV-2, both fixed and polymorphic, likely highly influenced by C to U deamination [2701]. Therefore, the obtained results were aligned with the biological specificities of this protein.

Nsp1, nsp3, nsp4, and N also resulted in Nucleotide Consistency Index values lower than 1%. In nsp1, mutations effectively occurred resulting in variations in structural stability [2702] and deletions leading to new strains [2703]. Nsp3 experienced high numbers of unique mutations [2704], and hotspots of insertion/deletion were detected in it [2705]. The genomic region corresponding to nsp4 is sensitive to non-synonymous mutations (G9802U/C9803U) [2706]. N has accumulated up to 1,034 unique nucleotide mutations [2707].

The highest values of the Nucleotide Consistency Index were for the 3'UTR pseudoknot stemLoop 1 and 2, with values of 38.89% and 41.38% respectively, S/ORF3a intergenic region with 60.00%, ORF3a/E intergenic region with 53.85%, and M/ORF6 intergenic region with 75.00%. It was notable that the regions with the most nucleotide coincidence were intergenic zones. In all these fragments, the homology in lengths was 100%. These regions are typically short genomic fragments that play a crucial role in viral replication, making it logical for them to be stable zones to maintain their functionality [2708, 2709].

10.7 Conclusions

In this Chapter, I aimed to analyze the degree of consistency of each genomic element in SARS-CoV-2, both as guidance for the generation of synthetic sequences and as a measure of validation (for reference) of predictions generated by a generative model.

The genomic region with the lowest consistency indices was precisely the S protein, which, according to expert literature, accumulates the highest mutation percentages. The function of this protein allows the virus to penetrate the cell. It is precisely in this protein where the highest number of mutations concentrates resulting in new variants, which are discriminated precisely by their implicit modifications in S. Therefore, these genomic consistency study was coincident with the results obtained

in the microbiological and genomic fields in terms of the degree of mutational profile per genomic element.

I detected a correlation between the degree of mutation by genomic region and my consistency measures, which is logical. Indeed, the nsp7 and nsp8 proteins showed high degrees of conservation. However, the four structural proteins (S, E, M, and N), as well as ORF7a and 7b, nsp3, nsp4, and nsp12, exhibited high degrees of mutation.

My results were consistent with the detected mutation rates, especially in the case of nsp7 and nsp8, which in my case were the most consistent regions in terms of both length and nucleotide matching. Regarding the less conserved regions, similarly, especially in S, N, nsp3, and nsp4, my results were coherent in terms of nucleotide matching degree.

Exploring each SARS-CoV-2 genomic element more thoroughly involves examining its conservation level and tendency for modifications (such as mutations, deletions, insertions, etc.). These analyses can provide a probabilistic framework for understanding the phylogenetic consequences and viral evolution.

War is the only proper school for a surgeon.

— Hippocrates, *a strong advocate for the importance of keeping the patient clean and sterile. This practice was lost until 1834 when Semmelweis advocated for treating the patient with clean hands. Four hundred years BC, Hippocrates already knew this.*

11

SARS-CoV-2 Genetic Recombination Analysis

Contents

11.1 Introduction 590
11.2 Coronaviruses Microbiological Considerations 592
11.3 Materials and Methods 599
11.3.1 SARS-CoV-2 complete genomic sequences compendium 599
11.3.2 Dataset Design 601
11.3.3 Generation of Genomic Spectrograms 602
11.3.4 Two-Stage Transfer Learning 604
11.3.5 Test Bench 605
11.3.6 Performance Measurement 605
11.3.7 Explainability Analysis 605
11.4 Results 609
11.4.1 Results per Subsampling 609
11.4.2 Results per Number of Epochs 611
11.5 Discussion 612
11.5.1 Optimal Configuration Selection 614
11.5.2 Reference Sequence 614
11.5.3 Analysis of Non-Recombinant Results 616
11.5.4 Analysis of Recombinant Results 617
11.6 Conclusions 621

11.1 Introduction

Understanding the evolution of viruses remains a significant challenge in pandemic control even today. The two main mechanisms associated with viral evolution are mutation and genetic recombination. Both occur more frequently during viral replication [93, 2710]. The phenomenon of mutation consists of the introduction of random errors into the genetic material of a virus, which can result in the emergence of genetic variants of the same virus [2711]. Genetic recombination is a completely distinct phenomenon that involves the exchange of genetic information between two viral genomes, whether they are of the same or different viruses, leading to the formation of a new hybrid genome [9, 117]. Both mechanisms have the potential to produce viable variants that hinder the prevention and treatment of infectious diseases.

Genetic recombination is observed in numerous RNA viruses, and it has been identified frequently in picornaviruses [116], coronaviruses [2475, 2712], and retroviruses [2500]. The viable viruses resulting from this recombination can vary greatly in their characteristics and, therefore, their consequences on public health. In some situations, genetic recombination can lead to the generation of viruses that retain their behavior or pathogenicity without significant changes, or that remain relatively benign. However, in other scenarios, the recombinant virus can pose a health risk due to the emergence of a new strain with new or unique harmful characteristics, such as increased transmission capacity or virulence [2441, 2713–2718]. Even in rare instances, failures in attenuated virus vaccines have led to events of genetic recombination [2719, 2720]. If genetic recombination occurs between two distinct viruses and produces a new viable virus, the implications regarding viral evolution and the impact on public health can be magnified [2721]. Such is the case of the emergence of SARS [2722] and MERS [2723].

SARS-CoV-2, an abbreviation for Severe Acute Respiratory Syndrome Coronavirus 2, is a newly identified coronavirus that surfaced in late 2019 and rapidly evolved into a worldwide pandemic [2724]. Its high contagion capacity contributed to the pandemic tragedy of Coronavirus Disease 2019 (COVID-19). The symptoms

of SARS-CoV-2 vary in severity, ranging from mild respiratory issues to severe pneumonia and, in some cases, death [2725]. According to World Health Organization (WHO), as of September 1st, 2023, COVID-19 has caused 770,437,327 confirmed cases, including 6,956,900 deaths [2726].

Since the outbreak of the pandemic in January 2020 [2727], several variants emerged with particular evolutionary characteristics leading to either increased transmissibility or enhanced vaccine evasion. Among the Variants of Concern (VOCs), Alpha (B.1.1.7) stood out, reaching its zenith around April 2021 [2728]. Another VOC of significant impact was Delta (B.1.617.2), reaching its peak around September 2021 [2729, 2730]. From early 2022 to the present day, Omicron and its different subvariants (B.1.1.529, BA.1 - BA.5, BQ.1, etc.) have dominated the epidemiological landscape [2731, 2732].

The recombination events detected in SARS-CoV-2 have not led to substantial modifications in terms of morbidity, mortality, etc. [2733]. Most of these recombinants have occurred between co-circulating Omicron sublineages, such as BA.1 (or BA.1.1), and the Delta variant or BA.2 [2734]. However, the combination of Omicron and Delta through genetic recombination could conceivably yield a virus possessing Omicron's transmissibility alongside Delta's heightened potential for severe illness, presenting a fresh and more worrisome scenario for public health [2735]. Japan reported a small set of recombinant sequences between Alpha and Delta, which ultimately did not survive and had a merely anecdotal impact [2731].

Comprehending genetic recombination phenomena is of utmost importance in the monitoring and management of the emergence of novel viruses or variants, as well as in foreseeing and tackling future pandemics [2736].

Deep Learning tools can provide new insights into the study of genetic recombination [1372, 1375] and the anticipation of new pandemics or emerging viruses [2737]. By analyzing large genomic datasets and identifying patterns, Deep Learning tools can detect signs of genetic recombination in viruses more quickly than traditional methods. In this way, we can not only reduce response times to emerging viruses but also anticipate their very emergence. Likewise, we can unravel the mysteries

of the genetic code from new perspectives, such as the search for mathematical patterns within the genome itself.

This Chapter begins with an explanatory section on some microbiological aspects of interest regarding coronaviruses (Section 11.2). In Section 11.3, I detail the methodology used and the equipment employed. The Chapter continues with the presentation of results (Section 11.4), followed by a discussion (Section 11.5). The chapter concludes with the main conclusions (Section 11.6).

11.2 Coronaviruses Microbiological Considerations

The *Coronaviridae* family belongs to the order *Nidovirales* along with the families *Roniviridae* and *Arteriviridae* [74]. The *Coronaviridae* family consists of two subfamilies, the *Coronavirinae* and the *Torovirinae*.

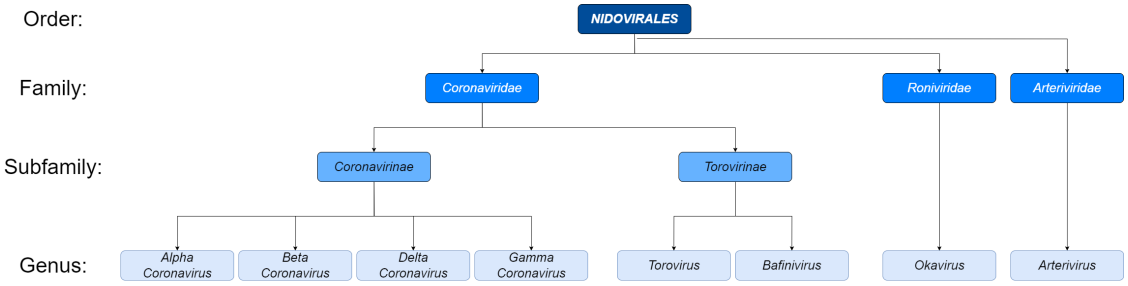


Figure 11.1: *Nidovirales* Taxonomy. This taxonomic classification is primarily based on phenotypic characteristics according to the International Committee on Taxonomy of Viruses (ICTV) [2738]. This image is an excerpt from "Fields Virology" [74].

The viruses of the *Roniviridae* family primarily infect arthropods, especially crustaceans [2739] [2739]. The *Arteriviridae* family and *Torovirinae* subfamily do not infect humans [74]. However, subfamily *Coronavirinae* cause common respiratory and enteric diseases in humans and domestic animals.

Coronaviruses are single-stranded positive-sense RNA viruses, meaning their genome consists of a single RNA strand that can be directly read to produce proteins. The distinctive feature of coronaviruses is their viral envelope, which is adorned with spike-shaped projections giving them a crown-like appearance when viewed under an electron microscope.

Coronavirus genomes are monopartite, positive-sense RNA, polyadenylated, and capped, ranging from 27 to 32 kb in length. They are the longest among known positive-sense RNA viruses. They can function as mRNA, and purified genomic RNA is infectious. Their organization is in a nested set structure, meaning they contain a main set of coding genes that are transcribed into messenger RNA (mRNA), as well as smaller subsets of coding genes that are transcribed from subgenomic mRNA [2740].

Some of the features specific to Coronaviridae include the nested set structure of 3'-coterminal mRNAs, the unique RNA transcription strategy, genome organization, nucleotide sequence homology, and properties of their structural proteins [2741].

Coronaviruses have a complex replication cycle involving virus binding to specific receptors on the surface of the host cell, virus entry into the cell, release of its genome into the cytoplasm, synthesis of viral proteins, and replication of the viral genome. This process can be modulated by a variety of factors, including the interaction between viral proteins and host cell proteins [2742].

Most coronaviruses infect only cells of their natural host species and some closely related species. In this regard, SARS-CoV-2 is entirely anomalous, as it infects multiple and varied different species, including humans, felines, canines, ferrets, bovines, equines, porcines, hominoids, etc. This phenomenon has never been documented before [2743–2745].

Some strains of coronavirus can infect a broader range of cells under certain experimental conditions; for example, intracerebral inoculation of some strains of Murine Hepatitis Virus (MHV), a betacoronavirus, in owl monkeys and African green monkeys can result in central nervous system infection [2746, 2747].

In their natural host species, coronaviruses exhibit marked tissue tropism; for instance, Transmissible Gastroenteritis Virus (TGEV) primarily causes gastrointestinal diseases in pigs, while a closely related strain, Porcine Respiratory Coronavirus (PRCoV), predominantly causes respiratory infections in pigs. The species and tissue specificity of the virus is at least partially determined by the nature and

distribution of cellular receptor molecules and variations in the sequence of the viral Spike (S) glycoprotein [2748, 2749].

The primary determinant of viral tropism is likely the binding of the S protein to a specific receptor glycoprotein on the cell surface. Receptors for various coronaviruses have been identified [2750].

Like most RNA viruses, coronaviruses are believed to mutate at a high frequency due to the high error rates of RNA polymerases [2751].

Deletion mutations also occur frequently. The most prominent example of the biological significance of deletion mutations is the emergence of PRCoV from TGEV. While TGEV causes enteric epizootic infections in pigs, in the early 1980s, PRCoV emerged in Europe as a new virus causing widespread epizootic respiratory infections in pigs. Nucleotide sequence comparisons of the two viruses showed that PRCoV was a derivative of TGEV with a large deletion within the S protein. Similar PRCoV strains have independently emerged in the United States as deletion mutants in the TGEV S gene [2752, 2753].

Likewise, Feline Infectious Peritonitis Virus (FIPV) probably arose from deletion mutations of Feline Enteric Coronavirus (FECoV) in a similar manner [2754].

A unique feature of coronavirus genetics is a high frequency of RNA recombination, especially in Murine Hepatitis Virus (MHV). Although the non-segmented genomes of RNA viruses generally exhibit very low or undetectable recombination frequencies, recombination frequencies for the entire coronavirus genome have been calculated to be up to 25% [2755].

The high frequency of RNA recombination in coronaviruses is likely the result of the unique coronavirus RNA synthesis mechanism, which involves discontinuous transcription and polymerase hopping. It is possible that the viral polymerase associated with incomplete nascent RNAs dissociates from its template at a random point and switches to a homologous site on a different RNA template to complete RNA synthesis through a copy-choice mechanism [2756, 2757].

RNA recombination is an important mechanism in the natural evolution of coronaviruses. For example, new strains of Infectious Bronchitis Virus (IBV) in

poultry flocks result from natural recombination between different field strains [2758]. A plausible hypothesis is that an inversion in the order of the M gene and the S gene in the genomes of IBV and MHV may have occurred due to homologous recombination in consensus intergenic regions between the two genes in an ancestor of one of the viruses [2759, 2760]. Recombination may also explain the acquisition of the hemagglutinin-esterase (HE) gene from an influenza C virus mRNA by an ancestor of group II coronaviruses [2761–2763]. Some biotypes of FECoV appear to have arisen through recombination between FECoV and Canine Coronavirus (CCoV) [2764].

Coronaviruses are unique among RNA viruses in many aspects of their viral structure and replication. They are characterized by exceptionally large RNA genomes, nested subgenomic mRNAs, a mechanism of discontinuous transcription, high-frequency RNA recombination, and an unusually large number of gene products associated with RNA synthesis [117, 2765, 2766].

The RNA of coronaviruses incorporates two different gene expression strategies: polyprotein processing for gene 1 (ORF1a in SARS-CoV-2) and expression of a gene to an mRNA for the other genes, likely reflecting their uncommon evolutionary origin [2767, 2768]. Many questions about the replication cycle of coronaviruses remain unresolved, including the mechanism of mRNA transcription, the functions of gene 1 proteins and other non-structural proteins, and the mechanisms of virus entry and assembly [2769, 2770]. Recent findings suggest that the structure of virions and the role of cellular factors in viral replication must be reevaluated [2771, 2772]. Research on coronaviruses has been hindered by the lack of infectious complementary DNA (cDNA) or RNA transcripts, rendering standard reverse genetics approaches inapplicable [2773, 2774].

The construction of a functional full-length genomic cDNA has been challenging due to the large size of the coronavirus RNA genome [2775]. The construction of an infectious coronavirus cDNA (for TGEV) was only recently achieved [2776]. These infectious recombinant genomes will undoubtedly facilitate research on coronaviruses. Furthermore, studies of related but less complex toroviruses and arteriviruses should

also enhance our understanding of the exceptionally complex replication cycle of coronaviruses.

Orthocoronavirinae, commonly known as coronaviruses, is a subfamily of positive-sense single-stranded RNA viruses belonging to the family Coronaviridae. It is subdivided into the genera Alphacoronavirus, Betacoronavirus, Deltacoronavirus, and Gammacoronavirus [2777, 2778].

AUTHOR'S NOTE:

The list of viruses presented below is the result of painstaking search and data aggregation from multiple incomplete sources. I was unpleasantly surprised by the absence of a comprehensive and unequivocal coronavirus report that is easily accessible to the research community and the general public.

This list is likely one of the most comprehensive available. Therefore, while it may not be specifically relevant to the research undertaken, I believe it is a contribution to science that should be considered.

Consequently, I decided to retain this list in this document.

Some of the major coronaviruses classified within the genus Alphacoronavirus are [2779–2783]:

- Canine coronavirus type I (CCoV Type I) (host: dog) [2784].
- Canine coronavirus type II (CCoV Type II) (host: dog) [2785].
- Feline coronavirus type I (FCoV Type I) (host: cat) [2786].
- Feline coronavirus type II (FCoV Type I) (host: cat) [9].

I must note that within FCoV there are two types of viruses: The first type is Feline Enteric Coronavirus (FECoV), which infects the gastrointestinal epithelial cells of the feline [2787]. The second type is the Feline Infectious Peritonitis Virus (FIPV) that causes a severe and often fatal disease in cats. Mortality data for FIPV vary widely, but the mortality rate is estimated to be very high, up to 95-100% of diagnosed cases. The disease is complex and can affect various systems of the cat's body, including the gastrointestinal, respiratory, renal, and nervous systems [2788]. It appears to originate from a mutation of FECoV [2789, 2790].

- Transmissible gastroenteritis virus (TGEV) (host: pig) [2791].

- Microhyla letovirus 1 (MLeV) (host: frog) [2792].

- Porcine respiratory coronavirus (PRCoV) (host: pig) [2793].

- Lucheng Rn rat coronavirus (LRNV) (host: rodents) [2794].

- Human coronavirus 229E (HCoV-229E) (host: human) [2795].

- Human coronavirus NL63 (HCoV-NL63) (host: human) [2796].

- NL63-related bat coronavirus strain BtKYNL63-9b (BtKYNL63) (host: bat) [2797].

- Bat coronavirus CDPHE15 (BtCoV-CDPHE15) (host: bat) [2798].

- Bat coronavirus HKU10 (host: bat) [2799].

- Miniopterus bat coronavirus HKU8 (BtMiCoV-HKU8) (host: bat) [2800].

- Rhinolophus bat coronavirus HKU2 (BtRhCoV-HKU2 (SADS)) (host: bat) [2801].

- Miniopterus bat coronavirus 1 (BtMiCoV-1) (host: bat) [2802].

- Rhinolophus ferrumeguinum alphacoronavirus HuB-2013 (BtRfCoV-HuB13) (host: bat) [2803].

- Porcine respiratory coronavirus (PRCoV) (host: pig) [2804].

- Porcine epidemic diarrehoea virus (PEDV) (host: pig) [2805].

- Scotophilus bat coronavirus 512 (BtScCoV-512) (host: bat) [2806].

- Mink coronavirus I (MCoV) (host: mink) [2807].

- Ferret coronavirus (FRCoV) (host: ferret) [2808].

- Myotis ricketti alphacoronavirus Sax-2011 (BtMy-Sax11) (host: bat) [2809].

- Nyctalus velutinus alphacoronavirus SC-2013 (BtNy-Sc13) (host: bat) [2810].

The genus Betacoronavirus is indeed reported to be divided into five subgenera: Sarbecovirus, Hibecovirus, Nobecovirus, Merbecovirus, and Embecovirus. The main betacoronaviruses are [2779–2783]:

- Bovine coronavirus (BCoV) (host: cow) [2811].

- Equine coronavirus (ECoV) (host: horse) [2812].

- Human coronavirus OC43 (HCoV-OC43) (host: human) Embecovirus [2813].

Some specialized texts refer to Betacoronavirus I as the group formed by BCoV, ECoV, and HCoV-OC43 [2814].

- Alpaca CoV (host: alpaca). This is BCoV-like coronavirus [2815].

- Human coronavirus HKU1 (HCoV-HKU1) (host: human) Embecovirus [2795].

- Porcine hemagglutinating encephalomyelitis virus (PHEV) (host: pig) [2816].

- China Rattus coronavirus HKU24 (RtCoV-HKU24) (host: rat) [2817].

- Murine coronavirus, also called Murine Hepatitis Virus (MHV) (host: mouse and rat) [2818, 2819].

- Rabbit coronavirus HKU14 (RbCoV HKU14) (host: rabbit) [2820].

- Tylonycteris bat coronavirus HKU4 (BtTyCoV-HKU4) (host: bat) [2821].

- Pipistrellus Bat coronavirus HKU5 (BtPiCoV-HKU5) (host: bat) [2822].

- Roussetus bat coronavirus GCCDC1 (BtEoCoV-GCCDC1) (host: bat) [2823].

- Rousettus bat coronavirus HKU9 (BtRoCoV-HKU9) (host: bat) [2824].

- Severe Acute Respiratory Syndrome CoV (SARS-CoV)(host: human) [2825].

- Severe Acute Respiratory Syndrome CoV 2 (SARS-CoV-2) (host: multiple species) [27].

- Middle East respiratory syndrome-related coronavirus (MERSr-CoV) (host: camelids and human) [2826, 2827].

- Hedgehog coronavirus 1 (EriCoV-1) (host: hedgehog) [2828].

- Canine Respiratory Coronavirus (CRCoV) (host: dog) [2829, 2830].

- Bat Hp-betacoronavirus Zhejiang2013 (BtHpCoV-ZJ13) (host: bat) [2831].

- Pangolin coronavirus (Pangolin-CoV) (host: pangolin) [2832].

The main deltacoronaviruses are [2780–2783]:

- Bulbul coronavirus HKU11 (BuCoV-HKU11) (host: bulbul) [2833].

- Common moorhen coronavirus HKU21 (CMCoV-HKU21) (host: moorhen) [2778].

- Coronavirus HKU15 (PoCoV-HKU15) (host: pig) [2834].

- Munia coronavirus HKU13 (MuCoV-HKU13) (host: munia) [2835].

- Night heron coronavirus HKU19 (NHCoV-HKU19) (host: night heron) [2836].

- Thrush coronavirus HKU12 (host: thrush) [2833].
- White-eye coronavirus HKU13 (WECoV-HKU13) (host: white-eye) [2837].
- Wigeon coronavirus HKU20 (WiCoV-HKU20) (host: wigeon) [2838].
- Porcine coronavirus HKU15 (host: pig) [2834].

The main gammacoronaviruses are [2780–2783]:

- Infectious bronchitis virus (IBV) (host: chicken and turkey) [2781].
- Beluga whale coronavirus SW1 (BWCoV-SW1) (host: marine mammals) [2839].

SARS-CoV-2 falls within the betacoronaviruses. Historically, betacoronaviruses have been responsible for outbreaks and pandemics of severe respiratory diseases in humans, such as Severe Acute Respiratory Syndrome (SARS), Middle East Respiratory Syndrome (MERS), and recently, the Coronavirus Disease 2019 (COVID-19) pandemic [2840, 2841].

In the veterinary field, betacoronaviruses are also a significant concern as they can affect a variety of animal species, including cattle, pigs, and poultry. These infections can lead to respiratory, gastrointestinal syndromes, and in some cases, severe systemic diseases, posing a challenge to livestock and poultry production [2842–2844].

11.3 Materials and Methods

All experiments were run in this equipment:

- Processing Unit: Intel(R) Core(TM) i7-4770K CPU. 3.5 GHz.
- Installed RAM: 32 GB usable.
- Operative System: Windows 10 Education. Version: 22H2.
- GPU: NVIDIA GeForce RTX 3090. Total memory: 40 GB.

11.3.1 SARS-CoV-2 complete genomic sequences compendium

In March 2023, I downloaded all complete SARS-CoV-2 sequences per variant. I performed the downloads from the public repository National Center for Biotechnology Information, Virus Database (NCBI Virus Database) [40]. The total number of

downloaded sequences was 1,541,293. Of these, 1,539,728 were categorized as non-recombinant, and 1,565 were recombinant. Tables 11.1 and 11.2 detail the number of each variant sequences classified as non-recombinant and recombinant, respectively.

The first Variant of Concern (VOC) was the Alpha variant, whose emergence was established around November 2020 [2845]. For classification purposes, I designated variants identified from January 2020 to November 2020 as pre-VOC variants. Table 11.1 shows the compilation of non-recombinant sequences by variants (pre-VOC and VOC).

Table 11.1: Non-recombinant SARS-CoV-2 sequences compendium. Obtained from the NCBI Virus Database (National Center for Biotechnology Information, Virus Database) [40], the total of 1,539,728 sequences corresponded to an approximate date of March 2023 (Release Date of NCBI Virus Database). The variants were arranged based on the approximate date of appearance (data retrieved from GISAID Initiative - Tracking of hCoV-19 Variants) [2846]. The "Variant" column denotes the WHO Name of the SARS-CoV-2 variant [2726].

Non-recombinant SARS-CoV-2 Sequences		
Variant	**No.**	**Percentage**
pre-VOC	218,198	14.17%
Alpha	198,722	12.91%
Beta	856	0.06%
Gamma	11,937	0.78%
Delta	325,285	21.13%
Epsilon	14,781	0.96%
Eta	738	0.05%
Iota	19,361	1.26%
Kappa	145	0.01%
Lambda	456	0.03%
Mu	49	0.00%
Theta	12	0.00%
Zeta	553	0.04%
Omicron	748,635	48.62%
TOTAL	**1,539,728**	**100%**

The predominant variants (in chronological order) included, primarily, the group referred to as pre-VOC, consisting of 218,198 sequences, followed by Alpha, with 198,722 sequences, Delta, with 325,285 sequences, and most notably, Omicron, along with its various sub-variants, totaling 748,635 variant sequences. The influence of

the remaining variants was comparatively limited, given the dominance of these prevalent ones that acquired an evolutionary edge. As a result, the rest of the variants represented only 48,888 sequences.

Table 11.2 displays the compilation of recombinant sequences of SARS-CoV-2 categorized by variants. Compared to non-recombinant sequences, the compilation of recombinant sequences exhibited a more balanced distribution, albeit with a slight predominance of XBB sub-variants followed by XE and XC.

Table 11.2: Recombinant SARS-CoV-2 sequences compendium. Obtained from the NCBI Virus Database [40], the total of 1,565 sequences corresponded to an approximate date of March 2023 (Release Date of NCBI Virus Database). The nomenclature for recombinant variants begins with an "X" in the Pango nomenclature [2847]. The "Variant" column denotes the WHO Name of the SARS-CoV-2 variant [2726].

Recombinant SARS-CoV-2 sequences					
Variant	**No.**	**Percentage**	**Variant**	**No.**	**Percentage**
XA	2	0.13%	XBB.3	62	3.96%
XAA	47	3.00%	XBB.4	8	0.51%
XAB	2	0.13%	XBD	30	1.92%
XAC	37	2.36%	XBE	32	2.04%
XAD	1	0.06%	XC	227	14.50%
XAE	18	1.15%	XD	5	0.32%
XAF	16	1.02%	XE	256	16.36%
XAG	8	0.51%	XF	4	0.26%
XAH	1	0.06%	XG	1	0.06%
XAJ	52	3.32%	XH	4	0.26%
XAK	21	1.34%	XL	11	0.70%
XAM	59	3.77%	XM	10	0.64%
XAN	5	0.32%	XN	16	1.02%
XAP	41	2.62%	XP	7	0.45%
XAT	2	0.13%	XQ	14	0.89%
XAU	4	0.26%	XS	24	1.53%
XB	19	1.21%	XW	36	2.30%
XBB	67	4.28%	XY	63	4.03%
XBB.1	88	5.62%	XZ	59	3.77%
XBB.2	206	13.16%	**TOTAL**	**1,565**	**100%**

11.3.2 Dataset Design

Throughout the SARS-CoV-2 pandemic, non-recombinant variants have been prevalent, especially during its most severe phases [2848]. This dominance of non-

recombinant variants resulted in a substantially larger number of sequences compared to recombinants. To address this imbalance, I employed a subsampling technique within the larger non-recombinant dataset. This approach involved randomly selecting representative subsets from the non-recombinant dataset, equalizing the number of data points between both categories. This strategy helped mitigate potential biases in my analysis and enhanced the reliability of my findings [2849]. I randomly selected 1,565 non-recombinant sequences to create a balanced dataset. To ensure the robustness of my results, I conducted multiple subsamplings. Specifically, I performed 10 subsamplings labeled with sequential numbers from 01 to 10 (SUB_01 - SUB_10).

After completing the subsampling of non-recombinant sequences, they were randomly allocated to the Training, Validation, and Test Sets [38], as outlined in Table 11.3, which provides an overview of the composition of each dataset resulting from the subsampling process.

Table 11.3: Dataset structure. All experiments were conducted using three balanced datasets encompassing both categories, with 60% randomly allocated to the Training Set, 20% to the Validation Set, and 20% to the Test Set.

Dataset	Non-recombinant	Recombinant	TOTAL
Training	939	939	1,878 (60%)
Validation	313	313	626 (20%)
Test	313	313	626 (20%)
TOTAL	**1,565**	**1,565**	**3,130**

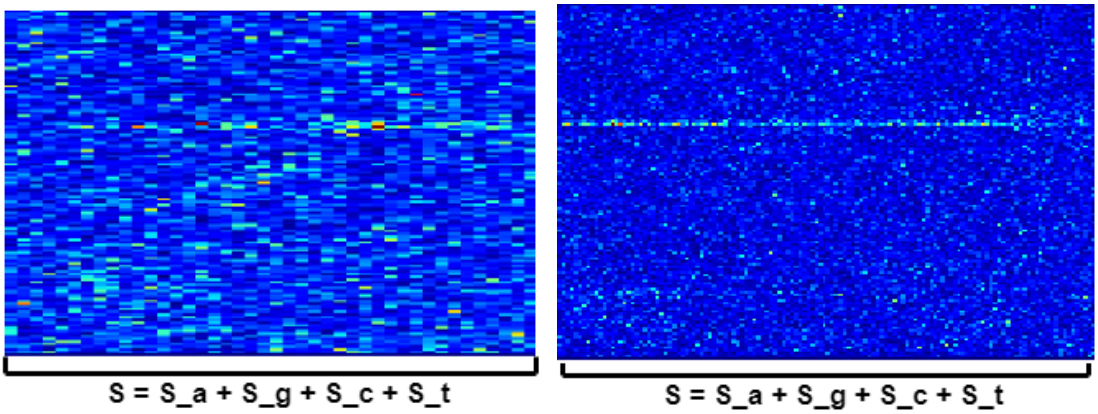
11.3.3 Generation of Genomic Spectrograms

The process of generating spectrograms adheres to the methodology outlined in my prior study on superposed spectrograms [2850]. In this method, I implemented Transfer Learning with a pre-trained Convolutional Neural Network (CNN) VGG-16 [197], originally trained on the ImageNet dataset [2851]. The network was initially trained to identify recombinant features within complete HIV-1 sequences.

In this spectrogram representation, the z-axis represents the summation of values from each of the four nucleotide types along the z-axis.

$$S = S_a + S_g + S_c + S_t \tag{11.1}$$

Figure 11.2 displays examples of the genomic spectrogram of HIV-1 (Figure 11.2a) and SARS-CoV-2 (Figure 11.2b).



(a) HIV-1 **(b)** SARS-CoV-2

Figure 11.2: Genomic Spectrogram Schemes. The length of the x-axis corresponds to the genome's length. For HIV-1, the x-axis spans approximately 10,000 nucleotides (Figure 11.2a), while for SARS-CoV-2, it is around 30,000 nucleotides (Figure 11.2b). The y-axis denotes a frequency range from 0 to 0.5 Hz. The z-axis signifies the spectrogram computation with a jet colormap scale applied (Figure 4.3).

Given that the length of the SARS-CoV-2 genome is approximately three times greater than that of the HIV-1 genome, this disparity is reflected in the spectrogram. This is primarily due to the fixed value of 256 used as the length of each time segment for FFT calculation (nperseg). Consequently, in the spectrogram of SARS-CoV-2, the x-axis appears three times longer than in the case of HIV-1, resulting in smaller color points along the z-axis [2389].

In both scenarios, the horizontal line at $f = 1/3$ is clearly visible. This observation may be related to the conversion of three nucleotides into a single amino acid in the coding regions of the genome [355]. Consequently, it is expected that this line will be sharp in these coding regions but less pronounced in non-coding regions. In the case of viruses, a significant portion of the genome is coding [2852]. Given that a substantial portion of viral genomes comprises coding segments [2852], this line is distinctly visible across almost the entirety of the genomic spectrogram.

This graphical representation of genome in the frequency spectrum allows for a more accurate identification of the genome regions crucial for the recombinant feature.

11.3.4 Two-Stage Transfer Learning

The imbalance between the number of recombinant and non-recombinant sequences in SARS-CoV-2 made it advisable for the CNN to learn from a balanced and well-documented case study, such as HIV-1. This knowledge acquired from HIV-1 could then be applied to the specific case of SARS-CoV-2.

Therefore, I performed a Two-Stage Transfer Learning process as indicated in 11.3. The first of these stages started from the network derived from Chapter 5 for superposed spectrograms.

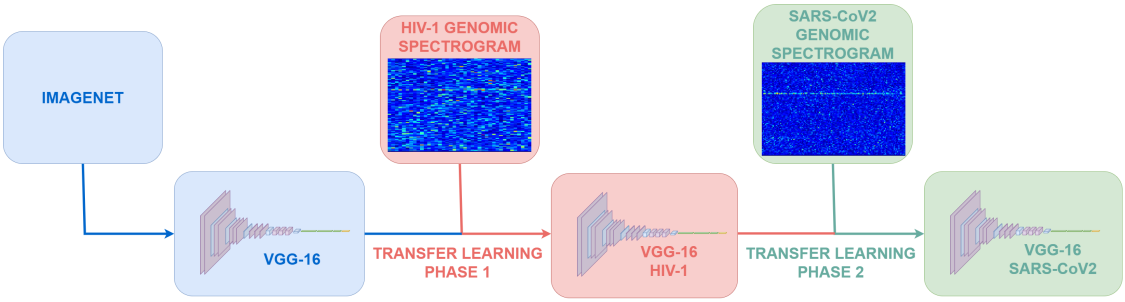


Figure 11.3: Two-Stage Transfer Learning Methodology

My starting point was a pre-trained VGG-16 trained with the ImageNET dataset. In Phase 1, I applied Transfer Learning to this pre-trained model on the genomic spectrogram dataset of complete HIV-1 sequences to detect the recombinant feature. In Phase 2, I applied Transfer Learning once again to the resulting network from Step 1 (VGG-16 HIV-1) using a genomic spectrogram dataset of complete SARS-CoV-2 sequences to also detect the recombinant feature (VGG-16 SARS-CoV-2).

All the experiments were performed using the MATLAB2021b App Deep Learning Designer.

11.3.5 Test Bench

I performed a test bench on each of the 10 subsamplings, assessing hyperparameter values with a fixed Learning Rate of 0.0001 and a fixed Batch Size of 52, while adjusting the number of Epochs to 10, 25, 50, 75, 100, 150, 200, and 250.

As a result, I executed a total of 80 experiments, yielding 80 VGG-16 models trained through Two-Stage Transfer Learning.

Initially, I conducted preliminary tests with a Batch Size of 128 and a Learning Rate of 0.01, but obtained suboptimal results. Subsequently, I determined that the optimal values for Batch Size and Learning Rate in this second stage of Transfer Learning were consistent with those utilized in the first stage (HIV-1).

11.3.6 Performance Measurement

Our performance measurement criteria were based on the following ratios:

- Validation Accuracy.
- Test Accuracy.
- Area Under the Curve (AUC) in the Test Set.
- Confusion Matrix in the Test Set.

I calculated the four ratios in all experiments.

I based the criteria for determining optimal configurations on selecting those that exhibit exceptional values of AUC and Validation Accuracy. Additionally, I sought configurations that not only attain the highest Test Accuracy but also maintain a balance between both categories (recombinant and non-recombinant) [2853].

I calculated all performance measurements using MATLAB2021b.

11.3.7 Explainability Analysis

I utilized explainability techniques, via heatmaps, to identify the crucial factors influencing the outputs of each generated model. My preferred tool for pinpointing genome regions where the network makes decisions was Grad-CAM [2854]. It provides enhanced visual clarity compared to other tools like LIME [231] or Gradient Attribution [2855], although it may sacrifice some precision.

The color scale employed in these heatmaps was a jet map, which represents the color distribution based on the scoremap value at each point, as depicted in Figure 4.3.

To analyze the Grad-CAM results, I conducted a three-step image processing to systematically identify the pertinent Total Hot Zones within the recombinant feature. Figure 11.4 visually depicts the methodology employed for explainability analysis.

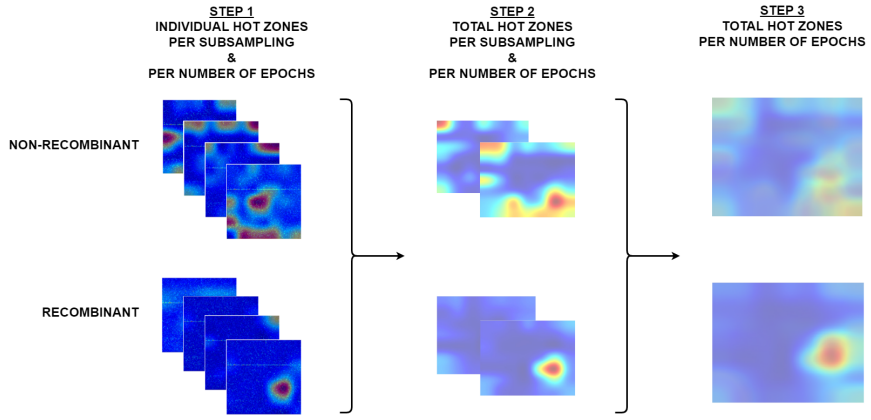


Figure 11.4: Three-step Explainability Scheme. Given that I conducted 80 experiments (test bench applied to 10 subsamplings), the initial step encompassed a total of 50,080 images, considering that each complete test set contains 626 sequences. The second step involved 160 images (80 experiments and 2 categories). Finally, the third step comprised a total of 16 images per category and number of Epochs.

In the first step, I obtained the scoremaps for each sequence, in each subsampling, and for each hyperparameter configuration. In the second step, I computed the total hot zones per category in each subsampling and for each hyperparameter configuration. In the third step, I determined the total scoremap image per category for each hyperparameter configuration, encompassing all ten subsamplings. The outcome of this third step signifies the significant hot zones across the collection of subsamplings by category.

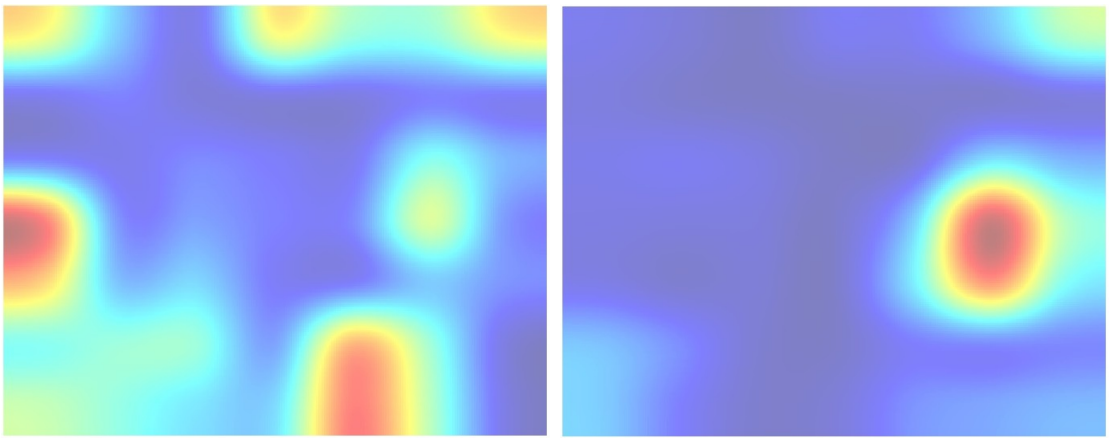
To compute the images generated in Steps Two and Three, I employed two distinct techniques, selected based on their input data. For the determination of the Total Hot Zones in Step Two, I analyzed the scoremaps (numerical matrices corresponding to each sequence) and summed the numerical values at each position within the matrix.

$$H(STEP2) = \sum_{i=1}^{n} H(STEP1)_i \tag{11.2}$$

The scalar summation of Grad-CAM (Class Activation Maps) and the generation of an average heatmap represent two distinct methods for summarizing and visualizing the significance of regions of interest in images. Each approach offers unique advantages.

The arithmetic summation (Step Two) allowed, in a straightforward and computationally efficient manner, to calculate the total hot zone for each of the 80 experiments conducted, without compromising image processing accuracy. By subsequently applying the same color map to this resulting numerical matrix, I obtained the total hot zone image per category for each subsampling and experiment within the test bench.

This way, I obtained a clear and illustrative heatmap highlighting the hot zones within the dataset. As an example, the total hot zones per category in Subsampling 05 for 200 Epochs are depicted in Figure 11.5. Figure 11.5a displays the total hot zones for non-recombinants, while Figure 11.5b illustrates those for recombinants.



(a) Non-Recombinant **(b)** Recombinant

Figure 11.5: Total Hot Zones (SUB_05, 200 Epochs). Non-recombinant hot zones exhibit greater variability and diffusion (Figure 11.5a). Conversely, in the case of recombinant sequences, a distinct, well-defined hot zone emerges, concentrated in the same area (Figure 11.5b).

In Step Three, I generated composite hot zone images for each hyperparameter configuration across the 10 subsamplings. These images represent the arithmetic mean of the hot zones obtained from each subsampling.

$$H(STEP3) = \frac{1}{n}\sum_{i=1}^{n} H(STEP2)_i \tag{11.3}$$

Where n represents the number of subsamplings, in this instance, 10.

Step Three visually illustrates the prevalent and significant hotspots for the recombinant feature across the 10 subsamplings. Given that I had only 10 images per category (non-recombinant and recombinant) in the initial data, I applied image processing techniques to compute a weighted average for each pixel in the image set. This resulted in a consolidated overall image depicting the hotspots across the 10 subsamplings. This method enabled me to reduce the impact of noisy or outlier regions in the individual maps, resulting in a generalized overview of the significant areas in each category.

The utilization of both image processing techniques allowed me to achieve a more comprehensive view of the total hot zones throughout the entire process.

After obtaining the total hot zone in Step Three, I adjusted the color scale to visually accentuate the hot zones. This involved normalizing the average matrix, setting the lowest value to 0 and the highest value to 255, as follows [2856]:

$$\text{Norm. Avg. Matrix} = \left(\frac{\text{Avg. Matrix} - \text{Avg. Matrix Min.}}{\text{Avg. Matrix Max.} - \text{Avg. Matrix Min.}}\right) \times 255 \tag{11.4}$$

Where:

- Avg. Matrix represents the resulting average matrix from the Step Three.
- Avg. Matrix Min. represents the minimum value contained in Avg. Matrix.
- Avg. Matrix Max. represents the maximum value contained in Avg. Matrix.

Finally, in a last adjustment, I generated the negative of the resulting matrix, ensuring that the minimum value appears as a light color and the maximum as a dark color. This was carried out with the intention of enhancing visualization and emphasizing the location of the total hot zones per epoch [2857].

All of these processes were conducted using MATLAB 2021b along with Python library functions, utilizing cv2 for image processing and numpy for multidimensional array manipulation and algebraic operations [2858].

11.4 Results

11.4.1 Results per Subsampling

Appendix F contains the comprehensive results per subsampling regarding performance metrics. These metrics were elaborated upon in Section 11.3.6. Configurations (designated by the number of epochs) that produced optimal outcomes, signified by the highest Test Accuracy values and a more balanced distribution of hit rates between the two categories, were highlighted in green.

I assessed the balance between the two categories by calculating the Standard Deviation (SD) between the Test Accuracy values for recombinants and non-recombinants.

Consequently, I identified optimal configurations as those exhibiting the highest hit rate in the test set and the most balanced distribution, indicated by a lower SD between the Test Accuracies of both categories.

Figure 11.6 visually depicts the confusion matrix scheme delineated in Table 11.4.

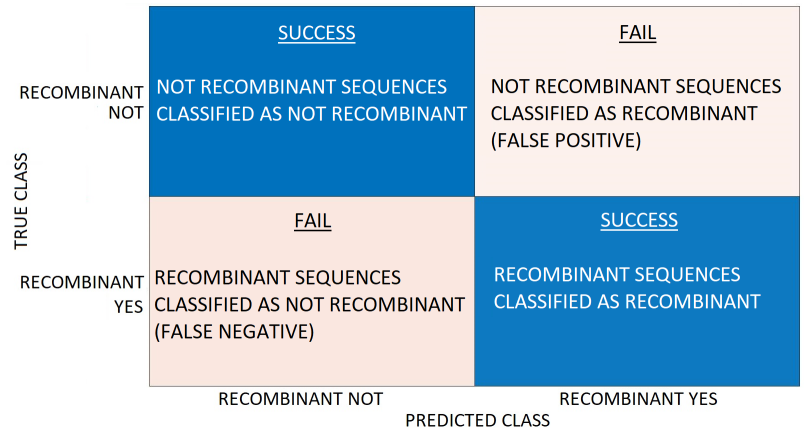


Figure 11.6: Confusion Matrix Scheme. The top row corresponds to the Non-recombinant Category, and the bottom row to the Recombinant Category.

Table 11.4 summarizes the optimal configurations identified for each of the 10 generated subsamplings. These configurations are characterized by the highest Test Accuracy values and a more balanced distribution between both categories. In 80% of the subsamplings, the optimal configurations involved 200 Epochs.

Table 11.4: Best results per subsampling

SUBSAMPLING 01				
Epochs	**AUC**	**Validation Accuracy**	**Test Accuracy**	**Confusion Matrix**
150	0.9818	94.25%	96.01%	302 11 14 299
200	0.9823	94.57%	96.01%	302 11 14 299
SUBSAMPLING 02				
Epochs	**AUC**	**Validation Accuracy**	**Test Accuracy**	**Confusion Matrix**
150	0.9806	94.57%	95.85%	303 10 16 297
250	0.9796	95.05%	95.53%	298 15 13 300
SUBSAMPLING 03				
Epochs	**AUC**	**Validation Accuracy**	**Test Accuracy**	**Confusion Matrix**
200	0.9824	95.21%	94.73%	297 16 17 296
SUBSAMPLING 04				
Epochs	**AUC**	**Validation Accuracy**	**Test Accuracy**	**Confusion Matrix**
200	0.9808	96.17%	95.53%	304 9 19 294
SUBSAMPLING 05				
Epochs	**AUC**	**Validation Accuracy**	**Test Accuracy**	**Confusion Matrix**
200	0.9698	95.53%	93.13%	294 19 24 289
250	0.9732	95.21%	93.13%	288 25 18 295
SUBSAMPLING 06				
Epochs	**AUC**	**Validation Accuracy**	**Test Accuracy**	**Confusion Matrix**
200	0.9782	94.57%	94.57%	297 16 18 295
SUBSAMPLING 07				

Epochs	AUC	Validation Accuracy	Test Accuracy	Confusion Matrix
150	0.9824	94.73%	95.37%	309 4 25 288
200	0.9806	94.73%	95.21%	305 8 22 291
SUBSAMPLING 08				
Epochs	**AUC**	**Validation Accuracy**	**Test Accuracy**	**Confusion Matrix**
150	0.9769	91.05%	91.69%	275 38 14 299
200	0.9754	91.69%	91.69%	275 38 14 299
SUBSAMPLING 09				
Epochs	**AUC**	**Validation Accuracy**	**Test Accuracy**	**Confusion Matrix**
100	0.9842	94.89%	96.17%	305 8 16 297
SUBSAMPLING 10				
Epochs	**AUC**	**Validation Accuracy**	**Test Accuracy**	**Confusion Matrix**
200	0.9728	93.45%	95.05%	302 11 20 293
250	0.9729	93.77%	95.05%	302 11 20 293

11.4.2 Results per Number of Epochs

Complete results are provided in Appendix G. Table 11.5 summarizes the most relevant data.

The configurations associated with 10 and 25 Epochs produced unsatisfactory results in terms of Test Accuracy for both categories, and the outcomes were unbalanced. Consequently, these configurations are deemed inappropriate due to insufficient training with such low values of the number of epochs. The qualitative advantage of 200 Epochs over 150 or 250 lies in its superior balance between the Test Accuracy of both categories. Therefore, although all three configurations demonstrate high hit rates on the test set, 200 Epochs attained the minimum value of Inter-Category SD.

Table 11.5: Best Results per Number of Epochs. I present the mean value and standard deviation of AUC, along with the mean Test Accuracy for both non-recombinant and recombinant sequences, and the standard deviation between the mean Test Accuracy values of recombinants and non-recombinants. The total number of sequences in each category within the Test Set is 313, serving as a measure of the CNN's accuracy.

	AUC		Non-rec. Test Acc.	Rec. Test Acc.	Inter Category
EPOCHS	**Avg.**	**SD**	**Avg.**	**Avg.**	**SD**
10	0.9013	0.0201	249.2	263.8	10.32
25	0.9535	0.0112	289.6	253.9	25.24
50	0.9696	0.0096	291.7	285.2	4.60
75	0.9729	0.0053	295.8	286.6	6.51
100	0.9761	0.0070	295.5	290.1	3.82
150	0.9787	0.0043	300.0	292.2	5.52
200	0.9785	0.0046	297.6	294.6	2.12
250	0.9785	0.0038	298.6	292.5	4.31

Table 11.6 presents the summary of the total hot zones per category for each subsampling by number of epochs (Step Two). I excluded the configurations of 10 and 25 Epochs because their performance metrics were inadequate, potentially due to insufficient training. From the images shown in Table 11.6, I generated the associated images for Step Three, namely, the weighted average hot zones for each configuration.

Table 11.7 showcases the weighted average hot zones for each configuration (Step Three) and their enhanced counterparts. Regarding non-recombinant category, its main hot zones are more diffuse, as opposed to the greater sharpness observed in recombinants. For the latter, all configurations clearly converge towards a single area.

11.5 Discussion

I analyzed 17,215 complete sequences of SARS-CoV-2, encompassing nearly all accessible complete recombinant sequences at the outset of the experimentation. Although I acknowledge that processing 10 subsamplings of the total non-recombinant pool involved handling only approximately 1% of the available non-recombinant sequences, the obtained results remain significant. This is particularly evident in the recombinant category, suggesting that the sample is representative. My findings affirm this assertion.

Table 11.6: Summary Table of Total Hot Zones Step Two. "N" symbolizes the non-recombinant category, and "R" symbolizes the recombinant one.

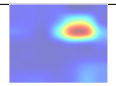
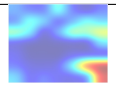
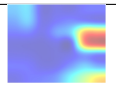
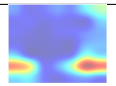
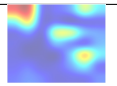
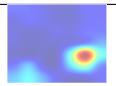
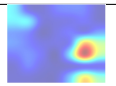
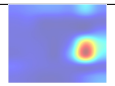
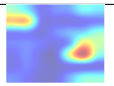
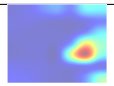
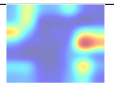
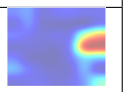
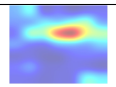
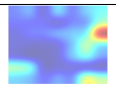
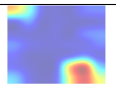
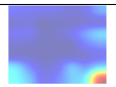
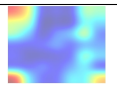
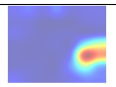
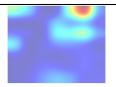
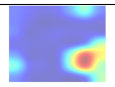
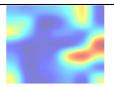
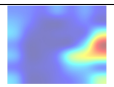
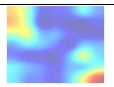
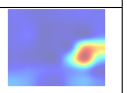
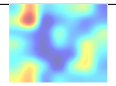
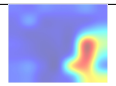
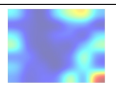
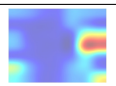
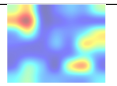
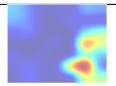
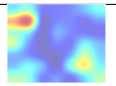
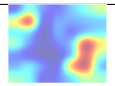
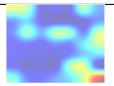
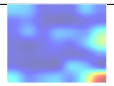
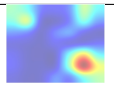
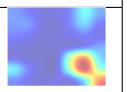
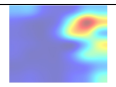
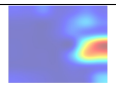
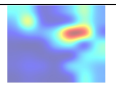
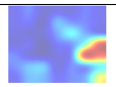
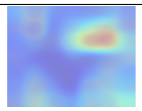
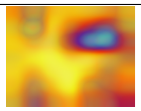
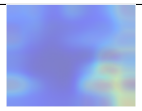
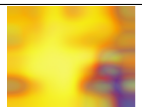
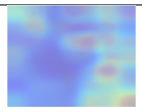
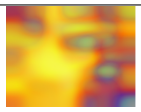
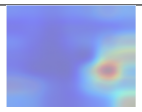
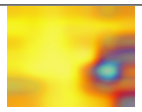
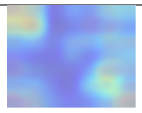
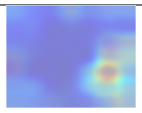
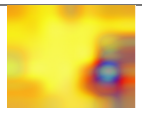
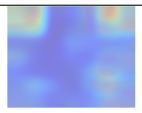
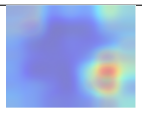
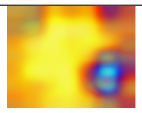
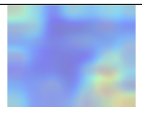
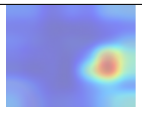
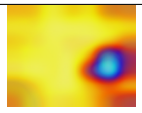
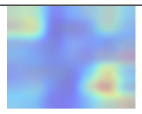
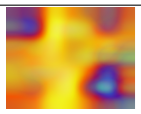
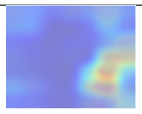
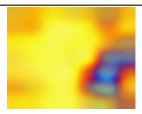
	50 EPOCHS		75 EPOCHS		100 EPOCHS		150 EPOCHS		200 EPOCHS		250 EPOCHS	
	N	**R**	**N**	**R**	**N**	**R**	**N**	**R**	**N**	**R**	**N**	**R**
SUB_01												
SUB_02												
SUB_03												
SUB_04												
SUB_05												
SUB_06												
SUB_07												
SUB_08												
SUB_09												
SUB_10												

Table 11.7: Total Hot Zones per Configuration. The average hot zones represent the hot zones for each number of epochs across the 10 subsamplings. The enhanced figures are the average hot zones with color scale modifications to clarify the relevant hot zones in each category.

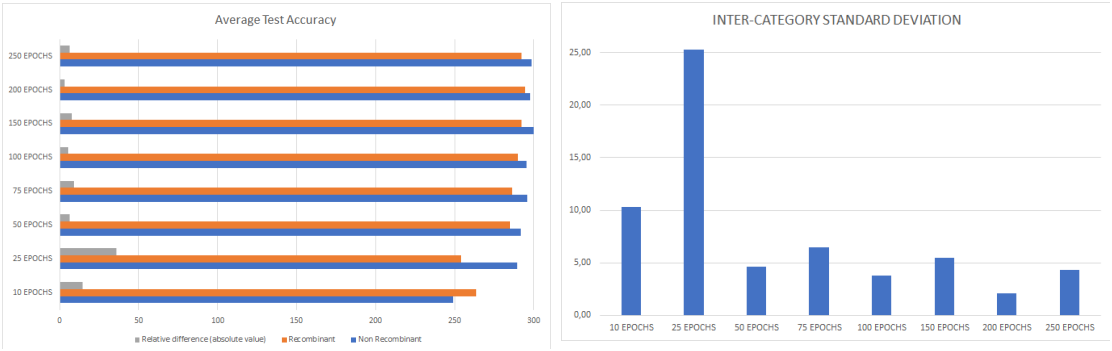
Total Hot Zones per Configuration (Average and Enhanced)				
	Non Recombinant		Recombinant	
EPOCHS	Average	Enhanced	Average	Enhanced
50				
75				
100				
150				
200				
250				

11.5.1 Optimal Configuration Selection

As I outlined in earlier sections, my decision criteria could not rely solely on mere measure of total accuracy within a single category or across both categories combined. I prioritized balanced outcomes, hence the inclusion of inter-category SD in my decision criteria. In this scenario, as depicted in Figure 11.7, the configuration with 200 Epochs attained notable accuracy rates in both categories (Figure 11.7a), with minimal relative differences in absolute terms. Indeed, the inter-category SD was the lowest (Figure 11.7b).

11.5.2 Reference Sequence

To pinpoint the genomic regions housing the hot zones, I referenced the complete genome of the Severe Acute Respiratory Syndrome Coronavirus 2 isolate Wuhan-Hu-



(a) Average Test Accuracy **(b)** Inter-category Standard Deviation

Figure 11.7: Statistical Graphics. Figure 11.7a illustrates the Average Test Accuracy for recombinants (in orange) and non-recombinants (in blue). Additionally, the relative difference between these mean absolute values is depicted in gray. Figure 11.7b showcases the Standard Deviation (SD) values between the average Test Accuracy values for recombinants and non-recombinants.

1, with the NCBI Reference Sequence: NC_045512.2 [2590]. Table 11.8 delineates the location of each structural, non-structural, and accessory protein.

Table 11.8: Location of protein coding regions in the SARS-CoV2 Wuhan-Hu-1 reference genome sequence (NC_045512.2) [2590]. The "Beginning" column specifies the first nucleotide for the corresponding protein, while the "End" column indicates the last nucleotide.

Protein coding region in NC_045512.2 genome sequence		
	Beginning	**End**
5'UTR	1	265
ORF1ab	266	21555
S	21563	25384
ORF3a	25393	26220
E	26245	26472
M	26523	27191
ORF6	27202	27387
ORF7a	27394	27759
ORF7b	27756	27887
ORF8	27894	28259
N	28274	29533
ORF10	29558	29674
3'UTR	29675	29903

Utilizing the consensus reference sequence, I developed a scaled graphical representation of the SARS-CoV-2 genome. This representation will act as a crucial

tool for accurately identifying the regions associated with the recombinant feature.

11.5.3 Analysis of Non-Recombinant Results

After determining that the optimal configuration corresponded to 200 Epochs, the subsequent task involved identifying the influential hot zones crucial for categorizing a sequence as non-recombinant. To achieve this, I computed the comprehensive average image (Step 3), the refined image, and pinpointed the epicenters of the primary hot zones along the x-axis (indicating genomic region involvement) and the y-axis (identifying frequency range), adhering to the directives outlined in Section 11.3.7.

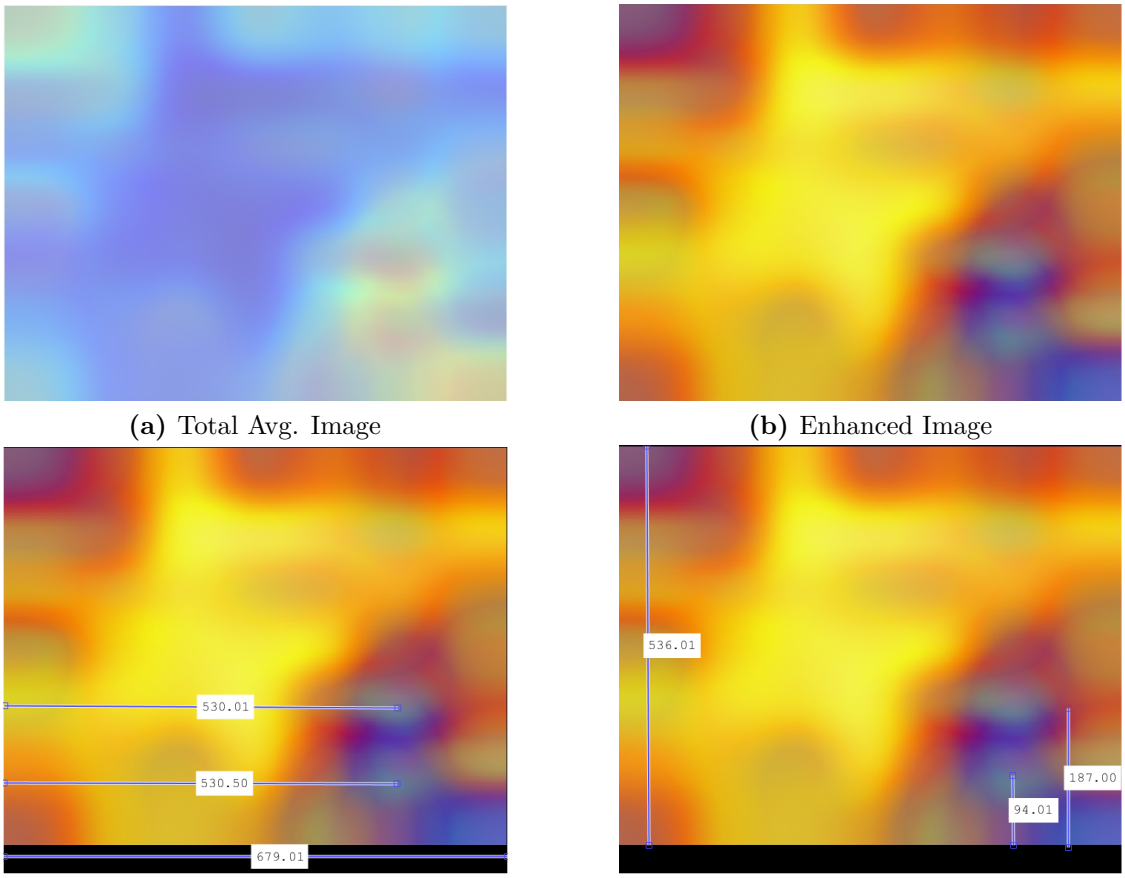
Figure 11.8 illustrates the graphical analysis of the Total Hot Zones (Step 3) in non-recombinants for 200 Epochs.

As depicted in Figure 11.8a, non-recombinant sequences lacked a discernible hot zone, and the hot zone boundaries are somewhat indistinct. This could be attributed to the greater diversity of sub-lineages and strains among non-recombinant variants, leading to increased variability due to subsampling. As highlighted more distinctly in Figure 11.8b, subtle hotspots were suggested around the Spike protein region. Another potential hotspot might exist towards the end of the genome, although its significance seemed less pronounced.

By extrapolating directly from the calculations depicted in Figure 11.8c, the epicenters of the primary hot zones lay within nucleotide positions ranging from 23,341 to 23,362. The central position of the Spike protein (S) aligns with nucleotide 23,473. Consequently, we can position the epicenters of both zones within the central region of the S protein.

By directly extrapolating from the calculations shown in Figure 11.8d, focusing on the vertical axis, which spans from 0 to 0.5 Hz, the critical frequencies fall approximately at $f = 1/12$ and $f = 1/6$ respectively.

Comparing the main hot zones while considering a representative scale of the SARS-CoV-2 genome, it became evident that the two primary hotspots were located in the vicinity of the Spike protein (Figure 11.9).



(a) Total Avg. Image

(b) Enhanced Image

(c) Horizontal Positioning

(d) Vertical Positioning

Figure 11.8: Identification of Total Hot Zones (Step 3) in Non-recombinants for 200 Epochs. Figure 11.8a represents the total hot zones resulting from the 10 subsamplings for 200 Epochs. Figure 11.8b showcases the image post color scale enhancement to facilitate hot zone identification. Figures 11.8c and 11.8d indicate the horizontal and vertical positioning, respectively, of the epicenters of the main hot zones.

Hence, despite the subtlety of these regions, it is notable that the primary decision areas were situated around the Spike protein for the frequencies $f = 1/12$ and $f = 1/6$.

11.5.4 Analysis of Recombinant Results

For recombinant SARS-CoV-2 sequences, the primary hot zone, where the CNN focuses on detecting the recombinant feature, is distinctly outlined. Figure 11.10 presents the graphical analysis of the Total Hot Zones (Step 3) in Recombinants for 200 Epochs.

The distinct sharpness observed in both Figures 11.10a and 11.10b indicates

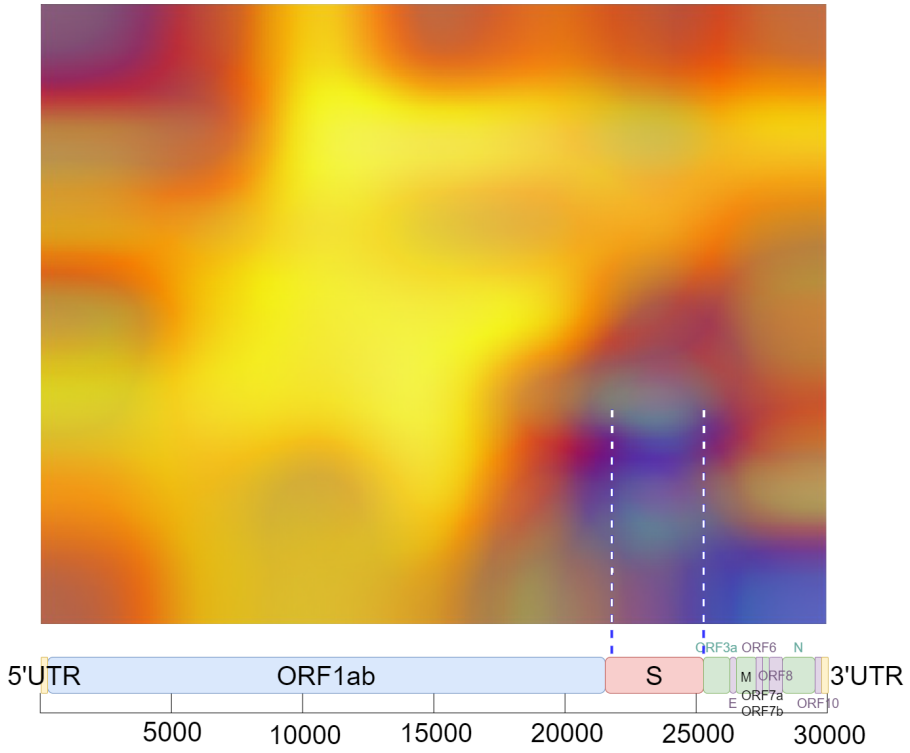


Figure 11.9: Localization of Hot Zones in Non-recombinant Sequences. Positioned in relation to a true-to-scale schematic representation of the composition of the SARS-CoV-2 genome. All significant proteins are duly labeled.

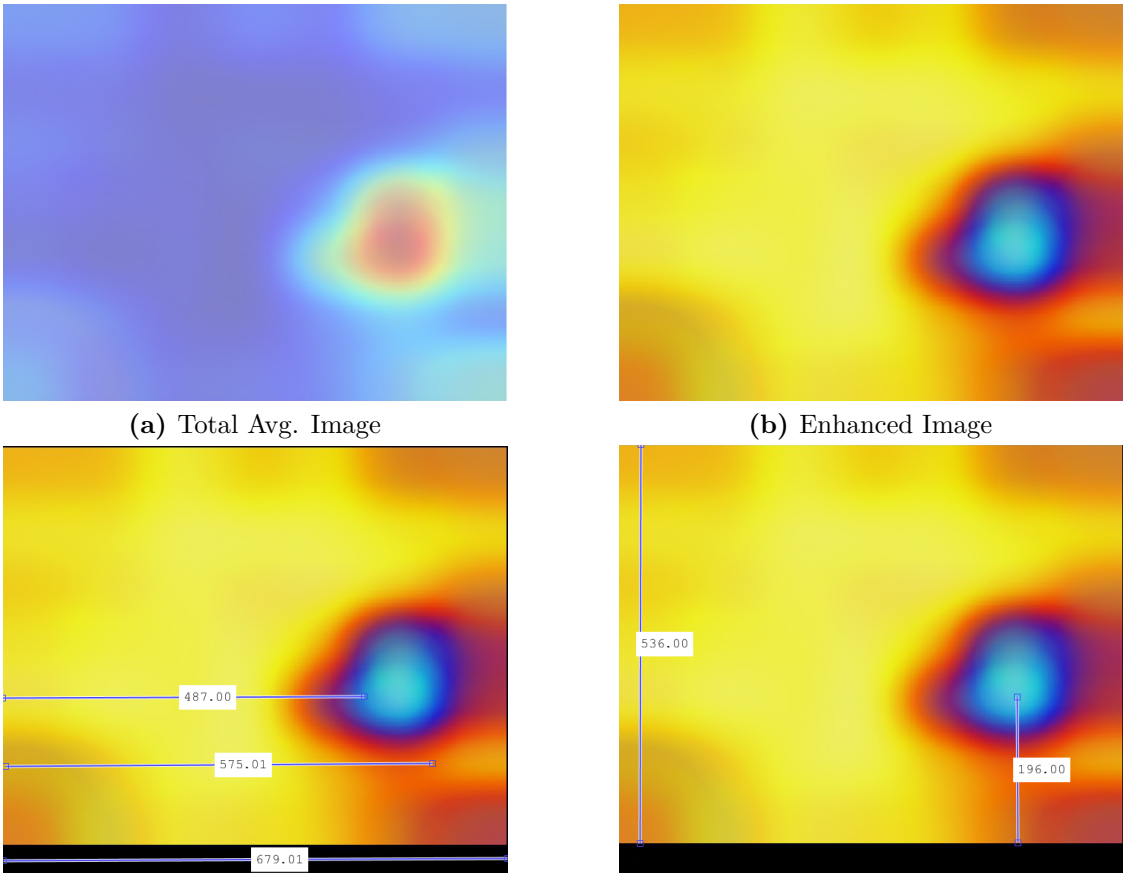
the consistent presence of this hot spot across all subsamplings for 200 epochs.

Upon extrapolating the boundaries to the length of the SARS-CoV-2 genome (29,903 nts), the hot zone crucial for identifying a SARS-CoV-2 sequence as recombinant spans approximately from positions 21,448 to 25,323 (Figure 11.10c). This region closely aligns with the Spike protein (ranging from approximately nucleotide 21,550 to around 25,400).

Regarding the vertical axis, whose total range is 0.5 Hz, as interpreted from the result in Figure 11.10d, the preliminary identification of the epicenter of these total hot zones is located around 0.183 Hz, that is, around $f = 1/6$.

Acknowledging the potential imprecision of Grad-CAM's explainability, alongside the inherent variability in protein locations across variants, the alignment of the hot zone closely enough suggests that the neural network predominantly concentrates on the S protein to discern the recombinant feature within the sequence (Figure 11.11).

Through my approach, I identified that the predominant hot zone resides within



(a) Total Avg. Image **(b)** Enhanced Image

(c) Horizontal Positioning **(d)** Vertical Positioning

Figure 11.10: Identification of Total Hot Zones (Step 3) in Recombinants for 200 Epochs. Figure 11.10a shows the total hot zones resulting from 10 subsamplings for 200 Epochs. Figure 11.10b is the image resulting from enhancing the color scale for better hot zone identification. Figures 11.10c and 11.10d point to the horizontal and vertical positioning, respectively, of the epicenters of the main hot zones.

the S protein. This method enabled me to precisely identify the regions within the genomic spectrogram image where the pre-trained CNN focuses to categorize a sequence as either recombinant or non-recombinant.

Within coronaviruses, the S protein has been a focal point for numerous genetic recombination events [2859]. This phenomenon was also detected in SARS-CoV-2 [30, 2860].

At this juncture, it is crucial to differentiate between my research findings and the occurrence of genetic recombination within the S protein. In my investigation of HIV-1 [2850], I identified a distinct mathematical signature within the genome that influenced the pre-trained CNN's classification of sequences as either recombinant

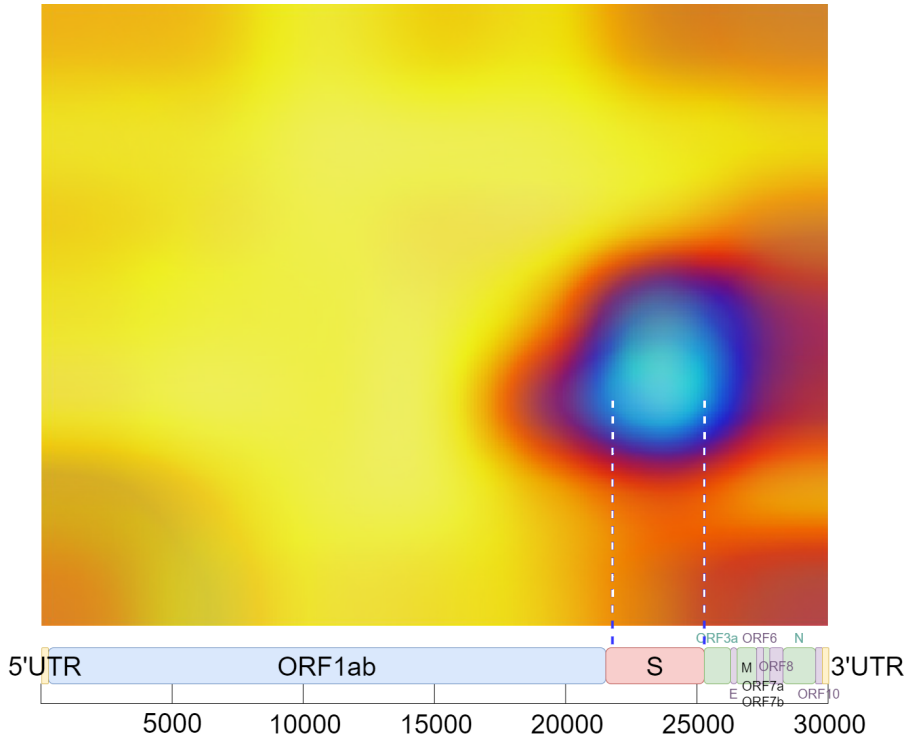


Figure 11.11: Localization of Hot Zones in Recombinant Sequences. Similarly to non-recombinant sequences, the main hot zones (Step 3) are positioned in relation to a scaled schematic representation of the SARS-CoV-2 genome.

or non-recombinant. This signature was predominantly situated in regions near the LTRs, occurring at a frequency of $f = 1/3$. Notably, this signature didn't necessarily align with the specific genomic regions where genetic recombination typically transpires among the various pure subtypes of HIV-1.

For SARS-CoV-2, my analysis unveiled that the mathematical signature aligns closely with the regions where multiple genetic recombination events have been observed, particularly within the S protein.

Future investigations should explore the correlation between the detected mathematical pattern, the mathematical signature identified by the pre-trained CNN, and the high incidence of genetic recombination events in the S protein of SARS-CoV-2. This exploration aims to unveil any potential relationship between these factors and decipher the biological significance of this phenomenon.

Given that the y-axis range spans from 0 to 0.5 Hz, the frequency of 0.18 Hz corresponds to approximately $f = 1/6$. Due to the inherent limitations of Grad-

CAM and the cumulative errors introduced during the computational processes involved in determining the total hot zones, it is reasonable to approximate the vertical epicenter to frequencies close to $f = 1/6$. For a more accurate assessment of the epicenter frequency, I conducted measurements on one of the experiments where this region is most prominently depicted: specifically, 200 Epochs in Subsampling 06.

Figure 11.12 illustrates the measurement taken precisely at the epicenter of the hot zone. This figure underwent minimal mathematical transformations. Indeed, the vertical epicenter of the hot zone aligns closely with frequencies around $f = 1/6$, suggesting that this frequency may indeed play a significant role in determining the recombinant feature.

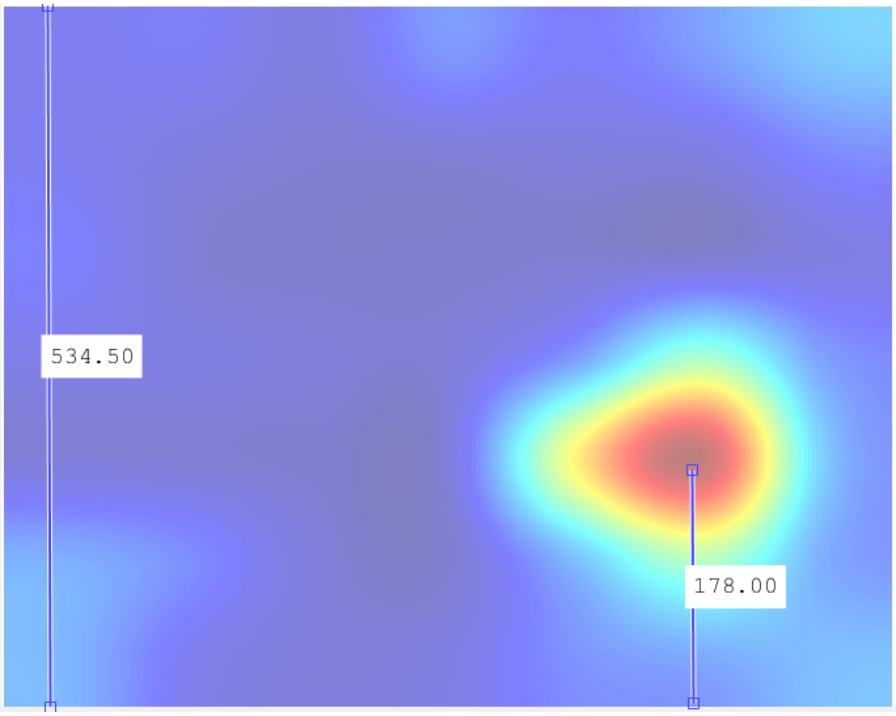


Figure 11.12: Vertical Positioning at 200 Epochs Configuration Subsampling 06. Indeed, considering the total length of the y-axis is 534.5 points, spanning from 0 to 0.5 Hz, the epicenter's position at 178 points precisely corresponds to 0.167 Hz. This value closely aligns with $f = 1/6$ in the most pronounced case observed across the entire test bench.

11.6 Conclusions

Employing genomic spectrograms with 10 random subsamplings to mitigate the size discrepancy between non-recombinants and recombinants, I devised a test

bench to uncover the optimal hyperparameter configurations. I utilized Transfer Learning in two phases, leveraging a pre-trained VGG-16 model trained on the ImageNet dataset. Phase 1 focused on HIV-1 genomic spectrograms, while Phase 2 targeted those of SARS-CoV-2. My objective was to identify the recombinant feature within a SARS-CoV-2 genomic sequence. Following this, I employed the Grad-CAM explainability tool in three steps to pinpoint the hot zones (areas where the CNN focuses for classification) within each sequence, within each subsample for every configuration, and collectively within each configuration. I utilized image processing techniques to refine the localization of these hot zones. These three steps entail not just the straightforward application of Grad-CAM, but also the mathematical analysis of its results to extrapolate the obtained outcomes. The image processing techniques employed enabled me to clearly delineate the pertinent regions for the recombinant feature.

Consistent and clearly delineated outcomes were achieved in each category. Notably, for SARS-CoV-2, the Spike protein emerges as a critical genomic region in both recombinant and non-recombinant categories.

The evident significance of the S protein in identifying the recombinant feature in SARS-CoV-2 aligns with the comprehensive research conducted by Nikolaidis et al. [2859]. They revealed numerous instances of double-crossover genetic recombination events across different coronaviruses, with a notable concentration within this particular protein. Thus, my findings complement and reinforce their conclusions through a distinct analytical approach.

For non-recombinant sequences, the hot zones (Step 3) exhibit a more scattered distribution, yet they seem to center around the region of the spike protein, particularly within the frequency range of $f = 1/12$ and $f = 1/6$.

However, the distinctness of the primary hot zones in Steps 2 and 3 is notably evident in the case of recombinant sequences. A specific area corresponding to the Spike protein is clearly highlighted, typically around an approximate frequency of $f = 1/6$.

Based on the results obtained, I identified two aspects that required further investigation. The first of these was related to the dissipation of hot zones in

non-recombinant sequences and the exploration of whether there is any relationship between the genetic variability associated with different non-recombinant variants and the lower specificity in the determined hot zones in the case of non-recombinant sequences. To this end, I chose to prospect the possible existence of patterns in the hot zones that identify the variant to which each sequence belongs. The results of this experimentation were detailed in the Chapter 12.

The second aspect under investigation is related to the recombinant category. The clear hot zone in the case of recombinant sequences led me to delve into identifying the mathematical signature associated with the recombinant feature in SARS-CoV-2. To this end, I further explored the search for mathematical patterns associated with this hot zone. The results of this exploration are detailed in Chapter 13, which outlines the hypothesis resulting from an analysis of the findings with a mathematical and microbiological perspective.

In conclusion, in this Chapter, I applied Deep Learning tools with high power in detecting patterns not visible to the human eye [2861, 2862]. I successfully pinpointed the critical regions responsible for the recombinant feature within genomic spectrograms of SARS-CoV-2. My approach allowed me to attain impressive levels of Test Accuracy while clearly delineating robust hot zones in both the recombinant and non-recombinant categories. This achievement underscores the potential of Deep Learning techniques in uncovering intricate patterns within complex biological data.

At this point, we know where the CNN looks to classify a SARS-CoV-2 sequence as recombinant. Now, I want to understand what it sees. What is the mathematical pattern embedded in the frequency spectrum of the genome of the Spike protein that causes a sequence to be classified as recombinant?

You know my method. It is founded upon the observation of trifles.

— Edgar Allan Poe (Sherlock Holmes), *he loved cats. He always had one nearby, and when writing, his inspiration and creativity increased when the cat nestled on his shoulder. In 1848, he wrote an essay titled "Eureka: A Prose Poem", in which he presented something very similar to what would later be termed the Big Bang theory.*

12 Analysis of SARS-CoV-2 Non-recombinant Variants

Contents

12.1 Introduction 625
12.2 Equipment 627
12.3 Methods 628
12.3.1 Spectrogram Generation 628
12.3.2 Dataset of Non-recombinant SARSCoV-2 Variants . . . 628
12.3.3 Two-Stage Transfer Learning 631
12.3.4 Best Performance Based Decision Criteria 633
12.3.5 Explainability Tools' Result Generation 633
12.4 Results and Discussion 635
12.4.1 Training Curves 635
12.4.2 Performance Measurement Ratios 639
12.4.3 Total Hot Zones per Variant 642
12.5 Conclusions 651

12.1 Introduction

In Chapter 11, I identified a clear hot zone that was crucial in classifying SARS-CoV-2 sequences as recombinant. This determinant hot zone in the recombinant feature was clearly located at $f = 1/6$ in the S protein.

However, the hot zones corresponding to the non-recombinant feature turned

out to be less clear. This fact could be related to greater diversity in terms of variants, while the recombinant sequences detected in SARSCoV-2 were mostly recombinants related to Omicron.

During the period of overlap between Delta and Omicron, recombinant variants between both emerged. The most significant of these recombinant versions was a combination of AY.4 (Delta) and BA.1 (Omicron). It was named as the BA.1 x AY.4 recombinant by the World Health Organization (WHO) [2726], first detected in France in January 2022 and then nicknamed as Deltacron" [2863]. Other variants between Omicron and Delta were called XF and XD (Pango designation) [2847]. These recombinant variants, as well as Delta itself, eventually became extinct due to the dominance of Omicron and its derivatives. Logically, the disappearance of Delta resulted in the absence of new recombinants with this variant as an ancestor or parental virus. Leaving the scenario dominated by the entire array of Omicron lineages since approximately September 2022 [2864, 2865].

The remaining recombinant variants detected were precisely derived from genetic recombination events between different sublineages of Omicron, such as XE (BA.1-BA.2). Other variants included, among others, XA, XB, XS, XY, XZ [2866], XBB and all its derivatives, recombinant of two cocirculating BA.2 lineages, BJ.1 and BM.1.1.1 (a progeny of BA.2.75), during the summer of 2022 [2867].

Therefore, the high influence of Omicron on these recombinants implies a closer phylogenetic proximity among all recombinant variants than may exist among the set of different non-recombinant variants that have emerged over time, such as Alpha, Delta, etc.

Indeed, within non-recombinant variants, variability is higher. This is due to multiple factors such as, firstly, a longer temporal space until the absolute dominance of XBB and its derivatives (Omicron recombinant), with the consequent elimination of all variants, not only less imposing lineages of Omicron but of every non-Omicron variant. This period spans from December 2019 until the appearance of the first descendant XBB (called XBB.1) around December 2022 [2868].

Within these non-recombinant variants, there is a high imbalance. Variants such as Beta, Epsilon, Eta, Gamma, Iota, Kappa, Lambda, Mu, Theta, or Zeta exhibit a limited number of sequences compared to other more predominant variants such as the compilation before the appearance of the first Variant Of Concern (referred to as pre-VOC), Alpha, Delta, and Omicron. The evolutionary advantages of these variants with their consequent imposition justify why there are many sequences of these and few of the others. These others turned out to be minority and lost out to others that prevailed.

This pre-VOC compilation encompasses all variants prior to the emergence of Alpha. It is not truly a clearly defined variant but rather a compilation of all lineages originating from the onset of the pandemic in Wuhan (January 2020) and undergoing genetic diversifications until the emergence and dominance of Alpha (detected in the last quarter of 2020 and peaking in April 2021).

I deemed it necessary to delve deeper into the reasons explaining the limited clarity in the hot zones corresponding to the non-recombinant category in Chapter 11. With this purpose in mind, I conducted all the experiments described in this Chapter.

This Chapter is structured into several key sections. It begins with Section 12.2, where the equipment used is specified, followed by a detailed description of the methodology employed (Section 12.3). The subsequent section presents the results and discussion (Section 12.4), analyzing the findings in depth. Finally, this Chapter concludes with a summary of the key conclusions drawn from the study (Section 12.5).

12.2 Equipment

All experiments were run in this equipment:

- Processing Unit: Intel(R) Core(TM) i7-4770K CPU. 3.5 GHz.
- Installed RAM: 32 GB usable.
- Operative System: Windows 10 Education. Version: 22H2.
- GPU: NVIDIA GeForce RTX 3090. Total memory: 40 GB.

12.3 Methods

12.3.1 Spectrogram Generation

Following the line of work outlined in Chapter 11, Section 11.3.3, I utilized the superposed spectrograms of the complete sequences of SARS-CoV-2 belonging to non-recombinant variants.

12.3.2 Dataset of Non-recombinant SARSCoV-2 Variants

The total compilation of SARS-CoV-2 sequences I gathered around March 2023 consists of the number of sequences per variant indicated in Table 11.1. As can be seen, the only variants that accounted for more than 10% of the total were pre-VOC, Alpha, Delta, and Omicron. The remaining variants did not exceed 1.3%, with almost all of them even below 1%.

This compilation shows a high imbalance between variants that have not thrived, such as Beta, Gamma, Epsilon, Eta, Iota, Kappa, Lambda, Mu, Theta, and Zeta; versus the pre-VOC, Alpha, Delta, and Omicron variants. The latter four variants demonstrated a series of evolutionary advantages that have allowed them to prevail over other less advantaged variants (Table 12.1).

Table 12.1: Clustering of Complete Non-Recombinant SARS-CoV-2 Sequences

Variant	**Number of Sequences**	**Percentage**
pre-VOC	218,198	14.17%
Alpha	198,722	12.91%
Delta	325,285	21.13%
Omicron	748,635	48.62%
Minority Variants	48,888	3.18%
TOTAL	**1,539,728**	**100.00%**

The number of Omicron sequences represents practically half of the total dataset, followed by Delta. Variants Alpha and pre-VOC (compendium of all lineages before the appearance of Alpha) contain a similar number of sequences. The total number of minority variants does not exceed 50,000 total sequences. This

minority category includes variants Beta, Gamma, Epsilon, Eta, Iota, Kappa, Lambda, Mu, Theta, and Zeta.

The compilation of minority variants, besides representing a small amount of data (3.18%) compared to the dominant variants, suffers from high variability because it encompasses 10 variants with notable differences.

This type of imbalance in the data can cause difficulties in the application of Deep Learning technologies, affecting the model's performance. The imbalance in the distribution of categories will cause the model to learn biases towards the majority classes, ignoring the minority classes. As a consequence, the following detrimental effects can occur:

- Overfitting towards the Majority Classes. The model may tend to overfit to the majority classes due to the larger amount of data available for them. This can lead to a low ability of the model to generalize correctly on the minority classes, resulting in poor performance on unseen data.

- Decreased Accuracy. In the presence of imbalance, the accuracy obtained can be misleading, as the model may have high accuracy for the majority classes but poor results for the minority classes.

- Sensitivity to Decision Thresholds. Imbalance can make the model sensitive to the decision thresholds used for classification, which can affect the interpretation of prediction probabilities and decision-making based on them.

- Difficulty in Finding Meaningful Patterns. In the presence of minority classes with few data, it can be difficult for the model to find meaningful patterns that distinguish these classes from others, limiting the model's ability to learn discriminative representations.

To eliminate the noise introduced by considering the minority classes, I removed all those unrepresentative variants in the set of complete SARS-CoV-2 sequences analyzed, as indicated in Table 12.2. The exclusion of the set of minority variants entails an adjustment in the weighting of each category. Omicron accounts for half of the total dataset, Delta around 20%, and pre-VOC and Alpha together represent slightly less than 15%.

Table 12.2: Clustering of Complete Non-Recombinant SARS-CoV-2 Sequences After Excluding Minority Variants

Variant	**Number of Sequences**	**Percentage**
pre-VOC	218,198	14.64%
Alpha	198,722	13.33%
Delta	325,285	21.82%
Omicron	748,635	50.22%
TOTAL SUBSAMPLE	**1,490,840**	**100.00%**

It is evident that Omicron is the broadest variant. First identified in Botswana in November 2021 [2869], it has become the dominant circulating variant that has ultimately prevailed. Currently, since approximately September 2022, 100% of the sequencing corresponds to Omicron [2865], so this variant has erased all others from the epidemic stage.

The Omicron sequence set represents more than double that of Delta and about 3.5 times more than pre-VOC and Alpha, respectively. This imbalance between Omicron and the other three variants under study necessitates the implementation of mitigation measures to work with a more balanced dataset. This way, I prevented the over-representation of Omicron from affecting suboptimal performances resulting from biases towards the majority class and achieved a more comprehensive representation of the feature space, leading to better generalization capacity.

To mitigate the imbalance between Omicron and the rest of the categories, I divided the Omicron dataset into two subsamples. I randomly split the total set of Omicron samples into two equally sized groups so that the samples included in VARIANT SUBSAMPLING 01 are not present in VARIANT SUBSAMPLING 02, and vice versa.

So, in each subsampling, I worked with all samples from Alpha, Delta, and pre-VOC; and with a random half of Omicron. Thus, the structure of each of these two subsamplings is as indicated in Table 12.3.

This subdivision in Omicron balanced the weighting of the four categories. Despite a clear predominance of Delta and Omicron over Alpha and pre-VOC, the arrangement of hundreds of thousands of samples in all categories allowed for

Table 12.3: Number of Sequences per Category per Subsampling. Although the distance between Omicron (the most abundant) and Alpha (the least abundant) is practically double, the difference between them has been reduced to almost half.

	Number of Sequences	**Percentage**
pre-VOC	218,198	19.54%
Alpha	198,722	17.80%
Delta	325,285	29.13%
Omicron	374,318	33.53%
TOTAL SUBSAMPLE	**1,116,523**	**100.00%**

sufficient information to be available for proper model learning and to minimize the impact that these variations in the number of samples may have.

Regarding the organization of these sequences for dataset creation, the Training Set comprised 60% of the total sequences, while the Validation Set and the Test Set each represented 20%.

Table 12.4 shows the composition of the dataset in each of the two subsamplings.

Table 12.4: Dataset Composition per Subsampling. For each category, I allocated 60% sequences to the Training Set, 20% to the Validation Set, and 20% to the Test Set.

Variant	**Training Set (60%)**	**Validation Set (20%)**	**Test Set (20%)**	**Total**
pre-VOC	130,919	43,640	43,640	218,198
Alpha	119,233	39,744	39,744	198,722
Delta	195,171	65,057	65,057	325,285
Omicron	224,591	74,864	74,864	374,318
TOTAL SUBSAMPLE	**669,914**	**223,305**	**223,305**	**1,116,523**

In each subsampling, I worked with half of the 748,635 Omicron sequences, which were randomly distributed among both subsamplings.

12.3.3 Two-Stage Transfer Learning

Similar to the approach used in Chapter 11, I conducted a Two-Phase Transfer Learning.

As explained, I used the HIV-1 case study as the first Transfer Learning phase for genetic recombination, given its sufficient, balanced, and thoroughly documented

number of existing sequences. The knowledge gained by the CNN during this initial phase was then applied to the specific case of SARS-CoV-2 in a second phase of Transfer Learning to detect the recombinant feature in this virus.

In the case study presented in this Chapter, I introduced several methodological modifications to this Two-Phase Transfer Learning to address the objectives of this Chapter.

Figure 12.1 illustrates the Transfer Learning methodology applied for the classification of non-recombinant variants.

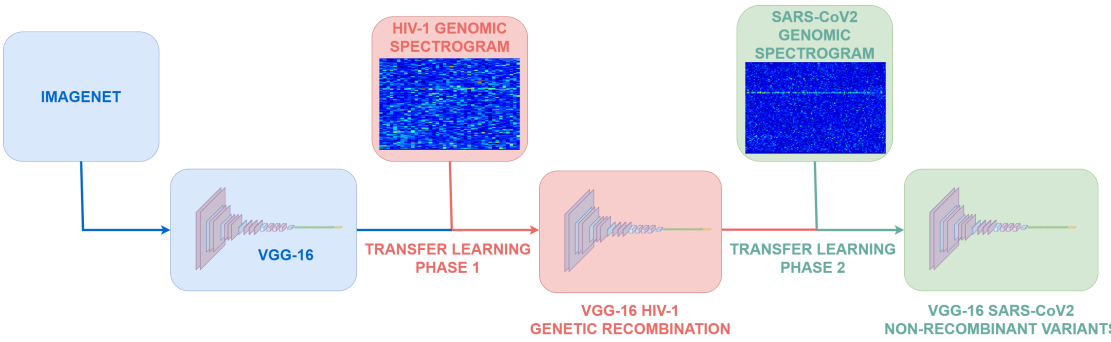


Figure 12.1: Two-Stage Transfer Learning Methodology Non-Recombinant Variant Detection. I started with a pre-trained VGG-16 using the ImageNET dataset. In Phase 1, I applied Transfer Learning to the genomic spectrogram dataset of complete HIV-1 sequences to detect the recombinant feature. In Phase 2, I applied Transfer Learning once again to the resulting network from Step 1 (VGG-16 HIV-1 GENETIC RECOMBINATION) using a genomic spectrogram dataset of complete SARS-CoV-2 sequences to also detect the non-recombinant variant to which each SARS-CoV-2 sequence belongs (VGG-16 SARS-CoV-2 NON-RECOMBINANT VARIANTS).

In the Figure 11.3 scheme, the first phase of Transfer Learning aimed to detect the recombinant feature in complete sequences of HIV-1 while the objective of the second phase was to detect the recombinant feature in complete sequences of SARS-CoV-2.

In this case, as depicted in Figure 12.1, the first phase of Transfer Learning was exactly the same. However, the objective in the second phase was not to detect the recombinant feature in SARS-CoV-2, but to classify complete sequences of non-recombinant SARS-CoV-2 into one of these four variants: pre-VOC, Alpha, Delta, or Omicron.

Therefore, the complexity of this experiment was greater, as I introduced two changes in the second phase of Transfer Learning. Not only did I change the

virus under study (from HIV-1 to SARS-CoV-2), but I also changed the detection target. I transitioned from classifying the genomic spectrograms as recombinant or non-recombinant to classifying the non-recombinant genomic spectrograms into four different variants.

12.3.4 Best Performance Based Decision Criteria

My performance measurement metrics were:

- Validation Accuracy. I employed it as a measure of a model's performance during the validation phase, to adjust the model's hyperparameters and optimize its performance during training.

- Training Time. I employed it as a measure of the computational cost associated with model training.

- Confusion Matrix. I employed a 4x4 Confusion Matrix in this classification problem with four different classes. Each row represented the true class of instances from the dataset, while each column represented the class predicted by the model. In this matrix, the diagonal elements represented true positives (correctly classified instances), while the off-diagonal elements indicated classification errors. For instance, if the row corresponding to Class 1 and the column corresponding to Class 2 have a high value, it means that many instances of Class 1 were misclassified as belonging to Class 2. The matrix provides a detailed insight into the model's performance in classifying each class, aiding in identifying where errors occur and improving the model in future iterations.

- Test Accuracy. This metric served as an assessment of the model's genuine performance, evaluated against previously unseen data. It was computed for each individual category as well as collectively.

12.3.5 Explainability Tools' Result Generation

Similar to Chapter 11, I applied Grad-CAM to identify the critical hot zones for classifying a SARS-CoV-2 sequence into its corresponding non-recombinant variant. Grad-CAM is a highly visual tool, which makes it deductive in its interpretation,

albeit at the cost of precision in detecting the hot zones [231, 2855]. I also applied image processing techniques to obtain the total hot zones per variant.

The methodology for interpreting results was similar to that presented in Section 11.3.7, albeit with modifications tailored to the case study. In this instance, I conducted a two-step processing of the explainability score maps. Figure 12.2 schematically summarizes the explainability processing methodology.

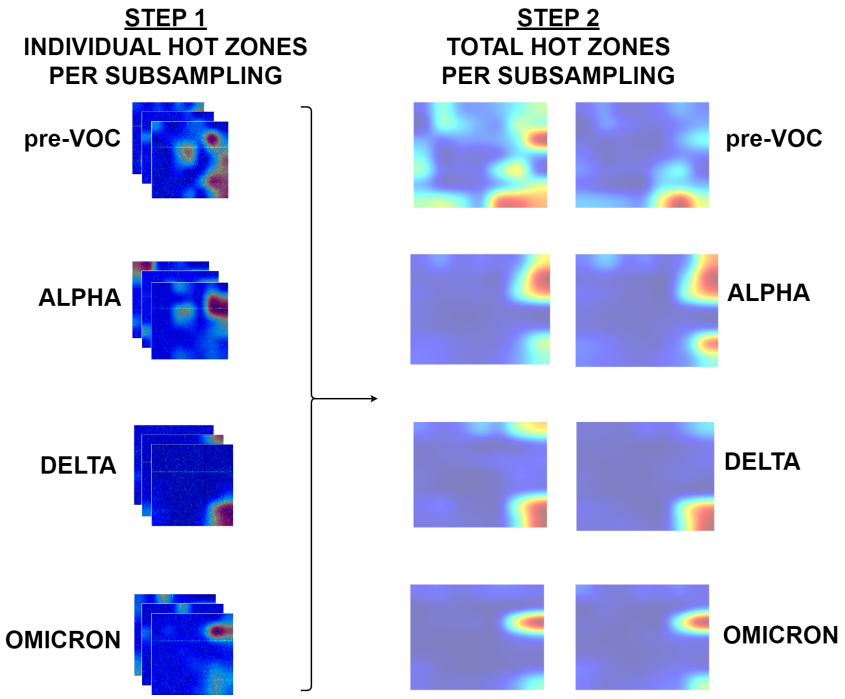


Figure 12.2: Two-Step Explainability. The first step involved a total of 223,305 images (total number of data points in the Test Set). The second step involved eight images in total, four categories in two subsamplings.

Similar to Section 11.3.7, in the first step, I computed the score maps for each of the non-recombinant sequences from the four categories in the Test Set and generated the corresponding Grad-CAM images using a jetmap color scale (Figure 4.3). I performed this process for each of the two subsamplings.

The second step of explainability involved calculating the total score maps for each of the four categories. Each total score map was the summation of the individual score maps per category (matrix summation).

$$H(STEP2) = \sum_{i=1}^{n} H(STEP1)_i \tag{12.1}$$

Once the total score map matrix of numbers was obtained, I generated the Grad-CAM image representing the total hot zones using the jet map color scale.

This processing was carried out for both subsamplings, resulting in a total of eight images corresponding to the four categories in the two generated subsamplings.

12.4 Results and Discussion

12.4.1 Training Curves

The computational requirements associated with processing the 1,116,523 images that composed the dataset in each of the two subsamplings have largely determined the optimal hyperparameters considered. The selected hyperparameters were:

- Learning Rate = 0.0001
- Batch size = 52
- Epochs = 1

By setting the Learning Rate to 0.0100 and/or Batch size to 128, the system demonstrated insufficient capacity to perform the necessary calculations, resulting in training interruptions due to collapse. As for the number of epochs, I set it to the minimum since tests conducted with Epochs = 2 resulted in clear overfitting.

Figures 12.3 and 12.4 display the Training Curves of both subsamplings, denoted as VARIANT SUBSAMPLING 01 and VARIANT SUBSAMPLING 02, respectively. Both curves exhibited the same behavior.

As can be observed, the Training Times were high. In the case of Figure 12.3, it took 1,190 minutes and 10 seconds, practically 20 hours per Epoch. In the case of Figure 12.4, it took 1,626 minutes and 16 seconds, which is 27 hours per Epoch.

The behavior of the training curve was remarkable. In one-third of an Epoch, both curves stabilized, maintaining a Validation Accuracy close to 99.50% in both datasets (Figure 12.3 99.42% and Figure 12.4 99.52%).

Therefore, the graphs were similar in appearance and in the values of Validation Accuracy.

The fact that only one epoch was used for training can be attributed to the large volume of available data. With a Training Set composed of a total of 669,914

sequences and a Validation Set including 223,305 sequences, the abundance of data allowed for a comprehensive representation of the dataset in just one epoch.

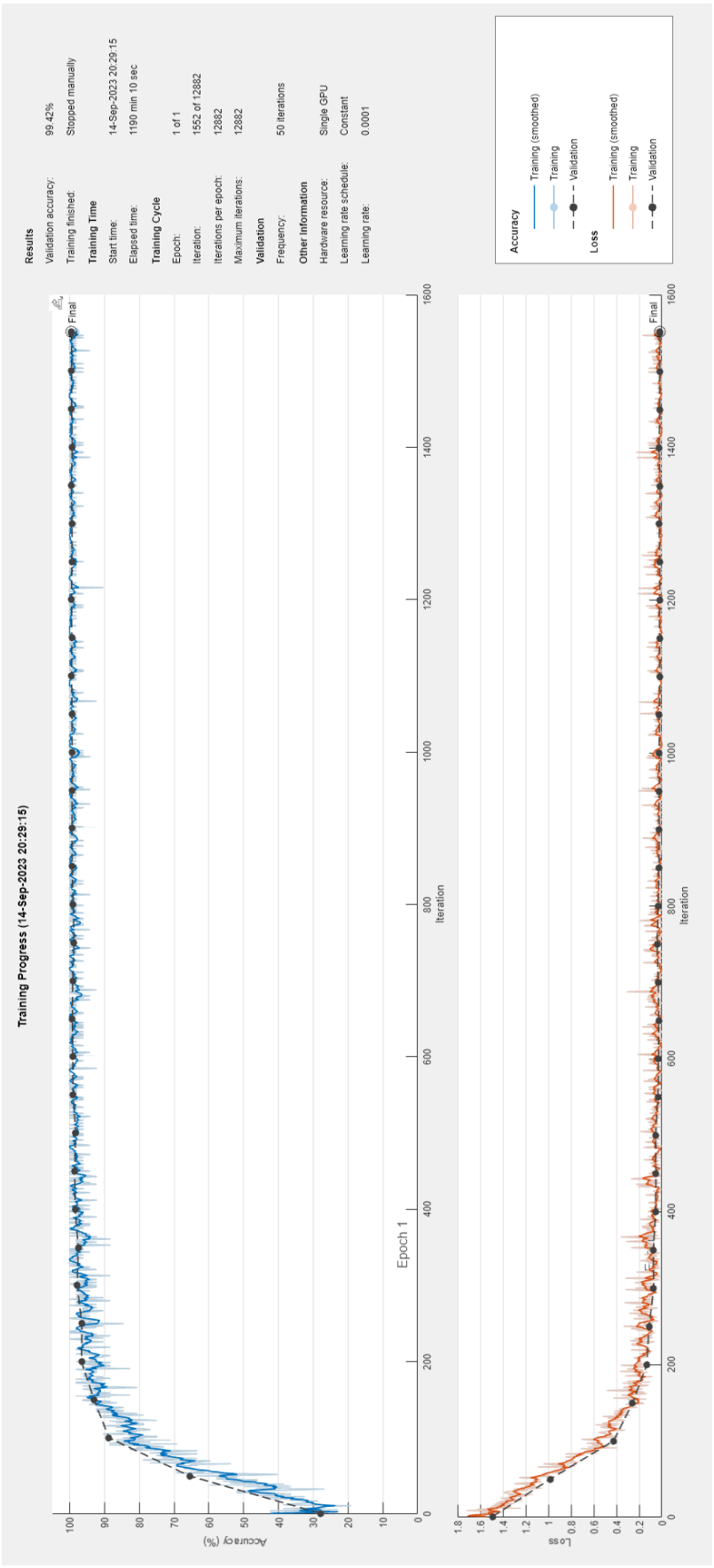


Figure 12.3: VARIANT_SUBSAMPLING_01 Training Curve. The upper graph displays the Accuracy Curves in %, the blue curve corresponding to the Training Set and the black curve to the Validation Set. The lower graph shows the Loss Curves, the orange curve representing the Training Set and the black curve representing the Validation Set.

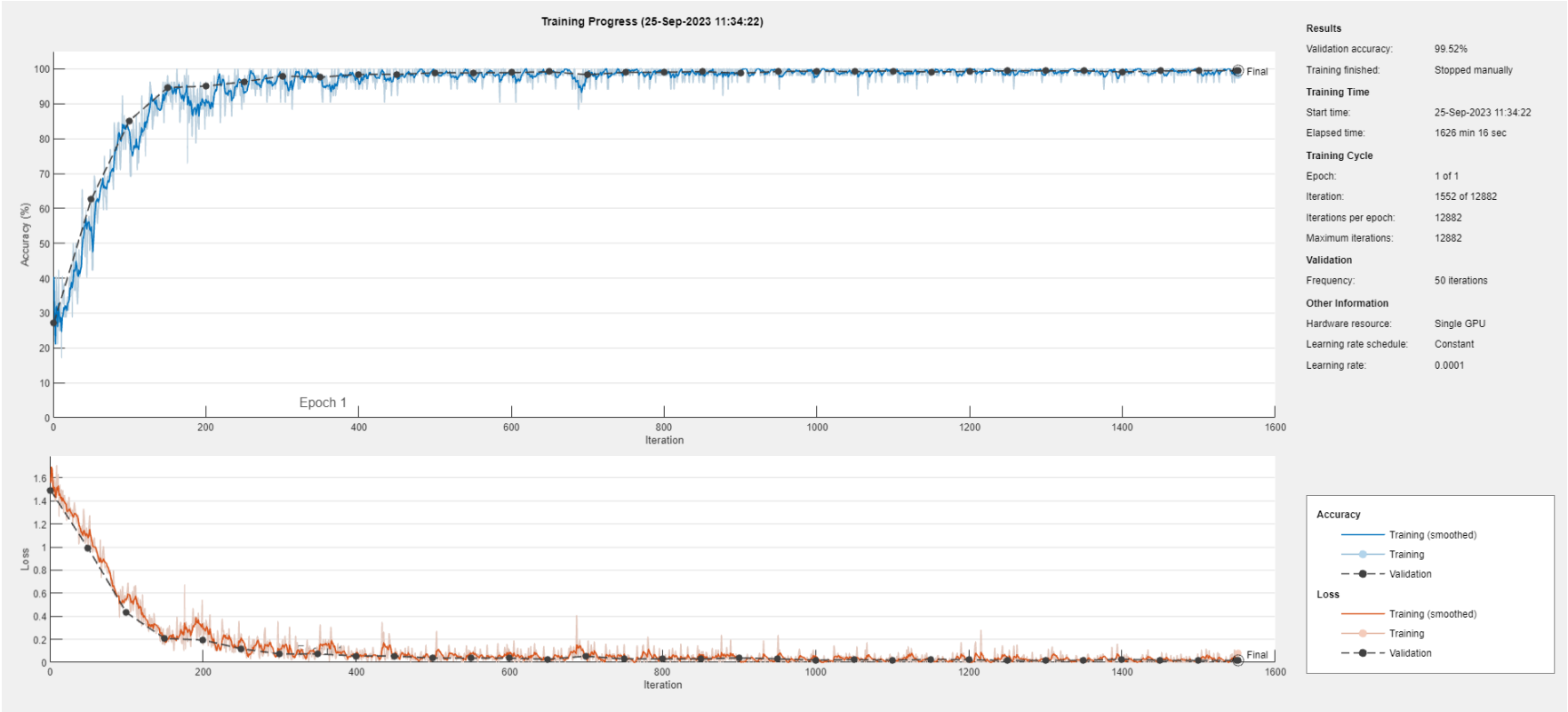


Figure 12.4: VARIANT_SUBSAMPLING_02 Training Curve. The upper graph displays the Accuracy Curves in %, the blue curve corresponding to the Training Set and the black curve to the Validation Set. The lower graph shows the Loss Curves, the orange curve representing the Training Set and the black curve representing the Validation Set.

Validation Loss Curves were also nearly identical and stabilized with a value around 0.80%.

12.4.2 Performance Measurement Ratios

Table 12.5 compiles the results of the performance measurement ratios in the case of the VARIANT SUBSAMPLING 01.

Table 12.5: Performance Measurement Ratios Non-Recombinant VARIANT SUBSAMPLING 01. The column "Description" displays the values of the experimental hyperparameters. "Test Accuracy" for each of the four categories is expressed as percentage of correctly classified sequences *(Correct Predictions per Category / Total Number of Samples in that Category)*, as well as the "Test Accuracy Global" *(Total Correct Predictions / Total Number of Samples)*.

VARIANT SUBSAMPLING 01			
Description	**Training Time**	**Validation Accuracy**	**Test Accuracy**
Learning Rate: 0.0001 Batch Size: 52 Epochs: 1	1,190 min 10 sec	99.42%	pre-VOC: 97.73% Alpha: 99.79% Delta: 99.88% Ómicron: 99.74% **Global: 99.40%**

The time required to conduct the training was 19.84 hours, given the computational load derived from the immense dataset and the limited hardware resources available. Despite the temporal cost of training, the accuracy values obtained in the Validation and Test Sets were excellent and virtually identical. The total test accuracy was only 0.02%, lower than the Validation Accuracy.

The results of the performance measurement ratios in the case of the VARIANT SUBSAMPLING 02 subsampling are available in Table 12.6.

Table 12.6: Performance Measurement Ratios Non-Recombinant VARIANT SUBSAMPLING 02. The column "Description" displays the values of the experimental hyperparameters. "Test Accuracy" for each of the four categories is expressed as percentage of correctly classified sequences *(Correct Predictions per Category / Total Number of Samples in that Category)*, as well as the "Test Accuracy Global" *(Total Correct Predictions / Total Number of Samples)*.

VARIANT SUBSAMPLING 02			
Description	**Training Time**	**Validation Accuracy**	**Test Accuracy**
Learning Rate: 0.0001 Batch Size: 52 Epochs: 1	1,626 min 16 sec	99.52%	pre-VOC: 98.12% Alpha: 99.86% Delta: 99.73% Ómicron: 99.85% **Global: 99.48%**

In the case of VARIANT SUBSAMPLING 02, The training Time reached 27.10 hours. This represented an increase of 36.64% compared to the training time corresponding to VARIANT SUBSAMPLING 01. In any case, both trainings required considerable training times.

Regarding the accuracy values, they were excellent in all cases, with the overall performance slightly superior to that obtained in VARIANT SUBSAMPLING 01. However, this slight improvement was anecdotal due to its minimal magnitude.

In conclusion, the results in the two subsamplings were excellent and very similar.

The high number of samples in the dataset could be related to these good results. The breadth of the dataset might provide a robust and diverse representation of phenomena, a more solid validation, or a more exhaustive pattern detection, such that the model learning could have developed quasi-optimally.

In both subsamplings, a slight decrease in Test Accuracy was detected for the pre-VOC category. This phenomenon could be associated with the fact that the pre-VOC category is not a specific variant, but rather a compilation of all lineages of SARS-CoV-2 existing until the emergence of Alpha, the first variant classified as a VOC.

Special mention is required for the confusion matrix. The Confusion Matrix results for VARIANT SUBSAMPLING 01 and VARIANT SUBSAMPLING 02 that are shown in Figures 12.5 and 12.6, respectively.

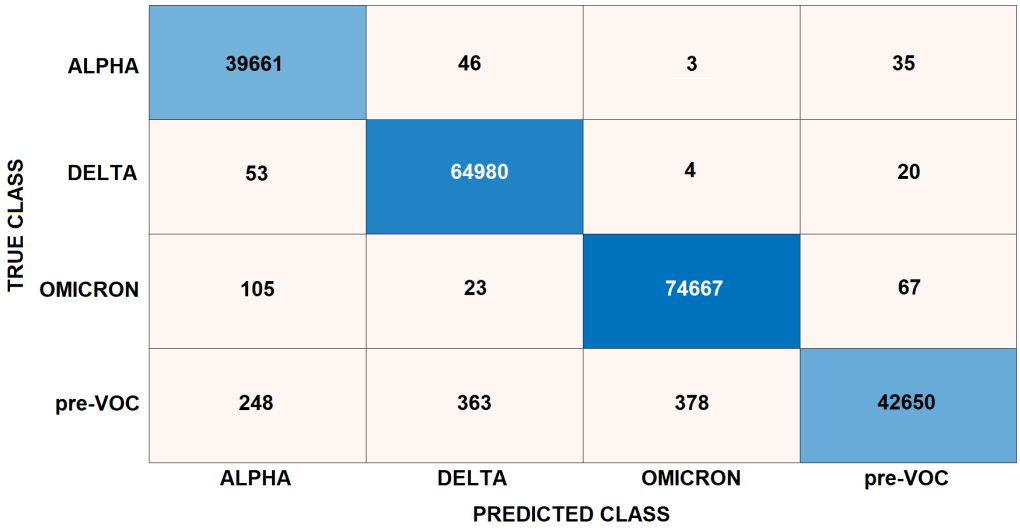


Figure 12.5: VARIANT_SUBSAMPLING_01 Confusion Matrix. The correct predictions are shaded in the grid in blue color, ranging from deep blue denoting a high accuracy ratio to light blue denoting a medium to low accuracy degree. Misclassifications are shaded in orange, ranging from very light orange denoting a low or very low error rate to intense orange indicating a medium to high error rate. Given the high accuracy rates, in this case, the blues are dark and the oranges are light.

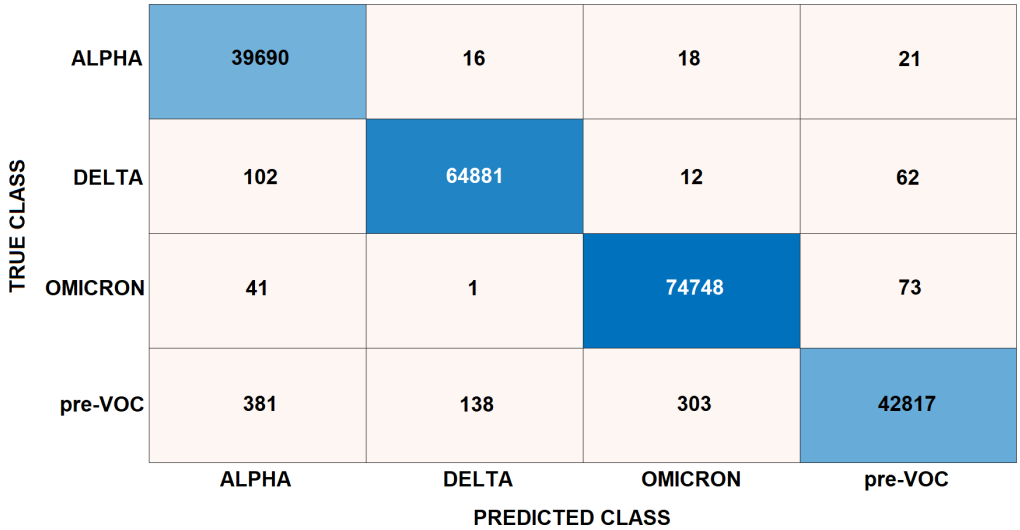


Figure 12.6: VARIANT_SUBSAMPLING_02 Confusion Matrix. The correct predictions are shaded in the grid in blue color, ranging from deep blue denoting a high accuracy ratio to light blue denoting a medium to low accuracy degree. Misclassifications are shaded in orange, ranging from very light orange denoting a low or very low error rate to intense orange indicating a medium to high error rate. Given the high accuracy rates, in this case, the blues are dark and the oranges are light.

In both subsamplings, the accuracy rate was excellent in all categories, with classification errors being practically anecdotal.

In both subsamplings, Omicron had the highest accuracy rate, followed by Delta and Alpha, although pre-VOC had the lowest accuracy rates, all of them were very high in any case. One influencing factor in these accuracy rates could be the quantity and homogeneity of samples in each variant. Omicron is the most numerous variant, while Alpha is precisely the variant with the fewest complete genomes sequenced. Additionally, pre-VOC represents the compilation of lineages before Alpha globally, so a greater variability compared to greater homology in Omicron cannot be ruled out.

In any case, I observed very positive accuracy ratios in all categories, as well as very similar behavior in both subsamplings. These factors indicate that the CNN adequately detects differentiating patterns of each of the four categories.

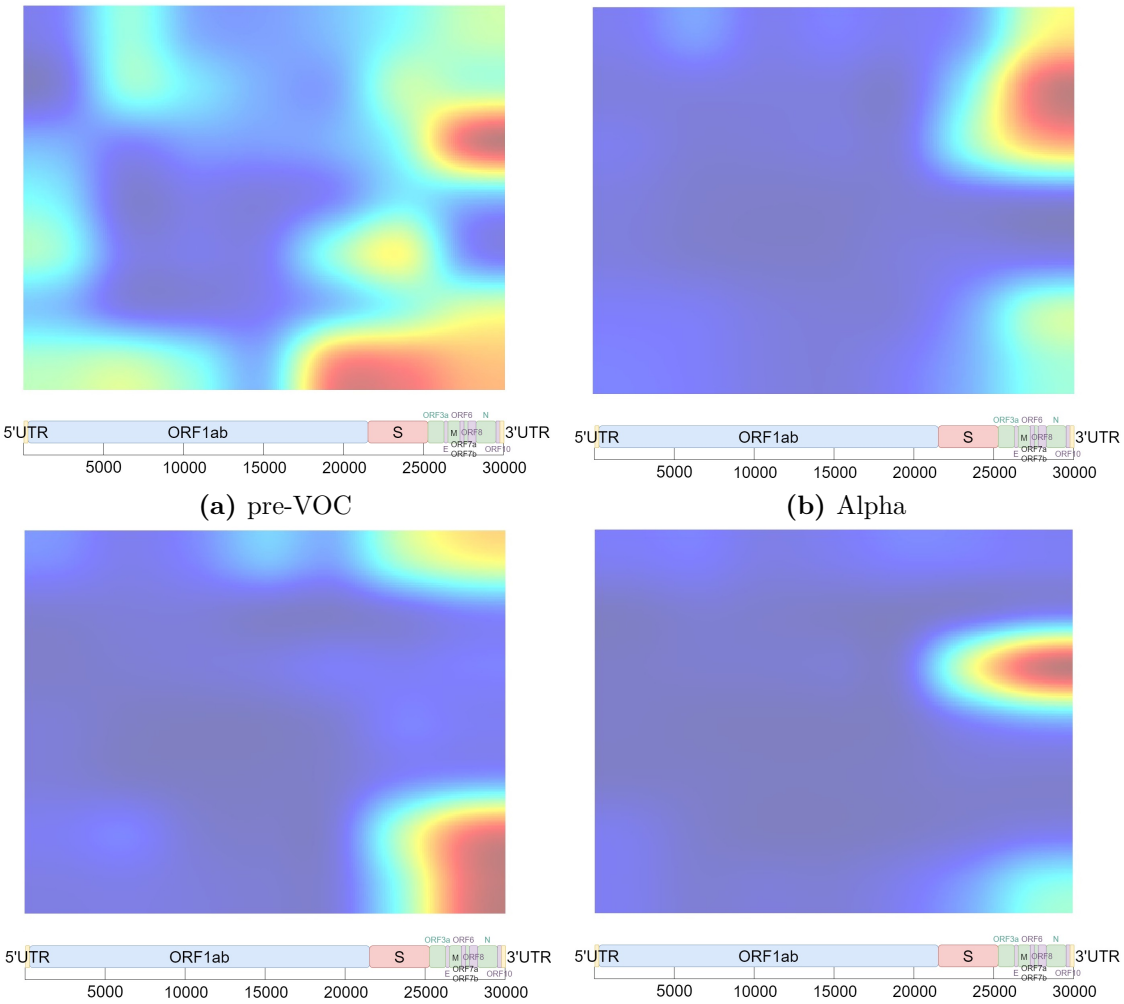
12.4.3 Total Hot Zones per Variant

The term "hot zone" refers to where CNN looks to classify a genomic spectrogram into the four categories of non-recombinant variants in SARS-CoV-2. All total hot zones were calculated as the sum of the scoremaps of the sequences in the Test Set.

In Section 12.4.2, the high Test Accuracy per category was the first indication that the CNN effectively detected patterns inherent to each of the non-recombinant categories in SARS-CoV-2. By calculating the Grad-CAM total hot zones, I identified the genomic regions where the CNN looked to determine which variant each sequence belongs to.

Figure 12.7 compiles the total hot zones per variant in VARIANT SUBSAMPLING 01.

In the case of VARIANT SUBSAMPLING 01, the total hot zones corresponding to the pre-VOC compilation (Figure 12.7a)were diverse and spread across the entire genomic spectrogram. The most relevant hot zones were located in the low-frequency range from the last quarter of ORF1ab to 3'UTR. However, a higher intensity was perceived from the last quarter of ORF1ab to the end of the S protein



(a) pre-VOC **(b)** Alpha

(c) Delta **(d)** Omicron

Figure 12.7: Total Hot Zones per Variant (VARIANT_SUBSAMPLING_01). Figure 12.7a shows the total hot zones corresponding to the pre-VOC classification. Figure 12.7b, those corresponding to the Alpha variant, Figure 12.7c those for the Delta case, and Figure 12.7d those corresponding to Omicron.

approximately. Another main total hot zone was clearly located at $f = 1/3$ from the start of ORF3a to 3'UTR. The zone of greatest intensity, however, was located from N to the end of the sequence (end of 3'UTR).

In addition to these main areas, a secondary hot zone was identified in the low-frequency range, spanning approximately over the first half of ORF1ab, up to nucleotide 10,200 approximately. The epicenter of this zone was visible in the vicinity around nucleotide 5,000 approximately.

Two secondary total hot zones were identified at $f = 1/6$. The first one spanned

the 5'UTR and a first fragment of ORF1ab, extending to approximately nucleotide 2,500. The second hot zone, located at $f = 1/6$, was identified around the S protein region. In Chapter 11, Section 11.5.4, I identified a distinct hot zone crucial for classification as recombinant, precisely in S, at $f = 1/6$. I could hypothesize that this coincidence between hot zones indicates the existence of pre-VOC evolutionary ramifications with hidden genetic recombination phenomena. In that case, it would be deeply interesting to conduct further research that includes the exact localization of the involved sequences and their characteristics.

Another secondary zone in the total hot zones corresponding to pre-VOC was located from the high frequency range (0.5 Hz) to $f = 1/3$ around nucleotide 7,500, in ORF1ab. The last secondary total hot zone also corresponded to the frequency range from $f = 1/3$ to 0.5 in the genomic region from S to the end of the sequence.

The total hot zones of Alpha were sharper and more centered (Figure 12.7b). The main hot zone was clearly located at $f = 1/3$ spanning to 0.5 Hz from S protein to the end of 3'UTR. The epicenter of this hot zone spanned from N to the end of 3'UTR. A secondary zone was detected at $f = 1/9$ from approximately E to the end of 3'UTR.

It was striking that the main hot zone of Delta (Figure 12.7c) showed similarities with the secondary hot zone of Alpha, both in its location and its geometric composition. This phenomenon requires further research, primarily to determine whether this coincidence could be related to common ancestors between both variants. In the case of Delta, this main hot zone was detected spanning from ORF3a to the end of 3'UTR, ranging from low frequencies up to $f = 1/9$. Another main hot zone was located in the high frequency range from S to the end of 3'UTR. However, the epicenter seems to have been located in 3'UTR, radiating up to approximately M.

In the case of Omicron (Figure 12.7d), the main hot zone was very sharp and clearly located at $f = 1/3$, spanning from S to the end of 3'UTR. However, the epicenter of this hot zone was clearly located in 3'UTR and radiated up to ORF8. There was also a secondary zone in the low frequency range, with its epicenter in 3'UTR and radiating up to M.

The frequency $f = 1/3$ in 3'UTR played a significant role in classifying a sequence as pre-VOC, Alpha, and Omicron. However, in the case of Delta, it was clearly neither a main nor a secondary hot zone. This is an interesting phenomenon that requires further investigation. It would be necessary to identify the characteristics shared by the Alpha, pre-VOC, and Omicron variants and subsequently analyze the differences between these common characteristics. The goal is to identify the distinctive elements within these common zones that clearly sway the CNN towards a particular variant.

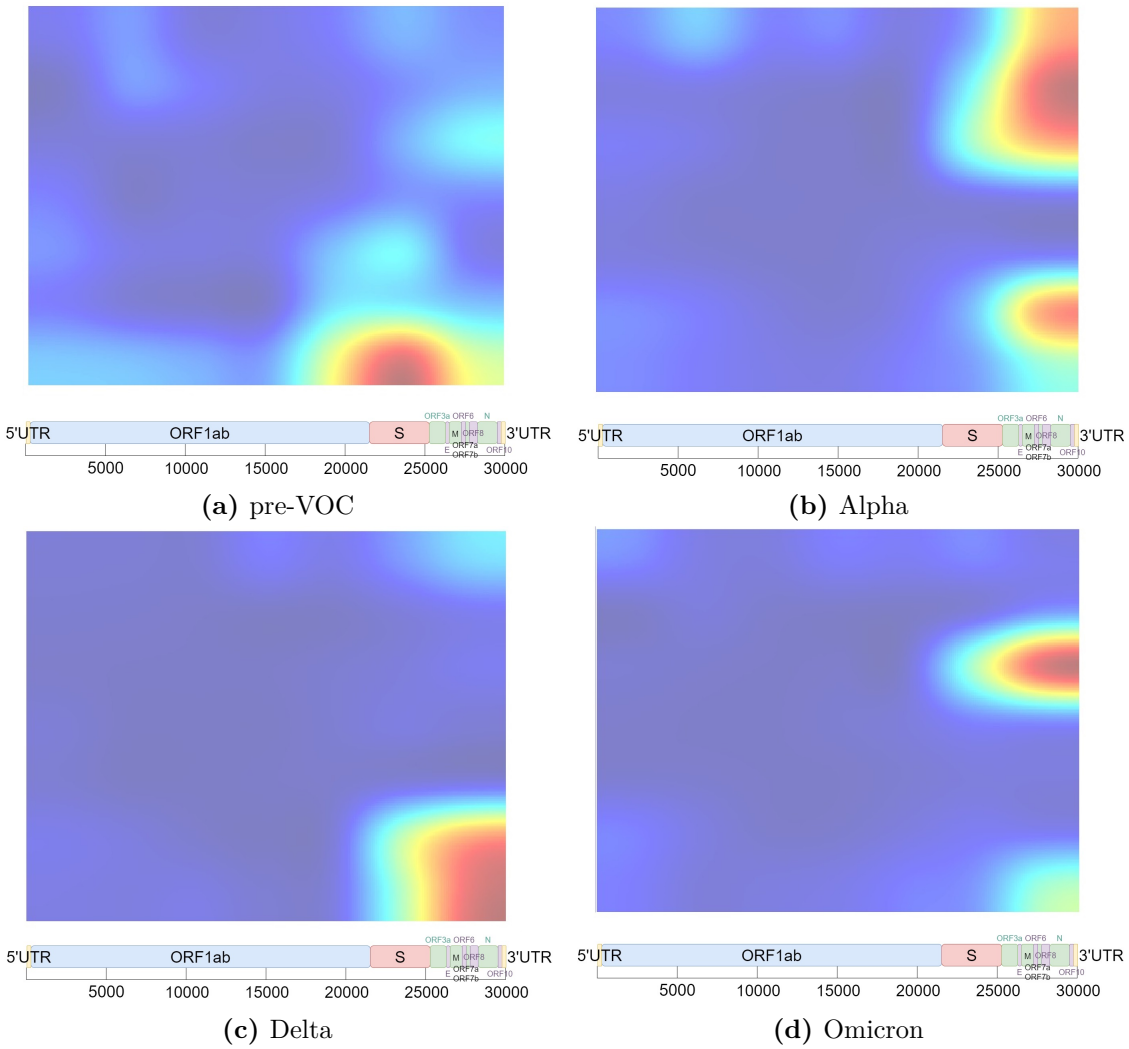
Regarding Delta, it is the only variant that did not include $f = 1/3$ in 3'UTR among its total hot zones, not even in a very secondary manner. At this point, it was the only non-recombinant variant that exhibited this behavior in the CNN. It will be interesting to analyze the differences between 3'UTR in Delta and the other three variants to try to identify why this genomic region does not harbor a pattern affecting Delta, what distinguishes it from the rest.

Figure 12.8 compiles the total hot zones by category in the case of VARIANT SUBSAMPLING 02.

Regarding pre-VOC variant (Figure 12.8a), there was a single sharp and intense main hot zone in the low frequency range. Its epicenter was clearly centered entirely on S, although it radiated to approximately nucleotide 17,500 (ORF1ab) on one side, and to the end of 3'UTR on the other.

The rest were secondary zones that exactly coincided with the remaining hot zones of VARIANT SUBSAMPLING 01 but with much lower power. The first secondary total hot zone was located at $f = 1/3$ from the start of ORF3a to 3'UTR. In the low frequency range, there was a secondary hot zone spanning approximately over the first half of ORF1ab, up to approximately nucleotide 10200.

At $f = 1/6$, two secondary total hot zones were perceived. The first hot zone located at $f = 1/6$ was identified approximately in the vicinity of the S protein. Similarly to VARIANT SUBSAMPLING 01, there was a secondary total hot zone located at $f = 1/6$ for S, the characteristic total hot zone of recombinant SARS-CoV-2 sequences. This phenomenon occurred in the pre-VOC category in both



(a) pre-VOC **(b)** Alpha

(c) Delta **(d)** Omicron

Figure 12.8: Total Hot Zones per Variant (VARIANT SUBSAMPLING 02). Figure 12.7a shows the total hot zones corresponding to the pre-VOC classification. Figure 12.7b, those corresponding to the Alpha variant, Figure 12.7c those for the Delta case, and Figure 12.7d those corresponding to Omicron.

subsamplings, which emphasizes the interest in delving into this phenomenon and analyzing the hypothesis of the existence of hidden recombination phenomena in certain pre-VOC lineages.

The second secondary hot zone at $f = 1/6$ was barely perceptible and encompassed the 5'UTR and a first fragment of ORF1ab, extending up to approximately nucleotide 2,500.

Another barely perceptible secondary zone in the total hot zones corresponding to pre-VOC was located from the high frequency range (0.5 Hz) to $f = 1/3$

around nucleotide 7,500, in ORF1ab. The last secondary total hot zone also corresponded to the frequency range from $f = 1/3$ to 0.5 in the genomic region from S to the end of the sequence.

In the case of Alpha (Figure 12.8b), the most potent relevant hot zone had its epicenter at $f = 1/3$ in 3'UTR up to N. However, it radiated to S, and in terms of frequencies, it even reached the high frequency zone.

Another main hot zone of Alpha was located around $f = 1/9$ from approximately ORF3a to the end of 3'UTR.

The main total hot zone of Delta (Figure 12.8c) was located from the low frequency range to $f = 1/9$, spanning from S to the end of 3'UTR. However, the epicenter was in 3'UTR, and the greatest power reached up to ORF3a.

A secondary zone is in the high frequency range from 3'UTR to ORF3a. The epicenter also appears to be in 3'UTR.

In the case of Omicron (Figure 12.8d), the main hot zone was very sharp and clearly located at $f = 1/3$, spanning from S to the end of 3'UTR. However, the epicenter of that hot zone was clearly located in 3'UTR and radiated up to ORF8.

There was a secondary zone in the low frequency range, with its epicenter in 3'UTR and radiating up to M or perhaps E.

Similar to VARIANT SUBSAMPLING 01, Delta was the only variant where $f = 1/3$ in 3'UTR had absolutely no relevance in terms of classification.

To establish a more direct comparison between the total hot zone results corresponding to VARIANT SUBSAMPLING 01 and VARIANT SUBSAMPLING 02, Table 12.7 compiles the results of both subsamplings by category.

The most visually relevant phenomenon is the difference between the total hot zones in both subsamplings for the category named pre-VOC. The zones corresponding to VARIANT SUBSAMPLING 01 were more diffuse and expanded than in the case of VARIANT SUBSAMPLING 02. In the latter subsampling, the main total hot zone was very centered and more delimited in the low frequency range around the S protein. Although the rest of the zones were located approximately

in the same regions and frequencies as in VARIANT SUBSAMPLING 01, their power was noticeably lower.

The pre-VOC category turned out to be the least similar of the four analyzed. It would be interesting to delve into the relationship between the sequences that compose both subsamplings and attempt to establish if there is a connection between these sequences and the disparity of results between both subsamplings. Just as determining the causes that provoke that sharpness in VARIANT SUBSAMPLING 02 and that variability in VARIANT SUBSAMPLING 01.

Homology in the case of Alpha, Delta, and especially Omicron, was higher between both subsamplings. In both subsamplings, they were highly similar and consistent with each other.

In the case of Alpha, the main primary hot zone, located around $f = 1/3$ in the final genomic region, turned out to be highly similar. In the case of VARIANT SUBSAMPLING 02, there was more power in the expansion from $f = 1/3$ to $f = 1/2$ (high frequencies).

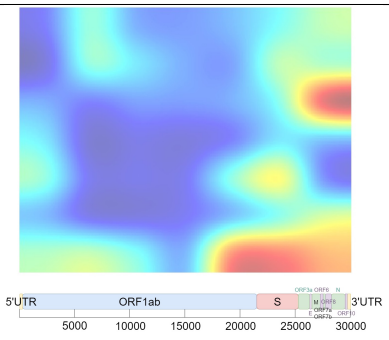
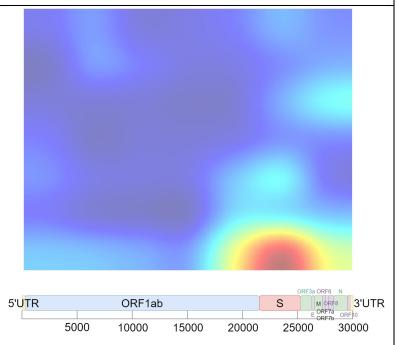
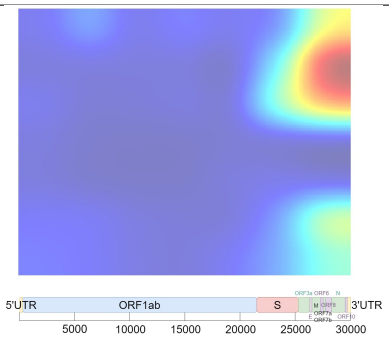
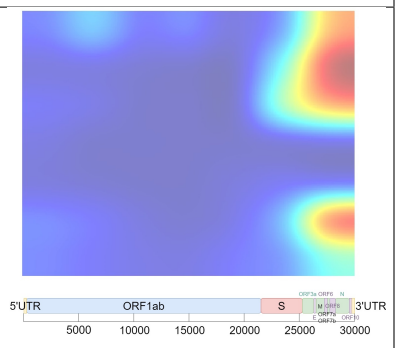
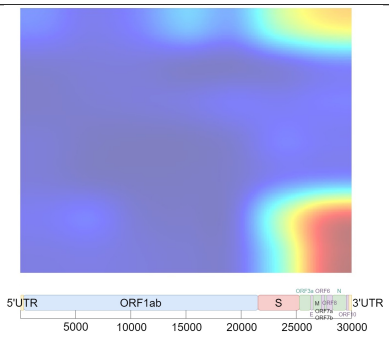
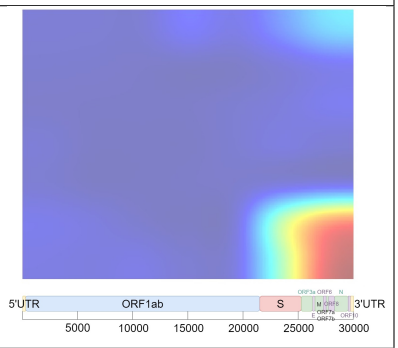
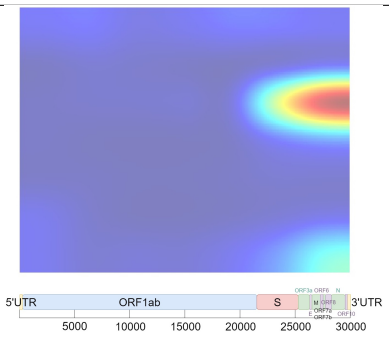
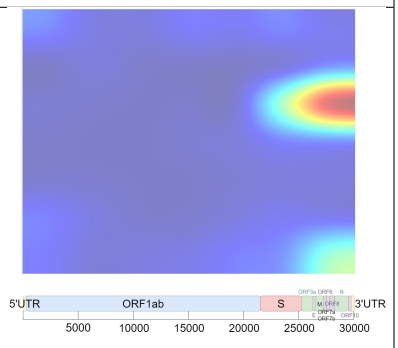
The secondary region located in the final stretch of the genome in VARIANT SUBSAMPLING 01, $f = 1/9$, had more power in VARIANT SUBSAMPLING 02.

I detected an inverse phenomenon in the case of Delta. In this category, VARIANT SUBSAMPLING 01 exhibited slightly higher power in the very secondary hot zone of VARIANT SUBSAMPLING 02, located in the high frequencies of the final stretch of the genome. The main hot zone was extremely sharp, centered, and similar. In both subsamplings, it was located in the range from low frequencies to $f = 1/9$ from the end of S to the end of 3'UTR.

Omicron was the category with the highest homology between both subsamplings. Both total hot zones were practically identical. The main hot zone located at $f = 1/3$ in the final genomic region was nearly identical, perhaps slightly more expanded towards S in VARIANT SUBSAMPLING 01 compared to VARIANT SUBSAMPLING 02. The secondary zone located in the low frequency range in the bottom right corner of the genomic spectrogram was virtually the same in

both subsamplings. There was a slightly higher power, practically undetectable, in the case of VARIANT SUBSAMPLING 02.

Table 12.7: Comparative Table of Total Hot Zones in Both Subsamplings

Variant	VARIANT SUBSAMPLING 01	VARIANT SUBSAMPLING 02
pre-VOC		
Alpha		
Delta		
Omicron		

Once the high degree of accuracy by the CNN in both subsamplings was

quantified and the total hot zones were calculated, the high degree of similarity in the total hot zones in both subsamplings (more so in Alpha, Delta, and Omicron than in pre-VOC) indicated that the behavior of the CNN was robust and consistent in determining intrinsic and differential patterns of each of the four categories analyzed.

The results regarding total hot zones in the pre-VOC category had two particularities compared to the rest of the analyzed variants. On the one hand, the total hot zones were more varied, less centered on specific points. And on the other hand, the hot zones between both subsamplings were notably different, not so much in position but in power.

This variability could be aligned with the fact that the category known as pre-VOC is not really a variant in itself. It is a compilation of everything that happened until the emergence of Alpha, the first VOC variant. However, several studies determined that during the period from the pandemic outbreak until the emergence of Alpha, the genetic variability of SARS-CoV-2 was limited [2731]. Some research determined the co-existence of two lineages simultaneously in Wuhan during the early stages of the virus spread [2870].

Curiously, pre-VOC appears to include a typical hot zone of the variant later named Omicron. The Omicron variant (B.1.1.529) of SARS-CoV-2 does not directly descend from the Alpha (B.1.1.7) or Delta (B.1.617.2) variants. Omicron is a variant identified as part of a distinct evolutionary lineage and has developed its own genetic characteristics. Among the multiple theories regarding its origin, three seem the most plausible [2871–2873].

First, Omicron could have circulated and evolved in a hidden population. A possible explanation is that Omicron, or its ancestral variant, evolved in a remote area of the world where nucleic acid tests are rarely conducted. In fact, nucleic acid tests in South Africa might have already captured a glimpse of Omicron's ancestor.

Second, Omicron might have evolved from a cat-and-mouse game between the virus and the host in some immunocompromised patients, such as AIDS patients infected with SARS-CoV-2.

Third, Omicron could have originated from adaptation in animal reservoirs (possibly rodents) and then transmitted back to humans.

More research is needed at this point to elucidate the possible reasons that may be aligned with the apparent variability in the mathematical signatures associated with the pre-VOC category, as well as the possible evolutionary relationships between pre-VOC and Omicron.

In any case, the existence of perfectly distinguishable mathematical patterns in each of the four categories studied is noteworthy.

12.5 Conclusions

Training the model with a single epoch yielded positive results in all performance measurement parameters, with particularly high Test Accuracy in both subsamplings. The large size of the dataset for each of the subsamplings, totaling 1,116,523 images, allowed the CNN to access a significant number of training examples, making it very likely that it could effectively learn patterns present in the data using only one epoch.

In the experiment presented in this chapter, I introduced a double change in the application of Two-Stage Transfer Learning. In Chapter 11, I extrapolated the knowledge of the CNN in genetic recombination in HIV-1 to genetic recombination in SARS-CoV-2, so the only change was related to the virus under study. In this case, I not only moved to another virus but also the object of study itself. I used a CNN that detects genetic recombination in HIV-1 to detect the type of non-recombinant variant in SARS-CoV-2. I obtained excellent and robust results in this double change, achieving test accuracies of 99.40% in VARIANT SUBSAMPLING 01 and 99.48% in VARIANT SUBSAMPLING 02. The Test Accuracies per category ranged between 97.73% and 99.88%.

Regarding the total hot zones by category, the patterns inherent to each category proved to be absolutely distinguishable, robust, and with high (in the case of pre-VOC) and very high homology (in the case of Alpha, Delta, and Omicron) in both subsamplings.

In the case of pre-VOC, the differences between both subsamplings referred to the loss of power in VARIANT SUBSAMPLING 02 compared to VARIANT SUBSAMPLING 01. However, the physical location of the total hot zones turned out to be coincident.

The Omicron variant showed total hot zones practically unaltered in both subsamplings.

I consider it extremely interesting to delve into the relationships between the phylogenetic reality and the evolutionary diversity associated with the pre-VOC, Alpha, Delta, and Omicron compilation, and the mathematical signatures that determine the CNN classifying them into these variants.

Although the total hot zones corresponding to each variant are unequivocal and perfectly distinguishable, there are certain similarities that make it highly interesting to delve into the phylogenetic relationships among these four main variants. For example, the similarity in the importance (though not in the pattern shape) of $f = 1/3$ in the final segment of the genome for pre-VOC and Omicron may be related to the phylogeny of SARS-CoV-2 or even to the origin of the Omicron variant from some pre-VOC branch. This hypothesis requires further research.

From a computational standpoint, I demonstrated a successful application of Two-Stage Transfer Learning with two simultaneous changes: the type of virus and the object of study. The CNN's learning on genomic spectrograms is promising not only for the detection of the recombinant feature but also for other characteristics, in this case, regarding the variant to which a sequence belongs.

Therefore, the positive results of this double application of Transfer Learning implicitly demonstrate the versatility in the utilization of genomic spectrograms in various virological analyses.

From a biological perspective, the robust and specific total hot zones in each type of variant denote the importance of the genomic region from the beginning of S to the end of 3'UTR in the identification of the analyzed variants. The frequencies $f = 1/9$ and especially $f = 1/3$ in this genomic region appeared to play a relevant role in such identification. It will be necessary to delve into the search for the

exact relationship between the nucleotide periodicities relative to these frequencies and the belonging to one variant or another.

Just like in the case of characterizing a sequence as recombinant, in this instance, I demonstrated the existence of a mathematical signature associated with the pre-COV, Alpha, Delta, and Omicron variants. Therefore, the existence of other mathematical signatures associated with other viral characteristics cannot be ruled out.

Mathematics is the language with which God has written the Universe.

— Galileo Galilei, *in 1581, he began to study medicine. Four years later, he dropped out of university and pursued his true passions: mathematics and physics.*

13

Exploring the Influence of the Third Codon Nucleotide on Genetic Recombination in the SARS-CoV-2 Spike Protein

Contents

13.1 Outcome Hypothesis Analysis 656
13.2 Equipment 657
13.3 Methods 657
13.3.1 Selection of Study Subsampling 657
13.3.2 Selection of Study Sequences 658
13.3.3 Download of EMBL Files 658
13.3.4 Extraction of Relevant Protein Genomes 659
13.3.5 Preliminary Identification of Visual Patterns in Arithmetic Sequences of 6 Nucleotides in S 660
13.3.6 Calculation of Nucleotide Content 661
13.3.7 Search for Common Nucleotides in S among Sequences of the Same Category 662
13.3.8 Biological Interpretation of Results 662
13.4 Results and Discussion 663
13.4.1 Subsampling Under Study 663
13.4.2 Sequences Under Study 663
13.4.3 Preliminary Verification of the Actual Position in S of the First Sixth Nucleotide 667
13.4.4 Preliminary Identification of Visual Patterns in Arithmetic Sequences of 6 Nucleotides in S 670
13.4.5 Nucleotide Content 672

13.4.6 Common Nucleotides 678
13.4.7 Microbiological Analysis 686
13.5 Conclusions 687
13.6 Future Research 691

13.1 Outcome Hypothesis Analysis

In Chapter 11, I identified a clear hot zone in the recombinant sequences of SARS-CoV-2, absolutely centered at $f = 1/6$ in the S protein. Its clarity was striking, not only in its sharpness but also in its precise positioning in the S protein. It is in this protein where numerous events of genetic recombination have occurred in SARS-CoV-2, and it is a fundamental protein in the virus's cell infection process.

In Chapter 5, I detected the presence of the most relevant mathematical signature determining that a complete sequence of HIV-1 was recombinant at $f = 1/3$ in 5'LTR and perhaps part of *gag* (see Figures 5.9 and 8.1b). The location of that signature did not coincide with the result of that genetic recombination.

However, in this case, the mathematical signature coincided with a genomic region where these phenomena have commonly occurred. And, as I saw in Chapter 10, it was a region of low conservation, highly subject to variability, especially in terms of nucleotide match. The degree of nucleotide match was only 0.23%, making it the most variable protein in terms of nucleotide conservation.

The clear detection by the CNN of the potential position of the characteristic mathematical signature of the recombinant feature in the S protein, frequency $f = 1/6$, made it necessary to deepen our understanding of this signature, especially given its evident positioning in such a hot and relevant region. I know where the CNN looks, and I want to understand more about what it sees. I have a case study that is an excellent opportunity for this. I accept the challenge and proceeded to explore what this mathematical signature entails.

This Chapter gathers the results of my approach to the mathematical signature that characterizes a SARS-CoV-2 sequence as recombinant. It begins with a description of the equipment used and the methodology employed (Sections 13.2

and 13.3 respectively). Subsequently, I present the results of the analysis conducted along with a discussion of these results (Section 13.4), concluding with the main conclusions from the experimentation covered in this Chapter (Section 13.5) and the suggestions for future research derived from this analysis (Section 13.6).

13.2 Equipment

All experiments were run in this equipment:

- Processing Unit: Intel(R) Core(TM) i7-4770K CPU. 3.5 GHz.
- Installed RAM: 32 GB usable.
- Operative System: Windows 10 Education. Version: 22H2.
- GPU: NVIDIA GeForce RTX 3090. Total memory: 40 GB.

13.3 Methods

All computer analyses were developed in a Python environment, using primarily the following libraries:

- Pandas, for data manipulation in the form of DataFrames. A DataFrame in Python is a labeled, mutable two-dimensional data structure that can contain data of different types [2676].
- Numpy, for implementing numerical operations [2677].
- Math, for applying certain mathematical functions [2678].
- Os, for operations related to the operating system [2679].
- Re, for regular expression operations [2680].
- The Selenium WebDriver library (Python), for automating the web download of EMBL files using scripts [2681].

13.3.1 Selection of Study Subsampling

Within the optimal performance configurations of the CNN expressed in Chapter 11, I selected a completely random subsampling. The choice of this optimal configuration ensures that the input data are the result of the CNN's best pattern detection,

with higher and more balanced accuracy levels. In this optimal configuration, the hot zones were more precise and accurate, and my input data in this study were more reliable.

For this same optimal configuration, the results per subsampling were highly similar, so the choice of a particular subsampling was indifferent.

Thus, my analysis data was based on the results obtained for 200 Epochs (optimal configuration), Subsampling 05.

13.3.2 Selection of Study Sequences

In that subsampling and for that configuration of epochs, out of the 313 recombinant sequences in the Test Set, I selected all those correctly classified as recombinant by the CNN whose hot zone for such classification was exclusively $f = 1/6$ in S. In these sequences, the CNN looked exclusively at S to correctly classify them as recombinant, so it is plausible that the pattern is more deterministic in these sequences than in others where the CNN looked at another zone in the spectrogram.

Based on the number of sequences resulting from this selection, I chose the same number of non-recombinant sequences correctly classified by the CNN whose hot zones do not include $f = 1/6$ in S. That is, this zone was expressly excluded from the hot zones in each non-recombinant sequence.

My objective with this selection criterion was to work with sequences whose patterns in the superposed spectrogram were as different as possible for the CNN, thus approximating a search for differential elements in an environment where structural differences were apparently more evident. In this way, pinpointing those differences between one study group and the other a priori may be simpler.

13.3.3 Download of EMBL Files

Once the complete SARS-CoV-2 recombinant and non-recombinant sequences under study were identified, the first step in the search for more data about that mathematical signature related to the recombinant feature in SARS-CoV-2 was

the identification of both the S proteins in each sequence under study and other proteins with which to compare potential results.

The FASTA file contains the genomic sequence expressed in nucleotides. These files provide a detailed representation of genetic information, showing the sequence of nitrogenous bases that make up DNA or RNA.

The EMBL file is a type of file used to store nucleic acid sequences and their associated information in the European Molecular Biology Laboratory (EMBL) Sequence Database format. EMBL files contain metadata about the sequence, such as accession numbers, descriptions, gene annotations, coding regions, DNA features (such as promoters, transcription factor binding sites, etc.), and any other relevant information [2682]. This format allows supplementing the mere nucleotide content information of the genome with supplementary data of high added value.

Through an automated detection and download system, I collected all the EMBL files of the involved sequences from the public repository of the European Nucleotide Archive (ENA) [39].

13.3.4 Extraction of Relevant Protein Genomes

The first step was the selection of the proteins under study. In the case of the recombinant sequences, whose exclusive hot zone was $f = 1/6$ in S, the proteins of interest were:

- S protein.

And, in order to compare any significant results I might find in this structural protein with other structural proteins, I selected:

- M protein.
- N protein.

Both structural proteins were more volatile in terms of nucleotide consistency in Chapter 10. In the case of N, its Nucleotide Consistency Index was 0.95%, and for M, it was 7.03%. The fourth structural protein E proved to be more stable with a Nucleotide Consistency Index of 23.25%. I preferred to work with less conserved structural proteins to establish comparisons in the worst-case scenario. The high

conservation of E compared to the rest would result in comparisons with a less variable element, and therefore more repetitive in its comparison. In other words, by changing less, we are always comparing with something similar. I found it more challenging to work with more volatile structural proteins.

In the case of non-recombinant sequences where $f = 1/6$ in S was irrelevant, the protein under study was limited to:

- S protein.

In this initial prospecting phase, I considered it sufficient to study these proteins to make a first approach to understanding that mathematical signature. Of course, in future research, a more comprehensive study will be necessary.

Once the proteins of interest were determined, with the EMBL files already downloaded, I proceeded to process their content to identify the positions of all the proteins based on the tags in their respective EMBL files. I located in the EMBL files the start and end positions of each protein in all the sequences under study. From these positions, I extracted from the FASTA file the genome of each protein, storing it in a new CSV file named "ID_protein.csv". Here, "ID" is the identifier of the sequence, and "protein" is the extracted protein.

13.3.5 Preliminary Identification of Visual Patterns in Arithmetic Sequences of 6 Nucleotides in S

In the case of genomic spectrograms, a frequency $f = 1/6$ refers to an arithmetic sequence of 6 nucleotides. That is, the nucleotides included in this sequence correspond to positions that are separated by intervals of 6, starting from a defined initial position. The possible defined initial positions were from 6, 12,... up to 11, 17,...

I made a very simple first approach to detect any particularity in some of the possible arithmetic sequences of 6 nucleotides. Starting from the series 6, 12,... and ending at 11, 17,...

For all target proteins, to this end, I first extracted in each of their corresponding CSV files the nucleotides corresponding to S in each of the sequences of each

series and stored them in separate text files called "ID_protein_from-initial-position.txt". Where:

- "ID" is the sequence identifier.

- "Protein" is the protein from which the nucleotides of the arithmetic sequence were extracted.

- "From-initial-position" is the arithmetic sequence, defined by the initial position (from 6 to 11).

Then I transferred the data from each text file into an XLS table with the following format: each column represented one of the recombinant sequences. Each row contained the nucleotides extracted every 6 positions from the indicated starting position. The first row contained the first sixth nucleotides, the second row the second sixth nucleotides, and so on.

Finally, I assigned a color to each type of nucleotide. Adenine (A) is represented in green, Thymine (T) in blue, Cytosine (C) in yellow, and Guanine (G) in red. I colored each cell according to the nucleotide it contained. This way, I was able to observe, in a very simple manner, visible patterns for the human eye in terms of nucleotide diversity, as well as the degree of similarity between the selected sequences.

This process was applied to the S proteins in both recombinant and non-recombinant sequences, as well as to the M and N proteins of the recombinant sequences.

13.3.6 Calculation of Nucleotide Content

In each of the arithmetic series extracted in the previous step, I calculated the average occurrence of A, T, C, and G for each of the analyzed proteins, as well as the Standard Deviation (SD), with the aim of obtaining information regarding the abundance of each nucleotide in each protein, as well as the variability found in the sequences under analysis. These data may allow us to search for peculiarities among the arithmetic series in each analyzed protein and provide clues about the mathematical signature embedded in the genome.

13.3.7 Search for Common Nucleotides in S among Sequences of the Same Category

The recombinant sequences in the test set under study were those where the hot zone was exclusively $f = 1/6$ in the S protein. The non-recombinant ones in the test set were those where precisely this hot zone did not appear in the Grad-CAM hot map result.

It seems plausible that this common feature in the hot zones of each of the two groups has some kind of correspondence or relationship with the nucleotides that are common in each of those two groups. It is possible that the common nucleotides may provide us with more information about the mathematical signature. Therefore, my next step was to determine exactly which nucleotides were common in all the selected recombinant sequences and which nucleotides were common in the selected non-recombinant sequences.

With this aim, I implemented a nucleotide detection program that returns their position within the protein and the type of nucleotide (A, T, C, G, or others).

In this case, I focused solely on the S protein of both sequence groups since the mathematical signature, according to the CNN, would be located in this protein. The study of M and N in the case of the selected recombinant sequences was limited to comparative terms with S, but the common nucleotides are not relevant in M and N in this case study.

These common elements were stored in an XLS file named "Group_commonnts.xls", where "Group" indicates whether it is the group of recombinant or non-recombinant sequences. This XLS file provides, in its first column, the position of each common nucleotide of S starting from position 1. The second column indicates, for that position, the corresponding common nucleotide (A, T, C, G, or others).

13.3.8 Biological Interpretation of Results

Once I obtained data from the previous steps, I analyzed it thoroughly in search of information that brings us closer to that mathematical signature, while also conducting a biological interpretation based on existing expert literature.

13.4 Results and Discussion

13.4.1 Subsampling Under Study

According to the criterion outlined in Section 13.3.1, I chose the configuration corresponding to 200 Epochs, whose performance ratios and balance between categories were optimal. See Section 11.4.2. The randomly selected subsampling was Subsampling 05.

13.4.2 Sequences Under Study

As indicated in Section 13.3.2, I selected the recombinant sequences within the test set whose hot zones were exclusively $f = 1/6$ in the S protein.

As a result of this criterion, 39 recombinant sequences out of the total of 313 recombinant sequences that composed the Test Set met this criterion.

Based on this number of sequences, I selected 39 non-recombinant sequences in the test set (composed of a total of 313 non-recombinant sequences) whose hot zones did not include either primarily or secondarily $f = 1/6$ in the S protein.

Appendix L is the compilation of all the hot zones corresponding to each sequence under study, grouped into recombinants and non-recombinants, respectively.

Table 13.1 provides the compilation of sequences selected according to this criterion, as well as the lineage to which each of the selected sequences belongs.

Table 13.1: Sequences under Study and Their Lineages. Grouped into recombinant and non-recombinant.

Recombinants			Non-Recombinants	
Sequence	**Lineage**		**Sequence**	**Lineage**
MZ280957	XB		BS002380	BA.1.1.2
MZ558275	XB		BS003228	BA.1.1.2
MZ817897	XC		HG997023	B.1.1.7
OK039706	XC		MT972409	D.3
OL791595	XC		MZ244586	B.1.1.7
OL920359	XC		MZ700508	B.1.1.7
OM059434	XC		OA982375	B.1.177.7
ON144454	XN		OD966578	B.1.1.7
ON330593	XAA		OD969387	B.1.1.7
ON594981	XAC		OD979502	B.1.1.7

Table 13.1 – Continuation from the previous page

Recombinants			Non-Recombinants	
Sequence	**Lineage**		**Sequence**	**Lineage**
ON650495	XAC		OL751379	AY.44
ON800622	XAJ		OL778149	AY.44
ON980049	XE		OL856645	AY.103
OP129015	XAK		OL868832	AY.103
OP294029	XAJ		OL993467	AY.25
OP311263	XAG		OM058523	AY.44
OP313193	XAJ		OM235729	BA.1.1.18
OP313256	XAJ		OM263706	B.1.1
OP313383	XAJ		OM336312	BA.1.20
OP359817	XAJ		OM356528	AY.39
OP768851	XE		OM396820	BA.1
OP934590	XBD		OM412022	BA.1.18
OQ094825	XBB		OM569369	BA.1
OQ111935	XBB		OM857717	BA.1
OQ112611	XBB		ON660231	BA.2.12.1
OQ112616	XBB		ON893479	AY.25
OQ112619	XBB		OP055674	BA.5.1
OQ112624	XBB		OP125987	BA.4.1
OQ112639	XBB		OP200039	BA.4.1
OQ112647	XBB		OP468268	BA.5.5
OQ112649	XBB		OQ085662	BQ.1
OQ112673	XBB		OU448371	AY.5
OQ112686	XBB		OU796188	B.1.1.7
OQ112691	XBB		OV006668	AY.122
OQ112730	XBB		OV008054	AY.4
OQ112737	XBB		OW409375	BA.1.17.2
OQ112740	XBB		OW464011	AY.4
OQ112771	XBB		OX073652	BA.1
OQ112893	XBB		OX248233	BA.5.2.1

From these 39 recombinant sequences, I extracted the specific genomes of S, M, and N. From the 39 non-recombinant sequences, I only extracted their S gene sequence. The exact identification of the genomic regions of all genes under study was automatically performed based on the annotations in the EMBL files of each sequence.

The lineage distribution of the 39 recombinant sequences whose hot zone was exclusively $f = 1/6$ is shown in Table 13.2.

Table 13.2: Lineage Distribution of Recombinant Sequences under Study

Lineage	**No.**	**Percentage**
XBB	17	43.59%
XAJ	6	15.38%
XC	5	12.82%
XB	2	5.13%
XAC	2	5.13%
XE	2	5.13%
XN	1	2.56%
XAA	1	2.56%
XAK	1	2.56%
XAG	1	2.56%
XBD	1	2.56%
Total	**39**	**100.00%**

The lineage XBB was notably the most abundant in this selection of 39 sequences, representing nearly half of the selection. The recombinant lineage XBB has been predominant since 2023, alongside BA.2.86 (Pirola) in 2024. XBB was a recombinant lineage with higher transmissibility, which led to its dominance over other lineages in 2023. It is striking that, despite the limited relevance of XBB in the total dataset (4.28%), this lineage represented 43.59% of the total selected recombinant sequences with a clearly defined hot zone at $f = 1/6$ in S. Further investigation is needed to understand the relationship between this hot zone and the genotypic characteristics of this lineage.

The next most abundant recombinant lineages were XAJ, representing 15.38% of the total, and XC, which accounted for 12.82%. XAJ is a recombinant between BA.2.12.1 and BA.4, both Omicron lineages. The parental lineages of XAC are BA.1 and BA.2., also derived from Omicron [2874].

The remaining eight lineages had specific weights of 5.13% and 2.56%, representing a total of 28.21%. Therefore, considering that the overwhelming majority of recombinants are derived from Omicron lineages, the predominant presence of XBB prompts an evaluation of its influence on the clarity of the hot zone. In other words, whether this hot zone is more representative of XBB than of other lineages. More sequences from various lineages need to be analyzed. It is

necessary to establish the distinguishing elements of each lineage with the specific hot zone areas of those lineages.

Table 13.3 shows the distribution of lineages for the 39 non-recombinant sequences under study.

Table 13.3: Lineage Distribution of Non-Recombinant Sequences under Study

Lineage	**No.**	**Percentage**
B.1.1.7	7	17.95%
BA.1	4	10.26%
AY.44	3	7.69%
BA.1.1.2	2	5.13%
AY.103	2	5.13%
AY.25	2	5.13%
BA.4.1	2	5.13%
AY.4	2	5.13%
D.3	1	2.56%
B.1.177.7	1	2.56%
BA.1.1.18	1	2.56%
B.1.1	1	2.56%
BA.1.20	1	2.56%
AY.39	1	2.56%
BA.1.18	1	2.56%
BA.2.12.1	1	2.56%
BA.5.1	1	2.56%
BA.5.5	1	2.56%
BQ.1	1	2.56%
AY.5	1	2.56%
AY.122	1	2.56%
BA.1.17.2	1	2.56%
BA.5.2.1	1	2.56%
Total	**39**	**100.00%**

In the case of non-recombinant sequences, intervening lineages are much more abundant and exhibit a more evenly sequence distribution. This fact is consistent with the greater variety of non-recombinant lineages compared to recombinant ones in SARS-CoV-2. My dataset consisted of a total of 1,539,728 complete sequences from 1,506 non-recombinant lineages compared to only 1,565 sequences from 39 recombinant lineages. The most abundant lineage in this selection of 39 non-recombinant sequences is B.1.1.7 (nearly 18%) followed by BA.1 (10.26%) and

AY.44 (7.69%). The remaining 20 lineages range from 5.13% to 2.56%. These remaining sequences account for a total of 64.10%, indicating greater variability, perhaps aligned with the greater number of non-recombinant lineages. It is necessary to delve into the possible phylogenetic relationship among these 23 lineages and the absence of the hot zone $f = 1/6$ in S in determining their non-recombinant character.

The compilation of lineages AY refers to the Delta variant. Lineages BA and BQ refer to the Omicron variant, while B.1.1.7 refers to the Alpha variant. To determine the prevalence of the variants labeled Omicron, Alpha, and Delta among the 39 non-recombinant study sequences, I calculated the specific weight of these three designations out of the total. Table 13.4 displays the results.

Table 13.4: Variant Distribution of Non-Recombinant Sequences under Study. The "Variant" column displays the non-recombinant variant, the "No." column indicates the total number of sequences out of the 39 non-recombinants grouped under each variant. The "Percentage" column shows the percentage of each variant relative to the total of the 39 sequences.

Variant	**No.**	**Percentage**
Omicron	17	43.59%
Delta	12	30.77%
Alpha	7	17.95%
Other	3	7.69%
Total	**39**	**100.00%**

Nearly half of the sequences resulting from the screening criteria belong to Omicron, followed by Delta, with a difference of five sequences. Alpha sequences account for only 18%, and only three sequences belonged to other types of variants. The distribution of variants in this study group is similar to the distribution of the total set of non-recombinant sequences (Table 12.1 in Chapter 12).

13.4.3 Preliminary Verification of the Actual Position in S of the First Sixth Nucleotide

The significance of the frequency $f = 1/3$, aligning with the coding regions of the genome in the Fourier analysis representation (Figure 11.2), was previously explained as a result of nucleotides clustering in triplets [2387]. These triplets determine the sequence of amino acids that will constitute the proteins encoded in

different regions of the genome. In a similar vein, my model's identification of a frequency $f = 1/6$ would correspond to a system operating with pairs of amino acids, with the significance being directly attributed to the third nucleotide of the triplet encoding the second amino acid in each pair. Plant and Ye observed a tendency for certain virus and host genomes to display bias at the third nucleotide of codon pairs, suggesting that my results could support or add to their research [2875]. Further exploration into the importance of the third nucleotide of codon pairs and its association with viral traits would be valuable. I hypothesize that the third letter of the triplet may contain encoded information within the frequency domain that could dictate specific viral characteristics, particularly regarding recombinant features.

In a preliminary analysis of 39 representative sequences, accurately classified as recombinant by the CNN, and whose hot zone was perfectly contained within the S protein proximal region at $f = 1/6$, I identified certain peculiarities. First, I verified, sequence by sequence, whether the position corresponding to the frequency 1/6 aligns with the third letter of the second amino acid in S. Upon thorough examination of the spectrograms generated with a window length of 256, I confirmed that the sixth letter indeed matches in most cases (Table 13.5).

Table 13.5: Sixth Nucleotide Position for the Selected 39 Successfully Identified Recombinant Sequences Using the Pretrained CNN. Considering that the genomic spectrograms were generated with a segment length of 256, this Table shows the position of the first sixth nucleotide of S from the beginning for each analyzed sequence. The "Relative Position" column indicates the position of this sixth nucleotide from the start of S (not from the start of the entire sequence). $Position = 6$ means that the first sixth nucleotide of S (counting from the start of the sequence) aligns with the third nucleotide of the second amino acid in S protein.

Sequence	**Lineage**	**Confidence Score**	**S Protein Coding Region**		**First 6th nt**	**Relative Position**
MZ558275	XB	99.99%	21,512	25,333	21,516	5
MZ280957	XB	69.42%	21,519	25,340	21,522	4
MZ817897	XC	68.77%	21,533	25,348	21,534	2
OK039706	XC	80.39%	21,532	25,353	21,534	3
OL791595	XC	99.99%	21,542	25,363	21,546	4
OL920359	XC	99.99%	21,533	25,354	21,534	2
OM059434	XC	99.99%	21,533	25,354	21,534	2
ON144454	XN	98.70%	21,529	25,341	21,534	6
ON330593	XAA	99.98%	21,530	25,342	21,534	5

Table 13.5 – Continuation from the previous page

Sequence	**Lineage**	**Confidence Score**	**S Protein Coding Region**		**First 6th nt**	**Position**
ON594981	XAC	99.98%	21,529	25,341	21,534	6
ON650495	XAC	99.99%	21,529	25,341	21,534	6
ON800622	XAJ	65.51%	21,529	25,350	21,534	6
ON980049	XE	99.94%	21,526	25,338	21,528	3
OP129015	XAK	99.99%	21,529	25,341	21,534	6
OP294029	XAJ	99.88%	21,529	25,341	21,534	6
OP311263	XAG	74.12%	21,526	25,338	21,528	3
OP313193	XAJ	99.86%	21,529	25,341	21,534	6
OP313256	XAJ	99.99%	21,529	25,341	21,534	6
OP313383	XAJ	96.92%	21,526	25,338	21,528	3
OP359817	XAJ	99.84%	21,529	25,341	21,534	6
OP768851	XE	99.91%	21,526	25,338	21,528	3
OP934590	XBD	99.99%	21,529	25,341	21,534	6
OQ094825	XBB	99.97%	21,529	25,338	21,534	6
OQ111935	XBB	99.99%	21,529	25,338	21,534	6
OQ112611	XBB	99.99%	21,529	25,338	21,534	6
OQ112616	XBB	99.99%	21,542	25,351	21,546	4
OQ112619	XBB	99.99%	21,529	25,338	21,534	6
OQ112624	XBB	99.99%	21,529	25,338	21,534	6
OQ112639	XBB	99.99%	21,529	25,338	21,534	6
OQ112647	XBB	99.99%	21,529	25,338	21,534	6
OQ112649	XBB	99.99%	21,529	25,338	21,534	6
OQ112673	XBB	99.99%	21,529	25,338	21,534	6
OQ112686	XBB	99.99%	21,529	25,338	21,534	6
OQ112691	XBB	99.99%	21,529	25,338	21,534	6
OQ112730	XBB	99.99%	21,529	25,338	21,534	6
OQ112737	XBB	99.99%	21,529	25,338	21,534	6
OQ112740	XBB	99.99%	21,529	25,338	21,534	6
OQ112771	XBB	99.99%	21,529	25,338	21,534	6
OQ112893	XBB	99.99%	21,545	25,354	21,546	2

In 25 out of the 39 sequences, the first sixth nucleotide in S, considered from the start of the complete genome, coincided with the third nucleotide of the second triplet, meaning the first element of the arithmetic series with a ratio of 6 from the start of the genome matched the first sixth nucleotide of S. This coincidence occurred in 64.10%. Out of these 25 sequences, 15 belonged to XBB.

13.4.4 Preliminary Identification of Visual Patterns in Arithmetic Sequences of 6 Nucleotides in S

According to what was exposed in Section 13.3.5, I performed a basic conversion, extracting nucleotides in groups of 6 and assigning a color code to each of these nucleotides in the 39 analyzed sequences. Adenine (A) is represented in green, Thymine (T) in blue, Cytosine (C) in yellow, and Guanine (G) in red.

Figure 13.1 shows the results from this conversion. Vertically, the Spike protein is depicted at intervals, frames, of 6 nucleotides. Upon initial observation, I discern a visual distinction between the arithmetic series 6, 12, ... and 9, 15, ... compared to the rest. Both series exhibit a prevalence of cool colors (T and A) not found in the others, along with a significant level of similarity between the two excerpts. The significant congruence between arithmetic series that are multiples of 3 compared to those that are not, reveals a striking and distinctive pattern, highlighting their remarkable homology.

At a glance, it is striking that the series 6, 12,... and the series 9, 15,... had a predominance of cool colors (T and A), especially blue (T). The color mixture was more uniform in the rest of the sequences, tending more towards warm colors (C and G). Therefore, at first sight, I observed a clear tendency towards a high content of T compared to that of A, C, and G.

The similarity of patterns among arithmetic series multiples of 3 in their positions was striking, compared to other patterns that were clearly different and more uniform among them in those mathematical series whose positions were not multiples of 3.

The initial conclusion I drew from observing this Figure 13.1, through the superposition of both time series, was that indeed, this differentiating pattern occurred in the time series corresponding to the third nucleotides, that is, the nucleotides occurring every 3 positions, with the first position being the first multiple of 3.

Without a specific way to know why the CNN focused on $f = 1/6$ in the S protein, I could at least grasp that the primary attention frequency motivating the recombinant feature, based on homology between patterns and nucleotide contents, was actually $f = 1/3$.

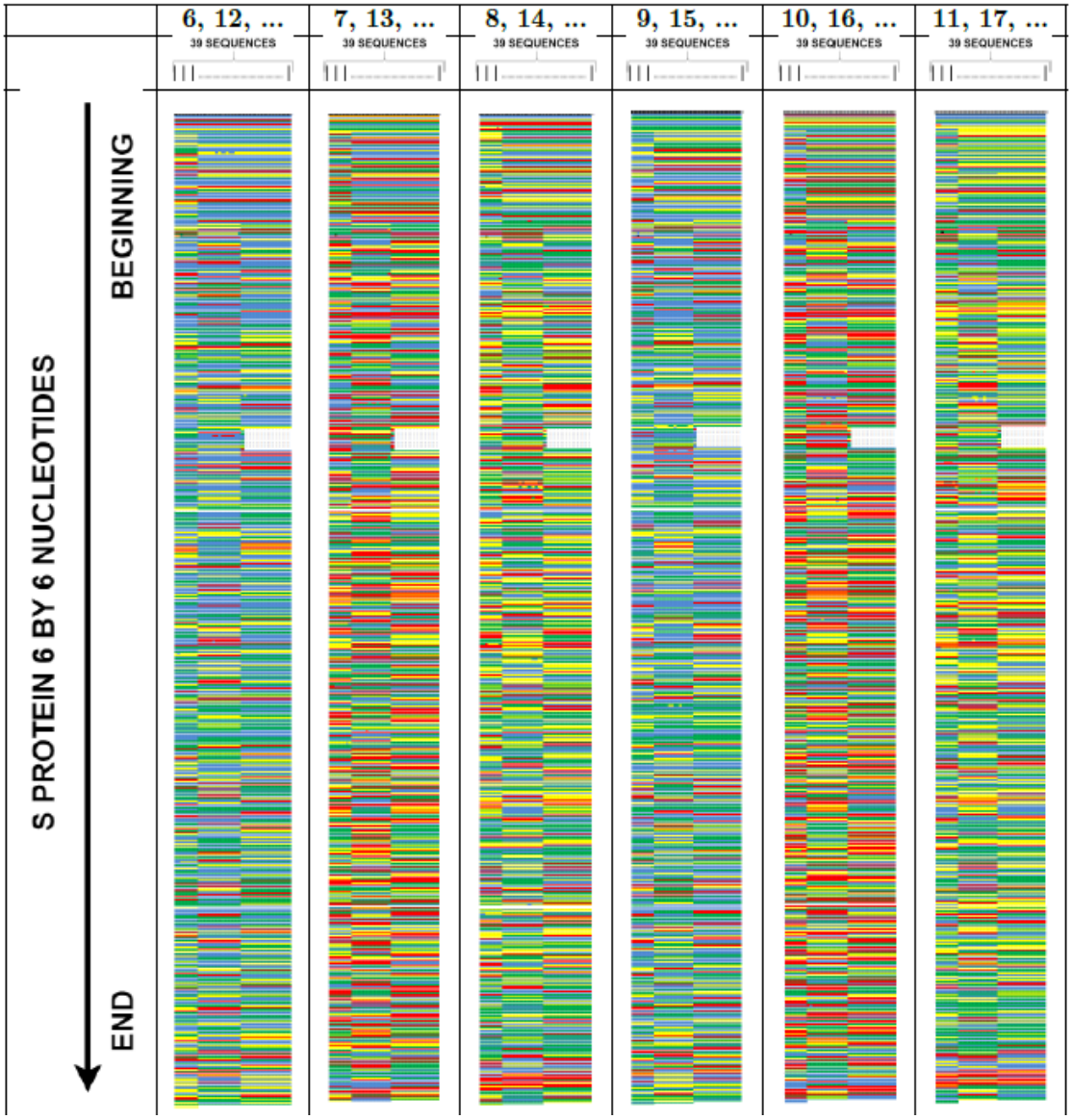


Figure 13.1: Graphic Representation of the Nucleotide Excerpt According to Arithmetic Series (Frames) with a Ratio of 6. The width of each image is divided into the 39 analyzed recombinant sequences. The height represents, from top (first element of the series) to bottom (last element of the series), the sixth nucleotides of the Spike protein. Each nucleotide is represented by a color (T = blue, A = green, C = yellow, G = red).

I could speculate about why the CNN looked at $f = 1/6$ instead of $f = 1/3$. One possible assumption is the existence of a simpler, more easily detectable pattern at $f = 1/6$. Therefore, the CNN tended to observe that area for performance efficiency reasons. However, this is a mere assumption that, unfortunately, I had no way to verify.

In any case, this first detected homology made me consider the hypothesis that $f = 1/6$ might be a 'harmonic' of $f = 1/3$, and that this mathematical signature could indeed be associated with events related to nucleotide periodicity in triplets.

13.4.5 Nucleotide Content

As outlined in Section 13.3.6, in light of the visual results in Figure 13.1, the next step was to calculate the nucleotide content corresponding to each of the arithmetic series, as well as the average content of each nucleotide type and its standard deviation.

I began with the 39 recombinant sequences in their S protein. In this case, analyzing the quantity of each nucleotide in each of the configurations (Table 13.6), both in series 6, 12,... and in 9, 15,..., indeed, the quantity of Thymine (T) was significantly higher.

The amount of Thymine in multiples of 3 arithmetic series rises above 45%. The remaining 55% is distributed among the other three types of nucleotides, with a slight predominance of adenine. In non-multiples of 3 series, the distribution percentages among the four types of nucleotides are more balanced. In all cases, the Standard Deviation (SD) of each nucleotide is less than 2%. This means that in the 39 analyzed sequences, nucleotides at each position are generally similar with minor variations.

So, there is a clearly distinct distribution pattern between the arithmetic series that are multiples of 3 and those that are not.

In the case of S protein of recombinant sequences, Appendix H displays the complete compilation of data regarding the presence of each type of nucleotide in each arithmetic series, along with complete statistical information. Table 13.6 provides a statistical summary of the relevant data by series and nucleotide type.

Table 13.6: SARS-CoV-2 Recombinant Sequences Study Case. Incidence of Each Nucleotide in the Frames Described in Figure 13.1. "No. Avg." refers to the average incidence of appearance of each nucleotide type in the S protein in each frame. "Avg. ± SD (%)" indicates "No. Avg", expressed as a percentage of the total number of nucleotides composing the S protein ± the standard deviation.

	6, 12, ...		**9, 15, ...**	
	No. Avg.	**Avg. ± SD (%)**	**No. Avg.**	**Avg. ± SD (%)**
A	169.54	26.62 ± 1.75	170.92	26.87 ± 1.96
T	293.87	46.13± 1.69	288.90	45.42 ± 1.51
C	99.95	15.69 ± 0.71	99.46	15.64 ± 0.74
G	66.28	10.41 ± 1.21	70.13	11.03 ± 1.14
	7, 13, ...		**10, 16, ...**	
	No. Avg.	**Avg. ± SD (%)**	**No. Avg.**	**Avg. ± SD (%)**
A	188.59	29.65 ± 1.05	195.18	30.69 ± 0.63
T	155.72	24.48 ± 0.53	150.18	23.61 ± 1.02
C	106.54	16.75 ± 1.37	100.41	15.79 ± 1.51
G	178.15	28.01 ± 1.82	183.28	28.82 ± 1.24
	8, 14, ...		**11, 17, ...**	
	No. Avg.	**Avg. ± SD (%)**	**No. Avg.**	**Avg. ± SD (%)**
A	193.79	30.47 ± 0.69	196.21	30.85 ± 0.25
T	182.69	28.73 ± 1.75	183.79	28.90 ± 1.45
C	150.33	23.64 ± 0.99	155.03	24.38 ± 0.77
G	102.59	16.13 ± 1.05	93.62	14.72± 1.42

In those arithmetic series that are multiples of 3, the presence of Thymine is significantly higher compared to the other three types of nucleotides combined. Thymine's presence alone accounts for almost half of the entire set. This phenomenon does not occur in sequences that are not multiples of 3.

In the context of genetics and molecular biology, the third letter of a codon refers to the position within a triplet of nucleotides (codon) in a DNA or RNA molecule that encodes a specific amino acid in a protein chain [2876].

The third letter of a codon is often less critical in determining the encoded amino acid compared to the first two letters of the codon. This is because the genetic code is degenerate, which means that multiple triplets can encode the same amino acid. The third letter of a codon is often flexible and can vary without altering the translated amino acid. This degeneracy in the third position of the codon is due to what is known as the "wobble hypothesis" [2877], proposed by

Francis Crick in 1966. According to this hypothesis, the third base of a codon (usually the outermost base) is the most flexible and can pair less strictly with the corresponding base in transfer RNA (tRNA) during translation. This allows for some variability in the third base without changing the amino acid incorporated into the protein chain. Mutations in the third base of a codon often may not have a significant effect on the final protein produced.

Performing the same analysis on the structural proteins M and N in the same sequences, I did not detect the pattern identified in S protein. Therefore, these phenomenon is only observed in the S protein, precisely where the CNN indicates the presence of the recombinant nature.

For recombinant sequences, the M protein, as outlined in Appendix I, presents a comprehensive compilation of data on the presence of each nucleotide type within each arithmetic series, along with detailed statistical information. Table 13.7 summarizes the relevant data by series and nucleotide type.

Table 13.7: SARS-CoV-2 Recombinant Sequences Study Case. Statistical Results Concerning Each Frame of 6 Nucleotides (M Protein). "No. Avg." refers to the average incidence of appearance of each nucleotide type in the M protein in each frame. "Avg. ± SD (%)" indicates "No. Avg", expressed as a percentage of the total number of nucleotides composing the M protein ± the standard deviation, also in %.

	6, 12, ...		**9, 15, ...**	
	No. Avg.	**Avg. ± SD (%)**	**No. Avg.**	**Avg. ± SD (%)**
A	22.97	20.70 ± 0.14	31.00	27.93 ± 0.00
T	36.56	32.94± 0.45	46.03	41.46 ± 0.14
C	28.44	25.62± 0.45	21.97	19.80 ± 0.14
G	23.03	20.74 ± 0.14	12.00	10.81 ± 0.00
	7, 13, ...		**10, 16, ...**	
	No. Avg.	**Avg. ± SD (%)**	**No. Avg.**	**Avg. ± SD (%)**
A	31.67	28.53 ± 0.47	35.00	31.82 ± 0.00
T	28.00	25.23 ± 0.00	21.00	19.09 ± 0.00
C	21.33	19.22 ± 0.42	27.00	24.55 ± 0.00
G	30.00	27.03 ± 0.29	27.00	24.55 ± 0.00
	8, 14, ...		**11, 17, ...**	
	No. Avg.	**Avg. ± SD (%)**	**No. Avg.**	**Avg. ± SD (%)**
A	30.95	27.88 ± 0.20	19.00	17.27 ± 0.00
T	37.00	33.33 ± 0.00	43.92	39.93 ± 0.25
C	24.00	21.62 ± 0.00	21.08	19.16 ± 0.25
G	19.05	17.16 ± 0.20	26.00	23.64 ± 0.00

In light of the results, the distribution patterns by nucleotide type do not match those detected in the case of the S Protein.

In the case of recombinant sequences, N protein, Appendix J displays the complete compilation of data regarding the presence of each type of nucleotide in each arithmetic series, along with complete statistical information. Table 13.8 provides a statistical summary of the relevant data by series and nucleotide type.

Table 13.8: SARS-CoV-2 Recombinant Sequences Study Case. Statistical Results Concerning Each Frame of 6 Nucleotides (N Protein). "No. Avg." refers to the average incidence of appearance of each nucleotide type in the N protein in each frame. "Avg. ± SD (%)" indicates "No. Avg", expressed as a percentage of the total number of nucleotides composing the N protein ± the standard deviation, also in %.

	6, 12, ...		**9, 15, ...**	
	No. Avg.	**Avg. ± SD (%)**	**No. Avg.**	**Avg. ± SD (%)**
A	61.54	29.53 ± 0.45	65.51	31.47 ± 0.49
T	71.69	34.42 ± 2.55	61.33	29.46 ± 2.37
C	40.74	19.54 ± 3.83	53.41	25.66 ± 3.77
G	34.38	16.51 ± 1.73	27.92	13.41 ± 1.88
	7, 13, ...		**10, 16, ...**	
	No. Avg.	**Avg. ± SD (%)**	**No. Avg.**	**Avg. ± SD (%)**
A	60.46	29.04 ± 2.50	68.85	33.21 ± 2.36
T	33.62	16.15 ± 0.49	28.54	13.76 ± 0.51
C	49.82	23.93 ± 0.23	52.00	25.08 ± 0.09
G	64.28	30.88 ± 1.81	57.97	27.95 ± 1.95
	8, 14, ...		**11, 17, ...**	
	No. Avg.	**Avg. ± SD (%)**	**No. Avg.**	**Avg. ± SD (%)**
A	70.26	33.75 ± 0.23	69.64	33.59 ± 0.50
T	36.49	17.53 ± 0.57	32.46	15.65 ± 0.41
C	57.18	27.47 ± 0.14	59.92	28.90 ± 0.20
G	44.26	21.26 ± 0.27	45.33	21.86 ± 0.27

In light of the results, the distribution patterns by nucleotide type do not match those detected in the case of the S Protein.

The occurrence rates of each nucleotide in the temporal series with a ratio of 6 in N and in M did not coincide with those detected in S. The percentage of predominance of T in the series with a multiple of 3 ratio did not reach such high levels with respect to the other nucleotide types in either of these two proteins.

Therefore, it appears that the predominance of over 45% of T was limited to the S protein, rather than other proteins.

The next step was to verify the existence of this pattern in non-recombinant sequences correctly classified by the CNN, where their hot zones did not include $f = 1/6$ in S. For the selected non-recombinant sequences, the S protein, as detailed in Appendix K, presents a thorough compilation of data concerning the presence of each nucleotide type within each arithmetic series, along with comprehensive statistical information. Table 13.9 offers a statistical overview of the pertinent data by series and nucleotide type.

Table 13.9: SARS-CoV-2 Non-Recombinant Sequences Study Case. Incidence of Each Nucleotide in Each Series of 6 Nucleotides (S Protein). "No. Avg." refers to the average incidence of appearance of each nucleotide type in the S protein in each frame. "Avg. ± SD (%)" indicates "No. Avg", expressed as a percentage of the total number of nucleotides composing the S protein ± the standard deviation, also in %.

	6, 12, ...		**9, 15, ...**	
	No. Avg.	**Avg. ± SD (%)**	**No. Avg.**	**Avg. ± SD (%)**
A	167,05	26.59 ± 1.91	173.59	27.66 ± 1.96
T	290.74	46.29± 1.50	290.00	46.18 ± 1.69
C	99.15	15.78 ± 0.87	100.49	16.02 ± 0.88
G	71.21	11.35 ± 1.33	63.62	10.14 ± 1.17
	7, 13, ...		**10, 16, ...**	
	No. Avg.	**Avg. ± SD (%)**	**No. Avg.**	**Avg. ± SD (%)**
A	196.69	31.33 ± 1.03	185.74	29.62 ± 0.93
T	146.26	23.30 ± 0.91	157.11	25.05 ± 0.93
C	100.74	16.04 ± 1.30	105.85	16.89 ± 1.38
G	184.05	29.33 ± 1.37	178.38	28.44 ± 1.36
	8, 14, ...		**11, 17, ...**	
	No. Avg.	**Avg. ± SD (%)**	**No. Avg.**	**Avg. ± SD (%)**
A	196.67	31.32 ± 0.70	191.44	30.52 ± 0.70
T	182.00	29.00 ± 1.67	184.36	29.39 ± 1.49
C	151.97	24.20 ± 0.97	151.59	24.18 ± 0.80
G	97.21	15.48 ± 1.41	99.72	15.91 ± 1.29

Similarly to the case of the 39 recombinant sequences, the amount of T in the series whose initial position is a multiple of 3 turned out to be around 45%. The patterns between the data corresponding to the recombinant and non-recombinant sequences were similar, suggesting that the high quantity of

thymine in the third nucleotides in S is a broadly independent characteristic of the sequence, in both study groups

Table 13.10 shows the results of the comparison in the average occurrence values of each nucleotide type in S, in % across all arithmetic series in both study groups.

Table 13.10: Comparison of Average Nucleotide Contents between Recombinant and Non-Recombinant Sequences. The column "Rec. Avg. (%)" collects the average percentage value of each nucleotide type in each arithmetic series for the 39 recombinant sequences. Similarly, "Non-Rec Avg. (%)" in the case of the 39 non-recombinant sequences. The "Deviation" column indicates the difference between the two previous columns.

	6, 12, ...			**9, 15, ...**		
	Rec. Avg.(%)	**Non-rec. Avg.(%)**	**Deviation**	**Rec. Avg.(%)**	**Non-rec. Avg.(%)**	**Deviation**
A	26.62	26.59	0.03	26.87	27.66	-0.79
T	46.13	46.29	-0.16	45.42	46.18	-0.76
C	15.69	15.78	-0.09	15.64	16.02	-0.38
G	10.41	11.35	-0.94	11.03	10.14	0.89
	7, 13, ...			**10, 16, ...**		
	Rec. Avg.(%)	**Non-rec. Avg.(%)**	**Deviation**	**Rec. Avg.(%)**	**Non-rec. Avg.(%)**	**Deviation**
A	29.65	31.33	-1.68	30.69	29.62	1.07
T	24.48	23.30	1.18	23.61	25.05	-1.44
C	16.75	16.04	0.71	15.79	16.89	-1.10
G	28.01	29.33	-1.32	28.82	28.44	0.38
	8, 14, ...			**11, 17, ...**		
	Rec. Avg.(%)	**Non-rec. Avg.(%)**	**Deviation**	**Rec. Avg.(%)**	**Non-rec. Avg.(%)**	**Deviation**
A	30.47	31.32	-0.85	30.85	30.52	0.33
T	28.73	29.00	-0.27	28.90	29.39	-0.49
C	23.64	24.20	-0.56	24.38	24.18	0.20
G	16.13	15.48	0.65	14.72	15.91	-1.19

I did not detect any significant difference between the recombinant and non-recombinant groups in terms of nucleotide occurrence percentages. All arithmetic series are similar in this regard. Therefore, in both groups, it happened that in the S protein, the content of Thymine was notably higher compared to the rest of nucleotide types. Therefore, this was a common characteristic in both analyzed groups and did not represent a distinctive element solely of the recombinants.

Between both groups, the most similar time series was 6, 12,... That is, the

deviations between their nucleotide contents were generally lower. The series 9, 15,... was slightly more variable. Perhaps this similarity in 6, 12,... in both groups holds some kind of relationship with the tendency shown by the CNN to look at $f = 1/6$ instead of $f = 1/3$ to classify a sequence as recombinant.

13.4.6 Common Nucleotides

Following the methods outlined in Section 13.3.7, the common nucleotides in the case of the 39 recombinant sequences were identified in Table 13.11.

Table 13.11: Common Nucleotides of S Gene in the 39 Analyzed Recombinant Sequences. The column "Position" indicates the position of the common nucleotide, starting from the first nucleotide of S.

Position	Nucleotide
3	G
6	T
9	T
12	T
15	T
18	T
21	A
24	G
27	A
30	A
36	T
39	T
42	G
45	T
48	T
51	T
54	T
57	A
63	A
66	T
69	A
72	A
78	T
93	T
102	T
108	T
117	T
120	C

Table 13.11 – Continuation from the previous page

Position	**Nucleotide**
123	A
165	C
177	T
189	T
198	T
318	T
363	T
369	T
399	T
405	T
582	T
1329	T
2382	T

Surprisingly, all the common nucleotides in those 39 recombinant sequences were located at positions that were multiples of 3. In other words, all of them were the third nucleotide of the triplet. This fact is fascinating because, out of the 313 recombinant sequences compiled in the dataset, the selected 39 sequences had all the common elements precisely as third nucleotides. There was not a single common nucleotide at positions 1 and 2 of the triplets.

This third nucleotide can tolerate changes without necessarily affecting the amino acid that is incorporated. This phenomenon is known as genetic code degeneracy, where several codons can encode the same amino acid because differences in the third position of the codon often do not alter the amino acid sequence in the final protein. This third codon is more susceptible to mutations compared to the first two positions due to this effect of genetic code degeneracy [2878].

In logic, one would therefore expect greater variability in these third codon positions among the 39 sequences (and consequently, less variability in the first two positions). However, in this study, the opposite phenomenon occurred.

For the identification of the common nucleotides, I relied on Figure 13.2, created by Zhang et al. [2631]. In this figure, the sub-elements composing the complete Spike protein in the genome of Wuhan-Hu-1 NC_045512.2 [2590] are represented, expressed in the 1273 amino acid residues it consists of.

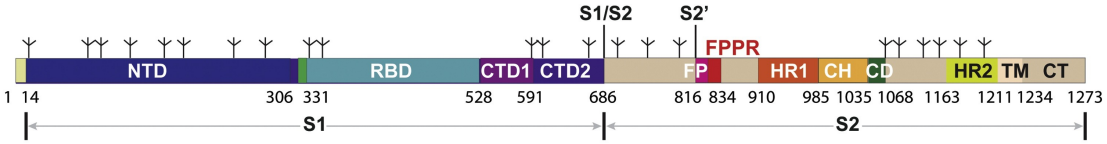


Figure 13.2: Schematic Representation of the SARS-CoV-2 Spike Protein Sub-elements. Image extracted from Zhang et al. [2631].

The conversion from amino acid positioning to nucleotide positions was simply done by multiplying the amino acid position by 3, since each amino acid is coded by three nucleotides.

The common elements from nucleotide 1 to 42 are framed within the initial sub-element of S up to the start of NTD. The starting pattern in this fragment in the common nucleotides is GTTTTTAGAATTG. The common nucleotides from position 45 to position 582 are located in the Spike region corresponding to the N-Terminal Domain (NTD), which ends at nucleotide 918. The common nucleotide at position 1329 is located in the central area of RBD, delimited from nucleotide 993 to nucleotide 1584. The common nucleotide at position 2382 is located in the last third of the genomic region from S1/S2 (nucleotide 2058) to S2' (nucleotide 2448), approximately.

I highlighted the predominant presence of T in the common nucleotides, accounting for 68.29% compared to 19.51% of A, 7.32% of G, and 4.88% of C. The common nucleotides framed in NTD from position 177 to position 582 were all T.

The common nucleotide located in RBD was T, as well as the one located in the genomic region from S1/S2 to S2'.

This phenomenon requires in-depth investigation.

Table 13.12 gathers the common nucleotides in the 39 non-recombinant sequences under study.

Table 13.12: Common Nucleotides of S Gene in the 39 Analyzed Non-Recombinant Sequences. The column "Position" indicates the position of the common nucleotide, starting from the first nucleotide of S.

Position	Nucleotide
3	G
6	T
9	T
12	T
15	T
18	T
24	G
27	A
42	G
45	T
48	T
57	A
189	T
198	T
2388	T

Similar to the case of the 39 recombinant sequences, it was striking that, once again, all the common elements in the S proteins of the 39 non-recombinant sequences analyzed were multiples of 3, the third nucleotides of the codons.

The 39 selected sequences belong to very diverse lineages (Table 13.3), with a greater number of insertions, deletions, and mutations given the greater phylogenetic distance between these lineages. These facts lead to the expectation of greater variability in the group of non-recombinant sequences vs. the group of recombinant sequences.

Indeed, the number of common nucleotides was lower than in the 39 recombinant sequences.

It is striking once again that in 39 sequences chosen from the total of 1,539,728 non-recombinant sequences, common elements emerged and that all of them were multiples of 3.

Furthermore, some of these common elements were also found in the recombinant sequences. All these preliminary findings are fascinating and statistically surprising.

According to the scheme in Figure 13.2, the common elements from nucleotide 1 to 42 are framed within the initial sub-element of S up to the start of NTD. The starting pattern in this fragment in the common nucleotides is GTTTTTGAG. The common nucleotides from position 45 to position 198 are located in NTD, which ends at nucleotide 918. The common nucleotide at position 2388 is located in the last third of the genomic region from S1/S2 (nucleotide 2058) to S2' (nucleotide 2448), approximately.

Similarly to the case of the recombinant sequences, I highlighted the predominant presence of T in the common nucleotides, with 66.67%, a percentage very similar to that of the recombinant sequences. The content of A represented 13.34%, and that of G was 20%. No common elements of type C were detected.

Similarly to the case of the recombinants, the common nucleotide located in the genomic region from S1/S2 to S2' was T; however, in the non-recombinants, it was located at position 2388, whereas in the recombinants, it was located at position 2382, two triplets earlier. This common element is peculiar in terms of its location distant from the rest of the common elements in both cases. It is surprising that, being T in both cases, in the recombinant group it was situated two triplets earlier than in the non-recombinant group. This phenomenon requires further research.

The common nucleotides shared between the 39 recombinant sequences and the 39 non-recombinant sequences are listed in Table 13.13.

Table 13.13: Common Nucleotides in the S Protein Sequences, Shared between Recombinants and Non-Recombinants. All shared elements are identified by their nucleotide position and the associated nucleotide type.

Position	**Nucleotide**
3	G
6	T
9	T
12	T
15	T
18	T
24	G
27	A
42	G
45	T
48	T
57	A
189	T
198	T

All shared nucleotides between recombinants and non-recombinants were located in NTD.

All common nucleotides in the case of non-recombinant sequences were also common in the recombinants, except for the nucleotide located at position 2388. This is the only exclusive common nucleotide of the non-recombinants.

The recombinant sequences, however, contained multiple common nucleotides that were specific to these recombinant sequences, and the are shown in Table 13.14.

Table 13.14: Common Nucleotides in the S Protein Sequences, Specific to Recombinants. All elements are identified by their nucleotide position and the type of nucleotide contained in it.

Position	Nucleotide
21	A
30	A
36	T
39	T
51	T
54	T
63	A
66	T
69	A
72	A
78	T
93	T
102	T
108	T
117	T
120	C
123	A
165	C
177	T
318	T
363	T
369	T
399	T
405	T
582	T
1329	T
2382	T

The common nucleotides from position 318 to position 582 were exclusive to the recombinants. They seem to be concentrated in a specific genomic region with a width of only 264 nucleotides, corresponding to 88 amino acids. This area appeared to be a hot spot for common elements in recombinants, unlike in non-recombinants, where no common nucleotide was detected in this fragment of S, corresponding to the central region of NTD [2631].

The NTD is a structural region of the Spike protein found at its N-terminal end that plays a significant role in virus binding to host cells and viral entry.

The overall structure of the NTD in SARS-CoV-2 is similar to that of the Spike proteins of Middle East respiratory syndrome coronavirus (MERS-CoV) [2879] and bovine coronavirus [2880].

The exact function of the NTD in the SARS-CoV-2 Spike protein remains unknown, although it has been shown that the NTDs of other coronaviruses recognize sugars in the initial stage of binding to specific protein receptors, or play a role in the transition from pre-fusion to post-fusion. However, neutralizing antibodies (nAbs) targeting the NTD have been isolated from patients infected with SARS-CoV-2, with potency at the nanomolar level, suggesting a critically functional role for this domain [2881]. Most antibodies target the region called NTD-1, which is therefore referred to as the superantigenic site NTD-1 [2882, 2883]. It is located at the edge of the upper-central region of the NTD, and it includes five surface loops that rearrange upon binding to different antibodies [2631]. The region called NTD-2 is recognized by non-neutralizing antibodies. The Alpha (lineage B.1.1.7), Beta (B.1.351), Gamma (B.1.1.28), and Delta (B.1.617.2) variants all have mutations and/or deletions within the superantigen NTD-1, which confers resistance to nAbs targeting the NTD [2884, 2885].

This presence of mutations and deletions in this area makes it even more exciting that, despite this, there are multiple common nucleotides precisely in such a poorly conserved region.

It seems plausible that the mathematical signature or some related factor to it is located in this area and related to the specific common elements in the group of 39 recombinant sequences. The high load of common nucleotides detected in the NTD in the case of the recombinant group requires further investigation to determine if there is any relationship between the recombinant characteristic and undiscovered functions expressed in this genomic region.

The results obtained continued to be fascinating at this point.

13.4.7 Microbiological Analysis

According to Section 13.3.8.

In those arithmetic series that are multiples of 3, the presence of Thymine was significantly higher compared to the other three types of nucleotides combined. Thymine's presence alone accounted for almost half of the entire set. This phenomenon did not occur in sequences that are not multiples of 3.

Hassanin (2022) already found that SARS-CoV2 related coronaviruses genomes contain less cytosines and more uracils at third codon-position than other sarbecoviruses [2886], in agreement with Rice et al. [2887] and Tort et al. [2888], whose reports pointed out a SARS-CoV-2 codon usage bias derived from mutations driving an A+U or U enrichment, among other factors. Saldivar-Espinoza et al. [2889] also reported that C>U mutations are the most prevalent in SARS-CoV2 genome, and suggested that host deaminases, deaminating unpaired nitrogenous bases, were responsible for a high proportion of mutations.

In the context of genetics and molecular biology, the third letter of a codon is often less critical in determining the encoded amino acid compared to the first two letters of the codon, as already pointed out by Francis Crick with his "wobble hypothesis" in 1966 [2877]. In this line, according to the human genetic code [2890], the third codon nucleotide can be C or U, without changing the encoded amino acid; therefore, if the high frequency of U in the third codon position derives from a C>U transition, it would not alter the amino acid sequence of the resulting Spike protein.

Nevertheless, it is interesting to highlight that in my analysis, other structural proteins such as M and N did not show the same frequency distribution pattern as for S. Therefore, I understand that the higher frequency of U at a certain position may have some biological significance that would require further studies for interpretation.

Although these phenomena do not directly explain the relationship between the detected mathematical pattern that classifies a sequence of SARS-CoV-2 as recombinant and the percentages of appearance of Uracil (Thymine) in the S protein, I cannot rule out that there is some relationship between them.

A possible additional factor that may influence the role played by the third nucleotide in the recombinant feature is related to the limitation of human DNA itself in recognizing information to generate transfer RNA. Without the necessary information to generate transfer RNA, the human ribosome cannot generate the corresponding amino acid. This happens, for example, in the case of CGC (Arginine), where the human ribosome does not know how to encode it. However, it does happen in the case of CGG, CGA, and CGU, where it is capable of correctly encoding this triplet as arginine. In the research by Zabolotskii et al. [2891], it is explained that the human ribosome, like ribosomes in other organisms, has limitations in its ability to recognize and translate all codons present in messenger RNA. This selectivity is determined by factors such as the availability of corresponding transfer RNA and codon recognition mechanisms. Therefore, there are codons that may not be efficiently recognized by the ribosome, which can affect the efficiency and accuracy of protein synthesis.

In situations where the human ribosome encounters a codon it cannot efficiently recognize, it may opt for an alternative codon that is similar in its sequence but still fulfills the required function in mRNA translation (known as "wobbling" or "genetic code degeneracy") [2892], where certain codons can be exchanged for others that encode the same amino acid. However, the efficiency and accuracy of this process may vary depending on the availability of corresponding transfer RNAs and specific translation contexts. This is the case with the substitution of U for C, which often encodes the same amino acid without altering the protein.

These limitations may represent an evolutionary advantage for the virus by introducing mutations in human ribosome intervention if the result is viable and provides a competitive advantage over other variants.

The analysis of these factors in transcription should be studied deeply.

13.5 Conclusions

From my perspective, this Chapter has been the most exciting and promising.

In summary, in Chapter 11, I detected a mathematical signature that characterizes a genomic sequence of SARS-CoV-2 as recombinant. This signature was located in the S protein, with its epicenter clearly situated at a frequency of $f = 1/6$.

The prominent hot zone in this protein, which is highly relevant in terms of viral evolution, prompted me to analyze this area in greater detail in order to extract more information about the inherent mathematical signature of the recombinant feature. Starting from a cursory visual analysis of the arithmetic series of nucleotides every 6 positions, I identified a specific pattern in the arithmetic series whose initial position was a multiple of 3, which did not occur in the rest of the mathematical series. Consequently, I inferred the study of the third nucleotides of triplets in S in complete sequences of SARS-CoV-2 as an area of interest. I selected all recombinant sequences from this set whose hot zone was exclusively $f = 1/6$ in the S protein. Out of the 313 sequences, 39 met this criterion. Subsequently, I selected an equal number of non-recombinant sequences from the test set whose hot zones did not include $f = 1/6$ either primarily or secondarily.

Once the sequences under study were selected, I extracted the third nucleotides from the S, M, and N proteins in the recombinant sequences to compare the patterns that could emerge in S with other structural proteins exhibiting sufficient variability.

I also compared the patterns in the third nucleotides of S in the recombinant sequences with their non-recombinant counterparts to check for possible differences between sequences in one category and the other.

I detected a high thymine content compared to the rest of the nucleotide types in the third nucleotides in S, exceeding 45%. I did not observe this phenomenon in M or N in the selected 39 recombinant sequences. The third nucleotides of the triplet in S in the non-recombinant sequences showed similar amounts of thymine, as in the case of the recombinants.

Finally, I determined the common nucleotides (content and position within S) in the 39 recombinant sequences and in the 39 selected non-recombinant sequences. Understanding that those common elements between one group and the other, as

well as their differences, could provide information about the inherent mathematical signature that determines the recombinant feature.

Surprisingly, I was able to confirm that, both in the recombinant and non-recombinant sequences, all their common elements were multiples of 3. That is, their common elements were third nucleotides of the triplet. Considering the high variability of S in terms of length and nucleotide content (see Chapter 11), this fact is especially relevant and motivating.

From a statistical point of view, attributing this coincidence to mere chance is complex. I selected 39 sequences out of a total of 1,565 recombinant sequences and 39 non-recombinant sequences out of a total of 15,650. That is, 78 selected sequences out of a total of 17,215. And if we consider the total number of sequences before the subsampling done in Chapter 11, there are a total of 1,541,293 recombinant and non-recombinant sequences in total. The S protein showed high variability in terms of nucleotide content (see Chapter 10), with only 0.23% of common nucleotides.

In my analysis presented in this chapter, all these nucleotides were multiples of 3, meaning the third nucleotide of the triplet. Additionally, due to the degeneracy of the genetic code (where several codons can encode for the same amino acid because differences in the third position of the codon often do not alter the amino acid sequence in the final protein) and perhaps due to the limitations of the human ribosome in correctly interpreting all codons, it may be more prone to mutation without necessarily resulting in a modification in the resulting protein. However, in certain circumstances, it may lead to such modifications, and if viable and advantageous from an evolutionary standpoint, it could give rise to new subvariants, variants, or even lineages.

This search for common nucleotides was conducted without applying any type of multiple alignment. In other words, in sequences of varying lengths in S, I found that the common elements were precisely the third nucleotides of the triplet without performing any prior "artificial" adjustment before comparison.

Finally, after identifying the common elements in both categories, I determined which elements are common to recombinant and non-recombinant sequences to try

to shed some light on that mathematical signature encrypted in the genome. All the evidence found suggests that this encrypted mathematical signature is located in the third nucleotide of the triplets in S (in the case of SARS-CoV-2), reinforcing the hypothesis that this third nucleotide, less relevant in terms of amino acid coding, may contain additional information in an encrypted manner that characterizes a SARS-CoV-2 sequence as recombinant.

There is still much to learn about the intricate rules contained within the genome, but I hope that the content of this document marks the beginning of a new way of viewing and analyzing the genome. In light of the results obtained, the identification of mathematical signatures in the virus genome through genomic spectrogram analysis opens up new avenues to investigate potential functions associated with these mathematical patterns.

I consider this chapter the culmination of a precious and fascinating research from start to finish.

The results presented in this chapter not only provide insight into what could be the beginning of the mathematical signature embedded in the genome that determines a sequence as recombinant. The mathematical signature that determines whether a sequence is recombinant or not is intricately linked to the third nucleotides of the triplet.

I found sufficient evidence to hypothesize an unknown role in the third nucleotide of the triplet. While the third nucleotide of the triplet is of little relevance from the perspective of amino acid coding, my results support the hypothesis that this third codon nucleotide contains encrypted information that can provide relevant insights into genome characterization, specifically whether a sequence of SARS-CoV-2 is recombinant or not.

The evidence presented, although not conclusive and requiring further extensive research to establish more robust conclusions, is extremely promising.

Perhaps within the genome, there are layers of encrypted information that must be interpreted according to specific codes. And maybe, that third nucleotide is not as trivial as it seems in terms of amino acid coding. Maybe the exact presence of

certain types of amino acids related to this third nucleotide contains or determines information regarding various organism characteristics.

13.6 Future Research

So far, I have been able to make a mere approximation to that mathematical signature for the purpose of formulating a hypothesis with some support. My findings were exciting and intriguing. But they are not sufficient. My next step should be to delve deeper into the unequivocal detection of the mathematical signature hidden in the third nucleotides of the triplets in the S protein of SARS-CoV-2.

To this end, the first step should be to replicate the tests performed in this chapter on a massive scale, preferably on all available complete SARS-CoV-2 sequences to date. In this way, we can verify if the detected patterns remain as the sample size increases, or if they are modified and in what way.

It would also be necessary to try to establish relationships or associations between common elements and variants, both recombinant and non-recombinant, in an attempt to extract inherent characteristics for each of them, if possible.

And, of course, we must investigate the biological significance of this mathematical signature embedded in the genome.

Considering that I demonstrated the existence of a mathematical signature in the S protein that determines the recombinant feature in SARS-CoV-2, and that I extracted enough evidence to relate it to the third nucleotide of the triplet, we must necessarily continue with what has been discovered so far. And perhaps, we can uncover the information encoded in that seemingly less important third nucleotide.

We may discover a new paradigm of reading and interpreting the genome.

Undoubtedly, this is the beginning of an exciting journey...

Y así, del poco dormir y del mucho leer, se le secó el cerebro.

— Miguel de Cervantes ("El Quijote"), *su tartamudez es un un divertimento para el cotilla. Sin embargo, nunca nadie ha escrito condensando tantos sentimientos simultáneamente en tan pocas palabras.*

14

Is There a Role Associated with the Third Codon Nucleotide in the Characterization of a Genomic Sequence? The Case of HIV-1 and SARS-CoV-2

Contents

14.1 Introduction 694
14.2 Compilation of Evidence 696
14.2.1 Recombinant Feature Hot Zones HIV-1 and SARS-CoV-2 (Superposed spectrograms) 696
14.2.2 Genetic Recombination in HIV-1 698
14.2.3 HIV-1 Hot Zones per Subtype 700
14.2.4 Genetic Recombination in SARS-CoV-2 709
14.2.5 SARS-CoV-2 Variant Hot Zones 710
14.3 Microbiological and Genomic Analysis 712
14.3.1 Differences between HIV-1 and SARS-CoV-2 712
14.3.2 The Role of Nucleotides A and T in Periodicities of 3 and their Multiples 715
14.3.3 LTRs of HIV-1 718
14.3.4 Differences between the LTRs in HIV-1 and the UTRs in SARS-CoV-2 720
14.3.5 The Spike Protein in SARS-CoV-2 722
14.4 Hypotheses 726

14.1 Introduction

In Chapter 4 (Genomic Spectrograms), I noted, in accordance with the existing literature, the horizontal line at $f = 1/3$ in the coding regions and its relation to the fact that in these coding regions, nucleotides are organized in triplets, with each triplet encoding an amino acid.

In Chapter 5 (HIV-1 Genetic Recombination Analysis), I explained the detection of a mathematical signature indicative of the recombinant feature in HIV-1. This signature was primarily located at $f = 1/3$ in the 5' UTR and possibly around the first half of *gag*.

In Chapter 6 (Analysis of Total Hot Zones in HIV-1 by Subtype), I delved into the hot zones associated with significant recombinant and non-recombinant subtypes of HIV-1 and their relation to their category equivalents, as well as the peculiarities of these subtypes as highlighted by the scientific community. Periodicities of 3 and their multiples were recurrent throughout this analysis.

In Chapter 8 (Analysis of HIV-1 Subtype U), I presented a detailed analysis of sequences classified as U with respect to the expert literature, examining the recombinant nature of each sequence in detail and possibly detecting hidden recombination phenomena in some sequences of this subtype. Once again, periodicities related to 3 were involved in multiple case studies.

In Chapter 9 (Analysis of HIV-1 Subtype G), I developed the analysis of the possible recombinant origin affecting subtype G based on the analysis of the hot zones of the sequences belonging to this subtype. The relationship between the periodicity 3 and the recombinant feature in HIV-1 allowed for a deeper exploration of additional information on this recombinant origin.

In Chapter 11 (SARS-CoV-2 Genetic Recombination Analysis), I explained the detection of a distinct mathematical signature at $f = 1/6$ for S protein that determines whether a SARS-CoV-2 sequence is recombinant. By delving into the detection of patterns related to this signature, I was able to find remarkable evidence that this signature is related to the third nucleotides of the triplet in that protein.

All these findings are detailed in Chapter 13 (Exploring the Influence of the Third Codon Nucleotide on Genetic Recombination in the SARS-CoV-2 Spike Protein).

I delved into the identification of mathematical signatures associated with genetic recombination in two viruses as different as HIV-1 and SARS-CoV-2. In both, I detected mathematical signatures associated with the recombinant feature. But not only that. In Chapter 12 (Analysis of SARS-CoV-2 Non-recombinant Variants), I adapted the methodology for detecting the recombinant feature to classify non-recombinant sequences into the pre-VOC, Alpha, Delta, and Omicron categories. This yielded notably positive results with specific mathematical signatures associated with each of these variants, with a clear presence of $f = 1/3$ among their associated hot zones.

In all cases, I found a relationship with the frequency $f = 1/3$, either directly or indirectly ($f = 1/6$ and $f = 1/9$).

Given that frequency is the inverse of the period, understood in this case as nucleotide periodicity, the equivalent period of $f = 1/3$ is 3. In the coding regions of the genome, where three nucleotides encode an amino acid, this periodicity of 3 is equivalent to the third codon nucleotides. For $f = 1/6$, the nucleotides related to this frequency will be the third codon nucleotides every 2 amino acids. In the case of $f = 1/9$, the nucleotides involved will be the third codon nucleotides every 3 amino acids.

Now I would like to pose the big question that will be a future research line as exciting as or even more than the one already conducted.

Is there a role associated with the third codon nucleotide in the characterization of a genomic sequence?

The purpose of this Chapter is an analysis, more reflective than quantitative, that can serve as guidance or inspiration for future research focused on determining the degree of universality of the role of the third nucleotide in the genetic recombination of viruses.

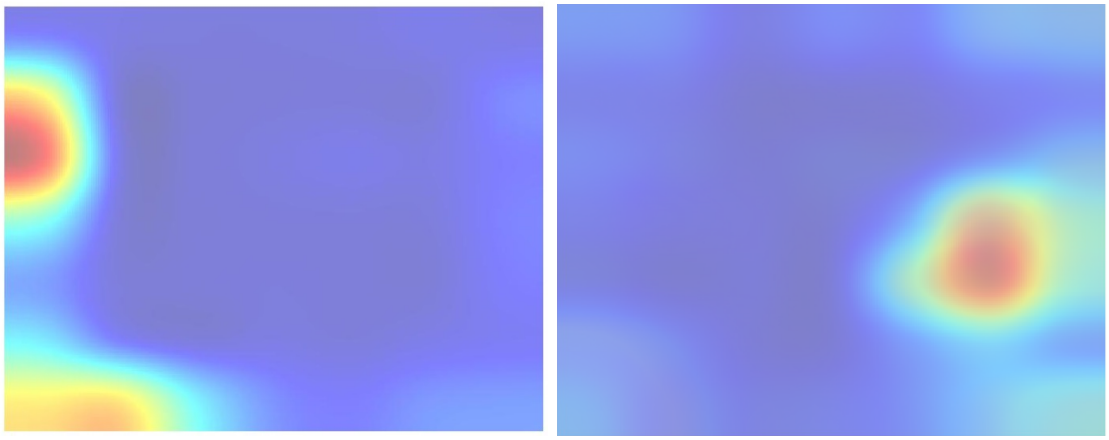
This Chapter consists of a section where I compiled the evidence presented throughout this document (Section 14.2). Subsequently, in Section 14.3 I present the results of the correlation between the evidence presented in this Chapter and the

scientific findings in the field of microbiology and virology. The Chapter concludes with the proposal of a hypothesis based on the information presented (Section 14.4).

14.2 Compilation of Evidence

14.2.1 Recombinant Feature Hot Zones HIV-1 and SARS-CoV-2 (Superposed spectrograms)

In a preliminary comparison between the total hot zones of the recombinant feature in HIV-1 and SARS-CoV-2, in superposed spectrograms (Figure 14.1), the role of periodicities of 3 or their multiples is relevant at first glance.



(a) HIV-1 Recombinant Total Hot Zones **(b)** SARS-CoV-2 Recombinant Total Hot Zones

Figure 14.1: Total Hot Zones Recombinant Feature of Superposed Spectrograms for HIV-1 and SARS-CoV-2. The detailed information of these active zones is available in Chapter 6, Section 6.6.2 for HIV-1, and in Chapter 11, Section 11.4.2 for SARS-CoV-2.

In superposed spectrograms, the length of the x-axis coincides with the length of the entire genome. In the case of HIV-1, the range of the x-axis is around 10,000 nucleotides, and in the case of SARS-CoV-2, it is around 30,000 nucleotides. The frequency range of the y-axis is 0.5 Hz. By direct proportionality, in the case of HIV-1, the most powerful main hot zone for the recombinant feature corresponded to a periodicity of 3 (i.e., $f = 1/3$) (Figure 14.1a). In the case of SARS-CoV-2 (Figure 14.1b), direct proportionality with respect to the upper limit on the y-axis (0.5 Hz) determined that the clear main hot zone that determines a SARS-CoV-2 sequence to be recombinant was located at $f = 1/6$, i.e., a periodicity of 6.

The periodicities of 3 and/or their multiples were indeed determinants in the recombinant feature in two different viruses, belonging to phylogenetically distant families.

The common ancestor between the *Retroviridae* family (to which HIV-1 belongs) and the *Coronaviridae* family (to which SARS-CoV-2 belongs) is of ancestral RNA. Their phylogenetic relationship is extremely distant [2893], this is, their evolutionary paths likely diverged hundreds of millions of years ago [2894–2896]. The origin of retroviruses was estimated to occur during the early Palaeozoic Era, between 460 and 550 million years ago [2897]. This moment is long before the appearance of mammals [2898].

All coronaviruses originated from a common ancestor about 293 million years ago [2899], this is, in the Late Permian [2900]. In the Late Permian, although mammals did not exist in their modern form, it is believed that there were ancestors of primitive mammals. These proto-mammals belonged to groups such as therapsids, which eventually gave rise to modern mammals during later periods, such as the Triassic. Therapsids were synapsid reptiles that exhibited characteristics of both reptiles and mammals and played a significant role in mammalian evolution [2901].

Despite their phylogenetic distance, in both viruses, the mathematical signature characterizing a sequence as recombinant maintains a close and clear relationship with the third codon nucleotides in specific genomic regions.

I find it remarkable that two viruses belonging to families that diverged during a time long before the appearance of mammals in their modern form [2902], exhibited similar evident signs in their patterns related to the mathematical signature determining the recombinant feature. These common signs were associated with a possible role of the third codon nucleotides.

The subsequent questions are: since when are these patterns in the third nucleotide imbued in the viral genome? Where does this encrypted information come from and for what reason?

The search for answers at this point is absolutely thrilling.

14.2.2 Genetic Recombination in HIV-1

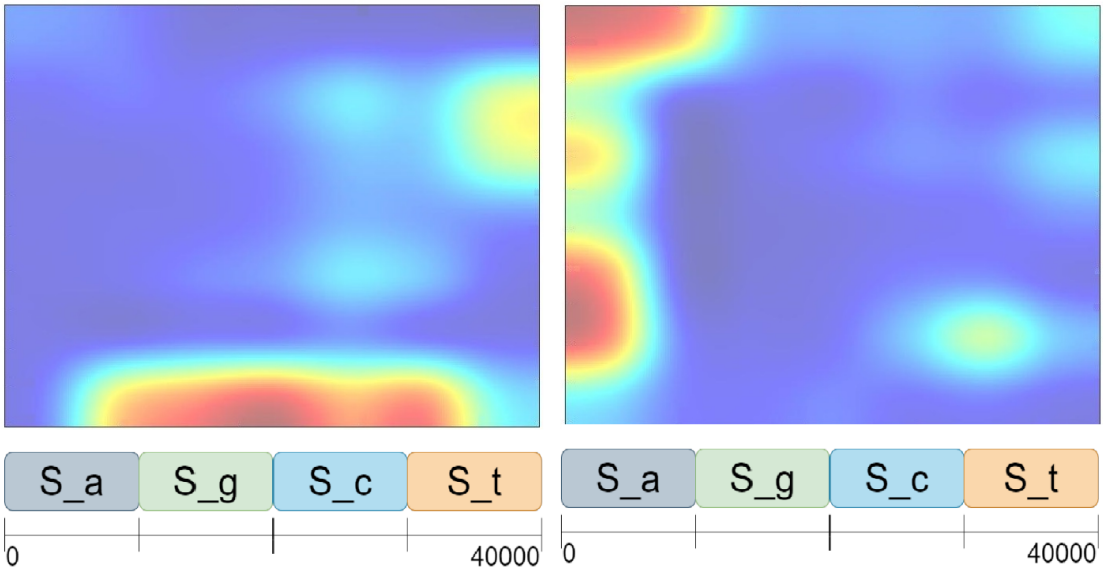
The small size of the HIV-1 genome (around 10,000 nts) allowed for a very visual approach to working with concatenated genomes, as well as superposed genomes. Both representations of genomic spectrograms proved to be complementary.

In the case of concatenated spectrograms, the spectrograms of each nucleotide type were concatenated in a single image, allowing us to obtain information about the types of nucleotides relevant to the recombinant feature. The total hot zones per category are shown in Figure 14.2.

As seen in Figure 14.2a, in the non-recombinant category, there was a clear secondary hot zone at $f = 1/3$ in the second half of the T spectrogram. That is, a mathematical signature related to every third nucleotide was detected for the T nucleotide spectrogram. Very secondary zones were perceived in the spectrogram relative to C at frequencies $f = 1/3$ and $f = 1/9$. This frequency would correspond to the third position nucleotide every nine nucleotides. In the case of coding regions, the third position nucleotide every three triplets.

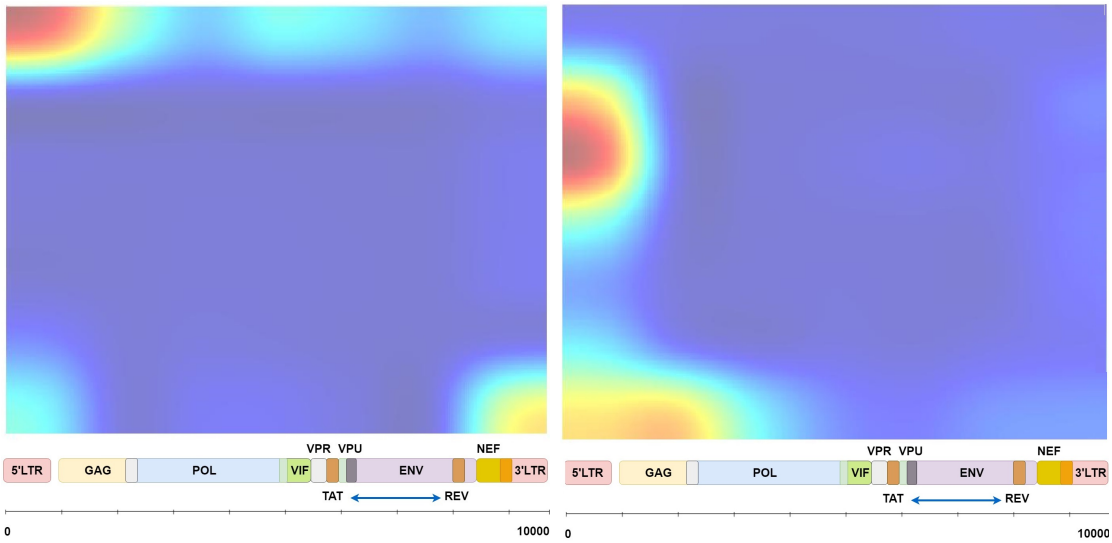
In the case of the recombinant category (Figure 14.2), there was a main hot zone in the vicinity between $f = 1/9$ and $f = 1/6$ in the first half of the A spectrogram. It is a very powerful but imprecise hot zone. However, the clear epicenter between both periodicities, multiples of 3, makes its relationship with every third nucleotides plausible. Another secondary zone was located at $f = 1/3$, also in the first half of the A spectrogram. Other very secondary zones were located at $f = 1/9$ between the spectrograms of C and T, as well as at $f = 1/3$ in T, and almost imperceptibly, in C. In some way, a mathematical signature associated with A was identified in the periodicities of 3 and its multiples 6 and 9; that is, the third position nucleotides every 3, 6, or 9 nucleotides. In the case of coding regions, the third position nucleotide every 1, 2, or 3 triplets.

In the case of superposed spectrograms, where the spectrogram value represented is the arithmetic sum of the spectrogram values for each nucleotide type, the hot zones in HIV-1 for both recombinant and non-recombinant categories are depicted in Figure 14.3.



(a) Non-recombinant **(b)** Recombinant

Figure 14.2: Total Hot Zones per Category (Concatenated Spectrograms) of HIV-1. The detailed information about these total hot zones is available in Chapter 6, Section 6.5.



(a) Non-recombinant Total Hot Zones **(b)** Recombinant Total Hot Zones

Figure 14.3: Total Hot Zones per Category (Superposed Spectrograms) of HIV-1. The exhaustive details concerning these total hot zones can be accessed in Chapter 6, Section 6.5.

In superposed spectrograms, no frequency related to one third had any degree of influence on the non-recombinant feature (Figure 14.3a). Conversely, for the recombinant feature (Figure 14.3b), the main hot zone was clearly located at $f = 1/3$ in 5'LTR and perhaps the first fragment of *gag*. According to these results, the third position nucleotides in 5'LTR and/or the third codon nucleotides at *gag* would contain fundamental encrypted information for detecting the recombinant characteristic of a HIV-1 genomic sequence.

14.2.3 HIV-1 Hot Zones per Subtype

After analyzing the hot zones corresponding to the recombinant and non-recombinant HIV-1 features by subtype, I identified numerous subtypes where hot zones for periodicities of 3 or their multiples, that is, related to the third popsition nucleotides, were involved. Table 14.1 is the compilation of all non-recombinant subtypes where I located among their main or secondary hot zones, areas related to the third position nucleotides.

Consistent with the total hot zones by category, for concatenated spectrograms, all representative subtypes, which were studied in Chapter 6, denoted their non-recombinant feature with the involvement of frequencies $f = 1/3$, $f = 1/6$, and $f = 1/9$ with varying degrees of potency.

In all non-recombinant subtypes analyzed, I detected a clear hot zone at $f = 1/3$ in the spectrogram of T. That is, the non-recombinant feature in all non-recombinant subtypes was related to the third position nucleotides (in case of coding regions, third codon nucleotides) in the spectrogram of T.

I detected two common secondary zones across all analyzed subtypes. The first was located at $f = 1/6$ (third position nucleotide every 2 triplets in coding regions, third nucleotides every six nucleotides in non-coding regions) around the spectrograms of C and T. The second at $f = 1/3$ in the spectrogram of C (third position nucleotides in non-coding regions, third codon nucleotides in coding regions).

In the case of the subtype U in superposed spectrograms, a secondary hot zone related to the third position nucleotides appeared in 5'LTR. In the rest of

the non-recombinant subtypes, no element related to a periodicity of 3 or its multiples was observed.

Table 14.1: Compilation of HIV-1 Non-Recombinant Subtypes with Hot Zones Associated with Nucleotide Periodicities of 3 or Their Multiples. Data extracted from Table 6.3.

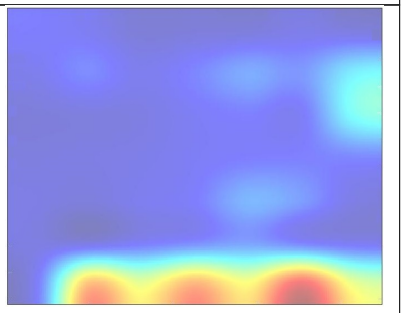
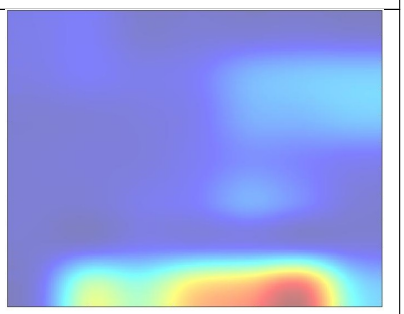
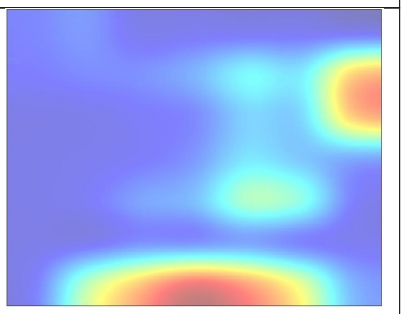
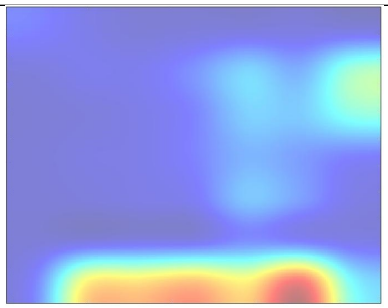
Subtype	**Spectrogram**	**Total Hot Zones**
A1	Concatenated	
A6	Concatenated	
B	Concatenated	
C	Concatenated	

Table 14.1 – Continuation from the previous page

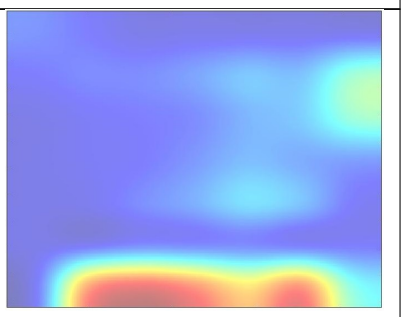
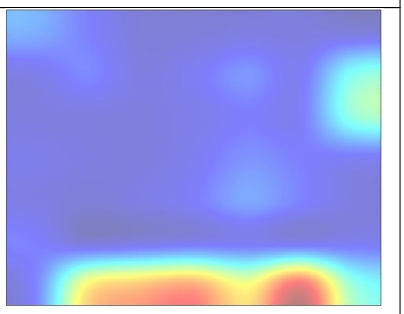
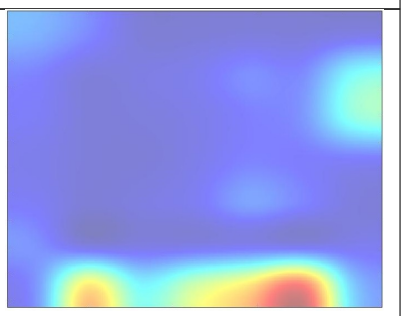
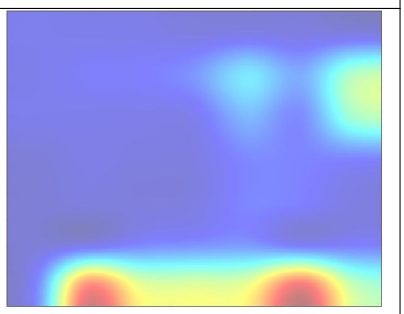
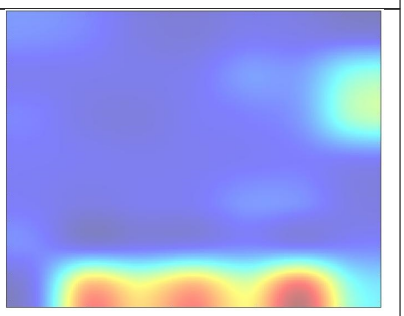
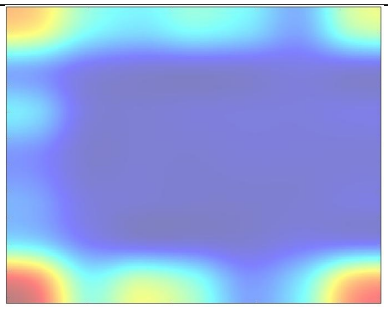
Subtype	Spectrogram	Total Hot Zones
D	Concatenated	
F1	Concatenated	
G	Concatenated	
O	Concatenated	
U	Concatenated	

Table 14.1 – Continuation from the previous page

Subtype	**Spectrogram**	**Total Hot Zones**
U	Superposed	

In the case of the recombinant subtypes analyzed in Chapter 6, the hot zones related to periodicity 3 or its multiples, by subtype, are compiled in Table 14.2.

Table 14.2: Compilation of HIV-1 Recombinant Subtypes with Hot Zones Associated with Nucleotide Periodicities of 3 or Their Multiples. Data extracted from Table 6.4.

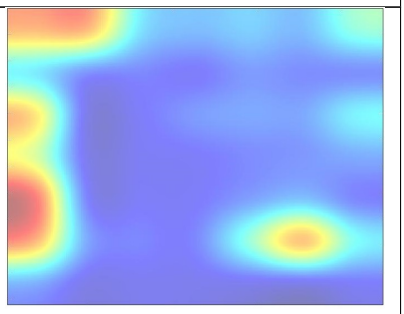
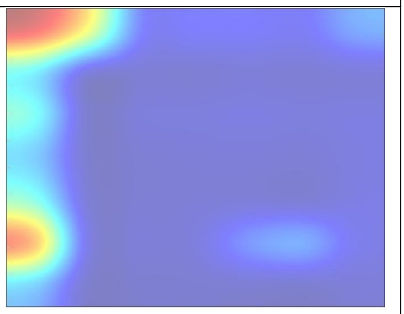
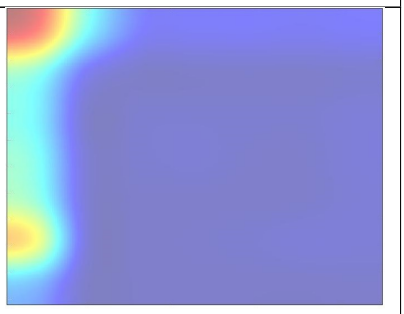
Subtype	**Spectrogram**	**Total Hot Zones**
CRF01_AE	Concatenated	
CRF02_AG	Concatenated	
CRF07_BC	Concatenated	

Table 14.2 – Continuation from the previous page

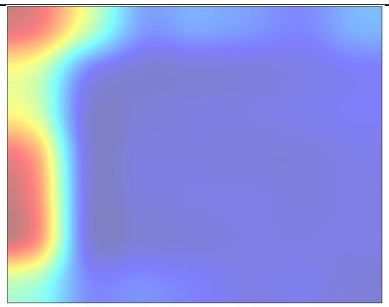
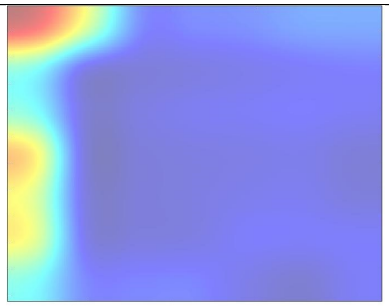
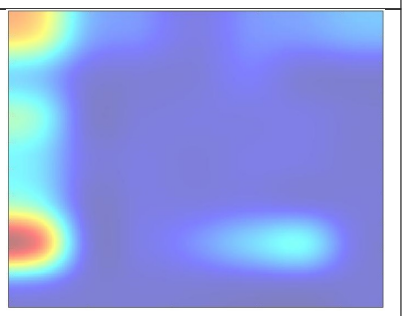
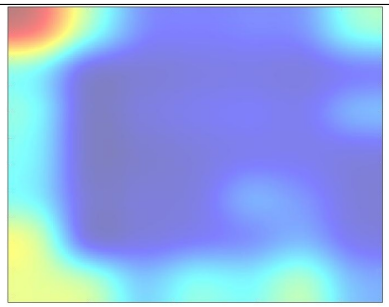
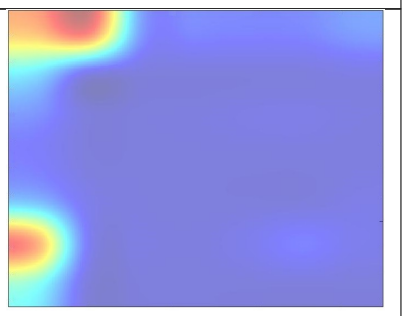
Subtype	Spectrogram	Total Hot Zones
CRF08_BC	Concatenated	
CRF11_cpx	Concatenated	
CRF20_BG	Concatenated	
CRF22_01A1	Concatenated	
CRF35_A1D	Concatenated	

Table 14.2 – Continuation from the previous page

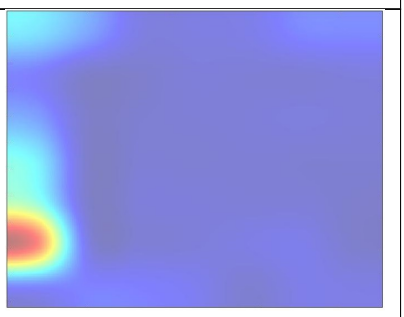
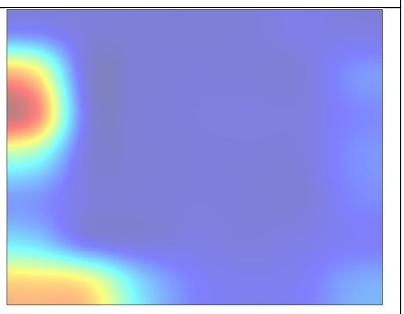
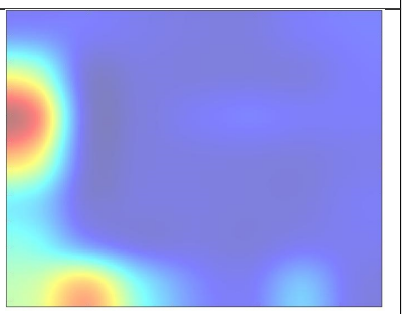
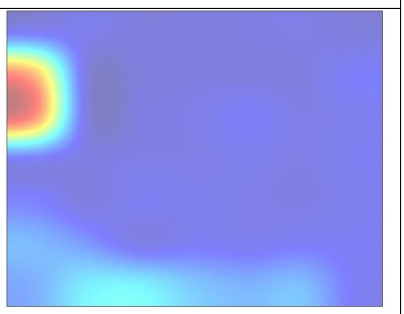
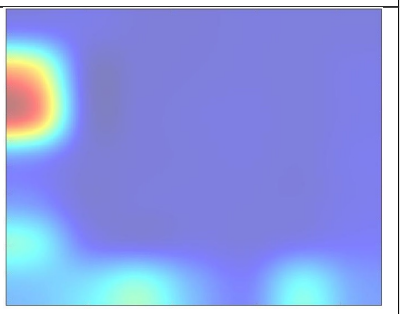
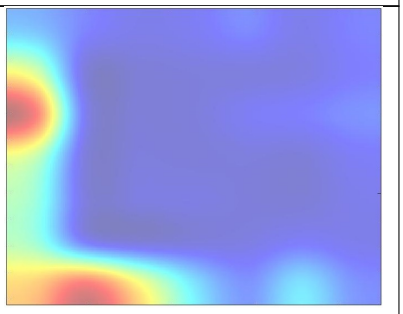
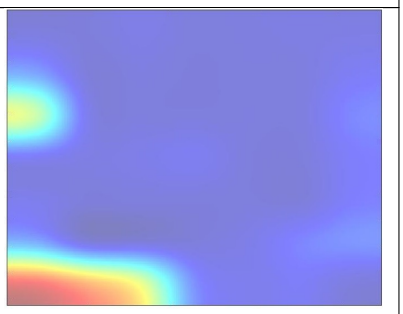
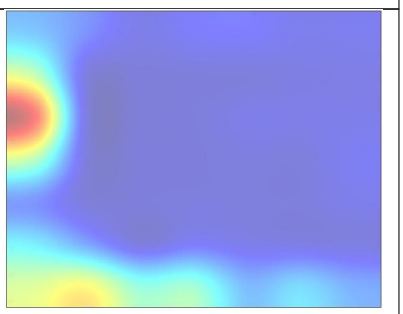
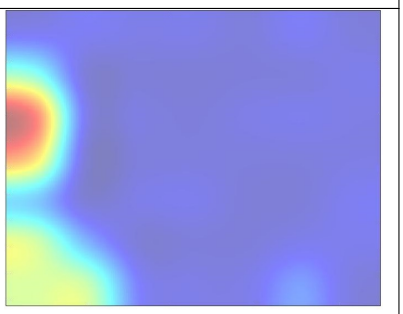
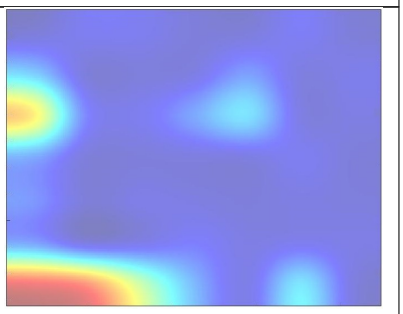
Subtype	**Spectrogram**	**Total Hot Zones**
CRF42_BF1	Concatenated	
CRF01_AE	Superposed	
CRF02_AG	Superposed	
CRF07_BC	Superposed	
CRF08_BC	Superposed	

Table 14.2 – Continuation from the previous page

Subtype	Spectrogram	Total Hot Zones
CRF11_cpx	Superposed	
CRF20_BG	Superposed	
CRF22_01A1	Superposed	
CRF35_A1D	Superposed	
CRF42_BF1	Superposed	

Regarding concatenated spectrograms, subtypes CRF01_AE, CRF02_AG,

CRF20_BG exhibited a hot zone related to a periodicity of 3 in the A spectrogram with variable potency, ranging from greater relevance in the case of subtype CRF01_AE to very minor relevance, as seen in subtype CRF02_AG.

The periodicity 6 in the A spectrogram played a role in the recombinant feature in the case of subtypes CRF01_AE, CRF08_BC, CRF11_cpx, and CRF42_BF1. This would be related to the third position nucleotide every two triplets (coding regions) or third position nucleotide every six nucleotides (non-coding regions).

I detected several hot zones related to a periodicity of 9. In the A spectrogram, I identified hot zones related to the third position nucleotide every 3 triplets (coding regions) or third position nucleotide every nine nucleotides (non-coding regions) in the recombinant subtypes CRF02_AG, CRF07_BC, CRF08_BC, CRF11_cpx, CRF20_BG, CRF22_01A1, CRF35_A1D, and CRF42_BF1. That is, in all subtypes except CRF01_AE. This may be related to the controversy existing in the scientific community around the so-called subtype E [120], and it should be investigated. In any case, this relationship between the third codon nucleotides every 3 triplets (coding regions) or third position nucleotide every nine nucleotides (non-coding regions) in the A spectrogram and the recombinant feature was common to virtually all analyzed subtypes. The strength of this hot zone varied, being more potent in subtypes CRF20_BG, CRF35_A1D, and CRF42_BF1 and notably less potent in CRF22_01A1 and CRF11_cpx.

Another hot zone with a periodicity of 9 in the context of the C and T spectrograms appeared with some strength in the CRF01_AE subtype, being the only subtype where the hot zone included orange tones. In subtypes CRF02_AG, CRF11_cpx, and CRF35_A1D, it was barely noticeable. The total hot zones of subtypes CRF20_BG and CRF22_01A1 also included this hot zone with a discreet but more easily noticeable strength.

The CRF01_AE subtype exhibited a secondary hot zone at $f = 1/3$ in the T spectrogram and, barely perceptible, at $f = 1/3$ in the C spectrogram. These hot zones were barely discernible in CRF11_cpx and CRF22_01A1.

The variety of hot zones related to the periodicity 3 or its multiples in the spectrograms, mainly of A and to a lesser extent of T and C (the latter being practically anecdotal), was striking. It will be deeply interesting to delve into the intrinsic relationships between the different periodicities related to the third position nucleotide and the subtypes involved in the recombinant feature, as well as their possible connections with the phylogenetic proximities among subtypes.

In the case of superposed spectrograms, the main hot zone corresponding to a periodicity of 3 in the 5'LTR was consistent, powerful, and common to all the recombinant subtypes under study. In subtype CRF20_BG, the power of this hot zone was lower. In subtype CRF42_BF1, its power was also limited, although somewhat higher than in the case of CRF20_BG.

Therefore, the relationship between the third position nucleotides in 5'LTR and the recombinant feature in HIV-1 seems clear.

A barely perceptible hot zone related to a periodicity of 3 was detected in the second half of the complete HIV-1 genome in subtypes CRF02_AG, CRF07_BC, and CRF11_cpx. In the case of CRF20_BG, this hot zone was more centered on 3'LTR.

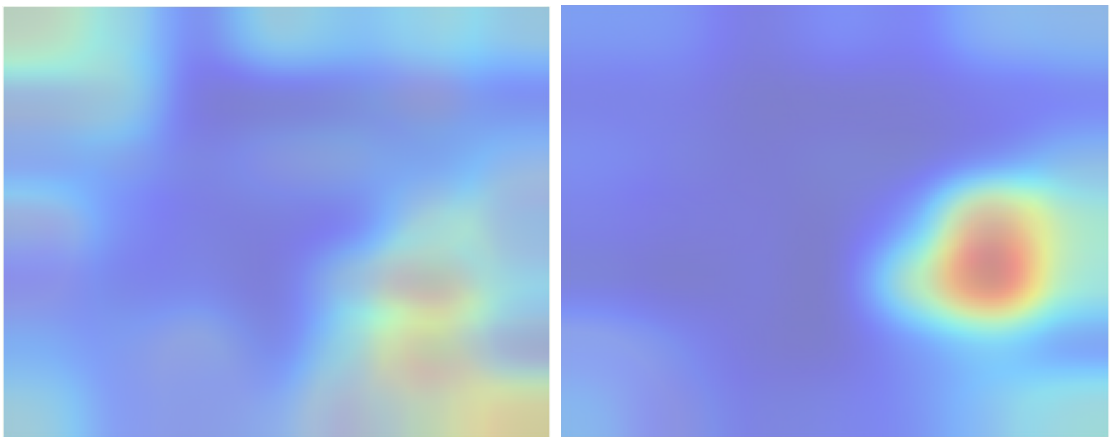
In subtype CRF42_BF1, a secondary hot zone perfectly visible at $f = 1/3$ was detected in the middle region of the genome, specifically around the end of *pol* and *vif*. This clearly visible hot zone was exclusive to this subtype, not present with the same intensity in any other of the analyzed recombinant subtypes.

In subtype CRF20_BG, a scarcely perceptible hot zone was detected at $f = 1/9$ in the last third of the genome.

In conclusion, all analyzed recombinant subtypes exhibited hot zones related to periodicity 3 and its multiples. The role of third position nucleotides in 5'LTR appeared particularly relevant, although most secondary regions were associated with multiples of 3. Therefore, in the obtained results, the role of every third nucleotides seemed to play a significantly relevant role, at least in the interpretation of the mathematical model.

14.2.4 Genetic Recombination in SARS-CoV-2

According to the detailed exposition in Chapter 11, based on the total hot zones generated in Step 3 of the interpretability results processing for 200 Epochs (optimal hyperparameter configuration), the total hot zones for recombinant and non-recombinant categories for complete sequences of SARS-CoV-2 were as shown in Figure 14.4.



(a) Non-recombinant **(b)** Recombinant

Figure 14.4: Identification of Total Hot Zones (Step 3) in SARS-CoV-2 Sequences for 200 Epochs (Superposed Spectrograms). The detailed information is available in Chapter 11, Section 11.4.2.

In the case of the non-recombinant category, the periodicities of ratio 6 and 9 in the S protein stood out for their potency. More secondary zones of periodicity 3 were detected in the 5'UTR and the first third of ORF1ab, in the second half of the genome. A barely perceptible secondary zone was detected for a periodicity of 6 in the first half of the genome.

In the case of recombinant sequences, the most striking result was the potent and focused hot zone for a periodicity of 6 in S. This would correspond to the third codon nucleotide every 2 amino acids. Additional analyses conducted in the pilot study of Chapter 13 redirected the determining periodicity in the recombinant feature of SARS-CoV-2 to a periodicity of 3.

I identified secondary zones barely perceptible related to a periodicity of 3 in the range from ORF3a to the end of 3'UTR and in 5'UTR and a first portion of

ORF1ab. In this genomic region, I also detected a barely perceptible hot zone relative to a periodicity of 9. However, it became evident that the epicenter in the periodicity of 6 in S protein was, by far, the most relevant zone for classifying a sequence as recombinant.

After these results, I conducted a pilot study detailed in Chapter 13. In this study, I extracted the common nucleotides between two groups of recombinant and non-recombinant sequences under study and discovered that all of them turned out to be third codon nucleotides. Therefore, in this pilot study, the mathematical signature associated with the recombinant feature was ultimately related to the periodicity of 3 nucleotides, specifically associated with the third codon nucleotides in Spike protein.

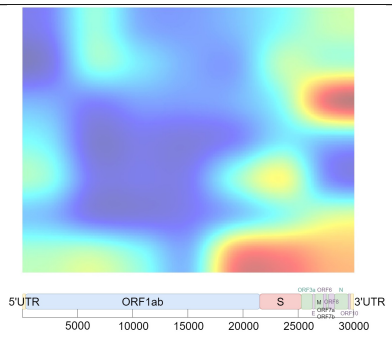
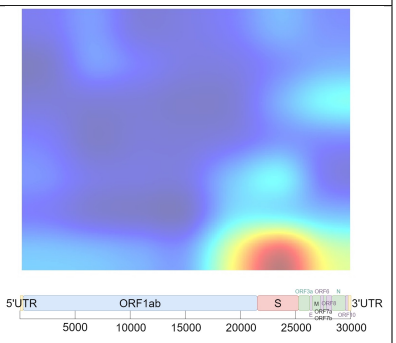
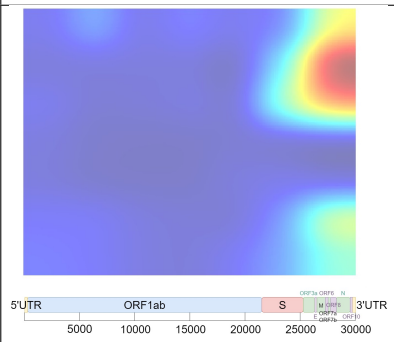
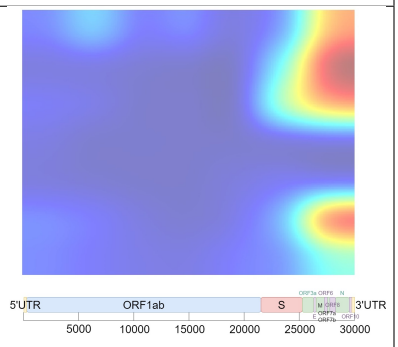
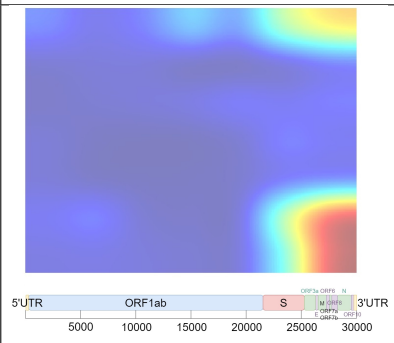
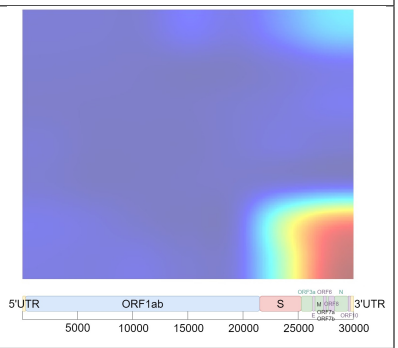
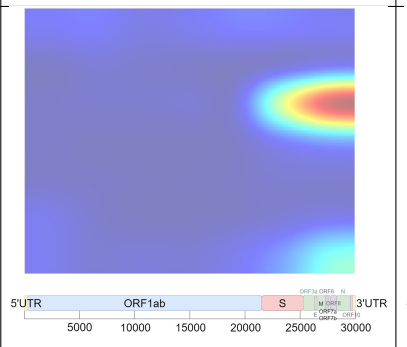
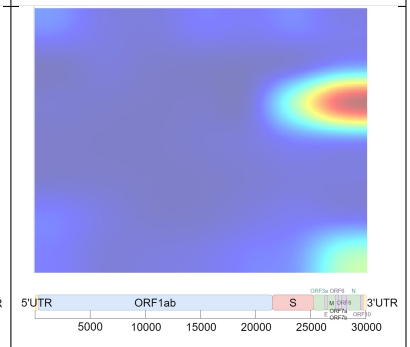
14.2.5 SARS-CoV-2 Variant Hot Zones

In order to delve into the variability of hot zones in the non-recombinant category of SARS-CoV-2, given the greater evolutionary diversity in sequences of this category, in Chapter 12, I aimed to deepen the recognition of the different major non-recombinant variants. In this case, the hot zones were not related to the recombinant or non-recombinant feature, but to the classification of a sequence into its corresponding variant, which in this case, were non-recombinant variants (pre-VOC, Alpha, Delta, and Omicron).

Table 14.3 compiles all the total hot zones of each variant for each of the two subsamples under study.

Multiple hot zones related to a nucleotide periodicity of 3 or its multiples, both primary and secondary, appeared. The pre-VOC category was the most dispersed, with a clear hot zone related to a nucleotide periodicity of 3 from ORF3a to the end of 3'UTR. In the case of VARIANT SUBSAMPLING 01, it was a highly potent hot zone, while in the case of VARIANT SUBSAMPLING 02, it was relegated to a secondary area. Secondary zones related to a nucleotide periodicity of 6 were detected in the S protein region and in 5'UTR, as well as in the first fragment of ORF1ab. Therefore, there is a relationship between the third codon nucleotides every 1 and every 2 amino acids and the classification of a sequence as pre-VOC.

Table 14.3: Comparative Table of Total Hot Zones of the Main Non-Recombinant Variants of SARS-CoV-2, Divided into Two Subsamplings

Variant	VARIANT SUBSAMPLING 01	VARIANT SUBSAMPLING 02
pre-VOC		
Alpha		
Delta		
Omicron		

The classification of a sequence as Alpha was strongly determined by a periodicity of 3, as well as, to a lesser extent, by the periodicity every 9 nucleotides in the range from ORF3a to the end of 3'UTR.

I must highlight that in the case of Delta, it seemed that the third codon nucleotides did not play a fundamental role, although in VARIANT SUBSAMPLING 01, an almost invisible hot zone was detected at $f = 1/9$ in the first third of the ORF1ab genome.

In the case of Omicron, a powerful and well-defined main hot zone at $f = 1/3$ was identified in the genomic region from the end of the S protein to the end of 3'UTR. Therefore, it seems that the third codon nucleotides in this region play a fundamental role in the mathematical signature associated with the classification of a sequence as Omicron.

14.3 Microbiological and Genomic Analysis

In this section, I correlated the evidence presented in this chapter with the scientific findings in the field of microbiology and virology.

14.3.1 Differences between HIV-1 and SARS-CoV-2

The HIV-1 and SARS-CoV-2 viruses differ significantly in multiple aspects. Broadly, the differences at various levels between these two viruses are as follows:

i) Classification and Viral Family:

- HIV-1: Belongs to the Retroviridae family, Lentivirus genus [2903].

- SARS-CoV-2: Belongs to the Coronaviridae family, Orthocoronavirinae subfamily, Betacoronavirus genus [2904].

ii) Viral Structure:

- HIV-1: Single-stranded RNA virus with an envelope. It has a protein capsid and a lipid envelope containing viral proteins, including the surface glycoprotein gp120 [2905].

- SARS-CoV-2: Single-stranded RNA virus with an envelope. It has a lipid envelope with spike (S) proteins, envelope, and matrix proteins [2906].

iii) Genome:

- HIV-1: It has a single-stranded RNA genome with approximately 9.8 kilobases in length [2907]. HIV-1 must first convert its RNA into DNA and integrate into the host's genome before it can initiate protein synthesis and viral replication [2908–2910]. Initially, the virus, which contains RNA as its genetic material, uses reverse transcriptase to transcribe its RNA into double-stranded DNA within the host cell [2911, 2912]. This step is crucial as the resulting viral DNA must integrate into the host's chromosomal DNA to establish a persistent infection [2913, 2914]. The enzyme integrase facilitates this integration by inserting the viral DNA into the host genome [2915, 2916]. Once integrated, the viral DNA is transcribed into messenger RNA (mRNA) by the host cell's transcription machinery [2917–2919]. This viral mRNA serves as a template for the synthesis of viral proteins and the production of new viral particles, marking the commencement of viral replication [2903, 2920]. This integration process is fundamental for the virus's proliferation and the persistence of the infection within the host organism [2921, 2922].

- SARS-CoV-2: It has a single-stranded RNA genome with approximately 29.9 kilobases in length [2923]. The RNA of SARS-CoV-2, a virus with a positive-sense RNA genome, acts directly as messenger RNA (mRNA) within the host cell [2670]. Upon entering the cell, the viral genome is immediately recognized by the cellular ribosomes and the host's translation machinery [2924, 2925]. This is because, in the context of SARS-CoV-2, the viral RNA can be directly translated into viral proteins without the need for transcription into DNA, unlike negative-sense RNA viruses [74, 2597]. Once the viral RNA is in the cytoplasm, it binds to ribosomes and is translated into viral proteins through the process of protein synthesis, utilizing the host cell's resources such as amino acids, tRNA, and ribosomes [2926]. This translation process produces the virus's structural proteins, including the spike protein (Spike), envelope protein (Envelope), membrane protein (Membrane), and nucleocapsid protein (Nucleocapsid), which are essential for the formation of new viral particles. Additionally, the positive-sense RNA also serves as a template for the synthesis of new viral genomes during replication, which are then assembled into

mature viral particles. The direct dependence on viral RNA for protein synthesis without additional transcription steps highlights the efficiency and speed with which SARS-CoV-2 can replicate and proliferate within the host [2610, 2927, 2928].

iv) Pathogenesis:

- HIV-1: Causes HIV infection and eventually Acquired Immunodeficiency Syndrome (AIDS), attacking and destroying the host's immune system [2929].

- SARS-CoV-2: Causes Coronavirus Disease 2019 (COVID-19), typically affecting the respiratory system and, in severe cases, leading to pneumonia, Acute Respiratory Distress Syndrome (ARDS), and other systemic symptoms [2930].

v) Transmission:

- HIV-1: Is primarily transmitted through body fluids such as blood, semen, vaginal secretions, and breast milk, as well as through unprotected sexual contact, sharing contaminated needles, and from mother to child during childbirth or breastfeeding [2931].

- SARS-CoV-2: Is primarily transmitted through respiratory droplets produced when an infected person coughs, sneezes, or talks. It can also spread by touching contaminated surfaces and then touching the mouth, nose, or eyes [2932].

vi) Infection Pattern:

- HIV-1: Selectively infects and destroys immune system cells, especially CD4+ T lymphocytes [2933].

- SARS-CoV-2: Infects epithelial cells of the upper and lower respiratory tract, as well as other tissues and organs expressing the Angiotensin-Converting Enzyme 2 (ACE2) receptor [2934].

vii) Immune Response:

- HIV-1: Suppresses and weakens the host's immune response, especially T-cell response and antibody production [2935].

- SARS-CoV-2: Triggers an immune response, which in some cases can be exaggerated and lead to a cytokine storm, contributing to the severity of the disease in some patients [2936].

viii) Treatments and Vaccines:

- HIV-1: Effective antiretroviral treatments exist that can control viral replication and delay progression to AIDS. There is still no preventive vaccine available [2937].

- SARS-CoV-2: Several vaccines have been developed and authorized for emergency use, and various antiviral treatments and monoclonal antibody therapies were authorized for use in COVID-19 patients [2938].

Therefore, HIV-1 and SARS-CoV-2 differ significantly in their structure, genome, pathogenesis, transmission, and immune response.

14.3.2 The Role of Nucleotides A and T in Periodicities of 3 and their Multiples

In HIV-1, the primary role of the nucleotide A, and to a lesser extent T, in the recombinant feature (and to some extent, in the non-recombinant feature) became evident.

Indeed, nucleotides A and T play a significant role in the generation of genetic recombination events in HIV-1.

In the HIV-1 genome, long stretches of Adenine (polyA) and Thymine (polyT) are frequently observed. These repetitive sequences possess a set of properties that render them particularly relevant to the recombination process:

i) RNA Structure:

- PolyA and polyT stretches can form less stable secondary structures due to the lower complexity in base pairing. This can lead to more flexible or "weak" zones in the RNA structure, which are more prone to errors during replication [2939–2941].

ii) Polymerase Skipping [2942, 2943]:

- PolyA and polyT regions are prone to omission by the reverse transcriptase enzyme, which is responsible for copying viral RNA into DNA. This skipping or "jumps" of the polymerase can lead to the introduction of breaks or recombinations in the viral genome.

- These "jumps" occur when the polymerase temporarily loses the RNA template and reunites with it at a different position, facilitating the mixing and pairing

of different regions of the viral genome. This results in the creation of new recombinant variants of the virus.

Furthermore, the prominent role of the adenine nucleotide in HIV-1 can be attributed to the abundant presence of adenine in the genome of lentiviruses (a subgroup of retroviruses to which HIV-1 belongs). This abundance has several implications:

i) Adenine Content in Lentiviruses:

- Lentiviruses, especially HIV-1, have a high adenine content in their genome [2944]. This not only influences the formation of polyadenine (polyA) tracts but also affects the stability and behavior of viral RNA during replication [2945].

ii) Evolutionary Implications:

- The high concentration of adenine may be an evolutionary adaptation that maximizes the genetic flexibility of the virus. This could facilitate greater genetic diversity and, therefore, a higher capacity to adapt to host defenses and antiretroviral treatments [2946].

iii) RNA Structure and Viral Function:

- Adenine-rich sequences can influence the secondary structure of viral RNA, affecting both viral RNA replication and translation. Flexibility in these regions may allow for greater plasticity in viral function and in the response to selective pressures [2947].

The combination of the aforementioned factors results in a higher propensity for genetic recombination in HIV-1. The outcome of these factors is:

- RNA Flexibility. The less stable structure of polyA and polyT regions facilitates the formation of loops and other structural elements that may interrupt the continuous replication of RNA [2948].

- Polymerase Skipping. The predisposition of reverse transcriptase to skip in these regions increases the frequency of recombinations. This is particularly important in a virus like HIV-1, where recombination is a key strategy for generating genetic diversity [2911].

- Genetic Diversity. The genetic recombination resulting from these processes is a crucial source of variation for HIV-1, allowing the virus to evade host immune pressure and adapt to pharmacological treatments [2949].

In summary, the presence of polyA and polyT stretches in the HIV-1 genome, along with the high concentration of adenine, significantly facilitates genetic recombination. This mechanism is essential for the virus's ability to evolve rapidly and persist in human populations despite efforts to control it through the immune system and antiretroviral therapy.

In the case of SARS-CoV-2, in the pilot study conducted in Chapter 13, I detected a high Thymine content in the third codon nucleotides in the Spike protein, around 45%. This high amount of Thymine compared to other nucleotide types was not detected in other structural proteins such as M and N.

Although SARS-CoV-2 is a virus with a positive-sense RNA genome, nucleotide sequence representations provided in FASTA files commonly use Thymine (T) instead of Uracil (U), which is present in RNA. This convention arises from the nature of the FASTA format and the tradition in bioinformatics of using DNA notation, which employs Thymine, to describe nucleotide sequences regardless of whether the nucleic acid is DNA or RNA. Therefore, in the case of SARS-CoV-2 RNA sequences, "T" in FASTA files should be interpreted as Uracil, acknowledging that this representation follows a standard practice in genomic data handling [2950].

This fact may be related to the reduction in the frequency of cytosines and a corresponding increase in the frequency of uracils at the third position of the codon detected in SARS-CoV-2. This phenomenon, known as codon usage bias, suggests a shift in codon preference over the virus's evolution, possibly driven by specific mutations and selective pressures in the host [2951–2953].

Cytosine to Uracil ($C > U$) mutations have been identified as the most prevalent in the genome of SARS-CoV-2. These mutations could be the result of host deaminase activity, enzymes that catalyze the conversion of cytosine to uracil in RNA. This observation underscores the importance of the interaction between the

virus and the host in determining the genetic diversity and evolutionary dynamics of SARS-CoV-2 [2701, 2954].

In terms of molecular genetics, the third position of the codon is known for its relative flexibility in terms of amino acid specificity. The wobble hypothesis, proposed by Francis Crick in 1966, postulates that the third base of the codon can pair non-conventionally with the corresponding base in transfer RNA (tRNA), allowing some tolerance in codon specificity. Therefore, changes in the third position of the codon, such as the C>U transition observed in the genome of SARS-CoV-2, may not necessarily have a direct impact on the amino acid sequence of the encoded protein, although they can influence translation efficiency and messenger RNA stability [2877].

In conclusion, based on the analysis of the hot zones resulting from the recombinant and non-recombinant categories, I determined the significant role of A and, to a lesser extent, T in the recombinant feature of HIV-1. Additionally, I identified the role of T in the recombinant feature of SARS-CoV-2. In both viruses, there was a relationship with these types of nucleotides, with T being the relevant nucleotide type in the classification of a sequence of both viruses as recombinant. This was the case despite these viruses being very different. Furthermore, in both cases, the interpretation was consistent with the bio-centric approach in line with the results obtained.

14.3.3 LTRs of HIV-1

The 5'LTR of HIV-1 plays a critical role in multiple aspects of the virus life cycle. Below are its main functions:

- Viral Promoter and Regulatory Sequences. The 5'LTR contains promoter sequences and regulatory elements that control the transcription of viral RNA from the viral DNA integrated into the host genome. These sequences include elements such as the U3 region (Unique 3'), which contains response elements to cellular and viral signals [2955], and the U5 region (Unique 5'), which contains transcription initiation sequences and termination and processing sequences [2956].

These sequences are recognized by the cellular and viral transcription machinery, and they are crucial for the initiation, regulation, and efficient processing of viral RNA transcription [2957].

- Transcription Initiation and Viral Promoter. At the 5' end of the 5'LTR there are transcription initiation sequences recognized by cellular RNA polymerase II [2958]. These sequences, along with viral promoter elements like the TAR (Trans-Activation Responsive) element, facilitate the binding of specific transcription factors and efficient initiation of viral transcription [2439]. Activation of this viral promoter is a crucial step in HIV-1 gene expression and is necessary for viral RNA production.

- Viral DNA Integration. The R (Repeat) sequences in the 5'LTR serve as binding sites for the viral integrase enzyme. The integrase catalyzes the integration of viral DNA into the host genome, an essential step for virus replication and persistence. Viral DNA integration allows HIV-1 to remain latent in host cells and may have significant implications for viral pathogenesis and host immune response [2959, 2960].

- Regulation of Viral Gene Expression. In addition to its role in transcription and viral DNA integration, the 5'LTR may also contain sequences responsive to cellular and viral signals that modulate HIV-1 gene expression. These sequences can influence the activity of the viral promoter and transcription of viral RNA in response to changes in the cellular environment or the presence of viral factors, affecting HIV-1 replication and pathogenesis [2961].

In summary, the 5'LTR of HIV-1 is a multifunctional genomic region that coordinates viral RNA transcription, viral DNA integration into the host genome, and HIV-1 gene expression. Its importance in the HIV-1 lifecycle makes it a key target for research into antiviral therapies directed against HIV-1.

The 3'LTR in HIV-1 is a specific region of the viral genome found at the 3' end of the viral DNA sequence. This region is extremely important in the virus life cycle and plays several critical functions:

- Promoter for viral genome transcription. The 3'LTR contains transcription start sequences recognized by the cellular transcription machinery, enabling the synthesis of viral RNA from the viral DNA integrated into the host genome.

These transcription start sequences are recognized by specific transcription factors that recruit RNA polymerase to initiate viral genome transcription [80].

- Site of viral DNA integration into the host genome. Viral DNA integration into the host genome is a crucial step in the HIV-1 lifecycle and is mediated by the viral enzyme integrase. The 3'LTR contains specific sequences recognized by integrase, facilitating the precise insertion of viral DNA into the host genome. This integration is essential for viral replication and infection persistence [2962].

- Signals for termination and processing of viral RNA transcription. After transcription of viral RNA from the integrated genome, the 3'LTR contains polyadenylation sequences that signal the end of transcription and facilitate the addition of a poly(A) tail to the 3' end of viral RNA. This poly(A) tail is important for the stability and translation of viral RNA [2963].

In summary, the 3'LTR of HIV-1 is a critical region of the viral genome that acts as a promoter for viral transcription, facilitates integration of viral DNA into the host genome, and provides necessary signals for processing and termination of viral transcription. Its importance in the HIV-1 life cycle makes it a key target for research and development of antiviral therapies directed against HIV-1.

14.3.4 Differences between the LTRs in HIV-1 and the UTRs in SARS-CoV-2

Given the relevance of the LTRs in the recombinant feature of HIV-1, one would expect a more significant role of the UTRs in the SARS-CoV-2 in its recombinant feature as an equivalent counterpart. However, in the case of SARS-CoV-2, it seemed to have a secondary and ambiguous role more in the classification as non-recombinant than in the case of the recombinant feature. The same applies to the classification of non-recombinant sequences in a specific variant. It was not possible to identify any unequivocal hot zone in the UTRs, without possible intervention from some adjacent genomic region.

It would seem reasonable to expect that the demonstrated importance of the LTRs in the case of HIV-1 would have a certain equivalence in the UTRs in the case of SARS-CoV-2. However, I did not detect clear indications of such equivalence.

Focusing on the differences between the LTRs in HIV-1 and the UTRs in SARS-CoV-2 to try to justify this lack of clear correlation, I delved deeper into both genomic regions and detected certain differences.

The LTRs of HIV-1 and the UTRs of SARS-CoV-2 exhibit substantial differences both in their composition and biological function.

Firstly, the LTRs of HIV-1 are genetic sequences approximately 634 base pairs in length that flank the viral genome at both ends, consisting of the U3, R, and U5 regions. These regions contain crucial regulatory elements for viral transcription and replication, including promoters, binding sites for transcription factors, and transcription initiation and termination sequences [2964].

On the other hand, the UTRs of SARS-CoV-2 comprise the untranslated regions located at the 5' and 3' ends of the viral RNA, essential for the regulation of translation and RNA stability [2965, 2966]. While crucial for viral biology, they do not possess the same organizational structure as the LTRs of HIV-1.

In terms of function, the LTRs of HIV-1 play a central role in the regulation of viral gene expression, acting as potent promoters that control the transcription of viral genes and the formation of the integrated viral DNA copy in the host genome [2439]. Meanwhile, the UTRs of SARS-CoV-2 play a fundamental role in the stability and efficiency of translating viral RNA into functional viral proteins [2594].Furthermore, the UTRs of SARS-CoV-2 are involved in interacting with host cell factors, thus modulating the immune response and viral pathogenicity [2967].

In summary, although both the LTRs of HIV-1 and the UTRs of SARS-CoV-2 are essential elements for viral biology and pathogenesis, they differ significantly in their structure, composition, and biological function. The LTRs of HIV-1 are crucial for viral transcription and replication, whereas the UTRs of SARS-CoV-2 are important for translation regulation and interaction with the host cell machinery.

As previously explained, the HIV-1 RNA must first be reverse-transcribed into DNA and integrated into the host genome, a process in which Long Terminal Repeats (LTRs) play a crucial role [2423, 2903, 2961]. In contrast, the SARS-CoV-2 RNA, being positive-sense RNA, is used directly for the synthesis of viral proteins [2670]. In this context, the UTRs, among other possible functions, must protect and stabilize this viral RNA [2968–2970].

Perhaps these differences in their functionalities are related to the different genomic regions involved in the mathematical signature associated with the recombinant feature in both viruses. This requires further research.

14.3.5 The Spike Protein in SARS-CoV-2

The Spike protein (S) of SARS-CoV-2 is a trimeric glycoprotein that plays a fundamental role in the infectivity and evolution of the virus. It is primarily responsible for the virus entering host cells and is one of the main targets for the body's immune response. Due to its importance, it has been the subject of intense studies to understand its structure, function, and how mutations in this protein can affect the virus transmissibility and immune evasion [2631].

Regarding its structure, the Spike protein is composed of approximately 1,273 amino acids and is organized into three main segments: the S1 subunit, the S2 subunit, and the Transmembrane Domain (TM) [2971]. The description of the structure of each segment is:

i) S1 Subunit:

- N-terminal Domain (NTD). This domain may be involved in immune modulation and interaction with cellular co-factors. The NTD is a flexible region that can undergo significant conformational changes, and mutations in this region can affect antibody responses [2972].

- Receptor Binding Domain (RBD). This is the critical region for interaction with the ACE2 receptor. The RBD can be found in two conformations: "up" (exposed) and "down" (hidden). This ability to alternate conformations influences the accessibility and binding affinity to the receptor [2973].

- Furin Cleavage Site (S1/S2). A specific cleavage site by the furin protease. This site is not present in close relatives of SARS-CoV-2 such as SARS-CoV, and its presence in SARS-CoV-2 has been associated with increased infectivity and transmission capacity [2974].

ii) Subunit S2:

- Fusion Peptide (FP). This peptide unfolds and inserts into the membrane of the host cell, initiating the membrane fusion process [2975].

- Heptad Repeat 1 (HR1) and Heptad Repeat 2 (HR2). These regions form a six-helix bundle (6-HB) structure during the fusion process, bridging the viral and cellular membranes [2976].

- S2' Cleavage Site. This site is processed by proteases such as TMPRSS2, facilitating the subsequent activation of subunit S2 for membrane fusion [2977].

iii) Transmembrane Domain (TM) and Intracellular Region (IC):

- Transmembrane Domain (TM). Anchors the Spike protein in the virus membrane [2978].

- Intracellular Region (IC). The cytoplasmic portion interacts with the viral and cellular machinery during virus assembly [2979].

Regarding the mechanisms of action of the Spike protein, the entry of SARS-CoV-2 into the host cell involves several coordinated steps, including receptor binding, proteolytic cleavage, and membrane fusion:

i) Binding to the ACE2 Receptor [2980]:

- The RBD in the S1 subunit binds to the ACE2 receptor on the host cell. This event is the first critical step in the viral entry process.

- The affinity of this interaction largely determines the infectivity of the virus. For example, the N501Y mutation, observed in the Alpha and Beta variants, increases the affinity of the RBD for ACE2, enhancing the virus ability to infect cells.

ii) Spike Cleavage and Activation [2981]:

- Furin Cleavage (S1/S2). Cellular furin cleaves the Spike protein at the S1/S2 site during virus biogenesis, preactivating the protein for more efficient entry.

- TMPRSS2 Cleavage (S2'). After initial binding, the cellular protease TMPRSS2 cleaves the Spike protein at the S2' site, which is crucial for activating the fusion peptide.

iii) Membrane Fusion [2982–2985]:

- Following activation by TMPRSS2, the fusion peptide in the S2 subunit inserts into the cell membrane. The HR1 and HR2 regions then reconfigure to form the six-helix bundle (6-HB) complex, bringing the viral and cellular membranes close together and facilitating their fusion, allowing the viral genome to enter the host cell.

Mutations in the Spike protein can have significant impacts on the virus biology. Here are some of the most important mutations and their impact:

i) D614G Mutation [2986, 2987]:

- This mutation changes an aspartic acid to a glycine at position 614 of the Spike protein.

- It increases the stability of the Spike trimer and favors the "up" conformation of the RBD, enhancing the virus's ability to infect cells.

- It has become the dominant form of the circulating virus due to its transmission advantage.

ii) Mutations in Variants of Concern (VOCs):

- Alpha (B.1.1.7). The N501Y mutation increases affinity for ACE2 [2988]. The amino acid deletion at positions 69-70 in the NTD may increase evasion of detection by some antibodies [2989].

- Beta (B.1.351) and Gamma (P.1). Mutations K417N/T, E484K, and N501Y increase affinity for ACE2 and reduce the effectiveness of neutralizing antibodies, decreasing vaccine efficacy [2990].

- Delta (B.1.617.2). Mutations L452R and T478K increase affinity for ACE2 and resistance to antibody neutralization, increasing transmissibility and virulence [2991].

- Omicron (B.1.1.529). This variant carries over 30 mutations in the Spike protein, including several in the RBD and NTD, enabling greater immune evasion and potentially altering infectivity [2992].

Mutations in the Spike protein allow SARS-CoV-2 to evade the immune response in several ways:

i) Antigenic Shift [2993]:

- Mutations in the RBD and other key regions can alter the epitopes that antibodies recognize, reducing neutralization efficacy [2994]. For example, mutations E484K and N501Y have been shown to reduce the effectiveness of various monoclonal antibodies and convalescent sera [2995].

ii) Glycosylation:

- The Spike protein is heavily glycosylated, with numerous glycosylation sites that can shield epitopes and evade antibody detection [2996].

- Changes in glycosylation patterns due to mutations can affect the visibility of certain epitopes to neutralizing antibodies [2997].

Regarding the role played by the S protein in natural selection and genetic recombination:

i) Selective Pressure [2998]:

- Selective pressure exerted by the host immune response and medical interventions such as vaccines drives the emergence of mutations beneficial to the virus.

- Mutations conferring adaptive advantages in terms of transmissibility and immune evasion tend to become fixed more rapidly in the viral population.

ii) Genetic Recombination:

The Spike protein is one of the most variable regions of the SARS-CoV-2 genome due to selective pressure from the host immune system and virus evolution. This variability in the Spike protein sequence can facilitate genetic recombination by providing breakpoints or "cross-over" points during the formation of new viral strains. Additionally, the RBD region, crucial for binding to the cellular receptor, may be particularly prone to genetic changes that could alter binding specificity and virus infectivity [2999, 3000].

The Spike protein of SARS-CoV-2 is a key target in the development of vaccines and in the diagnosis and treatment of COVID-19. Its continuous mutations pose challenges to the effectiveness of vaccines and to diagnostic and treatment strategies.

i) Vaccine Development [3001]:

- Most current vaccines against COVID-19 are designed to elicit an immune response against the Spike protein. The continuous evolution of this protein may require periodic updates to the vaccines to maintain their effectiveness.

- Variants such as Omicron, with multiple mutations in the Spike protein, have led to research on the need for booster doses and vaccine modifications to address these variants [3002].

ii) Diagnosis and Treatment:

- Mutations in the Spike protein can affect the sensitivity of diagnostic tests, especially those based on specific regions of the viral genome [3003].

- The effectiveness of monoclonal antibody treatments may be compromised by mutations affecting the binding epitopes of these antibodies. Continuous genomic monitoring is essential to adapt diagnostic and therapeutic strategies accordingly [3004].

Therefore, the importance of the Spike protein as a driver of evolution in SARS-CoV-2, as a highly variable region (as I observed in Chapter 10), its role in multiple genetic recombination events, and the tendency for $C > U$ hypermutations and genetic code degeneration, all of this may align with my results and the importance of third codon nucleotides in Spike. Further research into all these phenomena is necessary.

14.4 Hypotheses

Throughout all the experimentation presented, I identified various influences of nucleotides at the third position on the mathematical signatures inherent to the patterns detected by CNNs. Spectrographic frequencies $f = 1/3$, $f = 1/6$, and $f = 1/9$ were recurrently observed in the interpretability heatmaps studied, indicating the presence of periodicities with a ratio of 3 and its multiples. Therefore, it seems

that the third position nucleotides contain encrypted information regarding viral features (recombinant feature and variant) in viruses as distant as HIV-1 (every third codon in LTRs and maybe third codon nucleotides of *gag*) and SARS-CoV-2 (third codon nucleotides of S).

However, numerous clues established a correlation between phenomena related to arithmetic series with a ratio of 3 or its multiples and the mathematical signature associated with the recombinant feature in both viruses. In the case of HIV-1, the importance of the periodicity 3 in the 5'LTR region has been observed, with potent and stable associated hot zones, as detailed in Chapters 5 and 6. The possible influence of the third codon nucleotides of *gag* should also be considered. Similarly, in the case of SARS-CoV-2, I observed similar patterns in the third codon nucleotides of the Spike protein, specific to this protein and suggesting a determinant role in the recombinant feature, as developed mainly in Chapters 11 and 13.

These similarities between such different viruses lead me to consider the possible degree of universality in the mathematical signatures.

Furthermore, the bio-centric interpretation is aligned with the results obtained regarding the role of the A and T nucleotides in HIV-1 and the role of the LTRs in relation to their possible association with the location of the mathematical signatures associated with the viral feature. In the case of SARS-CoV-2, the role of the Spike protein and the phenomenon of hypermutations $C > U$ are consistent with the results obtained regarding the location of the mathematical signature relative to the recombinant feature precisely in that protein and with the high amounts of Thymine detected in the Spike protein.

The periodicities relative to 3 and its multiples were also clearly involved in the mathematical signatures associated with the classification of non-recombinant sequences in the pre-VOC, Alpha, Delta, and Omicron variants, although the correlation between this phenomenon and the genomic regions involved requires further investigation.

In this Chapter, I compiled the main evidence linking nucleotides in the third codon position (case of SARS-CoV-2 Spike and maybe HIV-1 *gag*) or every third

position in the case of HIV-1 LTRs; with activated zones where the CNN focused its attention to make decisions.

Based on this compilation of evidence, I strongly propose the hypothesis that nucleotides in the third position contain encrypted information that influences the classification of certain viral characteristics. Furthermore, I raise the possibility that this role of nucleotides in the third position may be common in different microorganisms, suggesting a degree of universality that needs to be further evaluated.

The validation of this hypothesis would represent a significant shift in the interpretation of the genetic code as we currently understand it. It would also confirm that the degeneracy of the genetic code is more complex than a simple conversion of nucleotides to amino acids, and that the identity of the nucleotide in the third codon position may be relevant for virus characterization.

If the genome indeed contains encrypted information in different layers, it would open up a wide field of research with significant implications for our understanding of molecular biology and viral evolution.

At this point, everything outlined in this Chapter requires further research, and my conclusion is nothing more than a mere hypothesis.

Hopefully, in time, we will have more data and can evolve our understanding of this wonderful universe that is the genome.

This is the beginning of an exciting journey.

When we are born, we cry that we are come to this great stage of fools.

— William Shakespeare ("King Lear"), *between 1592 and 1594, theaters were closed due to a plague epidemic. It was during this time that Shakespeare began writing poetry, as there was no demand for theatrical works.*

15

Final Conclusions

Contents

15.1 Introduction 730
15.2 Some Numbers 732
15.3 Main Contributions 733
15.3.1 Chapter 3. Technological Background 734
15.3.2 Chapter 4. Genomic Spectrogram 735
15.3.3 Chapter 5. HIV-1 Genetic Recombination Analysis . . . 736
15.3.4 Chapter 6. Analysis of Total Hot Zones in HIV-1 by Subtype 743
15.3.5 Chapter 7. A Brief Analysis of HIV-1 Subtypes F1, CRF22_01A1, and O 749
15.3.6 Chapter 8. Analysis of HIV-1 Subtype U 752
15.3.7 Chapter 9. Analysis of HIV-1 Subtype G 754
15.3.8 Chapter 10. Exhaustive analysis of SARS-CoV-2 Genomic Elements 758
15.3.9 Chapter 11. SARS-CoV-2 Genetic Recombination Analysis 762
15.3.10 Chapter 12. Analysis of SARSCoV-2 Non-recombinant Variants 764
15.3.11 Chapter 13. Exploring the Influence of the Third Codon Nucleotide on Genetic Recombination in the SARS-CoV-2 Spike Protein 767
15.3.12 Chapter 14. Is There a Role Associated with the Third Codon Nucleotide in the Characterization of a Genomic Sequence? The Case of HIV-1 and SARS-CoV-2 771
15.4 Fulfillment of Research Objectives 778
15.5 Main Conclusions 784
15.6 Implications 786

15.1 Introduction

The research that resulted in this document began with a research question that was unattainable with the current state of the art: Is it possible to predict the outcome of genetic recombination between SARS-CoV-2 and FCoV using Deep Learning tools?

The degree of unknowns in the genome and its functionalities, coupled with the limited number of complete FCoV sequences, made it impossible to efficiently apply Deep Learning algorithms to predict the outcome of recombination between both viruses. However, the undoubtedly demonstrated power of Deep Learning tools led me to continue considering these tools valuable for contributing knowledge in the genomics field. It simply involved redirecting the case study to the analysis of genetic recombination in SARS-CoV-2.

In a preliminary analysis of deep learning technologies, I determined a preference for mature technologies, whose effectiveness is well demonstrated, as an additional measure to ensure the quality of the results. Convolutional Neural Networks (CNNs) represent powerful tools for detecting patterns that may not be discernible to the human eye. This pattern detection capability is particularly pronounced in image analysis.

Applying Transfer Learning to a pre-trained CNN that has been trained on datasets like ImageNet, which consist of millions of images distributed across thousands of categories, allows leveraging the extensive knowledge of these networks on new sets of images with greater initial reliability.

The use of CNNs thus requires a representation of genomic sequences that precisely fits this image format. To this end, I selected the generation of genomic spectrograms as a possibility.

This representation system preserves biological coherence. Furthermore, the mathematical nature of the spectrogram provides information regarding the relationship of more mathematical aspects with certain features.

In order to locate the regions of the spectrogram that could be relevant for the recombinant feature, I decided to use explainability tools that provide us with where the CNN looks to make a decision. We do not know what it sees, but we will know where it looks.

Once I had pre-selected the technologies, my next step was to analyze SARS-CoV-2 as an interesting case study, being the most sequenced virus currently, by far surpassing others. Despite its high sequencing rate, I discovered a significant imbalance between available recombinant and non-recombinant sequences. There were 1,539,728 non-recombinant sequences compared to only 1,565 recombinant ones. This high imbalance seemed risky for a first approach to understanding genetic recombination.

Therefore, I decided to choose a virus case where there are abundant and well-documented cases of genetic recombination to use as a basis for obtaining knowledge in genetic recombination. This is the case of HIV-1, whose variability is much higher than that of other viruses, in part due to the multiple and well-documented cases of genetic recombination. In later stages, I extrapolated the knowledge gained from the most well-documented case (HIV-1) to other viruses, in this case, SARS-CoV-2.

With this set of premises resulting from prior documentation work, I was able to formulate my research hypothesis:

The application of advanced Deep Learning techniques and mathematical analysis to genomic spectrograms of viruses will enable not only the effective characterization of viral genomes, but also the identification of encrypted mathematical signatures. These signatures can facilitate the understanding of viral genotypic and phylogenetic relationships, providing a solid foundation for the design of generative architectures aimed at predicting genomic evolution in viruses.

From which the following research objectives were derived:

I. Evaluate genomic spectrograms as a method for representing genomic sequences.

II. Develop and evaluate Convolutional Neural Network (CNN) models for the categorization of genomic sequences.

III. Develop and evaluate explainable models to detect hidden mathematical signatures in the genome associated with the genotypic characterization of a sequence.

IV. Develop and evaluate a methodology for the identification of mechanisms associated with genetic recombination from a computational and bio-mathematical perspective.

V. Provide descriptive information on encrypted mathematical signatures in the genome associated with the genotypic characterization of sequences.

VI. Validate the developed methodology (genomic spectrograms + CNN + explainability) in applied cases for HIV-1 and SARS-CoV-2 viruses.

And with all of this, an exciting journey began.

15.2 Some Numbers

Roughly and approximately, I processed a total of 1,554,847 genomic sequences, generated a total of 1,568,401 genomic spectrograms, trained around 164 CNNs, with over 15,000 minutes of training, generated a phylogenetic tree of about 4,725 sequences, and produced a total of approximately 1,623,017 hot zones.

15.3 Main Contributions

I hereby present a compendium of the scientific contributions made to date.

Conference Papers:

Guerrero-Tamayo, Ana, Urquijo, B. S., Casado, C., Tosantos, M. D. M., Olivares, I., Pastor-López, I. (2023, August). Validating by Deep Learning an Efficient Method for Genomic Sequence Analysis: Genomic Spectrograms. In International Conference on Hybrid Artificial Intelligence Systems (pp. 281-292). Springer Nature Switzerland [3005].

Ana Guerrero-Tamayo, Borja Sanz Urquijo, Concepción Casado, María-Dolores Moragues Tosantos, Isabel Olivares, Iker Pastor-López. (2024). Unveiling HIV-1 U Sequences: Shedding Light Through Transfer Learning on Genomic Spectrograms. International Conference on Hybrid Artificial Intelligence Systems. Springer Nature Switzerland.

Ana Guerrero-Tamayo, Borja Sanz Urquijo, Concepción Casado, María-Dolores Moragues Tosantos, Isabel Olivares, Iker Pastor-López. (2024). The Third Codon Nucleotide's Role in Genetic Recombination Within SARS-CoV-2 Spike Protein: A Pilot Study. International Conference on Hybrid Artificial Intelligence Systems. Springer Nature Switzerland.

Research Papers:

Ana Guerrero-Tamayo, Borja Sanz Urquijo, Concepción Casado, María-Dolores Moragues Tosantos, Isabel Olivares, Iker Pastor-López. (2023). Discovering Mathematical Patterns Behind HIV-1 Genetic Recombination: a new methodology to identify viral features. IEEE Access [2850].

Ana Guerrero-Tamayo, Borja Sanz Urquijo, Concepción Casado, María-Dolores Moragues Tosantos, Isabel Olivares, Iker Pastor-López. (2024). Classification of SARS-CoV-2 sequences as recombinants via a pre-trained CNN and identification of a mathematical signature relative to recombinant feature at Spike, via interpretability. PLoS ONE 19(8): e0309391. https://-doi.org/10.1371/journal.pone.0309391 [3006].

Research Data:

Phylogenetic tree of all sequences comprising the HIV-1 Complete Consensus Multiple Sequence Alignment from the Los Alamos National Laboratory HIV-1 Database (until 2020). Ana Guerrero-Tamayo, Borja Sanz Urquijo, Concepción Casado, María-Dolores Moragues Tosantos, Isabel Olivares, Iker Pastor-López. (2023). Discovering Mathematical Patterns Behind HIV-1 Genetic Recombination: a new methodology to identify viral features - Supplementary Information. IEEE Dataport. https://dx.doi.org/10.21227/2917-1s55 [3007].

List of Coronaviruses in Section 11.2:

Outreach Activities:

Lecture entitled "Deep Learning Applied to Viral Evolution Prediction" para Computer Vision and Machine Learning at the University of Barcelona (CVMLUB). Delivered on 2023/06/16.

Lecture entitled "Modelos Deep Learning resilientes. Estratégias para uma gestão eficaz de viés em treinamentos de modelos Deep Learning". I - Workshop Internacional sobre Inteligência Artificial "Navegar, ou não, pelo Universo da Inteligência Artificial através de modelos Deep Learning. Universidade Politécnica de Coimbra (IPC), Portugal. Universidade Estadual do Paraná (Unespar), Brasil. Delivered on 2024/04/10.

Lecture entitled "Discovering Mathematical Patterns Behind HIV-1 Genetic Recombination: a new methodology to identify viral features" para Master's Degree in Computing and Intelligent Systems (MUCSI) at the University of Deusto. Delivered on 2024/05/16.

In the Appendix M, scientific evidence published during the course of this research is presented.

15.3.1 Chapter 3. Technological Background

In Chapter 2, I introduced a series of microbiological concepts to facilitate the understanding of subsequent chapters.

Building upon the preliminary analysis and deductive process described in Chapter 1, I proceeded to briefly analyze the state of the art in terms of analysis using genomic spectrograms, applications of Deep Learning in the biomedical

field, specifically in genomics, explainability in biomedical Deep Learning, and a critical analysis of both the current limitations of Multiple Sequence Alignment (MSA) algorithms and articles with questionable or outright incorrect application methodologies. The results of this analysis were presented in Chapter 3.

Thanks to this, besides verifying the robustness of the selected technologies, even in the biomedical field, I was able to establish the foundations of what should be a strict, rigorous, comprehensive, and robust methodology for the successful application of deep learning tools in genomics. I discovered the concept of bio-centric explainability, which, in my opinion, is essential in the field of applied science.

SCIENTIFIC CONTRIBUTION:

- Outreach Activity:

Lecture entitled "Modelos Deep Learning resilientes. Estratégias para uma gestão eficaz de viés em treinamentos de modelos Deep Learning". I - Workshop Internacional sobre Inteligência Artificial "Navegar, ou não, pelo Universo da Inteligência Artificial através de modelos Deep Learning. Universidade Politécnica de Coimbra (IPC), Portugal. Universidade Estadual do Paraná (Unespar), Brasil. Delivered on 2024/04/10.

SCIENTIFIC CONTRIBUTION:

- Popular Science Book:

Ana Guerrero-Tamayo, Fernando José Sadio-Ramos. (2024). Es posible saber Deep Learning sin ser ingeniero? Fundamentos y reflexiones sobre modelos generativos". Coimbra: Fernando José Sadio-Ramos (Editor). ISBN: 978-989-8486-52-3 DOI: http://dx.doi.org/10.5281/zenodo.12518081 [1961]

15.3.2 Chapter 4. Genomic Spectrogram

In Chapter 4, I analyzed with greater specificity the application of spectrograms (Sections 4.2), as well as their bio-mathematical implications (Sections 4.3 and 4.6). I reaffirmed the advantages of using this system for representing genomic sequences due to its alignment-free nature compared to traditional MSA algorithms (Section 4.4).

Based on this analysis, I designed two different types of spectrogram representations: concatenated and superposed. The specific characteristics of each representation system were specified in Section 4.5. Both representations provide complementary information. Concatenated spectrograms provide data about the relevant nucleotide types in the recombinant or non-recombinant feature, while superposed spectrograms allow for more precise identification of the genomic regions involved in these features.

15.3.3 Chapter 5. HIV-1 Genetic Recombination Analysis

In Chapter 5, I described the experiments conducted on all complete HIV-1 sequences from all subtypes obtained from the Los Alamos National Laboratory HIV Database repository (until 2020). HIV-1 was the specimen studied concerning genetic recombination.

In Section 5.4, I outlined the proposed research methodology, which primarily involved training a pre-trained CNN via Transfer Learning on the dataset of genomic spectrograms to classify a sequence (in this case, an HIV-1 sequence) as recombinant or non-recombinant. Subsequently, I applied explainability tools to identify the relevant areas in the genomic spectrograms where the CNN looked to make its classification (hot zones). These zones determined the location of the inherent mathematical signatures in the genome that the CNN was able to detect in the image.

The CNN classified complete HIV-1 sequences as recombinant and non-recombinant, in both concatenated and superposed spectrograms. Technical details regarding the battery of experiments conducted are available in sections 5.5, 5.6, 5.7, 5.8, 5.9, 5.10, 5.11, and 5.12.

In an initial set of experiments involving three different pre-trained CNNs (VGG-16, ResNET-101, and Inception V3), I observed consistently high performance ratios across all cases, with a slightly greater variability in Training Time. Notably, high Test Accuracy values were observed in all cases.

In the case of concatenated optimal hyperparameter configurations, the ranges of performance ratio results were those summarized in Table 15.1.

Table 15.1: Optimal Hyperparameter Configurations in Concatenated Spectrograms

Hyperparameter	**Optimal Values**
Validation Accuracy	94.82% - 94.92%
Total Test Accuracy	94.00% - 96.33%
Non-Recombinant Test Accuracy	96.67% - 99.33%
Recombinant Test Accuracy	89.33% - 95.33%
AUC	0.9772 - 0.9943
Training Time	17 min 36 sec - 148 min 1 sec

Analyzing the range of values in Training Time, in almost two and a half hours, in the case of the slowest architecture, I obtained specific training for classifying a complete HIV-1 sequence as recombinant or non-recombinant.

For superposed spectrograms, I observed similar performance metrics (Table 15.2).

Table 15.2: Optimal Hyperparameter Configurations in Superposed Spectrograms

Hyperparameter	**Optimal Values**
Validation Accuracy	92.90% - 96.83%
Total Test Accuracy	94.00% - 95.00%
Non-Recombinant Test Accuracy	95.33% - 99.33%
Recombinant Test Accuracy	90.00% - 92.67%
AUC	0.9879 - 0.9899
Training Time	15 min 59 sec - 149 min 6 sec

Training Time remains the most variable parameter, but in two and a half hours, I could have a trained model ready.

Once the VGG-16 was selected for having the highest balance in accuracy across both categories, and the robustness test was conducted on both datasets, with a positive result, the next step was to determine the hot zones in each category (Test Set) with the highest specificity possible.

The most relevant results from the VGG-16 on concatenated spectrograms were those of Table 15.3.

In the case of superposed spectrograms, the results were those of Table 15.4.

Table 15.3: Most Relevant Results of VGG-16 on Concatenated Spectrograms

Hyperparameter	**Optimal Values**
Validation Accuracy	94.92%
Training Time	17 min 36 sec
AUC	0.9772
Non-Recombinant Test Accuracy	96.67%
Recombinant Test Accuracy	91.33%
Total Test Accuracy	94.00%

Table 15.4: Most Relevant Results of VGG-16 on Superposed Spectrograms

Hyperparameter	**Optimal Values**
Validation Accuracy	92.90%
Training Time	15 min 59 sec
AUC	0.9888
Non-Recombinant Test Accuracy	95,33%
Recombinant Test Accuracy	92,67%
Total Test Accuracy	94.00%

As a measure of the balance degree in the results of both categories, I applied the total difference (concatenated + superposed) in the Inter-Category Test Accuracies. This ratio, called "Total Distance", was the lowest in the case of VGG-16, with only an 8.00%. This was the lowest compared to ResNET-101 (18.00%), and Inception V3 (19.65%). Therefore, it turned out to be the CNN with the highest balance in the accuracy degrees in the Test Set in both concatenated and superposed cases.

The CNN achieved a test accuracy of 94.00% when classifying complete HIV-1 sequences as recombinant or non-recombinant, and this level of accuracy is consistently maintained in both concatenated and superposed spectrograms.

To discern and extract the hot zones that occur most frequently with maximum precision, an individual scrutiny of interpretability results was conducted for each sequence in the Test Set, covering a total of 300 sequences. For sequences in the Test Set that were correctly classified by the CNN, I isolated the most common primary hot zones, as well as the secondary ones. In the case of sequences where the network predictions were incorrect, I identified the most recurrent erroneous primary hot zones, as well as the secondary ones.

I focused on Grad-CAM and LIME as they are more visually intuitive than Gradient Attribution.

Within the non-recombinant sequences, in the case of Grad-CAM, I identified the most relevant hot zones in the low-frequency range covering all four nucleotides, along with frequencies close to one-third for the nucleotide T.

In the case of recombinant concatenated spectrograms, the most relevant hot zone from Grad-CAM corresponded to the region of nucleotide A, extending to the lower third of the right edge, associated with the area of nucleotide T. Additionally, a shared area of interest was observed for both non-recombinant and recombinant sequences, located at $f = 1/3$ in the T nucleotide spectrogram.

In concatenated spectrograms, LIME consistently emphasized regions of high and low frequency associated with nucleotide A, for both non-recombinant and recombinant sequences. However, in several misclassification cases, it was observed that the CNN focused on locations that did not coincide with the established hot zones. This region of erroneous focus is situated at the average frequencies of nucleotide A.

For superposed spectrograms, the most potent hot zones in the non-recombinant feature corresponded to regions of high and low frequency, primarily matching the 5'LTR and 3'LTR. Secondary hot zones concentrated along the spectra of high and low frequencies.

In superposed spectrograms, for the recombinant feature, the main influential regions identified by Grad-CAM were situated within the low-frequency range and in proximity to $f = 1/3$ in the 5'LTR region.

In non-recombinant superposed spectrograms, LIME exhibited a pronounced focus on the low-frequency region in 5'LTR. To a lesser extent, LIME showed activity in the high-frequency area in 5'LTR, as well as in the low-frequency region in 3'LTR.

In recombinant superposed spectrograms, prominent regions were found around $f = 1/3$ in the region corresponding to 5'LTR and in regions closely associated with the *gag* and *env* genes.

The performance of the three explainability tools exhibits consistency in both concatenated and superposed spectrograms. The hot zones in each category in all cases were specific and distinguishable, indicating that the CNN detected discernible patterns for each analyzed category.

Furthermore, when the CNN made classification errors, both Grad-CAM and LIME precisely focused on the hot zones of the incorrect category. This finding suggests consistency in the CNN's behavior regarding the identification of areas of interest, even in situations where misclassification occurs.

I generated the phylogenetic tree of the total number of HIV-1 complete sequences via MATLAB 2021b - Computational Biology Apps. First, I downloaded the HIV-1 Complete Consensus Multiple Sequence Alignment from the Los Alamos National Laboratory HIV-1 Database (until 2020).

Subsequently, I conducted a thorough Biological Consistency Analysis. I was able to detect multiple examples that demonstrate the clear correlation between the degree of homology of the hot zones of two different sequences and their degree of phylogenetic proximity. The hot areas of sequences belonging to the same subtype were usually similar, and the degree of similarity was often greater with closer phylogenetic proximity. Intrapatient sequences were a clear example of this behavior.

This correlation between the degree of similarity in the highlighted areas of sequences of the same subtype and their phylogenetic proximity suggests that similarities in the regions of interest of the sequences are linked to their evolutionary relationship.

Furthermore, I detected that different sequences from different recombinant subtypes exhibit notably analogous critical points in both concatenated and superposed spectrograms. This phenomenon requires further research.

Interestingly, in multiple cases, the CNN misclassified the same sequence in both datasets. Specifically, sequence FJ389363, categorized as non-recombinant, and the recombinant sequences GU230127, HQ385479, JN864058, JX390977, MN172223, and MT559132 are incorrectly classified in both concatenated and superposed spectrograms. These findings could imply that these sequences exhibit distinct and

special biological characteristics that differentiate them from others. As a result, it is essential to conduct a comprehensive analysis of the integrity of their genome and their accurate classification within their designated subtypes. It is plausible that these double misclassifications are not random and may indicate underlying challenges in categorizing these sequences within their respective subtypes.

Since this research is entirely multidisciplinary, we should not limit ourselves to analyzing the methodology and results solely from a computational perspective. It is crucial to evaluate the results from microbiological and virological perspectives in accordance with the concept of bio-centric explainability. To this end, I conducted a comprehensive biological analysis comparing the results obtained with scientific evidence.

The relevance of the 5' and 3' LTR regions, along with their U3, R, and U5 regions, in classifying sequences as recombinant may be due to their crucial roles in viral transcription, regulating gene expression, and virus replication. These specific genomic elements in the frequency spectrum are essential for detecting recombinant features. Adjacent genes such as *gag*, *nef*, and *env* could also influence HIV-1 genetic recombination.

The importance of nucleotides A and T in recombination is linked to the high presence of polyA and polyT sequences in HIV-1, which are prone to polymerase skipping, thus facilitating genetic recombination. The prominence of nucleotide A may be related to the abundant adenine content in lentiviruses, especially in HIV-1.

The similarity in hot zones between sequences of different subtypes could be due to a higher incidence of genetic recombination than is currently recognized. It is plausible that certain viruses traditionally categorized as pure subtypes may harbor hidden recombination events.

In summary, the main hot zones for distinguishing between recombinant and non-recombinant sequences are the 5'LTR and 3'LTR regions, along with the nucleotide A, and to a lesser extent, nucleotide T. These findings are aligned with the biological characteristics of the HIV-1 genome structure and retrotranscription functions. This further supports the idea that the CNN identified mathematical

patterns in the frequency domain that define the recombinant or non-recombinant nature of a genomic sequence, along with their precise genomic locations.

SCIENTIFIC CONTRIBUTION:

- Outreach Activity:

Lecture entitled "Deep Learning Applied to Viral Evolution Prediction" for Computer Vision and Machine Learning at the University of Barcelona (CVMLUB). Delivered on 2023/06/16.

SCIENTIFIC CONTRIBUTION:

- Outreach Activity:

Lecture entitled "Discovering Mathematical Patterns Behind HIV-1 Genetic Recombination: a new methodology to identify viral features" for Master's Degree in Computing and Intelligent Systems (MUCSI) at the University of Deusto. Delivered on 2024/05/16.

SCIENTIFIC CONTRIBUTION:

- Research Data:

Phylogenetic tree of all sequences comprising the HIV-1 Complete Consensus Multiple Sequence Alignment from the Los Alamos National Laboratory HIV-1 Database (until 2020).

Ana Guerrero-Tamayo, Borja Sanz Urquijo, Concepción Casado, María-Dolores Moragues Tosantos, Isabel Olivares, Iker Pastor-López. (2023). Discovering Mathematical Patterns Behind HIV-1 Genetic Recombination: a new methodology to identify viral features - Supplementary Information. IEEE Dataport. https://dx.doi.org/10.21227/2917-1s55 [3007].

SCIENTIFIC CONTRIBUTION:

- Conference Paper:

Ana Guerrero-Tamayo, Borja Sanz Urquijo, Concepción Casado, María-Dolores Moragues Tosantos, Isabel Olivares, Iker Pastor-López. (2023). Validating by Deep Learning an Efficient Method for Genomic Sequence Analysis: Genomic Spectrograms. International Conference on Hybrid Artificial Intelligence Systems (pp. 281-292). Springer Nature Switzerland [3005].

SCIENTIFIC CONTRIBUTION:

- Research Paper:

Ana Guerrero-Tamayo, Borja Sanz Urquijo, Concepción Casado, María-Dolores Moragues Tosantos, Isabel Olivares, Iker Pastor-López. (2023). Discovering Mathematical Patterns Behind HIV-1 Genetic Recombination: a new methodology to identify viral features. IEEE Access [2850].

15.3.4 Chapter 6. Analysis of Total Hot Zones in HIV-1 by Subtype

In light of the results obtained in Chapter 5, I wanted to deepen our understanding of the inherent mathematical patterns in classifying sequences by subtype as recombinant. Furthermore, aligned with the aforementioned, my aim is to delve into our intrinsic understanding of how the CNN operates in classifying these genomic sequences as recombinant or non-recombinant.

I established a comparison between the classification error rates associated with each subtype and the total hot zones resulting, not only from the sequences comprising the Test Set but also from the 13,554 complete HIV-1 sequences identified up to 2020.

The results of all these are detailed in Chapter 6. In Section 6.3, I outlined the criteria for selecting representative subtypes, the procedures for calculating error rates, and the calculation and graphical representation of the total hot zones by category and by subtype. The results regarding error rates in representative subtypes were detailed in the results presented in Section 6.4.

Firstly, I computed the total Grad-CAM hot zones in concatenated and superposed spectrograms, as this explainability tool is the most visual and intuitive. To establish the total Grad-CAM hot zones, I performed a scalar summation of the score maps for each of the sequences using MATLAB 2021b. The results were presented in Section 6.5.

The calculation of the total hot zones per category involved scalar summation of the score maps of all sequences included in that category (11,163 non-recombinant and 2,391 recombinant). Likewise,the calculation of the total hot zones per subtype was performed by scalar summation of the score maps of all sequences included in each subtype.

Each category exhibited unequivocal patterns in its total hot areas, both in concatenated and superposed spectrograms.

In concatenated spectrograms, the primary hot zones of the non-recombinant category were located in the low-frequency range from the second half of the A, G, and C spectrograms, and the first half of the T spectrogram. Regarding secondary zones, the main total hot zone is found at $f = 1/3$ in the T spectrogram. Even more secondary zones are visually identified at $f = 1/3$ around the C spectrogram, in the frequency range from $f = 1/6$ to $f = 1/3$.

The main total hot areas in the recombinant category were primarily located around the range between $f = 1/9$ and $f = 1/6$ in the A spectrogram and in the high-frequency range around A, extending up to the first third of the G spectrogram. A lower-intensity total hot area was also detected at $f = 1/3$, also in the A spectrogram. Frequencies outside of these total hot areas corresponded to secondary total hot areas, likely due to the radiation from the more prominent hot areas, all within the A spectrogram. Other secondary areas were detected at $f = 1/9$ in the transition zone between the C and T spectrograms, as well as at $f = 1/3$ and high frequencies in the second half of the T spectrogram. There are very small, practically imperceptible areas in the high-frequency range from the second half of the G spectrogram to the second half of the T spectrogram, as well as at $f = 1/3$ from the C spectrogram to the second half of the T spectrogram.

Operating with superposed spectrograms, in the case of the non-recombinant category, the main total hot area was clearly identified in the high-frequency range centered on the 5'LTR, extending up to the first half of *gag*. This area radiated up to the first third of *pol* as a secondary area. The next most significant total hot area was located in the low-frequency range centered on the 3'LTR, extending towards the vicinity of the start of *nef*. I could observe more secondary areas in the low-frequency range from the 5'LTR to the first half of *gag*, and in the high-frequency range from the second half of *pol* to the beginning of the second half of *env*, as well as from the final area of *env* to the end of the 3'LTR.

Regarding the recombinant category in superposed spectrograms, I clearly identified the main total hot area in the vicinity of $f = 1/3$, with a distinctive epicenter in the 5'LTR, radiating up to the first half of *gag*. The second most significant total hot area was located in the low-frequency range in *gag*. This total hot area radiated up to, on one side, the first half of *pol*, and on the other side, to the initial stretch of the 5'LTR, connecting with another less powerful total hot area at the beginning of the 5'LTR. I perceived very secondary areas, barely visible, in the low-frequency range from the second half of *env* to the end of the 3'LTR, as well as at $f = 1/3$ in the 3'LTR.

The observation of a clear and significant distinction between the total hot areas across the dataset between the non-recombinant and recombinant categories suggests highly discriminative representation by the CNN. This phenomenon indicates that the CNN detected intrinsic and specific characteristics of each category. These clearly distinguishable patterns were generated both in concatenated and superposed spectrograms. The repetition of the phenomenon using two different representations reinforces the possible existence of mathematical signatures embedded in the genome that determine, in this case, whether a complete HIV-1 sequence is recombinant or not.

Regarding the total hot zones per subtype (Section 6.6), I analyzed the degree of homology of the total hot zones of each subtype with the corresponding total

hot zones of its category. I also contrasted the results with expert literature as a validation of the biological coherence of such results.

I determined the coherence between the homology of the hot zones in the individual sequences, the total hot zones per representative subtype, and between the total hot zones per category. From any Grad-CAM result image, it was possible to identify at first glance whether it belongs to the non-recombinant category or the recombinant category, whether it is an individual sequence or a subtype.

When analyzing the total hot zones per subtype, I detected greater stability and a clearer correlation between the degree of homology among the hot zones per subtype and the total hot zones for non-recombinant sequences. I also observed a correlation with the CNN's accuracy rate in each subtype under study.

In general terms, the differences between the hot zones per subtype and the corresponding total hot zones for their category were more easily noticeable in concatenated spectrograms than in superposed ones. A clear example of this is the case of subtype A6.

In the case of non-recombinant spectrograms, the hot zones using superposed spectrograms were very stable, both in relation to their own results in concatenated spectrograms and in relation to the results obtained in the case of recombinant subtypes.

Within the non-recombinant category, subtypes F1, G, O, and U exhibited peculiar results in their total hot zones.

G is classified by the HIV-1 expert community as problematic due to a potential recombinant origin, and U refers to all sequences that are simply unclassified or unclassifiable. The high error rates in these two cases and the discrepancies in the hot zones, far from indicating poor performance, clearly point to an even more coherent behavior than expected.

In the case of subtype G, certain patterns corresponding to the recombinant category were distinguished. This also occurred in the case of subtype U, where the hot zones were also noisy. Therefore, I found these two cases and the behavior of the CNN fascinating. As a result, I decided to study in depth the presence of

possible hidden genetic recombination events in sequences classified as U (Chapter 8) and explore the recombinant origin in subtype G through the analysis of the hot zones of the sequences comprising it (Chapter 9).

Regarding subtype F1, I did not find an apparent reason for the abnormal error rate observed in this subtype. As for subtype O, labeled as an "Outlier" among experts (initially it was considered an atypical strain of HIV-1 compared to subtypes A, B, C, etc.), further investigation will be necessary to determine if there is a connection between its atypical character and the limited correlation between the high accuracy rate and the degree of convergence between its total hot zones and those of its total category.

The hot zones of each of the analyzed recombinant subtypes were more variable with respect to the total hot zones in the recombinant category. The differences were more evident at first glance compared to the totals. This variability was again more noticeable in concatenated spectrograms than in superposed ones. This may be due to greater genomic variability in the set of sequences and recombinant subtypes compared to non-recombinants.

It is worth noting that there are over 100 types of recombinant CRFs, of which only nine can be considered representative recombinant subtypes (composed of more than 15 sequences). Therefore, the number of non-representative recombinant subtypes considered is quite high. This loss of information, combined with the greater genetic variability inherent in recombinant CRF subtypes, logically can affect the patterns in the total hotspots per subtype. Consequently, it will also affect the degree of homology between the total hotspots per subtype and the total recombinants, which include the genetic variability associated with all CRFs.

Therefore, once again, the results obtained by the CNN and interpretation tools were aligned with the biological reality affecting these subtypes.

The total number of sequences in the analyzed set of representative recombinant subtypes was 1,855, representing 77.58% of the total 2,391 complete recombinant sequences up to 2020. It is possible that the remaining 22.42%, although a minority in relation to the total, may have associated genetic variability that interferes

enough in the correlation between the total hotspots of the recombinant category and the total hotspots per subtype, especially considering the genetic variability among the analyzed representative recombinant subtypes.

This fact calls for further investigation, expanding this study beyond representative subtypes, including all recombinant and non-recombinant subtypes, and delving into comparisons of their hotspots.

It would also be intriguing to study the similarity relationships among the total hotspots of different subtypes and establish a comparison with the phylogenetic relationships among them, in order to deepen the virological plausibility of the Grad-CAM results.

In the case of recombinants, I did not observe a direct relationship between the CNN error rate and the degree of divergence between the hotspots and the total recombinants. Regardless of this error rate, there were slight differences, but no clear direct proportionality was perceived. For example, in the case of CRF11_cpx, the error rate reached 12%. However, its total hotspots were not particularly different. On the contrary, in the case of CRF20_BG, the error rate was 0%, yet the divergences between the hotspots were perfectly noticeable. Although all of them could be clearly considered as recombinants, the existing differences were mild.

In recombinants, the stability of the hotspot at $f = 1/3$ in the 5'LTR (superposed spectrograms) was particularly striking. This main hotspot remained virtually unchanged in all representative recombinant subtypes, except in the case of CRF20_BG. Considering the genetic variability associated with recombinant subtypes, as well as the variability observed in the total hotspots per subtype, this stability at $f = 1/3$ in the 5'LTR becomes even more noteworthy.

There are subtypes like CRF20_BG and CRF42_BF1 whose hotspots per subtype were notably similar. It would be interesting to dedicate possible future research to exploring potential relationships between these similarities in hotspot patterns of these subtypes and the existence of perhaps unknown phylogenetic relationships among them.

Overall, this investigation of the total hotspots per subtype provided additional information not only about the degree of coherence in the CNN's behavior but also intriguing clues to deepen the virological understanding of different subtypes of HIV-1.

The high error rates in the classification and/or the divergent hot zones were consistent with the problematic nature of the corresponding subtypes. This again demonstrates that the CNN operates cohesively.

15.3.5 Chapter 7. A Brief Analysis of HIV-1 Subtypes F1, CRF22_01A1, and O

Based on the results of the comparison presented in Chapter 6, Section 6.6, it was compelling to explore the relationship between the divergences identified among the total hot zones by category and those of specific subtypes.

In this Chapter, I conducted a brief analysis of three interesting subtypes:

- Subtype F1 with 82 complete sequences.
- Subtype CRF22_01A1 with 21 complete sequences.
- Subtype O with 57 complete sequences.

Subtypes F1 and CRF22_01A1 exhibited notable error rates, exceeding 15%. The total hot zones of these two subtypes reflected divergences from their respective categories. For CRF22_01A1, there were divergences relative to the recombinant category. However, in the case of subtype F1, despite its high error rate, the divergences from the non-recombinant category were not as apparent as in the case of CRF22_01A1.

Regarding subtype O, the difference between its total hot zones and those corresponding to the non-recombinant category was striking, despite its very low error rate and the high confidence scores in the CNN's classification of subtype O sequences as recombinant. Subtype O is termed "Outlier" because it is much less frequent and exhibits significant genetic divergence compared to the subtypes of group M (the most common).

My analysis involved examining the hot zones of sequences from each of these three subtypes, identifying possible groups with common features in their individual

hot zones, and comparing them with the total hot zones of the subtype and the corresponding category.

In this case, I focused solely on conducting a preliminary analysis, meaning that no phylogenetic analysis or comparison with existing literature was performed, as was done in Chapter 8 with subtype U and in Chapter 9 with subtype G.

In general terms, for subtypes F1 and CRF22_01A1, the differences between their total hot zones and those corresponding to their categories matched the patterns found in the hot zones of incorrectly classified sequences. This alignment highlights the consistent and robust performance of the CNN.

Additionally, by analyzing the hot zones and confidence scores of the sequences from these subtypes individually, I identified interesting patterns in the hot zones of some sequences that may provide insights into possible underlying phylogenetic relationships between sequences and subtypes.

The high error rates of subtype F1 and the specific characteristics observed in terms of variability and sub-patterns in the Grad-CAM hot zones of its sequences, along with certain similarities with sequences from other subtypes (such as the sequences subtype CRF22_01A1 EU743963 and JN864049 with subtype F1 sequences KJ883150, JN864059, and KJ849791), make F1 a case of great interest for understanding aspects such as the genetic evolution of HIV-1, the impact of genetic recombination on this evolution, and for expanding knowledge about phylogeny and the relationships among subtypes and their sequences. Future research will need to delve deeper into the connections between the obtained results and phylogeny to identify factors that might influence the observed phenomena, as well as assess their potential impact on revising the current classification of sequences and subtypes.

Regarding subtype CRF22_01A1, its high error rates and high confidence scores in both the classification of sequences as non-recombinant and recombinant suggest that the CNN identified clear patterns in multiple sequences, making a review of this subtype interesting. Its variability and the detected sub-patterns, as well as the fact that all sequences misclassified in superposed spectrograms were also misclassified in concatenated spectrograms, highlight the diversity and unique behavior of this

subtype in the experiments. After reviewing the consistency of the obtained results and the impact of the hot zones on the calculation of the total hot zones for this subtype, it is concluded that this subtype is extremely interesting for deepening the understanding of the relationships between different HIV-1 strains from the perspective of genetic recombination.

The case of subtype O (referred to as "Outlier") was even more peculiar. Despite its very low error rates and high confidence scores in correctly classifying most of its sequences, there were clearly notable differences between the hot zones of this subtype and those corresponding to the recombinant category. By analyzing the hot zones of the subtype O sequences, I detected certain peculiar sub-patterns that explained the divergences in the total hot zones of this subtype. Therefore, the peculiar Grad-CAM results showed a high consistency with the distinctive profile of this subtype compared to other HIV-1 subtypes. The identified sub-patterns may indicate not only phylogenetic relationships among the analyzed sequences, such as different evolutionary branches, but also connections with sequences from other subtypes. For example, there was a detected similarity between a sub-pattern in the hot zones of multiple Subtype O sequences and the hot zones of some sequences from the CRF22_01A1 subtype, such as EU743963 (which was also misclassified as non-recombinant). Similarities were also found with some subtype F1 sequences, such as AB485656, FJ771006, or JX140671. According to the results obtained by the CNN, the non-recombinant nature of this subtype seems clear. However, the peculiarities of its hot zones reveal a series of unique characteristics of this subtype. This phenomenon is fascinating as the results of my methodology appear to even provide complex phylogenetic data on this subtype.

Understanding this "Outlier," its origin and evolution can complete and clarify the evolution of HIV-1 itself, its genetic diversity, and the current classification of subtypes and sequences.

Certainly, my next step regarding these three subtypes will undoubtedly be to integrate phylogenetics and a thorough review of the existing scientific literature into this analysis. These are three undoubtedly fascinating subtypes.

15.3.6 Chapter 8. Analysis of HIV-1 Subtype U

In Chapter 6 I found that the total hot zones of subtype U were quite divergent, so I considered it of interest to carry out a thorough analysis of each of the sequences composing this subtype. As detailed in Section 8.3, in this case study, I focused solely on the Grad-CAM hotspots, given the visual clarity of the results obtained with this interpretation tool. Likewise, I exclusively worked with superposed spectrograms due to their greater visibility to the human eye compared to concatenated spectrograms.

Based on the phylogenetic position of all subtype U sequences in the complete HIV-1 sequence tree, I studied their proximity relationships and interpretability patterns. This analysis was complemented with data from samples recorded in the Los Alamos Sequence Database.

I analyzed each branch of the phylogenetic tree where the U sequences are located, both from a phylogenetic perspective and in terms of their similarity to the hotspots of phylogenetically close sequences.

I also examined the CNN classification and the degree of convergence of the hotspots with the total hotspots corresponding to their category. Additionally, I scrutinized in detail the level of homology of the hotspots both in subtype U sequences and in phylogenetically close sequences from other subtypes (Section 8.5).

I endeavored to ascertain the presence of hidden genetic recombination events in sequences classified as U, aiming to unearth potential ancestral relationships among U sequences and their phylogenetically close counterparts. Furthermore, I delved into exploring latent connections among non-U subtypes present in the regions of the phylogenetic tree under scrutiny.

In my quest for understanding, I embarked on a thorough investigation of existing microbiological evidence, striving to gauge the alignment between the results generated by the CNN and the nuanced expertise in microbiology and virology. This rigorous examination spanned across various sections, namely 8.6, 8.7, 8.8, and 8.9.

The CNN classifications, whether as recombinant or non-recombinant, and their corresponding patterns proved to be phylogenetically coherent, mirroring the data gleaned from the samples. Moreover, the level of similarity correlated with

the phylogenetic proximity, akin to how branches on a tree reflect the shared lineage of its leaves.

In several cases, most CRF subtypes phylogenetically close to U subtype sequences are recombinants that include fragments categorized as U. This is logical. However, delving into the specifics of genomic similarities between these U fragments of recombinants and complete U subtype sequences would be beneficial to determine the precise reasons behind this evolutionary proximity. Additionally, acquiring additional data could shed light on the origin of these U sequences.

I detected a particularly interesting case where, according to the CNN results, there appears to be a hidden recombination phenomenon in intra-patient sequences of U subtype labeled EF029066, EF029067, EF029068, and EF029069 (Patient Code: H10986).

The CNN classified EF029066 as non-recombinant with a confidence level that is almost random, at only 51.424%. However, the CNN classified the three remaining intra-patient sequences as recombinants, with moderate confidence levels ranging from 69% to 76%. The confidence levels of these three sequences were relatively consistent with those of the recombinant category. Their very high degrees of homology were in line with being intra-patient (maximum phylogenetic closeness). The hotspots of EF029066 were consistent with the non-recombinant feature and, consequently, differed from its three remaining intra-patient sequences.

It is possible that I detected a hidden genetic recombination phenomenon in these intra-patient sequences. A detailed study of the differences between recombinant and non-recombinant sequences, as well as the identification of genomic elements determining the recombinant feature, could provide insights into the mechanisms driving and producing genetic recombination in HIV-1. Therefore, it is an extremely interesting case study deserving further investigation.

My findings aligned with multiple microbiological evidence, as seen in EF029066, EF029067, EF029068, EF029069, FJ388921, HM215249, HM215251, JF683772, JX140679, KY392775, MN153489, MN153492, MN736701, and MN736702. Their interpretation patterns were consistent with phylogeny and multiple biological

evidence (60.870% of the total U sequences). 34.783% resulted rather inconclusive (AB485664, AB485665, KP718920, KY392773, KY392774, KY392780, MH078559, and MK095236). Sequence AY046058 was the most controversial case (4.348%).

I was also able to establish similarities between interpretation patterns of different subtypes whose sequences are phylogenetically close. This can help us search for common genomic elements in hotspots flagged by the CNN. My results also reflect the primary role of genetic recombination in the phylogeny of HIV-1.

Many of the CRF cpx involved in this research contain sections of subtype G. The relationship between subtypes U and G may denote some level of relatedness between them. And the possible recombinant origin of subtype G could influence the interpretation patterns of subtype U.

Sequence MN736702 is an excellent case study for making a direct comparison between the signatures associated with recombinant and non-recombinant categories in a very specific genomic area (low frequencies of 5'LTR) and obtaining more bio-mathematical information. Perhaps we can unravel those mathematical signatures that define a sequence as recombinant and better understand the genetic recombination of HIV-1.

SCIENTIFIC CONTRIBUTION:

- Conference Paper:

Ana Guerrero-Tamayo, Borja Sanz Urquijo, Concepción Casado, María-Dolores Moragues Tosantos, Isabel Olivares, Iker Pastor-López. (2024). Unveiling HIV-1 U Sequences: Shedding Light Through Transfer Learning on Genomic Spectrograms. International Conference on Hybrid Artificial Intelligence Systems. Springer Nature Switzerland.

15.3.7 Chapter 9. Analysis of HIV-1 Subtype G

Like Chapter 8, Chapter 9 was an in-depth exploration of the results obtained in Chapter 6. In this case, the subtype under study was subtype G.

Subtype G is classified among those considered "problematic." Up to this point, all characterized strains of subtype G share an equally ambiguous relationship with

subtype A in their accessory region, indicating a possible recombinant origin for this lineage. However, due to the absence of a representative "pure" strain of subtype G and the limited length of the region in question, the designation of the subtype G lineage is maintained rather than redefining it as recombinant. Although the decision is to maintain the classification of subtype G as non-recombinant, the exception of the accessory genes of subtype G as non-recombinant remains controversial.

Regardless of the naming decisions made, recent studies assert that subtype G has a recombinant origin, and clearly indicate the classification of subtype G as a Circulating Recombinant Form (CRF). Other studies suggest that the ability of subtype G to undergo recombination may stem from the formation of heterozygous viral particles or the creation of mixed functional chimeras with another strain of HIV-1, resulting in the emergence of a novel and complex CRF.

Through a CNN, I identified singularities in a specifically problematic subtype with a potential recombinant origin. In this Chapter, I went beyond mere classification and identification of areas of interest. I am substantiating the conclusions of multiple studies on the manifest peculiarities of subtype G from a completely innovative perspective.

The error rate exhibited by the CNN in subtype G may be related to the recombinant origin of the subtype, justifying an in-depth analysis of this issue. It is conceivable that this error rate goes beyond simple inaccuracies on the part of the CNN; rather, this neural network is detecting genomic components related to the frequency spectrum, indicating the potential recombinant origin in specific sequences.

There were multiple factors that caught attention besides the high error rate. These include the presence of numerous cases where the non-recombinant feature is associated with low confidence levels, the similarity between the hotspots of sequences identified in different categories, the phylogenetic clustering among sequences with similar CNN result peculiarities, and their hotspots, among others.

It is interesting to observe the significant variability in confidence levels, both in the classification as recombinant or non-recombinant. This subtype exhibited remarkable diversity in multiple aspects, not only confined to classifications.

Through a novel approach, I was able to provide complementary information to guide future research in unraveling the mysteries surrounding subtype G.

I explain the analysis methodology in Section 9.4, which is very similar to that outlined in Chapter 8.

Phylogenetically, subtype G sequences form two main groups, while sequence MN153484 remains isolated from the rest. I analyzed the phylogeny of these two groups and their respective interpretation patterns.

The phylogenetic analysis showed a clear tendency towards proximity of both recombinant and non-recombinant sequences. This pattern aligns with previous observations on the relationship between the similarity of areas of interest and phylogenetic proximity, as discussed in Chapter 5, Section 5.15.2.

Overall, sequences classified as recombinant by the CNN tend to cluster within the same phylogenetic zones. This trend may arise from a common, at least partial, origin among them. It would be interesting to delve into the relationship between this phenomenon and scientific studies proposing a recombinant origin in this subtype.

A curious and recurrent phenomenon occurs in multiple cases of subtype G, where the CNN classified phylogenetically related sequences into disparate categories. One of these sequences was classified as recombinant, while its counterpart was labeled as non-recombinant. The repeatability of this phenomenon requires exploration of the underlying reasons in these cases.

The analysis of intra-patient sequences yielded surprisingly diverse results. I identified highly coherent cases, where the hotspots are notably similar. However, I detected cases of non-recombinant sequences with markedly different hotspots. This variability in hotspots requires further investigation to attempt to unravel the reasons behind such divergence.

Additionally, I identified multiple case studies where the CNN produced conflicting classifications. These latter cases are particularly intriguing, especially considering their recurrent detection. This phenomenon suggests a possible association with hidden genetic recombination events and/or the recombinant origin of this subtype.

Through the pre-trained CNN I also pinpointed potential cases of latent genetic recombination even within intra-patient sequences. It would be of interest to note the exact location of the genomic regions defining the recombinant feature in this context and investigate their possible relationship with the recombinant origin of subtype G.

As for the isolated G sequence MN153484, its phylogenetic relation to the subtype U sequence MN736702 prompted an analysis in Chapter 8, Section 8.5.10. This sequence is exceptional from a phylogenetic standpoint. It is isolated from the rest of its subtype sequences but is found in a region of the tree predominantly occupied by recombinants related to the CRF02_AG recombinant subtype, whose parental subtypes are A and specifically G. We should also take into account the phylogenetic proximity of a subtype U sequence. It would be interesting to meticulously determine the homologies between their respective genomes to try to identify homologous genomic regions that provide new evolutionary perspectives among the various subtypes involved in this case study, as well as additional information about the origin of subtype G.

Based solely on the phylogenetic situation, the phylogenetic region where it is located is a very interesting case study that can provide information on undiscovered details about the origin and evolution of subtype G, as well as its relationships with other subtypes.

In this case study, there are several exciting factors to consider. First, the phylogenetic isolation of this sequence from the rest of its subtype sequences. This could suggest that this particular sequence possesses a set of genomic characteristics that make it divergent from other sequences within its subtype. A comprehensive analysis is crucial to uncover the reasons for this divergence between this sequence and the other subtype G sequences. Furthermore, delving into the exact relationship between this sequence and the subtype U sequence MN736702 is of particular interest.

On the other hand, it is surprising that the CNN classified this sequence as non-recombinant when its entire phylogenetic environment is composed of recombinant sequences. Even the subtype U sequence MN736702 is classified as recombinant, albeit with a confidence score of 52.86% (Chapter 8, Section 8.5.10, Table 8.26).

A comparative analysis between the MN153484 sequence of subtype G and the adjacent sequences may shed light on the genomic elements that are divergent from those in other sequences, determining its classification as non-recombinant.

The hot zones of MN153484 closely resemble those of the subtype G sequences AB287003, AB287004, and to a lesser extent, MN486044. These three sequences are located in the phylogenetic area corresponding to Group 1. It would be interesting to investigate the reasons behind these high similarities in their hot zones to uncover elements that might reveal unknown phylogenetic relationships.

Identifying the genetic elements that cause this evolutionary separation from the rest of its subtype sequences, as well as determining the exact relationship with the subtype U sequence MN736702, could provide valuable insights into the evolutionary ramifications affecting both subtype G and its relationship with sequences categorized under subtype U.

Another extremely intriguing case is the G sequence JN248591, classified by the CNN as recombinant with a confidence level of only 51.692%. Its hot zones differ notably from the interpretability patterns of all other G sequences. A detailed genomic analysis would be very interesting to investigate the reasons that make this sequence unique within its subtype regarding its hot zones.

The ability to accurately group areas of interest, identifying clearly discernible sets composed predominantly of recombinant sequences on one side and non-recombinant sequences on the other, is remarkable. The diverse and distinguishable interpretability patterns within these groups can provide valuable insights into the underlying evolutionary ramifications in this subtype, with the patterns observed in the recombinant sequence groups being particularly interesting.

15.3.8 Chapter 10. Exhaustive analysis of SARS-CoV-2 Genomic Elements

Once the research block on genetic recombination in HIV-1 was completed and before applying my methodology to SARS-CoV-2, I wanted to conduct a detailed analysis of the genomic elements of SARS-CoV-2. This analysis is based on

statistical approaches to determine the degree of conservation of each element as a measure of its stability.

This detailed analysis is merely an approach to the genomic regions most prone to undergo modifications, whether mutations or genetic recombination events. These measures can serve as a probabilistic reference for the phylogenetic ramifications produced (viral evolution analysis).

My first step consisted of conducting a comprehensive analysis of each genomic element present in the complete genome of SARS-CoV-2, as detailed in Section 10.2. This analysis is fundamental in the design of a specific architecture in the genomics field, especially for any interpretation of results.

As detailed in Section 10.4, using a Python environment, I processed complete SARS-CoV-2 sequences, detecting the tags of each genomic element in their EMBL file and breaking down the sequence into its parts according to those labels (Section 10.4.1).

I began with a dataset consisting of FASTA files of 97,236 complete SARS-CoV-2 sequences. From this sequence set and based on the data contained in the EMBL files, I carried out a two-phase filtering to avoid analysis bias due to incorrect data. Detailed information about this process is available in Section 10.4.2.

In the first phase, I removed all sequences in which structural and non-structural proteins were not labeled in the EMBL file, indicating the start and end positions for each of them. In other words, sequences in which at least one of the following occurred in their EMBL file:

- Absence of ORF labels.

Or:

- Partial labeling. That is, only the labeling of ORF1ab was present, with its start and end positions. But not for all the proteins included in this polyprotein.

In a second phase, I filtered out sequences where, in addition to all their elements being fully labeled, the start and end positions were correct. To do this, I worked by comparing the first three nucleotides and the last three nucleotides for each genomic element. I removed all sequences whose EMBL file is either unlabeled

or partially labeled, for example, if there is no breakdown in the non-structural proteins (nsps) within ORF1ab. After this elimination, the number of resulting operational complete sequences was reduced to 39,138.

Finally, I removed sequences whose labels contain errors. For example, errors in the start or end position of any of their genomic elements. An error in at least one genomic element results in the removal of the corresponding sequence. After this filtering, I was left with 38,047 sequences.

I determined the first and last triplet of each genomic element and checked their biological coherence, among others, matching the start of ORFs with codon start and the end with codon stop.

The next step was the breakdown of each genomic sequence into its genomic elements separately to facilitate mass comparison by element. All this without losing traceability between each element and the sequence to which it belongs (Section 10.4.3).

Then I conducted a comprehensive comparison of all the filtered complete sequences, for each genomic element, identifying the variations between the sequences per element, and calculating these parameters:

- Average Length for each genomic element.

- Standard Deviation of the length of each genomic element. Low values of this ratio indicate higher invariability.

- Length Consistency Index. Expressed in %, high levels of this ratio indicate greater consistency in terms of the length of the genomic sequence.

- Nucleotide Consistency Index. Similar to the previous ratio, high values denote greater consistency, this time in terms of nucleotide matching. It was also expressed in %.

In this way, I determined the conservation degrees from different approaches, as well as the genotypic peculiarities of each genomic region. This knowledge will contribute to the design of specific architectures for a coherent generation of evolutionary predictions, taking into consideration the microbiological particularities of each gene element, especially considering the limited depth of current genome knowledge.

All this analysis was carried out by comparing at each sub-phase with scientific literature as a measure of validation of the biological coherence of the obtained results. The detailed results are available in Sections 10.5 and 10.6.

The lowest value of the Length Consistency Index (excluding ORF8/N) was ORF8, with 14.96%, followed by ORF6 (38.71%).

The highest value of the Length Consistency Index was found in nsp5, nsp9, nsp10, nsp16, M, and the intergenic regions ORF1ab/S, S/ORF3a, ORF3a/E, M/ORF6, as well as in the pseudoknot stem-loop 1 of the 3'UTR and the pseudoknot stem-loop 2 of the 3'UTR. In all these cases, it reached 100%.

The lowest value of the Nucleotide Consistency Index was in S (0.23%). This is an extremely low value, especially considering the relative stability regarding the proportions related to length (length of 88.35 ± 32.21 versus a Length Consistency Index of 98.44%). Precisely, this protein has the highest mutation rate among the rest of the genes composing SARS-CoV-2, both fixed and polymorphic, likely highly influenced by the deamination of C to U.

Therefore, the obtained results were aligned with the biological specificities of this protein.

Nsp1, nsp3, nsp4, and N also resulted in Nucleotide Consistency Index values below 1%. In nsp1, mutations effectively occurred leading to variations in structural stability and deletions that led to new strains. Nsp3 experienced high numbers of unique mutations, and insertion/deletion hotspots were detected in it. The genomic region corresponding to nsp4 is sensitive to non-synonymous mutations (G9802U/C9803U). N has accumulated up to 1,034 unique nucleotide mutations.

The highest values of the Nucleotide Consistency Index were for the pseudoknot stemLoop 1 and 2 of the 3'UTR, with values of 38.89% and 41.38% respectively, the intergenic region S/ORF3a with 60.00%, the intergenic region ORF3a/E with 53.85%, and the intergenic region M/ORF6 with 75.00%. It was notable that the regions with the highest nucleotide consistency were intergenic zones. In all these fragments, the length homology was 100%.

My results were consistent with the detected mutation rates, particularly in nsp7 and nsp8, which in my analysis were highlighted as the most consistent regions in terms of length and nucleotide matching. Regarding the less conserved regions, similarly, especially in S, N, nsp3, and nsp4, the results demonstrated coherence in the degree of nucleotide matching.

This pre-analysis can serve as a foundation for future research aimed at the effective implementation of these algorithms in predicting the most likely viral evolutions in SARS-CoV-2, as well as in other viruses or microorganisms in future stages of research.

15.3.9 Chapter 11. SARS-CoV-2 Genetic Recombination Analysis

Chapters 5 to 9 were dedicated to the case study of genetic recombination in HIV-1, the virus specimen under study. Using pattern detection in HIV-1 genomic spectrograms as the learning basis, I began the study of genetic recombination in SARS-CoV-2.

As explained in Section 11.3.1, in March 2023, I downloaded from the public repository NCBI Virus Database all complete sequences of SARS-CoV-2 by variant [40]. The total number of sequences downloaded was 1,541,293. Of these, 1,539,728 were categorized as non-recombinant, and 1,565 were recombinant(Section 11.3.1).

To address the imbalance between categories, I performed random subsampling in the non-recombinant category. The details of this are developed in Section 11.3.2. Subsequently, I generated the superposed genomic spectrograms of all sequences involved in all subsamplings (Section 11.3.3). With these, I constructed the datasets (Training, Validation, and Test Sets).

Next, I applied a Two-Stage Transfer Learning. My starting point was a pre-trained VGG-16 network using the ImageNET dataset. In Phase 1, I applied Transfer Learning to this pre-trained model to detect the recombinant feature in the dataset of genomic spectrograms of complete HIV-1 sequences. In Phase 2, I applied Transfer Learning once again to the network resulting from Step 1

(VGG-16 HIV-1) using a dataset of genomic spectrograms of complete SARS-CoV-2 sequences to also detect the recombinant feature (VGG-16 SARS-CoV-2) using MATLAB2021b App Deep Learning Designer (Section 11.3.4) and conducted the designed Test Bench (Section 11.3.5).

After verifying the compliance with the performance measurement criteria indicated in Section 11.3.6, I conducted a Three-Step Interpretability method using image processing techniques. The initial step covered a total of 50,080 images, considering that each complete Test Set contains 626 sequences. The second step involved 160 images (80 experiments and two categories). Finally, the third step comprised a total of 16 images per category and number of epochs.

Based on the performance measurement ratios, I determined that in 80% of the subsamplings, the optimal configuration corresponded to 200 Epochs, with a Learning Rate of 0.0001 and a Batch Size of 52. For 200 Epochs, the AUC ranged from 0.9698 to 0.9824. Validation Accuracy ranged from 91.69% to 96.17%. Test Accuracy ranged from 91.69% to 96.01%. Regarding the balance of accuracy between categories, I included inter-category SD in my decision criteria because I prioritized the balanced performance of the CNN across both categories. The configuration with 200 epochs achieved the lowest inter-category SD, with a value of 2.12.

When analyzing the hot zones corresponding to the non-recombinant category, I observed a lack of a discernible hot zone, and the boundaries of the hot zone were somewhat indistinct. This could be attributed to a greater diversity of sublineages and strains among the non-recombinant variants, leading to greater variability due to subsampling. The hot zones were scattered, although I detected suggested hot spots around the Spike protein region, with critical frequencies falling approximately at $f = 1/12$ and $f = 1/6$, respectively.

The most spectacular result occurred in the recombinant category, where I identified a distinct hot zone with an epicenter at $f = 1/6$ in the S protein. The S protein (Spike) in SARS-CoV-2 is fundamental to the virus ability to infect human cells. This protein binds to the ACE2 receptor on the cell surface, facilitating the virus entry and subsequent replication. The S protein is also the main target of the

immune system. Additionally, it has been a determining factor in the viral evolution of SARS-CoV-2, concentrating multiple mutations and genetic recombination events.

This protein has generated numerous instances of double crossover genetic recombination events in different coronaviruses, with a notable concentration within this particular protein. Therefore, my findings complement and reinforce their conclusions through a different analytical approach.

The clarity of the hot zone associated with the recombinant category in SARS-CoV-2 in the S protein at $f = 1/6$ provided me with an opportunity to gain more information about the mathematical signature associated with the recombinant feature.

SCIENTIFIC CONTRIBUTION:

- Research Paper:

Ana Guerrero-Tamayo, Borja Sanz Urquijo, Concepción Casado, María-Dolores Moragues Tosantos, Isabel Olivares, Iker Pastor-López. (2024). Classification of SARS-CoV-2 sequences as recombinants via a pre-trained CNN and identification of a mathematical signature relative to recombinant feature at Spike, via interpretability. PLoS ONE 19(8): e0309391. https://-doi.org/10.1371/journal.pone.0309391 [3006].

15.3.10 Chapter 12. Analysis of SARSCoV-2 Non-recombinant Variants

In Chapter 11, I identified a clear hot zone that was crucial for classifying SARS-CoV-2 sequences as recombinant. This determinant hot zone for the recombinant feature was clearly located at $f = 1/6$ in the S protein. However, the hot zones corresponding to the non-recombinant feature turned out to be less clear. This could be related to a greater diversity in terms of variants.

In light of the results obtained in Chapter 11, I decided to deepen our understanding of non-recombinant variants. To this end, I adapted my research methodology to classify complete non-recombinant SARS-CoV-2 sequences within the four most numerous variants: pre-VOC (pre-Alpha), Alpha, Delta, and Omicron.

I generated the superposed genomic spectrograms for a total of 1,490,840 non-recombinant sequences (Section 12.3.1). Subsequently, I generated the dataset according to the specifications in Section 12.3.2. The majority presence of Omicron necessitated balancing the categories using two random subsamplings of Omicron.

Subsequently, I applied a Two-Stage Transfer Learning approach. I started with a VGG-16 model pre-trained using the ImageNET dataset. In Phase 1, I applied Transfer Learning to the genomic spectrogram dataset of complete HIV-1 sequences to detect the recombinant feature. In Phase 2, I applied Transfer Learning again to the network resulting from Phase 1 (VGG-16 HIV-1 GENETIC RECOMBINATION) using a genomic spectrogram dataset of complete SARS-CoV-2 sequences to detect the non-recombinant variant to which each SARS-CoV-2 sequence belongs (VGG-16 SARS-CoV-2 NON-RECOMBINANT VARIANTS).

Thus, I introduced a double change in this experiment. Not only did I switch the target virus of the study, but I also simultaneously changed the feature to be detected by the CNN. I moved from detecting genetic recombination in HIV-1 to classifying variants in SARS-CoV-2 (Section 12.3.3).

The performance measurement metrics were Validation Accuracy, Training Time, Test Accuracy, and Confusion Matrix.

The results obtained for the optimal hyperparameter configurations (Learning Rate of 0.0001, Batchsize of 52, and 1 Epoch) were in the following ranges (Table 15.5):

Table 15.5: Optimal Hyperparameter Configurations (Superposed Spectrograms)

Hyperparameter	**Optimal Values**
Validation Accuracy	99.42% - 99.52%
Global Test Accuracy	99.40% - 99.48%
Pre-VOC Test Accuracy	97.73% - 98.12%
Alpha Test Accuracy	99.79% - 99.86%
Delta Test Accuracy	99.88% - 99.73%
Omicron Test Accuracy	99.74% - 99.85%
Training Time	1,190 min 10 sec - 1,626 min 16 sec

Therefore, I obtained highly positive and very similar ratios between the two subsamplings. The large amount of data available for training likely influenced these results.

In both subsamplings, Omicron had the highest accuracy rate, followed by Delta. Alpha and pre-VOC had the lowest accuracy rates, although they were very high in any case. An influential factor in these accuracy rates could be the quantity and homogeneity of samples in each variant. Omicron is the most numerous variant, while Alpha is precisely the variant with the fewest complete genomes sequenced. Additionally, pre-VOC represents the collection of lineages before Alpha globally, so greater variability compared to greater homology in Omicron cannot be ruled out.

In any case, I observed very positive accuracy ratios in all categories, as well as very similar behavior in both subsamples. These factors indicate that the CNN is adequately detecting differentiating patterns for each of the four categories.

Regarding the total hot zones per category, the inherent patterns in each category proved to be absolutely distinguishable, robust, and with high (in the case of pre-VOC) and very high homology in both subsamples (in the case of Alpha, Delta, and Omicron).

In the case of pre-VOC, the differences between both subsamples related to a power loss in VARIANT SUBSAMPLING 02 compared to VARIANT SUBSAMPLING 01. However, the physical location of the total hot zones turned out to be coincident.

The Omicron variant showed virtually unchanged total hot zones in both subsamples. I consider it extremely interesting to delve into the relationships between the phylogenetic reality and the evolutionary diversity associated with the compilation of pre-VOC, Alpha, Delta, and Omicron, and the mathematical signatures that determine the CNN's classification into these variants.

From a computational standpoint, I demonstrated a successful application of Two-Stage Transfer Learning with two simultaneous changes: the type of virus and the object of study. CNN learning on genomic spectrograms is promising not

only for detecting the recombinant feature but also for other features, in this case, regarding the variant to which a sequence belongs.

Therefore, the positive results of this dual application of Transfer Learning implicitly demonstrate the versatility in using genomic spectrograms in various virological analyses.

From a biological perspective, the robust and specific total hot zones in each type of variant denote the importance of the genomic region from the start of S to the end of 3'UTR in identifying the analyzed variants. Frequencies $f = 1/9$ and especially $f = 1/3$ in this genomic region seemed to play a relevant role in such identification. Further investigation will be necessary to explore the exact relationship between nucleotide periodicities relative to these frequencies and membership in one variant or another.

Just as in the case of characterizing a sequence as recombinant, in this instance, I demonstrated the existence of a mathematical signature associated with the pre-VOC, Alpha, Delta, and Omicron variants. Therefore, the existence of other mathematical signatures associated with other viral features cannot be ruled out.

15.3.11 Chapter 13. Exploring the Influence of the Third Codon Nucleotide on Genetic Recombination in the SARS-CoV-2 Spike Protein

In this Chapter, I revisit the results related to the total hot zones concerning the recombinant feature in SARS-CoV-2 obtained in Chapter 11. In the recombinant zones, a clear total hot zone was detected with a distinct epicenter at $f = 1/6$ in the S protein.

It is in this protein where numerous events of genetic recombination have occurred in SARS-CoV-2, and it is a fundamental protein in the virus cell infection process.

In Chapter 5, I detected the presence of the most relevant mathematical signature determining that a complete sequence of HIV-1 was recombinant at $f = 1/3$ in the 5'LTR and possibly part of *gag* (Figures 5.9 and 8.1b). The location of this signature did not coincide with the result of that genetic recombination.

However, in this case, the mathematical signature coincided with a genomic region where these phenomena have commonly occurred. And, as I saw in Chapter 10, it was a region of low conservation, highly subjected to variability, especially in terms of nucleotide matching. The degree of nucleotide matching was only 0.23%, making it the most variable protein in terms of nucleotide conservation.

The clear detection by the CNN of the potential position of the characteristic mathematical signature of the recombinant feature in the S protein, frequency $f = 1/6$, made it necessary to deepen our understanding of this signature, especially given its apparent location in such a relevant and active region. We know where the CNN is looking, and I want to understand more about what it sees. I have a case study that is an excellent opportunity for this. I accepted the challenge and proceeded to explore what this mathematical signature implies.

As detailed in Section 13.3, the first step to obtain more information about this relevant mathematical signature in the recombinant feature of SARS-CoV-2 was the selection of sequence groups for study in the recombinant and non-recombinant categories (Section 13.3.2).

Based on the number of sequences resulting from this selection, I chose the same number of non-recombinant sequences correctly classified by the CNN whose hot zones did not include $f = 1/6$ in S. That is, this zone was expressly excluded from the hot zones in each non-recombinant sequence. This selection criterion resulted in 39 sequences out of a total of 313 in the Test Set. My objective with this selection criterion was to work with sequences whose patterns in the superposed spectrogram were as different as possible for the CNN, thus approaching a search for differential elements in an environment where structural differences were apparently more evident. In this way, identifying those differences between a study group and another beforehand could be simpler. Therefore, 39 sequences meeting this requirement were randomly selected.

Once these groups were established, I proceeded to extract the gene sequences of the study proteins based on the annotations in their EMBL files (Sections 13.3.3 and 13.3.4).

For the recombinant sequences, I extracted the S proteins and, in order to compare any significant findings I might encounter in this structural protein with other structural proteins, I selected M and N. For the non-recombinant sequences, I extracted the S protein.

After extracting these gene sequences, I proceeded to analyze inherent patterns in their sequences. The prominent hot zone in the S gene, which is highly relevant in terms of viral evolution, led me to analyze this area in more detail to extract more information about the inherent mathematical signature of the recombinant feature. From a superficial visual analysis of the arithmetic series of nucleotides every 6 positions, I identified a specific pattern in the arithmetic series whose initial position was a multiple of 3, which did not occur in the rest of the mathematical series. Consequently, I inferred the study of the third nucleotides of the triplets in S in complete sequences of SARS-CoV-2 as an area of interest. I selected all recombinant sequences from this set whose hot zone was exclusively $f = 1/6$ in the S protein. Out of the 313 sequences, 39 met this criterion. Subsequently, I selected an equal number of non-recombinant sequences from the Test Set whose hot zones did not include $f = 1/6$ either primarily or secondarily.

Once the sequences under study were selected, I extracted the third nucleotides of the S, M, and N proteins in the recombinant sequences to compare the patterns that might emerge in S with other structural proteins exhibiting sufficient variability.

I also compared the patterns in the third nucleotides of S in the recombinant sequences with their non-recombinant counterparts to verify potential differences between sequences in one category and the other.

I detected a high content of Thymine compared to the rest of the nucleotide types in the third nucleotides in S, exceeding 45%. I did not observe this phenomenon in M or N in the selected 39 recombinant sequences. The third nucleotides of the triplet in S in the non-recombinant sequences showed similar amounts of Thymine, as did the recombinant ones.

Finally, I determined the common nucleotides (content and position within S) in the 39 selected recombinant sequences and the 39 selected non-recombinant

sequences. Understanding these common elements between one group and the other, as well as their differences, could provide information about the inherent mathematical signature that determines the recombinant feature.

Surprisingly, I was able to confirm that, in both the recombinant and non-recombinant sequences, all their common elements were multiples of 3. That is, their common elements were third nucleotides of the triplet. Considering the high variability of S in terms of length and nucleotide content (Chapter 13), this fact is especially relevant and motivating.

From a statistical point of view, attributing this coincidence to mere chance is complex. I selected 39 sequences out of a total of 1,565 recombinant sequences and 39 non-recombinant sequences out of a total of 15,650. In other words, 78 selected sequences out of a total of 17,215. And if we consider the total number of sequences before the subsampling performed in Chapter 11, there is a total of 1,541,293 recombinant and non-recombinant sequences combined. The S protein showed high variability in terms of nucleotide content (Chapter 13), with only 0.23% of common nucleotides.

In the analysis presented in this Chapter, all of these nucleotides were multiples of 3, meaning they were the third nucleotide of the triplet. Additionally, due to the degeneracy of the genetic code (where multiple codons can encode for the same amino acid because differences in the third position of the codon often do not alter the amino acid sequence in the final protein) and perhaps due to limitations of the human ribosome to correctly interpret all codons, it may be more prone to mutation without necessarily resulting in a modification in the resulting protein. However, under certain circumstances, this can lead to such modifications, and if viable and advantageous from an evolutionary standpoint, it could give rise to new subvariants, variants, or even lineages.

This search for common nucleotides was conducted without applying any kind of multiple alignment. In other words, in sequences of variable lengths in S, I found that the common elements were precisely the third nucleotides of the triplet without performing any "artificial" adjustments prior to comparison.

Finally, after identifying the common elements in both categories, I determined which elements are common to both recombinant and non-recombinant sequences to try to shed some light on that mathematical signature encrypted in the genome.

All the evidence found suggests that this encrypted mathematical signature is located in the third nucleotide of the triplets in S (in the case of SARS-CoV-2).

Therefore, I detected enough evidence to reinforce the hypothesis that this third nucleotide, less relevant in terms of amino acid coding, might contain additional information in an encrypted manner that characterizes a SARS-CoV-2 sequence as recombinant.

SCIENTIFIC CONTRIBUTION:

- Conference Paper:

Ana Guerrero-Tamayo, Borja Sanz Urquijo, Concepción Casado, María-Dolores Moragues Tosantos, Isabel Olivares, Iker Pastor-López. (2024). The Third Codon Nucleotide's Role in Genetic Recombination Within SARS-CoV-2 Spike Protein: A Pilot Study. International Conference on Hybrid Artificial Intelligence Systems. Springer Nature Switzerland.

15.3.12 Chapter 14. Is There a Role Associated with the Third Codon Nucleotide in the Characterization of a Genomic Sequence? The Case of HIV-1 and SARS-CoV-2

In Chapter 4 (Genomic Spectrograms), I verified, in accordance with existing literature, the horizontal line at $f = 1/3$ in coding regions and its relationship with the fact that in these coding regions, nucleotides are organized in triplets, with each triplet encoding an amino acid.

In Chapter 5 (Analysis of Genetic Recombination in HIV-1), I explained the detection of a mathematical signature indicative of the recombinant feature in HIV-1. This signature was mainly located at $f = 1/3$ in the 5' LTR and possibly around the first half of *gag*.

In Chapter 6 (Analysis of Total Hot Zones in HIV-1 by Subtype), I delved into the hot zones associated with significant recombinant and non-recombinant subtypes

of HIV-1 and their relationship with their category counterparts, as well as the peculiarities of these subtypes highlighted by the scientific community. Periodicities of 3 and their multiples were recurrent throughout this analysis.

In Chapter 8 (Analysis of HIV-1 Subtype U), I presented a detailed analysis of sequences classified as U with respect to expert literature, examining the recombinant nature of each sequence in detail and possibly detecting hidden recombination phenomena in some sequences of this subtype. Once again, periodicities related to 3 were involved in multiple case studies.

In Chapter 9 (Analysis of HIV-1 Subtype G), I developed the analysis of the possible recombinant origin affecting subtype G based on the analysis of hot zones of the sequences belonging to this subtype. The relationship between the periodicity 3 and the recombinant feature in HIV-1 allowed for a deeper exploration of additional information about this recombinant origin.

In Chapter 11 (Analysis of Genetic Recombination in SARS-CoV-2), I explained the detection of a distinct mathematical signature at $f = 1/6$ that determines if a sequence of SARS-CoV-2 is recombinant. By delving into the detection of patterns related to this signature, I was able to find remarkable evidence that this signature is related to the third nucleotides of the triplet in that protein. All these findings are detailed in Chapter 13 (Exploring the Influence of the Third Nucleotide of the Codon on Genetic Recombination in the Spike Protein of SARS-CoV-2).

I delve into the identification of mathematical signatures associated with genetic recombination in two viruses as different as HIV-1 and SARS-CoV-2.

In both cases, I detected mathematical signatures associated with the recombinant feature. But not only that, in Chapter 12 (Analysis of Non-Recombinant Variants of SARS-CoV-2), I adapted the methodology to detect the recombinant feature and classify non-recombinant sequences into the pre-VOC, Alpha, Delta, and Omicron categories. This yielded remarkably positive results with specific mathematical signatures associated with each of these variants, with a clear presence of $f = 1/3$ among their associated hot zones.

In all cases, I found a relationship with the frequency $f = 1/3$, either directly or indirectly ($f = 1/6$ and $f = 1/9$).

Since frequency is the inverse of the period, understood in this case as the periodicity of nucleotides, the equivalent period of $f = 1/3$ is 3. In the coding regions of the genome, where three nucleotides encode an amino acid, this periodicity of 3 equates to the third nucleotides of the codon in coding regions. Similarly, for $f = 1/6$, the nucleotides related to this frequency will be the third nucleotides of the codon every 2 amino acids. In the case of $f = 1/9$, the involved nucleotides will be the third nucleotides of the codon every 3 amino acids.

Therefore, in the case of HIV-1, given that the primary mathematical signature associated with the recombinant feature was located at $f = 1/3$ in 5'LTR and perhaps the first half of *gag*, the underlying mathematical pattern would be related to every third nucleotide in 5'LTR (non-coding region) and possibly the third codon nucleotides of the first half of *gag*. In the case of SARS-CoV-2, the mathematical signature associated with the recombinant feature was clearly located at $f = 1/6$ in S, that is, the third codon nucleotides every 2 triplets. Subsequently, I detected that the relationship probably referred to the third codon nucleotides.

Now, I would like to pose the big question that will be a future research line as exciting as, or even more than, the research already conducted: What is the level of universality of the role of third nucleotide positions in non-coding regions or the third nucleotide of the codon in coding regions, in characterizing a genomic sequence?

The purpose of this Chapter was an analysis, more reflective than quantitative, that could serve as a guide or inspiration for future research focused on determining the degree of universality of the role of the third codon nucleotide in the genetic recombination of viruses.

In Section 14.2, I compiled all the evidence relating mathematical signatures of characteristics as different as genetic recombination and inclusion in a particular variant, as well as various detailed analyses by subtypes, degrees of homology between hot zones and phylogenetic proximity, detection of possible hidden genetic

recombination phenomena, complementary information on the possible recombinant origin of subtype G, etc.

The periodicities of 3 and their multiples consistently appeared in different features, especially in viruses as diverse as HIV-1 and SARS-CoV-2.

I correlated the evidence presented in this Chapter with scientific findings in the field of microbiology and virology (Section 14.2).

The similarities found are even more remarkable considering that the common ancestor between the *Retroviridae* family (to which HIV-1 belongs) and the *Coronaviridae* family (to which SARS-CoV-2 belongs) is of ancestral RNA type. Their phylogenetic relationship is extremely distant, indicating that their evolutionary paths likely diverged hundreds of millions of years ago. It is estimated that the origin of retroviruses occurred during the early Paleozoic Era, between 460 and 550 million years ago. This occurred long before the appearance of mammals.

All coronaviruses originated from a common ancestor approximately 293 million years ago, during the Late Permian period. Although modern mammals did not exist at that time, it is believed that there were ancestors of primitive mammals. These proto-mammals belonged to groups such as therapsids, which eventually gave rise to modern mammals during later periods, such as the Triassic. Therapsids were synapsid reptiles that exhibited characteristics of both reptiles and mammals and played a significant role in mammalian evolution.

Despite their phylogenetic distance, in both viruses, the mathematical signature that characterizes a sequence as recombinant maintains a close and clear relationship with the third nucleotide of the codon in specific genomic regions.

It is remarkable that two viruses belonging to families that diverged long before the appearance of modern mammals exhibited similar signs in their patterns related to the mathematical signature determining the recombinant feature. These common signs were associated with the role of third codon nucleotides.

On one hand, I must highlight that nucleotides A and T are significantly associated with the occurrence of genetic recombination events in HIV-1.

In the case of SARS-CoV-2, in the pilot study conducted in Chapter 13, I detected a high content of Thymine in the third codon nucleotides in the Spike gene, around 45%. This high amount of Thymine compared to other nucleotide types was not detected in other structural proteins like M and N.

This fact could be related to the reduction in the frequency of Cytosines and a corresponding increase in the frequency of Uracils in the third codon position detected in SARS-CoV-2. This phenomenon, known as codon usage bias, suggests a shift in codon preference throughout the virus's evolution, possibly driven by specific mutations and selective pressures in the host.

Mutations from Cytosine to Uracil ($C > U$) have been identified as the most prevalent in the SARS-CoV-2 genome. These mutations could result from the activity of host deaminases, enzymes that catalyze the conversion of Cytosine to Uracil in RNA. This observation underscores the importance of the interaction between the virus and the host in determining the genetic diversity and evolutionary dynamics of SARS-CoV-2.

In terms of molecular genetics, it is known that the third position of the codon is relatively flexible in terms of amino acid specificity. The "wobble" hypothesis, proposed by Francis Crick in 1966, postulates that the third base of the codon can pair non-conventionally with the corresponding base in transfer RNA (tRNA), allowing some tolerance in codon specificity. Therefore, changes in the third position of the codon, such as the $C > U$ transition observed in the SARS-CoV-2 genome, do not necessarily have a direct impact on the amino acid sequence of the encoded protein, although they may influence translation efficiency and messenger RNA stability.

In conclusion, based on the analysis of hot zones resulting from recombinant and non-recombinant categories, I determined the significant role of A and, to a lesser extent, T in the recombinant feature of HIV-1. Additionally, I identified the role of T in the recombinant feature of SARS-CoV-2. In both viruses, there was a relationship with these nucleotide types, with T being the relevant nucleotide type in classifying a sequence of both viruses as recombinant. This was the case despite

these viruses being very different. Furthermore, in both cases, the interpretation was consistent with the biocentric approach according to the results obtained.

Regarding the relevant genomic regions in the recombinant feature of HIV-1, Section 14.3.3 details the functions of the 5' and 3'LTR genomic regions. The 5'LTR of HIV-1 is a multifunctional genomic region that coordinates viral RNA transcription, viral DNA integration into the host genome, and HIV-1 gene expression. The 3'LTR of HIV-1 is a critical region of the viral genome that acts as a promoter for viral transcription, facilitates integration of viral DNA into the host genome, and provides necessary signals for processing and termination of viral transcription. Their importance in the HIV-1 life cycle makes them key targets for research and development of antiviral therapies directed against HIV-1.

Throughout all the experimentation presented, I identified several influences of nucleotides in the third position on the mathematical signatures inherent in the patterns detected by the CNNs. Spectrographic frequencies $f = 1/3$, $f = 1/6$, and $f = 1/9$ were recurrently observed in the studied heatmaps of interpretability, indicating the presence of periodicities with a relationship to 3 and its multiples. Therefore, it appears that nucleotides in the third codon position contain encrypted information about viral characteristics (recombinant feature and variant) in viruses as distant as HIV-1 and SARS-CoV-2.

However, numerous clues established a correlation between phenomena related to arithmetic series with a relationship to 3 or its multiples and the mathematical signature associated with the recombinant feature in both viruses. In the case of HIV-1, the importance of the periodicity 3 in the 5'LTR region has been observed, with potent and stable associated hot zones, as detailed in Chapters 5 and 6. It would be necessary to establish the possible influence of *gag* and its third codon nucleotides on that recombinant feature. Similarly, in the case of SARS-CoV-2, I observed similar patterns in the nucleotides of the third codon of the Spike protein, specific to this protein and suggesting a determinant role in the recombinant feature, as mainly developed in Chapters 11 and 13.

These similarities between such different viruses lead me to consider the possible degree of universality in mathematical signatures. This makes me think that there may be a certain degree of universality because there are similar patterns in two very different viruses.

Furthermore, the biocentric interpretation aligns with the results obtained regarding the role of nucleotides A and T in HIV-1 and the role of LTRs and maybe *gag* in relation to their potential association with the location of mathematical signatures associated with the viral feature. In the case of SARS-CoV-2, the role of the Spike protein and the phenomenon of $C > U$ hypermutations are consistent with the results obtained regarding the location of the mathematical signature relative to the recombinant feature precisely in that protein and with the high quantities of Thymine detected in the Spike protein gene.

Periodicities related to 3 and its multiples were also clearly involved in the mathematical signatures associated with the classification of non-recombinant sequences into pre-VOC, Alpha, Delta, and Omicron variants, although the correlation between this phenomenon and the involved genomic regions requires further investigation.

In this Chapter, I compiled the main evidence linking nucleotides in the third codon position with the activated zones where the CNN focused its attention to make decisions.

Based on this compilation of evidence, I strongly propose the hypothesis that nucleotides in the third codon position contain encrypted information that influences the classification of certain viral characteristics. Furthermore, I suggest the possibility that this role of nucleotides in the third codon position (coding regions) and every third nucleotide (non-coding regions), may be common across different microorganisms, implying a degree of universality that warrants further evaluation.

Validating this hypothesis would represent a significant shift in the interpretation of the genetic code as we currently understand it. It would also confirm that genetic code degeneracy is more complex than a simple conversion of nucleotides to amino acids, and that the identity of the nucleotide in the third codon position may be relevant for viral characterization.

If the genome indeed contains encrypted information in different layers, it would open a broad field of research with significant implications for our understanding of molecular biology and viral evolution.

At this point, everything presented in this Chapter requires further investigation, and my conclusion is nothing more than a mere hypothesis.

Hopefully, over time, we will gather more data and be able to evolve our understanding of this wonderful universe that is the genome.

This is the beginning of an exciting journey.

15.4 Fulfillment of Research Objectives

All the research objectives were met. This Section includes the main evidence of the fulfillment of the research objectives.

Achieving these objectives led to the validation of the research hypothesis proposed:

The application of advanced Deep Learning techniques and mathematical analysis to genomic spectrograms of viruses will enable not only the effective characterization of viral genomes, but also the identification of encrypted mathematical signatures. These signatures can facilitate the understanding of viral genotypic and phylogenetic relationships, providing a solid foundation for the design of generative architectures aimed at predicting genomic evolution in viruses.

RESEARCH OBJECTIVE:

I. Evaluate genomic spectrograms as a method for representing genomic sequences.

MAIN EVIDENCES:

- Chapters 3 and 4. Identification of multiple research articles describing the effective application of spectrograms in various fields unrelated to the biosanitary field.

- Chapters 3 and 4. Identification of multiple research articles describing the effective application of spectrograms in the biosanitary field.

- Chapters 3 and 4. Identification of multiple research articles describing the effective application of spectrograms in the genomics field.

RESEARCH OBJECTIVE:

II. Develop and evaluate Convolutional Neural Network (CNN) models for the categorization of genomic sequences.

MAIN EVIDENCES:

- Chapter 5. Implementation and evaluation of three pre-trained CNNs (VGG-16, ResNet-101, and Inception V3), applying Transfer Learning, for the classification of genomic spectrograms of complete HIV-1 sequences as recombinant or non-recombinant. Obtaining highly positive values in terms of Validation and Test Accuracy, AUC, Training Time, and Confusion Matrix.

- Chapter 11. Implementation and evaluation of pre-trained VGG-16 for the classification of genomic spectrograms of complete SARS-CoV-2 sequences as recombinant or non-recombinant. I used a Two-Stage Transfer Learning Methodology. My starting point was a pre-trained VGG-16 using the ImageNet dataset. In Phase 1, I applied this pre-trained model using Transfer Learning to the genomic spectrogram dataset of complete HIV-1 sequences to detect the recombinant feature. In Phase 2, I applied Transfer Learning once again to the resulting network from Step 1 (VGG-16 HIV-1) using a genomic spectrogram dataset of complete SARS-CoV-2 sequences to also detect the recombinant feature (VGG-16 SARS-CoV-2). I obtained highly positive values in terms of Validation and Test Accuracy, AUC, and Confusion Matrix.

- Chapter 12. Implementation and evaluation of pre-trained VGG-16 for the classification of genomic spectrograms of complete non-recombinant SARS-CoV-2 sequences into pre-VOC, Alpha, Delta, and Omicron variants. I started with a pre-trained VGG-16 using the ImageNet dataset. In Phase 1, I applied Transfer Learning to the genomic spectrogram dataset of complete HIV-1 sequences to detect the recombinant feature. In Phase 2, I applied Transfer Learning once again to the resulting network from Step 1 (VGG-16 HIV-1 GENETIC RECOMBINATION)

using a genomic spectrogram dataset of complete SARS-CoV-2 sequences to also detect the non-recombinant variant to which each SARS-CoV-2 sequence belongs (VGG-16 SARS-CoV-2 NON-RECOMBINANT VARIANTS). I obtained highly positive values in terms of Validation and Test Accuracy, Training Time, AUC, and Confusion Matrix.

RESEARCH OBJECTIVE:

III. Develop and evaluate explainable models to detect hidden mathematical signatures in the genome associated with the genotypic characterization of a sequence.

MAIN EVIDENCES:

- Chapter 5. Application of Grad-CAM, LIME, and Gradient Attribution explainability tools to determine the relevant zones in the CNN classification of genomic spectrograms of complete HIV-1 sequences as recombinant or non-recombinant. I obtained differentiated hot zones for each category in the 3 explainability tools. When the CNN makes classification errors, these tools focused on the hot zones of the incorrect category. The mathematical signatures associated with the recombinant feature were linked to 5' and 3' LTR zones and their U3, R, and U5 regions, as well as to nucleotides A and, to a lesser extent, T.

- Chapter 11. Application of the Grad-CAM explainability tool to determine the relevant zones in the CNN classification of genomic spectrograms of complete SARS-CoV-2 sequences as recombinant or non-recombinant. I obtained differentiated hot zones for each category. A clear mathematical signature associated with the recombinant feature in the S protein was identified.

- Chapter 12. Application of the Grad-CAM explainability tool to determine the relevant zones in the classification by the CNN of genomic spectrograms of complete SARS-CoV-2 sequences non-recombinant within the pre-VOC, Alpha, Delta, and Omicron variants. I obtained differentiated hot zones for each variant.

RESEARCH OBJECTIVES:

IV. Develop and evaluate a methodology for the identification of mechanisms associated with genetic recombination from a computational and bio-mathematical perspective.

VI. Validate the developed methodology (genomic spectrograms + CNN + explainability) in applied cases for HIV-1 and SARS-CoV-2 viruses.

MAIN EVIDENCES CASE STUDY GENETIC RECOMBINATION IN HIV-1:

- Chapter 5. Highlighted the role of the frequencies $f = 1/3$, $f = 1/6$, and $f = 1/9$ in the total hot zones by category. The mathematical signatures associated with the recombinant feature may be related to the functions of the LTRs during viral transcription, gene expression regulation, and virus replication. Additionally, a potential role was detected in the adjacent genes *gag*, *nef*, and *env*. The significance of A and T nucleotides in the recombinant feature could be linked to the high frequency of polyA and polyT sequences in the HIV-1 genome. These regions are particularly susceptible to polymerase skipping, promoting genetic recombination. The substantial role of the A nucleotide is likely associated with the high adenine content found in lentiviruses, especially HIV-1. Moreover, I detected multiple instances of homology between the degree of similarity of the hot zones and phylogenetic proximity, as well as potential hidden recombination events and controversial sequences.

- Chapter 6. Highlighted the role of the frequencies $f = 1/3$, $f = 1/6$, and $f = 1/9$ in the total hot zones by subtype. Analyzing all representative HIV-1 subtypes, I determined the homology between their individual sequences, the total hot zones by subtype, and those corresponding to their category. Through the comparison of the total hot zones by subtype and by category, I detected certain non-recombinant types whose total hot zones presented peculiarities. Except for F1, the rest of these subtypes (U, G, and O) indeed have special considerations in the scientific community expert in HIV-1. I detected a high analogy in the total hot zones of the CRF20_BG and CRF42_BF1 subtypes, a possible indicative of unknown phylogenetic relationships among them.

- Chapter 8. Analyzing the hot zones of subtype U (unknown) sequences and their phylogeny, I detected potential hidden genetic recombination events in intra-patient subtype U sequences EF029066, EF029067, EF029068, and EF029069 (Patient Code: H10986). I identified both phylogenetic similarities and degrees of similarity among the hot zones of various U, G sequences, and recombinant G subtypes, which may denote unknown parental relationships between G and U sequences. The interpretability patterns were consistent with phylogeny and multiple biological evidences in 60.870% of the cases, in 34.783% my analysis was somewhat inconclusive, and only Sequence AY046058 was a controversial case (4.348%).

- Chapter 9. I detected a certain tendency for subtype G sequences classified as recombinant and those classified as non-recombinant to be grouped into distinguishable phylogenetic zones. I observed a repeated phenomenon in this subtype where phylogenetically related sequences are classified into disparate categories. In multiple sequences, I detected hot zones scarcely coherent with those corresponding to their category. The subtype G sequence MN153484 was phylogenetically isolated from the rest of the G sequences and phylogenetically related to subtype U MN736702, which may indicate hidden parental relationships between subtype G and U sequences.

MAIN EVIDENCES CASE STUDY GENETIC RECOMBINATION IN SARS-CoV-2:

- Chapter 11. In the case of SARS-CoV-2, the Spike protein emerges as a critical genomic region in both recombinant and non-recombinant categories. This protein has been the focus of attention, revealing numerous instances of double-crossover genetic recombination events across different coronaviruses. The non-recombinant category stood out for its more diffuse hot zones, yet they seem to center around the region of the spike protein, particularly within the frequency range of $f = 1/12$ and $f = 1/6$. In the case of recombinants, the hot zone was sharp and clearly located in the S protein, centered at $f = 1/6$.

RESEARCH OBJECTIVE:

VI. Validate the developed methodology (genomic spectrograms + CNN + explainability) in applied cases for HIV-1 and SARS-CoV-2 viruses.

MAIN EVIDENCES IN THE CASE OF CLASSIFICATION OF NON-RECOMBINANT SARS-CoV-2 SEQUENCES INTO VARIANTS:

- Chapter 12. In addition to the excellent classification results, the total hot zones per variant were clear and clearly distinguishable for each category.

RESEARCH OBJECTIVES:

V. Provide descriptive information on encrypted mathematical signatures in the genome associated with the genotypic characterization of sequences.

VI. Validate the developed methodology (genomic spectrograms + CNN + explainability) in applied cases for HIV-1 and SARS-CoV-2 viruses.

MAIN EVIDENCES:

- Chapters 5, 6, 8, 9, 11, 12. The mathematical signatures associated with the recombinant or non-recombinant feature in HIV-1 and SARS-CoV-2, as well as the classification of a non-recombinant SARS-CoV-2 sequence into one of the four major variants, are all related to the frequency spectrum.

- Chapter 12. In broad terms, the mathematical signatures associated with each variant showed an importance of the genomic region from the beginning of S to the end of 3'UTR. The frequencies $f = 1/9$ and especially $f = 1/3$ in this genomic region appeared to play a relevant role in such identification.

- Chapters 12, 13. The most recurrent frequencies are $f = 1/3$, $f = 1/6$, and $f = 1/9$. The frequency $f = 1/12$ was also detected in certain cases analyzed.

- Chapters 10, 13. Analyzing SARS-CoV-2, I detected a high thymine content in the third codon nucleotides of the S protein compared to the rest of the nucleotide types in the third nucleotides of S, exceeding 45%. The third nucleotides of the triplet in S in the non-recombinant sequences showed similar amounts of Thymine, as in the case of the recombinants. In the selected study groups (39 recombinant and 39

non-recombinant sequences selected based on their hot zones), all common elements were multiples of 3, despite the high variability of S in terms of length and nucleotide content. That is, they were all third codon nucleotides, even though the expected variability in these nucleotides is higher due to the degeneracy of the genetic code.

- Chapter 14. The periodicities of 3 and their multiples played a striking role in all experiments conducted, in two viruses as phylogenetically distant as HIV-1 and SARS-CoV-2.

15.5 Main Conclusions

- The developed methodology (genomic spectrograms + CNN + explainability) demonstrated the ability to identify mathematical signatures for the classification of viral genomes as recombinant or non-recombinant, by variants, and even to highlight certain parameters for future research, such as the abundance of T/A in some of the identified signatures, as well as the role of the third nucleotide of the codon.

- The mathematical signature associated with genetic recombination in HIV-1 was located in the 5' and 3' LTR zones and their U3, R, and U5 regions, with high importance at $f = 1/3$. Indeed, these non-coding regions are crucial in viral transcription, in regulating gene expression, and in the virus replication process. The adjacent *gag*, *nef*, and *env* genes could also play a role in HIV-1 genetic recombination. The nucleotides A and T were found to be relevant in the recombinant feature. This may be associated with the elevated presence of polyA and polyT stretches in the HIV-1 sequence. These specific regions are prone to polymerase skipping, thereby facilitating genetic recombination. The prominent role of the A nucleotide might be linked to the abundant adenine content within lentiviruses, particularly in the case of HIV-1.

- The homology between the hot zones and the degree of phylogenetic proximity suggests that similarities in the regions of interest of the sequences are linked to their evolutionary relationship.

- By comparing the total hot zones and the phylogeny, I not only detected the special nature of subtype U but also detected a possible confirmation of the recombinant origin of subtype G. Likewise, I identified possible hidden genetic recombination events in certain genomic sequences of U. I also established relationships between G and U sequences that require further research to identify homologous genomic regions that provide new evolutionary perspectives among the various subtypes involved in this case study, as well as additional information about the origin of subtype G.

- I identified a clear mathematical signature associated with genetic recombination in SARS-CoV-2, located distinctly in the Spike protein, with an epicenter at $f = 1/6$. This protein is critical in the evolution of SARS-CoV-2, where multiple mutations and double-crossover genetic recombination events have occurred.

- Delving deeper into obtaining more information regarding this mathematical signature, I discovered a pattern in the third codon nucleotides in the S protein of SARS-CoV-2, with a higher amount of T than found in other structural proteins (over 45%). I extracted common nucleotides in S from two groups of recombinant and non-recombinant sequences under study. All common elements were multiples of 3, that is, third codon nucleotides. Considering that I calculated the nucleotide conservation degree in S to be 0.23%, this phenomenon is highly remarkable.

- I identified distinguishable mathematical signatures associated with the classification of a non-recombinant SARS-CoV-2 sequence into the pre-VOC, Alpha, Delta, and Omicron variants. The importance of the genomic region from the beginning of S to the end of 3'UTR in the identification of the analyzed variants. The frequencies $f = 1/9$ and especially $f = 1/3$ in this genomic region appeared to play a relevant role in such identification.

- I was able to observe the presence of periodicity 3 and its multiples throughout all the experiments conducted. Across phylogenetically distant viruses and also across different features. The final conclusion of this writing poses a wonderful research question: what is the role of the number 3 in the mathematical characterization of these encrypted signatures in the genome?

15.6 Implications

Analyzing the relationship with the third codon nucleotide in the case of SARS-CoV-2 (genetic recombination and variant) and every third nucleotide in the LTRs of HIV-1, as well as the coincidence in the relevance of A and T in both viruses, all these clues lead me to firmly propose the hypothesis that the genome contains encrypted information in the nucleotides in the third codon position that influences the classification of certain viral features.

The confirmation of this hypothesis would represent a new paradigm in genome interpretation, adding at least a new layer of reading overlaid on the currently demonstrated one.

La ciencia siempre vale la pena porque sus descubrimientos, tarde o temprano, siempre se aplican.

— Severo Ochoa, *tras el fallecimiento de su mujer en 1986, nunca volvió a publicar ningún trabajo científico.*

16 Future Research Lines

Contents

16.1 Introduction 788
16.2 Synthetic Sequence Generation 789
16.2.1 Preliminary Considerations 789
16.2.2 Potential Architecture 794
16.2.3 Technological Challenges 802
16.2.4 Some advantages 803
16.3 Genomic Size Delimitation for the Effective Application of Genomic Spectrograms 804
16.4 Deeper Understanding of the Detected Mathematical Signatures 806
16.5 Identification of Mathematical Signatures Associated with Antifungal Resistance 807
16.6 Identification of Mathematical Signatures Associated with Antibiotic Resistance 811
16.6.1 *Mycobacterium tuberculosis* 813
16.6.2 *Clostridium difficile* 815
16.6.3 Vancomycin-Resistant Enterococci 818
16.6.4 Methicillin-Resistant *Staphylococcus aureus* 821
16.6.5 *Neisseria gonorrhoeae* 823
16.6.6 Carbapenem-Resistant Enterobacteriaceae (*Escherichia coli*, *Klebsiella pneumoniae*,) 826
16.7 Bio-Mathematical Correlation between the HIV-1 Genome and the Host Genome 828
16.8 Deepening the Mathematical Signatures Characterizing the Phylogeny among Different Pathogens 830

16.1 Introduction

In this research, virtually every outcome led to a series of hypotheses and research questions that will require future research. Each positive result brings an urgent need to delve deeper and deeper until achieving a complete and unequivocal understanding of each of the leads that emerged in this investigation. I conducted my research in a virtually unknown field. Naturally, from each small experiment, exciting questions will arise that I want to try to answer. To meticulously detail all aspects that required future research in my investigation would entail an unmanageable extension of the document. In summary, everything presented requires further research.s

Some of the research lines are a direct continuation of those last explorations that I outlined in this document. I am referring mainly to the role of the third codon nucleotide in coronaviruses and its relationship with the mathematical signatures of various viral features. I am also referring to the third nucleotide position in certain non-coding regions of HIV-1 and the possible influence of third codon nucleotides of *gag* in recombinant feature. And I am also referring to whether that three-fold periodicity occurs in other organisms and is also related to mathematical signatures associated with traits, even in humans.

But in addition, and regardless of the course of the research (although related to it, no doubt), several research questions, hypotheses, emerged that would require individual investigations separate from the one described here. These new lines are not directly related to my research in the sense that they are not a mere continuation of it. The main continuations are described in various sections of conclusions and future research throughout the document.

In this Chapter, I wanted to outline those parallel lines with their own entity that I consider of high interest.

There are still many challenges for the effective application of Deep Learning tools in the field of genomics. We must develop architectures that integrate exceptionality into the generalization capacity of these models. Nature always surprises us.

Frequently, the exception, the option against all odds, all logic, and all statistics, emerges and prevails in nature. Nowadays, it is difficult for a neural network to even consider or contemplate such options.

But it does not matter, because humanity has been achieving what seemed impossible for thousands of years. So I have no doubt that we will achieve it, once again.

I would like to dedicate the rest of my professional life to further delve into this fascinating world of genomics. It is a book that we have in front of our eyes and we still do not fully know the language that will allow us to read it. Undoubtedly, we will succeed.

16.2 Synthetic Sequence Generation

16.2.1 Preliminary Considerations

Starting from the results regarding the variability of each genomic fragment of SARS-CoV-2 (Chapter 10, Section 10.5.2), I proposed two designs for the generation methodology:

i) Synthetically generating fragment by fragment according to the historical homology level. The proposal involves using generative text models to synthesize fragments according to their historical homology degree, followed by assembling these individual fragments. This methodology offers the potential advantage of generating more coherent and contextually relevant sequences. Additionally, it facilitates the verification and validation of the coherence and viability of the generated sequences. However, it is important to recognize that this approach could be considered trivial in certain aspects, as it simply involves the application of a sequential assembly process on segments of text generated synthetically. Figure 16.1 illustrates the schematic representation of this generation methodology.

The advantages of parallel generation of each fragment are:

- Gradual handling of complexity. Generating the genomic sequence piece by piece allows addressing the complexity of the genome in more manageable stages.

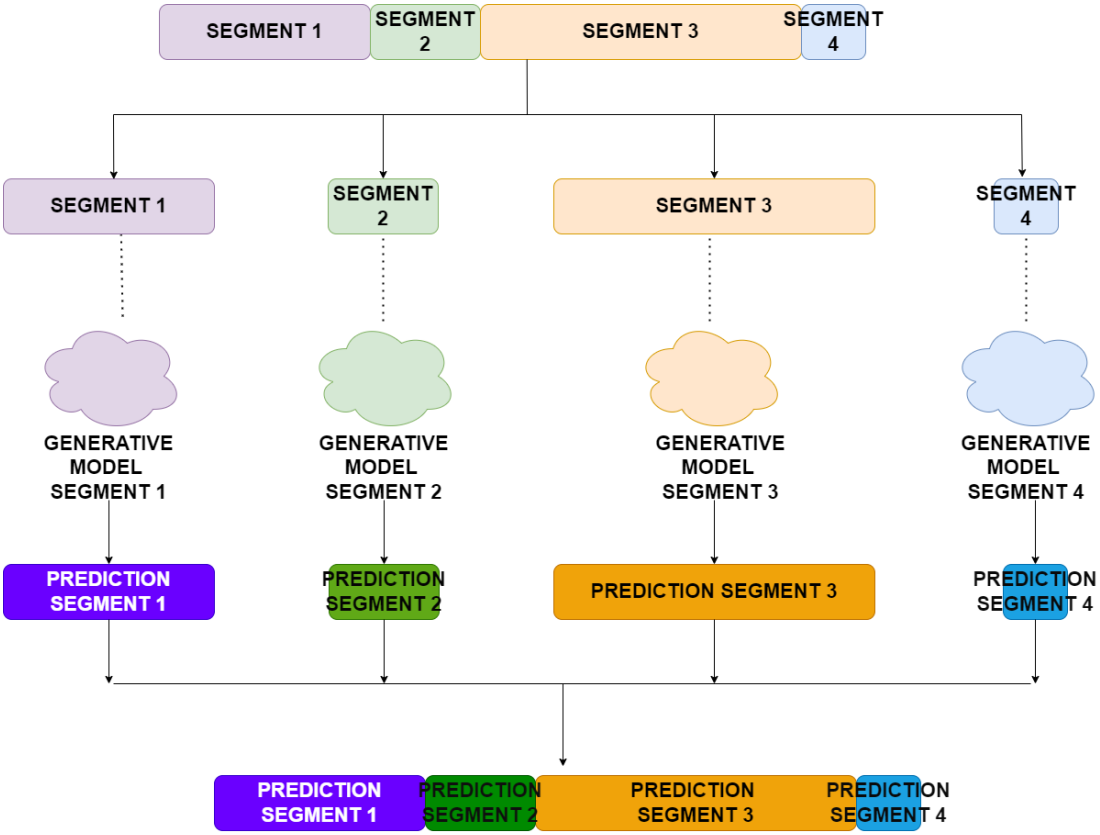


Figure 16.1: Generation of Genomic Segments Individually. Starting from the breakdown of the entire genome into its genomic elements, a generative model is trained for each genomic segment individually. The output of each of them will be the prediction of that genomic element, which will capture the particular characteristics learned by its corresponding generative model. Once the predictions for all genomic elements are obtained, these predictions are assembled. This assembly will be the predicted complete synthetic genome.

This is especially beneficial for large or complex genomes, where dividing the process into smaller pieces facilitates analysis and processing.

- Lower computational burden per iteration. By generating pieces of the genomic sequence individually, the computational load in each iteration can be reduced. This can improve the efficiency of the process and reduce computational resource requirements, enabling faster and more economical analyses.

- Greater flexibility in error resolution. Generating the genomic sequence piece by piece allows for the incremental identification and correction of errors at each stage of the process. This facilitates the early detection of potential issues and provides opportunities for adjustments and continuous improvements as progress

is made in generating the complete sequence.

- Ease of validation. Generating text pieces sequentially makes it easier to validate the coherence and quality of the generated text at each step of the process. This facilitates the identification and correction of potential problems or inconsistencies as progress is made in generating the complete text.

- Possibility of parallelization. Generating pieces of the genomic sequence individually facilitates parallelization of the process on distributed computing systems. This can significantly improve the efficiency and speed of the process, allowing for the generation of complete genomic sequences in a shorter time.

However, individual generation with subsequent assembly entails the following disadvantages:

- Potential loss of global context. Generating the genomic sequence piece by piece carries the risk of losing the global context of the genome. This can lead to coherence issues or lack of continuity in the final sequence, especially if the transition between individual pieces is not handled properly.

- Complexity in integrating pieces. Assembling individual pieces of the genomic sequence can be a complex and error-prone process. Variations in coverage, data quality, or the presence of repetitive regions can make it difficult to integrate the pieces properly, affecting the accuracy and reliability of the final sequence.

- Increased difficulty in detecting structural variations. Generating the genomic sequence piece by piece can make it challenging to detect and characterize structural variations in the genome, such as repetitions or inversions. These variations may be critical for understanding genomic variability and its association with phenotypic traits, so omitting them can limit the utility of the generated genomic sequence.

- Additional validation requirements. Validating the integrity and accuracy of the genomic sequence generated piece by piece may require additional efforts compared to generating the complete sequence. This is because each piece needs

to be validated individually, as well as the integrity and coherence of the final sequence after the integration of the pieces.

- Increased data management complexity. Generating the genomic sequence piece by piece may result in more complex data management, especially if multiple pieces are being generated simultaneously or in parallel. This can increase the complexity of the process and require more sophisticated data management systems to maintain the integrity and traceability of the generated data.

- Increased complexity in adjusting the model during the assembly phase. Generating piece by piece may require a more complex and sophisticated modeling to adequately capture the context and relationships between the generated pieces.

ii) Generating the entire genome simultaneously. This approach has the advantage of being technologically exciting as it challenges current limits of computing and bioinformatics. However, it faces significant challenges in terms of verification and validation due to the complexity and extent of the entire genome. Additionally, the intensive computational load associated with this task can result in additional challenges in terms of efficiency and processing time.

The advantages of simultaneous generation of the complete synthetic genome are as follows:

- Consistency in genomic structure. Generating the entire sequence at once helps maintain the integrity and coherence of the genomic structure as a whole. This is crucial for ensuring the accuracy and fidelity of the final genomic sequence, as any discontinuity or error in the sequence could have significant consequences in subsequent interpretation and analysis.

- Reduction of assembly errors. Generating the complete sequence helps reduce assembly errors that may arise when assembling individual fragments independently. By generating the complete sequence, potential discrepancies or inconsistencies between fragments that could lead to errors in genome reconstruction are avoided.

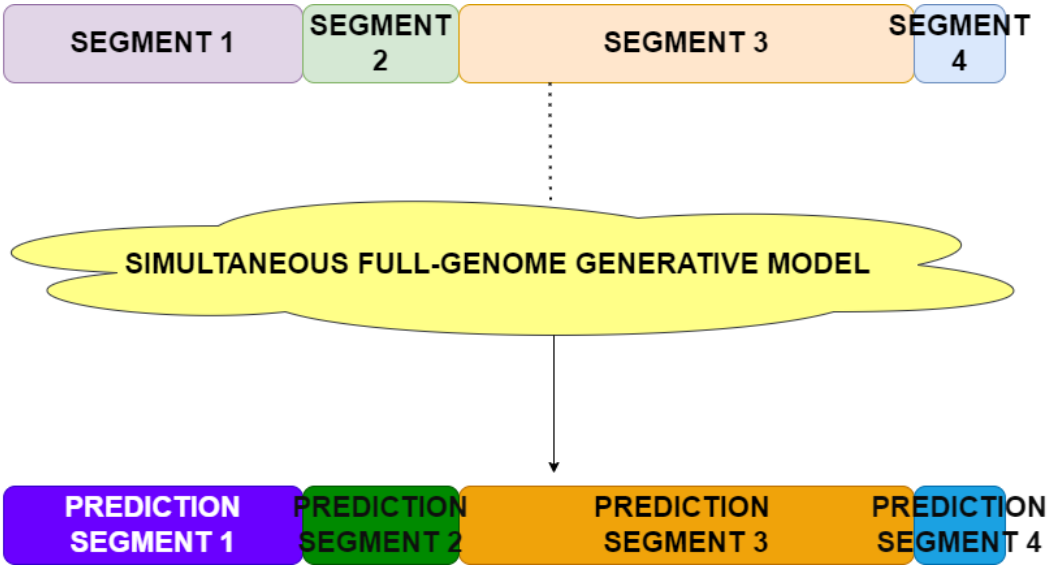


Figure 16.2: Simultaneous Generation of the Entire Genome. A single generative model will make a prediction of the entire genome simultaneously. If the model has been able to learn from the dataset of complete genomes the peculiarities of each genomic element, the synthetic samples generated will likely adequately capture those individual peculiarities, and the output (the complete genome) of that model will be coherent with the real samples from which it learned.

- Efficiency in analysis and interpretation. A complete genomic sequence facilitates the analysis and interpretation of genomic data by providing a holistic view of all genetic elements present in the genome. This can be especially useful in applications such as identifying genetic variants or predicting genomic function, where comprehensive analysis of the entire sequence is required.

- Ease of storage and sharing. Generating the complete sequence simplifies the storage and sharing of genomic data, as it is a single file or dataset containing all genetic information. This facilitates data management and exchange among different researchers or institutions.

- Greater accuracy in detecting genetic variants. By generating the complete sequence, coverage and resolution of the genomic sequence are maximized, increasing the accuracy in detecting genetic variants and mutations. This is crucial in applications such as personalized medicine or comparative genomics, where identifying relevant genetic differences between individuals or species is sought.

However, this methodology may present the following disadvantages:

- Computational complexity. Generating the complete genomic sequence may require a significant amount of computational resources, especially for large or complex genomes. This can be particularly challenging in terms of time and processing capacity, which may limit the scalability and accessibility of complete sequence generation.

- Difficulty in error correction. If errors occur during the process of generating the complete sequence, correcting them can be complicated and costly. Since the complete genomic sequence is an extensive and complex dataset, identifying and correcting errors may require a thorough and detailed review of the entire sequence.

- Risk of cumulative errors. Errors introduced in early stages of the complete sequence generation process can propagate and amplify throughout the entire sequence. This can result in a final genomic sequence with significant errors that may affect subsequent interpretation and analysis.

- Limitations in validation capacity. Validating the accuracy and integrity of the complete genomic sequence can be challenging, especially in the absence of reference data or independent validation methods. This can increase uncertainty about the quality and reliability of the generated sequence.

- Difficulty in detecting structural variations. Generating the complete genomic sequence may make it more difficult to detect and characterize structural variations or complex genomic regions, such as repeats or inversions. These variations can be critical for understanding genomic variability and its association with phenotypic traits, so their omission may limit the utility of the generated genomic sequence.

16.2.2 Potential Architecture

Some of the main generative architectures are:

i) Generative Adversarial Networks (GANs) [176, 177, 3008, 3009]: These consist of two neural networks, a generator and a discriminator. The generator creates synthetic samples attempting to deceive the discriminator, which in turn tries to

distinguish between real and synthetic samples. This adversarial competition drives the generator to continuously improve its results until they are indistinguishable from real data. GANs are widely used in the generation of images, videos, and music, as well as in the simulation of complex scenarios and the creation of creative content.

ii) Variational Autoencoders (VAEs) [2335, 3010, 3011]: These are generative models that learn to represent data in a continuous latent space. Unlike GANs, VAEs optimize a loss function that maximizes the similarity between generated samples and real data while minimizing the divergence between the distributions of real and generated data. They are particularly useful for generating data that follow complex distributions and for learning structured representations of data.

iii) Transformers [175, 3012, 3013]: Initially developed for natural language processing (NLP), Transformers have expanded to text generation and other types of sequential data. They use attention mechanisms to capture long-term relationships and model complex dependencies in sequential data. Transformers have proven effective in machine translation, text summarization, and more recently in image generation and the modeling of complex temporal sequences.

iv) Generative Recurrent Neural Networks (RNNs) [3014–3016]: These architectures are designed to handle sequential data where the output at each time step depends on the current input and previously computed hidden states. Although they have been surpassed in some aspects by Transformers, RNNs, including variants like LSTM and GRU, remain useful for generating sequences of text, music, and video, as well as in time series prediction applications.

v) Generative Flow Networks (Normalizing Flows) [3017–3019]: These architectures transform a simple probability distribution into a more complex distribution that models the training data. They allow for high-quality sample generation and precise probability density estimation. They are used in tasks such as generating high-resolution images and modeling text sequences, where the fidelity of the generated data is crucial.

Compared to Generative Adversarial Networks (GANs), which are powerful for generating synthetic samples but may be less efficient in long-term prediction applications and sequence modeling [3020], the Transformer offers a more robust and systematic alternative [3021]. GANs are suitable for generating genetic variants and simulating viral evolution but lack the structured and bidirectional approach of the Transformer in detailed analysis of complex sequential data such as genomes [3022]. In the case of generating synthetic genomic samples, Transformers present some advantages over GANs:

- Complex Sequence Modeling: The Transformer is designed to capture long-range relationships [3023] and model complex dependencies within sequences [3024], which are essential features for detailed analysis of viral genomes. This allows for a more accurate and in-depth representation of genetic sequences, facilitating the identification of significant mutations and the prediction of evolutionary changes with greater precision than GANs [3025].

- Flexibility and Adaptability: Unlike GANs, which often require a delicate balance between the generator and discriminator and can be sensitive to the quality of training data [176, 3026], the Transformer is more robust and adaptable to different types of genomic data [3027]. It can effectively handle sequences of varying lengths and structures, providing a more flexible platform for integrating multiple sources of genetic and clinical data [3028].

- Interpretation of Biological Data: The Transformer allows for a deeper and more biologically relevant interpretation of genomic data, thanks to its ability to learn hierarchical representations and capture complex patterns in genetic sequences [3029, 3030]. This is crucial in viral genomics, where understanding the functional implications of mutations and genetic interactions is essential for predicting viral evolution and developing therapeutic strategies [3031, 3032].

Compared to the generative flow network (Normalizing Flows), the Transformer also presents certain advantages:

- Complexity and Structure of Genomic Data: The Transformer is more suitable for analyzing genomic data with complex structure and variability in sequence length. Transformers have shown good performance in capturing long-term dependencies [3033] and modeling complex relationships in sequential data [3034], which is extremely important for detailed analysis of viral genomes and mutation prediction.

- Flexibility and Adaptability: Transformers are highly adaptable and can handle different sequence lengths and types of data with relative ease [3035]. Thanks to their ability to learn deep hierarchical representations, they can more easily integrate multiple sources of genetic, clinical, or epidemiological data for a comprehensive assessment of viral evolution [1388, 3036].

- Biological Interpretation of Results: If the goal is not only to generate synthetic genomic sequences but also to biologically interpret predictions [3037] and understand the functional implications of predicted mutations [3038], a Transformer could provide a more intuitive and biologically relevant interpretation of the data.

However, flow-based generative networks are known for their ability to generate high-quality samples [3039] and estimate precise probability densities [3040]. If it is necessary to generate new genomic sequences that closely adhere to the distribution of the training data [3018], maintain high fidelity in representing genetic features [3041], and achieve high accuracy in modeling the data distribution [3042], a flow-based network could be preferable. From my perspective, this approach might be more suitable for microorganisms with lower mutation rates or more predictable and less disruptive genetic variability. In the case of SARS-CoV-2 [26, 3043] and especially with HIV-1 [3044–3046], multiple evolutionary leaps have been surprising, indicating that an architecture tightly constrained to training data, which ultimately reflects historical data, may produce predictions that are insufficiently disruptive and therefore poorly aligned with the current reality of viral evolution.

The Transformer allows for a more flexible approach that may better suit virological reality.

Regarding the Variational Autoencoder (VAE), the Transformer also presents some particularly interesting advantages for genomic sequence prediction:

- Capturing Long-Term Dependencies: The advantage of Transformers lies in their high capacity to capture long-term dependencies within sequences [3033, 3047]. In viral genomics, where the structure and relationships between genetic components are crucial [3034], this feature is especially valuable for modeling complex genetic evolution and interactions between genes and proteins in viruses [3048]. This enables a deep understanding of how mutations can spread and phenotypically affect the virus over time [2710].

While VAEs can also learn useful latent representations for generating genomic sequences, their ability to capture long-term dependencies is limited compared to Transformers. This may result in a less precise representation of complex relationships between genetic sequences, which could affect accuracy in predicting viral evolution [2236].

- Flexibility and Adaptability to Different Sequence Lengths: Transformers, due to their attention-based structure and ability to process sequences of variable length, are highly adaptable to different types of viruses with genomes of varying lengths [3049]. They can effectively handle variability in genetic structure and diversity among viruses, which is crucial for comparative analysis among different viral strains [3050].

However, while VAEs can be effective in generating fixed-length sequences in a latent space [3051], their adaptability to variable-length sequences and management of genomic heterogeneity may be more limited compared to Transformers [3052]. This could result in a reduced capacity to model the actual variability observed in viral genomic data [3053].

- Biological Interpretation of Predictions: Transformers not only generate synthetic sequences that mimic the distribution of real data but can also provide biologically meaningful interpretations of predictions [711, 3054]. This is crucial for understanding the functional implications of predicted mutations and guiding therapeutic and preventive strategies in viral genomics.

In the case of VAEs, the direct biological interpretation of predictions may be less clear compared to Transformers [711, 3055], which are designed to explicitly model relationships between genetic components [1908].

To generate and predict sequences in the context of viral genomics, the advantages of Transformers over Recurrent Neural Networks (RNNs) focus on several key aspects:

- Capturing Long-Term Dependencies: Transformers use self-attention mechanisms that enable capturing long-range relationships among elements within genomic sequences [3056]. This capability allows modeling complex interactions between genes and proteins in viruses and predicting mutation evolution over time with high accuracy [3057, 3058].

RNNs, while capable of capturing temporal dependencies, struggle with long-term dependencies due to the vanishing or exploding gradient problem [3059, 3060]. This limitation can hinder their ability to model complex genetic phenomena and predict viral evolution with the same precision as Transformers.

- Efficiency in Processing Long and Variable Sequences: Thanks to their attention-based structure and parallel processing capability, Transformers are more efficient in handling variable-length sequences and managing genomic variability compared to RNNs [3061]. This is crucial in viral genomics, where viral genomes can vary considerably in length and structure.

In contrast, RNNs process sequences sequentially, which can lead to computational and efficiency limitations when dealing with long sequences or significant variations in viral genome length. This limitation can affect their ability to generate sequences effectively and accurately [3062, 3063].

- Biological Interpretation and Flexibility in Representation: In addition to generating synthetic sequences that mimic the distribution of real data, Transformers can provide biologically meaningful interpretations of predictions. This is essential for understanding the functional implications of mutations and guiding therapeutic and preventive strategies in viral genomics [3064].

While RNNs can be effective for sequence generation, the biological interpretation of predictions may be more complex due to the lack of explicit and structured representations of complex genetic relationships [1897].

Compared to other generative architectures, the Transformer not only excels in capturing complex patterns in sequential data but also offers significant advantages in terms of performance, flexibility, and adaptability to diverse biological and epidemiological applications [3065].

In summary, the Transformer uses attention mechanisms to effectively capture long-range relationships. This is crucial in viral genomics, where accurate detection of mutations and prediction of evolutionary changes require models that can handle genetic sequences comprehensively with coherence and precision [3066, 3067].

Furthermore, the Transformer naturally adapts to the nature of genomic data, where accurate representation of patterns and prediction of non-linear relationships are critical [3068, 3069]. Its ability to learn deep hierarchical representations and model complex dependencies through layers of self-attention positions it as a highly suitable tool for extracting meaningful features in genetic sequences, identifying sites of interest, and assessing genomic variability over time [3049, 3070].

In terms of practical applications, the Transformer facilitates the automatic translation of genomic information across different species [3071] and predicts emerging mutations with considerable accuracy [3072]. Its flexible and scalable architecture also enables efficient integration of multiple data sources, including clinical, epidemiological, and genetic data, for a holistic assessment of viral evolution and the formulation of adaptive preventive and therapeutic strategies [3073, 3074].

Therefore, based on this brief analysis, the architecture of choice a priori is likely the Transformer.

This architecture was initially developed for Natural Language Processing (NLP) tasks [3075], but has proven to be highly effective in a variety of sequential applications, such as automatic translation [3076], text generation [3077], voice synthesis [3078], and speech recognition [3079]. Even in computer vision [3080] or in handling sequential, temporal data [3081].

It is a promising architecture especially for text generation. But it can also be used for discrimination tasks, such as text classification or information extraction [3082].

All thanks to its high capacity to capture long-term dependencies in the input data [3083].

Generative Pre-trained Transformers (GPT) are a class of artificial intelligence models specifically designed for text generation. These models are based on the Transformer architecture, which is highly efficient at processing and understanding long sequences of text. The distinctive feature of GPT models is that they are pre-trained on large amounts of textual data, enabling them to capture complex patterns and understand the structure of language deeply [3084].

During pre-training, GPT models absorb a wide range of linguistic information from various domains, allowing them to acquire a general understanding of language and its usage in different contexts. This pre-training stage provides them with a solid foundation of natural language understanding, which can then be leveraged for specific text generation tasks [3085].

Once pre-trained, GPT models can be fine-tuned for specific text generation tasks, such as automatic article writing, machine translation, or dialogue generation. During this fine-tuning process, the model adapts to a narrower and more specific dataset to produce even more precise and relevant results for the given task [3086].

In summary, Generative Pre-trained Transformers represent a powerful tool in the field of text generation, thanks to their ability to understand natural language and generate coherent and relevant content in a variety of contexts and applications [3087, 3088].

The major innovation of this architecture is the introduction of the attention mechanism [175].

The use of transformers in synthetic genome generation is still in its infancy (synthetic genome generation itself is). The application of transformers is still more oriented towards discrimination, classification, or detection of "linguistic" and/or "grammatical" patterns in the genome [1213, 3089–3091].

Although I identified several developments aimed at various fields, such as predicting gene expression-based phenotypes or pre-miRNAs, protein functions, or DNA methylation prediction [3058, 3092–3094], I did not find any relevant or prestigious research specifically dedicated to generating synthetic genomic sequences using generative Deep Learning models. Despite the scarcity of developments in generating synthetic genomic samples, it could be a suitable architecture.

16.2.3 Technological Challenges

The big question that arose for me was: How do we manage insertions and deletions in the genome? A mere statistical approach may not adequately account for less probable but more determining possibilities. Especially when it is extremely difficult to quantify the imposition capacity of genomic variations. That is, it is difficult to quantify, for example, infectivity, expansion capacity, etc. Would these systems have been able to anticipate Omicron? It imposed itself over all other variants in a brutal way.

What are the most effective methods and approaches for validating the functional coherence, viability, and biological characterization of an artificially generated genome?

It is possible to consider several approaches such as:

i) Comparative analysis with natural genomes. Although we may not know much about the artificially generated genome, we could compare it with well-known natural genomes. We could conduct comparative analyses to identify conserved regions, protein coding patterns, regulatory elements, and other features that suggest functionality or similarity to natural genomes. The presence of these characteristics could indicate that the generated genome has the potential to be viable.

ii) Simulations and computational models. By using computational models and simulations, we could assess the structural and functional stability of the generated genome. We could simulate how the genome interacts with the cellular environment, how it is transcribed and translated into proteins, and how it affects other cellular processes. If these simulations suggest that the genome

is capable of functioning coherently within a cellular environment, it could be an indication of its viability.

iii) Gene expression assays. We could conduct experiments to assess the ability of the generated genome to express genes and produce functional proteins. This could involve introducing the genome into host cells and analyzing gene expression and protein production. If we observe significant gene expression and the production of functional proteins, this would support the viability of the generated genome.

iv) Cellular function evaluation. In addition to gene expression, we could evaluate how the generated genome affects host cells in terms of growth, viability, and overall cellular function. If we observe that cells with the generated genome are able to grow and function similarly to normal cells, this would be a strong indication of its viability.

However, the complexity of this matter is undeniable, especially considering that the available tools often still provide suboptimal solutions and that the level of understanding of the genome is still insufficient.

16.2.4 Some advantages

I would like to highlight the importance of initiatives like the one presented in this Section for several reasons:

i) Aligned with the results obtained in Chapter 10 regarding the variability analysis of each genomic element of SARS-CoV-2, delve deeper into understanding the genome of each genomic fragment of SARS-CoV-2, analyzing its degree of conservation, propensity to undergo modifications (mutations, deletions, insertions, etc.). These measures can serve as a probabilistic reference for the phylogenetic ramifications produced (viral evolution analysis).

ii) Increase predictive capacity for the emergence of new variants. If it were possible to predict the next harmful variant of any virus, it would be possible, among other things, to anticipate the early implementation of infection control measures, the development of vaccines and specific treatments, and the allocation

of healthcare resources based on parameters such as infectivity, transmissibility, mortality, morbidity, etc. Likewise, in the case of variants resistant to treatments or evasive to vaccines, it is possible to plan strategies for reducing spread.

iii) Optimization of response times to new pathogens. As a consequence of the previous point, it would be possible for response times to be minimized due to pre-designed vaccines and treatments, as well as control and care strategies. Therefore, minimal impact on society as a whole would be expected.

iv) Development of methodologies devoid of biological risk. The use of generative models in predicting new variants or pathogens eliminates the need for direct manipulation of biological samples in high-risk laboratories. Virtual sequence generation allows for the analysis and evaluation of potential scenarios without exposure to dangerous biological agents. This significantly reduces the biological risk associated with direct sample manipulation, thereby mitigating the possibility of accidents or leaks that could have serious consequences for public health and the environment. Additionally, by eliminating the need for experimental work in test tubes, costs and logistical requirements associated with handling and storing biological samples are reduced, while accelerating the research and development process of countermeasures against potential emerging biological threats.

Therefore, although currently it may not be viable, or even possible, to develop models that effectively and robustly predict variants or pathogens, the use of Deep Learning models is undoubtedly a promising alternative that must be worked on and developed.

There is still much to be done.

But with each passing day, we are getting closer to achieving it.

The ocean is a vast collection of water droplets.

16.3 Genomic Size Delimitation for the Effective Application of Genomic Spectrograms

A future line of research will be precisely to delineate the maximum lengths of complete genomes with which one can work correctly and, therefore, to which

organisms the scope of application of genomic spectrograms may be limited.

Concatenated genomic spectrograms retain more genomic information compared to superposed spectrograms. The summation of values in the spectrogram in the case of superposed spectrograms results in the loss of information regarding the degree of relevance or involvement of each nucleotide type in the mathematical signatures. However, concatenated spectrograms entail a loss of precision in the detailed location of the genomic regions involved. For this reason, I simultaneously worked with both types of representation.

If it were technologically feasible to overcome the disadvantage of precision loss in concatenated spectrograms, it would be a richer representation than superposed spectrograms, as we would be able to precisely determine the relevant genomic regions while also preserving information regarding the relevant nucleotide types. And all of this would be achieved using a single type of image.

In this research, I worked with viruses, whose genome is the smallest on the planet, with no more than thirty thousand nucleotides. However, the genome of other organisms is significantly larger. The bacterium *Salmonella enterica* subsp. Enterica Serovar Typhimurium strain, which is a common cause of food poisoning, has a circular genome approximately 4.9 million nucleotides in length. The genome of *Saccharomyces cerevisiae*, which is a fungus commonly used in research and food production, has approximately 12 million DNA nucleotides. The genome of the long-tailed lizard *Anolis carolinensis*, which is a species commonly used in research, has approximately 1.8 billion DNA nucleotides. And the entire human genome is composed of approximately 3 billion DNA nucleotides.

Working with complete genomes allows us to fully utilize all the information, without possible losses of apparently unrelated data due to ignorance. Is it possible to apply genomic spectrograms to complete genomes composed of millions of nucleotides? How do we computationally handle this burden?

The x-axis of concatenated spectrograms is 4 times the length of the represented genome. Imagine the size of a concatenated genomic spectrogram of any mammal,

for example. The superposed spectrogram, where the x-axis coincides with the length of the genome, would already have a considerable dimension.

Where is the limit in genomic size for pattern detection by a CNN? What requirements should the images have in terms of clarity to interpret the results if the CNN properly detects patterns and explainability indicates their location?

We will need to perform multiple test benches with different organisms to determine those maximum sizes. Likewise, we will need to research new models and architectures, new technologies, and new applications that allow us to adapt the methodology outlined in this research to large genomes, to the human genome. And to be able to replicate these experiments with genomes of billions of nucleotides, not just a few tens of thousands.

For now, we will limit ourselves to establishing the limits of application of existing models.

16.4 Deeper Understanding of the Detected Mathematical Signatures

The research question is very simple and constantly repeated throughout the document:

We know where the CNN looks to make a decision, but what exactly does it see?

The culmination of this line of research would be to capture the mathematical formulas that determine whether a sequence of HIV-1 or SARS-CoV-2 is recombinant or not, or whether a non-recombinant sequence of SARS-CoV-2 falls within the pre-VOC, Alpha, Delta, or Omicron variant.

But let's take it step by step. I detected indications that the mathematical signature associated with the recombinant feature in SARS-CoV-2 may be related to some pattern in the periodicity of the third codon nucleotide. These nucleotides, therefore, may contain encrypted information regarding virus characterization. In the case of variant characterization, the significant role of periodicities 3 and their multiples in coding regions may also lead me to think the same. In the case of

HIV-1, since the LTRs are non-coding regions, we cannot talk about codons, but the nucleotide positions muktiple of 3 also had their relevance. What does all this mean?

Simply put, the accurate demonstration of the role of these third nucleotides would represent a paradigm shift in genome interpretation. Just as discovering that the ORF1ab polyprotein in SARS-CoV-2 and other coronaviruses, through a frameshift of one nucleotide back, switched from reading ORF1a to reading ORF1b, thereby compressing maximum information into minimal genomic space.

If the third nucleotides contain encrypted information, this represents a new layer of genome reading. Are there more layers?

So many questions...

Humbly, for now, I will continue to investigate the knowledge of the patterns already identified, the clues already detected. And I will gradually approach what may someday be an unequivocal mathematical formula for the recombinant feature, membership in a variant, or who knows how many more characteristics.

I will start by conducting a comprehensive bio-mathematical approach to potential patterns associated with the genetic recombination of HIV-1 and SARS-CoV-2 sequences in their frequency spectrum representation. The same applies to the variants of SARS-CoV-2.

16.5 Identification of Mathematical Signatures Associated with Antifungal Resistance

Just as I conducted my research by working on a specific case study, I aim to analyze the existence of mathematical signatures in the genome of a fungus associated with antifungal resistance.

The proliferation of antifungal-resistant fungal strains is currently a public health issue [3095, 3096].

Antifungal resistance is a complex phenomenon involving several mechanisms at the molecular, cellular, and clinical levels [3097].

i) Some of the main mechanisms of resistance are [3098]:

- Alteration of the target. Fungi can develop mutations in the genes that encode the targets of antifungals, such as enzymes involved in the synthesis of the cell wall or plasma membrane. These mutations can reduce the affinity of the antifungal for its target, thereby decreasing its efficacy.

- Increased efflux efficiency. Some fungi can increase the expression of efflux transporters, proteins that actively pump antifungals out of the cell. This decreases the intracellular concentration of the antifungal, reducing its effectiveness.

- Production of inactivating enzymes. Some fungi can produce enzymes capable of degrading or chemically modifying antifungals, which reduces their activity.

ii) Some of the main factors contributing to resistance are [3099]:

- Excessive and improper use of antifungals. The overuse or inappropriate use of antifungals can select for resistant strains of fungi.

- Immunocompromised patients. Individuals with weakened immune systems, such as patients with HIV/AIDS or transplant recipients, are at higher risk of fungal infections and developing resistance to antifungals due to prolonged exposure to these compounds.

- Hospital environments. Hospitals and other healthcare settings can be reservoirs for resistant strains, especially in intensive care units and oncology wards.

iii) Some of the clinical consequences are [3100, 3101]:

- Difficulty in treatment. Antifungal resistance can make fungal infections more difficult to treat, potentially resulting in higher morbidity and mortality.

- Increased risk of complications. Resistant fungal infections can increase the risk of complications, such as the spread of the infection to other organs or systems of the body.

- Higher healthcare costs. Treating resistant fungal infections may require longer and more expensive therapies, leading to increased healthcare costs.

iv) This issue has been managed through various approaches [3102, 3103]. Some of the main ones are:

- Development of new antifungals. New antifungals with different mechanisms of action are needed to overcome resistance.

- Prudent use of antifungals. It is important to use antifungals appropriately and avoid overuse to reduce selective pressure on fungal strains.

- Surveillance of resistance. Continuous monitoring of antifungal resistance is essential to detect and respond quickly to changes in resistance patterns.

The methodology based on genomic spectrogram + CNN + explainability tools can be extremely useful and efficient in detecting which strains or sequences are resistant to antifungals. This not only allows for rapid surveillance but also provides information that can be useful in the design of treatments or preventive systems.

I will conduct this study on genomic sequences of *Candida albicans* (the fungus responsible for candidiasis).

C. albicans is a commensal fungus commonly found in the gastrointestinal tract and on the skin of humans. However, under certain conditions, it can become an opportunistic pathogen and cause infections in various tissues and organs of the body, especially in individuals with compromised immune systems or those undergoing invasive medical treatments [3104, 3105].

C. albicans is one of the main fungal pathogens affecting humans. Infections caused by *C. albicans* represent a significant burden on healthcare systems, both in terms of direct costs associated with treating infections and in terms of the morbidity and mortality associated with them. It is responsible for a wide range of infections, from superficial mucocutaneous infections to potentially life-threatening systemic infections, especially in vulnerable populations such as immunocompromised patients, patients hospitalized in intensive care units, and neonates [3106].

C. albicans has the ability to spread easily in healthcare settings, which can result in nosocomial outbreaks and the transmission of resistant strains between patients. This can increase the disease burden and pressure on public health resources [3107].

The treatment of infections caused by *C. albicans* and other species of the *Candida* genus generally involves the use of antifungals, which are drugs designed to inhibit growth or eliminate fungi. However, in recent decades, there has been a concerning increase in *C. albicans* resistance to various antifungals, posing significant challenges in the management of fungal infections [3108].

C. albicans has developed significant resistance to several antifungals commonly used in clinical practice, including azoles [3109], echinocandins [3110], and amphotericin B [3111]. This resistance can make the treatment of fungal infections more difficult and increase the morbidity and mortality associated with these infections.

Antifungal resistance in *C. albicans* can occur due to various mechanisms, including:

i) Modifications in the target. *C. albicans* can develop mutations in genes encoding the target proteins of antifungals, such as enzymes targeted by azoles (e.g., 14-α-demethylase) or cell wall proteins targeted by echinocandins (e.g., β-glucan synthase). These mutations can decrease the antifungal's affinity for its target, thus reducing its effectiveness [3112, 3113].

ii) Increased efflux expression. *C. albicans* can overexpress efflux pumps, such as ATP-binding cassette (ABC) family proteins or multidrug resistance-associated superfamilies (MRP), which actively eject antifungals from fungal cells, thereby reducing their intracellular concentration and therapeutic effectiveness [3114, 3115].

iii) Biofilm formation. *C. albicans* has the ability to form biofilms, which are microbial communities adhered to biotic or abiotic surfaces. Biofilms provide a protected environment that can confer resistance to antifungals by limiting their access to fungal cells and reducing their effectiveness [3116, 3117].

iv) Heteroresistance. Some strains of *C. albicans* may contain subpopulations with varying levels of sensitivity to antifungals, known as heteroresistance. This can make complete eradication of the infection challenging, as resistant subpopulations may persist and proliferate even after antifungal treatment [3118, 3119].

C. albicans resistance to antifungals is a significant clinical issue that can complicate the treatment of fungal infections, especially in immunocompromised patients or in intensive care units where the prevalence of nosocomial infections by *C. albicans* is high. To address this challenge, it is crucial to develop more effective diagnostic and treatment strategies, as well as promote rational use practices of antifungals to prevent the development and spread of resistance [3120].

To the extent possible, I will work with complete genomes. In case it is unfeasible, I will design fragmentation and parallelization systems to overcome this limitation.

My goal will be to identify genomic areas relevant to antifungal resistance in strains of this fungus to, in the future, improve the management of fungal infections, develop prevention strategies, and ensure that antifungals remain effective for treating these diseases in the future.

16.6 Identification of Mathematical Signatures Associated with Antibiotic Resistance

Similar to the study of antifungal resistance, it is imperative to extrapolate my methodology to genome detection of bacterial signatures associated with antibiotic resistance. Antibiotic resistance is a complex phenomenon that arises as a result of selective pressure exerted by the inappropriate or excessive use of antibiotics. Some data:

i) Some of the main mechanisms of antibiotic resistance are [3121]:

- Genetic Mutations. Bacteria can acquire mutations in their genes that confer resistance to antibiotics. These mutations can directly affect the antibiotic's target, alter the permeability of the cell membrane, or modify the enzymes that metabolize the drug [3122].

- Formation of biofilms. Bacterial biofilms can provide a protective environment that reduces the effectiveness of antibiotics by limiting their access to the bacteria [3123, 3124].

- Horizontal gene transfer. Bacteria can exchange genetic material with other bacteria through processes such as conjugation, transformation, and transduction. This can enable the transfer of resistance genes between different bacterial species, contributing to the spread of resistance [3125, 3126].

The first two mechanisms share general characteristics with antifungal resistance. However, horizontal gene transfer is more specific to bacteria.

ii) Some factors contributing to antibiotic resistance are [3127]:

- Inappropriate use of antibiotics. Excessive or incorrect use of antibiotics, such as unnecessary prescription, self-medication, and improper dosing administration, can foster the development of bacterial resistance [3128].

- Use in agriculture and livestock. The use of antibiotics in agriculture, both as therapeutic agents and growth promoters, can increase selective pressure on bacteria in the environment, thus contributing to resistance [3129].

- Hospital environments. Hospitals and other healthcare settings can be reservoirs of resistant bacteria, especially in intensive care units and areas where invasive procedures are performed [3130].

iii) The clinical consequences of this resistance include, among others [3131, 3132]:

- Increased mortality. Infections caused by antibiotic-resistant bacteria can be more challenging to treat and are associated with higher rates of morbidity and mortality.

- Increased morbidity. Antibiotic resistance can prolong the duration of illness and increase the risk of complications, resulting in a greater burden on both the patient and the healthcare system.

- Economic costs. Treating infections caused by resistant bacteria may require longer and more expensive therapies, leading to increased healthcare costs and potentially significant economic impact worldwide.

iv) Several approaches have been developed to address resistance [3133–3135]:

- Prudent use of antibiotics. It is crucial to promote the prudent use of antibiotics through the education of healthcare professionals and the public, as well as the implementation of responsible prescription policies.

- Development of new antibiotics. Continuous research and development are needed to identify and develop new antibiotics with different mechanisms of action that can overcome bacterial resistance.

- Surveillance of resistance. Continuous surveillance of antibiotic resistance is essential for monitoring resistance patterns, identifying emerging outbreaks, and guiding control and prevention strategies.

Similar to Section 16.5, effective development of my methodology enables rapid surveillance of bacteria that have developed these resistances. Identification of signatures associated with this characteristic allows for quick detection, reducing the processing time of a genomic spectrogram by the CNN. This minimization of detection times will provide experts with margin and additional information for the development of strategies to combat these resistant bacteria.

My goal will be the identification of differentiating genomic areas in strains of bacteria resistant to antibiotics in the following detailed case studies.

16.6.1 *Mycobacterium tuberculosis*

Mycobacterium tuberculosis is the etiological agent of tuberculosis (TB), an infectious disease that primarily affects the lungs but can spread to other organs. It is an acid-resistant, aerobic, slow-growing bacterium, with the following distinctive characteristics [3136]:

- Rich lipid cell wall. Its cell wall contains a high concentration of mycolic acids, which confers resistance to desiccation, many disinfectants, and staining with conventional dyes.

- Latency capacity. It can remain in a latent state in the human body for years, reactivating when the host's immune system is weakened.

- Intracellularity. It is capable of surviving and replicating within macrophages, evading the host's immune response.

We can find different degrees of antibiotic resistance of this bacterium [3137]:

a) Multidrug-resistant (MDR-TB) [3138]:

Strains resistant to isoniazid and rifampicin, the two most potent first-line drugs [3139].

Mechanisms of resistance:

- Isoniazid. Mutations in the **katG** gene (which encodes catalase-peroxidase) [3140], responsible for activating the drug, and in the **inhA** gene (related to mycolic acid biosynthesis) [3141].

- Rifampicin. Mutations in the **rpoB** gene, which encodes the beta subunit of RNA polymerase, the drug's target [3142].

b) Extensively drug-resistant (XDR-TB) [3143]:

Strains resistant to isoniazid and rifampicin, plus resistance to any fluoroquinolone and at least one of the second-line injectable drugs (amikacin, kanamycin, or capreomycin) [3144].

Mechanisms of resistance:

- Fluoroquinolones. Mutations in the **gyrA** and **gyrB** genes, which encode DNA gyrase [3145].

- Aminoglycosides. Mutations in the **rrs** and **eis** genes [3146].

c) Extremely drug-resistant (XXDR-TB) [3147]:

Strains Resistant to virtually all available drugs, including newer agents such as bedaquiline and delamanid [3148].

- Mechanisms of resistance: Various mutations in additional genes that are still being investigated [3149].

In addition to specific mutations in target genes, *Mycobacterium tuberculosis* employs other mechanisms of resistance:

- Efflux pump. Increases the expression of efflux pumps that expel antibiotics out of the bacterial cell [3150].

- Target site modification. Alterations in the target proteins of antibiotics, which decrease the drug's affinity [3151].

- Enzymatic inactivation. Production of enzymes that inactivate antibiotics, although this mechanism is less common in *Mycobacterium tuberculosis* compared to other bacteria [3152].

Antibiotic resistance in *Mycobacterium tuberculosis* has significant implications for public health:

- Longer and more costly treatments. MDR-TB and XDR-TB require treatment with multiple second-line drugs, which are more toxic, less effective, and require a longer treatment period (up to 24 months or more) [3153].

- Severe side effects. Second-line drugs have high toxicity profiles, which can lead to serious adverse effects and adherence issues to the treatment [3154].

- Higher mortality and morbidity. Resistant strains are associated with worse clinical outcomes and a higher mortality rate [3155].

- Spread of resistant strains. The transmission of resistant strains in the community and healthcare settings poses a significant threat to tuberculosis control [3156].

To address antibiotic resistance in *Mycobacterium tuberculosis*, several strategies are being implemented [3157]:

- Development of new drugs. Research and development of new antimicrobial agents, such as bedaquiline and delamanid, which show activity against resistant strains.

- Rapid diagnostic tests. Implementation of rapid diagnostic methods, such as GeneXpert MTB/RIF, which allows for the quick detection of rifampicin resistance.

- Combination therapies. Use of drug combinations to reduce the likelihood of resistance development.

- Treatment adherence. Patient follow-up programs and support to improve treatment adherence, such as Directly Observed Therapy (DOT).

- Infection control. Strict infection control measures in hospitals and communities to prevent the spread of resistant strains.

In summary, antibiotic resistance in *Mycobacterium tuberculosis* is a complex challenge that requires a multidisciplinary approach including the development of new drugs, improvements in diagnosis, combined treatment strategies, and effective public health measures.

16.6.2 *Clostridium difficile*

Clostridium difficile is a Gram-positive, anaerobic, spore-forming bacterium, which is the main causative agent of pseudomembranous colitis and other severe intestinal

infections, especially associated with antibiotic use. Below is a detailed and expert review of *Clostridium difficile* and its antibiotic resistance [3158].

The main characteristics of *Clostridium difficile* are as follows:

- Spore formation. *Clostridium difficile* spores are highly resistant to adverse conditions, allowing them to persist in the hospital environment and facilitate transmission [3159].

- Toxin production. Pathogenic strains of *Clostridium difficile* produce toxins, mainly toxin A (enterotoxin) and toxin B (cytotoxin), which are responsible for inflammation and damage to the intestinal mucosa [3160].

- Strict anaerobe. *Clostridium difficile* grows in oxygen-free environments, allowing it to colonize and proliferate in the large intestine [3161].

Antibiotic resistance in *Clostridium difficile* is a significant concern and can manifest through various mechanisms:

i) Resistance to common antibiotics:

- Clindamycin. Resistance mediated by mutations in the **ermB** gene, which encodes a methylase that modifies the antibiotic's binding site on the bacterial ribosome [3162].

- Fluoroquinolones. Mutations in the **gyrA** and **gyrB** genes, which encode DNA gyrase, decreasing the affinity for fluoroquinolones [3163].

- Metronidazole and vancomycin. Although significant clinical resistance to these antibiotics is rare, cases of reduced susceptibility have been reported [3164].

ii) Epidemics associated with hypervirulent strains:

- Hypervirulent strains (e.g., NAP1/BI/027). These strains exhibit increased antibiotic resistance, produce more toxins, and cause severe outbreaks in hospital settings [3165, 3166].

Some of the antibiotic resistance mechanisms of *Clostridium difficile* discovered are as follows:

- Modification of the target site. Alterations in target enzymes or ribosomes that reduce the affinity of antibiotics [3167].

- Efflux pumps. Although less studied in *Clostridium difficile*, increased expression of efflux pumps may contribute to resistance [3168].

- Inactivating enzymes. Production of enzymes that deactivate antibiotics, although this mechanism is less common in *Clostridium difficile* [3169, 3170].

Antibiotic resistance in *Clostridium difficile* has significant implications for public health and clinical practice:

- Nosocomial infections. *Clostridium difficile* is a leading cause of healthcare-associated infections, especially in patients who have received antibiotic therapy [3171].

- Mortality and morbidity. Infections caused by *Clostridium difficile* can be severe and are associated with high mortality rates, especially in elderly patients and those with comorbidities [3172].

- Healthcare costs. Managing *Clostridium difficile* infections is expensive due to the need for prolonged treatments, patient isolation, and infection control measures [3173].

To address the antibiotic resistance of *Clostridium difficile*, several mitigation strategies are implemented:

- Rational use of antibiotics. Reducing unnecessary antibiotic use to prevent intestinal dysbiosis that favors the proliferation of *Clostridium difficile* [3174].

- Targeted treatments. Use of drugs such as fidaxomicin, which has greater specific activity against *Clostridium difficile* and less impact on the intestinal microbiota [3175].

- Alternative therapies. Fecal Microbiota Transplantation (FMT) has proven effective in restoring the intestinal microbiota and preventing recurrences [3176].

- Infection control. Implementation of strict infection control measures in hospital settings, including hand hygiene, the use of gloves and gowns, and environmental disinfection with products effective against spores [3177].

- Surveillance and monitoring. Continuous monitoring of antibiotic resistance and epidemiological patterns of *Clostridium difficile* to quickly identify and control outbreaks [3178].

iii) Research and Development:

The ongoing study of *Clostridium difficile* and its antibiotic resistance is crucial for the development of new therapeutic and preventive strategies. Research focuses on:

- Development of new antibiotics. Investigation of new compounds with specific activity against *Clostridium difficile* [3179].

- Vaccines. Development of vaccines to prevent colonization and infection by *Clostridium difficile* [3180].

- Immunotherapy. Use of monoclonal antibodies against *Clostridium difficile* toxins as an adjuvant therapy [3181].

16.6.3 Vancomycin-Resistant Enterococci

Vancomycin-Resistant Enterococci (VRE) are one of the most significant concerns in the field of nosocomial infections due to their ability to resist standard treatments and their potential to cause serious infections [3182].

Enterococcus is a genus of Gram-positive bacteria with the following key characteristics [3183]:

- Commensal inhabitants. They are part of the normal intestinal microbiota in humans and other animals [3184].

- Relevant species. The most important species from a clinical perspective are *Enterococcus faecalis* and *Enterococcus faecium* [3185].

- Intrinsic resistance. They possess natural resistance to many antibiotics, including aminoglycosides and cephalosporins [3186].

Vancomycin resistance in enterococci has become a significant problem. It is characterized by the acquisition of resistance genes that modify the site of action of vancomycin [3187], such as [3188]:

- vanA. It confers high resistance to vancomycin and teicoplanin. It is found on plasmids that can be horizontally transferred between bacteria.

- vanB. It confers resistance to vancomycin but not to teicoplanin. It is also found on mobile elements that facilitate horizontal transfer.

- vanC, vanD, vanE, vanG. These genes are less common and generally associated with low-level resistance.

All these genes encode enzymes that alter the D-alanine-D-alanine termination of the cell wall pentapeptide, changing it to D-alanine-D-lactate or D-alanine-D-serine, which reduces vancomycin's affinity for its binding site [3189–3191].

VRE have significant clinical and public health implications:

- Nosocomial infections. VRE are responsible for serious hospital-acquired infections, including bacteremia, endocarditis, urinary tract infections, and wound infections [3192].

- Treatment complexity. Treatment options for infections caused by VRE are limited and typically include less effective and more toxic drugs such as linezolid, daptomycin, and quinupristin/dalfopristin [3193].

- Isolation and infection control. The spread of VRE in hospitals requires strict infection control measures, such as patient isolation and the use of protective barriers [3194].

- Horizontal transmission of resistance genes. The ability to transfer resistance genes to other Gram-positive pathogens, such as Staphylococcus aureus, represents a significant threat to public health [3187, 3195].

To address vancomycin resistance in enterococci, several strategies are implemented:

i) Prevention and Infection Control [3194]:

- Hand hygiene. Rigorous adherence to hand hygiene practices with alcohol-based solutions or antimicrobial soap.

- Patient isolation. Use of private rooms and contact precautions for patients infected or colonized with VRE.

- Environmental disinfection. Thorough cleaning and disinfection of surfaces and medical equipment.

ii) Antibiotic Treatment:

- Use of alternative antibiotics. Utilization of drugs such as linezolid, daptomycin, tigecycline, and quinupristin/dalfopristin to treat VRE infections [3196].

- Combination strategies. Use of combinations of antibiotics to overcome resistance and reduce the risk of further resistance development [3197].

iii) Surveillance and Monitoring [3198]:

- Surveillance programs. Continuous monitoring of VRE incidence in hospitals and communities to detect and control outbreaks rapidly.

- Susceptibility testing. Performing antimicrobial susceptibility testing to guide appropriate therapy and promptly detect resistant strains.

iv) Research and Development:

- New antibiotics. Research and development of new antimicrobial agents with activity against VRE [3199].

- Vaccines. Development of vaccines that can prevent colonization and infection by enterococci [3200].

- Resistance inhibitors. Development of compounds that can inhibit the enzymes responsible for vancomycin resistance [3201].

Vancomycin resistance in enterococci has a significant impact on public health because [3202]:

- Increased mortality. VRE infections are associated with higher mortality due to limited treatment options.

- Prolonged hospital stays. Patients with VRE infections often have longer hospital stays, increasing healthcare costs.

- Nosocomial transmission. VRE spreads easily in hospital environments, making infection control challenging and increasing the burden on healthcare systems.

16.6.4 Methicillin-Resistant *Staphylococcus aureus*

Methicillin-resistant *Staphylococcus aureus* (MRSA) is one of the most significant and concerning bacteria in the realm of both hospital-acquired and community-acquired infections due to its ability to resist multiple antibiotics. Below is a detailed and expert review of MRSA and its antibiotic resistance [3203].

Staphylococcus aureus is a Gram-positive bacterium with the following key characteristics [3204]:

- Clustered cocci. Morphologically, they appear in clusters resembling bunches of grapes [3205].

- Catalase-positive and coagulase-positive. These biochemical tests are important for its identification [3206, 3207].

- Opportunistic pathogen. Capable of causing a wide range of infections, from mild skin infections to severe and potentially life-threatening infections such as sepsis, endocarditis, and pneumonia [3208].

Methicillin resistance in *S. aureus* is a critical characteristic defining MRSA. This resistance is primarily mediated by the **mecA** gene, which encodes an altered Penicillin-Binding Protein 2a (PBP2a) that reduces affinity for methicillin and other β-lactams [3209, 3210], although allowing for the synthesis of the cell wall in the presence of these antibiotics [3211]. The **mecA** gene is located on a mobile genetic element called the Staphylococcal Chromosomal Cassette mec (SCCmec), which can be transferred between strains [3212].

Moreover, *S. aureus* can acquire of additional genes that confer resistance to other antibiotics, such as aminoglycosides, macrolides, and fluoroquinolones [3213].

MRSA has significant clinical and public health implications:

- Nosocomial infections. MRSA is a leading cause of hospital-acquired infections, including surgical site infections, pneumonia, bacteremia, and sepsis [3214].

- Community-acquired infections. Community-acquired MRSA (CA-MRSA) strains are responsible for skin and soft tissue infections and can affect healthy individuals [3215].

- Complicated treatment. Treatment options are limited due to resistance to multiple antibiotics, necessitating the use of reserve antibiotics such as vancomycin, linezolid, and daptomycin [3216].

- Control of infections. The transmission of MRSA in hospital and community settings requires strict infection control measures, including hand hygiene, the use of personal protective equipment, and the isolation of infected or colonized patients [3217].

To address MRSA resistance, several strategies are implemented:

i) Prevention and Infection Control [3218]:

- Hand hygiene. Rigorous hand hygiene practices to prevent transmission.

- Patient isolation. Use of private rooms and contact precautions for infected or colonized patients.

- Environmental disinfection. Regular cleaning and disinfection of surfaces and medical equipment.

ii) Antibiotic Treatment [3219, 3220]:

- Use of alternative antibiotics. Utilization of drugs such as vancomycin, linezolid, daptomycin, tigecycline, and ceftaroline to treat MRSA infections.

- Combination strategies. Employment of antibiotic combinations to enhance efficacy and reduce the risk of further resistance.

iii) Surveillance and Monitoring [3221, 3222]:

- Surveillance programs. Continuous monitoring of MRSA incidence in hospitals and communities to detect and control outbreaks promptly.

- Susceptibility testing. Conducting antimicrobial susceptibility testing to guide appropriate therapy and promptly detect resistant strains.

iv) Research and Development:

- New antibiotics. Research and development of new antimicrobial agents with activity against MRSA [3223].

- Vaccines. Development of vaccines that can prevent colonization and infection by MRSA [3204].

- Resistance inhibitors. Development of compounds that can inhibit the enzymes responsible for antibiotic resistance [3224].

Methicillin resistance in *Staphylococcus aureus* has a significant impact on public health because [3225]:

- Increased mortality. MRSA infections are associated with higher mortality due to limited treatment options.

- Prolonged hospital stays. Patients with MRSA infections often have longer hospital stays, increasing healthcare costs.

- Nosocomial and community transmission. MRSA spreads easily in hospital and community settings, making infection control challenging and increasing the burden on healthcare systems.

16.6.5 *Neisseria gonorrhoeae*

Antibiotic-resistant *Neisseria gonorrhoeae* is a critical concern in global public health due to its ability to develop resistance to multiple treatments, complicating its management and control. Below is a detailed and expert review of *N. gonorrhoeae* and its antibiotic resistance. *Neisseria gonorrhoeae* is a Gram-negative bacterium with the following key characteristics [3226]:

- Diplococci. Morphologically, it presents as pairs of cocci [3227].

- Oxidase positive and catalase positive. Important biochemical tests for its identification [3228].

- Human-exclusive pathogen. It is the causative agent of gonorrhea, a Sexually Transmitted Infection (STI) [3229].

Antibiotic resistance in *N. gonorrhoeae* has evolved rapidly, leading to a reduction in effective therapeutic options [3230]. The bacterium has developed resistance

to nearly all antibiotics used to treat it since the introduction of sulfonamides in the 1930s [3231].

The origin of some of the main resistance mechanisms identified are:

- Target modification. PenA. Mutations in the **penA** gene alter the PBP2 protein, decreasing affinity for penicillins and cephalosporins [3232, 3233].

- Efflux pumps. MtrR. Mutations in the **mtrR** gene increase the expression of the MtrCDE efflux system, which pumps multiple classes of antibiotics out of the bacterial cell [3234].

- Reduced permeability. PorA and PorB. Alterations in these genes affect the permeability of the outer membrane, reducing the entry of antibiotics [3235, 3236].

- Acquisition of plasmids and mobile elements. Horizontal transfer of resistance genes through plasmids and transposons has contributed to the spread of resistance [3237, 3238]

- Ribosome alteration. Modifications in ribosomes can confer resistance to macrolides and tetracyclines [3239, 3240].

Antibiotic-resistant *N. gonorrhoeae* has significant clinical and public health implications [3241–3243]:

- Clinical Complications. Resistance limits treatment options, making infections more difficult to treat and increasing the risk of complications such as pelvic inflammatory disease, infertility, and systemic dissemination.

- Continued Transmission. Resistance can lead to persistent infections and facilitate ongoing transmission within the population.

- Limited Treatment Options. With the decreasing efficacy of traditional antibiotics, new therapies and alternative approaches are required.

To address *N. gonorrhoeae* resistance, several strategies are implemented:

i) Prevention and Infection Control [3244]:

- In addition to awareness campaigns, increasing access to rapid and accurate tests to detect gonorrhea and its resistance profile is fundamental.

ii) Antibiotic Treatment:

- Use of antibiotic combinations. To reduce the likelihood of resistance development, the use of antibiotic combinations such as ceftriaxone with azithromycin is recommended [3245, 3246].

- Development of new antibiotics. Research and development of new antimicrobial agents with activity against resistant *N. gonorrhoeae* [3247].

iii) Surveillance and Monitoring [3244]:

- Surveillance Programs. Continuous monitoring of antibiotic resistance in *N. gonorrhoeae* worldwide to rapidly detect resistance outbreaks.

- Epidemiological Data. Collection and analysis of data to inform treatment guidelines and public health policies.

iv) Research and Development:

- New Therapeutic Approaches. Research into alternative therapies, such as immunotherapy and resistance inhibitors [3248, 3249].

- Vaccines. Development of vaccines that can prevent *N. gonorrhoeae* infection [3250].

Antibiotic resistance in *N. gonorrhoeae* has a significant impact on public health because [3251]:

- Increased morbidity. Resistant infections can lead to higher morbidity due to complications and the need for prolonged and costly treatments.

- Prolonged transmission. Untreated or poorly treated resistant infections can continue to spread within the community.

- Economic burden. Treating resistant infections and associated complications represent a significant economic burden for healthcare systems.

Given all the above, antibiotic-resistant *N. gonorrhoeae* is likely one of the most critical challenges for global public health currently.

16.6.6 Carbapenem-Resistant Enterobacteriaceae (*Escherichia coli*, *Klebsiella pneumoniae*,)

Carbapenem-resistant enterobacteriaceae, including *Escherichia coli* and *Klebsiella pneumoniae*, represent a significant threat to global public health due to their ability to resist some of the most potent antibiotics available. Enterobacteriaceae are a group of Gram-negative bacteria that include *E. coli*, *K. pneumoniae*, *Enterobacter* spp., among others. They have the following key characteristics [3252]:

- Bacillus morphology. Rod form [3253].

- Facultative anaerobes. Can grow in the presence or absence of oxygen [3254].

- Glucose fermenters. Able to ferment glucose producing acid and gas [3255].

Carbapenems, such as imipenem, meropenem, and ertapenem, are broad-spectrum antibiotics used as a last line of defense against serious bacterial infections. Resistance to carbapenems in Enterobacteriaceae is particularly concerning due to the limited therapeutic options available [3256].

The main mechanisms of resistance in these bacteria are:

i) Production of Carbapenemases. Carbapenemases can hydrolyze or inactivate carbapenems, allowing the bacteria to survive and multiply even in the presence of these antibiotics [3257].Some of the most relevant carbapenemases are:

- Klebsiella pneumoniae carbapenemase (KPC). One of the most common and globally disseminated carbapenemases [3258].

- New Delhi metallo-β-lactamase (NDM). Confers resistance to almost all β-lactams [3259].

- OXA-48. Found mainly in *K. pneumoniae* and other Enterobacteriaceae [3260].

- Verona integron-encoded metallo-β-lactamase (VIM) [3261] and **Imipenemase metallo-β-lactamase (IMP) [3262].** Metallo-β-lactamases that confer high resistance to carbapenems.

ii) Modification of Porins. Reduction in the expression or mutation of porins that limit the entry of antibiotics into the bacterial cell [3263].

iii) Efflux Systems. Increase in the expression of efflux pumps that expel antibiotics out of the cell [3264].

iv) Alterations in Penicillin-Binding Proteins (PBPs). Mutations in PBPs that decrease the affinity of β-lactam antibiotics [3265].

Carbapenem-resistant Enterobacteriaceae have significant clinical and public health implications [3266]:

- Hospital-acquired Infections. They are a leading cause of nosocomial infections, including urinary tract infections, pneumonia, bacteremia, and sepsis.

- Increased Morbidity and Mortality. Associated with high rates of morbidity and mortality due to limited effective treatment options.

- Complicated Treatment. Require the use of alternative antibiotics and therapeutic combinations that are often less effective and more toxic.

- Rapid Transmission and Spread. Capable of spreading rapidly in hospital and community settings, facilitating the transmission of resistance genes to other bacteria.

To address carbapenem-resistant Enterobacteriaceae, several strategies are implemented:

i) Infection Prevention and Control [3267]:

- Hand Hygiene. Strict hand hygiene practices to prevent transmission.

- Patient Isolation. Use of contact precautions and isolation for infected or colonized patients.

- Environmental Disinfection. Regular cleaning and disinfection of surfaces and medical equipment.

ii) Antibiotic Treatment:

- Use of Reserve Antibiotics. Use of agents such as colistin, tigecycline, and fosfomycin, although these may have significant side effects [3268].

- Antibiotic Combinations. Use of therapeutic combinations to enhance effectiveness and reduce the risk of further resistance [3269].

iii) Surveillance and Monitoring [3270]:

- Surveillance Programs. Continuous monitoring of the incidence of carbapenem resistance in hospital and community settings.

- Susceptibility Testing. Conducting antimicrobial susceptibility testing to guide appropriate therapy.

iv) Research and Development:

- New Antibiotics. Research and development of new antimicrobial agents with activity against carbapenem-resistant Enterobacteriaceae [3271].

- β-lactamase inhibitors. Development of new β-lactamase inhibitors that can be used in combination with β-lactams [3272].

Carbapenem resistance in Enterobacteriaceae has a significant impact on public health due to [3273, 3274]:

- Increased mortality. Resistant infections are associated with higher mortality rates.

- Prolonged hospital stays. Patients with infections caused by these bacteria often have longer hospital stays.

- High healthcare costs. Treating these infections and implementing infection control measures represent a considerable economic burden.

16.7 Bio-Mathematical Correlation between the HIV-1 Genome and the Host Genome

Currently, HIV-1 is considered a chronic disease, as current Antiretroviral Treatments (ART) have successfully controlled the progression to AIDS. However, these treatments have a range of long-term effects [3275]. Some of them include:

i) Organ Toxicity. Some antiretroviral medications may cause long-term toxicity in various organs and systems of the body, such as the liver, kidneys, heart, and bones. For example, hepatic toxicity can lead to chronic liver disease, while renal toxicity may result in chronic kidney failure [3276, 3277].

ii) Metabolic Alterations. ART can cause long-term metabolic changes, such as insulin resistance, dyslipidemia (elevated levels of cholesterol and triglycerides), and redistribution of body fat (lipodystrophy). These changes may increase the risk of cardiovascular diseases, diabetes, and metabolic disorders [3278, 3279].

iii) Cardiovascular Complications. Some studies have suggested that certain antiretroviral medications may increase the risk of long-term cardiovascular diseases, such as coronary artery disease and strokes. These effects may be related to dyslipidemia and other metabolic changes induced by ART [3280].

iv) Effects on the Central Nervous System. Some antiretroviral medications can cross the blood-brain barrier and affect the central nervous system. This can lead to neurological symptoms such as peripheral neuropathy, sleep disorders, depression, and anxiety, which may persist in the long term [3281, 3282].

v) Drug Resistance. Prolonged use of ART can increase the risk of developing resistance to antiretroviral drugs, which can limit effective treatment options in the future [107].

Many lines of HIV-1 research are aimed at determining whether it is possible to optimize medication strategies to delay and/or reduce the long-term effects of ARTs.

Given that several cases have been documented where certain individuals have successfully eliminated the viral load from their bodies [3283–3285], there may be genetic factors that facilitate the elimination of the virus from the host organism, or that promote autonomous disease control. If we were able to identify these profiles, personalized treatments could be designed to minimize the long-term impact of ARTs on organisms that may not require such heavy medication burdens. This could represent a significant advancement in precision medicine.

This line of research would have a dual approach. On the one hand, I will investigate whether the disease control demonstrated by certain individuals depends

on subtype, strain, or specific sequences of HIV-1. Therefore, in this aspect, my goal will be to apply my methodology to identify genomic components in HIV-1 sequences that promote disease control in certain individuals.

A second, more complex aspect, will involve the identification, using Deep Learning tools, of genomic components in the individuals' own genome that may influence this autonomous control. Undoubtedly, this research aspect will be extremely complex for various reasons, including the size of the human genome. However, as always, it will be exciting and may provide some additional information that could save lives and improve them.

16.8 Deepening the Mathematical Signatures Characterizing the Phylogeny among Different Pathogens

In this investigation, I detected a clear homology between the degree of similarity of hot zones regarding genetic recombination in HIV-1 and phylogenetic proximity. Obviously, this phenomenon requires further research.

Throughout the conducted research, I found a case study that may be particularly interesting and on which I would like to focus the investigation:

Analysis of similarity between the genome of the Omicron variant of SARS-CoV-2 and the Murine Hepatitis Virus (MHV), and characterization of their possible interactions with adenoviruses causing acute gastroenteritis, including the possibility of horizontal transfer of genetic information between both pathogens.

Why this case study? History repeats itself...

During the first half of 2022, several cases of acute childhood hepatitis of unknown origin emerged [3286]. The coincidence with waves of Omicron contagion led to the consideration of the possibility of some type of relationship between this variant and the unknown childhood hepatitis.

From May 2022 to September 2022, a total of 1,296 cases of severe acute hepatitis of unknown origin were reported in children across 37 countries. Among them, 449 cases were identified in 21 EU countries. 55 children required liver

transplantation, and 29 died. The affected children ranged from 1 month to 10 years of age. Adenovirus was detected in 62.2% of the analyzed samples [3287].

Adenoviruses (belonging to the *Adenoviridae* family) are double-stranded DNA viruses, non-enveloped, and approximately 70-90 nm in size. They are transmitted through direct contact, fecal-oral route, respiratory droplets, and contaminated surfaces [3288].

Although the cause of these cases remains unknown and there are multiple hypotheses, hepatitis viruses A to E and COVID-19 vaccines have been ruled out [3289, 3290]. A possible hypothesis suggests that these cases of severe hepatitis could be the result of an adenovirus infection in the intestines of previously healthy children who were infected with SARS-CoV-2 [3291]. This is based on the observation that many affected children had recently had an infection with SARS-CoV-2, which could have altered their immune response in a way that made them more susceptible to a severe second infection by adenovirus [3292].

So far, no other clear epidemiological risk factors have been identified [3293].Therefore, the etiology of the current cases of hepatitis remains under active investigation.

It happened again. Just like in the case of my research idea linking SARS-CoV-2 and FCoV, I related Omicron and that childhood hepatitis to Murine Hepatitis Virus (MHV).

MHV [3294] is a positive-sense single-stranded RNA virus belonging to the family *Coronaviridae*, genus Betacoronavirus [3295]. MHV primarily infects mice and causes a variety of diseases ranging from asymptomatic infections to respiratory, gastrointestinal, and neurological diseases. The pathogenesis of MHV depends on various factors, including the specific viral strain, the age and immune status of the host, and the route of infection [3296, 3297].

In mice, MHV can cause diseases similar to viral hepatitis, encephalitis, and acute or chronic respiratory diseases, depending on the viral strain and animal model used [3298]. Although MHV is an animal pathogen, it shares structural and genetic similarities with human coronaviruses, including Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV), Middle East Respiratory Syndrome

Coronavirus (MERS-CoV), and Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) [3299, 3300].

Of significant interest, however, is the fact that MHV infects the lymphatic tissue and therefore has significant and prolonged effects on the immune system, even in immunocompentent mice [3301]. Co-infection with MHV can modulate the course of several viral, bacterial, and parasitic infections, resulting in a more severe infection, resistance to infection, or a more or less severe course of infection, depending on the particular agent [3302]. MHV may also modify the activity of hepatic enzymes, reduce liver regeneration after partial hepatectomy, produce anaemia, leukopenia, and thrombocytopenia, and decrease the incidence of diabetes in some susceptible strains of mice [3303].

MHV exhibits two types of tropism during initial infection. Respiratory strains, also known as poli-tropic strains, are uncommon and replicate initially in the nasal mucosa, later disseminating to other organs through the bloodstream. The typical lesion in MHV disseminated respiratory infection is white foci found in the liver. Histopathological analysis reveals focal necrotizing hepatitis with syncytial cells. These lesions can also be found in lymphoid organs. Poli-tropic strains of MHV are those that frequently contaminate cell lines, tumors, and other materials of murine origin. Infection of neonates with no prior exposure to poli-tropic strains of MHV can lead to disseminated infection [3304, 3305].

Most strains of MHV found in animal facilities are of the enterotropic type. These strains primarily target the intestine and are excreted through feces. Due to this high fecal excretion level, enterotropic strains tend to spread rapidly from one mouse to another, leading to significant environmental contamination. In an infection with enterotropic MHV, lesions are primarily confined to the intestine in mice with a competent immune system. Various levels of epithelial cell lysis and villous atrophy in the terminal small intestine can be observed. Additionally, giant multinucleated syncytial cells may be found in the crypts and intestinal villi. Similar lesions may also be identified in the cecum and ascending colon [3306–3310].

Horizontal transfer of genetic information between two viruses is a process by which viruses exchange genetic material with each other [3311]. Horizontal gene transfer between viruses can occur in several ways:

i) Transduction. In this process, a virus acts as a vector to transfer genetic material from one virus to another. During viral replication, the host virus's DNA or RNA can be mistakenly packaged along with the viral genome and then be transferred to a new host cell, where it can integrate into a recipient virus's genome [3312, 3313].

ii) Viral Conjugation. Some viruses, especially bacteriophages (viruses that infect bacteria), can transfer genetic material directly from one cell to another through a process similar to bacterial conjugation [3314, 3315].

iii) Gene Capture. Viruses can acquire genes from their hosts by capturing genes from infected cells. These genes can provide new functions to the virus, such as resistance to antibiotics or antivirals, the ability to infect new types of cells, or adaptation to new environments [3316, 3317].

Horizontal transfer of genetic information among viruses is important for several reasons [3318, 3319]:

i) Viral Evolution. Horizontal gene transfer among viruses can lead to the emergence of new viral strains with different characteristics, such as increased virulence, resistance to antivirals, or the ability to infect new types of cells or hosts [3320].

ii) Adaptation to New Environments. Gene capture from hosts can allow viruses to adapt to new environments or environmental conditions, enabling them to survive and replicate more effectively [3321].

iii) Genetic Diversification. Horizontal gene transfer can increase genetic diversity within a viral population, which may enhance the viruses' ability to evade the host's immune system and persist in the population long-term [3322, 3323].

And, with all of this, I identified an interesting case study that can provide us with a lot of information to be prepared for future viral evolutions or new viruses.

I end as I began, happy and looking forward to continuing working.

Appendices

To know is to remember.

— Aristotle, *Plato used to refer to him as "the reader" and "the mind of the school".*

A

Compendium of Subtype F1 CNN Classification Results and Grad-CAM Hot Zones, in Alphanumeric Order

A.1 Subtype F1 Sequences

Table A.1: Subtype F1 CNN Classification Results and Interpretability Patterns. This table includes both the CNN classification results and the hot zones (concatenated and superposed spectrograms) of the 82 complete HIV-1 sequences within subtype F1, listed in alphanumeric order. "Sequence ID" corresponds to the unique identifier of each sequence. "Confidence Score" indicates the level of confidence demonstrated by the CNN in its non-recombinant classification. Values above 50% indicate that the classification is correct (non-recombinant), and values below 50% indicate a failure of the CNN in classification, that is, the CNN classified the sequence as recombinant. High values indicate a high degree of certainty by the CNN in its diagnosis. "Grad-CAM" provides the heat map resulting from the application of Grad-CAM.

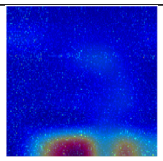
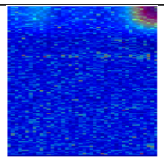
Non-Recombinant Category				
	Concatenated		**Superposed**	
Sequence ID	**Confidence Score**	**Grad-CAM**	**Confidence Score**	**Grad-CAM**
AB485656	0.97126138		0.86337656	

Table A.1 – Continuation from the previous page

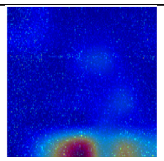
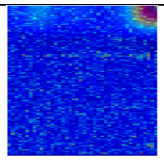
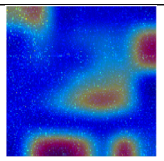
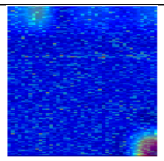
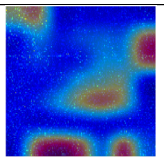
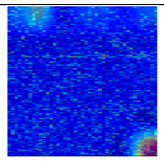
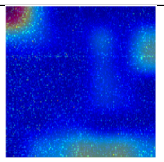
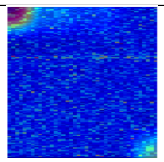
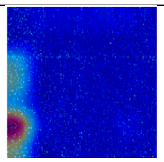
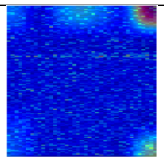
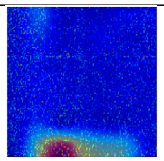
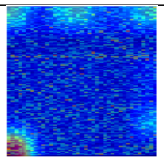
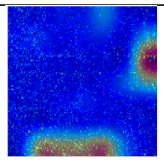
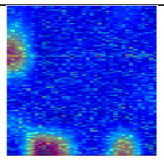
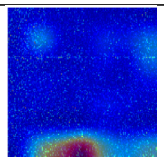
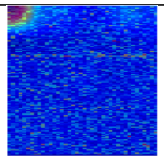
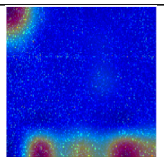
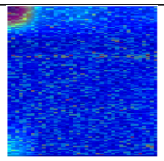
Non-Recombinant Category				
	Concatenated		**Superposed**	
Sequence ID	**Confidence Score**	**Grad-CAM**	**Confidence Score**	**Grad-CAM**
AB485657	0.97564381		0.71860975	
AB485658	0.99716210		0.97130698	
AB485659	0.99716210		0.97130698	
AF005494	0.98673069		0.99218082	
AF075703	0.47632891		0.92978340	
AF077336	0.62676984		0.53128004	
AJ249238	0.93734807		0.21464163	
AY173957	0.98636252		0.99943846	
AY173958	0.66884786		0.99181050	

Table A.1 – Continuation from the previous page

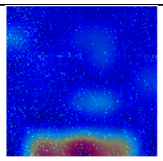
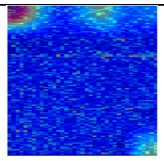
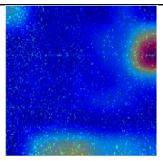
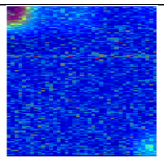
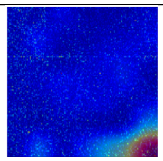
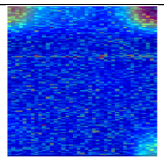
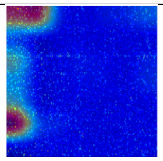
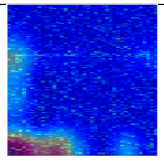
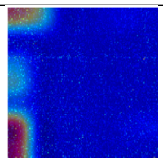
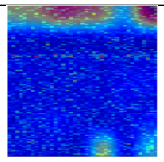
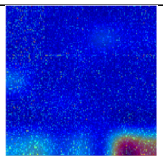
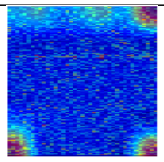
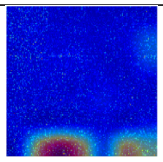
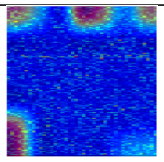
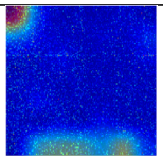
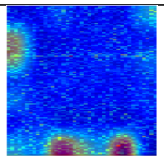
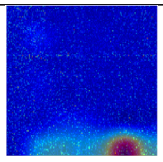
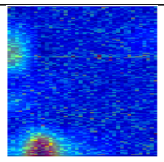
Non-Recombinant Category				
	Concatenated		**Superposed**	
Sequence ID	**Confidence Score**	**Grad-CAM**	**Confidence Score**	**Grad-CAM**
DQ189088	0.91134757		0.80327213	
DQ979023	0.84386384		0.83786082	
DQ979024	0.94219297		0.87136734	
DQ979025	0.34399492		0.20889668	
EU446022	0.35458723		0.56270295	
FJ670516	0.69824451		0.87515742	
FJ771006	0.53531951		0.84371614	
FJ771008	0.92909813		0.40365908	
FJ771009	0.59062725		0.39664441	

Table A.1 – Continuation from the previous page

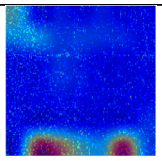
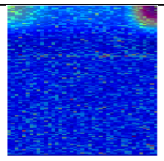
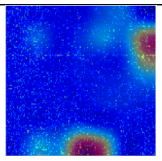
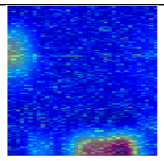
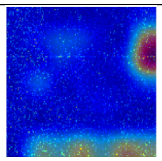
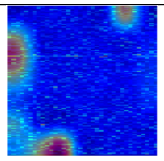
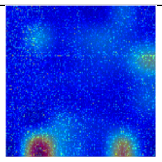
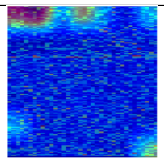
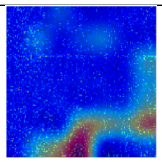
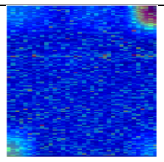
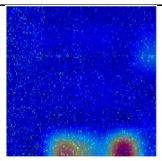
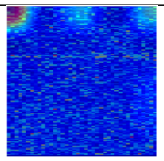
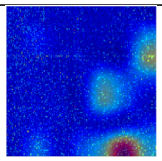
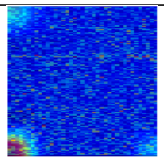
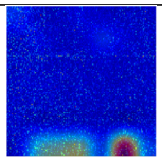
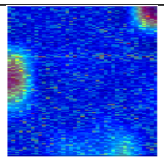
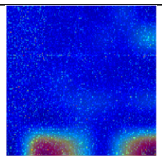
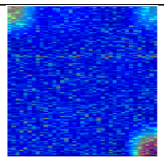
Non-Recombinant Category				
	Concatenated		**Superposed**	
Sequence ID	**Confidence Score**	**Grad-CAM**	**Confidence Score**	**Grad-CAM**
FJ771010	0.60734957		0.88615108	
FJ900266	0.88960779		0.38207427	
FJ900267	0.52454239		0.21767899	
FJ900268	0.95488888		0.93622065	
JF683771	0.98674184		0.80479991	
JX140671	0.99903536		0.99953818	
KC113009	0.99790549		0.97341305	
KJ849782	0.99814367		0.34296542	
KJ849791	0.99428540		0.98963320	

Table A.1 – Continuation from the previous page

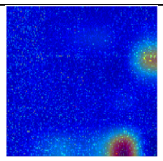
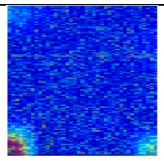
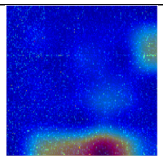
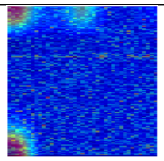
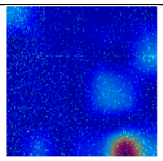
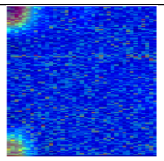
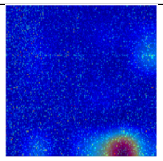
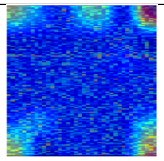
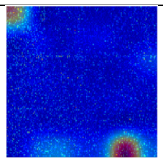
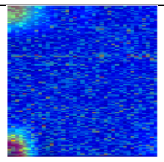
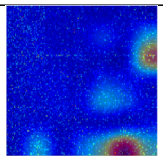
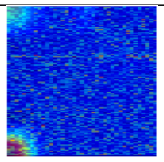
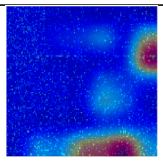
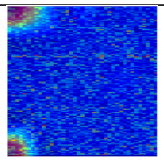
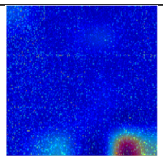
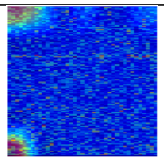
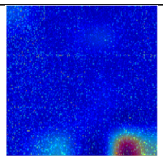
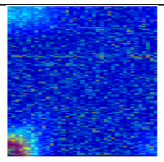
Non-Recombinant Category				
	Concatenated		**Superposed**	
Sequence ID	**Confidence Score**	**Grad-CAM**	**Confidence Score**	**Grad-CAM**
KJ883137	0.99396616		0.99370015	
KJ883138	0.63806075		0.96232021	
KJ883139	0.99662578		0.99578285	
KJ883140	0.96537137		0.89492697	
KJ883141	0.97730917		0.99214596	
KJ883142	0.98620671		0.93870032	
KJ883143	0.99202955		0.99570090	
KJ883144	0.99928987		0.98756963	
KJ883145	0.96053547		0.56175154	

Table A.1 – Continuation from the previous page

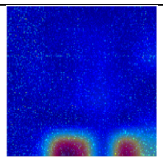
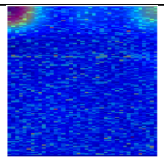
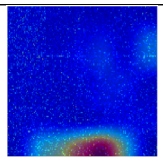
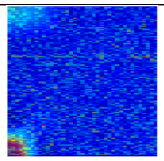
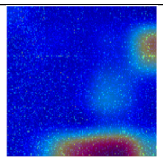
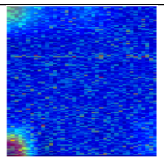
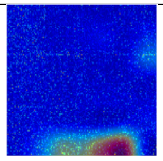
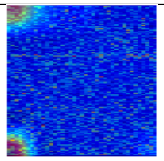
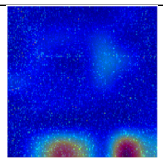
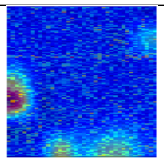
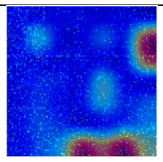
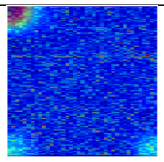
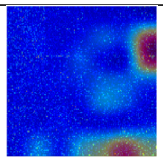
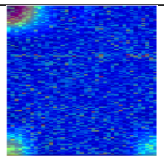
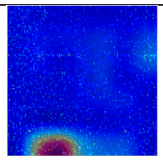
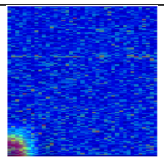
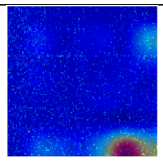
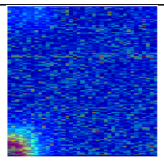
Non-Recombinant Category				
	Concatenated		**Superposed**	
Sequence ID	**Confidence Score**	**Grad-CAM**	**Confidence Score**	**Grad-CAM**
KJ883146	0.97278959		0.99960250	
KJ883147	0.75896615		0.95498258	
KJ883148	0.99589890		0.92667770	
KJ883149	0.98480254		0.99303508	
KJ883150	0.65480053		0.49418384	
KJ883151	0.98516744		0.99717110	
KJ883152	0.96697491		0.99047381	
KT276257	0.98942530		0.97901309	
KT276260	0.67440534		0.69034636	

Table A.1 – Continuation from the previous page

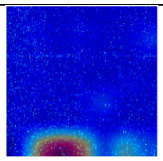
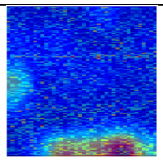
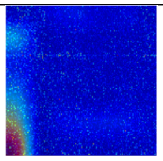
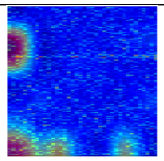
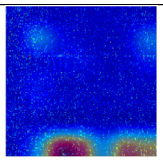
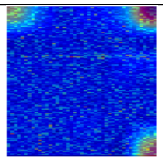
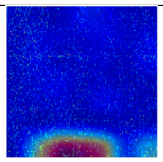
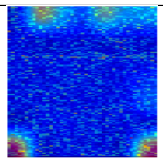
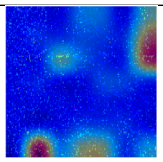
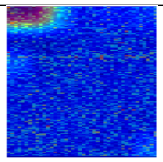
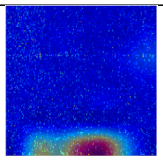
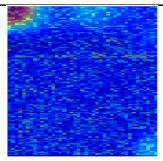
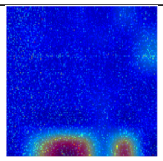
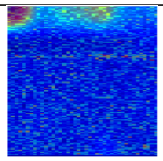
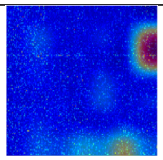
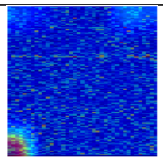
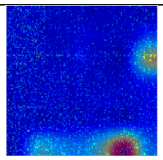
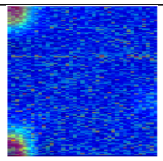
Non-Recombinant Category				
	Concatenated		**Superposed**	
Sequence ID	**Confidence Score**	**Grad-CAM**	**Confidence Score**	**Grad-CAM**
KT427774	0.77205658		0.31397808	
KT427814	0.39004901		0.012349190	
KU168274	0.94078940		0.98875976	
KU168275	0.97266847		0.98760313	
KU168276	0.65043950		0.66593117	
KU749395	0.72366655		0.79233700	
KU749396	0.98291439		0.61182326	
KY514084	0.99781597		0.99184263	
KY514085	0.99957353		0.99274033	

Table A.1 – Continuation from the previous page

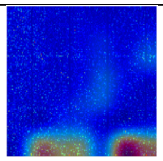
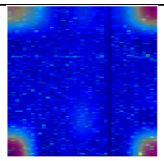
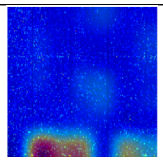
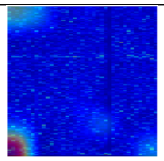
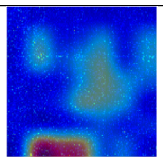
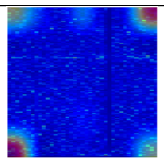
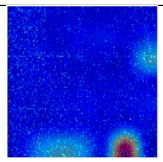
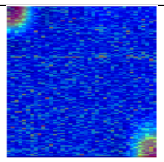
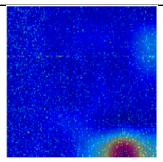
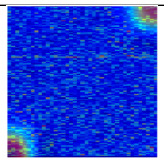
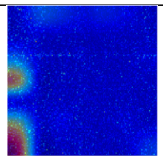
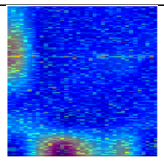
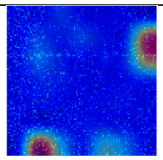
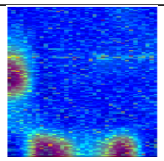
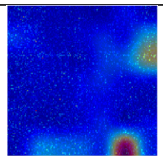
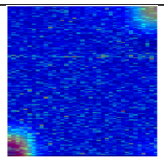
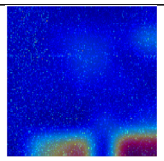
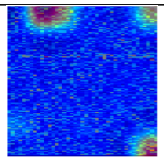
Non-Recombinant Category				
	Concatenated		**Superposed**	
Sequence ID	**Confidence Score**	**Grad-CAM**	**Confidence Score**	**Grad-CAM**
KY639271	0.88291824		0.79526740	
KY639272	0.89017004		0.97767538	
KY639273	0.81883091		0.93598765	
KY989955	0.99525923		0.99510574	
KY989957	0.99448121		0.98087388	
MF109434	0.48077449		0.36552623	
MF109526	0.85129678		0.29212782	
MF109640	0.80146289		0.65761119	
MG365762	0.64324737		0.97378516	

Table A.1 – Continuation from the previous page

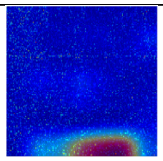
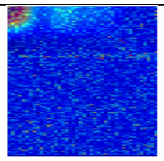
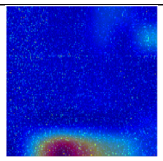
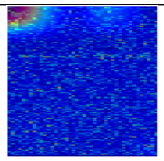
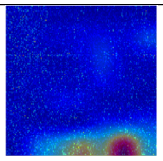
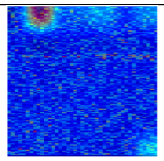
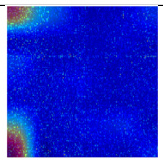
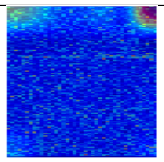
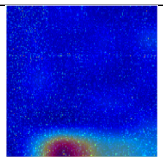
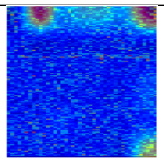
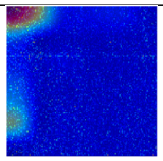
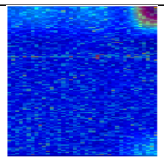
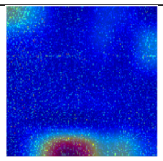
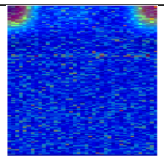
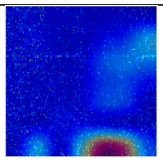
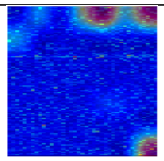
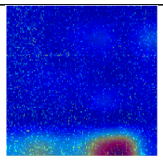
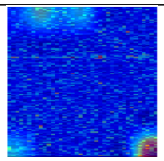
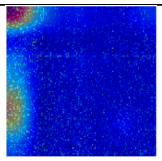
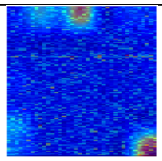
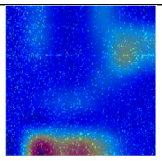
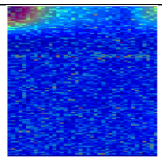
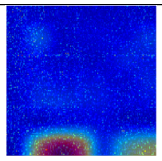
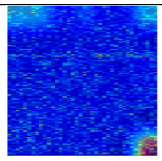
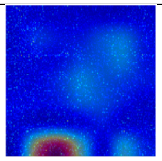
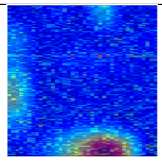
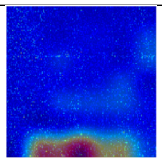
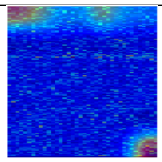
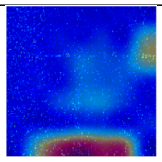
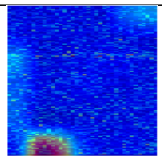
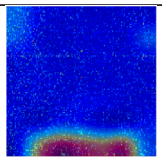
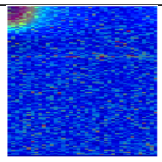
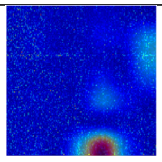
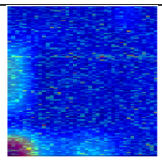
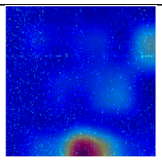
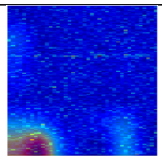
Non-Recombinant Category				
	Concatenated		**Superposed**	
Sequence ID	**Confidence Score**	**Grad-CAM**	**Confidence Score**	**Grad-CAM**
MG365763	0.83651102		0.77941793	
MG365764	0.99214339		0.98222381	
MG365766	0.61434537		0.80349880	
MG365767	0.14730887		0.61304283	
MG365768	0.54682845		0.61507100	
MH746263	0.053940132		0.97540295	
MK041565	0.98091447		0.93857902	
MK177827	0.95731449		0.53774112	
MN485989	0.67678934		0.62142700	

Table A.1 – Continuation from the previous page

Non-Recombinant Category				
	Concatenated		**Superposed**	
Sequence ID	**Confidence Score**	**Grad-CAM**	**Confidence Score**	**Grad-CAM**
MN486030	0.43544948		0.99667943	
MT222953	0.80201364		0.84729683	
MT417747	0.76550716		0.88939387	
MT417748	0.99254322		0.14745644	
MT417762	0.99634862		0.68152523	
MT417763	0.98270667		0.29784909	
MT417764	0.94145989		0.95977890	
MT417765	0.98940897		0.18301243	
MT417769	0.91214973		0.015416335	

I have not failed. I've just found 10,000 ways that won't work.

— Thomas Alva Edison, *three months after he started school, his teacher labeled him as a sterile and unproductive student. His mother Nancy, a very educated woman, decided to withdraw him from school and educated him at home.*

B

Compendium of Subtype CRF22_01A1 CNN Classification Results and Grad-CAM Hot Zones, in Alphanumeric Order

B.1 Subtype CRF22_01A1 Sequences

Table B.1: Subtype CRF22_01A1 CNN Classification Results and Interpretability Patterns. This table includes both the CNN classification results and the hot zones (concatenated and superposed spectrograms) of the 21 complete HIV-1 sequences within subtype CRF22_01A1, listed in alphanumeric order. "Sequence ID" corresponds to the unique identifier of each sequence. "Confidence Score" indicates the level of confidence demonstrated by the CNN in its recombinant classification. Values above 50% indicate that the classification is correct (recombinant), and values below 50% indicate a failure of the CNN in classification, that is, the CNN classified the sequence as non-recombinant. High values indicate a high degree of certainty by the CNN in its diagnosis. "Grad-CAM" provides the heat map resulting from the application of Grad-CAM.

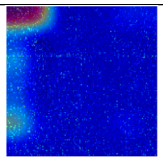
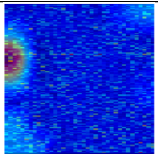
Recombinant Category				
	Concatenated		**Superposed**	
Sequence ID	**Confidence Score**	**Grad-CAM**	**Confidence Score**	**Grad-CAM**
AY371159	0.96656334		0.99985778	

Table B.1 – Continuation from the previous page

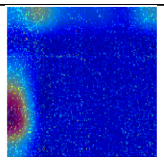
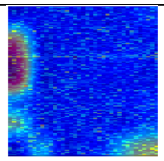
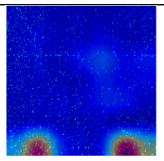
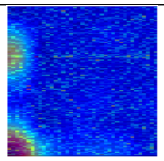
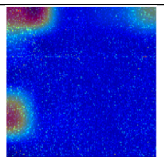
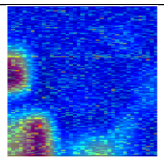
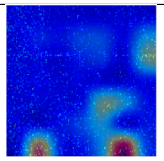
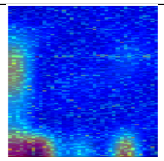
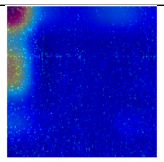
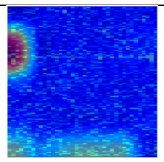
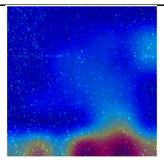
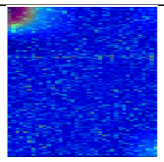
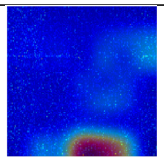
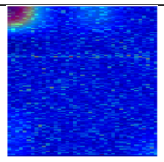
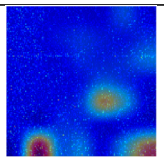
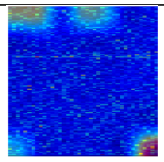
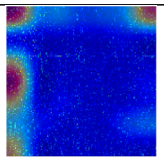
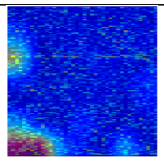
Recombinant Category				
	Concatenated		**Superposed**	
Sequence ID	**Confidence Score**	**Grad-CAM**	**Confidence Score**	**Grad-CAM**
AY371165	0.70579261		0.97759801	
EU743963	0.095893860		0.95634651	
GQ229529	0.99928892		0.86511111	
JN864049	0.19887774		0.63258851	
JN864050	0.94469112		0.82171077	
JN864051	0.099829063		0.49274105	
JN864058	0.071222097		0.081556499	
JN864059	0.26577246		0.39441884	
KF716460	0.99614245		0.99985349	

Table B.1 – Continuation from the previous page

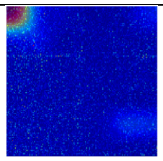
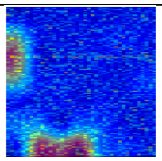
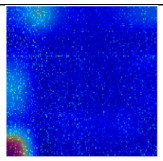
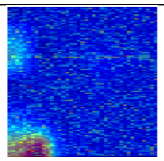
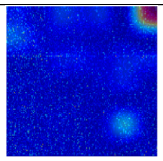
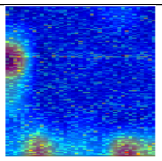
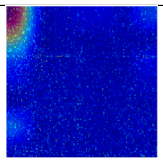
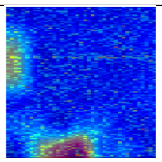
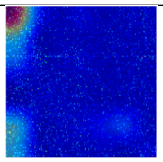
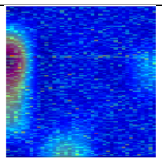
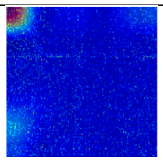
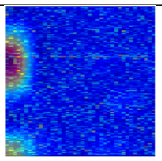
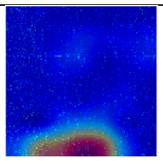
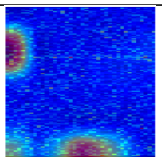
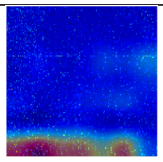
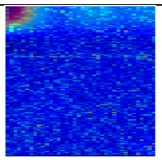
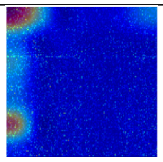
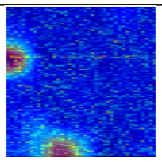
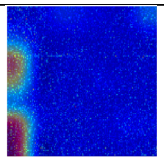
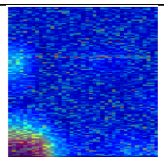
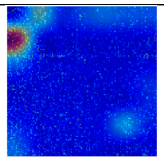
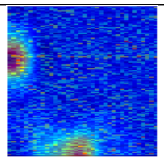
Recombinant Category				
	Concatenated		**Superposed**	
Sequence ID	**Confidence Score**	**Grad-CAM**	**Confidence Score**	**Grad-CAM**
KF716461	0.99901760		0.99980301	
KF716462	0.99236321		0.99981803	
KF716463	0.99983811		0.99392575	
KP109498	0.99646217		0.99925226	
KP109499	0.55124038		0.86139542	
KP109500	0.99889731		0.99980980	
KX228816	0.43857804		0.67217767	
KX228817	0.17623627		0.38101846	
MN153487	0.99250811		0.98742801	

Table B.1 – Continuation from the previous page

Recombinant Category				
	Concatenated		**Superposed**	
Sequence ID	**Confidence Score**	**Grad-CAM**	**Confidence Score**	**Grad-CAM**
MT349410	0.99835867		0.99716872	
MT349413	0.99722987		0.93800962	

The man with a new idea is a cranck until the idea succeeds.

— Mark Twain, *he was born in 1835 during the passage of Halley's Comet and predicted that he would die when the comet returned. On April 21, 1910, while Halley's Comet was passing close to Earth, Twain died of a heart attack.*

C

Compendium of Subtype O CNN Classification Results and Grad-CAM Hot Zones, in Alphanumeric Order

C.1 Subtype O Sequences

Table C.1: Subtype O CNN Classification Results and Interpretability Patterns. This table includes both the CNN classification results and the hot zones (concatenated and superposed spectrograms) of the 57 complete HIV-1 sequences within subtype O, listed in alphanumeric order. "Sequence ID" corresponds to the unique identifier of each sequence. "Confidence Score" indicates the level of confidence demonstrated by the CNN in its non-recombinant classification. Values above 50% indicate that the classification is correct (non-recombinant), and values below 50% indicate a failure of the CNN in classification, that is, the CNN classified the sequence as recombinant. High values indicate a high degree of certainty by the CNN in its diagnosis. "Grad-CAM" provides the heat map resulting from the application of Grad-CAM.

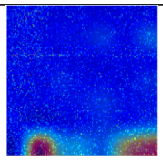
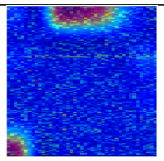
Non-Recombinant Category				
	Concatenated		**Superposed**	
Sequence ID	**Confidence Score**	**Grad-CAM**	**Confidence Score**	**Grad-CAM**
AB485666	0.99854183		0.99952483	

Table C.1 – Continuation from the previous page

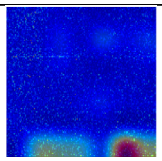
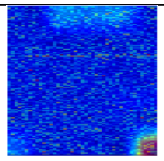
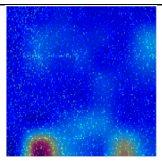
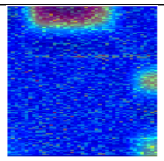
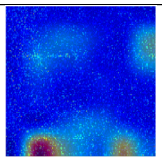
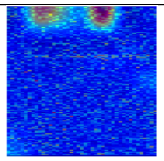
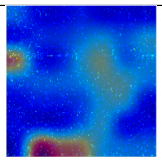
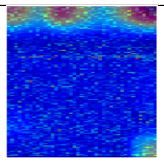
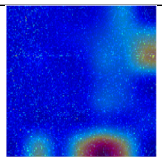
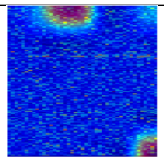
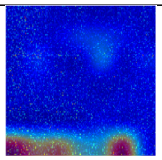
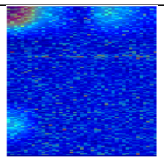
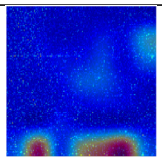
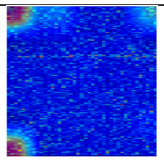
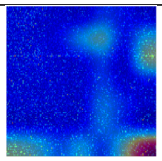
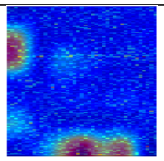
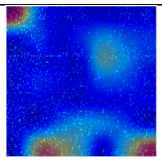
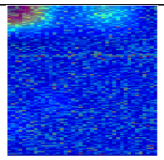
Non-Recombinant Category				
	Concatenated		**Superposed**	
Sequence ID	**Confidence Score**	**Grad-CAM**	**Confidence Score**	**Grad-CAM**
AB485667	0.99626046		0.99785477	
AB485668	0.94492567		0.95757324	
AB485669	0.96329194		0.95260042	
AF407418	0.78692532		0.98849803	
AJ302646	0.99786025		0.99822253	
AJ302647	0.99987161		0.99909902	
AY169802	0.99038130		0.99958569	
AY169803	0.98940021		0.49227378	
AY169804	0.98750579		0.90259385	

Table C.1 – Continuation from the previous page

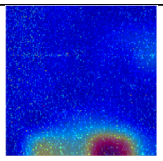
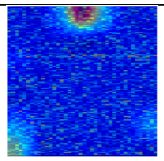
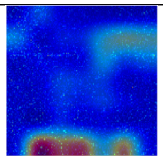
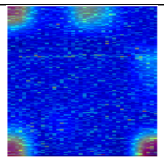
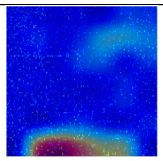
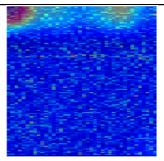
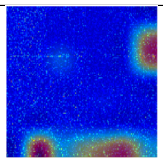
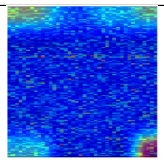
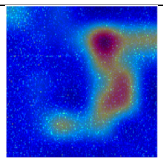
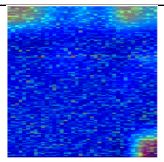
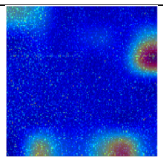
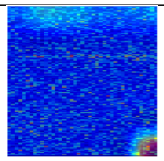
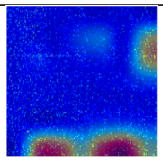
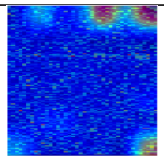
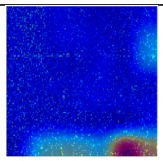
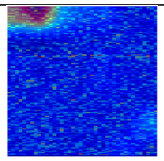
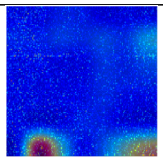
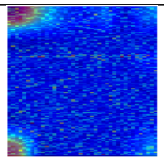
Non-Recombinant Category				
	Concatenated		**Superposed**	
Sequence ID	**Confidence Score**	**Grad-CAM**	**Confidence Score**	**Grad-CAM**
AY169805	0.96536005		0.99978799	
AY169806	0.80012530		0.99805754	
AY169807	0.98317564		0.99897480	
AY169808	0.99990940		0.94435722	
AY169809	0.99912959		0.99481046	
AY169810	0.99721354		0.99311793	
AY169811	0.95042008		0.98550040	
AY169812	0.98030341		0.98134768	
AY169813	0.96824318		0.99572456	

Table C.1 – Continuation from the previous page

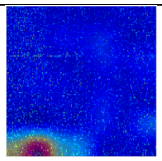
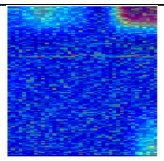
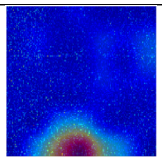
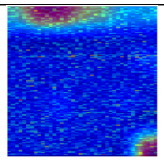
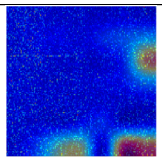
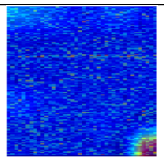
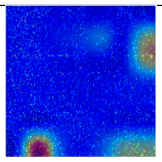
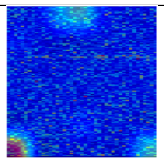
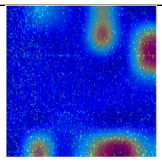
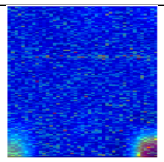
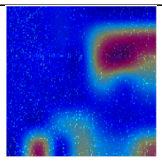
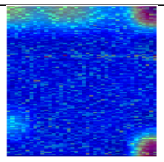
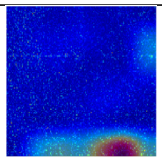
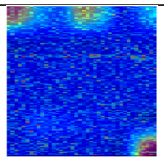
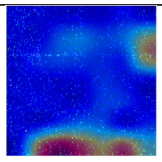
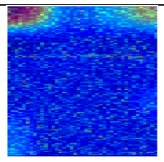
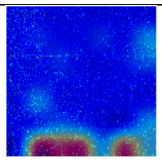
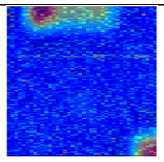
Non-Recombinant Category				
	Concatenated		**Superposed**	
Sequence ID	**Confidence Score**	**Grad-CAM**	**Confidence Score**	**Grad-CAM**
AY169814	0.99618173		0.99442297	
AY169815	0.99258476		0.99446338	
AY169816	0.99049020		0.99820375	
AY618998	0.97496450		0.99994183	
AY623602	0.99964368		0.99855655	
JN571034	0.99498367		0.99845481	
JX245014	0.91018569		0.99405020	
JX245015	0.98681462		0.99542987	
KF859742	0.84006888		0.98930490	

Table C.1 – Continuation from the previous page

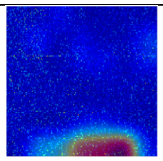
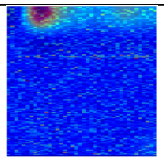
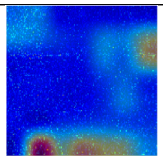
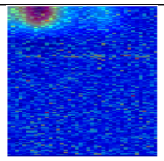
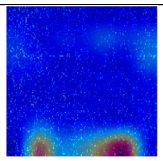
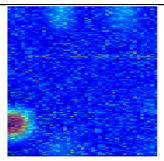
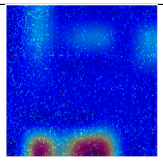
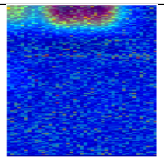
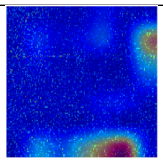
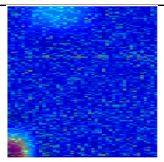
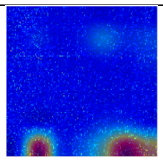
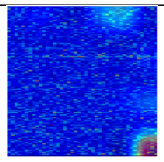
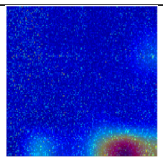
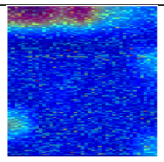
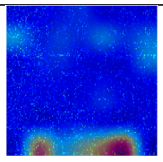
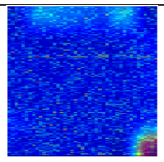
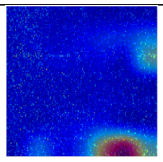
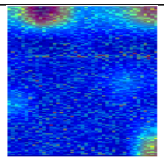
Non-Recombinant Category				
	Concatenated		**Superposed**	
Sequence ID	**Confidence Score**	**Grad-CAM**	**Confidence Score**	**Grad-CAM**
KF859743	0.99150473		0.97085339	
KF859744	0.94214135		0.98728174	
KU168281	0.91542292		0.99951732	
KU168282	0.99890673		0.97326827	
KU168283	0.99699128		0.96098906	
KU168284	0.91554320		0.99761105	
KU168285	0.98266107		0.99845552	
KU168286	0.95827639		0.99852991	
KU168288	0.99835664		0.99952126	

Table C.1 – Continuation from the previous page

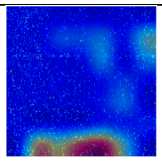
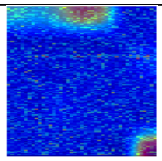
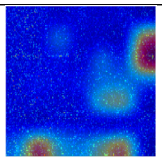
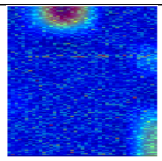
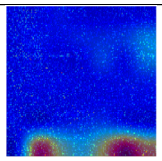
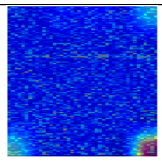
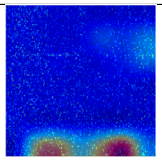
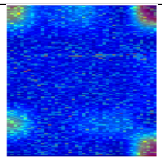
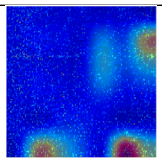
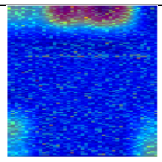
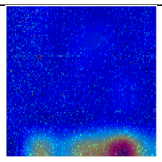
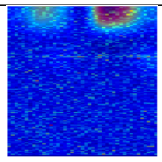
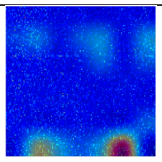
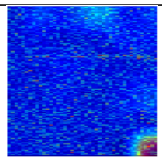
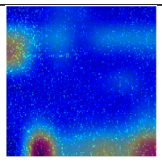
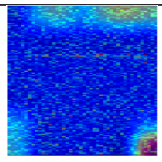
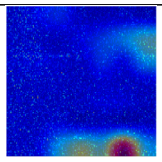
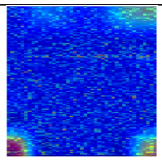
Non-Recombinant Category				
	Concatenated		**Superposed**	
Sequence ID	**Confidence Score**	**Grad-CAM**	**Confidence Score**	**Grad-CAM**
KU168289	0.97666639		0.99635351	
KU168290	0.99255329		0.99797899	
KU168291	0.97698820		0.99969530	
KU168292	0.96091431		0.99970585	
KU168293	0.88265246		0.98878479	
KU168294	0.99952984		0.99887353	
KU168295	0.78010982		0.99981850	
KU168296	0.95705414		0.98576689	
KU168297	0.99640274		0.99945492	

Table C.1 – Continuation from the previous page

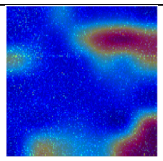
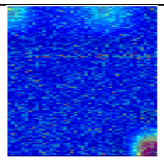
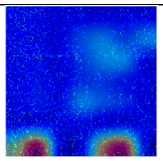
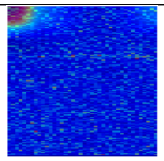
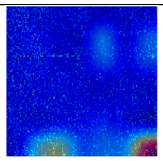
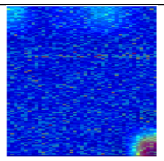
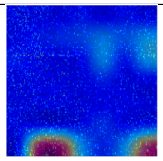
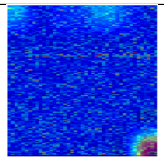
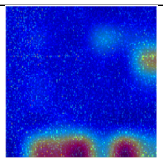
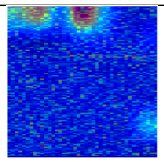
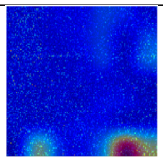
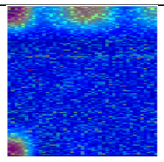
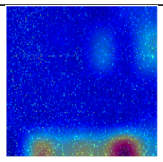
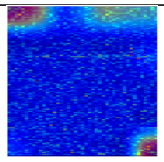
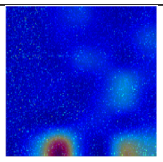
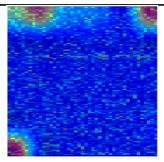
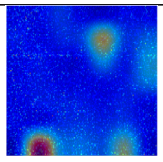
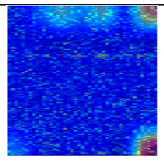
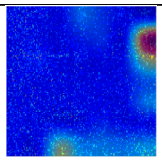
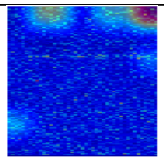
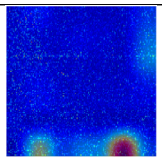
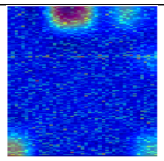
Non-Recombinant Category				
	Concatenated		**Superposed**	
Sequence ID	**Confidence Score**	**Grad-CAM**	**Confidence Score**	**Grad-CAM**
KU168298	0.96031851		0.99612576	
KX228804	0.98936754		0.94765824	
KY112585	0.99319106		0.99622190	
KY953205	0.99129784		0.99622190	
KY953206	0.95345253		0.99682850	
L20571	0.98114705		0.95797539	
L20587	0.98487145		0.97713041	
MH705147	0.98381597		0.95675308	
MH705148	0.95619524		0.94489014	

Table C.1 – Continuation from the previous page

Non-Recombinant Category				
	Concatenated		**Superposed**	
Sequence ID	**Confidence Score**	**Grad-CAM**	**Confidence Score**	**Grad-CAM**
MH705149	0.99949920		0.85092968	
MH705150	0.98717475		0.95876294	

A small truth is better than a great lie.

— Leonardo da Vinci, *designer of the viola organista, a musical instrument that uses rotating wheels to individually vibrate the strings thanks to a keyboard that achieves this movement by pressing its keys.*

D

Compendium of Subtype U Sequences and their Phylogenetically Related Sequences, in Alphanumeric Order

"Sequence ID" corresponds to the unique identifier of each sequence. "Classification" indicates the result of the classification performed by the CNN. "Confidence Score" indicates the level of confidence demonstrated by the CNN in its classification. High values indicate a high degree of certainty by the CNN in its diagnosis. "Grad-CAM" provides the heat map resulting from the application of Grad-CAM on the superposed genomic spectrogram of each sequence.

D.1 Subtype U Sequences

Table D.1: Subtype U Interpretability Patterns. This table includes both the CNN classification results and the hot zones (superposed spectrograms) of the 23 complete HIV-1 sequences within subtype U, listed in alphanumeric order.

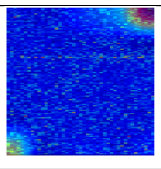
Sequence ID	Classification	Confidence Score	Grad-CAM
AB485664	Non-recombinant	90.331%	

Table D.1 – Continuation from the previous page

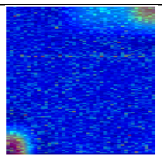
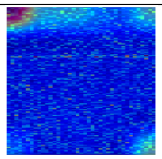
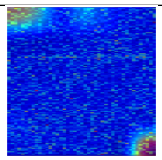
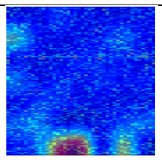
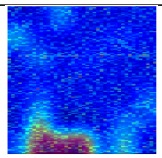
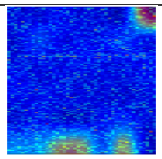
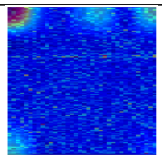
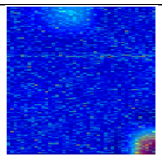
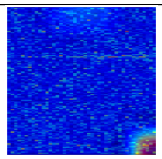
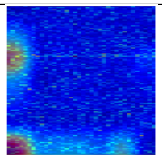
Sequence ID	Classification	Confidence Score	Grad-CAM
AB485665	Non-recombinant	92.995%	
AY046058	Non-recombinant	90.426%	
EF029066	Non-recombinant	51.424%	
EF029067	Recombinant	76.708%	
EF029068	Recombinant	69.344%	
EF029069	Recombinant	73.163%	
FJ388921	Non-recombinant	91.853%	
HM215249	Non-recombinant	99.553%	
HM215251	Non-recombinant	99.289%	
JF683772	Recombinant	75.854%	

Table D.1 – Continuation from the previous page

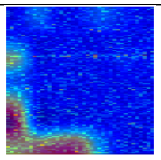
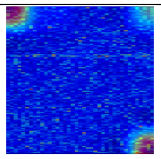
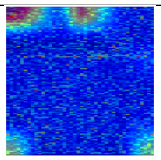
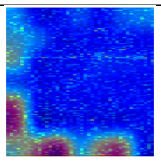
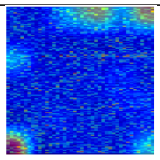
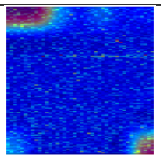
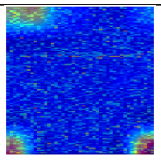
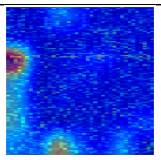
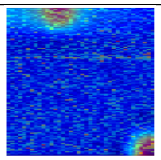
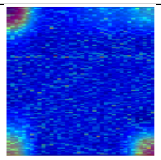
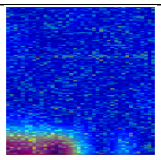
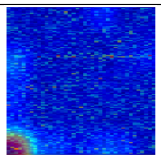
Sequence ID	**Classification**	**Confidence Score**	**Grad-CAM**
JX140679	Recombinant	77.945%	
KP718920	Non-recombinant	84.895%	
KY392773	Non-recombinant	69.660%	
KY392774	Recombinant	58.879%	
KY392775	Non-recombinant	89.344%	
KY392780	Non-recombinant	62.076%	
MH078559	Non-recombinant	91.760%	
MK095236	Recombinant	82.815%	
MN153489	Non-recombinant	93.586%	
MN153492	Non-recombinant	91.835%	

Table D.1 – Continuation from the previous page

Sequence ID	Classification	Confidence Score	Grad-CAM
MN736701	Recombinant	74.420%	
MN736702	Recombinant	52.857%	

D.2 Subtype A1 Sequences

Table D.2: Subtype A1 Interpretability Patterns. This table includes both the CNN classification results and the hot zones (superposed spectrograms) of the complete HIV-1 subtype A1 sequences mentioned in Chapter 8, listed in alphanumeric order.

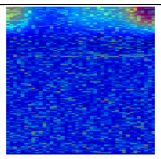
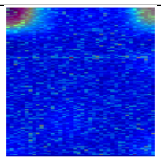
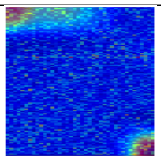
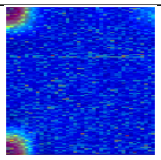
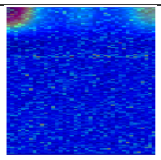
Sequence ID	Classification	Confidence Score	Grad-CAM
KX232594	Non-recombinant	88.062%	
KX232595	Non-recombinant	82.135%	
KX232597	Non-recombinant	99.701%	
KX232600	Non-recombinant	98.494%	
KX232607	Non-recombinant	93.285%	

Table D.2 – Continuation from the previous page

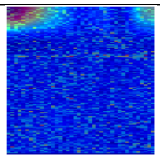
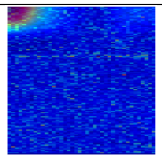
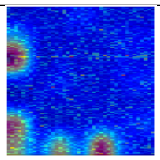
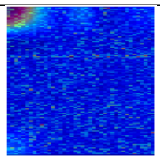
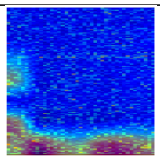
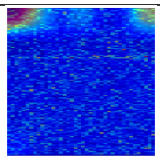
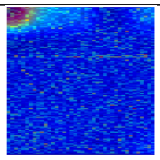
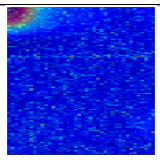
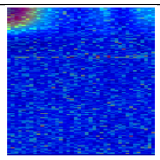
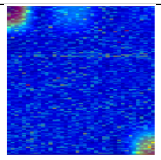
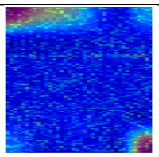
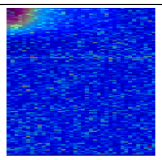
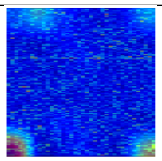
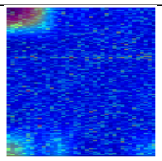
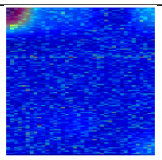
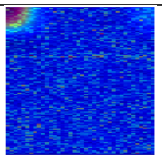
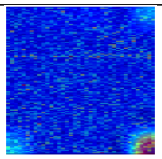
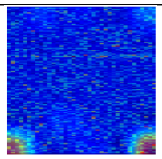
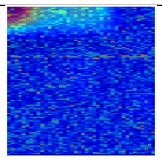
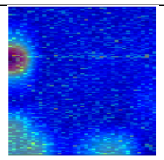
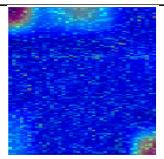
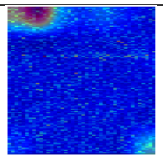
Sequence ID	**Classification**	**Confidence Score**	**Grad-CAM**
KX232609	Non-recombinant	0.90.666%	
KX232610	Non-recombinant	97.200%	
KX232611	Recombinant	96.440%	
KX232614	Non-recombinant	84.851%	
KX232618	Recombinant	69.297%	
KX232620	Non-recombinant	99.969%	
KX232624	Non-recombinant	99.239%	
KX232626	Non-recombinant	93.556%	
KU749411	Non-recombinant	87.714%	
MN153491	Non-recombinant	95.920%	

Table D.2 – Continuation from the previous page

Sequence ID	Classification	Confidence Score	Grad-CAM
MN791169	Non-recombinant	99.987%	
MN791209	Non-recombinant	99.703%	
MN791279	Non-recombinant	99.817%	
MN791333	Non-recombinant	99.864%	
MN791710	Non-recombinant	99.580%	
MN791853	Non-recombinant	99.926%	
MN791873	Non-recombinant	99.975%	
MN791964	Non-recombinant	99.852%	

D.3 Subtype A2 Sequences

Table D.3: Subtype A2 Interpretability Patterns. This table includes both the CNN classification results and the hot zones (superposed spectrograms) of the complete HIV-1 subtype A2 sequences mentioned in Chapter 8, listed in alphanumeric order.

Sequence ID	Classification	Confidence Score	Grad-CAM
AF286237	Non-recombinant	98.597%	
AF286238	Recombinant	99.801%	
GU201516	Non-recombinant	92.866%	
MH705163	Non-recombinant	53.498%	

D.4 Subtype A6 Sequences

Table D.4: Subtype A6 Interpretability Patterns. This table includes both the CNN classification results and the hot zones (superposed spectrograms) of the complete HIV-1 subtype A6 sequences mentioned in Chapter 8, listed in alphanumeric order.

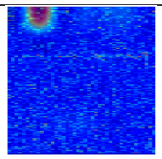
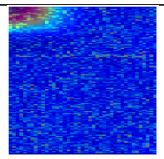
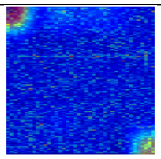
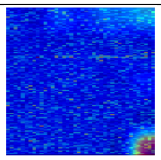
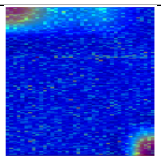
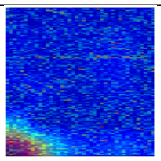
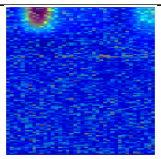
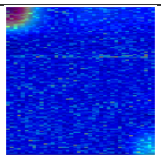
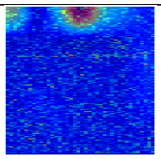
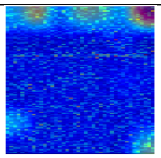
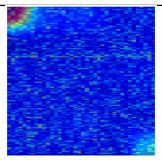
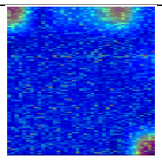
Sequence ID	Classification	Confidence Score	Grad-CAM
MT222946	Non-recombinant	99.926%	
MT222949	Non-recombinant	94.863%	

Table D.4 – Continuation from the previous page

Sequence ID	Classification	Confidence Score	Grad-CAM
MN703135	Non-recombinant	99.075%	
MN703136	Non-recombinant	91.241%	
MN703137	Non-recombinant	81.689%	
MN703138	Recombinant	56.614%	
MN703139	Non-recombinant	94.302%	
MN736699	Non-recombinant	94.463%	
MT222947	Non-recombinant	99.856%	
MT222948	Non-recombinant	99.002%	

D.5 Subtype A7 Sequences

Table D.5: Subtype A7 Interpretability Patterns. This table includes both the CNN classification results and the hot zones (superposed spectrograms) of the complete HIV-1 subtype A7 sequences mentioned in Chapter 8, listed in alphanumeric order.

Sequence ID	Classification	Confidence Score	Grad-CAM
KX389622	Non-recombinant	86.314%	
MH078558	Non-recombinant	84.715%	

D.6 Subtype F1 Sequences

Table D.6: Subtype F1 Interpretability Patterns. This table includes both the CNN classification results and the hot zones (superposed spectrograms) of the complete HIV-1 subtype F1 sequences mentioned in Chapter 8, listed in alphanumeric order.

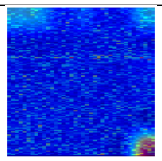
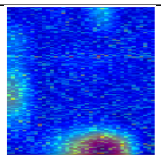
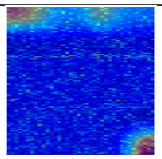
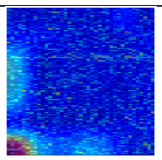
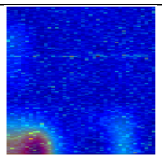
Sequence ID	Classification	Confidence Score	Grad-CAM
MT417747	Non-recombinant	88.939%	
MT417748	Recombinant	85.254%	
MT417762	Non-recombinant	68.153%	
MT417765	Recombinant	81.698%	

Table D.6 – Continuation from the previous page

Sequence ID	Classification	Confidence Score	Grad-CAM
MT417769	Recombinant	98.458%	

D.7 Subtype F2 Sequences

Table D.7: Subtype F2 Interpretability Patterns. This table includes both the CNN classification results and the hot zones (superposed spectrograms) of the complete HIV-1 subtype F2 sequences mentioned in Chapter 8, listed in alphanumeric order.

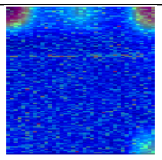
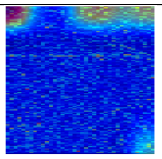
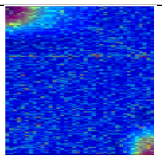
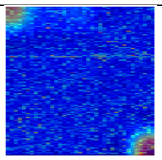
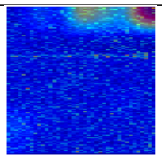
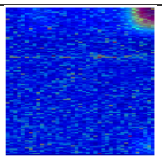
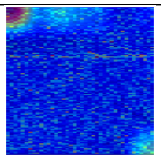
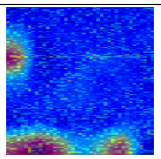
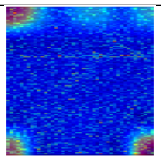
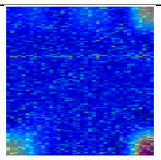
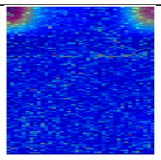
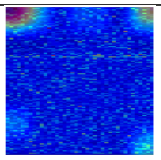
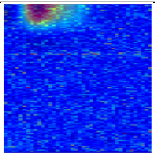
Sequence ID	Classification	Confidence Score	Grad-CAM
AF377956	Non-recombinant	99.293%	
AJ249236	Non-recombinant	97.015%	
AJ249237	Non-recombinant	98.966%	
AY371158	Non-recombinant	87.008%	
JX140672	Non-recombinant	96.758%	
JX140673	Non-recombinant	99.456%	

Table D.7 – Continuation from the previous page

Sequence ID	Classification	Confidence Score	Grad-CAM
KU749420	Non-recombinant	91.191%	
KU749422	Recombinant	60.523%	
MH705144	Non-recombinant	91.671%	
MK086131	Non-recombinant	90.423%	
MN153483	Non-recombinant	95.832%	
MN153485	Non-recombinant	84.152%	

D.8 Subtype G Sequences

Table D.8: Subtype G Interpretability Patterns. This table includes both the CNN classification results and the hot zones (superposed spectrograms) of the complete HIV-1 subtype G sequences mentioned in Chapter 8, listed in alphanumeric order.

Sequence ID	Classification	Confidence Score	Grad-CAM
MN153484	Non-recombinant	98.570%	

D.9 Subtype H Sequences

Table D.9: Subtype H Interpretability Patterns. This table includes both the CNN classification results and the hot zones (superposed spectrograms) of the complete HIV-1 subtype H sequences mentioned in Chapter 8, listed in alphanumeric order.

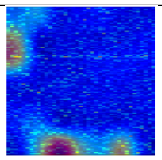
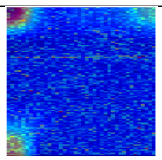
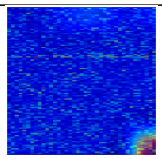
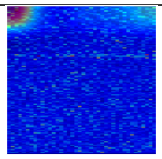
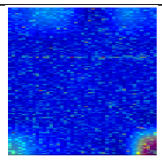
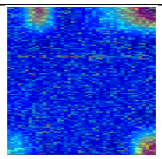
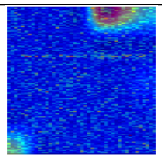
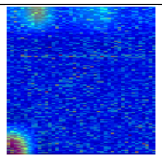
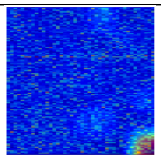
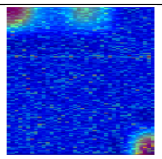
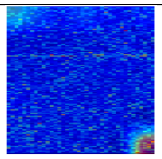
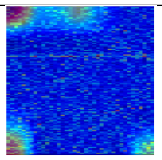
Sequence ID	Classification	Confidence Score	Grad-CAM
AF005496	Recombinant	77.936%	
AF190127	Non-recombinant	92.135%	
AF190128	Non-recombinant	98.441%	
FJ711703	Non-recombinant	98.114%	
KU168273	Non-recombinant	89.265%	
KU168279	Non-recombinant	93.215%	
KY392777	Non-recombinant	72.514%	
KY392778	Non-recombinant	84.282%	

Table D.9 – Continuation from the previous page

Sequence ID	Classification	Confidence Score	Grad-CAM
KY392779	Non-recombinant	99.995%	
MN486046	Non-recombinant	89.830%	

D.10 Subtype K Sequences

Table D.10: Subtype K Interpretability Patterns. This table includes both the CNN classification results and the hot zones (superposed spectrograms) of the complete HIV-1 subtype K sequences mentioned in Chapter 8, listed in alphanumeric order.

Sequence ID	Classification	Confidence Score	Grad-CAM
AJ249235	Non-recombinant	97.356%	
AJ249239	Non-recombinant	99.312%	

D.11 Subtype CRF02_AG Sequences

Table D.11: Subtype CRF02_AG Interpretability Patterns. This table includes both the CNN classification results and the hot zones (superposed spectrograms) of the complete HIV-1 subtype CRF02_AG sequences mentioned in Chapter 8, listed in alphanumeric order.

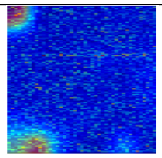
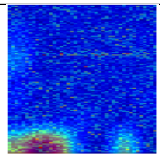
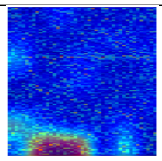
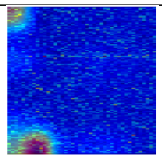
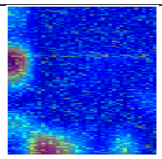
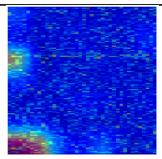
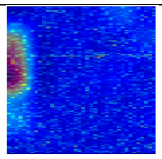
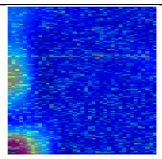
Sequence ID	Classification	Confidence Score	Grad-CAM
AB049811	Recombinant	99.952%	
AB286857	Recombinant	99.732%	
AB286862	Recombinant	99.905%	
AB485636	Recombinant	99.579%	
AF063223	Recombinant	99.830%	
AF107770	Recombinant	99.934%	
AF377955	Recombinant	99.984%	
AJ251057	Recombinant	99.999%	

Table D.11 – Continuation from the previous page

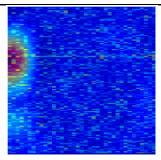
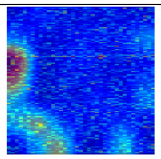
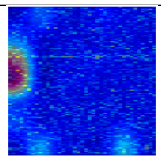
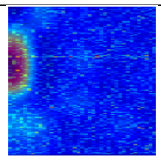
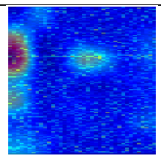
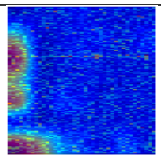
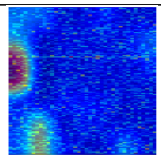
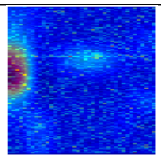
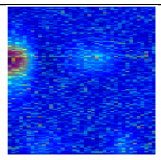
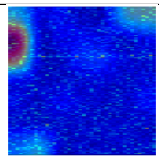
Sequence ID	**Classification**	**Confidence Score**	**Grad-CAM**
AY371122	Recombinant	99.966%	
AY371140	Recombinant	99.737%	
DQ168577	Recombinant	96.880%	
DQ168578	Recombinant	99.996%	
GU201494	Recombinant	99.984%	
GU201514	Recombinant	99.978%	
JN248580	Recombinant	99.992%	
JN248589	Recombinant	99.997%	
JN248590	Recombinant	99.968%	
KF561437	Recombinant	99.599%	

Table D.11 – Continuation from the previous page

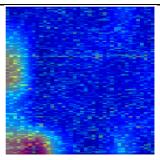
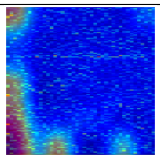
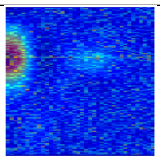
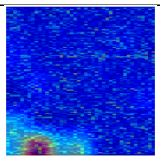
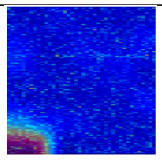
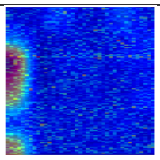
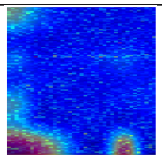
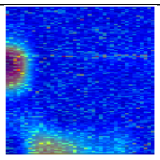
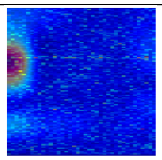
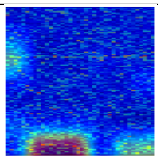
Sequence ID	Classification	Confidence Score	Grad-CAM
KX232616	Recombinant	99.957%	
KX232622	Recombinant	99.974%	
KX389610	Recombinant	99.845%	
KY658712	Recombinant	99.527%	
L39106	Recombinant	98.921%	
MF157740	Recombinant	99.960%	
MH078539	Recombinant	98.766%	
MN153477	Recombinant	99.926%	
MN153478	Recombinant	99.934%	
MN153479	Recombinant	99.590%	

Table D.11 – Continuation from the previous page

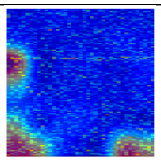
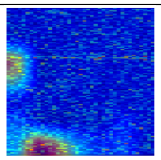
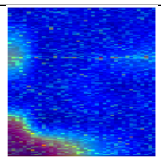
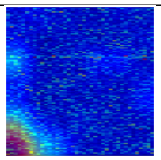
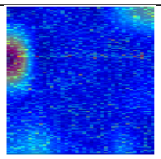
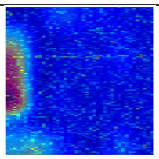
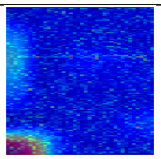
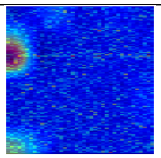
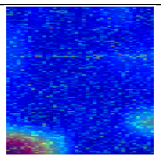
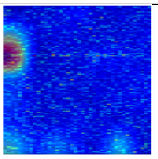
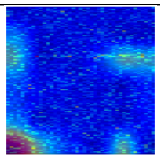
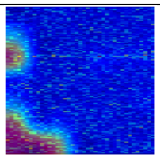
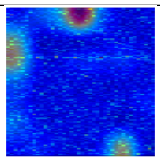
Sequence ID	**Classification**	**Confidence Score**	**Grad-CAM**
MN153480	Recombinant	99.960%	
MN153486	Recombinant	97.607%	
MN153494	Recombinant	97.899%	
MN153495	Recombinant	99.134%	
MN703128	Recombinant	99.644%	
MN703129	Recombinant	99.931%	
MN703130	Recombinant	99.573%	
MN703131	Recombinant	99.991%	
MN703132	Recombinant	99.799%	
MT222937	Recombinant	99.623%	

Table D.11 – Continuation from the previous page

Sequence ID	Classification	Confidence Score	Grad-CAM
MT222961	Recombinant	99.834%	
MT417738	Recombinant	98.258%	
MT877448	Recombinant	99.914%	

D.12 Subtype CRF04_cpx Sequences

Table D.12: Subtype CRF04_cpx Interpretability Patterns. This table includes both the CNN classification results and the hot zones (superposed spectrograms) of the complete HIV-1 subtype CRF04_cpx sequences mentioned in Chapter 8, listed in alphanumeric order.

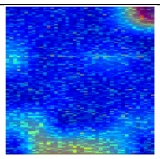
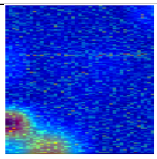
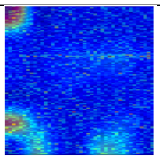
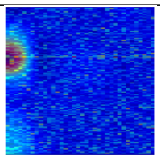
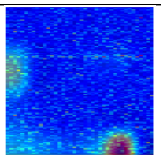
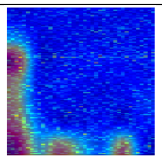
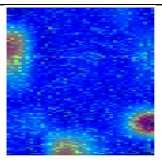
Sequence ID	Classification	Confidence Score	Grad-CAM
AF049337	Recombinant	88.671%	
AF119819	Recombinant	96.615%	
AF119820	Recombinant	92.580%	
FJ388917	Recombinant	96.297%	

Table D.12 – Continuation from the previous page

Sequence ID	Classification	Confidence Score	Grad-CAM
JX140648	Recombinant	97.716%	
MN153475	Recombinant	99.120%	

D.13 Subtype CRF06_cpx Sequences

Table D.13: Subtype CRF06_cpx Interpretability Patterns. This table includes both the CNN classification results and the hot zones (superposed spectrograms) of the complete HIV-1 subtype CRF06_cpx sequences mentioned in Chapter 8, listed in alphanumeric order.

Sequence ID	Classification	Confidence Score	Grad-CAM
MN153481	Recombinant	99.772%	

D.14 Subtype CRF09_cpx Sequences

Table D.14: Subtype CRF09_cpx Interpretability Patterns. This table includes both the CNN classification results and the hot zones (superposed spectrograms) of the complete HIV-1 subtype CRF09_cpx sequences mentioned in Chapter 8, listed in alphanumeric order.

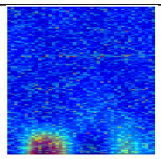
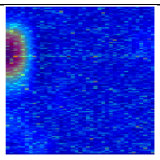
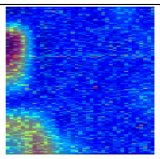
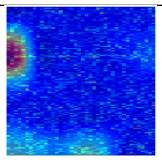
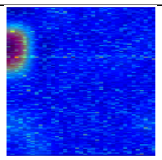
Sequence ID	Classification	Confidence Score	Grad-CAM
AJ866553	Recombinant	98.759%	

Table D.14 – Continuation from the previous page

Sequence ID	Classification	Confidence Score	Grad-CAM
AY093603	Recombinant	99.972%	
AY093604	Recombinant	99.894%	
AY093605	Recombinant	99.189%	
AY093607	Recombinant	99.688%	

D.15 Subtype CRF11_cpx Sequences

Table D.15: Subtype CRF11_cpx Interpretability Patterns. This table includes both the CNN classification results and the hot zones (superposed spectrograms) of the complete HIV-1 subtype CRF11_cpx sequences mentioned in Chapter 8, listed in alphanumeric order.

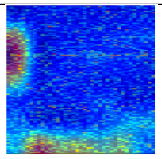
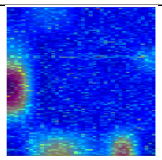
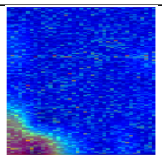
Sequence ID	Classification	Confidence Score	Grad-CAM
AF492623	Recombinant	85.659%	
AF492624	Recombinant	99.352%	
AF179368	Recombinant	98.542%	

Table D.15 – Continuation from the previous page

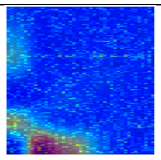
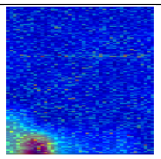
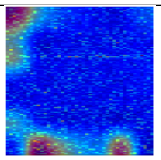
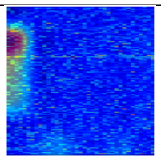
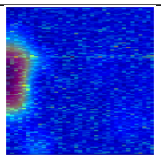
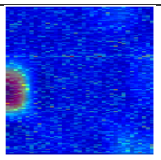
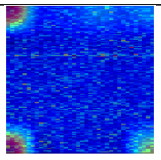
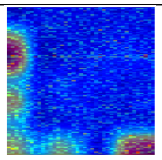
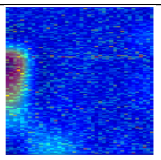
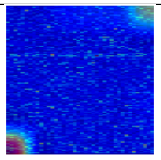
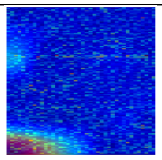
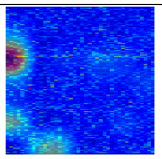
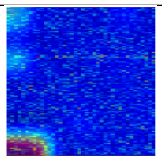
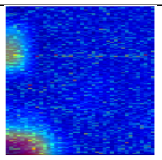
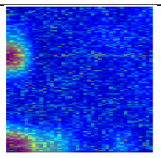
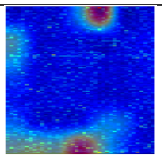
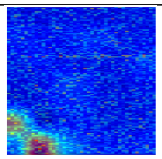
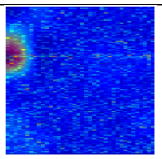
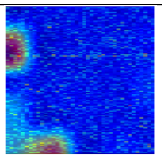
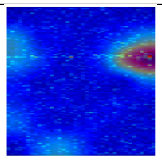
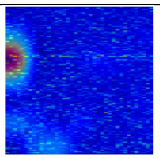
Sequence ID	**Classification**	**Confidence Score**	**Grad-CAM**
AJ291718	Recombinant	52.328%	
AJ291719	Recombinant	98.856%	
AJ291720	Recombinant	93.096%	
AY371149	Recombinant	92.801%	
AY371150	Recombinant	99.859%	
AY371151	Recombinant	99.686%	
JF683802	Non-recombinant	98.782%	
KP109503	Recombinant	96.425%	
KP718914	Recombinant	99.799%	
KP718929	Non-recombinant	59.627%	

Table D.15 – Continuation from the previous page

Sequence ID	Classification	Confidence Score	Grad-CAM
KP718934	Recombinant	99.956%	
KP718935	Recombinant	95.160%	
KP718936	Recombinant	98.496%	
KP718937	Recombinant	99.862%	
KP718938	Recombinant	96.437%	
KR822830	Recombinant	84.367%	
KU168270	Recombinant	91.791%	
KX389633	Recombinant	99.706%	
KY658700	Recombinant	99.995%	
MF373206	Recombinant	69.889%	

D.16 Subtype CRF12_BF Sequences

Table D.16: Subtype CRF12_BF Interpretability Patterns. This table includes both the CNN classification results and the hot zones (superposed spectrograms) of the complete HIV-1 subtype CRF12_BF sequences mentioned in Chapter 8, listed in alphanumeric order.

Sequence ID	Classification	Confidence Score	Grad-CAM
MT712270	Recombinant	99.571%	

D.17 Subtype CRF13_cpx Sequences

Table D.17: Subtype CRF13_cpx Interpretability Patterns. This table includes both the CNN classification results and the hot zones (superposed spectrograms) of the complete HIV-1 subtype CRF13_cpx sequences mentioned in Chapter 8, listed in alphanumeric order.

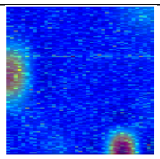
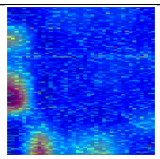
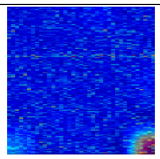
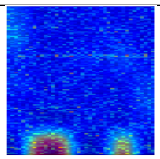
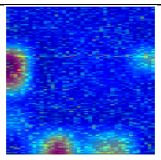
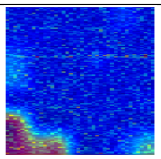
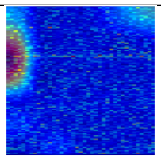
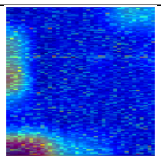
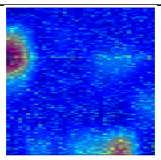
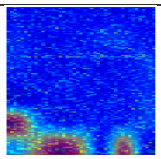
Sequence ID	Classification	Confidence Score	Grad-CAM
AF460972	Recombinant	76.672%	
AF460974	Recombinant	63.107%	
AY371154	Non-recombinant	54.231%	
DQ845386	Recombinant	98.412%	

Table D.17 – Continuation from the previous page

Sequence ID	Classification	Confidence Score	Grad-CAM
DQ845387	Recombinant	70.371%	
DQ845388	Recombinant	93.773%	
KP718924	Recombinant	99.760%	
KP718926	Recombinant	96.912%	
MF109417	Recombinant	51.066%	
MH705141	Recombinant	99.350%	

D.18 Subtype CRF16_A2D Sequences

Table D.18: Subtype CRF16_A2D Interpretability Patterns. This table includes both the CNN classification results and the hot zones (superposed spectrograms) of the complete HIV-1 subtype CRF16_A2D sequences mentioned in Chapter 8, listed in alphanumeric order.

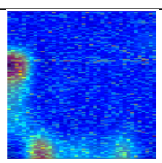
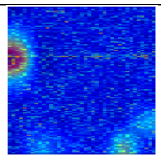
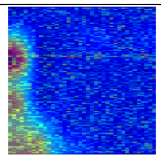
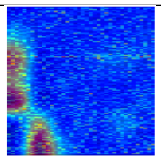
Sequence ID	Classification	Confidence Score	Grad-CAM
AF286239	Recombinant	95.870%	

Table D.18 – Continuation from the previous page

Sequence ID	Classification	Confidence Score	Grad-CAM
AY945736	Recombinant	99.652%	
KT022403	Recombinant	99.389%	
KT022406	Recombinant	94.974%	

D.19 Subtype CRF18_cpx Sequences

Table D.19: Subtype CRF18_cpx Interpretability Patterns. This table includes both the CNN classification results and the hot zones (superposed spectrograms) of the complete HIV-1 subtype CRF18_cpx sequences mentioned in Chapter 8, listed in alphanumeric order.

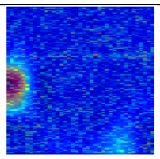
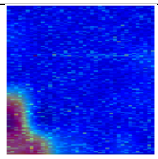
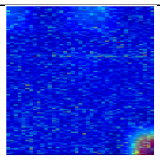
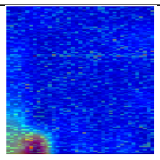
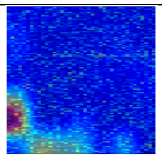
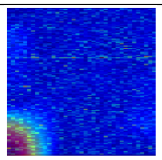
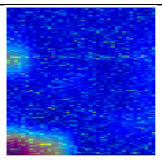
Sequence ID	Classification	Confidence Score	Grad-CAM
AF377959	Recombinant	96.008%	
AY586540	Recombinant	96.447%	
AY586541	Non-recombinant	74.161%	
AY894993	Recombinant	98.893%	

Table D.19 – Continuation from the previous page

Sequence ID	Classification	Confidence Score	Grad-CAM
KP718931	Recombinant	99.569%	
MF109378	Recombinant	98.499%	

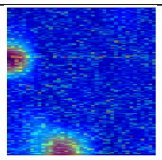
D.20 Subtype CRF20_BG Sequences

Table D.20: Subtype CRF20_BG Interpretability Patterns. This table includes both the CNN classification results and the hot zones (superposed spectrograms) of the complete HIV-1 subtype CRF20_BG sequences mentioned in Chapter 8, listed in alphanumeric order.

Sequence ID	Classification	Confidence Score	Grad-CAM
MW062981	Recombinant	99.958%	

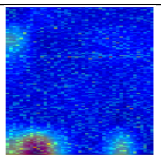
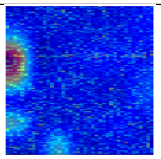
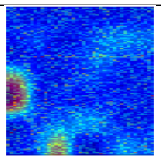
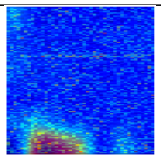
D.21 Subtype CRF22_01A1 Sequences

Table D.21: Subtype CRF22_01A1 Interpretability Patterns. This table includes both the CNN classification results and the hot zones (superposed spectrograms) of the complete HIV-1 subtype CRF22_01A1 sequences mentioned in Chapter 8, listed in alphanumeric order.

Sequence ID	Classification	Confidence Score	Grad-CAM
MN153487	Recombinant	98.743%	

D.22 Subtype CRF26_A5U Sequences

Table D.22: Subtype CRF26_A5U Interpretability Patterns. This table includes both the CNN classification results and the hot zones (superposed spectrograms) of the complete HIV-1 subtype CRF26_A5U sequences mentioned in Chapter 8, listed in alphanumeric order.

Sequence ID	Classification	Confidence Score	Grad-CAM
FM877777	Recombinant	99.886%	
FM877780	Recombinant	99.931%	
FM877781	Recombinant	99.733%	
FM877782	Recombinant	90.827%	

D.23 Subtype CRF27_cpx Sequences

Table D.23: Subtype CRF27_cpx Interpretability Patterns. This table includes both the CNN classification results and the hot zones (superposed spectrograms) of the complete HIV-1 subtype CRF27_cpx sequences mentioned in Chapter 8, listed in alphanumeric order.

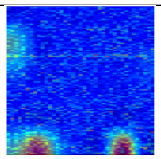
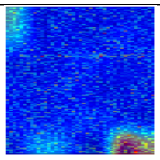
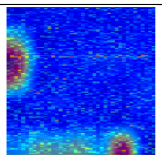
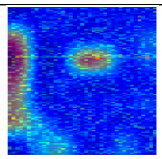
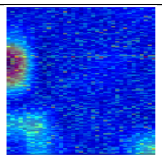
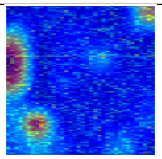
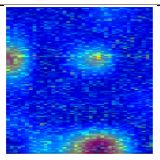
Sequence ID	Classification	Confidence Score	Grad-CAM
AJ404325	Recombinant	99.275%	

Table D.23 – Continuation from the previous page

Sequence ID	Classification	Confidence Score	Grad-CAM
AM851091	Recombinant	99.769%	
MN116204	Recombinant	99.750%	

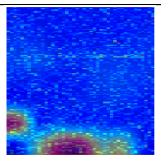
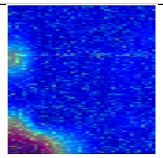
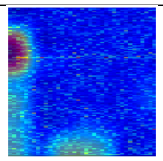
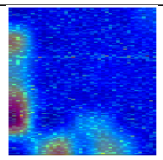
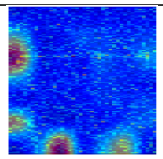
D.24 Subtype CRF37_cpx Sequences

Table D.24: Subtype CRF37_cpx Interpretability Patterns. This table includes both the CNN classification results and the hot zones (superposed spectrograms) of the complete HIV-1 subtype CRF37_cpx sequences mentioned in Chapter 8, listed in alphanumeric order.

Sequence ID	Classification	Confidence Score	Grad-CAM
AF377957	Recombinant	99.943%	
EF116594	Recombinant	99.397%	
JF683745	Recombinant	94.574%	
KP718917	Recombinant	58.757%	

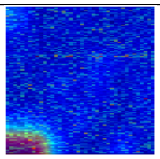
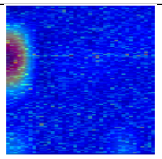
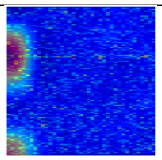
D.25 Subtype CRF45_cpx Sequences

Table D.25: Subtype CRF45_cpx Interpretability Patterns. This table includes both the CNN classification results and the hot zones (superposed spectrograms) of the complete HIV-1 subtype CRF45_cpx sequences mentioned in Chapter 8, listed in alphanumeric order.

Sequence ID	Classification	Confidence Score	Grad-CAM
EU448295	Recombinant	83.951%	
FN392874	Recombinant	99.806%	
FN392875	Recombinant	98.699%	
FN392876	Recombinant	95.330%	
FN392877	Recombinant	98.648%	

D.26 Subtype CRF63_02A6 Sequences

Table D.26: Subtype CRF63_02A6 Interpretability Patterns. This table includes both the CNN classification results and the hot zones (superposed spectrograms) of the complete HIV-1 subtype CRF63_02A6 sequences mentioned in Chapter 8, listed in alphanumeric order.

Sequence ID	Classification	Confidence Score	Grad-CAM
JN230353	Recombinant	99.935%	
KJ197201	Recombinant	99.830%	
KJ197202	Recombinant	99.967%	

Rara avis in terris, nigroque simillima cygno (A rare bird in the earth, and very like a black swan).

— Juvenal, *he insulted an actor with high levels of influence in the upper echelons of Rome and had to go into exile. It is not known for certain whether his exile was in Egypt or in Scotland.*

E

Compendium of Subtype G Sequences in Alphanumeric Order

"Sequence ID" corresponds to the unique identifier of each sequence. "Classification" indicates the result of the classification performed by the CNN. "Confidence Score" indicates the level of confidence demonstrated by the CNN in its classification. High values indicate a high degree of certainty by the CNN in its diagnosis. "Grad-CAM" provides the heat map resulting from the application of Grad-CAM on the superposed genomic spectrogram of each sequence.

E.1 Complete Subtype G Interpretability Patterns

Table E.1: Subtype G Interpretability Patterns. This table includes both the CNN classification results and the hot zones (superposed spectrograms) of the 99 complete HIV-1 sequences included within subtype G. The G sequences are listed in alphanumeric order.

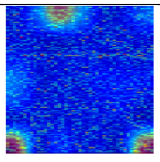
Sequence ID	Classification	Confidence Score	Grad-CAM
AB231893	Non-recombinant	98.911%	

Table E.1 – Continuation from the previous page

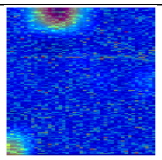
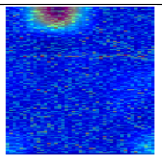
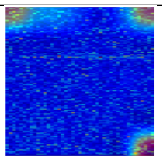
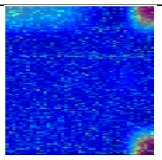
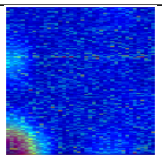
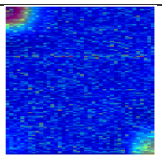
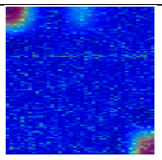
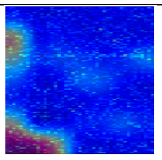
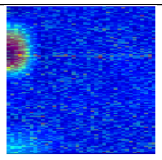
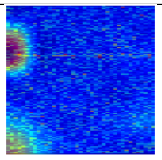
Sequence ID	Classification	Confidence Score	Grad-CAM
AB287003	Non-recombinant	97.376%	
AB287004	Non-recombinant	96.559%	
AB485662	Non-recombinant	87.516%	
AB485663	Non-recombinant	67.105%	
AF061640	Recombinant	52.583%	
AF061641	Non-recombinant	78.448%	
AF061642	Non-recombinant	91.067%	
AF084936	Recombinant	98.735%	
AF423760	Recombinant	95.373%	
AF450098	Recombinant	79.366%	

Table E.1 – Continuation from the previous page

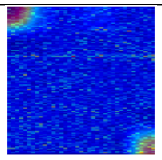
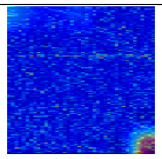
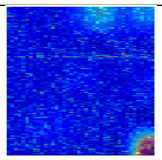
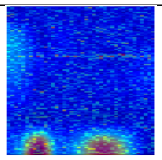
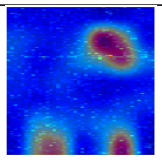
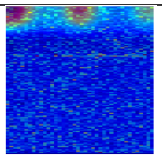
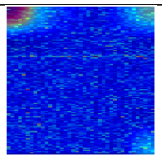
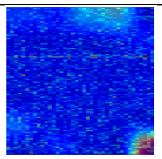
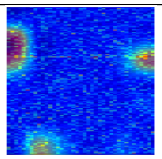
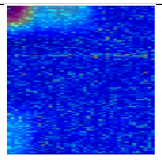
Sequence ID	**Classification**	**Confidence Score**	**Grad-CAM**
AY371121	Non-recombinant	89.034%	
AY586547	Non-recombinant	52.077%	
AY586548	Non-recombinant	59.551%	
AY586549	Recombinant	69.026%	
AY612637	Recombinant	51.362%	
AY772535	Non-recombinant	80.442%	
DQ168573	Non-recombinant	98.712%	
DQ168575	Non-recombinant	99.994%	
DQ168576	Recombinant	95.409%	
DQ168579	Non-recombinant	98.203%	

Table E.1 – Continuation from the previous page

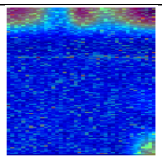
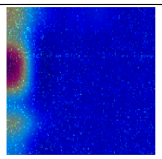
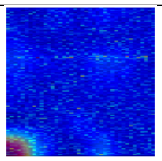
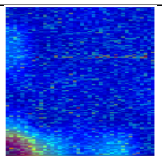
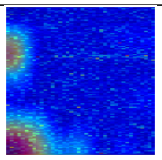
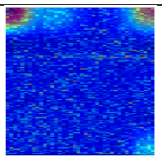
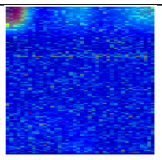
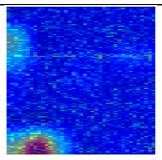
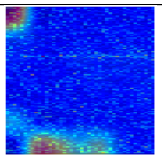
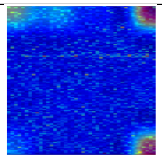
Sequence ID	Classification	Confidence Score	Grad-CAM
EU786670	Non-recombinant	59.745%	
FJ389363	Recombinant	98.285%	
FJ389364	Recombinant	97.485%	
FJ389365	Recombinant	75.938%	
FJ389366	Recombinant	79.741%	
FJ389367	Non-recombinant	88.919%	
FJ670520	Non-recombinant	87.638%	
FJ670530	Recombinant	60.648%	
FR846408	Recombinant	53.208%	
FR846409	Non-recombinant	98.008%	

Table E.1 – Continuation from the previous page

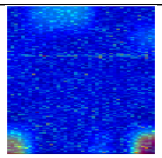
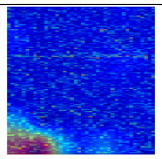
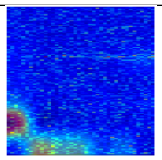
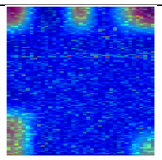
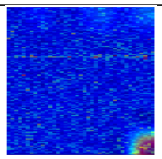
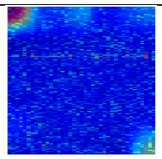
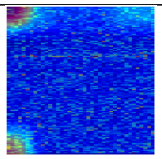
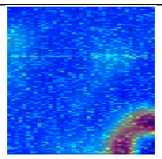
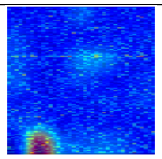
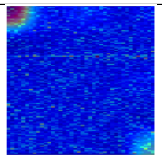
Sequence ID	**Classification**	**Confidence Score**	**Grad-CAM**
FR846410	Recombinant	54.472%	
GU362882	Recombinant	65.706%	
HM067749	Recombinant	92.267%	
JN106043	Non-recombinant	57.452%	
JN248582	Non-recombinant	99.939%	
JN248584	Non-recombinant	94.672%	
JN248586	Non-recombinant	84.494%	
JN248591	Recombinant	51.692%	
JN248593	Recombinant	61.615%	
JX140674	Non-recombinant	67.745%	

Table E.1 – Continuation from the previous page

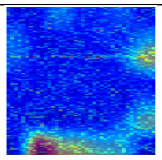
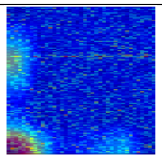
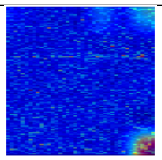
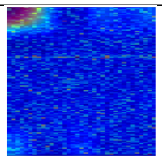
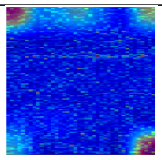
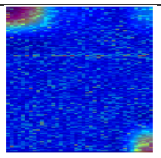
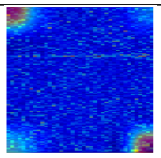
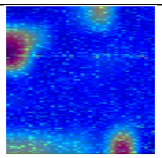
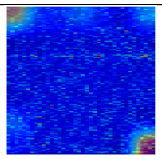
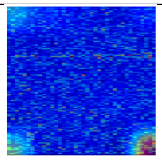
Sequence ID	Classification	Confidence Score	Grad-CAM
JX140675	Recombinant	70.169%	
JX140676	Recombinant	98.180%	
KF716477	Non-recombinant	99.266%	
KJ948662	Non-recombinant	99.914%	
KP109502	Non-recombinant	66.358%	
KP718915	Non-recombinant	99.821%	
KP718923	Non-recombinant	99.697%	
KP718925	Recombinant	96.602%	
KR017776	Non-recombinant	99.521%	
KT022379	Non-recombinant	97.569%	

Table E.1 – Continuation from the previous page

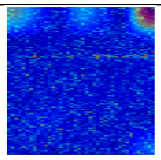
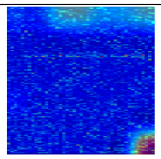
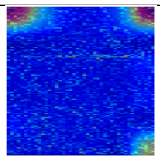
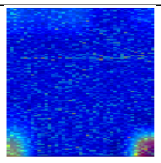
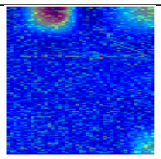
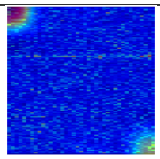
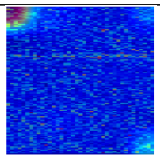
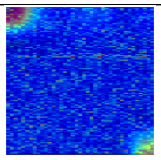
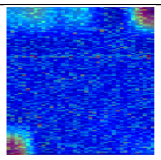
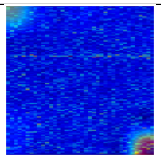
Sequence ID	**Classification**	**Confidence Score**	**Grad-CAM**
KT276261	Non-recombinant	56.534%	
KU168277	Non-recombinant	71.045%	
KU168278	Non-recombinant	74.893%	
KU168300	Non-recombinant	86.844%	
KU168302	Non-recombinant	86.075%	
KU685592	Non-recombinant	60.288%	
KX228801	Non-recombinant	99.562%	
KX228812	Non-recombinant	88.905%	
KX389615	Non-recombinant	99.948%	
KX389618	Non-recombinant	91.092%	

Table E.1 – Continuation from the previous page

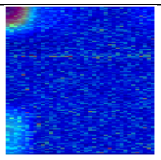
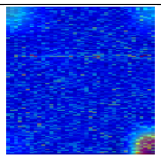
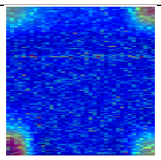
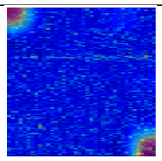
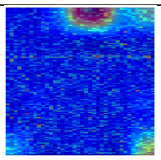
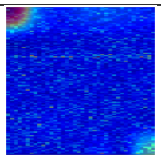
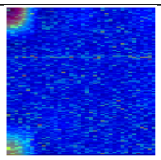
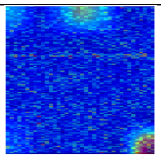
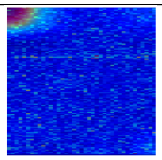
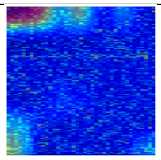
Sequence ID	Classification	Confidence Score	Grad-CAM
KX389619	Non-recombinant	87.054%	
KX389620	Non-recombinant	99.147%	
KX389625	Non-recombinant	99.771%	
KX389626	Non-recombinant	58.888%	
KX389628	Non-recombinant	96.122%	
KX389631	Non-recombinant	89.406%	
KX389635	Non-recombinant	81.198%	
KX389636	Non-recombinant	97.562%	
KX389641	Non-recombinant	57.324%	
KX389642	Non-recombinant	61.288%	

Table E.1 – Continuation from the previous page

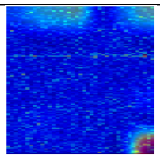
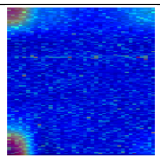
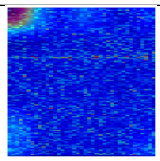
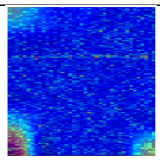
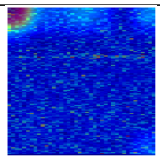
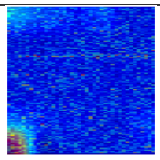
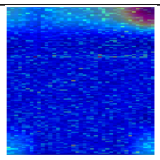
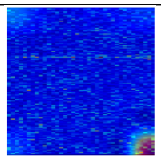
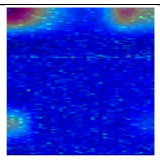
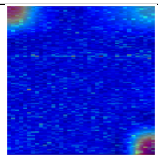
Sequence ID	**Classification**	**Confidence Score**	**Grad-CAM**
KX389645	Non-recombinant	98.702%	
KX389646	Non-recombinant	99.690%	
KX389648	Non-recombinant	88.867%	
KY275364	Non-recombinant	99.843%	
KY953200	Non-recombinant	75.915%	
KY953201	Non-recombinant	60.065%	
KY953202	Non-recombinant	99.913%	
MF109568	Non-recombinant	95.335%	
MF109598	Non-recombinant	98.178%	
MF109647	Non-recombinant	69.453%	

Table E.1 – Continuation from the previous page

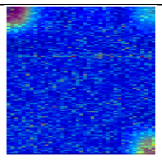
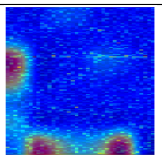
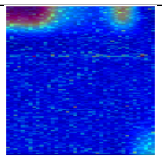
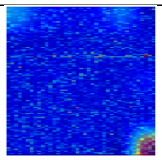
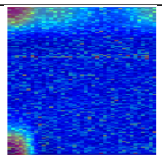
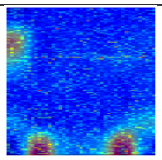
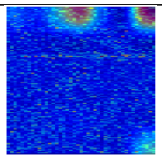
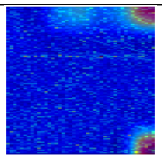
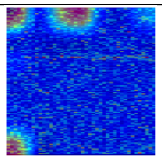
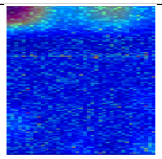
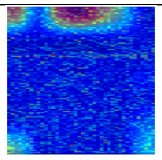
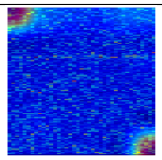
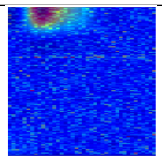
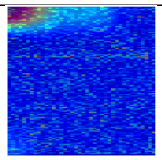
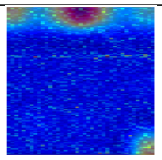
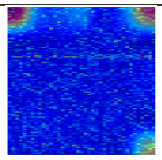
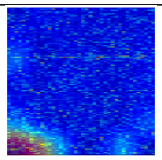
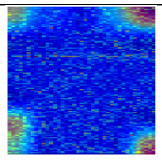
Sequence ID	Classification	Confidence Score	Grad-CAM
MF157743	Non-recombinant	82.400%	
MF614605	Recombinant	59.337%	
MF614606	Non-recombinant	99.152%	
MF614607	Non-recombinant	97.817%	
MH431770	Non-recombinant	88.232%	
MH705134	Recombinant	61.419%	
MH705145	Non-recombinant	55.218%	
MH705155	Non-recombinant	59.600%	
MH705162	Non-recombinant	70.117%	
MH746255	Non-recombinant	70.052%	

Table E.1 – Continuation from the previous page

Sequence ID	**Classification**	**Confidence Score**	**Grad-CAM**
MK254637	Non-recombinant	98.884%	
MK254639	Non-recombinant	93.354%	
MN153484	Non-recombinant	98.570%	
MN486044	Non-recombinant	95.916%	
MT334914	Non-recombinant	85.527%	
MT349406	Non-recombinant	86.871%	
MT349407	Recombinant	61.462%	
U88826	Non-recombinant	97.288%	

Start by doing what's necessary, then do what's possible, and suddenly you are doing the impossible.

— San Francesco d'Assisi, *legend has it that he made a pact with a wolf that was terrorizing Gubbio. In exchange for ceasing its mischief, the townsfolk would provide him with the sustenance he needed. The beast placed a front paw on Francis's hand in a sign of agreement. It managed to live among the people and died two years later in old age.*

F

SARS-CoV-2 Complete Results per Subsampling

The results for each subsampling are presented in the following tables. The rows highlighted in green correspond to configurations with the highest test accuracy and a better balance between both categories.

In all cases, Learning Rate is 0.0001 and Batch Size is 52.

Table F.1: Results in Subsampling 01 by the Number of Epochs. The configurations with 150 and 200 Epochs share optimal accuracy percentages in both categories. They exhibit the highest Test Accuracy values and the minimum inter-category SD. With the exception of the 10 and 25 Epoch configurations, which perform poorly compared to the other configurations, the rest demonstrate good accuracy rates.

SUBSAMPLING 01					
Epochs	**AUC**	**Validation Accuracy**	**Test Accuracy**	**Confusion Matrix**	**Inter Category SD**
10	0.9071	86.26%	84.03%	257 56 44 269	8.49
25	0.9529	87.54%	83.23%	294 19 86 227	47.38
50	0.9735	93.29%	93.13%	290 23 20 293	2.12
75	0.9769	91.37%	93.77%	287 26 13 300	9.19
100	0.9786	94.41%	94.89%	301 12 20 293	5.66
150	0.9818	94.25%	96.01%	302 11 14 299	2.12
200	0.9823	94.57%	96.01%	302 11 14 299	2.12
250	0.9830	94.41%	95.37%	301 12 17 296	3.54

Table F.2: Results in Subsampling 02 by the Number of Epochs. The highest test accuracy values are achieved at 150 and 250 Epochs, both exceeding 95%. While the 150 Epoch configuration is slightly less balanced than the 250 Epoch one, we consider both configurations as optimal, with only 26 and 28 total errors in the test set.

SUBSAMPLING 02					
Epochs	**AUC**	**Validation Accuracy**	**Test Accuracy**	**Confusion Matrix**	**Inter Category SD**
10	0.8941	81.95%	80.19%	213 100 24 289	53.74
25	0.9371	89.94%	85.46%	289 24 67 246	30.41
50	0.9714	94.41%	92.49%	279 34 13 300	14.85
75	0.9684	93.77%	92.33%	303 10 38 275	19.80
100	0.9814	94.73%	94.89%	298 15 17 296	1.41
150	0.9806	94.57%	95.85%	303 10 16 297	4.24
200	0.9784	94.73%	94.57%	293 20 14 299	4.24
250	0.9796	95.05%	95.53%	298 15 13 300	1.41

Table F.3: Results in Subsampling 03 by the Number of Epochs. The 200 Epoch configuration exhibits the optimal test accuracy value with a high balance between the two categories (Inter-Category SD lower than 1%). These results are significantly better than those of the remaining configurations.

SUBSAMPLING 03					
Epochs	**AUC**	**Validation Accuracy**	**Test Accuracy**	**Confusion Matrix**	**Inter Category SD**
10	0.9224	84.66%	82.11%	222 91 21 292	49.50
25	0.9692	89.62%	90.58%	277 36 23 290	9.19
50	0.9745	93.93%	92.33%	297 16 32 281	11.31
75	0.9696	93.77%	91.21%	295 18 37 276	13.44
100	0.9812	95.37%	94.57%	299 14 20 293	4.24
150	0.9807	94.89%	94.57%	301 12 22 291	7.07
200	0.9824	95.21%	94.73%	297 16 17 296	0.71
250	0.9805	94.41%	94.09%	301 12 25 288	9.19

Table F.4: Results in Subsampling 04 by the Number of Epochs. The 200 Epoch configuration achieves the highest test accuracy value. Although the Inter-Category SD is slightly higher than at 150 Epochs, the total of 28 errors compared to 31 errors in the case of 150 Epochs lead us to select this configuration as the best for this subsampling.

SUBSAMPLING 04					
Epochs	**AUC**	**Validation Accuracy**	**Test Accuracy**	**Confusion Matrix**	**Inter Category SD**
10	0.9039	88.18%	83.23%	246 67 38 275	20.51
25	0.9678	92.97%	89.78%	286 27 37 276	7.07
50	0.9784	94.57%	93.45%	297 16 25 288	6.36
75	0.9786	94.57%	93.77%	295 18 21 292	2.12
100	0.9769	94.25%	92.49%	294 19 28 285	6.36
150	0.9811	95.69%	95.05%	302 11 20 293	6.36
200	0.9808	96.17%	95.53%	304 9 19 294	7.07
250	0.9790	95.69%	94.57%	303 10 24 289	9.90

Table F.5: Results in Subsampling 05 by the Number of Epochs. The configurations of 200 and 250 Epochs share the same test accuracy value, with a slight penalty in the inter-category standard deviation for 250 Epochs compared to 200 Epochs.

SUBSAMPLING 05					
Epochs	**AUC**	**Validation Accuracy**	**Test Accuracy**	**Confusion Matrix**	**Inter Category SD**
10	0.8755	84.35%	79.23%	254 59 71 242	8.49
25	0.9481	91.53%	88.18%	268 45 29 284	11.31
50	0.9486	92.49%	85.30%	297 16 76 237	42.43
75	0.9680	94.73%	92.17%	285 28 21 292	4.95
100	0.9613	93.93%	90.58%	291 22 37 276	10.61
150	0.9722	94.73%	92.97%	293 20 24 289	2.83
200	0.9698	95.53%	93.13%	294 19 24 289	3.54
250	0.9732	95.21%	93.13%	288 25 18 295	4.95

Table F.6: Results in Subsampling 06 by the Number of Epochs. The 200 Epoch configuration is clearly the optimal one compared to the other configurations, both in terms of test accuracy values and, above all, Inter-Category SD.

SUBSAMPLING 06					
Epochs	**AUC**	**Validation Accuracy**	**Test Accuracy**	**Confusion Matrix**	**Inter Category SD**
10	0.8746	85.30%	79.71%	257 56 71 242	10.61
25	0.9390	89.30%	87.22%	277 36 44 269	5.66
50	0.9644	92.97%	91.69%	290 23 29 284	4.24
75	0.9665	93.93%	92.17%	289 24 25 288	0.71
100	0.9737	94.41%	94.25%	299 14 22 291	5.66
150	0.9744	93.93%	93.93%	304 9 29 284	14.14
200	0.9782	94.57%	94.57%	297 16 18 295	1.41
250	0.9788	94.73%	94.25%	300 13 23 290	7.07

Table F.7: Results in Subsampling 07 by the Number of Epochs. The 150 and 200 configurations are the only ones that surpass 95% test accuracy. Although the Inter-Category SD is high in both configurations, especially when compared to 100 Epochs, we consider that, in this subsampling, even though the accuracy between both categories is imbalanced, these configurations are the most stable in overall performance.

SUBSAMPLING 07					
Epochs	**AUC**	**Validation Accuracy**	**Test Accuracy**	**Confusion Matrix**	**Inter Category SD**
10	0.9108	84.50%	83.55%	274 39 64 249	17.68
25	0.9589	92.17%	87.50%	299 14 63 250	34.65
50	0.9795	93.93%	94.57%	301 12 22 291	7.07
75	0.9793	93.93%	92.49%	311 2 45 268	30.41
100	0.9813	94.89%	92.97%	293 20 24 289	2.83
150	0.9824	94.73%	95.37%	309 4 25 288	14.85
200	0.9806	94.73%	95.21%	305 8 22 291	9.90
250	0.9806	94.57%	93.77%	297 16 23 290	4.95

Table F.8: Results in Subsampling 08 by the Number of Epochs. In this case, we consider the 150 and 200 Epoch configurations as the optimal ones due to their total hit rate in the test set, which is slightly superior to the 100 Epoch configuration, even though the Inter-Category SD is slightly lower.

SUBSAMPLING 08					
Epochs	**AUC**	**Validation Accuracy**	**Test Accuracy**	**Confusion Matrix**	**Inter Category SD**
10	0.9136	84.82%	81.95%	235 78 35 278	30.41
25	0.9461	85.78%	81.63%	306 7 108 205	71.42
50	0.9735	93.61%	93.13%	290 23 20 293	2.12
75	0.9714	93.93%	93.77%	297 16 23 290	4.95
100	0.9682	92.81%	90.26%	272 41 20 293	14.85
150	0.9769	91.05%	91.69%	275 38 14 299	16.97
200	0.9754	91.69%	91.69%	275 38 14 299	16.97
250	0.9739	93.93%	93.29%	292 21 21 292	0.00

Table F.9: Results in Subsampling 09 by the Number of Epochs. In this case, we have chosen the 100 Epoch configuration as the optimal one due to its highest test accuracy value and lowest Inter-Category SD.

SUBSAMPLING 09					
Epochs	**AUC**	**Validation Accuracy**	**Test Accuracy**	**Confusion Matrix**	**Inter Category SD**
10	0.9324	83.55%	86.58%	266 47 37 276	7.07
25	0.9629	90.58%	89.46%	295 18 48 265	21.21
50	0.9735	86.10%	93.13%	290 23 20 293	2.12
75	0.9804	93.45%	94.73%	301 12 21 292	6.36
100	0.9842	94.89%	96.17%	305 8 16 297	5.66
150	0.9843	94.89%	95.69%	307 6 21 292	10.61
200	0.9845	94.89%	95.53%	307 6 22 291	11.31
250	0.9831	94.73%	95.21%	304 9 21 292	8.49

Table F.10: Results in Subsampling 10 by the Number of Epochs. The optimal configuration is clearly 200 Epochs, as it is the only one to exceed 96% test accuracy with low levels of Inter-Category SD.

SUBSAMPLING 10					
Epochs	**AUC**	**Validation Accuracy**	**Test Accuracy**	**Confusion Matrix**	**Inter Category SD**
10	0.8786	83.87%	78.91%	268 45 87 226	29.70
25	0.9525	89.62%	84.98%	305 8 86 227	55.15
50	0.9590	92.17%	92.33%	286 27 21 292	4.24
75	0.9703	92.33%	93.93%	295 18 20 293	1.41
100	0.9739	93.29%	94.41%	303 10 25 288	10.61
150	0.9726	93.45%	94.89%	304 9 23 290	9.90
200	0.9728	93.45%	95.05%	302 11 20 293	6.36
250	0.9729	93.77%	95.05%	302 11 20 293	6.36

The Papalagi are poor because they pursue things like madmen. Without things they cannot live at all.

— Tuiavii de Tiavea, *samoan chief. Erich Scheurmann was a great friend of his. He translated his speeches and compiled them into the book "The Speeches of the South Sea Chief Tuiavii of Tiavea" better known as "The Papalagi". This book is a great life lesson.*

G

SARS-CoV-2 Complete Results per Number of Epochs

Table G.1: Results at 10 Epochs. Comparing to the entire test dataset (313 genomic spectrograms per category), the average test accuracy can be improved. The hit rate is low, and the inter-category standard deviation is high, indicating an excessive imbalance between the accuracy in both categories. This configuration likely suffers from a lack of training.

10 EPOCHS				
SUBSAMPLING	**Non-rec. Test Acc.**	**Non-rec. Test Err.**	**Rec. Test Acc.**	**Rec. Test Err.**
SUB_01	257	56	269	44
SUB_02	213	100	289	24
SUB_03	222	91	292	21
SUB_04	246	67	275	38
SUB_05	254	59	242	71
SUB_06	257	56	242	71
SUB_07	274	39	249	64
SUB_08	235	78	278	35
SUB_09	266	47	276	37
SUB_10	268	45	226	87
Avg.	**249.2**		**263.8**	
SD	**20.16**		**22.44**	
Inter Category SD = 10.32				

Table G.2: Results at 25 Epochs. Similar to the previous case, this configuration is in need of improvement in terms of accuracy in both categories. It is particularly striking that there is a high imbalance between the accuracy of recombinants (which is also only 81%) and non-recombinants, exceeding even 25 points. This configuration lacks sufficient training and is therefore not suitable.

25 EPOCHS				
SUBSAMPLING	**Non-rec. Test Acc.**	**Non-rec. Test Err.**	**Rec. Test Acc.**	**Rec. Test Err.**
SUB_01	294	19	227	86
SUB_02	289	24	246	67
SUB_03	277	36	290	23
SUB_04	286	27	276	37
SUB_05	268	45	284	29
SUB_06	277	36	269	44
SUB_07	299	14	250	63
SUB_08	306	7	205	108
SUB_09	295	18	265	48
SUB_10	305	8	227	86
Avg.	**289.6**		**253.9**	
SD	**12.65**		**27.82**	
Inter Category SD = 25.24				

Table G.3: Results at 50 Epochs. The average test accuracy values are better compared to previous configurations in both categories, as well as the balance between both values. However, when compared to the subsequent configurations, we observe some potential for improvement by increasing the number of epochs.

50 EPOCHS				
SUBSAMPLING	**Non-rec. Test Acc.**	**Non-rec. Test Err.**	**Rec. Test Acc.**	**Rec. Test Err.**
SUB_01	290	23	293	20
SUB_02	279	34	300	13
SUB_03	297	16	281	32
SUB_04	297	16	288	25
SUB_05	297	16	237	76
SUB_06	290	23	284	29
SUB_07	301	12	291	22
SUB_08	290	23	293	20
SUB_09	290	23	293	20
SUB_10	286	27	292	21
Avg.	**291.7**		**285.2**	
SD	**6.46**		**17.74**	
Inter Category SD = 4.60				

Table G.4: Results at 75 Epochs. No significant improvements compared to 50 Epochs. While there are slight enhancements in test accuracy values in both categories, the inter-category standard deviation is slightly higher.

75 EPOCHS				
SUBSAMPLING	**Non-rec. Test Acc.**	**Non-rec. Test Err.**	**Rec. Test Acc.**	**Rec. Test Err.**
SUB_01	287	26	300	13
SUB_02	303	10	275	38
SUB_03	295	18	276	37
SUB_04	295	18	292	21
SUB_05	285	28	292	21
SUB_06	289	24	288	25
SUB_07	311	2	268	45
SUB_08	297	16	290	23
SUB_09	301	12	292	21
SUB_10	295	18	293	20
Avg.	**295.8**		**286.6**	
SD	**7.84**		**10.08**	
Inter Category SD = 6.51				

Table G.5: Results at 100 Epochs. Both test accuracy values surpass 290 correct predictions (over 92%), and the inter-category standard deviation drops below 4 points for the first time. This configuration may be sufficient, although there is potential for improvement.

100 EPOCHS				
SUBSAMPLING	**Non-rec. Test Acc.**	**Non-rec. Test Err.**	**Rec. Test Acc.**	**Rec. Test Err.**
SUB_01	301	12	293	20
SUB_02	298	15	296	17
SUB_03	299	14	293	20
SUB_04	294	19	285	28
SUB_05	291	22	276	37
SUB_06	299	14	291	22
SUB_07	293	20	289	24
SUB_08	272	41	293	20
SUB_09	305	8	297	16
SUB_10	303	10	288	25
Avg.	**295.5**		**290.1**	
SD	**9.36**		**6.14**	
Inter Category SD = 3.82				

Table G.6: Results at 150 Epochs. This configuration exhibits the highest test accuracy value in non-recombinants. However, it is slightly penalized by the imbalance with respect to the recombinants, where it performs well but is outperformed by the 200 Epoch configuration.

150 EPOCHS				
SUBSAMPLING	**Non-rec. Test Acc.**	**Non-rec. Test Err.**	**Rec. Test Acc.**	**Rec. Test Err.**
SUB_01	302	11	299	14
SUB_02	303	10	297	16
SUB_03	301	12	291	22
SUB_04	302	11	293	20
SUB_05	293	20	289	24
SUB_06	304	9	284	29
SUB_07	309	4	288	25
SUB_08	275	38	299	14
SUB_09	307	6	292	21
SUB_10	304	9	290	23
Avg.	**300.0**		**292.2**	
SD	**9.74**		**4.92**	
Inter Category SD = 5.52				

Table G.7: Results at 200 Epochs. Chosen as the optimal configuration due to its high accuracy rates in both categories and its optimal balance between both test accuracy.

200 EPOCHS				
SUBSAMPLING	**Non-rec. Test Acc.**	**Non-rec. Test Err.**	**Rec. Test Acc.**	**Rec. Test Err.**
SUB_01	302	11	299	14
SUB_02	293	20	299	14
SUB_03	297	16	296	17
SUB_04	304	9	294	19
SUB_05	294	19	289	24
SUB_06	297	16	295	18
SUB_07	305	8	291	22
SUB_08	275	38	299	14
SUB_09	307	6	291	22
SUB_10	302	11	293	20
Avg.	**297.6**		**294.6**	
SD	**9.24**		**3.66**	
Inter Category SD = 2.12				

Table G.8: Results at 250 Epochs. The hit rate in recombinants decreases slightly compared to the 200 Epoch configuration. Additionally, the inter-category standard deviation is slightly higher. While it may not be the optimal configuration, the obtained results are positive.

250 EPOCHS				
SUBSAMPLING	**Non-rec. Test Acc.**	**Non-rec. Test Err.**	**Rec. Test Acc.**	**Rec. Test Err.**
SUB_01	301	12	296	17
SUB_02	298	15	300	13
SUB_03	301	12	288	25
SUB_04	303	10	289	24
SUB_05	288	25	295	18
SUB_06	300	13	290	23
SUB_07	297	16	290	23
SUB_08	292	21	292	21
SUB_09	304	9	292	21
SUB_10	302	11	293	20
Avg.	**298.6**		**292.5**	
SD	**5.08**		**3.66**	
Inter Category SD = 4.31				

Statistics is a science that shows that if my neighbor has two cars and I have none, we both have one.

— George Bernard Shaw, *he is the only person in history to won a Nobel and an Oscar. He loved flowers, but he hated receiving them as gifts because for him, bouquets were like a collection of decapitated flowers.*

H

SARS-CoV-2 Recombinant Sequences Study Case. Statistical Data on the Types of Nucleotides Present in the Arithmetic Series (Arithmetic Ratio = 6) at S Protein

Each row in every table represents a recombinant sequence, accurately classified by VGG-16 SARS-CoV-2 with a high level of confidence and a clearly located hot zone in the Spike protein at $f = 1/6$.

Table H.1: Frame 6, 12, 18, ... In all analyzed sequences, it is noteworthy that at positions 6, 12, 18, ..., the number of T is significantly higher compared to the other nucleotide types, reaching nearly 50% of presence compared to the other three nucleotide types combined.

ID	A	T	C	G	A (%)	T (%)	C (%)	G (%)
MZ558275	185	283	107	62	29.04%	44.43%	16.80%	9.73%
MZ280957	185	285	105	62	29.04%	44.74%	16.48%	9.73%
MZ817897	185	283	107	61	29.04%	44.43%	16.80%	9.58%
OK039706	185	283	107	62	29.04%	44.43%	16.80%	9.73%
OL791595	185	283	107	62	29.04%	44.43%	16.80%	9.73%
OL920359	185	283	107	62	29.04%	44.43%	16.80%	9.73%
OM059434	185	283	107	62	29.04%	44.43%	16.80%	9.73%
ON800622	184	285	106	62	28.89%	44.74%	16.64%	9.73%
ON144454	156	308	95	76	24.49%	48.35%	14.91%	11.93%
ON330593	156	308	95	76	24.49%	48.35%	14.91%	11.93%

Table H.1 – Continuation from the previous page

ID	A	T	C	G	A (%)	T (%)	C (%)	G (%)
ON594981	156	308	95	76	24.49%	48.35%	14.91%	11.93%
ON650495	156	308	95	76	24.49%	48.35%	14.91%	11.93%
ON980049	156	308	95	76	24.49%	48.35%	14.91%	11.93%
OP129015	156	308	95	76	24.49%	48.35%	14.91%	11.93%
OP294029	156	308	94	77	24.49%	48.35%	14.76%	12.09%
OP311263	156	308	95	76	24.49%	48.35%	14.91%	11.93%
OP313193	156	308	94	77	24.49%	48.35%	14.76%	12.09%
OP313256	156	307	95	77	24.49%	48.19%	14.91%	12.09%
OP313383	156	308	94	77	24.49%	48.35%	14.76%	12.09%
OP359817	156	308	94	77	24.49%	48.35%	14.76%	12.09%
OP768851	156	308	95	76	24.49%	48.35%	14.91%	11.93%
OP934590	157	308	94	76	24.65%	48.35%	14.76%	11.93%
OQ094825	180	291	103	61	28.26%	45.68%	16.17%	9.58%
OQ111935	173	286	102	60	27.16%	44.90%	16.01%	9.42%
OQ112611	173	287	101	60	27.16%	45.05%	15.86%	9.42%
OQ112616	173	287	101	60	27.16%	45.05%	15.86%	9.42%
OQ112619	173	287	101	60	27.16%	45.05%	15.86%	9.42%
OQ112624	173	287	101	60	27.16%	45.05%	15.86%	9.42%
OQ112639	173	287	101	60	27.16%	45.05%	15.86%	9.42%
OQ112647	173	287	101	60	27.16%	45.05%	15.86%	9.42%
OQ112649	173	287	101	60	27.16%	45.05%	15.86%	9.42%
OQ112673	173	287	101	60	27.16%	45.05%	15.86%	9.42%
OQ112686	173	287	101	60	27.16%	45.05%	15.86%	9.42%
OQ112691	173	287	101	60	27.16%	45.05%	15.86%	9.42%
OQ112730	173	287	101	60	27.16%	45.05%	15.86%	9.42%
OQ112737	173	287	101	60	27.16%	45.05%	15.86%	9.42%
OQ112740	173	287	101	60	27.16%	45.05%	15.86%	9.42%
OQ112771	173	287	101	60	27.16%	45.05%	15.86%	9.42%
OQ112893	173	287	101	60	27.16%	45.05%	15.86%	9.42%

Table H.2: Frame 7, 13, 19, ... In contrast to the previous case, the phenomenon related to the high quantity of T at these positions does not occur. The proportions of the presence of all four types of nucleotides are balanced in all sequences.

ID	A	T	C	G	A (%)	T (%)	C (%)	G (%)
MZ558275	193	154	113	176	30.35%	24.21%	17.77%	27.67%
MZ280957	193	154	113	176	30.35%	24.21%	17.77%	27.67%
MZ817897	193	154	113	175	30.35%	24.21%	17.77%	27.52%
OK039706	193	155	113	175	30.35%	24.37%	17.77%	27.52%
OL791595	193	155	113	175	30.35%	24.37%	17.77%	27.52%
OL920359	193	155	113	175	30.35%	24.37%	17.77%	27.52%
OM059434	193	156	112	175	30.35%	24.53%	17.61%	27.52%
ON800622	192	156	113	175	30.19%	24.53%	17.77%	27.52%
ON144454	195	152	95	193	30.66%	23.90%	14.94%	30.35%
ON330593	195	152	95	193	30.66%	23.90%	14.94%	30.35%
ON594981	195	152	95	193	30.66%	23.90%	14.94%	30.35%
ON650495	195	152	95	193	30.66%	23.90%	14.94%	30.35%
ON980049	195	152	95	193	30.66%	23.90%	14.94%	30.35%
OP129015	195	151	95	194	30.66%	23.74%	14.94%	30.50%
OP294029	196	152	95	192	30.82%	23.90%	14.94%	30.19%
OP311263	195	152	95	193	30.66%	23.90%	14.94%	30.35%
OP313193	195	152	95	193	30.66%	23.90%	14.94%	30.35%
OP313256	196	152	95	192	30.82%	23.90%	14.94%	30.19%
OP313383	195	152	95	193	30.66%	23.90%	14.94%	30.35%
OP359817	195	152	95	193	30.66%	23.90%	14.94%	30.35%
OP768851	195	152	95	193	30.66%	23.90%	14.94%	30.35%
OP934590	195	152	96	192	30.66%	23.90%	15.09%	30.19%
OQ094825	184	163	113	174	28.93%	25.63%	17.77%	27.36%
OQ111935	181	159	113	167	28.46%	25.00%	17.77%	26.26%
OQ112611	181	159	113	167	28.46%	25.00%	17.77%	26.26%
OQ112616	181	159	113	167	28.46%	25.00%	17.77%	26.26%
OQ112619	181	159	113	167	28.46%	25.00%	17.77%	26.26%
OQ112624	181	159	113	167	28.46%	25.00%	17.77%	26.26%
OQ112639	181	159	113	167	28.46%	25.00%	17.77%	26.26%
OQ112647	181	159	113	167	28.46%	25.00%	17.77%	26.26%
OQ112649	181	159	113	167	28.46%	25.00%	17.77%	26.26%
OQ112673	181	159	113	167	28.46%	25.00%	17.77%	26.26%
OQ112686	181	159	113	167	28.46%	25.00%	17.77%	26.26%
OQ112691	181	159	113	167	28.46%	25.00%	17.77%	26.26%
OQ112730	181	159	113	167	28.46%	25.00%	17.77%	26.26%
OQ112737	181	159	113	167	28.46%	25.00%	17.77%	26.26%
OQ112740	181	159	113	167	28.46%	25.00%	17.77%	26.26%
OQ112771	181	159	113	167	28.46%	25.00%	17.77%	26.26%
OQ112893	181	159	113	167	28.46%	25.00%	17.77%	26.26%

Table H.3: Frame 8, 14, 20, ... Similar distribution as in the previous case. The presence of all four types of nucleotides is relatively balanced in all analyzed sequences.

ID	A	T	C	G	A (%)	T (%)	C (%)	G (%)
MZ558275	195	173	160	108	30.66%	27.20%	25.16%	16.98%
MZ280957	195	173	160	108	30.66%	27.20%	25.16%	16.98%
MZ817897	194	170	162	109	30.50%	26.73%	25.47%	17.14%
OK039706	194	171	163	108	30.50%	26.89%	25.63%	16.98%
OL791595	194	171	163	108	30.50%	26.89%	25.63%	16.98%
OL920359	194	171	163	108	30.50%	26.89%	25.63%	16.98%
OM059434	194	171	162	109	30.50%	26.89%	25.47%	17.14%
ON800622	194	170	162	110	30.50%	26.73%	25.47%	17.30%
ON144454	199	197	145	94	31.29%	30.97%	22.80%	14.78%
ON330593	199	197	145	94	31.29%	30.97%	22.80%	14.78%
ON594981	199	197	145	94	31.29%	30.97%	22.80%	14.78%
ON650495	199	197	145	94	31.29%	30.97%	22.80%	14.78%
ON980049	199	197	145	94	31.29%	30.97%	22.80%	14.78%
OP129015	200	197	145	93	31.45%	30.97%	22.80%	14.62%
OP294029	199	197	145	94	31.29%	30.97%	22.80%	14.78%
OP311263	199	197	145	94	31.29%	30.97%	22.80%	14.78%
OP313193	198	197	146	94	31.13%	30.97%	22.96%	14.78%
OP313256	198	198	144	95	31.13%	31.13%	22.64%	14.94%
OP313383	199	198	145	93	31.29%	31.13%	22.80%	14.62%
OP359817	199	198	144	94	31.29%	31.13%	22.64%	14.78%
OP768851	199	197	145	94	31.29%	30.97%	22.80%	14.78%
OP934590	199	196	147	93	31.29%	30.82%	23.11%	14.62%
OQ094825	194	179	153	108	30.50%	28.14%	24.06%	16.98%
OQ111935	190	176	149	106	29.87%	27.67%	23.43%	16.67%
OQ112611	189	176	149	107	29.72%	27.67%	23.43%	16.82%
OQ112616	189	176	149	107	29.72%	27.67%	23.43%	16.82%
OQ112619	189	176	149	107	29.72%	27.67%	23.43%	16.82%
OQ112624	189	176	149	107	29.72%	27.67%	23.43%	16.82%
OQ112639	189	176	149	107	29.72%	27.67%	23.43%	16.82%
OQ112647	189	176	149	107	29.72%	27.67%	23.43%	16.82%
OQ112649	189	176	149	107	29.72%	27.67%	23.43%	16.82%
OQ112673	189	176	149	107	29.72%	27.67%	23.43%	16.82%
OQ112686	189	176	149	107	29.72%	27.67%	23.43%	16.82%
OQ112691	189	176	149	107	29.72%	27.67%	23.43%	16.82%
OQ112730	189	176	149	107	29.72%	27.67%	23.43%	16.82%
OQ112737	189	176	149	107	29.72%	27.67%	23.43%	16.82%
OQ112740	189	176	149	107	29.72%	27.67%	23.43%	16.82%
OQ112771	189	176	149	107	29.72%	27.67%	23.43%	16.82%
OQ112893	189	176	149	107	29.72%	27.67%	23.43%	16.82%

Table H.4: Frame 9, 15, 21, ... Just like in the case of the Arithmetic Series 6, 12, 18, ..., the presence of T at these sixth positions is significantly higher, accounting for nearly half compared to the other nucleotide types. We are observing the same phenomenon in both series.

ID	A	T	C	G	A (%)	T (%)	C (%)	G (%)
MZ558275	159	307	96	74	25,00%	48,27%	15,09%	11,64%
MZ280957	159	307	96	74	25.00%	48.27%	15.09%	11.64%
MZ817897	159	307	95	74	25.00%	48.27%	14.94%	11.64%
OK039706	159	307	96	74	25.00%	48.27%	15.09%	11.64%
OL791595	159	307	96	74	25.00%	48.27%	15.09%	11.64%
OL920359	159	307	96	74	25.00%	48.27%	15.09%	11.64%
OM059434	159	307	96	74	25.00%	48.27%	15.09%	11.64%
ON800622	160	305	96	75	25.16%	47.96%	15.09%	11.79%
ON144454	187	282	105	61	29.40%	44.34%	16.51%	9.59%
ON330593	187	281	106	61	29.40%	44.18%	16.67%	9.59%
ON594981	187	281	106	61	29.40%	44.18%	16.67%	9.59%
ON650495	187	281	106	61	29.40%	44.18%	16.67%	9.59%
ON980049	187	281	106	61	29.40%	44.18%	16.67%	9.59%
OP129015	187	282	106	60	29.40%	44.34%	16.67%	9.43%
OP294029	187	282	106	60	29.40%	44.34%	16.67%	9.43%
OP311263	187	282	105	61	29.40%	44.34%	16.51%	9.59%
OP313193	188	282	105	60	29.56%	44.34%	16.51%	9.43%
OP313256	187	282	106	60	29.40%	44.34%	16.67%	9.43%
OP313383	188	282	105	60	29.56%	44.34%	16.51%	9.43%
OP359817	188	282	105	60	29.56%	44.34%	16.51%	9.43%
OP768851	187	281	106	61	29.40%	44.18%	16.67%	9.59%
OP934590	187	280	106	62	29.40%	44.03%	16.67%	9.75%
OQ094825	164	296	97	77	25.79%	46.54%	15.25%	12.11%
OQ111935	163	286	96	76	25.63%	44.97%	15.09%	11.95%
OQ112611	163	286	96	76	25.63%	44.97%	15.09%	11.95%
OQ112616	163	286	96	76	25.63%	44.97%	15.09%	11.95%
OQ112619	163	286	96	76	25.63%	44.97%	15.09%	11.95%
OQ112624	163	286	96	76	25.63%	44.97%	15.09%	11.95%
OQ112639	163	286	96	76	25.63%	44.97%	15.09%	11.95%
OQ112647	163	286	96	76	25.63%	44.97%	15.09%	11.95%
OQ112649	163	286	96	76	25.63%	44.97%	15.09%	11.95%
OQ112673	163	286	96	76	25.63%	44.97%	15.09%	11.95%
OQ112686	163	286	96	76	25.63%	44.97%	15.09%	11.95%
OQ112691	163	286	96	76	25.63%	44.97%	15.09%	11.95%
OQ112730	163	286	96	76	25.63%	44.97%	15.09%	11.95%
OQ112737	163	286	96	76	25.63%	44.97%	15.09%	11.95%
OQ112740	163	286	96	76	25.63%	44.97%	15.09%	11.95%
OQ112771	163	286	96	76	25.63%	44.97%	15.09%	11.95%

Table H.4 – Continuation from the previous page

ID	A	T	C	G	A (%)	T (%)	C (%)	G (%)
OQ112893	163	286	96	76	25.63%	44.97%	15.09%	11.95%

Table H.5: Frame 10, 16, 22, ... The presence of all four types of nucleotides is once again balanced in all sequences.

ID	A	T	C	G	A (%)	T (%)	C (%)	G (%)
MZ558275	194	153	96	193	30.50%	24.06%	15.09%	30.35%
MZ280957	194	153	96	193	30.50%	24.06%	15.09%	30.35%
MZ817897	194	153	96	192	30.50%	24.06%	15.09%	30.19%
OK039706	195	153	96	192	30.66%	24.06%	15.09%	30.19%
OL791595	195	153	96	192	30.66%	24.06%	15.09%	30.19%
OL920359	195	153	96	192	30.66%	24.06%	15.09%	30.19%
OM059434	194	153	96	193	30.50%	24.06%	15.09%	30.35%
ON800622	194	153	96	193	30.50%	24.06%	15.09%	30.35%
ON144454	191	157	113	173	30.03%	24.69%	17.77%	27.20%
ON330593	191	158	113	172	30.03%	24.84%	17.77%	27.04%
ON594981	191	157	113	173	30.03%	24.69%	17.77%	27.20%
ON650495	191	157	113	173	30.03%	24.69%	17.77%	27.20%
ON980049	191	157	113	173	30.03%	24.69%	17.77%	27.20%
OP129015	190	156	114	174	29.87%	24.53%	17.92%	27.36%
OP294029	190	158	112	174	29.87%	24.84%	17.61%	27.36%
OP311263	191	157	113	173	30.03%	24.69%	17.77%	27.20%
OP313193	191	157	113	173	30.03%	24.69%	17.77%	27.20%
OP313256	190	158	112	174	29.87%	24.84%	17.61%	27.36%
OP313383	191	157	112	174	30.03%	24.69%	17.61%	27.36%
OP359817	190	157	113	174	29.87%	24.69%	17.77%	27.36%
OP768851	191	157	113	173	30.03%	24.69%	17.77%	27.20%
OP934590	190	157	114	173	29.87%	24.69%	17.92%	27.20%
OQ094825	204	145	95	190	32.08%	22.80%	14.94%	29.87%
OQ111935	199	143	92	187	31.29%	22.48%	14.47%	29.40%
OQ112611	199	143	92	187	31.29%	22.48%	14.47%	29.40%
OQ112616	199	143	92	187	31.29%	22.48%	14.47%	29.40%
OQ112619	199	143	92	187	31.29%	22.48%	14.47%	29.40%
OQ112624	199	143	92	187	31.29%	22.48%	14.47%	29.40%
OQ112639	199	143	92	187	31.29%	22.48%	14.47%	29.40%
OQ112647	199	143	92	187	31.29%	22.48%	14.47%	29.40%
OQ112649	199	143	92	187	31.29%	22.48%	14.47%	29.40%
OQ112673	199	143	92	187	31.29%	22.48%	14.47%	29.40%
OQ112686	199	143	92	187	31.29%	22.48%	14.47%	29.40%
OQ112691	199	143	92	187	31.29%	22.48%	14.47%	29.40%

Table H.5 – Continuation from the previous page

ID	A	T	C	G	A (%)	T (%)	C (%)	G (%)
OQ112730	199	143	92	187	31.29%	22.48%	14.47%	29.40%
OQ112737	199	143	92	187	31.29%	22.48%	14.47%	29.40%
OQ112740	199	143	92	187	31.29%	22.48%	14.47%	29.40%
OQ112771	199	143	92	187	31.29%	22.48%	14.47%	29.40%
OQ112893	199	143	92	187	31.29%	22.48%	14.47%	29.40%

Table H.6: Frame 11, 17, 23, ... Similar to all arithmetic series that are not multiples of 3, the presence of the four types of nucleotides is more balanced than in the case of multiples of 3.

ID	A	T	C	G	A (%)	T (%)	C (%)	G (%)
MZ558275	198	199	147	92	31.13%	31.29%	23.11%	14.47%
MZ280957	198	199	147	92	31.13%	31.29%	23.11%	14.47%
MZ817897	197	199	148	91	30.97%	31.29%	23.27%	14.31%
OK039706	199	198	147	92	31.29%	31.13%	23.11%	14.47%
OL791595	199	198	147	92	31.29%	31.13%	23.11%	14.47%
OL920359	200	198	147	91	31.45%	31.13%	23.11%	14.31%
OM059434	198	199	148	91	31.13%	31.29%	23.27%	14.31%
ON800622	198	199	148	91	31.13%	31.29%	23.27%	14.31%
ON144454	195	174	160	105	30.66%	27.36%	25.16%	16.51%
ON330593	195	174	160	105	30.66%	27.36%	25.16%	16.51%
ON594981	195	174	160	105	30.66%	27.36%	25.16%	16.51%
ON650495	195	174	160	105	30.66%	27.36%	25.16%	16.51%
ON980049	195	174	160	105	30.66%	27.36%	25.16%	16.51%
OP129015	193	173	162	106	30.35%	27.20%	25.47%	16.67%
OP294029	195	172	162	105	30.66%	27.04%	25.47%	16.51%
OP311263	195	174	160	105	30.66%	27.36%	25.16%	16.51%
OP313193	195	174	160	105	30.66%	27.36%	25.16%	16.51%
OP313256	194	172	162	106	30.50%	27.04%	25.47%	16.67%
OP313383	195	174	160	105	30.66%	27.36%	25.16%	16.51%
OP359817	195	174	160	105	30.66%	27.36%	25.16%	16.51%
OP768851	195	174	160	105	30.66%	27.36%	25.16%	16.51%
OP934590	196	174	160	104	30.82%	27.36%	25.16%	16.35%
OQ094825	201	188	157	88	31.60%	29.56%	24.69%	13.84%
OQ111935	196	185	154	85	30.82%	29.09%	24.21%	13.36%
OQ112611	196	185	154	85	30.82%	29.09%	24.21%	13.36%
OQ112616	196	185	154	85	30.82%	29.09%	24.21%	13.36%
OQ112619	196	185	154	85	30.82%	29.09%	24.21%	13.36%
OQ112624	196	185	154	85	30.82%	29.09%	24.21%	13.36%
OQ112639	196	185	154	85	30.82%	29.09%	24.21%	13.36%

Table H.6 – Continuation from the previous page

ID	**A**	**T**	**C**	**G**	**A** (%)	**T** (%)	**C** (%)	**G** (%)
OQ112647	196	185	154	85	30.82%	29.09%	24.21%	13.36%
OQ112649	196	185	154	85	30.82%	29.09%	24.21%	13.36%
OQ112673	196	185	154	85	30.82%	29.09%	24.21%	13.36%
OQ112686	196	185	154	85	30.82%	29.09%	24.21%	13.36%
OQ112691	196	185	154	85	30.82%	29.09%	24.21%	13.36%
OQ112730	196	185	154	85	30.82%	29.09%	24.21%	13.36%
OQ112737	196	185	154	85	30.82%	29.09%	24.21%	13.36%
OQ112740	196	185	154	85	30.82%	29.09%	24.21%	13.36%
OQ112771	196	185	154	85	30.82%	29.09%	24.21%	13.36%
OQ112893	196	185	154	85	30.82%	29.09%	24.21%	13.36%

Perfect numbers are rare just like perfect men.

— Descartes, *In his work "Meditationes de prima Philosophia" (1641), he developed the demonstration of the existence of God and the immortality of the soul. In 1663, this book was included in the "Index librorum prohibitorum" due to Church censorship.*

I

SARS-CoV-2 Recombinant Sequences Study Case. Statistical Data on the Types of Nucleotides Present in the Arithmetic Series (Arithmetic Ratio = 6) at M Protein

Each row in every table represents a recombinant sequence, accurately classified by VGG-16 SARS-CoV-2 with a high level of confidence and a clearly located hot zone in the Spyke protein at $f = 1/6$.

Table I.1: Frame 6, 12, 18, ... In all analyzed sequences, the proportions of the presence of all four types of nucleotides are balanced in all sequences.

ID	A	T	C	G	A (%)	T (%)	C (%)	G (%)
MZ558275	23	36	29	23	20.72%	32.43%	26.13%	20.72%
MZ280957	23	36	29	23	20.72%	32.43%	26.13%	20.72%
MZ817897	23	36	29	23	20.72%	32.43%	26.13%	20.72%
OK039706	23	36	29	23	20.72%	32.43%	26.13%	20.72%
OL791595	23	36	29	23	20.72%	32.43%	26.13%	20.72%
OL920359	23	36	29	23	20.72%	32.43%	26.13%	20.72%
OM059434	23	36	29	23	20.72%	32.43%	26.13%	20.72%
ON144454	23	37	28	23	20.72%	33.33%	25.23%	20.72%
ON330593	23	37	28	23	20.72%	33.33%	25.23%	20.72%
ON594981	23	36	29	23	20.72%	32.43%	26.13%	20.72%
ON650495	23	36	29	23	20.72%	32.43%	26.13%	20.72%
ON800622	23	36	29	23	20.72%	32.43%	26.13%	20.72%

Table I.1 – Continuation from the previous page

ID	A	T	C	G	A (%)	T (%)	C (%)	G (%)
ON980049	23	37	28	23	20.72%	33.33%	25.23%	20.72%
OP129015	23	36	29	23	20.72%	32.43%	26.13%	20.72%
OP294029	23	37	28	23	20.72%	33.33%	25.23%	20.72%
OP311263	23	37	28	23	20.72%	33.33%	25.23%	20.72%
OP313193	23	36	29	23	20.72%	32.43%	26.13%	20.72%
OP313256	23	36	29	23	20.72%	32.43%	26.13%	20.72%
OP313383	23	36	29	23	20.72%	32.43%	26.13%	20.72%
OP359817	23	36	29	23	20.72%	32.43%	26.13%	20.72%
OP768851	23	37	28	23	20.72%	33.33%	25.23%	20.72%
OP934590	22	36	29	24	19.82%	32.43%	26.13%	21.62%
OQ094825	23	36	29	23	20.72%	32.43%	26.13%	20.72%
OQ111935	23	37	28	23	20.72%	33.33%	25.23%	20.72%
OQ112611	23	37	28	23	20.72%	33.33%	25.23%	20.72%
OQ112616	23	37	28	23	20.72%	33.33%	25.23%	20.72%
OQ112619	23	37	28	23	20.72%	33.33%	25.23%	20.72%
OQ112624	23	37	28	23	20.72%	33.33%	25.23%	20.72%
OQ112639	23	37	28	23	20.72%	33.33%	25.23%	20.72%
OQ112647	23	37	28	23	20.72%	33.33%	25.23%	20.72%
OQ112649	23	37	28	23	20.72%	33.33%	25.23%	20.72%
OQ112673	23	37	28	23	20.72%	33.33%	25.23%	20.72%
OQ112686	23	37	28	23	20.72%	33.33%	25.23%	20.72%
OQ112691	23	37	28	23	20.72%	33.33%	25.23%	20.72%
OQ112730	23	37	28	23	20.72%	33.33%	25.23%	20.72%
OQ112737	23	37	28	23	20.72%	33.33%	25.23%	20.72%
OQ112740	23	37	28	23	20.72%	33.33%	25.23%	20.72%
OQ112771	23	37	28	23	20.72%	33.33%	25.23%	20.72%
OQ112893	23	37	28	23	20.72%	33.33%	25.23%	20.72%

Table I.2: Frame 7, 13, 19, ... The proportions of the presence of all four types of nucleotides are balanced in all sequences.

ID	A	T	C	G	A (%)	T (%)	C (%)	G (%)
MZ558275	31	28	22	30	27.93%	25.23%	19.82%	27.03%
MZ280957	31	28	22	30	27.93%	25.23%	19.82%	27.03%
MZ817897	31	28	22	30	27.93%	25.23%	19.82%	27.03%
OK039706	31	28	22	30	27.93%	25.23%	19.82%	27.03%
OL791595	31	28	22	30	27.93%	25.23%	19.82%	27.03%
OL920359	31	28	22	30	27.93%	25.23%	19.82%	27.03%
OM059434	31	28	22	30	27.93%	25.23%	19.82%	27.03%
ON144454	32	28	21	30	28.83%	25.23%	18.92%	27.03%
ON330593	32	28	21	30	28.83%	25.23%	18.92%	27.03%
ON594981	32	28	21	30	28.83%	25.23%	18.92%	27.03%
ON650495	32	28	21	30	28.83%	25.23%	18.92%	27.03%
ON800622	31	28	22	30	27.93%	25.23%	19.82%	27.03%
ON980049	32	28	21	30	28.83%	25.23%	18.92%	27.03%
OP129015	31	28	22	30	27.93%	25.23%	19.82%	27.03%
OP294029	32	28	22	29	28.83%	25.23%	19.82%	26.13%
OP311263	32	28	21	30	28.83%	25.23%	18.92%	27.03%
OP313193	31	28	21	31	27.93%	25.23%	18.92%	27.93%
OP313256	31	28	22	30	27.93%	25.23%	19.82%	27.03%
OP313383	31	28	22	30	27.93%	25.23%	19.82%	27.03%
OP359817	31	28	22	30	27.93%	25.23%	19.82%	27.03%
OP768851	32	28	21	30	28.83%	25.23%	18.92%	27.03%
OP934590	33	28	21	29	29.73%	25.23%	18.92%	26.13%
OQ094825	31	28	21	31	27.93%	25.23%	18.92%	27.93%
OQ111935	32	28	21	30	28.83%	25.23%	18.92%	27.03%
OQ112611	32	28	21	30	28.83%	25.23%	18.92%	27.03%
OQ112616	32	28	21	30	28.83%	25.23%	18.92%	27.03%
OQ112619	32	28	21	30	28.83%	25.23%	18.92%	27.03%
OQ112624	32	28	21	30	28.83%	25.23%	18.92%	27.03%
OQ112639	32	28	21	30	28.83%	25.23%	18.92%	27.03%
OQ112647	32	28	21	30	28.83%	25.23%	18.92%	27.03%
OQ112649	32	28	21	30	28.83%	25.23%	18.92%	27.03%
OQ112673	32	28	21	30	28.83%	25.23%	18.92%	27.03%
OQ112686	32	28	21	30	28.83%	25.23%	18.92%	27.03%
OQ112691	32	28	21	30	28.83%	25.23%	18.92%	27.03%
OQ112730	32	28	21	30	28.83%	25.23%	18.92%	27.03%
OQ112737	32	28	21	30	28.83%	25.23%	18.92%	27.03%
OQ112740	32	28	21	30	28.83%	25.23%	18.92%	27.03%
OQ112771	32	28	21	30	28.83%	25.23%	18.92%	27.03%
OQ112893	32	28	21	30	28.83%	25.23%	18.92%	27.03%

Table I.3: Frame 8, 14, 20, ... The proportions of the presence of all four types of nucleotides are balanced in all sequences.

ID	A	T	C	G	A (%)	T (%)	C (%)	G (%)
MZ558275	31	37	24	19	27.93%	33.33%	21.62%	17.12%
MZ280957	31	37	24	19	27.93%	33.33%	21.62%	17.12%
MZ817897	31	37	24	19	27.93%	33.33%	21.62%	17.12%
OK039706	31	37	24	19	27.93%	33.33%	21.62%	17.12%
OL791595	31	37	24	19	27.93%	33.33%	21.62%	17.12%
OL920359	31	37	24	19	27.93%	33.33%	21.62%	17.12%
OM059434	31	37	24	19	27.93%	33.33%	21.62%	17.12%
ON144454	31	37	24	19	27.93%	33.33%	21.62%	17.12%
ON330593	31	37	24	19	27.93%	33.33%	21.62%	17.12%
ON594981	30	37	24	20	27.03%	33.33%	21.62%	18.02%
ON650495	30	37	24	20	27.03%	33.33%	21.62%	18.02%
ON800622	31	37	24	19	27.93%	33.33%	21.62%	17.12%
ON980049	31	37	24	19	27.93%	33.33%	21.62%	17.12%
OP129015	31	37	24	19	27.93%	33.33%	21.62%	17.12%
OP294029	31	37	24	19	27.93%	33.33%	21.62%	17.12%
OP311263	31	37	24	19	27.93%	33.33%	21.62%	17.12%
OP313193	31	37	24	19	27.93%	33.33%	21.62%	17.12%
OP313256	31	37	24	19	27.93%	33.33%	21.62%	17.12%
OP313383	31	37	24	19	27.93%	33.33%	21.62%	17.12%
OP359817	31	37	24	19	27.93%	33.33%	21.62%	17.12%
OP768851	31	37	24	19	27.93%	33.33%	21.62%	17.12%
OP934590	31	37	24	19	27.93%	33.33%	21.62%	17.12%
OQ094825	31	37	24	19	27.93%	33.33%	21.62%	17.12%
OQ111935	31	37	24	19	27.93%	33.33%	21.62%	17.12%
OQ112611	31	37	24	19	27.93%	33.33%	21.62%	17.12%
OQ112616	31	37	24	19	27.93%	33.33%	21.62%	17.12%
OQ112619	31	37	24	19	27.93%	33.33%	21.62%	17.12%
OQ112624	31	37	24	19	27.93%	33.33%	21.62%	17.12%
OQ112639	31	37	24	19	27.93%	33.33%	21.62%	17.12%
OQ112647	31	37	24	19	27.93%	33.33%	21.62%	17.12%
OQ112649	31	37	24	19	27.93%	33.33%	21.62%	17.12%
OQ112673	31	37	24	19	27.93%	33.33%	21.62%	17.12%
OQ112686	31	37	24	19	27.93%	33.33%	21.62%	17.12%
OQ112691	31	37	24	19	27.93%	33.33%	21.62%	17.12%
OQ112730	31	37	24	19	27.93%	33.33%	21.62%	17.12%
OQ112737	31	37	24	19	27.93%	33.33%	21.62%	17.12%
OQ112740	31	37	24	19	27.93%	33.33%	21.62%	17.12%
OQ112771	31	37	24	19	27.93%	33.33%	21.62%	17.12%
OQ112893	31	37	24	19	27.93%	33.33%	21.62%	17.12%

Table I.4: Frame 9, 15, 21, ... The proportions of the presence of all four types of nucleotides are balanced in all sequences.

ID	A	T	C	G	A (%)	T (%)	C (%)	G (%)
MZ558275	31	46	22	12	27.93%	41.44%	19.82%	10.81%
MZ280957	31	46	22	12	27.93%	41.44%	19.82%	10.81%
MZ817897	31	46	22	12	27.93%	41.44%	19.82%	10.81%
OK039706	31	46	22	12	27.93%	41.44%	19.82%	10.81%
OL791595	31	46	22	12	27.93%	41.44%	19.82%	10.81%
OL920359	31	46	22	12	27.93%	41.44%	19.82%	10.81%
OM059434	31	46	22	12	27.93%	41.44%	19.82%	10.81%
ON144454	31	46	22	12	27.93%	41.44%	19.82%	10.81%
ON330593	31	46	22	12	27.93%	41.44%	19.82%	10.81%
ON594981	31	46	22	12	27.93%	41.44%	19.82%	10.81%
ON650495	31	46	22	12	27.93%	41.44%	19.82%	10.81%
ON800622	31	46	22	12	27.93%	41.44%	19.82%	10.81%
ON980049	31	47	21	12	27.93%	42.34%	18.92%	10.81%
OP129015	31	46	22	12	27.93%	41.44%	19.82%	10.81%
OP294029	31	46	22	12	27.93%	41.44%	19.82%	10.81%
OP311263	31	46	22	12	27.93%	41.44%	19.82%	10.81%
OP313193	31	46	22	12	27.93%	41.44%	19.82%	10.81%
OP313256	31	46	22	12	27.93%	41.44%	19.82%	10.81%
OP313383	31	46	22	12	27.93%	41.44%	19.82%	10.81%
OP359817	31	46	22	12	27.93%	41.44%	19.82%	10.81%
OP768851	31	46	22	12	27.93%	41.44%	19.82%	10.81%
OP934590	31	46	22	12	27.93%	41.44%	19.82%	10.81%
OQ094825	31	46	22	12	27.93%	41.44%	19.82%	10.81%
OQ111935	31	46	22	12	27.93%	41.44%	19.82%	10.81%
OQ112611	31	46	22	12	27.93%	41.44%	19.82%	10.81%
OQ112616	31	46	22	12	27.93%	41.44%	19.82%	10.81%
OQ112619	31	46	22	12	27.93%	41.44%	19.82%	10.81%
OQ112624	31	46	22	12	27.93%	41.44%	19.82%	10.81%
OQ112639	31	46	22	12	27.93%	41.44%	19.82%	10.81%
OQ112647	31	46	22	12	27.93%	41.44%	19.82%	10.81%
OQ112649	31	46	22	12	27.93%	41.44%	19.82%	10.81%
OQ112673	31	46	22	12	27.93%	41.44%	19.82%	10.81%
OQ112686	31	46	22	12	27.93%	41.44%	19.82%	10.81%
OQ112691	31	46	22	12	27.93%	41.44%	19.82%	10.81%
OQ112730	31	46	22	12	27.93%	41.44%	19.82%	10.81%
OQ112737	31	46	22	12	27.93%	41.44%	19.82%	10.81%
OQ112740	31	46	22	12	27.93%	41.44%	19.82%	10.81%
OQ112771	31	46	22	12	27.93%	41.44%	19.82%	10.81%
OQ112893	31	46	22	12	27.93%	41.44%	19.82%	10.81%

Table I.5: Frame 10, 16, 22, ... The proportions of the presence of all four types of nucleotides are balanced in all sequences.

ID	A	T	C	G	A (%)	T (%)	C (%)	G (%)
MZ558275	35	21	27	27	31.82%	19.09%	24.55%	24.55%
MZ280957	35	21	27	27	31.82%	19.09%	24.55%	24.55%
MZ817897	35	21	27	27	31.82%	19.09%	24.55%	24.55%
OK039706	35	21	27	27	31.82%	19.09%	24.55%	24.55%
OL791595	35	21	27	27	31.82%	19.09%	24.55%	24.55%
OL920359	35	21	27	27	31.82%	19.09%	24.55%	24.55%
OM059434	35	21	27	27	31.82%	19.09%	24.55%	24.55%
ON144454	35	21	27	27	31.82%	19.09%	24.55%	24.55%
ON330593	35	21	27	27	31.82%	19.09%	24.55%	24.55%
ON594981	35	21	27	27	31.82%	19.09%	24.55%	24.55%
ON650495	35	21	27	27	31.82%	19.09%	24.55%	24.55%
ON800622	35	21	27	27	31.82%	19.09%	24.55%	24.55%
ON980049	35	21	27	27	31.82%	19.09%	24.55%	24.55%
OP129015	35	21	27	27	31.82%	19.09%	24.55%	24.55%
OP294029	35	21	27	27	31.82%	19.09%	24.55%	24.55%
OP311263	35	21	27	27	31.82%	19.09%	24.55%	24.55%
OP313193	35	21	27	27	31.82%	19.09%	24.55%	24.55%
OP313256	35	21	27	27	31.82%	19.09%	24.55%	24.55%
OP313383	35	21	27	27	31.82%	19.09%	24.55%	24.55%
OP359817	35	21	27	27	31.82%	19.09%	24.55%	24.55%
OP768851	35	21	27	27	31.82%	19.09%	24.55%	24.55%
OP934590	35	21	27	27	31.82%	19.09%	24.55%	24.55%
OQ094825	35	21	27	27	31.82%	19.09%	24.55%	24.55%
OQ111935	35	21	27	27	31.82%	19.09%	24.55%	24.55%
OQ112611	35	21	27	27	31.82%	19.09%	24.55%	24.55%
OQ112616	35	21	27	27	31.82%	19.09%	24.55%	24.55%
OQ112619	35	21	27	27	31.82%	19.09%	24.55%	24.55%
OQ112624	35	21	27	27	31.82%	19.09%	24.55%	24.55%
OQ112639	35	21	27	27	31.82%	19.09%	24.55%	24.55%
OQ112647	35	21	27	27	31.82%	19.09%	24.55%	24.55%
OQ112649	35	21	27	27	31.82%	19.09%	24.55%	24.55%
OQ112673	35	21	27	27	31.82%	19.09%	24.55%	24.55%
OQ112686	35	21	27	27	31.82%	19.09%	24.55%	24.55%
OQ112691	35	21	27	27	31.82%	19.09%	24.55%	24.55%
OQ112730	35	21	27	27	31.82%	19.09%	24.55%	24.55%
OQ112737	35	21	27	27	31.82%	19.09%	24.55%	24.55%
OQ112740	35	21	27	27	31.82%	19.09%	24.55%	24.55%
OQ112771	35	21	27	27	31.82%	19.09%	24.55%	24.55%
OQ112893	35	21	27	27	31.82%	19.09%	24.55%	24.55%

Table I.6: Frame 11, 17, 23, ... The proportions of the presence of all four types of nucleotides are balanced in all sequences.

ID	A	T	C	G	A (%)	T (%)	C (%)	G (%)
MZ558275	19	44	21	26	17.27%	40.00%	19.09%	23.64%
MZ280957	19	44	21	26	17.27%	40.00%	19.09%	23.64%
MZ817897	19	44	21	26	17.27%	40.00%	19.09%	23.64%
OK039706	19	43	22	26	17.27%	39.09%	20.00%	23.64%
OL791595	19	43	22	26	17.27%	39.09%	20.00%	23.64%
OL920359	19	43	22	26	17.27%	39.09%	20.00%	23.64%
OM059434	19	44	21	26	17.27%	40.00%	19.09%	23.64%
ON144454	19	44	21	26	17.27%	40.00%	19.09%	23.64%
ON330593	19	44	21	26	17.27%	40.00%	19.09%	23.64%
ON594981	19	44	21	26	17.27%	40.00%	19.09%	23.64%
ON650495	19	44	21	26	17.27%	40.00%	19.09%	23.64%
ON800622	19	44	21	26	17.27%	40.00%	19.09%	23.64%
ON980049	19	44	21	26	17.27%	40.00%	19.09%	23.64%
OP129015	19	44	21	26	17.27%	40.00%	19.09%	23.64%
OP294029	19	44	21	26	17.27%	40.00%	19.09%	23.64%
OP311263	19	44	21	26	17.27%	40.00%	19.09%	23.64%
OP313193	19	44	21	26	17.27%	40.00%	19.09%	23.64%
OP313256	19	44	21	26	17.27%	40.00%	19.09%	23.64%
OP313383	19	44	21	26	17.27%	40.00%	19.09%	23.64%
OP359817	19	44	21	26	17.27%	40.00%	19.09%	23.64%
OP768851	19	44	21	26	17.27%	40.00%	19.09%	23.64%
OP934590	19	44	21	26	17.27%	40.00%	19.09%	23.64%
OQ094825	19	44	21	26	17.27%	40.00%	19.09%	23.64%
OQ111935	19	44	21	26	17.27%	40.00%	19.09%	23.64%
OQ112611	19	44	21	26	17.27%	40.00%	19.09%	23.64%
OQ112616	19	44	21	26	17.27%	40.00%	19.09%	23.64%
OQ112619	19	44	21	26	17.27%	40.00%	19.09%	23.64%
OQ112624	19	44	21	26	17.27%	40.00%	19.09%	23.64%
OQ112639	19	44	21	26	17.27%	40.00%	19.09%	23.64%
OQ112647	19	44	21	26	17.27%	40.00%	19.09%	23.64%
OQ112649	19	44	21	26	17.27%	40.00%	19.09%	23.64%
OQ112673	19	44	21	26	17.27%	40.00%	19.09%	23.64%
OQ112686	19	44	21	26	17.27%	40.00%	19.09%	23.64%
OQ112691	19	44	21	26	17.27%	40.00%	19.09%	23.64%
OQ112730	19	44	21	26	17.27%	40.00%	19.09%	23.64%
OQ112737	19	44	21	26	17.27%	40.00%	19.09%	23.64%
OQ112740	19	44	21	26	17.27%	40.00%	19.09%	23.64%
OQ112771	19	44	21	26	17.27%	40.00%	19.09%	23.64%
OQ112893	19	44	21	26	17.27%	40.00%	19.09%	23.64%

The best crimes for my novels have occurred to me while washing dishes.

— Agatha Christie, *she worked as a nurse, where she became an expert in poisons, including necessary doses, symptoms, consequences, etc. She applied this knowledge in many of her novels, using poisons in numerous literary murders.*

J SARS-CoV-2 Recombinant Sequences Study Case. Statistical Data on the Types of Nucleotides Present in the Arithmetic Series (Arithmetic Ratio = 6) at N Protein

Each row in every table represents a recombinant sequence, accurately classified by VGG-16 SARS-CoV-2 with a high level of confidence and a clearly located hot zone in the Spyke protein at $f = 1/6$.

Table J.1: Frame 6, 12, 18, ... In all analyzed sequences, the proportions of the presence of all four types of nucleotides are balanced in all sequences.

ID	A	T	C	G	A (%)	T (%)	C (%)	G (%)
MZ558275	64	61	58	27	30.48%	29.05%	27.62%	12.86%
MZ280957	64	61	58	27	30.48%	29.05%	27.62%	12.86%
MZ817897	64	61	58	27	30.48%	29.05%	27.62%	12.86%
OK039706	64	61	58	27	30.48%	29.05%	27.62%	12.86%
OL791595	64	61	58	27	30.48%	29.05%	27.62%	12.86%
OL920359	64	61	58	27	30.48%	29.05%	27.62%	12.86%
OM059434	64	61	58	27	30.48%	29.05%	27.62%	12.86%
ON144454	61	74	37	36	29.33%	35.58%	17.79%	17.31%
ON330593	61	74	37	36	29.33%	35.58%	17.79%	17.31%
ON594981	61	74	37	36	29.33%	35.58%	17.79%	17.31%
ON650495	61	74	37	36	29.33%	35.58%	17.79%	17.31%
ON800622	61	74	37	36	29.33%	35.58%	17.79%	17.31%

Table J.1 – Continuation from the previous page

ID	**A**	**T**	**C**	**G**	**A** (%)	**T** (%)	**C** (%)	**G** (%)
ON980049	61	74	37	36	29.33%	35.58%	17.79%	17.31%
OP129015	61	75	36	36	29.33%	36.06%	17.31%	17.31%
OP294029	61	74	37	36	29.33%	35.58%	17.79%	17.31%
OP311263	61	74	37	36	29.33%	35.58%	17.79%	17.31%
OP313193	61	74	37	36	29.33%	35.58%	17.79%	17.31%
OP313256	61	74	37	36	29.33%	35.58%	17.79%	17.31%
OP313383	61	74	37	36	29.33%	35.58%	17.79%	17.31%
OP359817	61	74	37	36	29.33%	35.58%	17.79%	17.31%
OP768851	61	74	37	36	29.33%	35.58%	17.79%	17.31%
OP934590	61	74	37	36	29.33%	35.58%	17.79%	17.31%
OQ094825	61	74	37	36	29.33%	35.58%	17.79%	17.31%
OQ111935	61	74	37	36	29.33%	35.58%	17.79%	17.31%
OQ112611	61	74	37	36	29.33%	35.58%	17.79%	17.31%
OQ112616	61	74	37	36	29.33%	35.58%	17.79%	17.31%
OQ112619	61	74	37	36	29.33%	35.58%	17.79%	17.31%
OQ112624	61	74	37	36	29.33%	35.58%	17.79%	17.31%
OQ112639	61	74	37	36	29.33%	35.58%	17.79%	17.31%
OQ112647	61	74	37	36	29.33%	35.58%	17.79%	17.31%
OQ112649	61	74	37	36	29.33%	35.58%	17.79%	17.31%
OQ112673	61	74	37	36	29.33%	35.58%	17.79%	17.31%
OQ112686	61	74	37	36	29.33%	35.58%	17.79%	17.31%
OQ112691	61	74	37	36	29.33%	35.58%	17.79%	17.31%
OQ112730	61	74	37	36	29.33%	35.58%	17.79%	17.31%
OQ112737	61	74	37	36	29.33%	35.58%	17.79%	17.31%
OQ112740	61	74	37	36	29.33%	35.58%	17.79%	17.31%
OQ112771	61	74	37	36	29.33%	35.58%	17.79%	17.31%
OQ112893	61	74	37	36	29.33%	35.58%	17.79%	17.31%

Table J.2: Frame Series 7, 13, 19, ... The proportions of the presence of all four types of nucleotides are balanced in all sequences.

ID	A	T	C	G	A (%)	T (%)	C (%)	G (%)
MZ558275	71	32	49	57	33.97%	15.31%	23.44%	27.27%
MZ280957	71	32	49	57	33.97%	15.31%	23.44%	27.27%
MZ817897	72	31	49	57	34.45%	14.83%	23.44%	27.27%
OK039706	72	31	49	57	34.45%	14.83%	23.44%	27.27%
OL791595	72	32	49	56	34.45%	15.31%	23.44%	26.79%
OL920359	72	31	49	57	34.45%	14.83%	23.44%	27.27%
OM059434	72	33	49	55	34.45%	15.79%	23.44%	26.32%
ON144454	58	34	50	66	27.88%	16.35%	24.04%	31.73%
ON330593	58	34	50	66	27.88%	16.35%	24.04%	31.73%
ON594981	58	34	50	66	27.88%	16.35%	24.04%	31.73%
ON650495	58	34	50	66	27.88%	16.35%	24.04%	31.73%
ON800622	58	34	50	66	27.88%	16.35%	24.04%	31.73%
ON980049	58	34	50	66	27.88%	16.35%	24.04%	31.73%
OP129015	58	35	50	65	27.88%	16.83%	24.04%	31.25%
OP294029	58	34	50	66	27.88%	16.35%	24.04%	31.73%
OP311263	58	34	50	66	27.88%	16.35%	24.04%	31.73%
OP313193	58	34	50	66	27.88%	16.35%	24.04%	31.73%
OP313256	58	34	50	66	27.88%	16.35%	24.04%	31.73%
OP313383	58	34	50	66	27.88%	16.35%	24.04%	31.73%
OP359817	58	34	50	66	27.88%	16.35%	24.04%	31.73%
OP768851	58	34	50	66	27.88%	16.35%	24.04%	31.73%
OP934590	58	34	50	66	27.88%	16.35%	24.04%	31.73%
OQ094825	58	34	50	66	27.88%	16.35%	24.04%	31.73%
OQ111935	58	34	50	66	27.88%	16.35%	24.04%	31.73%
OQ112611	58	34	50	66	27.88%	16.35%	24.04%	31.73%
OQ112616	58	34	50	66	27.88%	16.35%	24.04%	31.73%
OQ112619	58	34	50	66	27.88%	16.35%	24.04%	31.73%
OQ112624	58	34	50	66	27.88%	16.35%	24.04%	31.73%
OQ112639	58	34	50	66	27.88%	16.35%	24.04%	31.73%
OQ112647	58	34	50	66	27.88%	16.35%	24.04%	31.73%
OQ112649	58	34	50	66	27.88%	16.35%	24.04%	31.73%
OQ112673	58	34	50	66	27.88%	16.35%	24.04%	31.73%
OQ112686	58	34	50	66	27.88%	16.35%	24.04%	31.73%
OQ112691	58	34	50	66	27.88%	16.35%	24.04%	31.73%
OQ112730	58	34	50	66	27.88%	16.35%	24.04%	31.73%
OQ112737	58	34	50	66	27.88%	16.35%	24.04%	31.73%
OQ112740	58	34	50	66	27.88%	16.35%	24.04%	31.73%
OQ112771	58	34	50	66	27.88%	16.35%	24.04%	31.73%
OQ112893	58	34	50	66	27.88%	16.35%	24.04%	31.73%

Table J.3: Frame 8, 14, 20, ... The proportions of the presence of all four types of nucleotides are balanced in all sequences.

ID	A	T	C	G	A (%)	T (%)	C (%)	G (%)
MZ558275	72	34	58	45	34.45%	16.27%	27.75%	21.53%
MZ280957	72	34	58	45	34.45%	16.27%	27.75%	21.53%
MZ817897	71	34	58	46	33.97%	16.27%	27.75%	22.01%
OK039706	71	34	58	46	33.97%	16.27%	27.75%	22.01%
OL791595	71	34	58	46	33.97%	16.27%	27.75%	22.01%
OL920359	71	34	58	46	33.97%	16.27%	27.75%	22.01%
OM059434	72	35	58	44	34.45%	16.75%	27.75%	21.05%
ON144454	70	37	57	44	33.65%	17.79%	27.40%	21.15%
ON330593	70	37	57	44	33.65%	17.79%	27.40%	21.15%
ON594981	70	37	57	44	33.65%	17.79%	27.40%	21.15%
ON650495	70	37	57	44	33.65%	17.79%	27.40%	21.15%
ON800622	70	37	57	44	33.65%	17.79%	27.40%	21.15%
ON980049	70	37	57	44	33.65%	17.79%	27.40%	21.15%
OP129015	70	37	57	44	33.65%	17.79%	27.40%	21.15%
OP294029	70	37	57	44	33.65%	17.79%	27.40%	21.15%
OP311263	70	37	57	44	33.65%	17.79%	27.40%	21.15%
OP313193	70	37	57	44	33.65%	17.79%	27.40%	21.15%
OP313256	70	37	57	44	33.65%	17.79%	27.40%	21.15%
OP313383	70	37	57	44	33.65%	17.79%	27.40%	21.15%
OP359817	70	37	57	44	33.65%	17.79%	27.40%	21.15%
OP768851	70	37	57	44	33.65%	17.79%	27.40%	21.15%
OP934590	70	37	57	44	33.65%	17.79%	27.40%	21.15%
OQ094825	70	37	57	44	33.65%	17.79%	27.40%	21.15%
OQ111935	70	37	57	44	33.65%	17.79%	27.40%	21.15%
OQ112611	70	37	57	44	33.65%	17.79%	27.40%	21.15%
OQ112616	70	37	57	44	33.65%	17.79%	27.40%	21.15%
OQ112619	70	37	57	44	33.65%	17.79%	27.40%	21.15%
OQ112624	70	37	57	44	33.65%	17.79%	27.40%	21.15%
OQ112639	70	37	57	44	33.65%	17.79%	27.40%	21.15%
OQ112647	70	37	57	44	33.65%	17.79%	27.40%	21.15%
OQ112649	70	37	57	44	33.65%	17.79%	27.40%	21.15%
OQ112673	70	37	57	44	33.65%	17.79%	27.40%	21.15%
OQ112686	70	37	57	44	33.65%	17.79%	27.40%	21.15%
OQ112691	70	37	57	44	33.65%	17.79%	27.40%	21.15%
OQ112730	70	37	57	44	33.65%	17.79%	27.40%	21.15%
OQ112737	70	37	57	44	33.65%	17.79%	27.40%	21.15%
OQ112740	70	37	57	44	33.65%	17.79%	27.40%	21.15%
OQ112771	70	37	57	44	33.65%	17.79%	27.40%	21.15%
OQ112893	70	37	57	44	33.65%	17.79%	27.40%	21.15%

Table J.4: Frame 9, 15, 21, ... The proportions of the presence of all four types of nucleotides are balanced in all sequences.

ID	A	T	C	G	A (%)	T (%)	C (%)	G (%)
MZ558275	63	72	37	37	30.14%	34.45%	17.70%	17.70%
MZ280957	63	72	37	37	30.14%	34.45%	17.70%	17.70%
MZ817897	64	72	37	36	30.62%	34.45%	17.70%	17.22%
OK039706	64	72	37	36	30.62%	34.45%	17.70%	17.22%
OL791595	64	72	37	36	30.62%	34.45%	17.70%	17.22%
OL920359	64	72	37	36	30.62%	34.45%	17.70%	17.22%
OM059434	64	72	37	36	30.62%	34.45%	17.70%	17.22%
ON144454	66	59	57	26	31.73%	28.37%	27.40%	12.50%
ON330593	66	59	57	26	31.73%	28.37%	27.40%	12.50%
ON594981	65	59	57	27	31.25%	28.37%	27.40%	12.98%
ON650495	65	59	57	27	31.25%	28.37%	27.40%	12.98%
ON800622	66	59	57	26	31.73%	28.37%	27.40%	12.50%
ON980049	66	59	57	26	31.73%	28.37%	27.40%	12.50%
OP129015	66	59	57	26	31.73%	28.37%	27.40%	12.50%
OP294029	66	59	57	26	31.73%	28.37%	27.40%	12.50%
OP311263	66	59	57	26	31.73%	28.37%	27.40%	12.50%
OP313193	66	59	57	26	31.73%	28.37%	27.40%	12.50%
OP313256	66	59	57	26	31.73%	28.37%	27.40%	12.50%
OP313383	66	59	57	26	31.73%	28.37%	27.40%	12.50%
OP359817	66	59	57	26	31.73%	28.37%	27.40%	12.50%
OP768851	66	59	57	26	31.73%	28.37%	27.40%	12.50%
OP934590	65	59	57	27	31.25%	28.37%	27.40%	12.98%
OQ094825	66	59	57	26	31.73%	28.37%	27.40%	12.50%
OQ111935	66	59	57	26	31.73%	28.37%	27.40%	12.50%
OQ112611	66	59	57	26	31.73%	28.37%	27.40%	12.50%
OQ112616	66	59	57	26	31.73%	28.37%	27.40%	12.50%
OQ112619	66	59	57	26	31.73%	28.37%	27.40%	12.50%
OQ112624	66	59	57	26	31.73%	28.37%	27.40%	12.50%
OQ112639	66	59	57	26	31.73%	28.37%	27.40%	12.50%
OQ112647	66	59	57	26	31.73%	28.37%	27.40%	12.50%
OQ112649	66	59	57	26	31.73%	28.37%	27.40%	12.50%
OQ112673	66	59	57	26	31.73%	28.37%	27.40%	12.50%
OQ112686	66	59	57	26	31.73%	28.37%	27.40%	12.50%
OQ112691	66	59	57	26	31.73%	28.37%	27.40%	12.50%
OQ112730	66	59	57	26	31.73%	28.37%	27.40%	12.50%
OQ112737	66	59	57	26	31.73%	28.37%	27.40%	12.50%
OQ112740	66	59	57	26	31.73%	28.37%	27.40%	12.50%
OQ112771	66	59	57	26	31.73%	28.37%	27.40%	12.50%
OQ112893	66	59	57	26	31.73%	28.37%	27.40%	12.50%

Table J.5: Frame 10, 16, 22, ... The proportions of the presence of all four types of nucleotides are balanced in all sequences.

ID	A	T	C	G	A (%)	T (%)	C (%)	G (%)
MZ558275	59	31	52	67	28.23%	14.83%	24.88%	32.06%
MZ280957	59	31	52	67	28.23%	14.83%	24.88%	32.06%
MZ817897	59	31	52	67	28.23%	14.83%	24.88%	32.06%
OK039706	59	31	52	67	28.23%	14.83%	24.88%	32.06%
OL791595	59	31	52	67	28.23%	14.83%	24.88%	32.06%
OL920359	59	31	52	67	28.23%	14.83%	24.88%	32.06%
OM059434	59	31	52	67	28.23%	14.83%	24.88%	32.06%
ON144454	71	28	52	56	34.30%	13.53%	25.12%	27.05%
ON330593	71	28	52	56	34.30%	13.53%	25.12%	27.05%
ON594981	71	28	52	56	34.30%	13.53%	25.12%	27.05%
ON650495	71	28	52	56	34.30%	13.53%	25.12%	27.05%
ON800622	71	28	52	56	34.30%	13.53%	25.12%	27.05%
ON980049	71	28	52	56	34.30%	13.53%	25.12%	27.05%
OP129015	71	28	52	56	34.30%	13.53%	25.12%	27.05%
OP294029	71	28	52	56	34.30%	13.53%	25.12%	27.05%
OP311263	71	28	52	56	34.30%	13.53%	25.12%	27.05%
OP313193	71	28	52	56	34.30%	13.53%	25.12%	27.05%
OP313256	71	28	52	56	34.30%	13.53%	25.12%	27.05%
OP313383	71	28	52	56	34.30%	13.53%	25.12%	27.05%
OP359817	71	28	52	56	34.30%	13.53%	25.12%	27.05%
OP768851	71	28	52	56	34.30%	13.53%	25.12%	27.05%
OP934590	71	28	52	56	34.30%	13.53%	25.12%	27.05%
OQ094825	71	28	52	56	34.30%	13.53%	25.12%	27.05%
OQ111935	71	28	52	56	34.30%	13.53%	25.12%	27.05%
OQ112611	71	28	52	56	34.30%	13.53%	25.12%	27.05%
OQ112616	71	28	52	56	34.30%	13.53%	25.12%	27.05%
OQ112619	71	28	52	56	34.30%	13.53%	25.12%	27.05%
OQ112624	71	28	52	56	34.30%	13.53%	25.12%	27.05%
OQ112639	71	28	52	56	34.30%	13.53%	25.12%	27.05%
OQ112647	71	28	52	56	34.30%	13.53%	25.12%	27.05%
OQ112649	71	28	52	56	34.30%	13.53%	25.12%	27.05%
OQ112673	71	28	52	56	34.30%	13.53%	25.12%	27.05%
OQ112686	71	28	52	56	34.30%	13.53%	25.12%	27.05%
OQ112691	71	28	52	56	34.30%	13.53%	25.12%	27.05%
OQ112730	71	28	52	56	34.30%	13.53%	25.12%	27.05%
OQ112737	71	28	52	56	34.30%	13.53%	25.12%	27.05%
OQ112740	71	28	52	56	34.30%	13.53%	25.12%	27.05%
OQ112771	71	28	52	56	34.30%	13.53%	25.12%	27.05%
OQ112893	71	28	52	56	34.30%	13.53%	25.12%	27.05%

Table J.6: Frame 11, 17, 23, ... The proportions of the presence of all four types of nucleotides are balanced in all sequences.

ID	A	T	C	G	A (%)	T (%)	C (%)	G (%)
MZ558275	68	35	59	47	32.54%	16.75%	28.23%	22.49%
MZ280957	68	35	59	47	32.54%	16.75%	28.23%	22.49%
MZ817897	68	34	60	47	32.54%	16.27%	28.71%	22.49%
OK039706	68	35	60	46	32.54%	16.75%	28.71%	22.01%
OL791595	68	34	60	47	32.54%	16.27%	28.71%	22.49%
OL920359	68	34	60	47	32.54%	16.27%	28.71%	22.49%
OM059434	68	34	60	47	32.54%	16.27%	28.71%	22.49%
ON144454	70	32	60	45	33.82%	15.46%	28.99%	21.74%
ON330593	70	32	60	45	33.82%	15.46%	28.99%	21.74%
ON594981	70	32	60	45	33.82%	15.46%	28.99%	21.74%
ON650495	70	32	60	45	33.82%	15.46%	28.99%	21.74%
ON800622	70	32	60	45	33.82%	15.46%	28.99%	21.74%
ON980049	70	32	60	45	33.82%	15.46%	28.99%	21.74%
OP129015	70	32	60	45	33.82%	15.46%	28.99%	21.74%
OP294029	70	32	60	45	33.82%	15.46%	28.99%	21.74%
OP311263	70	32	60	45	33.82%	15.46%	28.99%	21.74%
OP313193	70	32	60	45	33.82%	15.46%	28.99%	21.74%
OP313256	70	32	60	45	33.82%	15.46%	28.99%	21.74%
OP313383	70	32	60	45	33.82%	15.46%	28.99%	21.74%
OP359817	70	32	60	45	33.82%	15.46%	28.99%	21.74%
OP768851	70	32	60	45	33.82%	15.46%	28.99%	21.74%
OP934590	70	32	60	45	33.82%	15.46%	28.99%	21.74%
OQ094825	70	33	59	45	33.82%	15.94%	28.50%	21.74%
OQ111935	70	32	60	45	33.82%	15.46%	28.99%	21.74%
OQ112611	70	32	60	45	33.82%	15.46%	28.99%	21.74%
OQ112616	70	32	60	45	33.82%	15.46%	28.99%	21.74%
OQ112619	70	32	60	45	33.82%	15.46%	28.99%	21.74%
OQ112624	70	32	60	45	33.82%	15.46%	28.99%	21.74%
OQ112639	70	32	60	45	33.82%	15.46%	28.99%	21.74%
OQ112647	70	32	60	45	33.82%	15.46%	28.99%	21.74%
OQ112649	70	32	60	45	33.82%	15.46%	28.99%	21.74%
OQ112673	70	32	60	45	33.82%	15.46%	28.99%	21.74%
OQ112686	70	32	60	45	33.82%	15.46%	28.99%	21.74%
OQ112691	70	32	60	45	33.82%	15.46%	28.99%	21.74%
OQ112730	70	32	60	45	33.82%	15.46%	28.99%	21.74%
OQ112737	70	32	60	45	33.82%	15.46%	28.99%	21.74%
OQ112740	70	32	60	45	33.82%	15.46%	28.99%	21.74%
OQ112771	70	32	60	45	33.82%	15.46%	28.99%	21.74%
OQ112893	70	32	60	45	33.82%	15.46%	28.99%	21.74%

Jamás se descubriría nada si nos considerásemos satisfechos con las cosas descubiertas.

— Séneca, *cordobés, dueño de plantaciones de olivos, determinista, estoico y desafortunado mentor de Nerón. Su obra inspiró, entre otros, a San Agustín, Erasmo de Rotterdam, Calvino y Descartes.*

K

SARS-CoV-2 Non-Recombinant Sequences Study Case. Statistical Data on the Types of Nucleotides Present in the Arithmetic Series (Arithmetic Ratio = 6) at S Protein

Each row in every table represents a recombinant sequence, accurately classified by VGG-16 SARS-CoV-2 with a high level of confidence and a clearly located hot zone in the Spike protein at $f = 1/6$.

Table K.1: Frame 6, 12, 18, ... In all analyzed sequences, it is noteworthy that at positions 6, 12, 18, ..., the number of T is significantly higher compared to the other nucleotide types, reaching nearly 50% of presence compared to the other three nucleotide types combined.

ID	A	T	C	G	A (%)	T (%)	C (%)	G (%)
OX248233	138	262	77	69	25,27%	47,99%	14,10%	12,64%
BS002380	162	298	97	78	25,51%	46,93%	15,28%	12,28%
BS003228	162	298	97	78	25,51%	46,93%	15,28%	12,28%
HG997023	159	300	97	79	25,04%	47,24%	15,28%	12,44%
MT972409	185	282	108	62	29,04%	44,27%	16,95%	9,73%
MZ244586	159	300	97	79	25,04%	47,24%	15,28%	12,44%
MZ700508	159	299	98	79	25,04%	47,09%	15,43%	12,44%
OA982375	185	283	107	62	29,04%	44,43%	16,80%	9,73%
OD966578	147	282	92	74	24,71%	47,39%	15,46%	12,44%
OD969387	159	300	97	79	25,04%	47,24%	15,28%	12,44%

Table K.1 – Continuation from the previous page

ID	**A**	**T**	**C**	**G**	**A** (%)	**T** (%)	**C** (%)	**G** (%)
OD979502	159	300	97	79	25,04%	47,24%	15,28%	12,44%
OL751379	185	283	107	61	29,09%	44,50%	16,82%	9,59%
OL778149	185	283	107	61	29,09%	44,50%	16,82%	9,59%
OL856645	185	284	106	61	29,09%	44,65%	16,67%	9,59%
OL868832	184	284	106	62	28,93%	44,65%	16,67%	9,75%
OL993467	185	283	107	61	29,09%	44,50%	16,82%	9,59%
OM058523	185	283	107	61	29,09%	44,50%	16,82%	9,59%
OM235729	162	298	97	78	25,51%	46,93%	15,28%	12,28%
OM263706	160	297	97	80	25,24%	46,85%	15,30%	12,62%
OM336312	162	298	97	78	25,51%	46,93%	15,28%	12,28%
OM356528	185	283	107	61	29,09%	44,50%	16,82%	9,59%
OM396820	162	298	97	78	25,51%	46,93%	15,28%	12,28%
OM412022	162	298	97	78	25,51%	46,93%	15,28%	12,28%
OM569369	162	297	97	79	25,51%	46,77%	15,28%	12,44%
OM857717	162	298	97	78	25,51%	46,93%	15,28%	12,28%
ON660231	155	309	94	77	24,41%	48,66%	14,80%	12,13%
ON893479	185	283	107	61	29,09%	44,50%	16,82%	9,59%
OP055674	156	306	96	76	24,61%	48,26%	15,14%	11,99%
OP125987	155	307	95	77	24,45%	48,42%	14,98%	12,15%
OP200039	156	307	95	76	24,61%	48,42%	14,98%	11,99%
OP468268	156	307	95	76	24,61%	48,42%	14,98%	11,99%
OQ085662	156	307	95	76	24,61%	48,42%	14,98%	11,99%
OU448371	185	284	107	60	29,09%	44,65%	16,82%	9,43%
OU796188	152	274	88	76	25,76%	46,44%	14,92%	12,88%
OV006668	185	283	107	61	29,09%	44,50%	16,82%	9,59%
OV008054	171	262	105	57	28,74%	44,03%	17,65%	9,58%
OW409375	159	278	93	74	26,32%	46,03%	15,40%	12,25%
OW464011	185	283	107	61	29,09%	44,50%	16,82%	9,59%
OX073652	159	278	93	74	26,32%	46,03%	15,40%	12,25%

Table K.2: Frame 7, 13, 19, ... In contrast to the previous case, the phenomenon related to the high quantity of T at these positions does not occur. The proportions of the presence of all four types of nucleotides are balanced in all sequences.

ID	A	T	C	G	A (%)	T (%)	C (%)	G (%)
OX248233	158	130	84	173	28,99%	23,85%	15,41%	31,74%
BS002380	207	142	95	191	32,60%	22,36%	14,96%	30,08%
BS003228	207	142	95	191	32,60%	22,36%	14,96%	30,08%
HG997023	205	141	97	192	32,28%	22,20%	15,28%	30,24%
MT972409	193	155	113	175	30,35%	24,37%	17,77%	27,52%
MZ244586	205	141	97	192	32,28%	22,20%	15,28%	30,24%
MZ700508	205	141	97	192	32,28%	22,20%	15,28%	30,24%
OA982375	193	155	113	175	30,35%	24,37%	17,77%	27,52%
OD966578	195	132	90	177	32,83%	22,22%	15,15%	29,80%
OD969387	205	141	97	192	32,28%	22,20%	15,28%	30,24%
OD979502	205	141	97	192	32,28%	22,20%	15,28%	30,24%
OL751379	193	154	113	175	30,39%	24,25%	17,80%	27,56%
OL778149	193	154	113	175	30,39%	24,25%	17,80%	27,56%
OL856645	193	154	113	175	30,39%	24,25%	17,80%	27,56%
OL868832	193	154	113	175	30,39%	24,25%	17,80%	27,56%
OL993467	193	154	113	175	30,39%	24,25%	17,80%	27,56%
OM058523	193	154	113	175	30,39%	24,25%	17,80%	27,56%
OM235729	207	142	95	191	32,60%	22,36%	14,96%	30,08%
OM263706	205	143	96	190	32,33%	22,56%	15,14%	29,97%
OM336312	207	142	95	191	32,60%	22,36%	14,96%	30,08%
OM356528	193	154	113	175	30,39%	24,25%	17,80%	27,56%
OM396820	207	142	95	191	32,60%	22,36%	14,96%	30,08%
OM412022	207	142	95	191	32,60%	22,36%	14,96%	30,08%
OM569369	206	142	95	192	32,44%	22,36%	14,96%	30,24%
OM857717	207	142	95	191	32,60%	22,36%	14,96%	30,08%
ON660231	195	152	95	193	30,71%	23,94%	14,96%	30,39%
ON893479	193	154	113	175	30,39%	24,25%	17,80%	27,56%
OP055674	195	151	95	193	30,76%	23,82%	14,98%	30,44%
OP125987	195	151	95	193	30,76%	23,82%	14,98%	30,44%
OP200039	195	152	95	192	30,76%	23,97%	14,98%	30,28%
OP468268	195	151	95	193	30,76%	23,82%	14,98%	30,44%
OQ085662	195	151	95	193	30,76%	23,82%	14,98%	30,44%
OU448371	193	154	113	175	30,39%	24,25%	17,80%	27,56%
OU796188	190	131	92	177	32,20%	22,20%	15,59%	30,00%
OV006668	193	154	113	175	30,39%	24,25%	17,80%	27,56%
OV008054	186	141	101	166	31,31%	23,74%	17,00%	27,95%
OW409375	189	137	91	187	31,29%	22,68%	15,07%	30,96%
OW464011	193	154	113	175	30,39%	24,25%	17,80%	27,56%
OX073652	189	137	91	187	31,29%	22,68%	15,07%	30,96%

Table K.3: Frame 8, 14, 20, ... Similar distribution as in the previous case. The presence of all four types of nucleotides is relatively balanced in all analyzed sequences.

ID	A	T	C	G	A (%)	T (%)	C (%)	G (%)
OX248233	169	170	126	82	30,90%	31,08%	23,03%	14,99%
BS002380	204	190	151	90	32,13%	29,92%	23,78%	14,17%
BS003228	204	190	151	90	32,13%	29,92%	23,78%	14,17%
HG997023	203	190	151	91	31,97%	29,92%	23,78%	14,33%
MT972409	194	171	163	108	30,50%	26,89%	25,63%	16,98%
MZ244586	203	190	151	91	31,97%	29,92%	23,78%	14,33%
MZ700508	203	191	150	91	31,97%	30,08%	23,62%	14,33%
OA982375	194	171	163	108	30,50%	26,89%	25,63%	16,98%
OD966578	194	175	139	87	32,61%	29,41%	23,36%	14,62%
OD969387	203	190	151	91	31,97%	29,92%	23,78%	14,33%
OD979502	203	190	151	91	31,97%	29,92%	23,78%	14,33%
OL751379	194	170	161	110	30,55%	26,77%	25,35%	17,32%
OL778149	194	169	162	110	30,55%	26,61%	25,51%	17,32%
OL856645	194	170	161	110	30,55%	26,77%	25,35%	17,32%
OL868832	194	170	161	110	30,55%	26,77%	25,35%	17,32%
OL993467	194	170	161	110	30,55%	26,77%	25,35%	17,32%
OM058523	194	170	161	110	30,55%	26,77%	25,35%	17,32%
OM235729	204	190	151	90	32,13%	29,92%	23,78%	14,17%
OM263706	202	190	151	91	31,86%	29,97%	23,82%	14,35%
OM336312	203	190	151	91	31,97%	29,92%	23,78%	14,33%
OM356528	194	171	160	110	30,55%	26,93%	25,20%	17,32%
OM396820	203	190	151	91	31,97%	29,92%	23,78%	14,33%
OM412022	203	190	151	91	31,97%	29,92%	23,78%	14,33%
OM569369	203	190	151	91	31,97%	29,92%	23,78%	14,33%
OM857717	203	190	151	91	31,97%	29,92%	23,78%	14,33%
ON660231	200	197	144	94	31,50%	31,02%	22,68%	14,80%
ON893479	194	170	161	110	30,55%	26,77%	25,35%	17,32%
OP055674	199	195	145	95	31,39%	30,76%	22,87%	14,98%
OP125987	199	194	145	96	31,39%	30,60%	22,87%	15,14%
OP200039	199	195	144	96	31,39%	30,76%	22,71%	15,14%
OP468268	199	196	144	95	31,39%	30,91%	22,71%	14,98%
OQ085662	198	195	146	95	31,23%	30,76%	23,03%	14,98%
OU448371	194	170	161	110	30,55%	26,77%	25,35%	17,32%
OU796188	189	178	144	80	31,98%	30,12%	24,37%	13,54%
OV006668	194	170	161	110	30,55%	26,77%	25,35%	17,32%
OV008054	178	163	149	104	29,97%	27,44%	25,08%	17,51%
OW409375	190	183	146	85	31,46%	30,30%	24,17%	14,07%
OW464011	194	171	160	110	30,55%	26,93%	25,20%	17,32%
OX073652	190	183	146	85	31,46%	30,30%	24,17%	14,07%

Table K.4: Frame 9, 15, 21, ... Just like in the case of the Arithmetic Series 6, 12, 18, ..., the presence of T at these sixth positions is significantly higher, accounting for nearly half compared to the other nucleotide types. We are observing the same phenomenon in both series.

ID	A	T	C	G	A (%)	T (%)	C (%)	G (%)
OX248233	161	236	98	52	29,43%	43,14%	17,92%	9,51%
BS002380	185	287	103	60	29,13%	45,20%	16,22%	9,45%
BS003228	185	287	103	60	29,13%	45,20%	16,22%	9,45%
HG997023	184	288	106	57	28,98%	45,35%	16,69%	8,98%
MT972409	159	307	96	74	25,00%	48,27%	15,09%	11,64%
MZ244586	184	288	106	57	28,98%	45,35%	16,69%	8,98%
MZ700508	184	288	106	57	28,98%	45,35%	16,69%	8,98%
OA982375	159	307	96	74	25,00%	48,27%	15,09%	11,64%
OD966578	170	274	100	52	28,52%	45,97%	16,78%	8,72%
OD969387	184	288	106	57	28,98%	45,35%	16,69%	8,98%
OD979502	184	288	106	57	28,98%	45,35%	16,69%	8,98%
OL751379	159	307	95	74	25,04%	48,35%	14,96%	11,65%
OL778149	159	307	95	74	25,04%	48,35%	14,96%	11,65%
OL856645	159	307	95	74	25,04%	48,35%	14,96%	11,65%
OL868832	159	307	95	74	25,04%	48,35%	14,96%	11,65%
OL993467	159	308	95	73	25,04%	48,50%	14,96%	11,50%
OM058523	159	307	95	74	25,04%	48,35%	14,96%	11,65%
OM235729	185	287	103	60	29,13%	45,20%	16,22%	9,45%
OM263706	185	289	103	57	29,18%	45,58%	16,25%	8,99%
OM336312	185	287	103	60	29,13%	45,20%	16,22%	9,45%
OM356528	159	307	95	74	25,04%	48,35%	14,96%	11,65%
OM396820	185	287	103	60	29,13%	45,20%	16,22%	9,45%
OM412022	185	287	103	60	29,13%	45,20%	16,22%	9,45%
OM569369	185	288	103	59	29,13%	45,35%	16,22%	9,29%
OM857717	185	287	103	60	29,13%	45,20%	16,22%	9,45%
ON660231	187	281	106	61	29,45%	44,25%	16,69%	9,61%
ON893479	159	308	95	73	25,04%	48,50%	14,96%	11,50%
OP055674	186	281	106	61	29,34%	44,32%	16,72%	9,62%
OP125987	186	281	106	61	29,34%	44,32%	16,72%	9,62%
OP200039	186	281	106	61	29,34%	44,32%	16,72%	9,62%
OP468268	186	281	106	61	29,34%	44,32%	16,72%	9,62%
OQ085662	187	280	106	61	29,50%	44,16%	16,72%	9,62%
OU448371	159	307	95	74	25,04%	48,35%	14,96%	11,65%
OU796188	168	267	103	53	28,43%	45,18%	17,43%	8,97%
OV006668	159	307	95	74	25,04%	48,35%	14,96%	11,65%
OV008054	153	283	86	71	25,80%	47,72%	14,50%	11,97%
OW409375	174	273	101	55	28,86%	45,27%	16,75%	9,12%
OW464011	159	307	95	74	25,04%	48,35%	14,96%	11,65%

Table K.4 – Continuation from the previous page

ID	A	T	C	G	A (%)	T (%)	C (%)	G (%)
OX073652	174	273	101	55	28,86%	45,27%	16,75%	9,12%

Table K.5: Frame 10, 16, 22, ... The presence of all four types of nucleotides is once again balanced in all sequences.

ID	A	T	C	G	A (%)	T (%)	C (%)	G (%)
OX248233	168	126	103	149	30,77%	23,08%	18,86%	27,29%
BS002380	181	164	115	174	28,55%	25,87%	18,14%	27,44%
BS003228	181	164	115	174	28,55%	25,87%	18,14%	27,44%
HG997023	181	166	112	175	28,55%	26,18%	17,67%	27,60%
MT972409	194	153	96	193	30,50%	24,06%	15,09%	30,35%
MZ244586	181	166	112	175	28,55%	26,18%	17,67%	27,60%
MZ700508	181	165	113	175	28,55%	26,03%	17,82%	27,60%
OA982375	194	154	95	193	30,50%	24,21%	14,94%	30,35%
OD966578	167	160	102	165	28,11%	26,94%	17,17%	27,78%
OD969387	181	166	112	175	28,55%	26,18%	17,67%	27,60%
OD979502	181	166	112	175	28,55%	26,18%	17,67%	27,60%
OL751379	194	153	96	192	30,55%	24,09%	15,12%	30,24%
OL778149	194	153	96	192	30,55%	24,09%	15,12%	30,24%
OL856645	194	153	96	192	30,55%	24,09%	15,12%	30,24%
OL868832	194	153	96	192	30,55%	24,09%	15,12%	30,24%
OL993467	194	154	96	191	30,55%	24,25%	15,12%	30,08%
OM058523	194	153	96	192	30,55%	24,09%	15,12%	30,24%
OM235729	181	164	115	174	28,55%	25,87%	18,14%	27,44%
OM263706	182	166	110	175	28,75%	26,22%	17,38%	27,65%
OM336312	181	164	115	174	28,55%	25,87%	18,14%	27,44%
OM356528	194	153	96	192	30,55%	24,09%	15,12%	30,24%
OM396820	181	164	115	174	28,55%	25,87%	18,14%	27,44%
OM412022	181	164	115	174	28,55%	25,87%	18,14%	27,44%
OM569369	181	164	115	174	28,55%	25,87%	18,14%	27,44%
OM857717	181	164	115	174	28,55%	25,87%	18,14%	27,44%
ON660231	191	157	113	173	30,13%	24,76%	17,82%	27,29%
ON893479	194	153	96	192	30,55%	24,09%	15,12%	30,24%
OP055674	191	157	112	173	30,17%	24,80%	17,69%	27,33%
OP125987	191	157	112	173	30,17%	24,80%	17,69%	27,33%
OP200039	191	157	112	173	30,17%	24,80%	17,69%	27,33%
OP468268	191	157	112	173	30,17%	24,80%	17,69%	27,33%
OQ085662	191	157	112	173	30,17%	24,80%	17,69%	27,33%
OU448371	194	153	96	192	30,55%	24,09%	15,12%	30,24%
OU796188	175	151	99	165	29,66%	25,59%	16,78%	27,97%

Table K.5 – Continuation from the previous page

ID	A	T	C	G	A (%)	T (%)	C (%)	G (%)
OV006668	194	154	96	191	30,55%	24,25%	15,12%	30,08%
OV008054	179	144	91	178	30,24%	24,32%	15,37%	30,07%
OW409375	176	153	111	162	29,24%	25,42%	18,44%	26,91%
OW464011	194	153	96	192	30,55%	24,09%	15,12%	30,24%
OX073652	176	153	111	162	29,24%	25,42%	18,44%	26,91%

Table K.6: Frame 11, 17, 23, ... Similar to all arithmetic series that are not multiples of 3, the presence of the four types of nucleotides is more balanced than in the case of multiples of 3.

ID	A	T	C	G	A (%)	T (%)	C (%)	G (%)
OX248233	169	151	137	88	31,01%	27,71%	25,14%	16,15%
BS002380	190	182	155	107	29,97%	28,71%	24,45%	16,88%
BS003228	190	182	155	107	29,97%	28,71%	24,45%	16,88%
HG997023	189	180	157	108	29,81%	28,39%	24,76%	17,03%
MT972409	198	199	148	91	31,13%	31,29%	23,27%	14,31%
MZ244586	189	180	157	108	29,81%	28,39%	24,76%	17,03%
MZ700508	189	180	157	108	29,81%	28,39%	24,76%	17,03%
OA982375	198	200	147	91	31,13%	31,45%	23,11%	14,31%
OD966578	176	171	144	103	29,63%	28,79%	24,24%	17,34%
OD969387	189	180	157	108	29,81%	28,39%	24,76%	17,03%
OD979502	189	180	157	108	29,81%	28,39%	24,76%	17,03%
OL751379	199	198	147	91	31,34%	31,18%	23,15%	14,33%
OL778149	199	198	147	91	31,34%	31,18%	23,15%	14,33%
OL856645	199	198	147	91	31,34%	31,18%	23,15%	14,33%
OL868832	200	198	147	91	31,45%	31,13%	23,11%	14,31%
OL993467	199	198	147	91	31,34%	31,18%	23,15%	14,33%
OM058523	199	198	147	91	31,34%	31,18%	23,15%	14,33%
OM235729	190	182	155	107	29,97%	28,71%	24,45%	16,88%
OM263706	190	180	156	107	30,02%	28,44%	24,64%	16,90%
OM336312	190	182	155	107	29,97%	28,71%	24,45%	16,88%
OM356528	199	198	147	91	31,34%	31,18%	23,15%	14,33%
OM396820	190	182	155	107	29,97%	28,71%	24,45%	16,88%
OM412022	190	182	155	107	29,97%	28,71%	24,45%	16,88%
OM569369	190	182	155	107	29,97%	28,71%	24,45%	16,88%
OM857717	190	182	155	107	29,97%	28,71%	24,45%	16,88%
ON660231	195	174	160	105	30,76%	27,44%	25,24%	16,56%
ON893479	199	198	147	91	31,34%	31,18%	23,15%	14,33%
OP055674	195	174	160	104	30,81%	27,49%	25,28%	16,43%
OP125987	195	174	160	104	30,81%	27,49%	25,28%	16,43%

Table K.6 – Continuation from the previous page

ID	**A**	**T**	**C**	**G**	**A** (%)	**T** (%)	**C** (%)	**G** (%)
OP200039	195	174	160	104	30,81%	27,49%	25,28%	16,43%
OP468268	195	174	160	104	30,81%	27,49%	25,28%	16,43%
OQ085662	195	175	159	104	30,81%	27,65%	25,12%	16,43%
OU448371	199	198	147	91	31,34%	31,18%	23,15%	14,33%
OU796188	173	169	145	102	29,37%	28,69%	24,62%	17,32%
OV006668	199	198	147	91	31,34%	31,18%	23,15%	14,33%
OV008054	184	189	142	79	30,98%	31,82%	23,91%	13,30%
OW409375	177	176	146	103	29,40%	29,24%	24,25%	17,11%
OW464011	198	199	146	91	31,23%	31,39%	23,03%	14,35%
OX073652	177	175	147	103	29,40%	29,07%	24,42%	17,11%

I am always doing what I cannot do yet, in order to learn how to do it.

— Vincent Van Gogh, *He used to represent the mathematical concept of turbulence, still unsolved, in his works.*

L Compendium of Hot Zones in SARS-CoV-2 Sequences Under Analysis

"Sequence ID" corresponds to the unique identifier of each sequence. "Grad-CAM" provides the heat map resulting from the application of Grad-CAM on the superposed genomic spectrogram of each sequence.

L.1 Recombinant Sequences Group

Table L.1: Case Study of Recombinant Sequences Group. Compilation of recombinant sequences from the Test Set that were correctly classified, whose Grad-CAM main hot zones were exclusively at $f = 1/6$ in the Spike protein. There were a total of 39 sequences. Each sequence is identified by its corresponding ID.

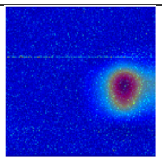
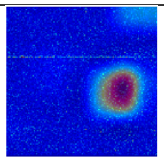
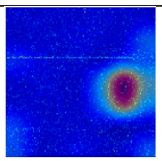
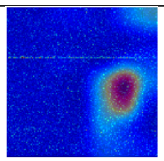
Sequence ID	Grad-CAM	Sequence ID	Grad-CAM
MZ280957		OP768851	
MZ558275		OP934590	

Table L.1 – Continuation from the previous page

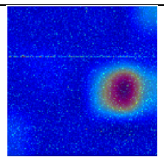
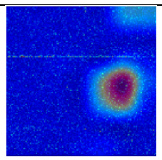
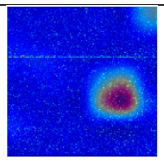
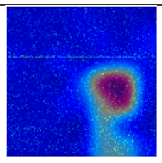
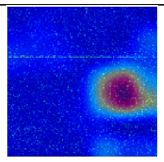
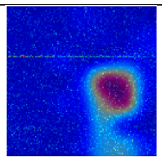
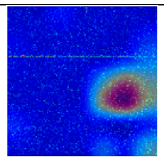
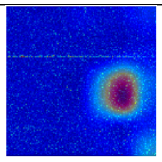
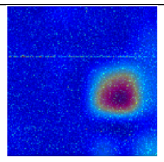
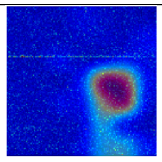
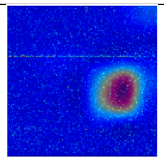
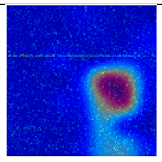
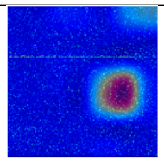
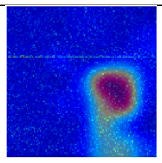
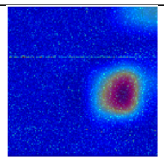
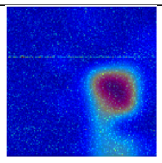
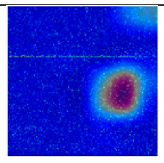
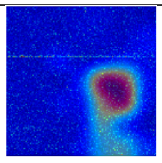
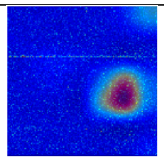
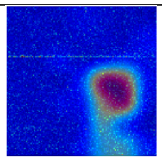
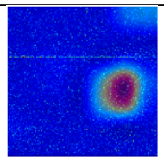
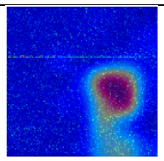
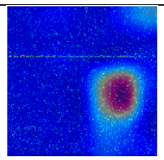
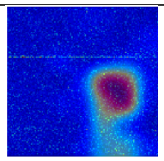
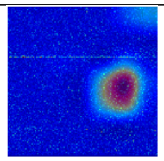
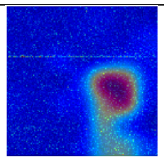
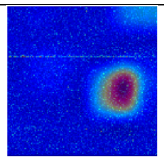
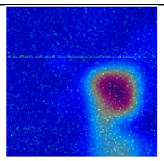
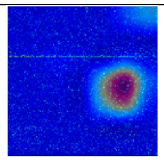
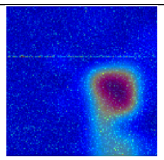
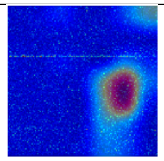
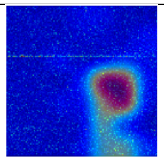
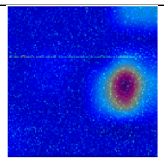
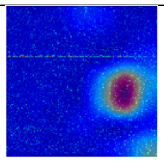
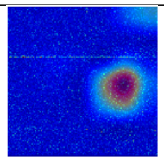
Sequence ID	Grad-CAM	Sequence ID	Grad-CAM
MZ817897		OQ094825	
OK039706		OQ111935	
OL791595		OQ112611	
OL920359		OQ112616	
OM059434		OQ112619	
ON144454		OQ112624	
ON330593		OQ112639	
ON594981		OQ112647	
ON650495		OQ112649	
ON800622		OQ112673	

Table L.1 – Continuation from the previous page

Sequence ID	Grad-CAM	Sequence ID	Grad-CAM
ON980049		OQ112686	
OP129015		OQ112691	
OP294029		OQ112730	
OP311263		OQ112737	
OP313193		OQ112740	
OP313256		OQ112771	
OP313383		OQ112893	
OP359817			

L.2 Non-Recombinant Sequences

Table L.2: Case Study of Non-Recombinant Sequences Group. Compilation of 39 non-recombinant sequences from the test set (the same number as the extracted recombinant sequences) that were correctly classified, whose Grad-CAM main hot zones explicitly exclude $f = 1/6$ in the Spike protein. Each sequence is identified by its corresponding ID.

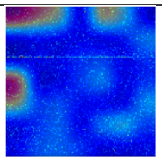
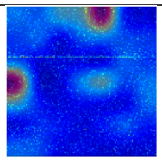
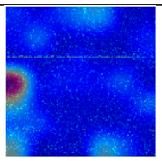
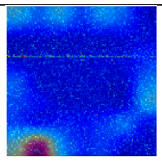
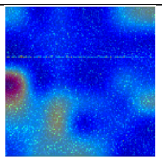
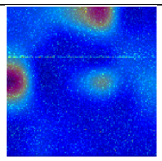
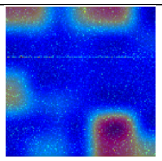
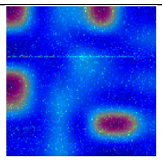
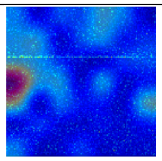
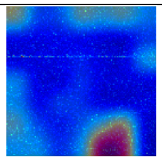
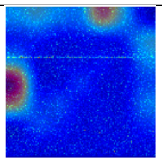
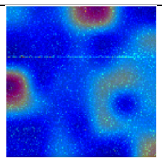
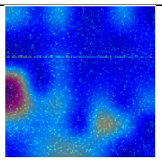
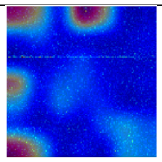
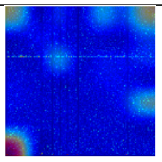
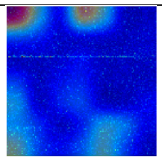
Sequence ID	Grad-CAM	Sequence ID	Grad-CAM
BS002380		OM396820	
BS003228		OM412022	
HG997023		OM569369	
MT972409		OM857717	
MZ244586		ON660231	
MZ700508		ON893479	
OA982375		OP055674	
OD966578		OP125987	

Table L.2 – Continuation from the previous page

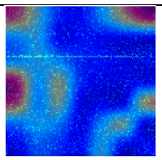
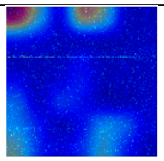
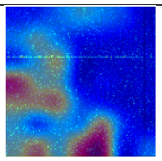
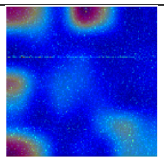
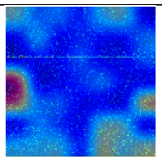
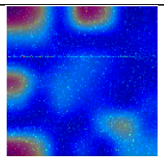
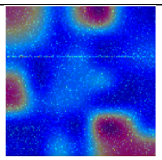
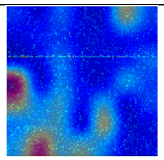
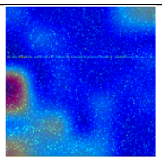
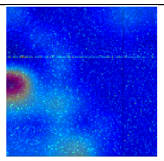
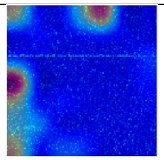
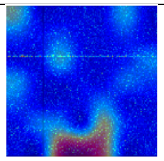
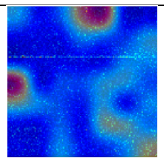
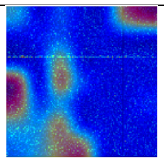
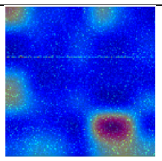
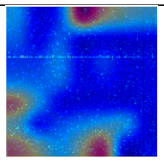
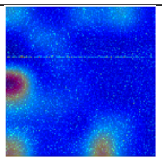
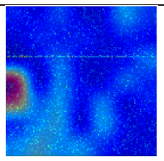
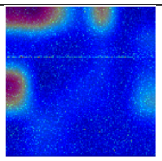
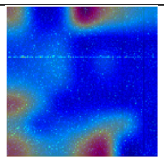
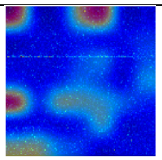
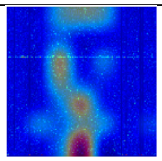
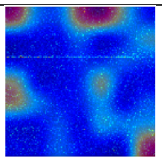
Sequence ID	Grad-CAM	Sequence ID	Grad-CAM
OD969387		OP200039	
OD979502		OP468268	
OL751379		OQ085662	
OL778149		OU448371	
OL856645		OU796188	
OL868832		OV006668	
OL993467		OV008054	
OM058523		OW409375	
OM235729		OW464011	
OM263706		OX073652	

Table L.2 – Continuation from the previous page

Sequence ID	Grad-CAM	Sequence ID	Grad-CAM
OM336312		OX248233	
OM356528			

If you want to go fast, go alone. If you want to go far, go together.

— African Proverb.

M
Scientific Evidence

Attached are the main scientific contributions produced during this research in order of publication.

Ana Guerrero-Tamayo, Borja Sanz Urquijo, Concepción Casado, María-Dolores Moragues Tosantos, Isabel Olivares, Iker Pastor-López. (2023). Validating by Deep Learning an Efficient Method for Genomic Sequence Analysis: Genomic Spectrograms. International Conference on Hybrid Artificial Intelligence Systems (pp. 281-292). Springer Nature Switzerland [3005].

Ana Guerrero-Tamayo, Borja Sanz Urquijo, Concepción Casado, María-Dolores Moragues Tosantos, Isabel Olivares, Iker Pastor-López. (2023). Discovering Mathematical Patterns Behind HIV-1 Genetic Recombination: a new methodology to identify viral features. IEEE Access [2850].

Ana Guerrero-Tamayo, Borja Sanz Urquijo, Concepción Casado, María-Dolores Moragues Tosantos, Isabel Olivares, Iker Pastor-López. (2023). Phylogenetic tree of all sequences comprising the HIV-1 Complete Consensus Multiple Sequence Alignment from the Los Alamos National Laboratory HIV-1 Database (until 2020). Discovering Mathematical Patterns Behind HIV-1 Genetic Recombination: a new methodology to identify viral features - Supplementary Information. IEEE Dataport. https://dx.doi.org/10.21227/2917-1s55 [3007]. Attached is the main screen of IEEE Dataport corresponding to Supplementary Information.

Lecture entitled "Modelos Deep Learning resilientes. Estratégias para uma gestão eficaz de viés em treinamentos de modelos Deep Learning". I - Workshop Internacional sobre Inteligência Artificial "Navegar, ou não, pelo Universo da Inteligência Artificial através de modelos Deep Learning. Universidade Politécnica de Coimbra (IPC), Portugal. Universidade Estadual do Paraná (Unespar), Brasil. Delivered on 2024/04/10.

Ana Guerrero-Tamayo, Fernando José Sadio-Ramos. (2024). Es posible saber Deep Learning sin ser ingeniero? Fundamentos y reflexiones sobre modelos generativos". Coimbra: Fernando José Sadio-Ramos (Editor). ISBN: 978-989-8486-52-3 DOI: http://dx.doi.org/10.5281/zenodo.12518081 [1961]. Due to space constraints, I am attaching the cover page, technical sheet, and general index.

Ana Guerrero-Tamayo, Borja Sanz Urquijo, Concepción Casado, María-Dolores Moragues Tosantos, Isabel Olivares, Iker Pastor-López. (2024). Classification of SARS-CoV-2 sequences as recombinants via a pre-trained CNN and identification of a mathematical signature relative to recombinant feature at Spike, via interpretability. PLoS ONE 19(8): e0309391. https://doi.org/10.1371/journal.pone.0309391 [3006].

Ana Guerrero-Tamayo, Borja Sanz Urquijo, Concepción Casado, María-Dolores Moragues Tosantos, Isabel Olivares, Iker Pastor-López. (2024). Unveiling HIV-1 U Sequences: Shedding Light Through Transfer Learning on Genomic Spectrograms. International Conference on Hybrid Artificial Intelligence Systems. Springer Nature Switzerland.

Ana Guerrero-Tamayo, Borja Sanz Urquijo, Concepción Casado, María-Dolores Moragues Tosantos, Isabel Olivares, Iker Pastor-López. (2024). The Third Codon Nucleotide's Role in Genetic Recombination Within SARS-CoV-2 Spike Protein: A Pilot Study. International Conference on Hybrid Artificial Intelligence Systems. Springer Nature Switzerland.

VNiVERSiDAD D SALAMANCA

Attendance/Presentation Certificate

The Organizing Committee of the **HAIS 2023 Conference**, held in Salamanca, Spain, on September 5th-7th, 2023, hereby

CERTIFIES that:

Ana Guerrero Tamayo has presented the following paper:

Validating by Deep Learning an Efficient Method for Genomic Sequence Analysis: Genomic Spectrograms
Ana Guerrero-Tamayo, Borja Sanz Urquijo, Concepción Casado, María-Dolores Moragues Tosantos, Isabel Olivares and Iker Pastor-López

In the:

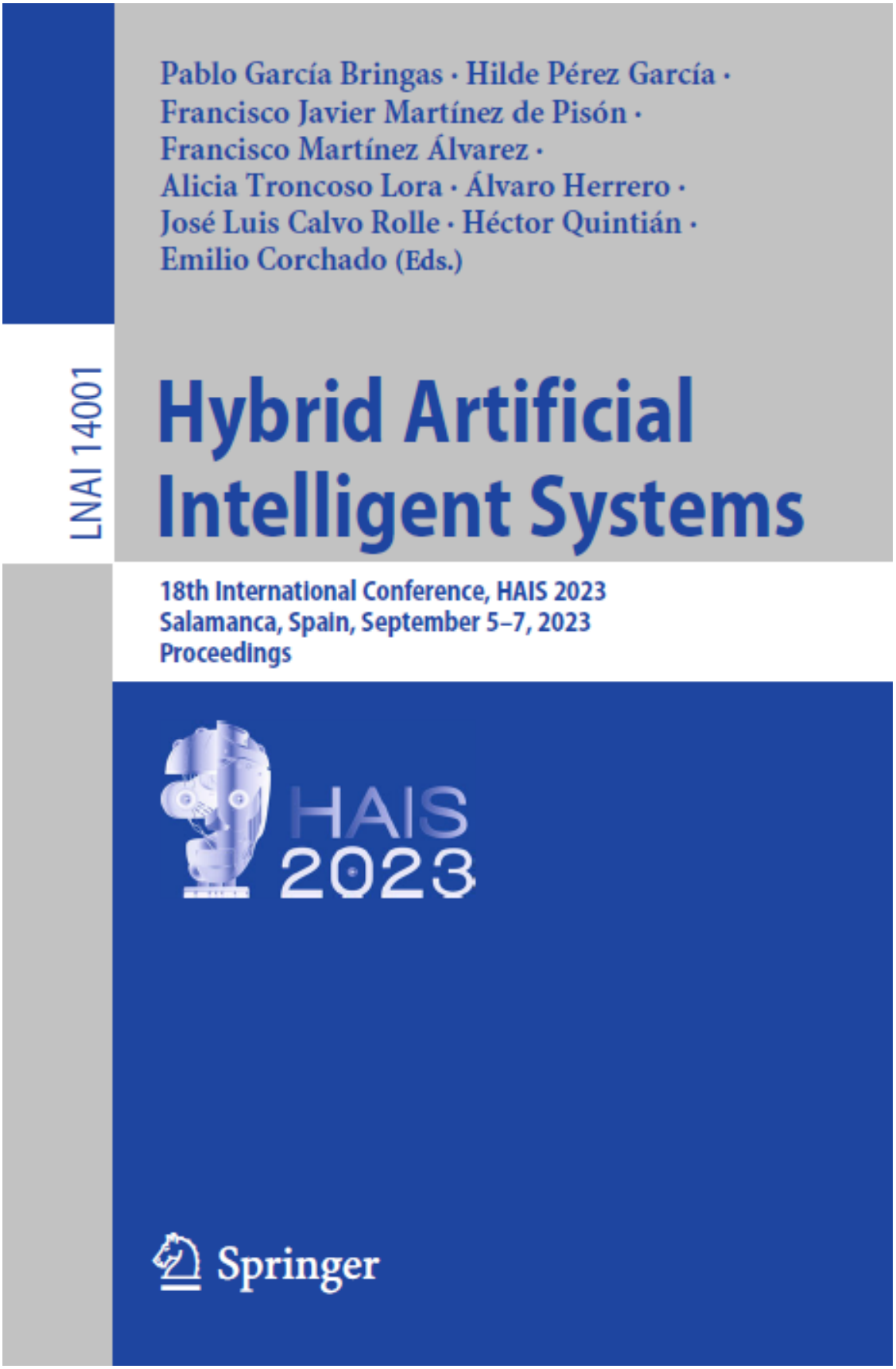
18th International Conference on Hybrid Artificial Intelligence Systems
HAIS 2023

September 5, 2023

Sincerely,



Emilio Corchado Rodríguez
General Chair


Pablo García Bringas · Hilde Pérez García ·
Francisco Javier Martínez de Pisón ·
Francisco Martínez Álvarez ·
Alicia Troncoso Lora · Álvaro Herrero ·
José Luis Calvo Rolle · Héctor Quintián ·
Emilio Corchado (Eds.)
LNAI 14001
Hybrid Artificial Intelligent Systems
18th International Conference, HAIS 2023
Salamanca, Spain, September 5–7, 2023
Proceedings
HAIS 2023
Springer

Analysis of Transformer Model Applications 231
M. I. Cabrera-Bermejo, M. J. Del Jesus, A. J. Rivera, D. Elizondo, F. Charte, and M. D. Pérez-Godoy

Real-Time Workflow Scheduling in Cloud with Recursive Neural Network and List Scheduling 244
Vahab Samandi, Peter Tiňo, and Rami Bahsoon

Robust Losses in Deep Regression 256
Adrián Rubio and Jose R. Dorronsoro

Structure Learning in Deep Multi-Task Models 269
Carlos Ruiz, Carlos M. Alaíz, and José R. Dorronsoro

Validating by Deep Learning an Efficient Method for Genomic Sequence Analysis: Genomic Spectrograms 281
Ana Guerrero-Tamayo, Borja Sanz Urquijo, Concepción Casado, María-Dolores Moragues Tosantos, Isabel Olivares, and Iker Pastor-López

Sentiment Analysis for Vietnamese – Based Hybrid Deep Learning Models 293
Cach N. Dang, María N. Moreno-García, Fernando De la Prieta, Kien V. Nguyen, and Vuong M. Ngo

Optimizing LIME Explanations Using REVEL Metrics 304
Ivan Sevillano-Garcia, Julian Luengo, and Francisco Herrera

Assessing the Impact of Noise on Quantum Neural Networks: An Experimental Analysis 314
Erik Terres Escudero, Danel Arias Alamo, Oier Mentxaka Gómez, and Pablo García Bringas

Varroa Mite Detection Using Deep Learning Techniques 326
Jose Divasón, Francisco Javier Martinez-de-Pison, Ana Romero, Pilar Santolaria, and Jesús L. Yániz

Evolutionary Computation and Optimization

Enhancing Evolutionary Optimization Performance Under Byzantine Fault Conditions 341
Carlos Cotta

A Hybrid Based Genetic Algorithm for Solving the Clustered Generalized Traveling Salesman Problem 352
Ovidiu Cosma, Petricǎ C. Pop, and Laura Cosma



Validating by Deep Learning an Efficient Method for Genomic Sequence Analysis: Genomic Spectrograms

Ana Guerrero-Tamayo[1(✉)], Borja Sanz Urquijo[1], Concepción Casado[2], María-Dolores Moragues Tosantos[3], Isabel Olivares[2], and Iker Pastor-López[1]

[1] Faculty of Engineering, University of Deusto, Avenida de las Universidades, 24, 48007 Bilbao, Biscay, Spain
ana.guerrero@deusto.es
[2] National Microbiology Center (NMC) - Instituto de Salud Carlos III (ISCIII), Ctra. de Pozuelo, 28, Majadahonda, 28222 Madrid, Spain
[3] Faculty of Medicine and Nursing, University of the Basque Country UPV/EHU, Barrio Sarriena s/n, 48940 Leioa, Biscay, Spain

Abstract. Traditionally, comparative and simultaneous analysis of multiple genomic sequences is performed using multiple sequence alignment algorithms. However, these algorithms are computationally complex, limiting the number of sequences that can be analyzed or requiring the comparison of only fragments, leading to a loss of information. Moreover, it is necessary to implement manual adjustments letter by letter after the alignment. Spectrograms offer a way to represent genomic sequences that enables comparison without relying on these traditional methods. Genomic spectrograms are alignment-free, thereby reducing sequence handling and processing time. Several studies have used this representation, and it appears promising. In this study, we validate this representation using a pre-trained Convolutional Neural Network (Inception V3) and transfer learning with a well-documented dataset of HIV-1 sequences to classify recombinant and non-recombinant sequences. As HIV-1 is a small organism, we can work with all fully sequenced genomes and validate the approach. In addition, we processed all HIV-1 complete genomes, not just fragments, handling a total of 134.54 million DNA nucleotides. This way, we did not lose any microbiological information. Sensitivity for classification is 96.33% in the best tested configuration and a validation accuracy over 87.81% in all tested configurations. AUC and Confusion Matrix demonstrate balanced accuracy rates. Our research provides further endorsement to the studies conducted to date and encourages the application of this representation method to genomic research.

Keywords: Genome · Spectrogram · Deep Learning · Convolutional Neural Network · Transfer Learning

University of Deusto - Research Training Grants Program.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
P. García Bringas et al. (Eds.): HAIS 2023, LNAI 14001, pp. 281–292, 2023.
https://doi.org/10.1007/978-3-031-40725-3_24

1 Introduction

In recent years, advances in sequencing technologies have revolutionized the field of genomics, allowing for the rapid and cost-effective generation of vast amounts of genomic data [15,21]. These technologies have a significant impact on the field of metagenomics, which involves the study of genetic material extracted directly from environmental samples [4]. Metagenomics provides a new perspective on the diversity and functions of microbial communities, revealing previously unknown species and metabolic pathways [20,32,34].

However, the state of the art is not enough to face the current challenges, like the interpretation of the complete human genome [7] or the sequencing of almost 2 million SARS-CoV2 sequences [5]. With the increasing amount of sequencing data, new challenges emerge in the field of bioinformatics, particularly in the area of multiple sequence alignment [33]. Assembling and aligning large and complex genomes, as well as metagenomic datasets, is a computationally intensive task that requires sophisticated algorithms and high-performance computing resources. Moreover, these datasets often contain errors and variations that make it challenging to achieve accurate alignments.

Multiple alignment techniques and algorithms are statistical approximations that require complex expert manual adjustments [12]. In general, the most common multiple sequence alignment algorithms, such as ClustalW or MUSCLE, tend to have a time complexity on the order of $\mathcal{O}(N^2L)$, where N is the number of sequences and L is the average sequence length. That is, the execution time will increase quadratically with N and L. This means that a multiple sequence alignment of 40 sequences with an average length of one million nucleotides can take hours or even days to complete.

Working with genomic spectrograms is alignment-free. Therefore, it is not necessary to compare all the sequences letter by letter and to make manual adjustments based on expert knowledge. Compared with alignment-dependent processes, processing time and computational requirements should be minimal.

Our objective was to validate the utility of spectrograms in the field of microbiology. Spectrograms offer advantages such as converting viral sequences into images, addressing issues like multiple sequence alignment. To accomplish this, we verified that the classification of complete HIV-1 genomic spectrograms into recombinant and non-recombinant forms using a pretrained CNN (Inception V3) aligned with microbiological evidence.

2 Background

The Fast Fourier Transform (FFT) is a computational algorithm that efficiently calculates the Discrete Fourier Transform (DFT) of a sequence. This is, the Fourier coefficients of a time-domain signal. It is widely used in signal processing, data analysis, and other related fields to analyze and manipulate signals in the frequency domain. By recursively breaking down the DFT into smaller subproblems, the FFT algorithm utilizes a divide-and-conquer approach that can

efficiently compute the DFT using simple operations. This results in a significant acceleration in computing time compared to direct computation of the DFT, particularly for larger input sizes.

$$X_k = \sum_{n=0}^{N-1} x_n \cdot e^{-\frac{2\pi i}{N} kn} \tag{1}$$

Spectrograms represent the time evolution of the Fast Fourier Transform. Common applications of spectrograms in various fields are in:

- Acoustics: Spectrograms are widely used in acoustics to analyze and visualize sound signals, such as speech, music, and animal vocalizations [23]. They can reveal important information about the frequency content, timing, and intensity of sounds, which can be used for various purposes, such as speech recognition, music analysis, and bioacoustics research [11,16].
- Signal processing: Spectrograms are used in signal processing to analyze and manipulate various types of signals, such as vibration signals [18], electromagnetic signals [10], and radar signals [14]. They can help identify patterns, anomalies, and other characteristics of signals that are difficult to detect in the time domain.
- Medical imaging: Spectrograms are used in medical imaging to analyze and visualize physiological signals, such as electrocardiograms (ECGs) [19] and electroencephalograms (EEGs). They can help diagnose and monitor various medical conditions, such as obstructive sleep apnea [22] or autism spectrum disorders [31].
- Astrophysics: Spectrograms are used in astrophysics to analyze and classify gravitational waves from core-collapse supernova [28].
- Seismology: Spectrograms are used in seismology to analyze and visualize seismic signals, such as avalanches [30] or seismic events [8]. They can provide information about the location, magnitude, and source of seismic events, which can help seismologists monitor and predict seismic activity.

Its successful application in such diverse fields suggests that this type of representation could be applied to new fields such as genomics. There is a growing trend in research to apply genomic spectrograms, which are closely linked with machine learning and deep learning techniques [2,9,24]. But most of them focus mainly on incomplete genome, fragments of coding regions. Usually, the size of the sequences does not allow working with complete genomes due to computational limitations. This can lead to loss of biological information contained in non-coding regions or other auxiliary elements. Validating this representation system using the complete genome of a given organism can provide additional support to the studies carried out and to future studies in this line.

3 Materials and Methods

All experiments ran in this equipment:

- Processing Unit: Intel(R) Xeon(R) Gold 5220R CPU @ 2.20 GHz.
- Installed RAM: 256 GB usable.
- Operative System: Windows 10 Education. Version: 21H1.
- GPU: NVIDIA RTX A6000. Total memory: 179451 MB. VRAM: 48571 MB.

3.1 Case of Study

The genome is the complete set of genetic instructions encoded in the DNA (deoxyribonucleic acid) of an organism. It contains all the information necessary for the development, function, and reproduction of an individual. The genome is composed of genes, which are specific sequences of DNA that encode for proteins or RNA (ribonucleic acid) molecules, as well as non-coding regions that regulate gene expression and other cellular processes. The genome is organized into chromosomes, which are long, linear or circular structures made up of DNA and associated proteins. Nucleotides are the building blocks of DNA and RNA, which are the two types of nucleic acids found in living organisms. Each nucleotide consists of a nitrogenous base, a sugar molecule, and a phosphate group. There are four types of nitrogenous bases found in DNA: adenine (A), thymine (T), guanine (G), and cytosine (C). In RNA, thymine is replaced by uracil (U) [6]. Depending on the type and complexity of the organism, the length of its genome can be extremely extensive, making it difficult to handle and compare. The entire human genome is composed of approximately 3 billion DNA nucleotides. The genome of the long-tailed lizard Anolis carolinensis, which is a species commonly used in research, has approximately 1.8 billion DNA nucleotides. The genome of Saccharomyces cerevisiae, which is a fungus commonly used in research and food production, has approximately 12 million DNA nucleotides. The bacterium Salmonella enterica subsp. Enterica Serovar Typhimurium strain, which is a common cause of food poisoning, has a circular genome approximately 4.9 million nucleotides in length [1]. The DNA of all these organisms is too long for a graphic representation of their complete genome. However, the genome size of viruses, which are the smallest "quasi-living" organisms, is much smaller and presents with a more simple structure.

Consequently, we focused on working with virus genomes, which are the most tractable due to their compact size and simple structure. Within this group, we selected Human Immunodeficiency Virus Type 1 (HIV-1) as our primary model organism. Its genome is small, even compared to other viruses. It consists of approximately 9.8 thousands of nucleotides in length. In contrast, the SARS-CoV2 genome is approximately 30 thousands of nucleotides, roughly three times the size of the HIV-1 genome. As a result, HIV-1 is computationally manageable.

Since the 1980s, the human immunodeficiency virus (HIV) has caused a global health crisis by triggering the AIDS pandemic, which attacks the immune system and leaves individuals vulnerable to various infections and cancers. Despite

progress in understanding, preventing, and treating HIV/AIDS, it remains a significant public health challenge. In 2021, it was estimated that around 38 million people worldwide were living with HIV/AIDS, with sub-Saharan Africa being the region most affected. However, interventions like antiretroviral therapy (ART) and pre-exposure prophylaxis (PrEP) have considerably improved the prognosis for those living with HIV/AIDS [3]. Nevertheless, ongoing efforts are necessary to address the pandemic and guarantee fair access to prevention, testing, and treatment services. The severity of this global health crisis led to a high number of complete sequenced samples (13554 complete genomes until 2020).

HIV-1 is known for its high mutation rate, which enables it to evade the host immune system and develop resistance to antiretroviral drugs. However, another important mechanism that contributes to the genetic diversity of HIV-1 is recombination. Recombination occurs when two genetically distinct viruses infect the same cell and exchange genetic material during the replication process.

In the case of HIV-1, recombination can occur between two different subtypes of the virus, which are prevalent in different regions of the world. This can result in the emergence of new recombinant forms that have unique genetic characteristics and may have different clinical outcomes compared to the parental viruses. Recombination can also occur within the same subtype, leading to the generation of diverse viral variants that can further evolve and adapt to their environment. The high level of genetic diversity of HIV-1 poses significant challenges to the development of effective vaccines and antiretroviral therapies. Understanding the mechanisms of HIV-1 recombination and the genetic factors that contribute to viral diversity is crucial for developing new approaches to combat the global HIV/AIDS epidemic [26].

The importance of genetic recombination in HIV-1 results in the existence of many recombinant sequences perfectly classified as such. So, we have a case study with a sufficient and adequate dataset to validate the experiment through Deep Learning, being the handling of information computationally affordable.

3.2 HIV-1 Sequence Labelling

Our first step was downloading HIV-1 all complete sequences from Los Alamos National Laboratory HIV Database [17]. This compendium is divided into two categories:

- Pure Groups. M, N, O, and P. Group M is the most widespread in the world and is divided into nine subtypes: A, B, C, D, F, G, H, J, and K. Those four pure groups are called non-recombinants.
- Circulating Recombinant Forms (CRF). Those are called recombinant sequences.

Following this criterion, we labelled each sequence as non-recombinant (pure groups) or recombinant (CRF sequences).

286 A. Guerrero-Tamayo et al.

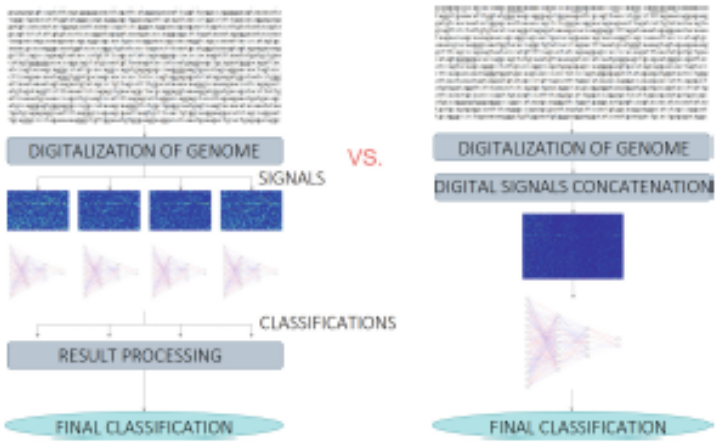


Fig. 1. Comparison of Procedures

3.3 Spectrogram Generation

The first step is to digitize the genome, decomposing it into 4 digital signals (one per nucleotide). At each position of the genome, the signal for a particular nucleotide type is activated only if that nucleotide is present at that position, while the signals corresponding to the other three nucleotides remain deactivated.

We aim to accelerate experimentation by simultaneously working with the spectrograms of individual nucleotides, rather than employing four parallel networks for each nucleotide. To achieve this, we concatenate the spectrograms of each nucleotide into a single image (Fig. 1).

It should be noted that this system could be only viable for analyzing small to medium genomes. For more complex organisms, a separate network would be required for each nucleotide.

Figure 2 presents the image resulting of concatenating the spectrograms of each of the 4 nucleotides into a single image. The x-axis denotes the nucleotide position, while the y-axis represents the frequency range from 0 to 0.5 Hz. Due to the concatenation of the four spectrograms, the length of the figure is four times the sequence length. The z-axis corresponds to the spectrogram calculation.

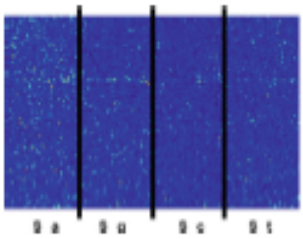


Fig. 2. Spectrogram Scheme

We generated the spectrograms using Python, Scipy library. In order to avoid biases and bad practices during the Convolutional Neural Network (CNN) training, we omitted axes, margins and any other elements that may distort the operation of the network.

3.4 Dataset Setup and Adjustments

Considering the total number of samples available in both categories, the dataset is structured as shown in Table 1.

Table 1. HIV-1 Dataset Structure

Dataset	Recombinant	Non-Recombinant	TOTAL
Train (83.13%)	1903 (16.89%)	9364 (83.11%)	11267
Validation (14.66%)	338 (17.01%)	1649 (82.99%)	1987
Test (2.21%)	150 (50.00%)	150 (50.00%)	300
	2391 (17.64%)	**11163 (82.36%)**	**13554**

The 82.36% of the sequenced HIV-1 virus are non-recombinant. Recombinant HIV-1 viruses are the 17.64% of the total sequenced. The train and validation folders are aligned with these percentages. As depicted, the weighting of the test dataset (50%/50%) intentionally differs from that of the training and validation datasets (17%/83%) in order to detect any under-fitting biases during the training stages.

Initial testing revealed that traditional data augmentation on datasets notably compromised the performance, likely due to the modification of the biological significance of the sequence. Using this technique would be detrimental, particularly if rotations are produced. This is because the precise orientation of the image determines its biological significance. As a result, we refrained from utilizing this approach.

3.5 Validation Method

Inception is a deep convolutional neural network (CNN) developed for the first time by Google researchers at 2014. Its novel Inception module allows the network to learn spatial and channel-wise features efficiently. This Inception module combines multiple filter sizes (1×1, 3×3, 5×5) and pooling operations within a single layer. And it reduces the number of parameters and computational cost through the use of 1×1 convolutions (bottleneck layers). Inception V3 is formed by symmetric and asymmetric building blocks, which include convolutions, average pooling, max pooling, concatenations, dropouts, and fully connected layers. Batch normalization is applied to activation inputs. Loss is calculated using softmax.Its architecture consists of:

- Convolutional and pooling layers.
- Inception modules (multiple, each with 1×1, 3×3, 5×5 convolutions, and pooling).
- Auxiliary classifiers (x2, used during training for intermediate supervision).
- Global average pooling. It reduces spatial dimensions to 1×1.
- Output layer: 1000 neurons (softmax, for ImageNet's 1000 classes).

Figure 3 shows a detailed scheme of Inception V3 architecture (extracted from Inception V3 Advanced Guide by Google [13]):

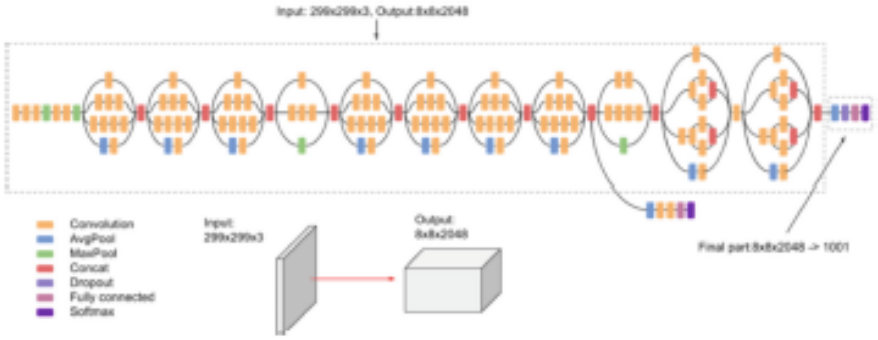


Fig. 3. Inception V3 Architecture

Inception v3 achieves higher accuracy on image recognition tasks compared to its predecessor, Inception v2 [27,29]. This is due to a better use of batch normalization and improved training techniques. Inception v3 uses efficient training with smaller models and techniques to reduce overfitting (e.g., dropout, weight decay). It outperforms other pretrained networks (VGG, ResNet, GoogLeNet) by capturing local and global features with multiple filter layers of different sizes.

Transfer learning uses a pre-trained model to start training a new model for a related task. It reuses learned feature maps to detect general visual features. This helps the new model learn specific features faster, with less data and time. Transfer learning is useful with limited training data, as it leverages knowledge from pre-trained models trained on large datasets. To apply transfer learning, we replaced the last fully connected layer and the classification layer. This approach has been successfully applied in many computer vision applications, including object detection, segmentation, and recognition. In addition, it is widely used in complex image typologies, such as X-ray images [25] or diabetic retinopathy detection [35], achieving consistent positive results.

We did all the experimentation using MATLAB2021b App Deep Learning Designer.

3.6 Performance Measurement

We did all the performance measurement using MATLAB2021b.

To measure and evaluate the performance of each CNN in the different preset configurations, the following parameters were considered: validation accuracy, training time, confusion matrix and test accuracy.

4 Results

To evaluate the performance of our model, we conducted a test bench with various hyperparameter configurations by empirical validation, as shown in the Table 2. Our purposes were to find the optimal combination of hyperparameters that would result in the best performance of the model and to test the robustness of the model.

Table 2. Inception V3 Results

Hyperparameters	Val. Acc.	Training Time	Confusion Matrix	AUC	Test Acc.
Learning Rate: 0.0001 Batchsize: 52 Epochs: 10	0.8938	00:48:28	143 (95%) – 7 (5%) 43 (29%) – 107 (71%)	0.9226	0.8333
Learning Rate: 0.0100 Batchsize: 52 Epochs: 10	**0.9260**	**00:49:07**	**149 (99%) – 1 (1%) 13 (9%) – 137 (91%)**	**0.9863**	**0.9533**
Learning Rate: 0.0001 Batchsize: 52 Epochs: 30	0.8908	02:08:18	145 (97%) – 5 (3%) 38 (25%) – 112 (75%)	0.9499	0.8567
Learning Rate: 0.0100 Batchsize: 52 Epochs: 30	**0.9482**	**02:28:01**	**146 (97%) – 4 (3%) 7 (5%) – 143 (95%)**	**0.9943**	**0.9633**
Learning Rate: 0.0001 Batchsize: 128 Epochs: 10	0.8868	00:33:54	145 (97%) – 5 (3%) 54 (36%) – 96 (64%)	0.8729	0.8033
Learning Rate: 0.0100 Batchsize: 128 Epochs: 10	0.9321	00:34:54	147 (98%) – 3 (2%) 21 (14%) – 129 (86%)	0.9850	0.9200
Learning Rate: 0.0001 Batchsize: 128 Epochs: 30	0.8782	01:42:06	141 (94%) – 9 (6%) 48 (32%) – 102 (68%)	0.9075	0.8100
Learning Rate: 0.0100 Batchsize: 128 Epochs: 30	**0.9351**	**01:49:35**	**147 (98%) – 3 (2%) 14 (9%) - 136 (91%)**	**0.9867**	**0.9433**

The optimal configurations, defined as those that achieve the best and most balanced results, are highlighted in bold in Table 2. Among them, the configuration with 10 Epochs stands out, as it significantly reduces the training time compared to other configurations, while maintaining a high hit rate without suffering excessive penalties. Despite the varying dataset weighting, all configurations achieved an accuracy rate exceeding 87%. However, the accuracy significantly deteriorated in the recombinant sequences category for Learning Rate = 0.0001. The validation of this representation model should not only be the measurement of the absolute accuracy level. We also need balanced hit rates in both categories of the virus. Consequently, the best configuration is for Learning Rate = 0.0100, Batchsize = 52 and Epochs = 30.

Furthermore, the computation times are manageable, especially when compared to the computation times handled by other bio programs. We are working simultaneously with 13554 complete HIV-1 sequences and obtaining results with a Test Accuracy of 0.9633 in the optimal configuration, with an Average Hit Rate of 0.9450 in the configurations corresponding to Learning Rate = 0.0100.

5 Conclusions

Our research demonstrates that utilizing genomic spectrograms is a valid approach for representing and managing genomic sequences. The alignment-free nature of this representation method allowed us to save time by avoiding multiple sequence alignment and manual adjustments. The use of a pre-trained CNN (Inception V3) with transfer learning achieved high accuracy rates for classifying recombinant and non-recombinant HIV-1 sequences. The AUC and Confusion Matrix analysis revealed that the accuracy was balanced, indicating that genomic spectrograms provide a reliable representation of genomic sequences. We were able to validate this representation model by using whole genomes instead of just fragments. As a result, we processed the complete information without any loss of data. Overall, our findings suggest that genomic spectrograms are an efficient and practical alternative to multiple sequence alignment for analyzing large-scale genomic data.

It is highly interesting to develop biological algorithms specifically designed for the analysis of genomic sequences represented through spectrograms. Developing algorithms tailored to this approach could lead to further advancements in genomic research and analysis. By taking advantage of the alignment-free nature of this representation method, we can reduce the computational complexity of traditional alignment-based approaches and potentially uncover novel biological insights. Overall, our study highlights the potential benefits of developing specialized algorithms for analyzing genomic sequences represented through spectrograms.

Our next steps will involve repeating the experiment with other viruses and different viral characteristics, such as specific organ involvement, mortality, morbidity, antiviral resistance, etc. Additionally, we will analyze the feasibility of these concatenated spectrograms in larger organisms, such as bacteria.

References

1. Archive, E.E.N.: ENA (2023). https://www.ebi.ac.uk/ena/browser/. Accessed 12 May 2023
2. Bonidia, R.P., et al.: Feature extraction approaches for biological sequences: a comparative study of mathematical features. Brief. Bioinform. **22**(5), bbab011 (2021)
3. Celum, C., Baeten, J.M.: Antiretroviral-based HIV-1 prevention: antiretroviral treatment and pre-exposure prophylaxis. Antivir. Ther. **17**(8), 1483–1493 (2012)
4. Chiu, C.Y., Miller, S.A.: Clinical metagenomics. Nat. Rev. Genet. **20**(6), 341–355 (2019)
5. Collaboration, I.I.N.S.D.: SARS-CoV2 database (2023). https://covid19.sfb.uit.no/. Accessed 23 May 2023
6. Corona, D.Y.: Brock. Biología de los microorganismos. Pearson, London (2015)
7. Cortés-Ciriano, I., Gulhan, D.C., Lee, J.J.K., Melloni, G.E., Park, P.J.: Computational analysis of cancer genome sequencing data. Nat. Rev. Genet. **23**(5), 298–314 (2022)
8. Curilem, M., et al.: Discriminating seismic events of the Llaima volcano (Chile) based on spectrogram cross-correlations. J. Volcanol. Geoth. Res. **367**, 63–78 (2018)
9. Dimitrova, N., Cheung, Y.H., Zhang, M.: Analysis and visualization of DNA spectrograms: open possibilities for the genome research. In: Proceedings of the 14th ACM International Conference on Multimedia, pp. 1017–1024 (2006)
10. Dong, X., et al.: Detection and identification of vehicles based on their unintended electromagnetic emissions. IEEE Trans. Electromagn. Compat. **48**(4), 752–759 (2006)
11. Elemans, C.P., Heeck, K., Muller, M.: Spectrogram analysis of animal sound production. Bioacoustics **18**(2), 183–212 (2008)
12. Gaschen, B., et al.: Diversity considerations in HIV-1 vaccine selection. Science **296**(5577), 2354–2360 (2002)
13. Google: Inception v3 advanced guide (2023). https://cloud.google.com/docs/inception-v3-advanced. Accessed 05 June 2023
14. Harmanny, R., De Wit, J., Cabic, G.P.: Radar micro-doppler feature extraction using the spectrogram and the cepstrogram. In: 2014 11th European Radar Conference, pp. 165–168. IEEE (2014)
15. Hu, T., Chitnis, N., Monos, D., Dinh, A.: Next-generation sequencing technologies: an overview. Hum. Immunol. **82**(11), 801–811 (2021)
16. Jothimani, S., Premalatha, K.: MFF-SAug: multi feature fusion with spectrogram augmentation of speech emotion recognition using convolution neural network. Chaos Solitons Fractals **162**, 112512 (2022)
17. Laboratory, L.A.N.: The HIV database (2023). https://www.hiv.lanl.gov/. Accessed 09 June 2023
18. Leonard, F., Lanteigne, J., Lalonde, S., Turcotte, Y.: Free-vibration behaviour of a cracked cantilever beam and crack detection. Mech. Syst. Signal Process. **15**(3), 529–548 (2001)
19. Li, H., Boulanger, P.: Structural anomalies detection from electrocardiogram (ECG) with spectrogram and handcrafted features. Sensors **22**(7), 2467 (2022)
20. Lu, Z.X., Wang, X.R., Lian, X.L., Liao, X.P., Liu, Y.H., Sun, J.: Advances in the discovery of novel antibiotic-resistant genes based on functional metagenomics. Biotechnol. Bull. **38**(9), 17 (2022)

21. McCombie, W.R., McPherson, J.D., Mardis, E.R.: Next-generation sequencing technologies. Cold Spring Harb. Perspect. Med. **9**(11), a036798 (2019)
22. McNames, J., Fraser, A.: Obstructive sleep apnea classification based on spectrogram patterns in the electrocardiogram. In: Computers in Cardiology 2000, vol. 27 (Cat. 00CH37163), pp. 749–752. IEEE (2000)
23. Miyamoto, S., Nose, T., Hiroshiba, K., Odagiri, Y., Ito, A.: Two-stage sequence-to-sequence neural voice conversion with low-to-high definition spectrogram mapping. In: Pan, J.-S., Ito, A., Tsai, P.-W., Jain, L.C. (eds.) IIH-MSP 2018. SIST, vol. 110, pp. 132–139. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-03748-2_16
24. Morales, J.A., et al.: Deep learning for the classification of genomic signals. Math. Probl. Eng. **2020**, 1–9 (2020)
25. Mujahid, M., Rustam, F., Álvarez, R., Luis Vidal Mazón, J., Díez, I.D.L.T., Ashraf, I.: Pneumonia classification from X-ray images with inception-v3 and convolutional neural network. Diagnostics **12**(5), 1280 (2022)
26. Nájera, R., Delgado, E., Pérez-Alvarez, L., Thomson, M.M.: Genetic recombination and its role in the development of the HIV-1 pandemic. AIDS **16**, S3–S16 (2002)
27. Nguyen, T.-T., Huynh, K.-T.: Spliced image forgery detection based on the combination of image pre-processing and inception V3. In: Dang, T.K., Küng, J., Chung, T.M., Takizawa, M. (eds.) FDSE 2021. LNCS, vol. 13076, pp. 308–322. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-91387-8_20
28. Roma, V., Powell, J., Heng, I.S., Frey, R.: Astrophysics with core-collapse supernova gravitational wave signals in the next generation of gravitational wave detectors. Phys. Rev. D **99**(6), 063018 (2019)
29. Saini, M., Susan, S.: Data augmentation of minority class with transfer learning for classification of imbalanced breast cancer dataset using inception-V3. In: Morales, A., Fierrez, J., Sánchez, J.S., Ribeiro, B. (eds.) IbPRIA 2019, Part I. LNCS, vol. 11867, pp. 409–420. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-31332-6_36
30. Suriñach, E., Flores-Márquez, E.L., Roig-Lafon, P., Furdada, G., Tapia, M.: Estimation of avalanche development and frontal velocities based on the spectrogram of the seismic signals generated at the vallée de la sionne test site. Geosciences **10**(3), 113 (2020)
31. Tawhid, M.N.A., Siuly, S., Wang, H., Whittaker, F., Wang, K., Zhang, Y.: A spectrogram image based intelligent technique for automatic detection of autism spectrum disorder from EEG. PLoS One **16**(6), e0253094 (2021)
32. Tong, L., et al.: Discovery of RNA and DNA viruses using next-generation sequencing: metagenomics. CVR Genom. (2023)
33. Wang, T., et al.: The human pangenome project: a global resource to map genomic diversity. Nature **604**(7906), 437–446 (2022)
34. Wani, A.K., et al.: Discovering untapped microbial communities through metagenomics for microplastic remediation: recent advances, challenges, and way forward. Environ. Sci. Pollut. Res. 1–24 (2023)
35. Yadav, S., Awasthi, P.: Diabetic retinopathy detection using deep learning and inception-v3 model. Int. Res. J. Mod. Eng. Technol. Sci. **4**, 1731–1735 (2022)

IEEE Access
Multidisciplinary | Rapid Reviews | Open Access Journal

Received 6 August 2023, accepted 29 August 2023, date of publication 4 September 2023, date of current version 8 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3311752

RESEARCH ARTICLE

Discovering Mathematical Patterns Behind HIV-1 Genetic Recombination: A New Methodology to Identify Viral Features

ANA GUERRERO-TAMAYO[1], BORJA SANZ URQUIJO[1], CONCEPCIÓN CASADO[2], MARÍA-DOLORES MORAGUES TOSANTOS[3], ISABEL OLIVARES[2], AND IKER PASTOR-LÓPEZ[1]

[1]Faculty of Engineering, University of Deusto, Bilbao, 48007 Biscay, Spain
[2]National Microbiology Center (NMC), Instituto de Salud Carlos III (ISCIII), Majadahonda, 81523 Madrid, Spain
[3]Faculty of Medicine and Nursing, University of the Basque Country (UPV/EHU), Leioa, 48940 Biscay, Spain

Corresponding author: Ana Guerrero-Tamayo (ana.guerrero@deusto.es)

This work was supported by the Research Training Grants Program, University of Deusto.

ABSTRACT In this article, we introduce a novel methodology for characterizing viral genetic features: the Unified Methodology of recombinant virus Identification (UMI). Our methodology converts genomic sequences into spectrograms, applies transfer learning using a pre-trained Convolutional Neural Network (CNN), and employs interpretability tools to identify the genomic regions relevant for characterizing a viral sequence as recombinant. The UMI methodology does not necessitate multiple sequence alignment or manual adjustments. As a result, it operates much faster, has low computational demands, and is capable of handling substantial amounts of data. To validate this, we applied UMI to one extensively studied and documented case: HIV-1 genetic recombination. We worked with all identified HIV-1 complete sequences (13554 sequences up to 2020), searching for mathematical patterns, signatures, that characterize an HIV-1 sequence as recombinant. CNN's hit rate (test accuracy) is 94%, with consistent and differentiated decision areas in each category. Using interpretability tools, we verified that the hot zones were similar for sequences of the same subtype and phylogenetic proximity. The leading areas for classifying a sequence as recombinant or non-recombinant are coincident with genomic regions that play a key role in genetic recombination processes. By applying UMI methodology we found that there is indeed a genome mathematical pattern that assesses an HIV-1 sequence as recombinant. In addition, we located its position. Considering expert knowledge, our results showed a substantial, robust and biologically-consistent hit rate. This type of solution can successfully guide the location and subsequent characterization of relevant areas, avoiding the heavy analysis of multiple sequence alignment and manual adjustments.

INDEX TERMS Convolutional neural network, deep learning, genetic recombination, genome mathematical pattern, genome mathematical signature, HIV-1.

I. INTRODUCTION

The capacity of viruses to vary severely complicates health care and prevention policies [1]. Mutation is one of the main mechanisms underlying viral diversity. It consists of introducing random errors during the replication process. These errors may result in differences in some viral characteristics [2]. Genetic recombination is another frequent phenomenon. This happens when two different strains of a virus, or two different viruses, co-infect a cell at the same time and the result of this replication is a new hybrid genome, called recombinant [3]. Recombinant viruses are usually life-incompatible. However, in some cases, they can produce new pandemic viruses. This is the case of SARS-CoV2, MERS-CoV, Influenza Virus or Human Immunodeficiency Virus Type 1 (HIV-1) [4].

Recombination in RNA viruses is an exchange of genetic information between genomic RNA molecules [5], which occurs during viral replication usually by a mechanism

The associate editor coordinating the review of this manuscript and approving it for publication was Kathiravan Srinivasan.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/



(a) HIV-1 env gp160 3D Protein (b) HIV-1 env gp160 FASTA File

FIGURE 1. 3D Graphical Representation of a Protein and its Coding Genomic Sequence. Fig. 1a shows the three-dimensional structure of the protein gp160 encoded by the *env* gene. Q79670 (ENV_HV1MV) Human immunodeficiency virus type 1 group O (isolate MVP5180) (HIV-1) Envelope glycoprotein gp160. See https://swissmodel.expasy.org/. Fig. 1b shows the FASTA file containing the genomic sequence of the HIV-1 *env* gene, complete cds (ID. L42371).

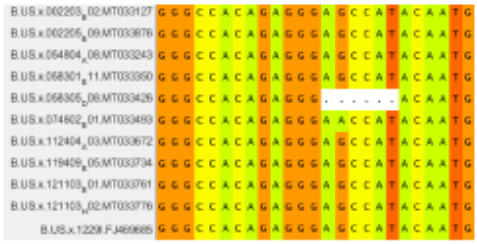


FIGURE 2. HIV-1 Multiple Sequence Alignment. By introducing GAPS (-), multiple sequence alignment algorithms search the better position for each nucleotide of each sequence so a direct comparison is possible. After applying these algorithms, experts usually must make manual adjustments to reach the optimal alignment. This figure was generated using MATLAB2021b Computational Biology Apps.

known as copy-choice which consists of a switch of the growing RNA from the genomic RNA template of a virus to the genomic RNA of another, resulting in a hybrid viral genome containing RNA fragments from both parental genomes [6]. Recombinant viruses have been detected with high frequency in various RNA viruses, such as picornaviruses [7] and coronaviruses [8]. But until now, HIV-1 is one of the viruses with the most recombinant sequences characterized, and a large amount of information on whole genome sequences is found in sequence databases.

The genome contains all the genetic information of an organism, coded in a sequence of four nucleotides (A, C, G, and T). It includes protein coding regions, regulatory regions and additional elements [9]. Fig. 1 shows an example of a genomic sequence (HIV-1 *env* gene) and its encoded protein.

Viruses resulting from genetic recombination can cause pandemics with new and potentially devastating effects [10], [11], [12], [13]. The HIV-1 pandemic caused a dramatic social impact in the 1980s and, consequently, high funding for the development of the AIDS cure. Despite all the efforts, HIV-1 variability severely affects the current lack of effective treatments and vaccines (genetic recombination is one of the determinant changeability factors). Therefore, it is necessary to understand the process of genetic recombination and identify all the possible mechanisms that determine its occurrence and its outcome. The importance of genetic recombination in viral diversification [14] renders the development of new methods to detect recombinant viruses very significant, and the search for mathematical patterns (inherent mathematical regularities and structures in the DNA sequence). Applying emerging technologies, such as Deep Learning, can help improve the current level of microbiological knowledge with new approaches. HIV-1 is an excellent study model because of its high number of identified and labelled recombinant sequences.

Traditional methodologies for the analysis and comparison of genome sequences are based on multiple sequence alignment. This requires introducing gaps to place each nucleotide, letter by letter, in the same position and then make a direct comparison of the aligned sequences (Fig. 2). There are several multiple sequence alignment algorithms, including Clustal [15], TCoffee [16], Probcons [17], MAFFT [18], Muscle [19], MSAprobs [20], and GLProbs [21]. These algorithms are complex, computationally expensive, and do not allow the handling of large amounts of data; therefore, they provide sub-optimal solutions [22].

There is still no efficient solution for the analysis of large numbers of sequences and no objective method has been approved by the entire scientific community. The best approaches are to apply several algorithms, select their common results, and then perform manual adjustments based on expert microbiological knowledge.

We propose a new methodology using a genomic representation independent of sequence length, and the implementation of Deep Learning algorithms to model the categorization of recombinant sequences [23] with interpretability tools that point to the genomic areas involved in genetic recombination. The Unified Methodology of recombinant virus Identification, UMI Methodology (see Section III-D), achieves excellent and robust results, with the added complexity that both recombinant and non-recombinant sequences also undergo mutations. Thus, the complexity is even greater. We applied several interpretability tools to identify where the CNN relies on to determine the classification of each sample.

- This article outlines the detection of mathematical patterns, mathematical signatures, embedded in the genome that characterize the Human Immunodeficiency Virus Type 1 (HIV-1) as recombinant.

- UMI allows not only the detection of the recombinant feature of each sample but also targets the genome areas where

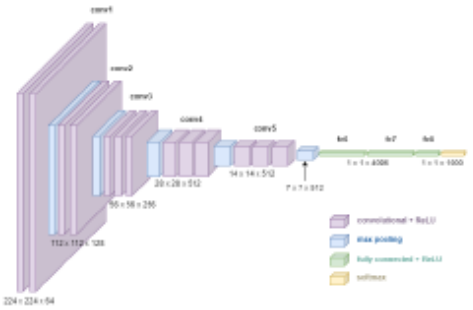


FIGURE 3. Pre-trained Convolutional Neural Network VGG-16 Architecture Schematic. Once the image "enters" the CNN, convolution phase begins. This consists of taking groups of nearby pixels from the input image and transforming them mathematically (scalar product) to filter different features. Pooling phase emphasizes the most important features detected by each convolutional filter. The output of pooling phase is the pooled feature map (numerical matrix). Next step converts this feature matrix to a vertical vector, which is the input to the fully connected layer. Its neurons detect certain features, then communicate their values to softmax neurons and they check out each feature and decide about its relevancy. Finally, each output neuron gives a classification result and a confidence score (how sure this neuron is about its decision). This figure is based on the diagram drawn by Kenneth Leung, which illustrates VGG-16 architecture [30].

this characterization is located. This will help to gain more knowledge about the processes of genetic recombination in HIV-1.

The remainder of this paper is organized as follows. Section II describes the background of the UMI methodology. Section III contains a complete description of the designed method. Section IV details the results. Section V (Biological Consistency Analysis) delves into the peculiarities of the HIV-1 replication cycle and the microbiological meaning of the obtained results, with multiple examples of coherence between them and expert microbiological knowledge. This document ends with Conclusions, summarizing the microbiological consistency of our results, and Future Studies.

II. BACKGROUND

Convolutional Neural Networks (CNN) have demonstrated high efficacy in detecting patterns in images not visible to the human eye [24], [25], [26], [27]. They are even applied in the detection of HIV outbreaks using genetic data [28]. Fig. 3 depicts the architecture of one of the most widely used CNNs, VGG-16 [29].

The optimal use of this potential requires biologically consistent graphical conversion of genomic sequences [31]. Genome frequency-domain analysis is a promising graphical representation for detecting biological patterns [32], [33], [34].

The application of signal processing tools to genomic sequences of different organisms has shed light on the relationship between nucleotide periodicity and various features [35], such as the relationship between coding areas and sensitivity at frequency $f = 1/3$ [36], non-coding RNA molecules location [37], GpC islands detection or microsatellites identification [38].

In addition, these tools have the advantage of not requiring multiple sequence alignments to compare different genomes. This reduces the computational resources needed and eliminates the useless information of aligned sequences [39].

Specifically, spectrogram imaging, as a graphical representation of the Fast Fourier Transform (FFT), is increasingly used to analyze genomic sequences [40]. The first step in generating the spectrogram of a sequence was to convert that sequence into four digital signals (U_a, U_g, U_c, U_t), which are activated by the presence of a nucleotide type (A, G, C, T) at each position.

$$U_\alpha(x_i) = \begin{cases} 1, & \text{if } x_i = \alpha \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

The next step was to calculate the spectrogram for each of the four binary signals obtained from this decomposition.

The Fast Fourier Transform (FFT) corresponding to each of these signals was performed, and this is a computationally efficient method for the calculation of the Discrete Fourier Transform (DFT) [41].

$$X(k) = \sum_{n=0}^{N-1} U_\alpha(x_i) W_N^{kn} \quad 0 \leq k \leq N-1 \tag{2}$$

where:

$$W_N = e^{-j2\pi/N} \tag{3}$$

FFT algorithms decompose an N-point DFT into smaller DFTs [42] using sliding windows.

Although the application of Deep Learning Tools in health is still controversial [43], multiple studies have demonstrated that the performance of Deep Learning Tools using spectrograms renders good results in the interpretation of medical results [44], [45], [46], [47], [48], [49].

Calculation of the spectrogram of a genomic sequence makes it possible to obtain a graphic representation of the significance of each periodicity (frequency) at each position (assimilating the concept of time to position) of the genome.

The x-axis represents the position of each nucleotide. The y-axis is the frequency range and the z-axis is the value of the FFT. The spectrogram is a two dimensional representation obtained by replacing the z-axis with a color palette according to the amplitude of the FFT of the position of a nucleotide (x-axis) at a given frequency (y-axis) [50].

Applying Transfer Learning on a pre-trained CNN allows extending its potential to new types of images and new domains such as medical diagnostics: prediction of multiple diseases from chest X-ray images [51], classification of abdominal ultrasound images [52], diabetic retinopathy images [53], cellular morphological changes to identify cellular phenotypes (incipient cancer process and so on) [54] or detection of Major Depressive Disorder (MDD) by observing

electroencephalogram (EEG) signals transformed to spectrograms [55].

Therefore, the application of pre-trained CNNs as a tool for the classification of genomic sequence spectrograms can yield interesting and positive results. In this case, the HIV-1 compete sequences were classified as recombinant or non-recombinant.

The spectrogram is a visual tool that represents the spectral composition of a signal as a function of time using the Fast Fourier Transform (FFT). This image contains a significant mathematical component.

The identification of invisible mathematical patterns that determine the recombination of a sequence can potentially offer increased robustness compared to traditional algorithms used for detecting genetic recombination events in nucleotide sequences. These traditional methods rely on statistical approaches, which may raise concerns regarding their performance and reliability [56].

Interpretability tools can provide information on areas that are determinants of subsequent classification [57], [58], [59]. The results of the interpretability analysis make it possible to evaluate which factors determine whether an HIV-1 sequence is recombinant or not.

Three different interpretability tools can identify different areas of interest, and the results can be combined to obtain the maximum information:

1) Grad-CAM (Gradient-weighted Class Activation Mapping). It uses gradients of the classification score with respect to the final convolutional feature map [60]. Grad-CAM is intuitive, but its resolution is low.

2) LIME (Local Interpretable Model-agnostic Explanations). It computes feature importance by segmenting an image into several features and generating "artificial" observations by randomly including or excluding these features [61], [62]. Its accuracy can vary but it exhibits good visual interpretability.

3) Gradient Attribution. It accurately identifies the specific pixels in the image that are crucial for the decision-making process of the Deep Learning tool. It returns a pixel map that highlights the most influential pixels on the class score when changed [63], [64]. Its accuracy is very high; however, visually, it is very noisy.

In terms of accuracy and ease of use for the human eye, the advantages and disadvantages of these three tools make them particularly interesting to work with simultaneously, enabling the comparison of results and obtaining complementary data on features and patterns.

III. MATERIALS AND METHODS

All experiments ran in this equipment:

- Processing Unit: Intel(R) Xeon(R) Gold 5220R CPU @ 2.20 GHz.
- Installed RAM: 256 GB usable.
- Operative System: Windows 10 Education. Version: 21H1.
- GPU: NVIDIA RTX A6000. Total memory: 179.451 GB. VRAM: 48.571 GB.

TABLE 1. Dataset Structure. Non-Recombinants account for 82.36% of the total (13554 complete HIV-1 sequences until 2020). Recombinants are the 17.64%. Percentages in parentheses are calculated with respect to the total number of sequences of training, validation and test sets (11267, 1987 and 300 sequences).

Dataset	Recombinant	Non-Recombinant	TOTAL
Train	1903 (16.89%)	9364 (83.11%)	11267
Validation	338 (17.01%)	1649 (82.99%)	1987
Test	150 (50.00%)	150 (50.00%)	300
TOTAL	2391	11163	13554

A. HIV-1 COMPLETE GENOMIC SEQUENCES COMPENDIUM

HIV-1 genome sequences were downloaded from the Los Alamos National Laboratory HIV Database (https://www.hiv.lanl.gov/), all subtypes, complete genome. Recombinant sequences are named as Circulating Recombinant Forms (CRFs). The entire genome of these viruses has been demonstrated to exhibit genetic recombination. This means that certain regions of the genome cluster with one subtype, whereas other regions cluster with a different subtype. To facilitate the quick identification of subtypes, CRF expression is omitted, referring to recombinants only by acronyms after the hyphen symbol. Therefore, CRF29-BF1 is referred to as BF1.

To construct the dataset, we labelled all CRF sequences as recombinant and non-CRFs as non-recombinant. Subsequently, we randomly distributed these sequences into training, validation, and test folders according to the guidelines outlined in Table 1.

As shown in Table 1, the test dataset ponderation (50 % / 50 %) deliberately differs from that corresponding to train and validation sets (17 % / 83 %) to detect under-fitting biases during the training phases [65].

B. SPECTROGRAM GENERATION

Working with two different graphical representations of the spectrogram of a sequence allows us to obtain complementary information that may clarify the possible pattern that determines whether an HIV-1 genome is recombinant.

The first one, called **Concatenated Spectrogram**, displays the concatenation of the spectrogram for each of the four types of nucleotides on the x-axis. A schematic is shown in Fig. 4a.

This representation system can determine the most influential nucleotide types for the characterization of a virus as recombinant. However, it is difficult to determine accurately the involved area of the genome.

The second one is called **Superposed Spectrogram**. In this representation, the z-axis is the arithmetic sum of the values along the z-axis for each of the four nucleotide types:

$$S = S_a + S_g + S_c + S_t \tag{4}$$

Fig. 4b shows the graphical concept of this type of representation, which allows us to more effectively specify the influential zones of the genome.

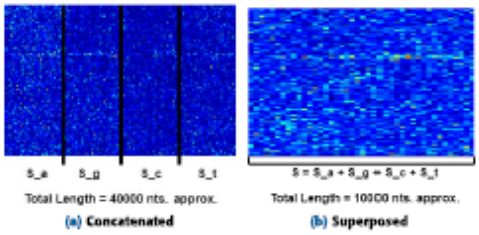


FIGURE 4. Spectrograms Schemes. The x-axis represents the position of each nucleotide. Therefore, in Fig. 4a the length is four times the sequence length owing to the concatenation of the four spectrograms. In Fig. 4b the length of x-axis matches the genome length. In both cases, the y-axis represents a frequency range from 0 to 0.5. The z-axis is the spectrogram calculation.



FIGURE 5. Jet Color Map. Blue color corresponds to lowest sensitivities and red to the highest ones. See jet colormap at https://matplotlib.org.

To work with two-dimensional spectrograms, the z-axis is assimilated to a "jet" color palette. See Fig. 5. Using the matplotlib.pyplot.pcolormesh function, we mapped the minimum value of the FFT to deep blue (0) and the maximum value to deep red (1), automatically generating color assignments for each intermediate value of the FFT. By default, the data range is mapped to the colorbar range using linear scaling.

Each spectrogram is identified by the reference of the corresponding sequence.

We generated spectrograms of both datasets using Python, the Scipy library, scipy.signal.spectrogram. To avoid biases and bad practices during CNN training, we omitted axes, margins and any other elements that may distort the operation of the network. Therefore, the image is limited to the spectrogram only [66], [67]. We performed all the experiments by splitting the entire dataset into training, validation and test folders. The complete dataset is available in the Supplementary Information - Dataset folder. CNNs trained with the same information in concatenated and superposed. This improves the comparability of results by eliminating GAPs related to different learning data.

C. PRE-TRAINED CNN SELECTION CRITERIA

Once both spectrogram datasets were generated, the next step was to perform a test bench to select not only the CNN with the best performance, but also the set of hyperparameters that achieved the best and most balanced accuracy levels between non-recombinant and recombinant viral genomes.

CNNs of choice were:

- VGG16 [68].
- ResNET - 101 [69].
- Inception V3 [70].

The experiments were performed using the MATLAB2021b App Deep Learning Designer. These CNNs are pre-trained with millions of images from the ImageNet database, and they can classify images into 1000 categories. More information about the pre-trained convolutional neural networks in MATLAB is available at https://www.mathworks.com.

Through Transfer Learning, the CNN learns the specific features of both datasets. Freezing the weights of all intermediate layers intact, we replaced the last fully connected layer, specifying the new number of categories to analyze. We also replaced the classification layer of the pre-trained model. In this way, we extract all previous knowledge from the other datasets and adapt it to our new problem.

As demonstrated by preliminary tests, Data Augmentation degraded the performance significantly, probably because it altered the biological meaning of the sequence. Applying this technique would be counterproductive, particularly if rotations are generated. This is because the biological meaning of the image is determined by its exact orientation. Therefore, we did not apply this technique.

The following test bench was performed to standardize the tests on concatenated and superposed spectrogram datasets with the most direct comparability possible between different CNNs (Table 2).

TABLE 2. Test Bench. It was applied under the same conditions to all the selected CNNs. It consists of different combinations of Learning Rate (0.0001 and 0.0100), Batchsize (52 and 128) and Epochs (10 and 30).

Test	Learning Rate	Batchsize	Epochs
1)	0.0001	52	10
2)	0.0100	52	10
3)	0.0001	52	30
4)	0.0100	52	30
5)	0.0001	128	10
6)	0.0100	128	10
7)	0.0001	128	30
8)	0.0100	128	30

1) BEST PERFORMANCE BASED DECISION CRITERIA

The criteria for determining the best results should not be solely based on the total number of hits. The four selected performance metrics of the CNN are as follows:

- Validation Accuracy.
- Training Time.
- Confusion Matrix Results.
- AUC.

Training Time is a measure of the required computation. For similar Validation Accuracy and AUC values, the decisive criterion was the maximum Test Accuracy in both recombinant and non-recombinant categories (Confusion Matrix Results). The best performance was achieved when these hit rates in both categories were balanced and optimal.

According to this criterion, the CNN that performs best in both datasets (concatenated and superposed) was the one selected. The balance should not only exist in both categories, but also in both datasets.

The results of all experiments and training curves are available at the Supplementary Information - Training Curves folder.

2) ROBUSTNESS CRITERIA

To determine the robustness of the results, the same optimized training with exactly the same dataset was repeated three times. If the results of these three attempts are similar, the system is considered to be robust.

3) INTERPRETABILITY TOOLS' RESULT GENERATION

To locate the positions of these mathematical patterns that enable the CNN to classify a genomic spectrogram as recombinant, we applied MATLAB functions such as gradCAM, imageLIME, and gradientMap.

These three functions generate a scoremap, which is a numerical matrix highlighting the relevant areas that contribute to the classification task.

These hot zones are output in image format. To achieve this, they overlaid a color map onto the original image, making the hot regions clearly visible.

In the case of Grad-CAM and LIME, a 256-color Jet Color Map is applied (see Fig. 5), from 0 (deep blue) to 1 (deep red). Gradient Attribution provides a pixel-resolution map that shows the most relevant pixels to the network's classification. This map is scaled with 255 colors from 0 (white) to 1 (black).

The hot maps of the complete dataset are available at the Supplementary Information - Hot Zones folder.

4) PHYLOGENETIC TREE AND COMPARISON WITH INTERPRETABILITY TOOLS' PATTERNS

A comparison was made between sequences with similar interpretability patterns and the phylogenetic tree of the 13554 HIV-1 complete sequences [71], [72]. First, we downloaded the HIV-1 complete consensus multiple sequence alignment from the Los Alamos National Laboratory HIV-1 Database (until 2020). These alignments contain a single sequence per patient [73]. Thus, we avoid biases due to comprehensive data from a few sources and to make the resultant phylogenetic tree more user-friendly.

An analysis was then performed between the similarity level of the patterns indicated by the interpretability tools, and the position of the sequences in the phylogenetic tree (the closeness between them). A phylogenetic tree of these multiple sequence alignments and phylogenetic distances between two sequences (Path Length), were calculated using MATLAB2021b - Computational Biology Apps.

When the analyzed sequence did not appear in the HIV-1 consensus multialignment, the representative sequence from the same patient was taken as a reference. Sometimes, it was not possible to locate the position of some sequences in the phylogenetic tree, because alignments lacked either the sequence itself or the patient's reference sequence. The patient was not identified in some sequences.

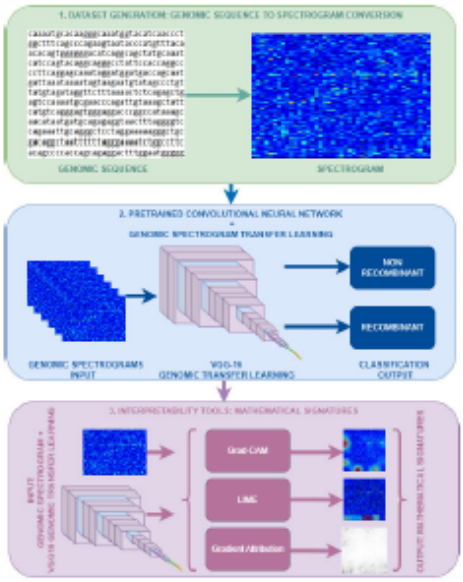


FIGURE 6. UMI Methodology Block Diagram. The first phase is a conversion of genomic sequences to biologically-coherent images, this is, genomic spectrograms. The second phase is applying transfer learning to a pre-trained Convolutional Neural Network (CNN) with this dataset of frequency-domain images, so the CNN can detect features in this type of images. Once the CNN is trained, interpretability tools allow to identify which areas the CNN looks at to classify each image. The result of the third and final phase is the location of mathematical signatures related with a viral feature.

The traceability of intra-patient sequences (virus sequences produced in the same patient's organism) was determined by the Patient Code (ID) field.

We loaded the HIV-1 complete consensus multiple sequence alignment into the Sequence Alignment Tool. We automatically generated the phylogenetic tree using the "View Tree" option. The phylogenetic path and distance are displayed by selecting two sequences on the tree itself.

This phylogenetic tree is available at the Supplementary Information - Phylogenetic Tree folder.

D. UMI METHODOLODY BLOCK DIAGRAM

Our new methodology, the Unified Methodology of recombinant virus Identification (UMI), consists of three phases, as described in Fig. 6.

IV. RESULTS

A. CNN SELECTION

To improve the reading and understanding of Confusion Matrix Results in the following sections, only the number of hits and failures are shown, according to Fig. 7).

TABLE 3. Best Results Abstract. Once all test combinations were performed, and considering the selection criteria, the best performance and the best hyperparameters for each of the pre-trained CNNs are shown.

Concatenated Spectrogram Dataset							
CNN	Hyperparameters	Validation Accuracy	Training Time	Training Time per Epoch	Test Accuracy	Confusion Matrix Results	AUC
VGG-16	Learning Rate: 0.0001 Batchsize: 52 Epochs: 10	94.92%	17 min 36 sec	1 min 45.60 sec	96.00%	145 – 5 13 – 137	0.9772
Resnet-101	Learning Rate: 0.0100 Batchsize: 52 Epochs: 10	94.82%	41 min 7 sec	4 min 6.70 sec	94.33%	149 – 1 16 – 134	0.9928
Inception-V3	Learning Rate: 0.0100 Batchsize: 52 Epochs: 30	94.82%	148 min 1 sec	4 min 56.03 sec	96.33%	146 – 4 7 – 143	0.9943
Superposed Spectrogram Dataset							
CNN	Hyperparameters	Validation Accuracy	Training Time	Training Time per Epoch	Test Accuracy	Confusion Matrix Results	AUC
VGG-16	Learning Rate: 0.0001 Batchsize: 52 Epochs: 10	92.90%	15 min 59 sec	1 min 35.90 sec	94.00%	143 – 7 11 – 139	0.9888
Resnet-101	Learning Rate: 0.0100 Batchsize: 52 Epochs: 30	96.80%	136 min 24 sec	4 min 32.80 sec	92.67%	143 – 7 15 – 135	0.9699
Inception-V3	Learning Rate: 0.0100 Batchsize: 52 Epochs: 30	96.68%	148 min 6 sec	4 min 56.20 sec	95.00%	149 – 1 14 – 136	0.9879



FIGURE 7. Confusion Matrix Scheme. The top row refers to non-recombinants. Hits are on the left and misses area on the right. The bottom row refers to recombinants. Failures are on the left and hits area on the right.

As shown in Table 3, both for concatenated and superposed spectrograms:

VGG-16 achieved better performances with a Learning Rate = 0.0001. The optimal Learning Rate of ResNET-101 and Inception V3 is 0.0100.

All the three CNNs improved with Batchsize = 52 vs. Batchsize = 128.

These three CNNs exhibited the same behavior as the number of Epochs increased. The success rate increased directly proportional to the number of Epochs classifying non-recombinant sequences, achieving 100%. However, for recombinant sequences, the success rate increased until it reaches its maximum, after which it began to decrease.

VGG-16 yielded better results with Epochs = 10 vs. Epochs = 30. ResNET-101 performed better with Epochs = 10 when working with concatenated spectrograms. However, if it worked with superposed spectrograms, Epochs = 30 was a better choice. For both datasets, Inception V3 slightly improved its performance for Epochs = 30.

VGG-16 training time was the lowest. Inception V3 developed the slowest learning process.

Because Res-NET 101 obtained the lowest accuracy for both datasets, it was discarded. Inception V3 provided better results for concatenated spectrograms but its accuracy was slightly lower for superposed recombinant spectrograms.

Therefore, in accordance with the performance measurement parameters defined in Section III-C1, VGG-16 was the CNN selected for this research. It yielded more balanced hit rates on the test dataset between non-recombinant and recombinant categories, and its training times were substantially lower. The best success rate and the most balanced results were achieved with 10 Epochs.

The best VGG-16 configuration for both concatenated and superposed spectrograms was:

- Learning Rate = 0.0001
- Batchsize = 52
- Epochs = 10

B. ROBUSTNESS

We performed the robustness test in accordance with the specifications described in Section III-C2. The same experiment, which was performed three times, achieved very similar results. Thus, the robustness of the method was demonstrated.

The results were similar in all three cases for both datasets (Tables 4 and 5); thus, the performance was robust. The most successful training was selected according to the criteria outlined in Section III-C1. In both datasets, Test 1 achieved the best hit rates and moreover the best balance between recombinants and non-recombinants.

C. HOT ZONES DETERMINATION

Interpretability tools allow us to identify where a CNN determines the recombinant character of a sequence. These spectrogram areas are called hot zones (where the CNN looks at to classify a sequence).

Figs. 8 and 9 represent the most repeated interpretability hot zones. In order to extract the most recurrent hot zones with the highest level of accuracy, we analyzed the interpretability results of each sequence in the test set individually (a total of 300 sequences). For the test set sequences that the network correctly predicted, we identified the most frequently occurring hot areas as well as the secondary ones. For the sequences where the network failed, we determined the most frequently occurring erroneous hot zones as well as the secondary ones.

The Grad-CAM and LIME results, although less accurate, easily render hot zones to the naked eye. So, they are more

TABLE 4. VGG-16 Concatenated Spectrograms Three Identical Test Results. These tests were performed with the best resulting configuration: Learning Rate= 0.0001, Batchsize=52 and N°Epochs=10.

Concatenated Spectrogram Dataset					
	Validation Accuracy	Training Time	Test Accuracy	Confusion Matrix Results	AUC
Test 1	94.92%	17 min 36 sec	94.00%	$\begin{matrix}145-5\\13-137\end{matrix}$	0.9772
Test 2	95.82%	15 min 51 sec	92.67%	$\begin{matrix}144-6\\16-134\end{matrix}$	0.9828
Test 3	95.02%	17 min 36 sec	90.33%	$\begin{matrix}142-8\\21-129\end{matrix}$	0.9714

TABLE 5. VGG-16 Superposed Spectrograms Three Identical Test Results. These tests were performed with the best resulting configuration (same as concatenated): Learning Rate= 0.0001, Batchsize=52 and N°Epochs=10.

Superposed Spectrogram Dataset					
	Validation Accuracy	Training Time	Test Accuracy	Confusion Matrix Results	AUC
Test 1	92.90%	15 min 59 sec	94.00%	$\begin{matrix}143-7\\11-139\end{matrix}$	0.9888
Test 2	92.25%	16 min 9 sec	93.33%	$\begin{matrix}144-6\\14-136\end{matrix}$	0.9904
Test 3	94.46%	16 min 39 sec	93.33%	$\begin{matrix}144-6\\14-136\end{matrix}$	0.9872

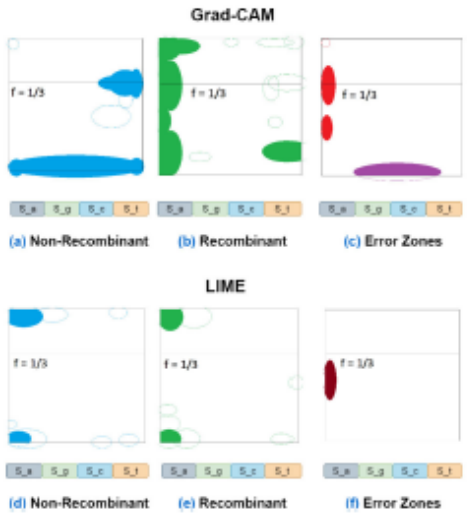


FIGURE 8. Concatenated Hot Zones. The most common hot areas are color-filled. The not filled ones represent hot areas with lower repeatability. In the case of Grad-CAM,(a) shows the hot areas where the CNN looked at to classify a sequence as non-recombinant. (b) shows those corresponding to the recombinant category. (c) indicates where did the CNN look at in misclassification cases. "Recombinant classified as non-recombinant" hot zones are purple-colored and "Non-recombinant classified as recombinant" hot zones are red-colored. In the case of LIME, (d) shows the hot areas where the CNN looked at to classify a sequence as non-recombinant. (e) shows those corresponding to the recombinant category. When the CNN fails, the LIME hot zones are indicated in (f) in brown.

suitable for the first approximation to the zones of interest. By contrast, Gradient Attribution provides a difficult visual interpretation.

Grad-CAM hot zones were analyzed, as shown in Fig. 8a., the low-frequency range of the four nucleotides plays an important role in determining a sequence as non-recombinant, as well as the frequencies close to the one-third for the T nucleotide. In recombinant concatenated spectrograms (Fig. 8b.), the leading Grad-CAM hot zone pivots around the left border (A nucleotide area) and in the lower third of the right border (T nucleotide area). There is a zone of interest common to non-recombinants and recombinants, located on the right border at the height of one-third of the T-nucleotide area. In the misclassification cases, the CNN looked at areas corresponding to the opposite category. The most common error zones are as shown in Fig. 8c.

In the case of LIME applied to concatenated spectrograms (8d. and 8e.), LIME targets the left corners (A nucleotide high and low frequencies) both for non-recombinant and recombinant. In several failure cases, the CNN was fixed in places that did not coincide with hot zones. This error zone is located around the middle of the left axis (average A nucleotide frequencies), as shown in 8f.

For non-recombinant superposed spectrograms, Grad-CAM hot zones are shown in Fig. 9a. The leading one corresponds to the high and low-frequency areas, corresponding mainly to 5'LTR and 3'LTR (the four corners). Secondary hot zones are concentrated at the upper and lower edges and at high and low frequencies.

Analyzing recombinant superposed spectrograms, Grad-CAM main influential areas (Fig. 9b.) are located in low frequencies and one-third frequency of 5'LTR.

In non-recombinant superposed spectrograms, LIME is overwhelmingly fixed in the lower-left corner (low-frequency area at the beginning of the sequence, 5'LTR primarily). To a lesser extent, in the upper-left corner (high frequency at the beginning of the sequence) and in the lower-right corner

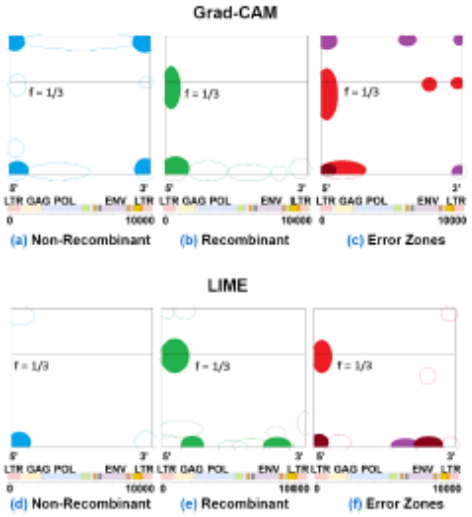


FIGURE 9. Superposed Hot Zones. The most common hot areas are color-filled. The not filled ones represent hot areas with lower repeatability. (a) shows the hot areas where the CNN looked at to classify a sequence as non-recombinant, and (d) presents the corresponding ones to LIME. (b) and (e) shows those corresponding to the recombinant category of Grad-CAM and LIME, respectively. (c) and (f) indicate where did the CNN look at in misclassification cases. "Recombinant classified as non-recombinant" hot zones are purple-colored and "Non-recombinant classified as recombinant" hot zones are red-colored. Brown-colored area refers to the common error zone between the two types of misclassification.

(low-frequency zone at the end of the sequence, 3'LTR). See Fig. 9d.

The significant regions in the superposed spectrograms of recombinant sequences are located at approximately one-third of the frequency range at the beginning of the sequence and some areas that approximate the *gag* and *env* genes. See Fig. 9e.

In superposed spectrograms, Grad-CAM and LIME indicate that the CNN misclassified looking at the hot zones of the opposite category. See Fig. 9c. and 9f.

The behaviors of the three interpretability tools are similar in concatenated and superposed spectrograms. Both methods obtain robust results and similar zones of interest in the nearby sequences.

V. BIOLOGICAL CONSISTENCY ANALYSIS

As explained in Section III-C4, after confirming that the CNN could differentiate between recombinant and non-recombinant sequences based on distinct decision patterns, the next step was to deepen the biological interpretation of this behavior. We generated the phylogenetic tree of all the complete HIV-1 sequences up to 2020. Next, we compared the similarity level of the hot zones with phylogenetic distance. We also compared hot zones with microbiological expert knowledge to validate the biological coherence or our results.

To facilitate the biological meaning of the hot areas, we provide a set of all the figures in Tables 6 to 11 with their scaled reference. See Supplementary Information -Hot Zones Reference Information.

A. HIV-1 MICROBIOLOGICAL CONSIDERATIONS

HIV-1 is a retrovirus, the genome of which contains two copies of single stranded RNA. The hallmark of retroviral replication is reverse transcription [74], that generates a linear DNA duplex from single stranded genomic RNA. Reverse transcription occurs via a series of steps, including two switches of nascent DNA, from one end to the opposite end of the RNA template [75]. As a result, two identical sequences called Long Terminal Repeats (LTRs), are formed at each end of DNA. After reverse transcription, the viral DNA is integrated into the genome of the host cell.

LTRs play a vital role in the initiation and regulation of viral transcription. Despite the identity of the sequence, each one has different functions. 5'LTR contains the signals for the initiation and regulation of viral transcription. 3'LTR performs in post-transcriptional processes. 3'LTR has an efficient polyadenylation site (adding a polyA tail to messenger RNA), and it is fundamental in the editing of viral transcripts [76], [77].

The two copies of contained RNA, as well as the switches occurring during reverse transcription, may be related to the high level of genetic recombination that occurs in an HIV infectious process [78]. Reverse transcriptase indistinctly uses portions of these two RNA molecules contained in each virion as a template [79].

HIV-1 is classified into four groups named M, N, O and P. Group M, which is the most widespread in the world, is divided into nine subtypes: A, B, C, D, F, G, H, J and K. Multiple recombinants between subtypes are currently circulating; they are called Circulating Recombinant Forms (CRF) [80].

B. ANALYSIS OF SUCCESSFUL CLASSIFICATIONS

The hot areas of sequences belonging to the same subtype are usually similar. In most cases, the closer the phylogenetic distance, the greater the similarity of the hot zones. A representative example is subtype 22-01A1 AY371165 and KF716462 (considering its reference sequence JN864058). See Table 6. The phylogenetic closeness between the two sequences is proven in Fig. 10. This path is highlighted in red.

The similarity level is higher for intra-patient sequences [81]. Therefore, interpretability tools are expected to yield even more similar results. As an example, Table 7 shows that there is a high similarity between the hot zones of concatenated and superposed spectrograms at sequences KR820306

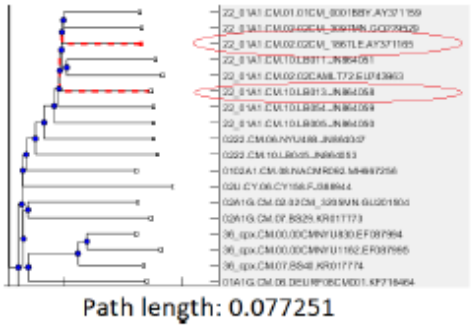
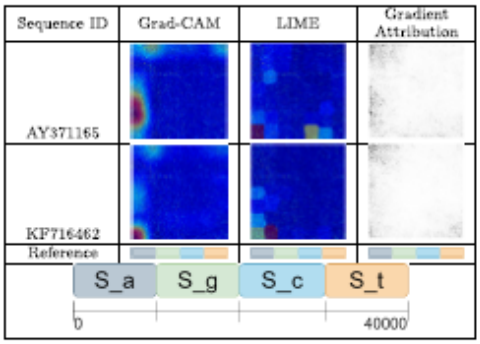


FIGURE 10. Phylogenetic Tree AY371165 and JN864058. These sequences are included in the recombinant subtype 22-01A1, which contains 21 sequences. As shown, both are phylogenetically close sequences, with a path distance of 0.077251. Their last common node results in a clade of seven sequences.

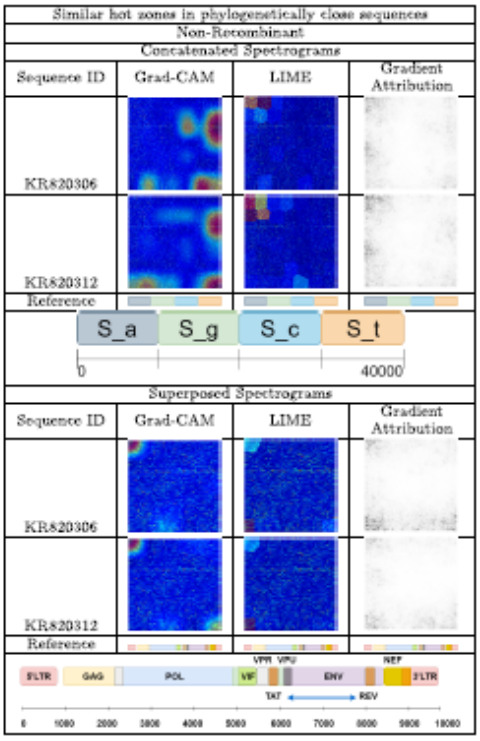
TABLE 6. Comparison of Interpretability Recombinant Concatenated Spectrograms - Subtype 22-01A1. AY371165 and KF716462 have close phylogenetic positions and their hot zones are similar. Grad-CAM points to low and one-sixth frequencies of A. Another hot zone is at high frequencies of A and the beginning of G. LIME basically points to low frequencies of A. Gradient Attribution is noisy but visually centered in A nucleotide area and T (low frequencies).



and KR820312, both belonging to subtype C and intra-patient (Patient ID ZM331F).

In some cases of recombinant spectrograms, hot areas are slightly more volatile, particularly in the concatenated graphical representation. In the case of AY371129 and AY371130, both from subtype 02-AG, their concatenated spectrograms (Table 8) show divergences between the interpretability hot zones despite being phylogenetically close sequences. Their path length is 0.085774. However, in the case of superposed spectrograms, their hot zones are similar. Given the phylogenetic closeness between both sequences, a higher level of similarity between hot zones in concatenated spectrograms is expected. Consistency occurs in the case of superposed spectrograms. This fact requires further research.

TABLE 7. Similar Hot Zones in Phylogenetically Close Non-Recombinant Sequences. KR820306 and KR820312 are intra-patient subtype C sequences. As expected, their decision areas are absolutely similar. In concatenated spectrograms, Grad-CAM highlights low and one-third frequencies of T nucleotide, with certain importance of one-third in C spectrogram. Another hot area is located at low frequencies of A and maybe G. LIME centers in high frequencies of A. Gradient Attribution shows more activity at A and T spectrograms. In superposed spectrograms, Grad-CAM fixes in high frequencies of 5'LTR in both cases. KR820312 has a secondary area at low frequencies of 3'LTR. A very peripheral area can be perceived at *pol* gene. LIME highlights low frequencies of 5'LTR, followed by high frequencies of this area. Gradient Attribution seems to emphasize high and low frequencies of 5'LTR and *gag*, and low frequencies of some area near *vif*, *env* and, specially, 3'LTR.



In general, there is a high similarity level of the hot zones as a function of phylogenetic closeness in concatenated and superposed spectrograms. In some cases, this match is even slightly higher in the superposed case.

An additional example is the case of intra-patient recombinant sequences MW443219 and MW443225 (Patient Code 40503), belonging to subtype 01-AE (Table 9).

As shown in Table 9, intra-patient (Patient ID 40503) MW443219 and MW443225 sequences (subtype 01-AE) develop the same high concordance in concatenated and superposed spectrograms.

Thus, based on the findings of the interpretability tools, the CNN exhibits consistent behavior for both datasets. This

TABLE 8. Comparison of Interpretability Recombinant Spectrograms - Subtype 02-AG. Grad-CAM hot zones are different in concatenated spectrograms. Hot zones in AY371129 are around high frequencies of A and one-ninth and high frequencies in T. However, AY371130 most relevant zones are one-sixth and high frequencies of A. LIME has no deviations. The most determinant areas are high and low frequencies in A. In Gradient Attribution, the most important region seems to be the high frequency range of A. In superposed, Grad-CAM relevant zone is located around one-third frequency of 5'LTR. Secondary areas are surrounding low frequencies of 5'LTR. LIME results are more dispersed. In AY371129, a zone close to one-ninth frequency at 5'LTR outlines. Less relevant sites are identified at low frequencies and one-third frequency at *gag*. In AY371130, the most determinant zone is one-third of 5'LTR. There is a secondary area in the low frequency range of the central part of the genome. Gradient Attribution point clouds are visually different, more noisy in the case of AY371129. In AY371129, the most relevant areas are low frequencies in 5'LTR and the first half of *gag*, *vif* and surroundings and 3'LTR. Another remarkable zone is between one-third and high frequencies of 5'LTR and *gag*. In AY371130, Gradient Attribution shows low activity, highlighting low frequencies of *gag* and maybe 5'LTR, around *vif* and, very weak, 3'LTR.

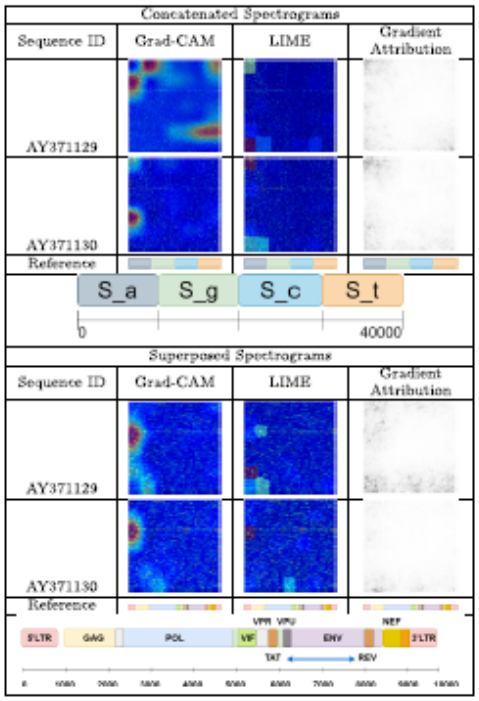
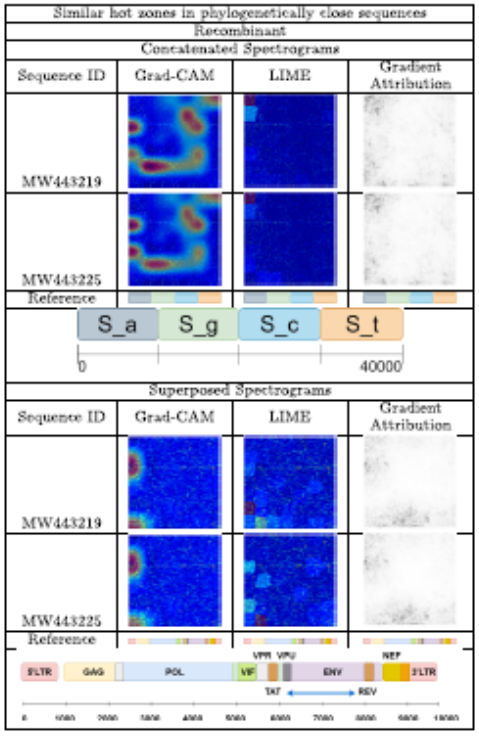


TABLE 9. Similar Hot Zones in Phylogenetically Close Recombinant Sequences. MW443219 and MW443225 are intra-patient recombinant sequences of subtype 01-AE. The most remarkable feature of Grad-CAM is the high similarity level of the hot zones in concatenated spectrograms. The main key areas are one-third, one-sixth and one-ninth frequencies in A, one-ninth frequency in G, some points near one-third, one-sixth and one-ninth in C, and one-sixth, one third and high frequencies of T. However, LIME fixes in high frequencies of A. Noisy results of Gradient Attribution seem to enhance low and high frequencies of A, some areas of low frequency range in G and C and low and one-third frequencies of T. In superposed spectrograms, Grad-CAM points to one-third and low frequencies of 5'LTR. LIME focuses at frequencies near one-ninth of 5'LTR and, low frequencies of *gag*. One minor zone is one-third in 5'LTR. Hot points in Gradient Attribution are grouped in one-third and high frequencies of 5'LTR, low frequencies in the central area of the genome and low frequencies around 3'LTR.



is particularly true for non-recombinant sequences, as hot zones tend to be more similar the closer the sequences are phylogenetically. In the case of recombinants there is greater volatility. Sometimes, a greater phylogenetic closeness does not correlate with a greater similarity in decision patterns. Compared to concatenated spectrograms, superposed spectrograms reflect the relationship between phylogenetic closeness and the level of similarity in interpretability tool results more accurately.

C. ANALYSIS OF FAILED CLASSIFICATIONS

A failed sequence is considered a sequence of the test dataset misclassified by the CNN. That is, a non-recombinant sequence is classified as recombinant and vice versa.

In several cases, the CNN fails the same sequence in both datasets. Specifically, non-recombinant FJ389363 and recombinant GU230127, HQ385479, JN864058, JX390977, MN172223 and MT559132 are double misclassified. These phenomena suggest that these sequences are biologically interesting, and have distinct features compared with the

TABLE 10. Failed Classification Analysis Example 1. FJ389363 (subtype G) is misclassified as recombinant. Comparing its hot areas with those of AJ508596 (recombinant sequence of subtype 30-0206) shows a high level of concordance between them, even in Gradient Attribution. Incorrect patterns are similar to those of the recombinant determinant areas. In concatenated spectrograms, Grad-CAM focuses on A spectrogram, LIME at low frequencies of A and Gradient Attribution at high frequencies of A and low frequencies of T and C. In superposed spectrograms, the Grad-CAM hot zones are low frequencies of 5'LTR. LIME is more dispersed throughout the low-frequency range, locating the main zones between 5'LTR and *gag* and at some points near *vif*. Gradient Attribution hot areas seem to focus on low and high frequencies.

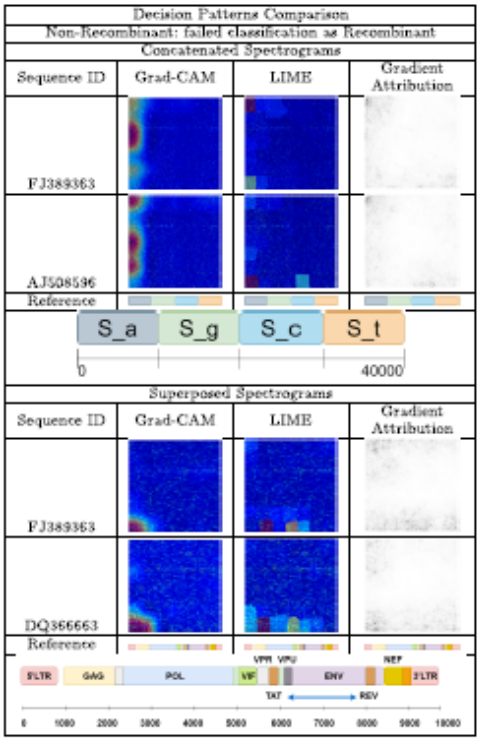
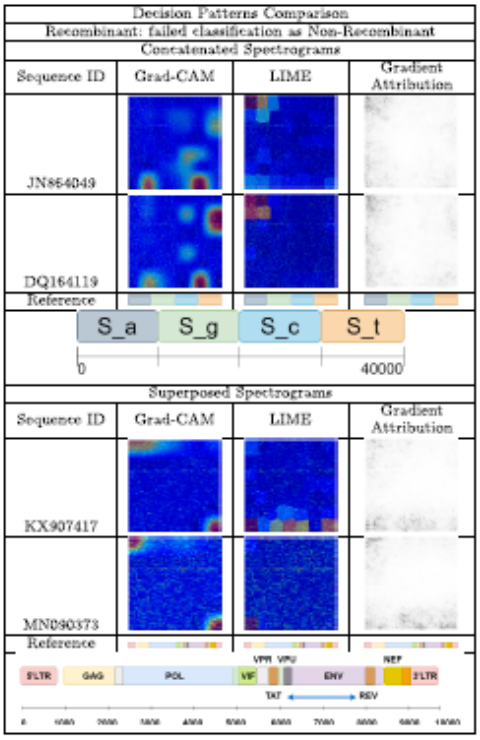


TABLE 11. Failed Classification Analysis Example 2. JN864049 (subtype 22-01A1) is misclassified as non-recombinant. Comparing its hot areas with those of DQ164119 (non-recombinant sequence of subtype C) shows again a high level of concordance between them, even in Gradient Attribution. In concatenated spectrograms, Grad-CAM highlights low frequencies between A and G and some zones near one-sixth, one-third and low frequencies in C and T. LIME focuses on high frequencies of A and Gradient Attribution is most active in the whole frequency range of A (specially high frequencies), followed by T (mostly one-third and low frequencies). In superposed spectrograms, Grad-CAM hot zones are high frequencies of 5'LTR and low frequencies of 3'LTR. LIME results are more scattered, but the determinant zone of LIME is low frequency range of 5'LTR. Additional areas are detected in low frequencies of the central part of the genome (end of *pol* and *vif*) and 3'LTR. Gradient Attribution determinant points are along the low frequency range and high frequencies of 5'LTR.



rest of the sequences. Therefore, a deep analysis is required regarding genome integrity and their correct classification in their subtypes. It is possible that these double misclassifications are not random, and that these sequences may have issues with their classification into their respective subtypes.

Analyzing non-recombinant sequences incorrectly classified as recombinant, the CNN seems to make a decision based on recombinant hot zones. Two representative examples are presented in Table 10.

The CNN incorrectly classifies the concatenated spectrogram of the non-recombinant sequence FJ389363 (subtype G) as a recombinant. Analyzing the Grad-CAM hot zones, we observe a high resemblance to AJ508596 hot zones (recombinant sequence of subtype 30-0206). To a lesser extent, the LIME hot points are also similar and Gradient Attribution is the most different, although they are still alike.

Similarly, the CNN also misclassifies the superposed spectrogram of the sequence FJ389363. In this case, the Grad-CAM and LIME hot zones in the two study sequences are very similar. In contrast, those of Gradient Attribution show appreciable differences to the naked eye. In any case, there is a direct link between the misclassification and hot zones of the incorrect category.

The same phenomenon occurs in the reverse direction. When the CNN fails to classify a recombinant sequence as non-recombinant, the decision zones look like those

associated with non-recombinants. Table 11 presents two graphical examples.

In Table 11, for concatenated spectrograms, recombinant JN864049 (subtype 22-01A1) is incorrectly classified as non-recombinant. Comparing the hot zones of its spectrograms with those of the non-recombinant sequence DQ164119 (subtype C), we observe that the decision patterns of the CNN when making the wrong decision are similar to those corresponding to DQ164119, both in concatenated and superposed. Even Gradient Attribution shows a higher degree of resemblance, which is greater than that seen in the previous example.

Similarly, the hot zones of the recombinant sequence KX907417 (subtype 41-CD), incorrectly classified as non-recombinant, are compared for the superposed spectrograms. High similarity is observed in the Grad-CAM hot zones corresponding to the non-recombinant MN090373 sequence (subtype B). In the case of LIME, the KX907417 hot zones are less specific than those in the case of MN090373. Similarities in the Gradient Attribution hot zones can be observed with the naked eye.

VI. CONCLUSION

Our methodology allows a rapid, robust and coherent identification of genomic areas that are determinant in the characterization of genetic recombination in HIV-1. This localization can significantly reduce genomic research timelines, and help microbiological experts to focus on the genomic areas. Our methodology allows for a complementary representation of genomic sequences, in addition to the traditional nucleotide sequence format. The conversion of genomic sequences to biologically consistent images such as spectrograms, enables the application of powerful deep learning and machine learning tools to the field of genomics, thus taking advantage of their full potential. The high hit rates on genomic spectrograms, despite their distinct nature from ImageNet images, demonstrate the transferability and applicability of the pattern detection capability from these tools to new domains. Therefore, we can efficiently identify areas of interest within the sequences through massive information comparison. This type of analysis is much more arduous to perform using the traditional system.

The CNN's total hit rate (test accuracy) for classifying viral sequences as recombinants or non-recombinants is 94% in both concatenated and superposed spectrograms. The decision areas in each category are consistent and differentiated. The hot zones are similar for sequences of the same subtype and phylogenetic proximity. When the CNN misclassifies, both Grad-CAM and LIME target the hot zones of the incorrect category.

The importance of the 5' and 3' LTR zones and their U3, R and U5 regions as hot zones for classifying a sequence as recombinant may be related to the role played by these terminal sections of the genome during viral transcription [82], [83], [84]. The neighboring *gag*, *nef* and *env* genes may also be involved in HIV-1 genetic recombination.

The importance of A and T nucleotides in recombinant feature correlates with the high content of polyA and polyT stretches in the HIV-1 sequence. These regions are susceptible to polymerase skipping and thus, genetic recombination. The key role of A nucleotide could be related to the high adenine content in lentivirus, especially in HIV-1 [85].

As Olabode et al. [86] suggested, the fact that sequences of different subtypes present very similar hot zones may be due to the existence of more cases of recombination than is currently described. It is possible that some viruses that are considered pure subtypes may exhibit hidden recombination phenomena.

In addition, the detected recombinant sequences are those which have prevailed because the resulting viruses have some selective advantages. Although numerous recombination phenomena can occur, they will not be selected if the resulting virus is non-viable or has lower fitness than the parental viruses. Through this research, we can gain insights into those genome regions that are more receptive to recombination and may confer a selective advantage for the virus [87].

Some sequences are misclassified in both concatenated and superposed spectrograms. The integrity and coherence of these published genomes must be verified. In addition, the criteria for classifying these sequences into their respective subtypes should be revised. This phenomenon may reinforce the correct operation of the CNN. If the subtypes of these sequences are incorrectly assessed or their genomes contain errors, it is biologically coherent that the CNN fails in both datasets.

In conclusion, 5'LTR and 3'LTR, and nucleotides A, followed by T, stand out as the leading hot zones for determining whether a sequence is recombinant or non-recombinant. These results seem biologically coherent with the genome structure and retrotranscription function. These facts validate that the CNN detects frequency-domain mathematical patterns that characterize a genomic sequence as recombinant or non-recombinant, including the location of these patterns within the sequence.

VII. FUTURE STUDIES

The application of methods based on multiple sequence alignment and manual adjustments did not allow for the discovery of patterns determining genetic recombination in viruses. Notwithstanding the information loss incurred through the conversion of sequences into spectrograms, it is possible to uncover previously unknown data embedded in the genome, thereby revealing patterns that are not visible in the nucleotide sequence itself.

The capacity of Deep Learning algorithms to find hidden patterns could greatly enhance research on this topic. To achieve this, it is necessary to delve into understanding the operation of these networks and how they detect these patterns. Detailed studies of existing pre-trained architectures and the development of interpretability tools are essential at this juncture. As more accurate neural networks are developed and applied to genomics, researchers will be able to

detect the specific regions of the genome that contribute to sequence characterization with enhanced precision and certainty. Consequently, the time required for sequence analysis and comparison will be significantly reduced. This holds particular significance for larger organisms with genomes spanning millions of nucleotides. Traditional analysis methods for such sequences are computationally intensive and extremely time-consuming (letter-by-letter manual adjustments). In such scenarios, it becomes crucial to possess a tool capable of pinpointing the genomic regions pertinent to the specific study feature. Our methodology can pave the way to address this need by offering precise identification, thereby greatly enhancing the efficiency of microbiological analyses.

Furthermore, it is necessary to feed these algorithms with curated datasets to help them identify these patterns. We needed to employ the entire dataset of available HIV-1 complete sequences for training and testing the neural network, which requires thousands of data samples to perform robustly. It is necessary to intensify the sequencing of diverse organisms to enhance the availability of the datasets.

This is the first time that a pre-trained CNN is applied to spectrogram datasets to detect and classify HIV-1 sequences as recombinant, with notable success rates and biologically supported results. Moreover, leveraging interpretability tools enabled us to identify specific genomic regions implicated in this feature. Our methodology contributes to the advancement of knowledge on genetic recombination in HIV-1, even though it may not have direct clinical application at present. Understanding the specific regions that play a key role in the selection of viable recombinant viruses can be useful for the search for new HIV-1 antiretroviral drugs. For example, the development of new antiretroviral treatments targeting specific genomic regions.

The UMI methodology is much faster and has lower computing costs than the traditional methods. This approach obviates the need for multiple sequence alignments and manual adjustments. These results indicate that there are probably mathematical patterns embedded in the genome that characterize a virus with different features. This encourages us to deepen their search and try to establish their mathematical formulations. Given its high accuracy, Gradient Attribution appears to be the most appropriate method for accurately locating these signatures.

We know where the CNN detects the recombinant feature. Our next step is to search for a signature or unique genetic identifier that allows to find similar variants or sequences in a much less computationally expensive manner.

Initiatives such as the one described, can successfully guide and reduce the heavy task of characterizing determining areas, which is probably applicable to more complex living organisms and their characteristics.

This research opens up new fields of application for biology.

Many exciting questions and fields emerge from this research.

VIII. DATA AVAILABILITY

All databases, codes and generated results are available from the corresponding author upon request. Supplementary Information is available at IEEE Dataport (Discovering Mathematical Patterns Behind HIV-1 Genetic Recombination: a new methodology to identify viral features - Supplementary Information).

ACKNOWLEDGMENT

The authors would like to thank a Professor Gerardo Mendizabal Ruiz with the Computer Science Department, University of Guadalajara, a Professor José Valentín Osuna Enciso with the Center for Exact Sciences and Engineering, University of Guadalajara, a Professor Ana Macarulla-Arenaza with the Faculty of Engineering, University of Deusto, a Research Associate Tony Castillo-Calzadilla with the DeustoTech Energy & Environment, a Research Assistant Maite Puerta-Beldarrain with the DeustoTech MoreLaboratory, a Research Assistant Armando Mendoza-Aguayo with DeustoTech Energy & Environment, and a Professor José Miguel Tamayo González with Loyola University Chicago for their support and assessment, also would like to thank the editors and reviewers for taking the time and effort necessary to review their manuscript, and also would like to thank all the valuable comments and suggestions, which helped them improve the quality of this article.

REFERENCES

[1] T. Hoenen et al., "Mutation rate and genotype variation of Ebola virus from Mali case sequences," *Science*, vol. 348, no. 6230, pp. 117–119, Apr. 2015.
[2] S. Duffy, "Why are RNA virus mutation rates so damn high?" *PLoS Biol.*, vol. 16, no. 8, Aug. 2018, Art. no. e3000003.
[3] J. Louten, "Virus replication," in *Essential Human Virology*, May 2016, ch. 4, pp. 49–70.
[4] M. M. C. Lai, "Genetic recombination in RNA viruses," in *Genetic Diversity of RNA Viruses*. Berlin, Germany: Springer, 1992, pp. 21–32.
[5] A. M. King, "Genetic recombination in positive strand RNA viruses," in *RNA Genetics*. Boca Raton, FL, USA: CRC Press, 2018, pp. 149–165.
[6] M. M. Lai, "RNA recombination in animal and plant viruses," *Microbiol. Rev.*, vol. 56, no. 1, pp. 61–79, Mar. 1992.
[7] C. Muslin, A. M. Kain, M. Bessaud, B. Blondel, and F. Delpeyroux, "Recombination in enteroviruses, a multi-step modular evolutionary process," *Viruses*, vol. 11, no. 9, p. 859, Sep. 2019.
[8] S. Su, G. Wong, W. Shi, J. Liu, A. C. K. Lai, J. Zhou, W. Liu, Y. Bi, and G. F. Gao, "Epidemiology, genetic recombination, and pathogenesis of coronaviruses," *Trends Microbiol.*, vol. 24, no. 6, pp. 490–502, Jun. 2016.
[9] A. D. Goldman and L. F. Landweber, "What is a genome?" *PLoS Genet.*, vol. 12, no. 7, 2016, Art. no. e1006181.
[10] R. G. Webster and W. Laver, "The origin of pandemic influenza," *Bull. World Health Org.*, vol. 47, no. 4, pp. 449–452, 1972.
[11] M. L. Kalish, K. E. Robbins, D. Pieniazek, A. Schaefer, N. Nzilambi, T. C. Quinn, M. E. S. Louis, A. S. Youngpairoj, J. Phillips, H. W. Jaffe, and T. M. Folks, "Recombinant viruses and early global HIV-1 epidemic," *Emerg. Infectious Diseases*, vol. 10, no. 7, pp. 1227–1234, Jul. 2004.
[12] P. Spreeuwenberg, M. Kroneman, and J. Paget, "Reassessing the global mortality burden of the 1918 influenza pandemic," *Amer. J. Epidemiol.*, vol. 187, no. 12, pp. 2561–2567, Dec. 2018.
[13] A. Mocroft, S. Vella, T. Benfield, A. Chiesi, V. Miller, P. Gargalianos, A. D. Monforte, I. Yust, J. Bruun, A. Phillips, and J. Lundgren, "Changing patterns of mortality across Europe in patients infected with HIV-1," *Lancet*, vol. 352, no. 9142, pp. 1725–1730, Nov. 1998.
[14] R. Nájera, E. Delgado, L. Pérez-Alvarez, and M. M. Thomson, "Genetic recombination and its role in the development of the HIV-1 pandemic," *AIDS*, vol. 16, pp. S3–S16, 2002.

[15] D. G. Higgins and P. M. Sharp, "CLUSTAL: A package for performing multiple sequence alignment on a microcomputer," *Gene*, vol. 73, no. 1, pp. 237–244, Dec. 1988.
[16] C. Notredame, D. G. Higgins, and J. Heringa, "T-Coffee: A novel method for fast and accurate multiple sequence alignment," *J. Mol. Biol.*, vol. 302, no. 1, pp. 205–217, Sep. 2000.
[17] C. B. Do, M. S. P. Mahabhashyam, M. Brudno, and S. Batzoglou, "ProbCons: Probabilistic consistency-based multiple sequence alignment," *Genome Res.*, vol. 15, no. 2, pp. 330–340, Feb. 2005.
[18] K. Katoh, G. Asimenos, and H. Toh, "Multiple alignment of DNA sequences with MAFFT," in *Bioinformatics for DNA Sequence Analysis*. Totowa, NJ, USA: Humana Press, 2009, pp. 39–64.
[19] R. C. Edgar, "MUSCLE: Multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Res.*, vol. 32, no. 5, pp. 1792–1797, Mar. 2004.
[20] Y. Liu, B. Schmidt, and D. L. Maskell, "MSAProbs: Multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities," *Bioinformatics*, vol. 26, no. 16, pp. 1958–1964, Aug. 2010.
[21] Y. Ye, D. W. Cheung, Y. Wang, S. M. Yiu, Q. Zhan, T. W. Lam, and H.-F. Ting, "GLProbs: Aligning multiple sequences adaptively," in *Proc. Int. Conf. Bioinf., Comput. Biol. Biomed. Informat.*, 2013, pp. 152–160.
[22] I. M. Wallace, O. Orla, and D. G. Higgins, "Evaluation of iterative alignment algorithms for multiple alignment," *Bioinformatics*, vol. 21, no. 8, pp. 1408–1414, Apr. 2005.
[23] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.
[24] C.-C.-J. Kuo, "Understanding convolutional neural networks with a mathematical model," *J. Vis. Commun. Image Represent.*, vol. 41, pp. 406–413, Nov. 2016.
[25] H. Wang, A. Cruz-Roa, A. Basavanhally, H. Gilmore, N. Shih, M. Feldman, J. Tomaszewski, F. Gonzalez, and A. Madabhushi, "Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features," *J. Med. Imag.*, vol. 1, no. 3, Oct. 2014, Art. no. 034003.
[26] H. Lu, Y. Zhu, M. Yin, G. Yin, and L. Xie, "Multimodal fusion convolutional neural network with cross-attention mechanism for internal defect detection of magnetic tile," *IEEE Access*, vol. 10, pp. 60876–60886, 2022.
[27] M. Rizwan, A. Shabbir, A. R. Javed, M. Shabbir, T. Baker, and D. A.-J. Obe, "Brain tumor and glioma grade classification using Gaussian convolutional neural network," *IEEE Access*, vol. 10, pp. 29731–29740, 2022.
[28] M. D. Kupperman, T. Leitner, and R. Ke, "A deep learning approach to real-time HIV outbreak detection using genetic data," *PLoS Comput. Biol.*, vol. 18, no. 10, Oct. 2022, Art. no. e1010598.
[29] S. Tammina, "Transfer learning using VGG-16 with deep convolutional neural network for classifying images," *Int. J. Sci. Res. Publ. (IJSRP)*, vol. 9, no. 10, pp. 143–150, 2019.
[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[31] S. Soffer, A. Ben-Cohen, O. Shimon, M. M. Amitai, H. Greenspan, and E. Klang, "Convolutional neural networks for radiologic images: A radiologist's guide," *Radiology*, vol. 290, no. 3, pp. 590–606, Mar. 2019.
[32] A. W.-C. Liew, H. Yan, and M. Yang, "Pattern recognition techniques for the emerging field of bioinformatics: A review," *Pattern Recognit.*, vol. 38, no. 11, pp. 2055–2073, Nov. 2005.
[33] A. E. Ouaslati, N. Ellouze, and Z. Lachiri, "3D spectrum analysis of DNA sequence: Application to Caenorhabditis elegans genome," in *Proc. IEEE 7th Int. Symp. Bioinf. Bioeng.*, Oct. 2007, pp. 864–871.
[34] N. Dimitrova, Y. H. Cheung, and M. Zhang, "Analysis and visualization of DNA spectrograms: Open possibilities for the genome research," in *Proc. 14th ACM Int. Conf. Multimedia*, Oct. 2006, pp. 1017–1024.
[35] A. Bucur, J. van Leeuwen, N. Dimitrova, and C. Mittal, "Alignment method for spectrograms of DNA sequences," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 1, pp. 3–9, Jan. 2010.
[36] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, "Prediction of probable genes by Fourier analysis of genomic sequences," *Bioinformatics*, vol. 13, no. 3, pp. 263–270, 1997.
[37] P. P. Vaidyanathan and B.-J. Yoon, "The role of signal-processing concepts in genomics and proteomics," *J. Franklin Inst.*, vol. 341, nos. 1–2, pp. 111–135, Jan. 2004.
[38] D. Sussillo, A. Kundaje, and D. Anastassiou, "Spectrogram analysis of genomes," *EURASIP J. Adv. Signal Process.*, vol. 2004, no. 1, Dec. 2004, Art. no. 700248.
[39] V. Kubicova and I. Provaznik, "Use of whole genome DNA spectrograms in bacterial classification," *Comput. Biol. Med.*, vol. 69, pp. 298–307, Feb. 2016.
[40] J. A. Morales, R. Saldaña, M. H. Santana-Castolo, C. E. Torres-Cerna, E. Borrayo, A. P. Mendizabal-Ruiz, H. A. Vélez-Pérez, and G. Mendizabal-Ruiz, "Deep learning for the classification of genomic signals," *Math. Problems Eng.*, vol. 2020, May 2020, Art. no. 7698500.
[41] H. J. Nussbaumer, "The fast Fourier transform," in *Fast Fourier Transform and Convolution Algorithms*. Cham, Switzerland: Springer, 1981, pp. 80–111.
[42] J. G. Proakis and D. G. Manolakis, *Introduction to Digital Signal Processing*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1988.
[43] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, "Deep learning for health informatics," *IEEE J. Biomed. Health Informat.*, vol. 21, no. 1, pp. 4–21, Jan. 2017.
[44] L. Yuan and J. Cao, "Patients' EEG data analysis via spectrogram image with a convolution neural network," in *Proc. Int. Conf. Intell. Decis. Technol.* Cham, Switzerland: Springer, 2017, pp. 13–21.
[45] K. N. Khan, F. A. Khan, A. Abid, T. Olmez, Z. Dokur, A. Khandakar, M. E. H. Chowdhury, and M. S. Khan, "Deep learning based classification of unsegmented phonocardiogram spectrograms leveraging transfer learning," *Physiol. Meas.*, vol. 42, no. 9, Sep. 2021, Art. no. 095003.
[46] G. Ruffini, D. Ibañez, M. Castellano, L. Dubreuil-Vall, A. Soria-Frisch, R. Postuma, J.-F. Gagnon, and J. Montplaisir, "Deep learning with EEG spectrograms in rapid eye movement behavior disorder," *Frontiers Neurol.*, vol. 10, p. 806, Jul. 2019.
[47] W. Zhang, J. Han, and S. Deng, "Heart sound classification based on scaled spectrogram and tensor decomposition," *Expert Syst. Appl.*, vol. 84, pp. 220–231, Oct. 2017.
[48] M. Qiao, Z. Fang, Y. Guo, S. Zhou, C. Chang, and Y. Wang, "Breast calcification detection based on multichannel radiofrequency signals via a unified deep learning framework," *Expert Syst. Appl.*, vol. 168, Apr. 2021, Art. no. 114218.
[49] M. Eltemadi, S. B. Abkenar, A. Ahmadzadeh, M. H. Kashani, P. Asghari, M. Akbari, and E. Mahdipour, "A systematic review of healthcare recommender systems: Open issues, challenges, and techniques," *Expert Syst. Appl.*, vol. 213, Mar. 2023, Art. no. 118823.
[50] E. Santo and N. Dimitrova, "Improvement of spectral analysis as a genomic analysis tool," in *Proc. IEEE Int. Workshop Genomic Signal Process. Statist.*, Jun. 2007, pp. 1–4.
[51] M. Nahiduzzaman, M. R. Islam, and R. Hassan, "ChestX-ray6: Prediction of multiple diseases including COVID-19 from chest X-ray images using convolutional neural network," *Expert Syst. Appl.*, vol. 211, Jan. 2023, Art. no. 118576.
[52] P. M. Cheng and H. S. Malhi, "Transfer learning with convolutional neural networks for classification of abdominal ultrasound images," *J. Digit. Imag.*, vol. 30, no. 2, pp. 234–243, Apr. 2017.
[53] I. Kandel and M. Castelli, "Transfer learning with convolutional neural networks for diabetic retinopathy image classification. A review," *Appl. Sci.*, vol. 10, no. 6, p. 2021, Mar. 2020.
[54] A. Kensert, P. J. Harrison, and O. Spjuth, "Transfer learning with deep convolutional neural networks for classifying cellular morphological changes," *SLAS Discovery, Advancing Life Sci. R&D*, vol. 24, no. 4, pp. 466–475, 2019.
[55] H. W. Loh, C. P. Ooi, E. Aydemir, T. Tuncer, S. Dogan, and U. R. Acharya, "Decision support system for major depression detection using spectrogram and convolution neural network with EEG signals," *Expert Syst.*, vol. 39, no. 3, Mar. 2022, Art. no. e12773.
[56] D. Posada and K. A. Crandall, "Evaluation of methods for detecting recombination from DNA sequences: Computer simulations," *Proc. Nat. Acad. Sci. USA*, vol. 98, no. 24, pp. 13757–13762, Nov. 2001.
[57] Q.-S. Zhang and S.-C. Zhu, "Visual interpretability for deep learning: A survey," *Frontiers Inf. Technol. Electron. Eng.*, vol. 19, no. 1, pp. 27–39, Jan. 2018.
[58] S. Chakraborty, R. Tomsett, R. Raghavendra, D. Harborne, M. Alzantot, F. Cerutti, M. Srivastava, A. Preece, S. Julier, R. M. Rao, T. D. Kelley, D. Braines, M. Sensoy, C. J. Willis, and P. Gurram, "Interpretability of deep learning models: A survey of results," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud & Big Data Comput., Internet People Smart City Innov. (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, Aug. 2017, pp. 1–6.
[59] J.-X. Mi, A.-D. Li, and L.-F. Zhou, "Review study of interpretation methods for future interpretable machine learning," *IEEE Access*, vol. 8, pp. 191969–191985, 2020.

[60] V. Jahmunah, E. Y. K. Ng, R.-S. Tan, S. L. Oh, and U. R. Acharya, "Explainable detection of myocardial infarction using deep learning models with grad-CAM technique on ECG signals," *Comput. Biol. Med.*, vol. 146, Jul. 2022, Art. no. 105550.
[61] Y. Zhang, K. Song, Y. Sun, S. Tan, and M. Udell, "'Why should you trust my explanation?' Understanding uncertainty in LIME explanations," 2019, *arXiv:1904.12991*.
[62] K. Mridha, M. M. Uddin, J. Shin, S. Khadka, and M. F. Mridha, "An interpretable skin cancer classification using optimized convolutional neural network for a smart healthcare system," *IEEE Access*, vol. 11, pp. 41003–41018, 2023.
[63] I. E. Nielsen, D. Dera, G. Rasool, N. Bouaynaya, and R. P. Ramachandran, "Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks," 2021, *arXiv:2107.11400*.
[64] Z. Salahuddin, H. C. Woodruff, A. Chatterjee, and P. Lambin, "Transparency of deep neural networks for medical image analysis: A review of interpretability methods," *Comput. Biol. Med.*, vol. 140, Jan. 2022, Art. no. 105111.
[65] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[66] O. D. T. Catalá, I. S. Igual, F. J. Pérez-Benito, D. M. Escrivá, V. O. Castelló, R. Llobet, and J.-C. Peréz-Cortés, "Bias analysis on public X-ray image datasets of pneumonia and COVID-19 patients," *IEEE Access*, vol. 9, pp. 42370–42383, 2021.
[67] L. Chen, A. Cruz, S. Ramsey, C. J. Dickson, J. S. Duca, V. Hornak, D. R. Koes, and T. Kurtzman, "Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening," *PLoS ONE*, vol. 14, no. 8, Aug. 2019, Art. no. e0220113.
[68] V. Gupta and R. Patel, "Lungs disease classification using VGG-16 architecture with PCA," in *Proc. Int. Conf. Advancement Comput. Comput. Technol. (InCACCT)*, May 2023, pp. 495–500.
[69] Z. Zhou, Y. Liu, Q. Wang, and T. T. Toe, "Detection of pneumonia based on ResNet improved by attention mechanism," in *Proc. IEEE 3rd Int. Conf. Power, Electron. Comput. Appl. (ICPECA)*, Jan. 2023, pp. 859–863.
[70] R. Singh, N. Sharma, and R. Gupta, "Detection of Alzheimer's risk level using inception V3 transfer learning model," in *Proc. Int. Conf. Distrib. Comput. Electr. Circuits Electron. (ICDCECE)*, Apr. 2023, pp. 1–6.
[71] K. D. Hyde, D. Udayanga, D. S. Manamgoda, L. Tedersoo, E. Larsson, K. Abarenkov, Y. J. K. Bertrand, B. Oxelman, M. Hartmann, H. Kauserud, M. Ryberg, E. Kristiansson, and H. R. Nilsson, "Incorporating molecular data in fungal systematics: A guide for aspiring researchers," 2013, *arXiv:1302.3244*.
[72] E. Castro-Nallar, M. Pérez-Losada, G. F. Burton, and K. A. Crandall, "The evolution of HIV: Inferences using phylogenetics," *Mol. Phylogenetics Evol.*, vol. 62, no. 2, pp. 777–792, Feb. 2012.
[73] C. Kuiken, B. Korber, and R. W. Shafer, "HIV sequence databases," *AIDS Rev.*, vol. 5, no. 1, pp. 52–61, 2003.
[74] E. Simon-Loriere and E. C. Holmes, "Why do RNA viruses recombine?" *Nature Rev. Microbiol.*, vol. 9, no. 8, pp. 617–626, Aug. 2011.
[75] J. M. Coffin, S. H. Hughes, and H. E. Varmus, *Retroviruses*. Cold Spring Harbor, NY, USA: Cold Spring Harbor Laboratory Press, 1997.
[76] A. Santana, C. Domínguez, A. Lemes, T. Molero, and E. Salido, "Biología celular y molecular del virus de inmunodeficiencia humana (VIH)," *Revista Diagnóstico Biológico*, vol. 52, no. 1, pp. 7–18, 2003.
[77] R. Gómez-Román and C. Soler-Claudín, "La importancia de la secuencia terminal repetida larga (LTR) en la patogenia del virus de la inmunodeficiencia humana," *Revista Biomédica*, vol. 11, no. 1, pp. 61–71, 2000.
[78] W. S. Hu and H. M. Temin, "Genetic consequences of packaging two RNA genomes in one retroviral particle: Pseudodiploidy and high rate of genetic recombination," *Proc. Nat. Acad. Sci. USA*, vol. 87, no. 4, pp. 1556–1560, Feb. 1990.
[79] G. Turk, M. Carobene, A. Monczor, A. E. Rubio, M. Gómez-Carrillo, and H. Salomón, "Higher transactivation activity associated with LTR and Tat elements from HIV-1 BF intersubtype recombinant variants," *Retrovirology*, vol. 3, no. 1, pp. 1–12, Dec. 2006.
[80] J. Hemelaar, R. Elangovan, J. Yun, L. Dickson-Tetteh, I. Fleminger, S. Kirtley, B. Williams, E. Gouws-Williams, and P. D. Ghys, "Global and regional molecular epidemiology of HIV-1, 1990–2015: A systematic review, global survey, and trend analysis," *Lancet Infectious Diseases*, vol. 19, no. 2, pp. 143–155, 2019.
[81] R. A. Neher and T. Leitner, "Recombination rate and selection strength in HIV intra-patient evolution," *PLoS Comput. Biol.*, vol. 6, no. 1, Jan. 2010, Art. no. e1000660.
[82] W.-S. Hu and H. M. Temin, "Retroviral recombination and reverse transcription," *Science*, vol. 250, no. 4985, pp. 1227–1233, Nov. 1990.
[83] B. C. Ramirez, E. Simon-Loriere, R. Galetto, and M. Negroni, "Implications of recombination for HIV diversity," *Virus Res.*, vol. 134, nos. 1–2, pp. 64–73, Jun. 2008.
[84] E. R. de Arellano, J. Alcamí, M. López, V. Soriano, and Á. Holguín, "Drastic decrease of transcription activity due to hypermutated long terminal repeat (LTR) region in different HIV-1 subtypes and recombinants," *Antiviral Res.*, vol. 88, no. 2, pp. 152–159, Nov. 2010.
[85] J. P. Vartanian, A. Meyerhans, B. Asjö, and S. Wain-Hobson, "Selection, recombination, and G—A hypermutation of human immunodeficiency virus type 1 genomes," *J. Virology*, vol. 65, no. 4, pp. 1779–1788, Apr. 1991.
[86] A. S. Olabode, G. T. Ng, K. E. Wade, M. Salnikov, H. E. Grant, D. W. Dick, and A. F. Y. Poon, "Revisiting the recombinant history of HIV-1 group M with dynamic network community detection," *Proc. Nat. Acad. Sci. USA*, vol. 119, no. 19, May 2022, Art. no. e2108815119.
[87] L. R. Banner, J. G. Keck, and M. M. C. Lai, "A clustering of rna recombination sites adjacent to a hypervariable region of the peplomer gene of murine coronavirus," *Virology*, vol. 175, no. 2, pp. 548–555, Apr. 1990.



ANA GUERRERO-TAMAYO received the Bachelor of Engineering degree in electronics and industrial automation from the Faculty of Engineering, University of Deusto, in 2001, the master's degree in occupational risk prevention with a focus on occupational safety, industrial hygiene and ergonomics, and applied psychosociology, and the M.B.A. Executive degree from the Faculty of Economics and Business Administration, University of the Basque Country (UPV/EHU), in 2006. She is currently pursuing the Ph.D. degree in the application of deep learning algorithms in genomics with the Faculty of Engineering, University of Deusto.

She received a Special Mention for the Best Master's Thesis and Best Record of the 2005–2006 Promotion. She also completed a course in international business management from the School of Industrial Organization of Madrid (EOI), where she received the Special Mention to the Best Record of the Promotion. She held positions of responsibility in private companies related to the field of electrical MV/HV networks and in strategic consulting for multinational corporations. She is also a Research Assistant with DeustoTech, University of Deusto. Her research interests include applied mathematics in genomics, the analysis of gene expression and regulation, the prediction of protein structure and function, the analysis of massive sequencing data, and the development of new tools for genomic data analysis and visualization.



BORJA SANZ URQUIJO received the Ph.D. degree (cum laude) in information systems, in 2012, with a focus on malware detection in android mobile devices.

He is currently a Professor with the Faculty of Engineering, University of Deusto. He has been a Researcher with the DeustoTech, Computing Research Unit, since 2008, becoming a Head Researcher, from 2015 to 2018. He contributed to over 100 projects, including H2020 and national and private initiatives. Throughout these projects, he served as a researcher and a project manager in several of them. Some of his standout projects are Detecting and Analyzing Terrorist-Related Online Content and Financing Activities (DANTE), Security and Trust in the Information Society (CENIT SEGURA), and Models for Malware Propagation through Online Social Networks (MARSOL). He has published several book chapters and more than 60 articles in specialized national and international impact journals, such as *Logic Journal of IGPL*, *Electronic Commerce Research and Applications*, *Expert Systems with Applications*, and IEEE Access. His primary research interests include machine learning, big data, knowledge discovery, natural language processing, information retrieval, and the application of these techniques in the field of social sciences. He is also working on fairness, ethics, accountability, and the impact of artificial intelligence in society.

Dr. Urquijo won the Award for the Best Oral Presentation at the International Congress on Human Rights, Emerging Challenges 2018, with the paper: "Hic Sunt draconis: Human Rights and Big Data, analysis of an unexplored collaboration." He is also a member of the Observatory on the

Social and Ethical Impact of Artificial Intelligence (ODISEIA). For more information visit the link: (http://paginaspersonales.deusto.es/bosanz).



CONCEPCIÓN CASADO received the Graduate degree in biological sciences, in 1985.

She is currently researching HIV-1/AIDS pathogenesis. She has been with Instituto de Salud Carlos III (ISCIII), since 1987. Her early work was focused on diagnosing retroviral infections and evaluating commercial reagents for detection. In 1995, she joined the Molecular Virology Group, where she completed her Ph.D. thesis in biologic sciences on the molecular characterization of the Spanish HIV-1 epidemic and the intra-patient evolution of HIV-1. One of her most important results was the development of a method for dating viral nucleotide sequences based on the application of a relaxed molecular clock to viral evolution, which has since been widely used in the laboratory. This methodology applied to the viral evolution of HIV-1 in patients without clinical progression (LTNPs) led to the identification of a group of patients, the LTNP-elite controllers (LTNP-EC), which are characterized by the presence of ancestral viruses and lack of evolution. Since then, she has focused on studying the viral and host factors associated with LTNPs, including the accumulation of defective viral forms and non-progressive host factors. She was a part of a research team that demonstrated the relevance of viral environment functionality in clinical progression. Recently, this team described, for the first time, the functional cure in 3 LTNP-EC patients infected for over 30 years without progression. She contributed to the creation of LTNP and elite controllers (EC-RIS) cohorts of the RETICS AIDS NETWORK. She has published collaborative works characterizing LTNP-EC patients, who represent a model of functional cure for HIV-1 infection. Since 2016, she has been combining her research work with participation in the diagnosis and reference of retrovirus infections with the National Center of Microbiology (ISCIII). Since 2020, she has been responsible for the Molecular Virology Group. She has participated in numerous research projects (14). She has published numerous articles in international journals (56). She has participated in numerous national and international conferences (40).



MARÍA-DOLORES MORAGUES TOSANTOS received the Graduate degree in biological sciences from the University of the Basque Country (UPV/EHU), in 1979, and the Ph.D. degree, in 1989. She is currently a Full Professor and a Honorary Collaborator with the Faculty of Medicine and Nursing, UPV/EHU. She has been teaching activity in the degree in nursing (human physiology, microbiology, research methodology, and public health), since 1983. She obtained a predoctoral grant from the Basque Government (1980–1982), which included a six-month stay with the Department of Microbiology, California College of Medicine, University of California at Irvine, Irvine (Mechanism of inducible phenotypic drug resistance in Mucor racemosus). From 1980 to 1990, her work focused on the dimorphism of the fungus Aureobasidium pullulans. Since 1991, she has been studying the diagnosis and pathogenesis of invasive candidiasis, including the production of polyclonal and monoclonal antibodies and the characterization of virulence factors. In 2010, she became responsible for the Invasive Fungal Infection Study Group (GEIFI). Her current research interests include the diagnosis of fungal infections and resistance to antifungal compounds. She has collaborated in more than 30 research projects subsidized in public calls, with at least ten of them as a PI. In total, she has published nine books/chapters and 78 research articles in indexed journals, and she has signed more than 150 communications/presentations in national and international scientific meetings. She has been a member of several scientific societies, most notably the Spanish Society of Microbiology (SEM), since 1978, Spanish Association of Mycology (AEM), since 1997, International Society for Human and Animal Mycology (ISHAM), since 2008, American Society of Microbiology (ASM), since 2010, and Spanish Society of Infectious Diseases and Clinical Microbiology (SEIMC), since 2011. She was an Assistant Editor (2002–2014) and a member of the editorial board of the *Revista Iberoamericana de Micología* (ISSN: 1130-1406), since 2015, and an evaluator of manuscripts for several scientific journals. Between 2008 and 2017, she was a member of the Ethics Committee for Research Related to Human Beings (CEISH) of the Commission on Ethics in Research and Teaching (CEID), and the Director of the Department of Nursing, School of Medicine and Nursing, UPV/EHU, from 2019 to 2020.



ISABEL OLIVARES received the bachelor's degree in general medicine and surgery, and the Ph.D. degree from Universidad Autónoma de Madrid (UAM).

She has been a Research Scientist with Instituto de Salud Carlos III (ISCIII), since 1990. She is actively involved in virological research. First, she studied herpes simplex virus type 2 during her Ph.D. degree. As a Postdoctoral Fellow, she contributed to studies on the African swine fever virus. Since 1990, her research has been devoted to study several facets of the Human Immunodeficiency Virus Type I (HIV-1). Initially, she was mainly involved in molecular epidemiology, variability, and genetic characterization studies of HIV-1 isolates. She has participated in the "WHO Network for HIV Isolation and Characterization" as it was the first WHO Laboratory Network organized for HIV-1 studies, with great relevance. Since 1996, she has been focusing on HIV-1 infection pathogenesis by "in vitro" experiments. She has generated and characterized two infectious molecular clones from a patient HIV-1 isolate. This is a remarkably useful tool for multiple "in vitro" studies. She has collaborated with Centro de Biología Molecular Severo Ochoa on three coordinated research projects to study the implications of the viral reverse transcriptase domain for reverse transcriptase copy fidelity, viral replicative capacity, and viral resistance to antiretrovirals. Since 2000, she has also been focused on studying HIV-1 persistence using various persistently infected cell lines obtained in the laboratory as a model to identify viral or cellular factors associated with viral persistence. She also explored lethal mutagenesis, a new antiviral strategy to extinguish HIV-1 through elevated mutation rate in a persistently infected cell line, in two coordinated research projects. Since 2003, she has been a member of the RETICS AIDS NETWORK, focusing on the viral and host features of LTNP patients. Recently, she has been interested in the search for new treatments against HIV-1 targeting host cell proteins, studying the antiviral effect of an Ataxia telangiectasia, and Rad3-related protein (ATR) inhibitor on HIV-1 replication. She has participated in numerous research projects and she is the author of 42 peer-reviewed articles.



IKER PASTOR-LÓPEZ received the bachelor's degree in computer engineering, in 2007, the master's degree in information security, in 2010, and the Ph.D. degree (Hons.) in computer science from the Faculty of Engineering, University of Deusto, in 2013.

He is currently an Assistant Professor and an Erasmus Coordinator with the Faculty of Engineering, University of Deusto, teaching courses, such as deep learning, computer vision, and information security, where he is also a Researcher with DeustoTech. He is a member of the Department of Information Technology, Electronics, and Communication. He is also the Director of the Master's Program in Computing and Intelligent Systems, Faculty of Engineering, University of Deusto, where he has been a member of the Big Data and Business Intelligence Program, since 2016. His research interests include big data analytics, deep learning, natural language processing, opinion mining, and computer vision. His published work includes more than 45 scientific articles and international conferences. He has participated in the conceptualization and scientific and technical development of numerous competitive projects and contracts with companies, the latter with several successful cases of knowledge transfer actions.

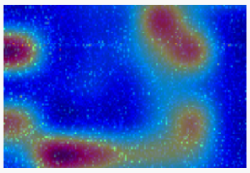
Dr. Pastor-López is a member of the Scientific Committee of several congresses, such as CISIS, SOCO, and ICEUTE. He is a Reviewer for many journals, including those indexed in JCR, such as *Neurocomputing*, *Logic Journal of IGPL*, and the *Engineering and Industry* magazine (DYNA), among others.

IEEEDataPort DATASETS SUBMIT A DATASET COMPETITIONS IEEE

Datasets

Standard Dataset

DISCOVERING MATHEMATICAL PATTERNS BEHIND HIV-1 GENETIC RECOMBINATION: A NEW METHODOLOGY TO IDENTIFY VIRAL FEATURES - SUPPLEMENTARY INFORMATION



Citation Author(s):
Ana Guerrero-Tamayo *(Faculty of Engineering - University of Deusto)*
Borja Sanz Urquijo *(Faculty of Engineering - University of Deusto)*
Concepción Casado *(National Microbiology Center (NMC) - Instituto de Salud Carlos III (ISCIII))*
María-Dolores Moragues Tosantos *(Faculty of Medicine and Nursing University of the Basque Country UPV/EHU)*
Isabel Olivares *(National Microbiology Center (NMC) - Instituto de Salud Carlos III (ISCIII))*
Iker Pastor-López *(Faculty of Engineering - University of Deusto)*

277 Views

Categories: Artificial Intelligence, Machine Learning, Image Processing, Biomedical and Health Sciences

Keywords: convolutional neural network, Deep Learning, Genetic Recombination, Genome Mathematical Pattern, Genome Mathematical Signature, HIV-1

Submitted by: Ana Guerrero-Tamayo

Last updated: Tue, 09/12/2023 - 06:02

DOI: 10.21227/2917-1s55

Research Article Link: Discovering Mathematical Patterns Behind HIV-1 Genetic Recombination: a new methodology to identify viral features

License: Creative Commons Attribution

☆☆☆☆☆ *0 ratings - Please login to submit your rating.*

ACCESS DATASET CITE SHARE/EMBED

ABSTRACT

This dataset contains the Supplementary Information of the article "Discovering Mathematical Patterns Behind HIV-1 Genetic Recombination: a new methodology to identify viral features" (Manuscript DOI: 10.1109/ACCESS.2023.3311752). It contains the two datasets used (concatenated and superposed). The results of all experiments and training curves are available at the Supplementary Information - Training Curves folder. The hot maps of the complete dataset are available at the Supplementary Information - Hot Zones folder. The complete HIV-1 phylogenetic tree is available at the Supplementary Information - Phylogenetic Tree folder. To facilitate the biological meaning of the hot areas, we provide a set of figures with their scaled reference. See Supplementary Information -Hot Zones Reference Information.

Funding Agency: Universidad de Deusto

DATASET FILES

- HOT ZONES REFERENCE INFORMATION.zip (42.46 MB)
- PHYLOGENETIC TREE.zip (101.31 kB)
- TRAINING CURVES.zip (5.54 MB)
- DATASET.zip (795.59 MB)
- HOT ZONES.zip (5.94 GB)

LOGIN TO ACCESS DATASET FILES



CERTIFICADO

I WORKSHOP INTERNACIONAL SOBRE INTELIGÊNCIA ARTIFICIAL

Certificamos que ANA GUERRERO TAMAYO, portador(a) do DNI 78884726Q, organizou o I Workshop Internacional sobre Inteligência Artificial, no dia 10 de abril de 2024, com carga horária de 10 horas.

Paranavaí, 21 de junho de 2024.

Profº. Dra. Marila Annibelli Vellozo
Diretora Pró-tempore do Escritório de Relações Internacionais da UNESPAR
Portaria N.º 344/2024

Profº. Dra. Aurea Andrade Viana de Andrade
Coordenadora do Workshop Internacional Sobre Inteligência Artificial
Universidade Estadual do Paraná - UNESPAR

Prof. João Paulo Martins Gouveia
Coordenador do Workshop Internacional Sobre Inteligência Artificial
Instituto Superior de Engenharia
Universidade Politécnica de Coimbra

Registro:
Livro: 1
Folha: 1
Data: 10/04/2024

Escola Superior de Educação
Politécnico de Coimbra

Instituto Superior de Engenharia
Politécnico de Coimbra

UNESPAR

SUScita

PPGSED

PALI

GRUPO GERA

ERI

Politécnico de Coimbra

Instituto de Investigação Aplicada
Politécnico de Coimbra

PROGRAMAÇÃO

I WORKSHOP INTERNACIONAL SOBRE INTELIGÊNCIA ARTIFICIAL
'Navegar, ou não, pelo Universo da Inteligência Artificial através de modelos Deep Learning'.
Data: 10 de abril de 2024

Palestra:
Modelos Deep Learning resilientes. Estratégias para uma gestão eficaz de viés emtreinamentos de modelos Deep Learning.

Palestrante convidada:
Profª. Ana Guerrero Tamayo, Facultad de Ingeniería de la Universidadde Deusto, Espanha.

Debatedores:
Prof. Fernando Sadio-Ramos (SUScita, ESEC, IPC), Portugal
Prof. Gustavo Matias (PALI, Unespar), Brasil
Profª. Teresa Fragoso (SUScita, ISEC, IPC), Portugal
Prof. Fernando Lermen (PALI, Unespar), Brasil

Moderador:
Prof. João Paulo Martins Gouveia (SUScita, ISEC, IPC), Portugal

Programação: horário de Brasília/horário de Portugal
Apresentação: 14h00 / 18h00
Palestra: 14h10 / 18h10
Debate: 15h30 / 19h30
Término: 17h00 / 21h00

Carga horária: 3h



CERTIFICADO

I WORKSHOP INTERNACIONAL SOBRE INTELIGÊNCIA ARTIFICIAL

Certificamos que ANA GUERRERO TAMAYO, portador(a) do DNI 78884726Q, proferiu palestra sobre o tema **"Modelos Deep Learning resilientes. Estratégias para uma gestão eficaz de viés em treinamentos de modelos Deep Learning"** no I Workshop Internacional sobre Inteligência Artificial, no dia 10 de abril de 2024, com carga horária de 3 horas.

Paranavaí, 21 de junho de 2024.

Profª. Dra. Marila Annibelli Vellozo
Diretora Pró-tempore do Escritório de Relações Internacionais da UNESPAR
Portaria N.º 344/2024

Profª. Dra. Aurea Andrade Viana de Andrade
Coordenadora do Workshop Internacional Sobre Inteligência Artificial
Universidade Estadual do Paraná - UNESPAR

Prof. João Paulo Martins Gouveia
Coordenador do Workshop Internacional Sobre Inteligência Artificial
Instituto Superior de Engenharia
Universidade Politécnica de Coimbra

Registro:
Livro: 1
Folha: 1
Data: 10/04/2024

PROGRAMAÇÃO

I WORKSHOP INTERNACIONAL SOBRE INTELIGÊNCIA ARTIFICIAL
'Navegar, ou não, pelo Universo da Inteligência Artificial através de modelos Deep Learning'.
Data: 10 de abril de 2024

Palestra:
Modelos Deep Learning resilientes. Estratégias para uma gestão eficaz de viés emtreinamentos de modelos Deep Learning.

Palestrante convidada:
Profª. Ana Guerrero Tamayo, Facultad de Ingeniería de la Universidadde Deusto, Espanha.

Debatedores:
Prof. Fernando Sadio-Ramos (SUScita, ESEC, IPC), Portugal
Prof. Gustavo Matias (PALI, Unespar), Brasil
Profª. Teresa Fragoso (SUScita, ISEC, IPC), Portugal
Prof. Fernando Lermen (PALI, Unespar), Brasil

Moderador:
Prof. João Paulo Martins Gouveia (SUScita, ISEC, IPC), Portugal

Programação: horário de Brasília/horário de Portugal
Apresentação: 14h00 / 18h00
Palestra: 14h10 / 18h10
Debate: 15h30 / 19h30
Término: 17h00 / 21h00

Carga horária: 3h



CERTIFICADO

I WORKSHOP INTERNACIONAL SOBRE INTELIGÊNCIA ARTIFICIAL

Certificamos que ANA GUERRERO TAMAYO, portador(a) do DNI 78884726Q, participou como ouvinte no I Workshop Internacional sobre Inteligência Artificial, no dia 10 de abril de 2024, com carga horária de 3 horas.

Paranavaí, 17 de junho de 2024.

Profº. Dra. Marila Annibelli Vellozo
Diretora Pró-tempore do Escritório de Relações Internacionais da UNESPAR
Portaria N.º 344/2024

Profº. Dra. Aurea Andrade Viana de Andrade
Coordenadora do Workshop Internacional Sobre Inteligência Artificial
Universidade Estadual do Paraná - UNESPAR

Prof. João Paulo Martins Gouveia
Coordenador do Workshop Internacional Sobre Inteligência Artificial
Instituto Superior de Engenharia
Universidade Politécnica de Coimbra

Registro:
Livro: 1
Folha: 1
Data: 10/04/2024

Escola Superior de Educação
Politécnico de Coimbra

Instituto Superior de Engenharia
Politécnico de Coimbra

UNESPAR

SUScita

PPGSED

PALI

GRUPO GERA

ERI

Politécnico de Coimbra

Instituto de Investigação Aplicada
Politécnico de Coimbra

P R O G R A M A Ç Ã O

I WORKSHOP INTERNACIONAL SOBRE INTELIGÊNCIA ARTIFICIAL
'Navegar, ou não, pelo Universo da Inteligência Artificial através de modelos Deep Learning'.
Data: 10 de abril de 2024

Palestra:
Modelos Deep Learning resilientes. Estratégias para uma gestão eficaz de viés emtreinamentos de modelos Deep Learning.

Palestrante convidada:
Profª. Ana Guerrero Tamayo, Facultad de Ingeniería de la Universidadde Deusto, Espanha.

Debatedores:
Prof. Fernando Sadio-Ramos (SUScita, ESEC, IPC), Portugal
Prof. Gustavo Matias (PALI, Unespar), Brasil
Profª. Teresa Fragoso (SUScita, ISEC, IPC), Portugal
Prof. Fernando Lermen (PALI, Unespar), Brasil

Moderador:
Prof. João Paulo Martins Gouveia (SUScita, ISEC, IPC), Portugal

Programação: horário de Brasília/horário de Portugal
Apresentação: 14h00 / 18h00
Palestra: 14h10 / 18h10
Debate: 15h30 / 19h30
Término: 17h00 / 21h00

Carga horária: 3h

¿Es Posible Saber Deep Learning Sin Ser Ingeniero?

Fundamentos y Reflexiones sobre Modelos Generativos



Ana Guerrero Tamayo
Fernando José Sadio-Ramos

Universidad de Deusto
Institulo Politécnico de Coimbra

Ficha Técnica

ISBN: 978-989-8486-52-3
Título: ¿Es Posible Saber Deep Learning Sin Ser Ingeniero? Fundamentos y Reflexiones sobre Modelos Generativos
Autores: Ana Guerrero Tamayo; Fernando José Sadio-Ramos
Edición: Fernando Ramos (Editor)® (Coimbra)
Layout y composición:: Ana Guerrero Tamayo
CDU: 001; 004.8; 17; 342.7; 37

Apoyo

O Instituto de Estudos Filosóficos é financiado por fundos nacionais através da FCT – Fundação para a Ciência e a Tecnologia, I.P., no âmbito do projecto UID/FIL/00010/2020



Índice General

Lista de Figuras xv

Abreviaturas xvii

1. Conceptos Básicos en Machine Learning 1
1.1. ¿Qué es Inteligencia Artificial Realmente? . . . 1
1.2. ¿Qué Necesitamos para Aplicar Algoritmos Machine Learning? . . . 4

2. Algunos Conceptos Técnicos 7
2.1. Nota de los Autores Relativa al Uso de Anglicismos . . . 7
2.2. Tipos de Algoritmo, a Grandes Rasgos . . . 8
2.3. Deep Learning. Principio de Diseño . . . 11
2.4. Ciclo de Entrenamiento . . . 14
2.5. Algunos Hiperparámetros y Otros Conceptos . . . 18
2.6. La Importancia de un Buen Entrenamiento . . . 24
2.7. Procedimiento para el Entrenamiento . . . 25
2.8. ¿Cómo Detectar un Buen Entrenamiento? . . . 28
2.9. Técnicas de Comprobación del Funcionamiento del Modelo . . . 35
2.9.1. Curva ROC y AUC (Area Under Curve) . . . 36
2.9.2. Confusion Matrix . . . 37
2.9.3. Herramientas de Interpretabilidad . . . 38

3. Algunas arquitecturas 43
3.1. Introducción . . . 43
3.2. Convolutional Neural Network (CNN) . . . 44
3.3. Recurrent Neural Network (RNN) . . . 47
3.4. (Long Short-Term Memory (LSTM) . . . 47
3.5. Gated Recurrent Unit (GRU) . . . 48
3.6. Generative Advesarial Network (GAN) . . . 49
3.7. Autoencoder . . . 51
3.8. Transformer . . . 53

4. Requisitos para Preparar un Buen Dataset **57**
4.1. ¿Cómo Debe Ser un Buen Dataset? . . . 57
4.2. Cuidado en el Uso de Transfer Learning . . . 60
4.3. Cuidado con la Generación de Muestras Sintéticas . . . 61
4.4. Las Redes Neuronales a Veces Son un Misterio . . . 63

5. El Método Científico **67**
5.1. ¿Qué es el Método Científico? . . . 67
5.2. Falacia . . . 69
5.2.1. Falacias de Relevancia . . . 69
5.2.2. Falacias de Ambigüedad . . . 71
5.2.3. Falacias de Falsa Causa . . . 71
5.2.4. Falacias de Presunción . . . 72
5.2.5. Falacias de Ignorancia . . . 73
5.2.6. Falacias de Inducción . . . 74
5.3. Falacias en Computación . . . 75
5.4. Las Dos Obsesiones del Método Científico . . . 77
5.5. Refutabilidad . . . 77
5.6. Refutabilidad en Computación . . . 79
5.7. Replicabilidad . . . 79
5.8. Replicabilidad en Deep Learning . . . 80
5.9. Algunos Consejos para Aproximar estos Algoritmos al Método Científico 82

6. Los Sesgos, Ese Gran Problema... **85**
6.1. Los Algoritmos Machine Learning No Entienden al Ser Humano . . . 85
6.2. Los Algoritmos No Tienen Sesgos. Los Seres Humanos, Sí . . . 86
6.3. ¿Somos Conscientes de Nuestros Propios Sesgos, de Nuestros Prejuicios? 86
6.4. El Reflejo de los Sesgos Humanos en Machine Learning . . . 90
6.5. Un Par de Agravantes al Problema de los Sesgos . . . 92
6.5.1. El Poder de la Influencia Social . . . 92
6.5.2. El anonimato en redes sociales . . . 94
6.6. El Resultado de Todo Esto . . . 95
6.7. Correlaciones Espurias . . . 96
6.8. Consejos para la Minimización de Sesgos . . . 99

7. ¿Funcionan Bien Estas Herramientas? ¿Son Seguras? **103**
7.1. Natural Language Processing (NLP) . . . 103
7.2. El Fenómeno de las Alucinaciones . . . 109
7.3. Desmitificamos los Grandes Modelos Generativos . . . 112
7.4. De Nuevo, la Importancia del Dataset . . . 112
7.5. ¿Funcionan Bien o No? . . . 113

8. Reflexiones ante un Fenómeno Transversal al Desarrollo de Modelos Generativos (2023-2024) 115
8.1. Suceso . . . 115
8.2. Biometría . . . 116
8.3. Enfermedades Detectables a Través del Ojo . . . 117
8.4. Reflexiones Relacionadas con Donar una Foto de Tu Iris, de Tu Ojo 118

9. Algunas Reflexiones Éticas 121
9.1. Qué Difícil... . . . 121
9.2. Pongamos en su Sitio lo que Son Realmente estos Modelos y Tal Vez No Sea Tan Difícil . . . 122
9.3. Contenidos de Redes Sociales como Dataset . . . 123
9.4. Decisiones Militares Tomadas por un Modelo Deep Learning . . . 123
9.5. Supuestos Éticos . . . 124
9.5.1. Supuesto 1: Condiciones Laborales . . . 124
9.5.2. Supuesto 2: Impacto Medioambiental . . . 125
9.5.3. Supuesto 3: Manipulación . . . 127

10. Reflexiones de Índole Jurídica 129
10.1. No Somos Juristas . . . 129
10.2. Datos Personales . . . 130
10.3. Derechos de Autor, Propiedad Intelectual . . . 131
10.4. ¿A Quién Damos el Óscar? . . . 133
10.5. Pederastia . . . 133
10.6. Desafíos Jurídicos . . . 134
10.7. Una Vez Más, Autocrítica . . . 135

11. Ventajas de un Uso Adecuado de Herramientas Deep Learning 137
11.1. Introducción . . . 137
11.2. Eliminación de Tareas de Bajo Valor Añadido . . . 138
11.3. Análisis Masivo de Datos . . . 140
11.4. Procesado de Estructuras de Datos Complejas . . . 141
11.5. Rentabilidad y Universalidad . . . 142
11.6. Los Algoritmos No Pueden Tener Sesgos . . . 143
11.7. Auto-conocimiento . . . 143

12. Y Terminamos 145

Bibliografía 149

PLOS ONE

RESEARCH ARTICLE

Classification of SARS-CoV-2 sequences as recombinants via a pre-trained CNN and identification of a mathematical signature relative to recombinant feature at Spike, via interpretability

Ana Guerrero-Tamayo[1]*, Borja Sanz Urquijo[1], Isabel Olivares[2], Maria-Dolores Moragues Tosantos[3], Concepción Casado[2], Iker Pastor-López[1]

1 Faculty of Engineering, University of Deusto, Bilbao, Biscay, Spain, **2** National Microbiology Center (NMC), Instituto de Salud Carlos III (ISCIII), Majadahonda, Madrid, Spain, **3** Faculty of Medicine and Nursing University of the Basque Country UPV/EHU, Leioa, Biscay, Spain

*ana.guerrero@deusto.es

Check for updates

OPEN ACCESS

Citation: Guerrero-Tamayo A, Sanz Urquijo B, Olivares I, Moragues Tosantos M-D, Casado C, Pastor-López I (2024) Classification of SARS-CoV-2 sequences as recombinants via a pre-trained CNN and identification of a mathematical signature relative to recombinant feature at Spike, via interpretability. PLoS ONE 19(8): e0309391. https://doi.org/10.1371/journal.pone.0309391

Editor: Kapil Gupta, UPES: University of Petroleum and Energy Studies, INDIA

Received: December 21, 2023

Accepted: August 10, 2024

Published: August 26, 2024

Copyright: © 2024 Guerrero-Tamayo et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All datasets of the experiments conducted are available in Zenodo, DOI 10.5281/zenodo.10848103 (https://doi.org/10.5281/zenodo.10848103).

Funding: This work was supported by the Research Training Grants Program - University of Deusto Ref. FPI UD_2021_10. The funders had no role in study design, data collection and analysis,

Abstract

The global impact of the SARS-CoV-2 pandemic has underscored the need for a deeper understanding of viral evolution to anticipate new viruses or variants. Genetic recombination is a fundamental mechanism in viral evolution, yet it remains poorly understood. In this study, we conducted a comprehensive research on the genetic regions associated with genetic recombination features in SARS-CoV-2. With this aim, we implemented a two-phase transfer learning approach using genomic spectrograms of complete SARS-CoV-2 sequences. In the first phase, we utilized a pre-trained VGG-16 model with genomic spectrograms of HIV-1, and in the second phase, we applied HIV-1 VGG-16 model to SARS-CoV-2 spectrograms. The identification of key recombination hot zones was achieved using the Grad-CAM interpretability tool, and the results were analyzed by mathematical and image processing techniques. Our findings unequivocally identify the SARS-CoV-2 Spike protein (S protein) as the pivotal region in the genetic recombination feature. For non-recombinant sequences, the relevant frequencies clustered around 1/6 and 1/12. In recombinant sequences, the sharp prominence of the main hot zone in the Spike protein prominently indicated a frequency of 1/6. These findings suggest that in the arithmetic series, every 6 nucleotides (two triplets) in S may encode crucial information, potentially concealing essential details about viral characteristics, in this case, recombinant feature of a SARS-CoV-2 genetic sequence. This insight further underscores the potential presence of multifaceted information within the genome, including mathematical signatures that define an organism's unique attributes.

1 Introduction

The evolution of viruses poses a significant challenge in pandemic control. Two main mechanisms are responsible for the high rate of viral evolution: mutation and genetic recombination.

decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Both mechanisms occur with great frequency during viral replication [1, 2]. Mutations introduce random errors into the genetic material of a virus, resulting in genetic variants of the same virus [3]. Genetic recombination is the exchange of genetic information between two viral genomes of the same or different viruses, resulting in a new hybrid genome [4, 5]. Both processes can generate variants that hinder the prevention and treatment of infectious diseases. Recombination can occur in many RNA viruses, having been detected with high frequency in picornaviruses [6], coronaviruses [7, 8], and retroviruses [9]. Recombinant viruses represent an enigma in the realm of infectious diseases, as their consequences can vary widely. In some instances, genetic recombination may result in viruses showing no significant changes in their behavior or pathogenicity, or remaining relatively harmless. However, in other scenarios, this recombination can lead to the emergence of new viral strains with unique characteristics, such as increased transmission capacity or virulence [10–16]. Even on rare occasions, genetic recombination has been the cause of failure of attenuated virus vaccines [17, 18]. The genetic recombination that occurs among different viruses, as observed in SARS [19] and MERS [20], can have significant implications for viral evolution and its impact on public health [21].

SARS-CoV-2, short for Severe Acute Respiratory Syndrome Coronavirus 2, is a novel coronavirus that emerged in late 2019 and quickly spread to become a global pandemic [22]. This highly contagious virus is responsible for the Coronavirus Disease 2019 (COVID-19), characterized by a range of symptoms from mild respiratory issues to severe pneumonia and, in some cases, death [23]. According to World Health Organization (WHO), as of September 1st, 2023, COVID-19 has caused 770,437,327 confirmed cases, including 6,956,900 deaths [24].

Since the outbreak of the pandemic in January 2020 [25], several variants have emerged with evolutionary characteristics leading to either increased transmissibility or enhanced vaccine evasion. Among the Variants of Concern (VOCs), Alpha (B.1.1.7) stood out, reaching its zenith around April 2021 [26]. Another VOC of significant impact was Delta (B.1.617.2), reaching its peak around September 2021 [27, 28]. From early 2022 to the present day, Omicron and its different subvariants (B.1.1.529, BA.1—BA.5, BQ.1, etc.) have dominated the epidemiological landscape [29, 30].

The recombination events detected in SARS-CoV-2 have occurred solely between different lineages of the virus, with no substantial modifications in terms of morbidity, mortality, etc [31]. Most of these recombinants have occurred between co-circulating Omicron sublineages, such as BA.1 (or BA.1.1), and the Delta variant or BA.2 [32]. However, recombination between Omicron and Delta could potentially result in a virus with Omicron's transmissibility and Delta's potentially increased risk of severe illness, leading to a new and more concerning scenario for public health [33].

Understanding genetic recombination phenomena holds paramount importance in anticipating and addressing pandemics, as well as in the surveillance and control of the emergence of new viruses or variants [34].

Deep Learning tools can provide new insights into the study of genetic recombination [35, 36] and the anticipation of new pandemics or emerging viruses [37]. By analyzing large genomic datasets and identifying patterns, Deep Learning tools can detect signs of genetic recombination in viruses more quickly than traditional methods. In this way, we can not only reduce response times to emerging viruses but also anticipate their emergence. Likewise, we can unravel the mysteries of the genetic code from new perspectives, such as the search for mathematical patterns within the genome itself.

2 Materials and methods

All experiments were ran in this equipment:

- Processing Unit: Intel(R) Core(TM) i7–4770K CPU. 3.5 GHz.
- Installed RAM: 32 GB usable.
- Operative System: Windows 10 Education. Version: 22H2.
- GPU: NVIDIA GeForce RTX 3090. Total memory: 40 GB.

2.1 SARS-CoV-2 complete genomic sequences compendium

We downloaded the complete collections of SARS-CoV-2 sequences by variant from the NCBI Virus Database (National Center for Biotechnology Information, Virus Database), in March 2023. Out of a total of 1,541,293 sequences, 1,539,728 were assigned as non-recombinant, and 1,565 recombinant, and their variant distribution are detailed in Tables 1 and 2, respectively.

The Variants of Concern (VOCs) began to emerge around November 2020, with the Alpha variant being the most prominent at that time [38]. For classification purposes, we refer to variants identified between January 2020 and November 2020 as pre-VOC variants. Table 1 shows the compilation of non-recombinant sequences by variants (pre-VOC and VOC).

The prevalent variants are, first and foremost, the collection known as pre-VOC, with 218,198 sequences, followed by Alpha, with 198,722 sequences, Delta, with 325,285 sequences, and, above all, Omicron, along with all its sub-variants, totaling 748,635 variant sequences. The impact of the remaining variants has been more limited, due to the dominance of those prevalent ones that gained an evolutionary advantage. Therefore, the rest of the variants represent only 48,888 sequences.

Table 2 shows the SARS-CoV-2 compilation of recombinant sequences by variants. The compilation of recombinant sequences is more balanced than in the case of non-recombinants, with a slight prevalence of XBB sub-variants over the others.

Table 1. Non-recombinant SARS-CoV-2 sequence compendium. Downloaded from the NCBI Virus Database (National Center for Biotechnology Information, Virus Database) [38], the total of 1,539,728 sequences corresponds to an approximate date of March 2023 (Release Date of NCBI Virus Database). The variants are sorted by the approximate date of appearance (data obtained from GISAID Initiative—Tracking of hCoV-19 Variants) [40]. The column "Variant" indicates the WHO Name of the SARS-CoV-2 variant. The column "No." indicates the total number of downloaded complete sequences. The "Percentage" column indicates the percentage that this total number of complete sequences represents out of the total downloaded sequences.

Non-recombinant SARS-CoV-2 Sequences		
Variant	**No.**	**Percentage**
PRE-VOC	218,198	14.17%
ALPHA	198,722	12.91%
BETA	856	0.06%
GAMMA	11,937	0.78%
DELTA	325,285	21.13%
EPSILON	14,781	0.96%
ETA	738	0.05%
IOTA	19,361	1.26%
KAPPA	145	0.01%
LAMBDA	456	0.03%
MU	49	0.00%
THETA	12	0.00%
ZETA	553	0.04%
OMICRON	748,635	48.62%
TOTAL	**1,539,728**	**100%**

https://doi.org/10.1371/journal.pone.0309391.t001

Table 2. Recombinant SARS-CoV-2 sequence compendium. Downloaded from the NCBI Virus Database (National Center for Biotechnology Information, Virus Database) [39], the total of 1,565 sequences corresponds to an approximate date of March 2023 (Release Date of NCBI Virus Database). The nomenclature for recombinant variants begins with an "X" in the Pango nomenclature [41]. The column "Variant" indicates the WHO Name of the SARS-CoV-2 variant. The column "No." indicates the total number of downloaded complete sequences. The "Percentage" column indicates the percentage that this total number of complete sequences represents out of the total downloaded sequences.

Recombinant SARS-CoV-2 sequences					
Variant	**No.**	**Percentage**	**Variant**	**No.**	**Percentage**
XA	2	0.13%	XBB.3	62	3.96%
XAA	47	3.00%	XBB.4	8	0.51%
XAB	2	0.13%	XBD	30	1.92%
XAC	37	2.36%	XBE	32	2.04%
XAD	1	0.06%	XC	227	14.50%
XAE	18	1.15%	XD	5	0.32%
XAF	16	1.02%	XE	256	16.36%
XAG	8	0.51%	XF	4	0.26%
XAH	1	0.06%	XG	1	0.06%
XAJ	52	3.32%	XH	4	0.26%
XAK	21	1.34%	XL	11	0.70%
XAM	59	3.77%	XM	10	0.64%
XAN	5	0.32%	XN	16	1.02%
XAP	41	2.62%	XP	7	0.45%
XAT	2	0.13%	XQ	14	0.89%
XAU	4	0.26%	XS	24	1.53%
XB	19	1.21%	XW	36	2.30%
XBB	67	4.28%	XY	63	4.03%
XBB.1	88	5.62%	XZ	59	3.77%
XBB.2	206	13.16%	**TOTAL**	**1,565**	**100%**

https://doi.org/10.1371/journal.pone.0309391.t002

2.2 Dataset design

The prevalent variants throughout the SARS-CoV-2 pandemic (and its worst moments) were non-recombinants [42]. Primarily for this reason, the number of non-recombinant variant sequences is substantially greater than that of recombinants. In response to this disparity, we opted to implement a subsampling technique in the larger non-recombinant dataset. This strategy involves selecting a random, representative subsample from the larger category, thereby equalizing the number of data points between both categories. This, in turn, helps mitigate potential biases in our analysis and enhances the validity of our results [43]. We randomly selected 1,565 non-recombinant sequences to work with a balanced dataset. To ensure the generalization of our results, we performed a significant and sufficient number of different subsamplings. In this case, 10 subsamplings, labeled with sequential numbers from 01 to 10 (SUB_01—SUB_10).

Once the subsampling of non-recombinant sequences was completed, they were randomly distributed among the Training, Validation, and Test sets [44], as detailed in Table 3, which illustrates the structure of each dataset generated by subsampling.

2.3 Generation of genomic spectrograms

The generation of spectrograms follows the procedure corresponding to the superposed spectrograms in our previous work [45], were we applied transfer learning to a pre-trained

Table 3. Dataset structure. We conducted all experiments using three balanced datasets between both categories, allocating 60% to the Training Set, 20% to the Validation Set, and 20% to the Test Set.

Dataset	Non-recombinant	Recombinant	TOTAL
Training	939	939	1,878 (60%)
Validation	313	313	626 (20%)
Test	313	313	626 (20%)
TOTAL	**1,565**	**1,565**	**3,130**

https://doi.org/10.1371/journal.pone.0309391.t003

Convolutional Neural Network (CNN) VGG-16 [46] using the ImageNet dataset [47]. This network was trained to detect the recombinant feature in complete HIV-1 sequences.

In this spectrogram representation, the z-axis represents the summation of values from each of the four nucleotide types along the z-axis.

$$S = S_a + S_g + S_c + S_t \tag{1}$$

Fig 1 shows an example of the genomic spectrogram of HIV-1, and Fig 2 shows that of SARS-CoV-2. The length of x-axis matches the genome length.

In both cases, the y-axis represents a frequency range from 0 to 0.5 Hz. The z-axis represents the spectrogram computation with an applied jet colormap scale. Lower values are represented in blue with a progressive scaling towards the color red, which represents higher values, transitioning through intermediate colors in the range of greens, yellows, and oranges.

We generated spectrograms of both datasets using Python, the Scipy library, scipy.signal.spectrogram. The length of the SARS-CoV-2 genome is approximately three times greater than that of the HIV-1 genome, and this contrast is evident in the spectrogram due to the fixed value of 256 used as the length of each time segment for FFT calculation (nperseg). In the

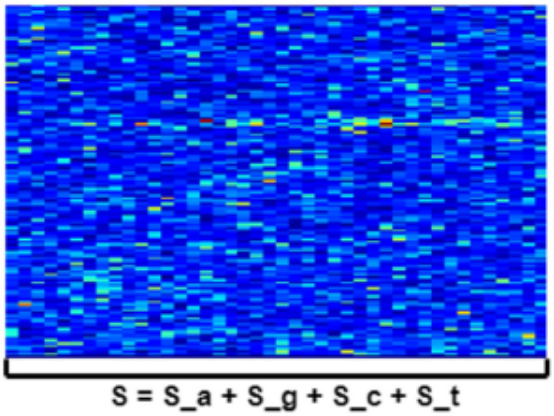


Fig 1. HIV-1 genomic spectrogram scheme. In the case of HIV-1, the length of the x-axis is approximately 10,000 nucleotides.

https://doi.org/10.1371/journal.pone.0309391.g001

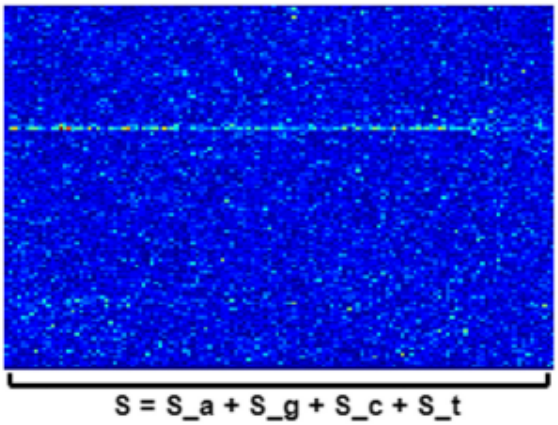


Fig 2. SARS-CoV-2 genomic spectrogram scheme. In the case of SARS-CoV-2, the length of x-axis is around 30,000 nucleotides.

https://doi.org/10.1371/journal.pone.0309391.g002

spectrogram of SARS-CoV-2, the x-axis is three times longer than in the case of HIV-1, resulting in smaller color points along the z-axis [48].

In both cases, the horizontal line at $f = 1/3$ is clearly visible. This observation may be related to the conversion of three nucleotides into a single amino acid in the coding regions of the genome [49]. It is expected that this line will be sharp in these coding regions but less pronounced in non-coding regions. In the case of viruses, a significant portion of the genome is coding [50]. Hence, this line is perfectly visible throughout nearly the entire genomic spectrogram. The appearance of this line at $f = 1/3$ is an indicative of a correct generation of genomic spectrograms.

This graphical representation of genome in the frequency spectrum allows for a more accurate identification of the genome regions crucial for the recombinant feature.

2.4 Two-stage transfer learning

We performed a two-stage transfer learning process as indicated in Fig 3. The first of these stages started from the network derived from [45] for superposed spectrograms.

All the experiments were performed using the MATLAB2021b App Deep Learning Designer.

2.5 Test bench

We conducted a Test Bench on each of the 10 subsamplings, evaluating hyperparameter values with a fixed Learning Rate of 0.0001, a fixed Batchsize of 52, and varying the number of Epochs at 10, 25, 50, 75, 100, 150, 200, and 250.

Consequently, we conducted a total of 80 experiments, resulting in 80 VGG-16 trained through Two-Stage Transfer Learning.

We conducted preliminary tests with a Batchsize of 128 and a Learning Rate of 0.01, which yielded suboptimal results. Consequently, we found that the optimal values for Batchsize and

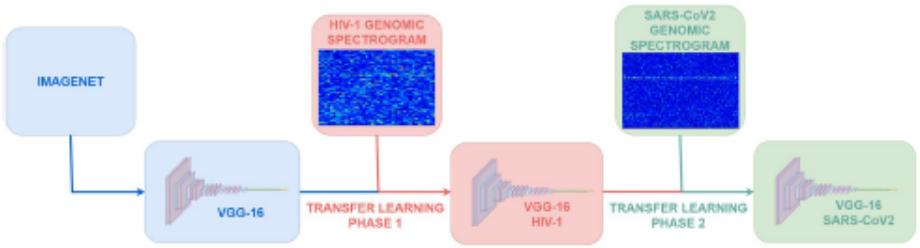


Fig 3. Two-stage transfer learning methodology. We started with a pre-trained VGG-16 using the ImageNET dataset. In Phase 1, we applied transfer learning to the genomic spectrogram dataset of complete HIV-1 sequences to detect the recombinant feature. In Phase 2, we applied transfer learning once again to the resulting network from Step 1 (VGG-16 HIV-1) using a genomic spectrogram dataset of complete SARS-CoV-2 sequences to also detect the recombinant feature (VGG-16 SARS-CoV-2).

https://doi.org/10.1371/journal.pone.0309391.g003

Learning Rate in this second stage of transfer learning align with those used in the first stage (HIV-1).

2.6 Performance measurement

As performance metrics, alongside Validation Accuracy and Test Accuracy, we calculated the Area Under the Curve (AUC) and the Confusion Matrix on the Test Set for all experiments.

Our criterion for determining the optimal configurations was based on selecting those that, with remarkable values of AUC and Validation Accuracy, not only achieve the highest Test Accuracy but also maintain a balance in both categories (recombinant and non-recombinant) [51].

We calculated all performance measurements using MATLAB2021b.

2.7 Interpretability analysis

We applied interpretability techniques to discern, via heatmaps, the critical influences on the outputs of each of the generated models. Our tool of choice for pinpointing the regions of the genome where the network looks to make decisions is Grad-CAM [52]. It offers greater visual clarity compared to other tools like LIME [53] or Gradient Attribution [54], albeit at the cost of some precision.

The color scale applied to these heatmaps is a jet map, whose color distribution based on the value of the scoremap at each point is shown in Fig 4.

To process the Grad-CAM results, we performed a three-step image processing to progressively determine the relevant Total Hot Zones in the recombinant feature. Fig 5 graphically illustrates the interpretability analysis methodology.

In the first step, we obtained the scoremaps for each sequence, in each subsampling, and for each hyperparameter configuration. In the second step, we calculated the total hot zones per



Fig 4. Jetmap color scale. Color values increase from blue to red, with blue indicating lower values and red indicating higher values, transitioning through intermediate colors such as greens, yellows, and oranges.

https://doi.org/10.1371/journal.pone.0309391.g004

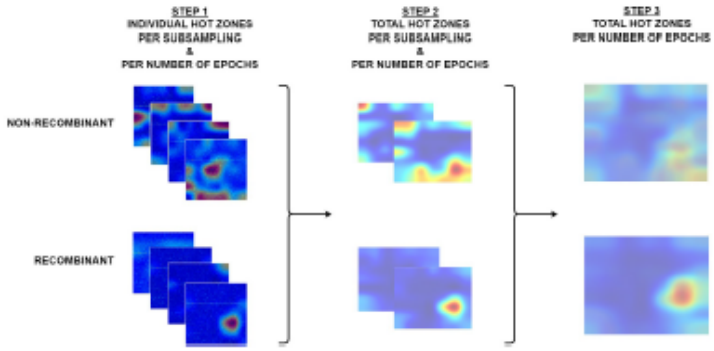


Fig 5. Three-step interpretability. Considering that we conducted 80 experiments (test bench applied to 10 subsamplings), the first step involved a total of 50,080 images, taking into account that each complete test set contains 626 sequences. The second step involves 160 images (80 experiments and 2 categories). And the third step involves a total of 16 images per category and number of Epochs.

https://doi.org/10.1371/journal.pone.0309391.g005

category in each subsampling and for each hyperparameter configuration. In the third step, we calculated the total scoremap image per category for each hyperparameter configuration, considering the ten subsamplings. The result of this third step represents the relevant hot zones across the set of subsamplings by category.

For the calculation of the images resulting from Steps Two and Three, we applied two different techniques, which were determined by their input data. For the calculation of the Total Hot Zones in Step Two, we processed the scoremaps (numerical matrices corresponding to each of the sequences) and summed the numerical values at each position in the matrix.

$$H(STEP2) = \sum_{i=1}^{n} H(STEP1)_i, \tag{2}$$

The scalar summation of Grad-CAM (Class Activation Maps) and the generation of an average heatmap are two different approaches to summarize and visualize the importance of regions of interest in the images. Each approach has specific advantages.

The arithmetic summation (Step Two) allowed, in a straightforward and computationally efficient manner, to calculate the total hot zone for each of the 80 experiments conducted, without any inherent loss of image processing accuracy. Subsequently applying the same color map to this resulting numerical matrix allowed us to obtain the total hot zone image per category for each subsampling and each experiment in the test bench.

This way, we obtained a clear and representative heatmap of the hot zones in the dataset. For example, the total hot zones per category in subsampling 05 for 200 Epochs are shown in Figs 6 and 7. Fig 6 shows the total hot zones in the case of non-recombinants and Fig 7 in the case of recombinants.

In Step Three, we generated composite hot zone images for each hyperparameter configuration across the 10 subsamplings. These images represent the arithmetic mean of the hot zones

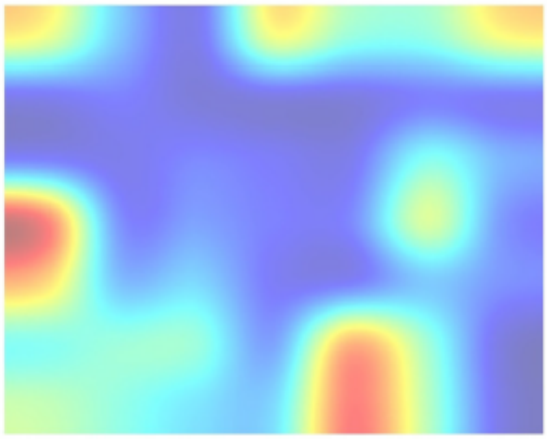


Fig 6. Total hot zones (SUB_05, 200 Epochs) in non-recombinants. The non-recombinant hot zones are more variable and diffuse.

https://doi.org/10.1371/journal.pone.0309391.g006

obtained from each subsampling.

$$H(STEP3) = \frac{1}{n}\sum_{i=1}^{n} H(STEP2)_i \tag{3}$$

Where n represents the number of subsamplings, in this case, 10.

Step Three visually represents the common and relevant hot zones for the recombinant feature considering the 10 subsamplings. Since we only had 10 images per category (non-recombinant and recombinant) in the initial data, we implemented image processing techniques to calculate a weighted average of each pixel in the image set, creating a comprehensive total image of the hot zones across the 10 subsamplings. This technique allowed us to diminish the significance of noisy or atypical regions in the individual maps, achieving a generalized view of the important areas in each category.

The application of both image processing techniques enabled us to attain a more comprehensive view of the total hot zones throughout the whole process.

After obtaining the total hot zone in Step Three, we modified the color scale to visually enhance the hot zones. We achieved this by normalizing the average matrix so that the lowest value equals 0, and the highest equals 255 as follows [55]:

$$\text{Norm. Avg. Matrix} = \left(\frac{\text{Avg. Matrix} - \text{Avg. Matrix Min.}}{\text{Avg. Matrix Max.} - \text{Avg. Matrix Min.}}\right) \times 255 \tag{4}$$

Where Avg. Matrix represents the resulting average matrix from the Step Three. Avg. Matrix Min. represents the minimum value contained in Avg. Matrix, and Avg. Matrix Max. represents the maximum value.

Finally, in one last adjustment, we generated the negative of the resulting matrix so that the minimum value appears as a light color and the maximum as a dark color. All of this was done

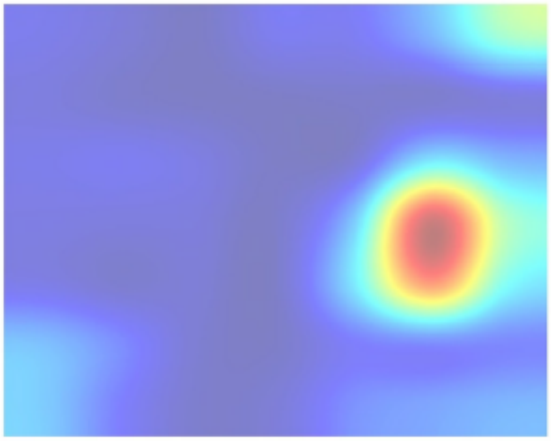


Fig 7. Total hot zones (SUB_05, 200 Epochs) in recombinants. However, in the case of recombinant sequences, a clear, well-defined hot zone appears, concentrated in the same area.

https://doi.org/10.1371/journal.pone.0309391.g007

with the aim of enhancing visualization and highlighting the location of the total hot zones per epochs [56].

All of these processes were conducted using MATLAB 2021b along with Python library functions, utilizing cv2 for image processing and numpy for multidimensional array manipulation and algebraic operations [57].

3 Results and discussion

3.1 Results per subsampling

S1 Appendix includes the complete set of results per subsampling in terms of performance. The performance metrics are detailed in Section 2.6. Those configurations (specified by the number of epochs) that yielded best results, meaning highest test accuracy values and a more balanced distribution of hit rates between the two categories, are highlighted in green.

We evaluated the balance between the two categories by computing the Standard Deviation (SD) between the test accuracy values for recombinants and non-recombinants.

Therefore, we considered the optimal configurations to be those with the highest hit rate in the test set and the most balanced distribution (lower SD between the test accuracies of both categories).

Fig 8 graphically represents the confusion matrix scheme outlined in Table 4.

Table 4 summarizes the best configurations for each of the 10 generated subsamplings. These are the ones that exhibit the highest test accuracy values with a greater balance between both categories.

3.2 Results per number of epochs

Complete results are provided in S2 Appendix Table 5 summarizes the most relevant data.

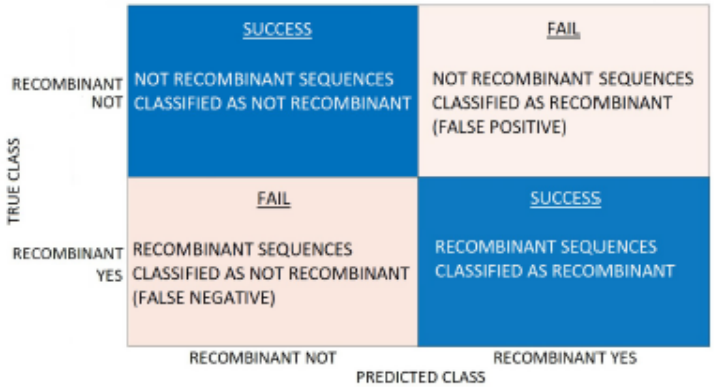


Fig 8. Confusion matrix scheme. The top row corresponds to the Non-recombinant Category, and the bottom row to the Recombinant Category.

https://doi.org/10.1371/journal.pone.0309391.g008

The configurations corresponding to 10 and 25 Epochs yielded deficient results in terms of test accuracy in both categories, and the results are unbalanced, making them inappropriate configurations due to the insufficient training with such low values of the number of epochs. The qualitative advantage of 200 Epochs over 150 or 250 is its higher degree of balance between the test accuracy of both categories. Therefore, although these three configurations exhibit high hit rates on the test set, in the case of 200 Epochs, the minimum value of Inter-Category SD was achieved.

Fig 9 displays the summary of the total hot zones per category for each subsampling by number of epochs (Step Two). From the images shown in Fig 9, we generated the corresponding images for Step Three, that is, the weighted average hot zones for each configuration.

We omitted 10 and 25 Epochs as their performance ratios were not suitable, possibly due to insufficient training.

Fig 10 displays the weighted average hot zones for each configuration (Step Three) and their enhanced counterparts. In the case of non-recombinants, the main hot zones are more diffuse, as opposed to the greater sharpness observed in recombinants. For the latter, all configurations clearly converge towards a single area.

We processed 17,215 complete sequences of SARS-CoV-2, utilizing virtually all available complete recombinant sequences at the beginning of the experimentation. We are aware that handling 10 subsamplings of the total non-recombinant pool involved processing only approximately 1% of the available non-recombinant sequences. Nevertheless, the results obtained are significant, especially in the recombinant category, indicating that it is a representative sample. Our results confirm this point.

3.3 Optimal configuration selection

The configurations with the highest number of correct predictions in the test set are 150 epochs and 200 epochs, with a total test accuracy in both cases of 94.60%. At similar test

Table 4. Best results per subsampling. The best configurations feature 200 Epochs in 80% of the subsamplings.

SUBSAMPLING 01				
Epochs	**AUC**	**Validation Accuracy**	**Test Accuracy**	**Confusion Matrix**
150	0.9818	94.25%	96.01%	302–11 14–299
200	0.9823	94.57%	96.01%	302–11 14–299
SUBSAMPLING 02				
Epochs	**AUC**	**Validation Accuracy**	**Test Accuracy**	**Confusion Matrix**
150	0.9806	94.57%	95.85%	303–10 16–297
250	0.9796	95.05%	95.53%	298–15 13–300
SUBSAMPLING 03				
Epochs	**AUC**	**Validation Accuracy**	**Test Accuracy**	**Confusion Matrix**
200	0.9824	95.21%	94.73%	297–16 17–296
SUBSAMPLING 04				
Epochs	**AUC**	**Validation Accuracy**	**Test Accuracy**	**Confusion Matrix**
200	0.9808	96.17%	95.53%	304–9 19–294
SUBSAMPLING 05				
Epochs	**AUC**	**Validation Accuracy**	**Test Accuracy**	**Confusion Matrix**
200	0.9698	95.53%	93.13%	294–19 24–289
250	0.9732	95.21%	93.13%	288–25 18–295
SUBSAMPLING 06				
Epochs	**AUC**	**Validation Accuracy**	**Test Accuracy**	**Confusion Matrix**
200	0.9782	94.57%	94.57%	297–16 18–295
SUBSAMPLING 07				
Epochs	**AUC**	**Validation Accuracy**	**Test Accuracy**	**Confusion Matrix**
150	0.9824	94.73%	95.37%	309–4 25–288
200	0.9806	94.73%	95.21%	305–8 22–291
SUBSAMPLING 08				
Epochs	**AUC**	**Validation Accuracy**	**Test Accuracy**	**Confusion Matrix**
150	0.9769	91.05%	91.69%	275–38 14–299
200	0.9754	91.69%	91.69%	275–38 14–299
SUBSAMPLING 09				
Epochs	**AUC**	**Validation Accuracy**	**Test Accuracy**	**Confusion Matrix**
100	0.9842	94.89%	96.17%	305–8 16–297
SUBSAMPLING 10				
Epochs	**AUC**	**Validation Accuracy**	**Test Accuracy**	**Confusion Matrix**
200	0.9728	93.45%	95.05%	302–11 20–293
250	0.9729	93.77%	95.05%	302–11 20–293

https://doi.org/10.1371/journal.pone.0309391.t004

Table 5. Best results per number of epochs. We include the mean value and standard deviation of AUC, the mean Test Accuracy for non-recombinant and recombinant sequences, and the standard deviation between the mean test accuracy values of recombinants and non-recombinants. As a measure of the network's accuracy, the total number of sequences in each category within the test set is 313.

	AUC		Non-rec. Test Acc.	Rec. Test Acc.	Inter Category
EPOCHS	Avg.	SD	Avg.	Avg.	SD
10	0.9013	0.0201	249.2	263.8	10.32
25	0.9535	0.0112	289.6	253.9	25.24
50	0.9696	0.0096	291.7	285.2	4.60
75	0.9729	0.0053	295.8	286.6	6.51
100	0.9761	0.0070	295.5	290.1	3.82
150	0.9787	0.0043	300.0	292.2	5.52
200	0.9785	0.0046	297.6	294.6	2.12
250	0.9785	0.0038	298.6	292.5	4.31

https://doi.org/10.1371/journal.pone.0309391.t005

accuracy values, maintaining the test accuracy constant, improving the balance between categories leads to superior model performance.

As we discussed in previous sections, the decision criteria cannot be based solely on the mere measure of total accuracy in a single category or in both categories combined. We require balanced results, hence the need to include inter-category SD in the decision criteria. In this case, the 200 Epochs configuration achieves remarkable accuracy rates in both categories (see Fig 11), and the relative difference in absolute terms is minimal. Indeed, the inter-category SD is the lowest (see Fig 12).

3.4 Reference sequence

To identify the genomic regions where the hot zones are located, we relied on the Severe Acute Respiratory Syndrome Coronavirus 2 isolate Wuhan-Hu-1, complete genome NCBI Reference Sequence: NC_045512.2 [58]. The location of each structural, non-structural, and accessory protein is indicated in Table 6.

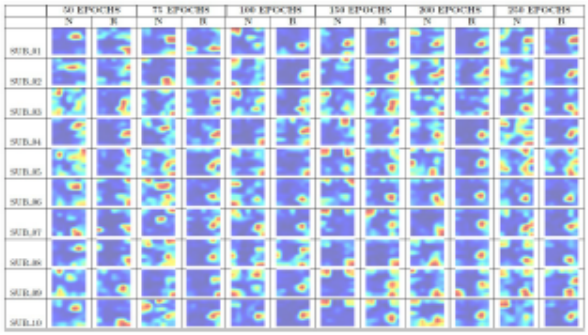


Fig 9. Summary table of total hot zones Step Two. N stands for the non-recombinant category, and R stands for the recombinant one.

https://doi.org/10.1371/journal.pone.0309391.g009

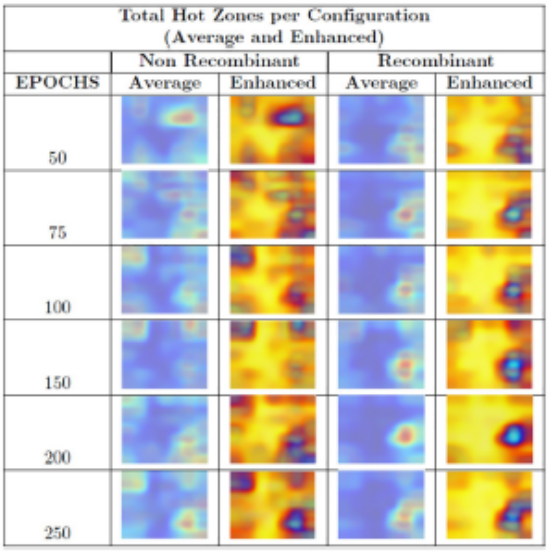


Fig 10. Total hot zones per configuration. The average hot zones represent the hot zones for each number of epochs across the 10 subsamplings. The enhanced figures are the average hot zones with color scale modifications to clarify the relevant hot zones in each category.

https://doi.org/10.1371/journal.pone.0309891.g010

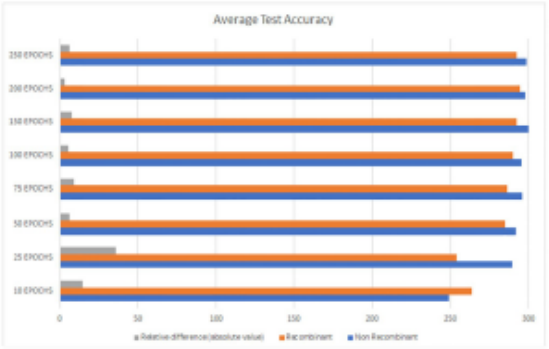


Fig 11. Average test accuracy for recombinants (in orange) and non-recombinants (in blue). In gray, we depicted the relative difference between these mean absolute values.

https://doi.org/10.1371/journal.pone.0309891.g011

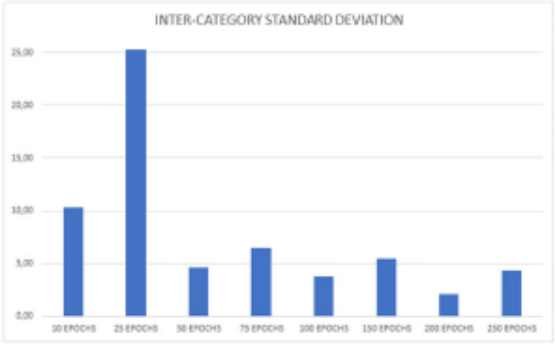


Fig 12. SD values between the average test accuracy values for recombinants and non-recombinants.

https://doi.org/10.1371/journal.pone.0309891.g012

Based on the consensus reference sequence, we constructed a scaled graphical representation of the SARS-CoV-2 genome, which will serve as a pivotal tool for the precise identification of regions involved in the recombinant feature.

3.5 Analysis of non-recombinant results

Once it is established that the optimal configuration corresponds to 200 Epochs, the next step is to identify the high-impact hot zones for classifying a sequence as non-recombinant. To do this, we calculate the overall average image (Step 3), the enhanced image, and the localization of the epicenters of the main hot zones on the x-axis (indicating genomic region involvement) and the y-axis (frequency range identification), all in accordance with the guidelines outlined in Section 2.7.

Table 6. Location of protein coding regions in the SARS-CoV2 Wuhan-Hu-1 reference genome sequence (NC_045512.2) [58]. The "Beginning" column specifies the first nucleotide of the corresponding protein, while the "End" column indicates the last nucleotide.

Protein coding region in NC_045512.2 genome sequence		
	Beginning	**End**
5'UTR	1	265
ORF1ab	266	21555
S	21563	25384
ORF3a	25393	26220
E	26245	26472
M	26523	27191
ORF6	27202	27387
ORF7a	27394	27759
ORF7b	27756	27887
ORF8	27894	28259
N	28274	29533
ORF10	29558	29674
3'UTR	29675	29903

https://doi.org/10.1371/journal.pone.0309891.t006

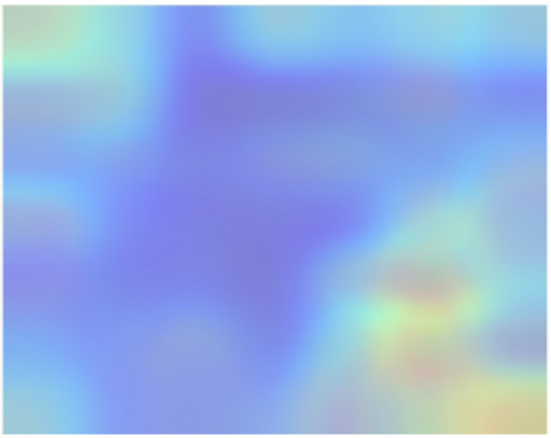


Fig 13. Non-recombinant total hot zones resulting from the 10 subsamplings for 200 Epochs.

https://doi.org/10.1371/journal.pone.0309391.g013

Figs 13–16 show the graphical analysis of the Total Hot Zones (Step 3) in Non-recombinants for 200 Epochs.

As seen in Fig 13, non-recombinant sequences do not exhibit a distinct hot zone, and the boundaries of hot zones are somewhat blurred. This phenomenon could be attributed to the greater diversity of sub-lineages and strains among non-recombinant variants, resulting in increased variability due to subsampling. As can be seen more clearly in Fig 14, subtle hotspots are hinted at around the Spike protein region. A potential third hotspot may exist towards the end of the genome, although its relevance appears to be less pronounced.

By direct extrapolation to the calculations shown in Fig 15, the epicenters of the main hot zones are situated at nucleotide positions between 23,341 and 23,362. The central position of the Spike protein (S) corresponds to nucleotide 23,473. Therefore, we can place the epicenters of both zones in the central region of the S protein.

By directly extrapolating from the calculations shown in Fig 16, regarding the vertical axis, considering its range is 0–0.5 Hz, the critical frequencies fall approximately at $f = 1/12$ and $f = 1/6$ respectively.

In a comparison between the main hot zones using a representative scale of the SARS-CoV-2 genome, it becomes evident that the two primary hotspots are located in the vicinity of the Spike protein. See Fig 17.

Therefore, even though the areas are subtle, we can observe that the main decision regions are at S protein for the frequencies $f = 1/12$ and $f = 1/6$.

3.6 Analysis of recombinant results

In the case of recombinant SARS-CoV-2 sequences, the main hot zone, where the CNN looks to detect the recombinant feature, is clearly delineated. Figs 18–21 compile the graphical analysis of the Total Hot Zones (Step 3) in Recombinants for 200 Epochs.

The sharpness of the total hot zone shown in Figs 18 and 19 denotes a prevalence of this hot spot across all subsamplings for 200 epochs.



Fig 14. Image resulting from color scale enhancement for better non-recombinant hot zone identification.

https://doi.org/10.1371/journal.pone.0309391.g014

After measuring the boundaries, when extrapolating to the length of the SARS-CoV-2 genome (29,903 nts.), the hot zone that determines a SARS-CoV-2 sequence to be recombinant is located roughly between positions 21,448 and 25,323 (see Fig 20). This location almost coincides with the position of the Spike protein (from approximately nucleotide 21,550 to roughly 25,400).

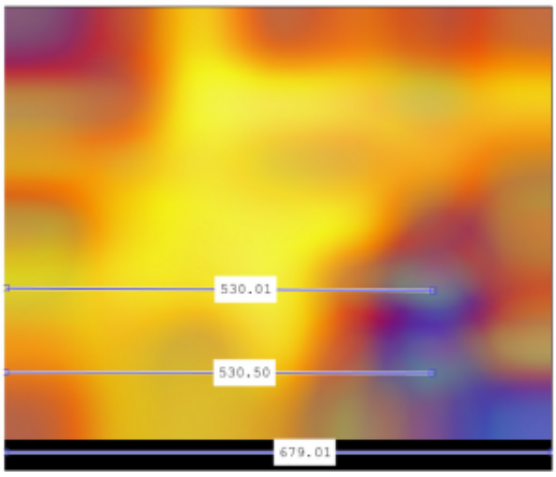


Fig 15. Horizontal positioning of the epicenters of the main non-recombinant hot zones.

https://doi.org/10.1371/journal.pone.0309391.g015

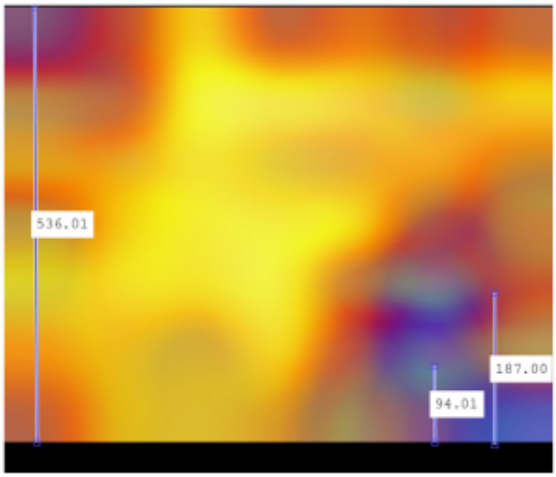


Fig 16. Vertical positioning of the epicenters of the main non-recombinant hot zones.

https://doi.org/10.1371/journal.pone.0309391.g016

Regarding the vertical axis, whose total range is 0.5 Hz, as interpreted from the result in Fig 21, the preliminary identification of the epicenter of these total hot zones is located around 0.183 Hz, that is, around $f = 1/6$.

Despite Grad-CAM's imprecise interpretability, and considering that the location of each protein may vary depending on the variant and inherent sequence variability, the hot zone closely aligns enough to infer that the neural network is focusing on the S protein to identify the recombinant feature within the sequence. See Fig 22.

Using our methodology, we determined that the main hot zone is clearly located in the S protein. The use of this methodology allowed us to pinpoint the areas of the genomic spectrogram image where the pre-trained CNN "looks" to classify a sequence as recombinant or non-recombinant.

In coronaviruses, multiple genetic recombination events occurred in the S protein [59]. This phenomenon also happened in SARS-CoV-2 itself [60, 61].

At this point, we must make a distinction between the results obtained in our research and the fact that genetic recombination occurs in the S protein. In the research conducted on HIV-1 [45], the mathematical signature embedded in the genome that caused the pre-trained CNN to classify a sequence as recombinant or non-recombinant, was predominantly located in areas near the LTRs at a frequency of $f = 1/3$, regardless of the genomic regions where genetic recombination actually occurs between the different pure subtypes of HIV-1.

In the case of SARS-CoV-2, the location of this mathematical signature was detected in the same region where multiple genetic recombination events occurred. That is, the S protein.

The coincidence in location between the mathematical pattern, the mathematical signature detected by the pre-trained CNN, and the fact that the S protein is where abundant genetic recombination events occurred in SARS-CoV-2 should be studied in future investigations to determine if there is any relationship between these factors, as well as to unravel the significance of this phenomenon from a biological perspective.

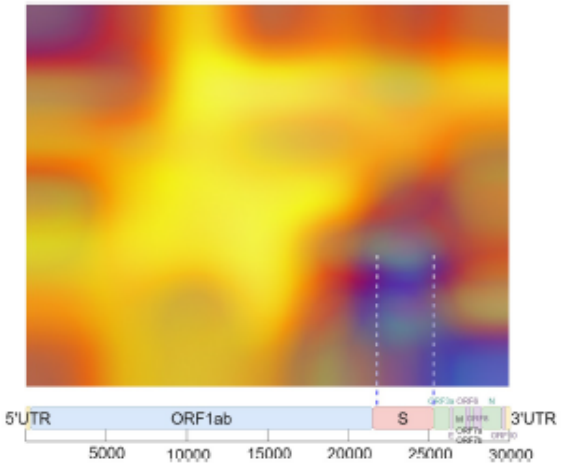


Fig 17. Localization of hot zones in non-recombinant sequences. Positioned in relation to a true-to-scale schematic representation of the composition of the SARS-CoV-2 genome. All notable proteins are appropriately marked.

https://doi.org/10.1371/journal.pone.0309391.g017

Considering that the y-axis range is 0–0.5 Hz, this equates to a frequency of 0.18 Hz. Given the limited accuracy of Grad-CAM and the additive errors in the successive mathematical transformations performed in calculating the total hot zones, it is not unreasonable to consider the vertical epicenter at frequencies close to $f = 1/6$. To perform the most reliable verification

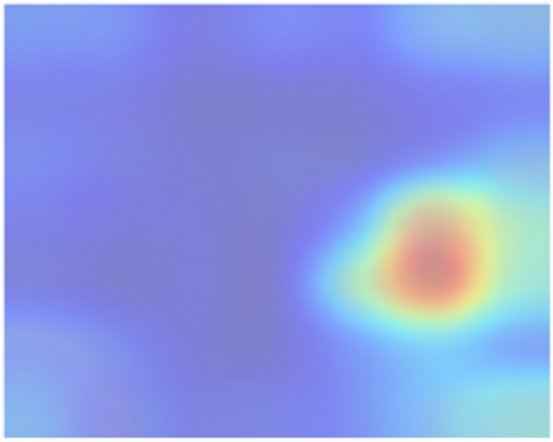


Fig 18. Recombinant total hot zones resulting from the 10 subsamplings for 200 Epochs.

https://doi.org/10.1371/journal.pone.0309391.g018

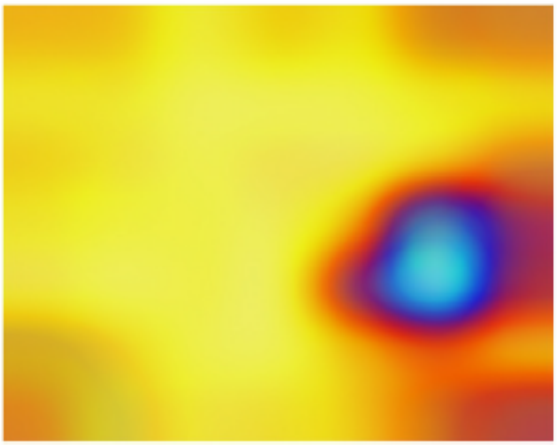


Fig 19. Image resulting from color scale enhancement for better recombinant hot zone identification.

https://doi.org/10.1371/journal.pone.0309891.g019

of the epicenter frequency, we measured it in one of the experiments where this area is depicted most prominently: 200 Epochs in subsampling 06.

Fig 23 displays the measurement obtained precisely at the epicenter of the hot zone. This figure underwent fewer mathematical transformations. Indeed, the vertical epicenter of the

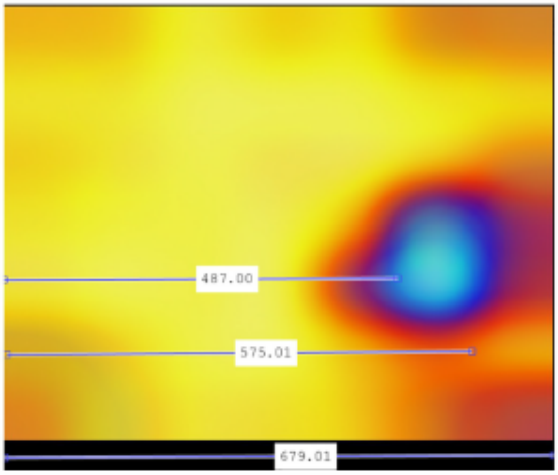


Fig 20. Horizontal positioning of the epicenters of the main recombinant hot zones.

https://doi.org/10.1371/journal.pone.0309891.g020

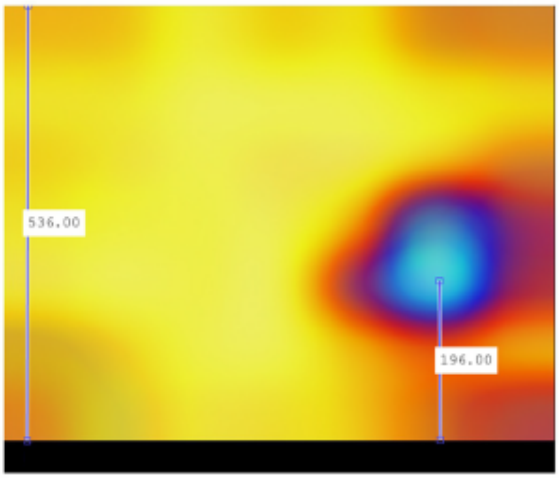


Fig 21. Vertical positioning of the epicenters of the main recombinant hot zones.

https://doi.org/10.1371/journal.pone.0309391.g021

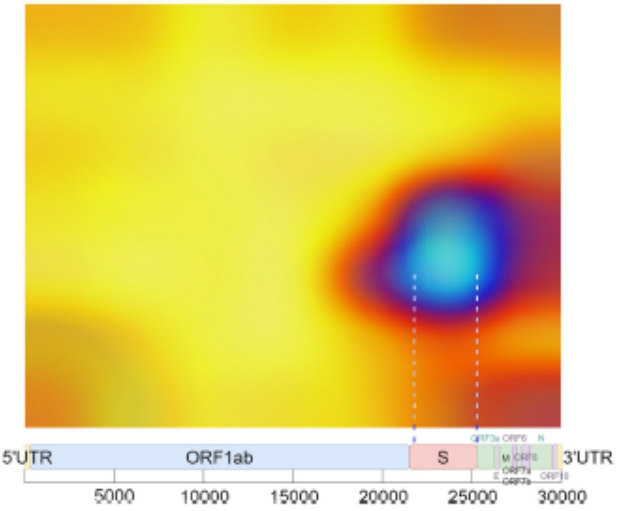


Fig 22. Localization of hot zones in recombinant sequences. Similarly to non-recombinant sequences, the main hot zones (Step 3) are positioned in relation to a scaled schematic representation of the SARS-CoV-2 genome.

https://doi.org/10.1371/journal.pone.0309391.g022

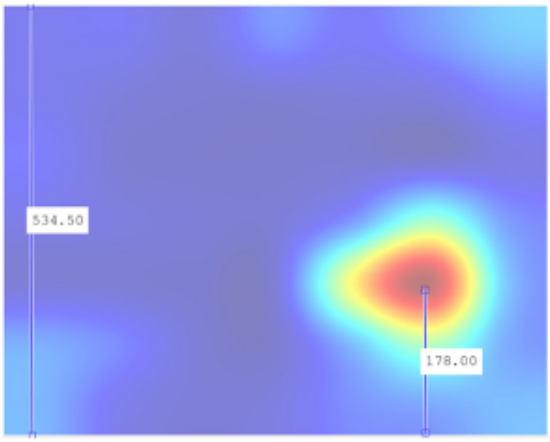


Fig 23. Vertical positioning at 200 Epochs configuration subsampling 06. Indeed, given that the total length of the y-axis is 534.5 points, with a range of 0–0.5 Hz, the epicenter's position at 178 points precisely corresponds to 0.167, a value aligning with $f = 1/6$ in the sharpest case across the entire test bench.

https://doi.org/10.1371/journal.pone.0309891.g023

hot zone is located at frequencies close to $f = 1/6$, so it is a plausible hypothesis to consider this frequency as influential in determining the recombinant feature.

4 Conclusions

Using genomic spectrograms with 10 random subsamplings to address the disparity in size between non-recombinants and recombinants, we designed a test bench to elucidate the optimal hyperparameter configurations. We applied transfer learning in 2 phases using a pre-trained VGG-16 model on the ImageNet dataset. Phase 1 was focused on HIV-1 genomic spectrograms, and Phase 2 on those of SARS-CoV-2. All of this with the goal of detecting the recombinant feature of a SARS-CoV-2 genomic sequence. Subsequently, we applied the Grad-CAM interpretability tool in 3 steps to identify the hot zones (where the CNN looks for classification) in each sequence, in each subsample for every configuration, and in total in each configuration. We applied image processing techniques to enhance the localization of the hot zones. These 3 steps involve not only the mere application of Grad-CAM but also the mathematical processing of its results to extrapolate the obtained outcomes. The image processing techniques used allowed us to delineate the relevant areas for the recombinant feature as clearly as possible.

We obtained consistent and well-defined results in each category. In the case of SARS-CoV-2, the spike protein emerges as a determinant in both recombinant and non-recombinant categories.

The evident significance of the S protein in identifying the recombinant feature in SARS-CoV-2 aligns with the excellent research conducted by Nikolaidis et al. [59]. They uncovered multiple instances of double crossover genetic recombination events across various CoVs, and

interestingly, the majority of these events are precisely located within this protein. Therefore, our work in a way reinforces their results by means of a different approach.

In the case of the non-recombinants, the hot zones (Step 3) are more diffuse, although they appear to pivot around the area of the spike protein within the frequency range of $f = 1/12$ and $f = 1/6$.

Nevertheless, the clarity of the main hot zones in Steps 2 and 3 is particularly striking in the case of recombinant sequences. A region corresponding to the Spike protein is clearly elucidated, at an approximate frequency of $f = 1/6$.

By utilizing Deep Learning tools, with their high potential in pattern recognition in images [62, 63], we were able to identify the determinant regions in the recombinant feature of genomic spectrograms of SARS-CoV-2. Achieving high test accuracy and robust, distinguishable hot zones in both categories.

5 Future research

In summary, we detected a mathematical signature that characterizes a genomic sequence of SARS-CoV-2 as recombinant. This signature is located in the S protein, with its epicenter at a frequency of $f = 1/6$. Consequently, the location of this mathematical signature is related to a nucleotide periodicity of 6, meaning that in the arithmetic series, every 6 nucleotides (two triplets) in S may encode crucial information related to the recombinant feature in SARS-CoV-2.

We know where the CNN looks to classify a SARS-CoV-2 sequence as recombinant. Now, we want to understand what it sees. What is the mathematical pattern embedded in the frequency spectrum of the genome of the Spike protein that causes a sequence to be classified as recombinant?

Our future research should focus on determining not only the formulation of this mathematical signature embedded in the genome but also its biological significance.

Another interesting line of research would be to determine the relationships between the dispersion detected in the hot zones in the non-recombinant category with the abundance and phylogenetic diversity of the set of non-recombinant variants in SARS-CoV-2.

In light of the results obtained, the identification of mathematical signatures in the virus genome through genomic spectrogram analysis opens up new avenues to investigate potential functions associated with these mathematical patterns.

Supporting information

S1 Appendix. Complete results per subsampling.
(ZIP)

S2 Appendix. Complete results per number of epochs.
(ZIP)

Acknowledgments

We would like to thank Professor Ana Macarulla-Arenaza (University of Deusto—Faculty of Engineering), Professor Nekane Ione Sainz Bedoya (University of Deusto—Faculty of Engineering), Head of the General Secretariat María Isabel Muñoz San Ildefonso (University of Deusto), Professor Juan María Collantes Metola (University of the Basque Country UPV/EHU—Faculty of Science and Technology), Professor Ibone Lizarraga Mallo (University of the Basque Country UPV/EHU—Faculty of Science and Technology), Professor Ignacio Angulo (University of Deusto—Faculty of Engineering), Research Associate Tony Castillo-Calzadilla (DeustoTech Energy & Environment), Research Assistant Maite Puerta-Beldarrain

(DeustoTech MoreLab), and Research Assistant Armando Mendoza-Aguayo (DeustoTech Energy & Environment) for their support and assessment. We would like to especially thank the Editors and Reviewers for taking the time and effort necessary to review our manuscript. We sincerely appreciate all the valuable comments and suggestions, which helped us improve the quality of this article.

Author Contributions

Conceptualization: Ana Guerrero-Tamayo, Borja Sanz Urquijo, Isabel Olivares.

Data curation: Ana Guerrero-Tamayo, Concepción Casado, Iker Pastor-López.

Formal analysis: Ana Guerrero-Tamayo, Isabel Olivares, María-Dolores Moragues Tosantos, Concepción Casado.

Funding acquisition: Borja Sanz Urquijo, Iker Pastor-López.

Investigation: Ana Guerrero-Tamayo, Isabel Olivares, María-Dolores Moragues Tosantos, Concepción Casado.

Methodology: Ana Guerrero-Tamayo, Isabel Olivares, María-Dolores Moragues Tosantos, Iker Pastor-López.

Project administration: Borja Sanz Urquijo, Iker Pastor-López.

Resources: Borja Sanz Urquijo, Isabel Olivares, María-Dolores Moragues Tosantos, Concepción Casado.

Software: Ana Guerrero-Tamayo.

Supervision: Borja Sanz Urquijo, Isabel Olivares, María-Dolores Moragues Tosantos, Concepción Casado, Iker Pastor-López.

Validation: Ana Guerrero-Tamayo, Borja Sanz Urquijo, Isabel Olivares, María-Dolores Moragues Tosantos, Concepción Casado, Iker Pastor-López.

Visualization: Ana Guerrero-Tamayo.

Writing – original draft: Ana Guerrero-Tamayo.

Writing – review & editing: Ana Guerrero-Tamayo, Borja Sanz Urquijo, Isabel Olivares, María-Dolores Moragues Tosantos, Concepción Casado, Iker Pastor-López.

References

1. Manrubia SC, Lázaro E. 2006. Viral evolution. Physics of Life Reviews, vol. 3, no. 2, pp. 65–92. https://doi.org/10.1016/j.plrev.2005.11.002
2. Domingo E, Holland JJ. 1997. RNA virus mutations and fitness for survival. Annual Review of Microbiology, vol. 51, no. 1, pp. 151–178. https://doi.org/10.1146/annurev.micro.51.1.151 PMID: 9343347
3. Sanjuán R, Domingo-Calap P. 2016. Mechanisms of viral mutation. Cellular and Molecular Life Sciences, vol. 73, pp. 4433–4448. https://doi.org/10.1007/s00018-016-2299-6 PMID: 27392606
4. Su S et al. 2016. Epidemiology, genetic recombination, and pathogenesis of coronaviruses. Trends in Microbiology, vol. 24, no. 6, pp. 490–502. https://doi.org/10.1016/j.tim.2016.03.003 PMID: 27012512
5. Herrewegh AAPM, Smeenk I, Horzinek MC, Rottier PJM, de Groot RJ. 1998. Feline Coronavirus Type II strains 79-1683 and 79-1146 originate from a double recombination between Feline Coronavirus Type I and Canine Coronavirus. Journal of Virology, vol. 72, no. 5, pp. 4508–4514. https://doi.org/10.1128/jvi.72.5.4508-4514.1998 PMID: 9557750
6. Muslin C, Mac Kain A, Bessaud M, Blondel B, Delpeyroux F. 2019. Recombination in enteroviruses, a multi-step modular evolutionary process. Viruses, vol. 11, no. 9, p. 859. https://doi.org/10.3390/v11090859 PMID: 31540135

7. Amoutzias GD, Nikolaidis M, Tryfonopoulou E, Chlichlia K, Markoulatos P, Oliver SG. 2022. The remarkable evolutionary plasticity of coronaviruses by mutation and recombination: insights for the COVID-19 pandemic and the future evolutionary paths of SARS-CoV-2. Viruses, vol. 14, no. 1, p. 78. https://doi.org/10.3390/v14010078 PMID: 35062282
8. Banner LR, Keck JG, Lai MC. 1990. A clustering of RNA recombination sites adjacent to a hypervariable region of the peplomer gene of murine coronavirus. Virology, vol. 175, no. 2, pp. 548–555. https://doi.org/10.1016/0042-6822(90)90439-X PMID: 2158184
9. Onafuwa-Nuga A, Telesnitsky A. 2009. The remarkable frequency of Human Immunodeficiency Virus Type 1 genetic recombination. Microbiology and Molecular Biology Reviews, vol. 73, no. 3, pp. 451–480. https://doi.org/10.1128/MMBR.00012-09 PMID: 19721086
10. Njai HF, Gali Y, Vanham G, Clybergh C, Jennes W, Vidal N, et al. 2006. The predominance of Human Immunodeficiency Virus Type 1 (HIV-1) circulating recombinant form 02 (CRF02_AG) in West Central Africa may be related to its replicative fitness. Retrovirology, vol. 3, no. 1, pp. 1–11. https://doi.org/10.1186/1742-4690-3-40 PMID: 16817969
11. Turk G, Carobene M, Monczor A, Rubio AE, Gómez-Carrillo M, Salomón H. 2006. Higher transactivation activity associated with LTR and Tat elements from HIV-1 BF intersubtype recombinant variants. Retrovirology, vol. 3, no. 1, pp. 1–12. https://doi.org/10.1186/1742-4690-3-14 PMID: 16483381
12. Combelas N., Holmblat B., Joffret M.-L., Colbère-Garapin F., and Delpeyroux F., "Recombination between poliovirus and coxsackie A viruses of species C: a model of viral genetic plasticity and emergence," *Viruses*, vol. 3, no. 8, pp. 1460–1484, 2011. https://doi.org/10.3390/v3081460 PMID: 21994791
13. Poonsin P, Wiwatviswakorn V, Chansaenroj J, Poovorawan Y, Plewbang C, Techangamsuwan S. 2023. Canine respiratory coronavirus in Thailand undergoes mutation and evidences a potential putative parent for genetic recombination. Microbiology Spectrum, e02268–23. https://doi.org/10.1128/spectrum.02268-23 PMID: 37707446
14. Tekes G, Hofmann-Lehmann R, Bank-Wolf B, Maier R, Thiel HJ, Thiel V. 2010. Chimeric Feline Coronaviruses that encode Type II Spike protein on Type I genetic background display accelerated viral growth and altered receptor usage. Journal of Virology, vol. 84, no. 3, pp. 1326–1333. https://doi.org/10.1128/JVI.01568-09 PMID: 19906918
15. Gordon K, Omar S, Nofemela A, Bandawe G, Williamson C, Woodman Z. 2016. A recombinant variant with increased envelope entry efficiency emerged during early infection of an HIV-1 subtype C dual infected rapid progressor. AIDS Research and Human Retroviruses, vol. 32, no. 3, pp. 303–310.
16. Vlasova AN, Diaz A, Damtie D, Xiu L, Toh TH, Lee JSY, et al. 2022. Novel canine coronavirus isolated from a hospitalized patient with pneumonia in East Malaysia. Clinical Infectious Diseases, vol. 74, no. 3, pp. 446–454. https://doi.org/10.1093/cid/ciab456 PMID: 34013321
17. Cuervo NS, Guillot S, Romanenkova N, Combiescu M, Aubert-Combiescu A, Seghier M, et al. 2001. Genomic features of intertypic recombinant Sabin poliovirus strains excreted by primary vaccinees. Journal of Virology, vol. 75, no. 13, pp. 5740–5751. https://doi.org/10.1128/JVI.75.13.5740-5751.2001 PMID: 11390576
18. Jegouic S, Joffret ML, Blanchard C, Riquet FB, Perret C, Pallansch I, et al. 2009. Recombination between polioviruses and co-circulating Coxsackie A viruses: role in the emergence of pathogenic vaccine-derived polioviruses. PLoS Pathogens, vol. 5, no. 5, p. e1000412. https://doi.org/10.1371/journal.ppat.1000412 PMID: 19412342
19. Hon CC, Lam TY, Shi ZL, Drummond AJ, Yip CW, Zeng F, et al. 2008. Evidence of the recombinant origin of a bat severe acute respiratory syndrome (SARS)-like coronavirus and its implications on the direct ancestor of SARS coronavirus. Journal of Virology, vol. 82, no. 4, pp. 1819–1826. https://doi.org/10.1128/JVI.01926-07 PMID: 18057240
20. Sabir JS M et al. 2016. Co-circulation of three camel coronavirus species and recombination of MERS-CoVs in Saudi Arabia. Science, vol. 351, no. 6268, pp. 81–84. https://doi.org/10.1126/science.aac8608 PMID: 26678874
21. Hu B, Ge X, Wang LF, Shi ZL. 2015. Bat origin of human coronaviruses. Virology Journal, vol. 12, no. 1, pp. 1–10. https://doi.org/10.1186/s12985-015-0422-1 PMID: 26689940
22. Atzrodt CL, Maknojia I, McCarthy RDP, Oldfield TM, Po J, Ta KTL, et al. 2020. A Guide to COVID-19: a global pandemic caused by the novel coronavirus SARS-CoV-2. The FEBS journal, vol. 287, no. 17, pp. 3633–3650, 2020. Wiley Online Library. https://doi.org/10.1111/febs.15375 PMID: 32446285
23. Marks KM, Gulick RM. 2023. COVID-19. Annals of Internal Medicine, vol. 176, no. 10, pp. ITC145–ITC160. https://doi.org/10.7326/AITC202310170 PMID: 37812774
24. World Health Organization. WHO COVID-19 Dashboard. 2020. Available online: https://covid19.who.int/ (Accessed on September 11, 2023).

25. Yu X, Li N. 2021. Understanding the beginning of a pandemic: China's response to the emergence of COVID-19. Journal of Infection and Public Health, vol. 14, no. 3, pp. 347–352. https://doi.org/10.1016/j.jiph.2020.12.024 PMID: 33618279

26. Davies NG, Jarvis CI, Edmunds WJ, Jewell NP, Diaz-Ordaz K, Keogh RH. 2021. Increased mortality in community-tested cases of SARS-CoV-2 lineage B.1.1.7. Nature, vol. 593, no. 7858, pp. 270–274. https://doi.org/10.1038/s41586-021-03426-1 PMID: 33723411

27. Twohig KA et al. 2022. Hospital admission and emergency care attendance risk for SARS-CoV-2 Delta (B.1.617.2) compared with Alpha (B.1.1.7) variants of concern: a cohort study. The Lancet Infectious Diseases, vol. 22, no. 1, pp. 35–42. https://doi.org/10.1016/S1473-3099(21)00475-8 PMID: 34461056

28. Zali A, Khodadoost M, Gholamzadeh S, Janbazi S, Piri H, Taraghikhah N, et al. 2022. Mortality among hospitalized COVID-19 patients during surges of SARS-CoV-2 Alpha (B. 1.1. 7) and Delta (B. 1.617. 2) variants. Scientific Reports, vol. 12, no. 1, p. 18918. https://doi.org/10.1038/s41598-022-23312-8 PMID: 36344540

29. Markov PV, Ghafari M, Beer M, Lythgoe K, Simmonds P, Stilianakis NI, et al. 2023. The evolution of SARS-CoV-2. Nature Reviews Microbiology, vol. 21, no. 6, pp. 361–379. https://doi.org/10.1038/s41579-023-00878-2 PMID: 37020110

30. Araf Y, Akter F, Tang YD, Fatemi R, Parvez MSA, Zheng C, et al. 2022. Omicron variant of SARS-CoV-2: genomics, transmissibility, and responses to current COVID-19 vaccines. Journal of Medical Virology, vol. 94, no. 5, pp. 1825–1832. https://doi.org/10.1002/jmv.27588 PMID: 35023191

31. Saksena NK, Reddy SB, Miranda-Saksena M, Cardoso THS, Silva EMA, Ferreira JC, et al. 2023. SARS-CoV-2 variants, its recombinants and epigenomic exploitation of host defenses. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, p. 166836. https://doi.org/10.1016/j.bbadis.2023.166836 PMID: 37549720

32. Wang Y, Long Y, Wang F, Li C, Liu W. 2023. Characterization of SARS-CoV-2 recombinants and emerging Omicron sublineages. International Journal of Medical Sciences, vol. 20, no. 1, pp. 151. https://doi.org/10.7150/ijms.79116 PMID: 36619228

33. Wang L, Gao GF. 2022. The "Wolf" is indeed coming: recombinant "Deltacron" SARS-CoV-2 detected. China CDC Weekly, vol. 4, no. 14, pp. 285. https://doi.org/10.46234/ccdcw2022.054 PMID: 35433090

34. Piret J, Boivin G. 2021. Pandemics throughout history. Frontiers in Microbiology, vol. 11, p. 631736. https://doi.org/10.3389/fmicb.2020.631736 PMID: 33584597

35. Adrion JR, Galloway JG, Kern AD. 2020. Predicting the landscape of recombination using Deep Learning. Molecular Biology and Evolution, vol. 37, no. 6, pp. 1790–1808. https://doi.org/10.1093/molbev/msaa038 PMID: 32077950

36. Li Y, Chen S, Rapakoulia T, Kuwahara H, Yip KY, Gao X. 2022. Deep learning identifies and quantifies recombination hotspot determinants. Bioinformatics, vol. 38, no. 10, pp. 2683–2691. https://doi.org/10.1093/bioinformatics/btac234 PMID: 35561158

37. Ajagbe SA, Adigun MO. 2023. Deep learning techniques for detection and prediction of pandemic diseases: a systematic literature review. Multimedia Tools and Applications, pp. 1–35. https://doi.org/10.1007/s11042-023-15805-z PMID: 37362693

38. Choi JY, Smith DM. 2021. SARS-CoV-2 variants of concern. Yonsei Medical Journal, vol. 62, no. 11, pp. 961. Yonsei University College of Medicine. https://doi.org/10.3349/ymj.2021.62.11.961 PMID: 34672129

39. National Library of Medicine (US). National Center for Biotechnology Information. 2023. NCBI Virus. [Online]. Available: https://www.ncbi.nlm.nih.gov/labs/virus/vssi/. [Accessed: Sep. 11, 2023].

40. GISAID Initiative. 2023. GISAID-Tracking of hCoV-19 Variants. [Online]. Available: https://gisaid.org/hcov19-variants/. [Accessed: November 21, 2023].

41. Rambaut A et al. 2020. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nature Microbiology, vol. 5, no. 11, pp. 1403–1407. Nature Publishing Group UK London. https://doi.org/10.1038/s41564-020-0770-5 PMID: 32669681

42. Carabelli AM et al. 2023. SARS-CoV-2 variant biology: immune escape, transmission and fitness. Nature Reviews Microbiology, vol. 21, no. 3, pp. 162–177. Nature Publishing Group UK London. https://doi.org/10.1038/s41579-022-00841-7 PMID: 36653446

43. Yu J, Ai M, Ye Z. 2023. A review on design inspired subsampling for big data. Statistical Papers, pp. 1–44. Springer.

44. Goodfellow I, Bengio Y, Courville A. 2016. Deep Learning. Cambridge, MA: MIT Press.

45. Guerrero-Tamayo A, Sanz Urquijo B, Casado C, Moragues-Tosantos MD, Olivares I, Pastor-López I. 2023. Discovering Mathematical Patterns Behind HIV-1 Genetic Recombination: a new methodology to identify viral features. IEEE Access. https://doi.org/10.1109/ACCESS.2023.3311752

46. Tammina S. 2019. ransfer learning using VGG-16 with deep convolutional neural network for classifying images. International Journal of Scientific and Research Publications (IJSRP), vol. 9, no. 10, pp. 143–150. https://doi.org/10.29322/IJSRP.9.10.2019.p9420
47. Kornblith S, Shlens J, Le QV. 2019. Do better ImageNet models transfer better?. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2661–2671.
48. Proakis JG, Manolakis DG. 1988. Introduction to Digital Signal Processing. Prentice Hall Professional Technical Reference.
49. Sussillo D, Kundaje A, Anastassiou D. 2004. Spectrogram analysis of genomes. EURASIP Journal on Advances in Signal Processing, vol. 2004, pp. 1–14. https://doi.org/10.1155/S1110865704310048
50. Howley PM, Knipe DM, Whelan S, Freed EO. 2022. Fields Virology: RNA Viruses. Lippincott Williams & Wilkins.
51. Huang J, Ling CX. 2005. Using AUC and accuracy in evaluating learning algorithms. IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 3, pp. 299–310. https://doi.org/10.1109/TKDE.2005.50
52. Marmolejo-Saucedo JA, Kose U. 2022. Numerical grad-CAM based explainable convolutional neural network for brain tumor diagnosis. Mobile Networks and Applications, pp. 1–10.
53. Hamilton N et al. 2022. Enhancing visualization and explainability of computer vision models with Local Interpretable Model-Agnostic Explanations (LIME). 2022 IEEE Symposium Series on Computational Intelligence (SSCI), pp. 604–611, IEEE.
54. Wargnier-Dauchelle Vet al. 2023. A weakly supervised gradient attribution constraint for interpretable classification and anomaly detection. IEEE Transactions on Medical Imaging. https://doi.org/10.1109/TMI.2023.3282789 PMID: 37276116
55. Friston KJ, Ashburner J, Frith CD, Poline JB, Heather JD, Frackowiak RSJ. 1995. Spatial registration and normalization of images. Human Brain Mapping, vol. 3, no. 3, pp. 165–189, 1995. https://doi.org/10.1002/hbm.460030303
56. Tan J, Xu L, Li T, Su B, Wu J. 2014. Image-Contrast Technology Based on the Electrochemiluminescence of Porous Silicon and Its Application in Fingerprint Visualization. Angewandte Chemie, vol. 126, no. 37, pp. 9980–9984. https://doi.org/10.1002/ange.201404948 PMID: 25044650
57. Bradski G. 2000. The OpenCV library. Dr. Dobb's Journal of Software Tools, vol. 2000.
58. Mercatelli D, Giorgi FM. 2020. Geographic and genomic distribution of SARS-CoV-2 mutations. Frontiers in Microbiology, vol. 11, p. 1800. https://doi.org/10.3389/fmicb.2020.01800 PMID: 32793182
59. Nikolaidis M, Markoulatos P, Van de Peer Y, Oliver SG, Amoutzias Gd. 2022. The neighborhood of the Spike gene is a hotspot for modular intertypic homologous and nonhomologous recombination in coronavirus genomes. Molecular Biology and Evolution, vol. 39, no. 1, p. msab292. https://doi.org/10.1093/molbev/msab292 PMID: 34638137
60. Jackson Bet al. 2021. Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic. Cell, vol. 184, no. 20, pp. 5179–5188. https://doi.org/10.1016/j.cell.2021.08.014 PMID: 34499854
61. Turakhia Yet al. 2022. Pandemic-scale phylogenomics reveals the SARS-CoV-2 recombination landscape. Nature, vol. 609, no. 7929, pp. 994–997. https://doi.org/10.1038/s41586-022-05189-9 PMID: 35952714
62. LI C, Li X, Chen M, Sun X. 2023. Deep learning and image recognition. 2023 IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT), pp. 557–562.
63. Cai L, Gao J, Zhao D. 2020. A review of the application of deep learning in medical image classification and segmentation. Annals of Translational Medicine, vol. 8, no. 11. https://doi.org/10.21037/atm.2020.02.44 PMID: 32617333

Deusto

Ana Guerrero Tamayo <ana.guerrero@deusto.es>

HAIS 2024: Submit source files by next 22nd July 2024

HAIS 2024 <hais2024@easychair.org> 19 de julio de 2024, 16:37
Para: Ana Guerrero-Tamayo <ana.guerrero@deusto.es>

Dear Ana Guerrero-Tamayo,

You submitted the following paper to be considered for publication at the HAIS 2024 Conference:

Paper ID: 1862
Title: Unveiling HIV-1 U Sequences: Shedding Light Through Transfer Learning on Genomic Spectrograms

We are pleased to inform you that your paper has been finally accepted, as you successfully addressed the reviewers' comments.

The next step is to upload the source files to Easychair so that the paper can be included in the conference Proceedings.
To do that, please STRICTLY FOLLOW THESE INSTRUCTIONS by July 22nd:

https://haisconference.eu/source-files

The source files submission deadline is 22nd July 2024.
The payment deadline is 25th July 2024.

Best regards,
HAIS 2024 Scientific Programme Committee

Unveiling HIV-1 U Sequences: Shedding Light Through Transfer Learning on Genomic Spectrograms*

Ana Guerrero-Tamayo[1][0000-0002-1828-360X], Borja Sanz Urquijo[1][0000-0003-2039-7773], Isabel Olivares[2][0000-0002-2347-5717], María-Dolores Moragues Tosantos[3][0000-0002-7272-8810], Concepción Casado[2][0000-0003-3412-2877], and Iker Pastor-López[1][0000-0002-3068-6248]

[1] Faculty of Engineering - University of Deusto, Avenida de las Universidades, 24, Bilbao, 48007 Biscay, Spain
[2] National Microbiology Center (NMC) - Instituto de Salud Carlos III (ISCIII), Ctra. de Pozuelo, 28, Majadahonda, 28222, Madrid, Spain
[3] Faculty of Medicine and Nursing - University of the Basque Country UPV/EHU, Barrio Sarriena s/n, Leioa, 48940, Biscay, Spain

Abstract. The human immunodeficiency virus type 1 (HIV-1) has caused one of the most pernicious pandemics in recent history. The current difficulty with effective treatments and vaccines is largely attributed to its high mutation rate and multiple complex genetic recombination events, resulting in various subtypes. Some fragments and complete sequences do not match any subtype (referred to as U, unclassified or unclassifiable). Further understanding of these sequences and fragments may provide insights into the phylogeny of HIV-1 and its viral evolution. By applying transfer learning to a pre-trained CNN (VGG-16) using all the genomic spectrograms of all complete HIV-1 genomic sequences, we achieved a total hit rate of 98.27%, with the CNN primarily focusing on detecting the recombinant features, mainly in the 5'LTR and 3'LTR regions. We also generated a phylogenetic tree of all complete HIV-1 sequences. Subsequently, we compared subtype U interpretability patterns with phylogenetic distance, yielding valuable information not only about subtype U but also about intersubtype parentage and their common and distinct genomic components. Interpretability patterns coincided with multiple microbiological evidences in 60.870% of the sequences, were inconclusive in 34.783%, and only one controversial sequence was detected (4.348%). Compared to multiple sequence alignment algorithms, our methodology is faster, requires low computing power, and does not necessitate manual adjustments.

Keywords: HIV-1 · Genetic Recombination · U Sequences · Deep Learning · Convolutional Neural Network · Interpretability Patterns · Mathematical Patterns

* University of Deusto - Research Training Grants Program.

1 Introduction

Human immunodeficiency virus type 1 (HIV-1) is responsible of one of the worst pandemics in recent story. Its strong variability due to its high rate of mutations and genetic recombination phenomena has a decisive influence on the lack of effective treatments and vaccines [24, 26]. This changeability results in multiple subtypes and classifying a sequence in any of them can be complex and controversial [3, 21].

HIV-1 is classified into four groups named M, N, O and P. Group M, the most globally widespread, is divided into nine pure subtypes (non-recombinant) designed as A, B, C, D, F, G, H, J and K. Circulating recombinant forms (CRFs) are those recombinant subtypes currently circulating [2], whose recombination points are common and perfectly described. These CRF sequences must be identified in at least three epidemiologically non-related individuals [20]. The nomenclature of each CRF subtype refers to the recombined subtypes involved. CRF02_AG means genetic recombination between subtypes A and G [12]. Complex recombination phenomena (cpx) is the designation of mosaic recombinants of multiple subtypes. CRF04_cpx is a mixture between A, G, H, K and U [18]. Non-CRF recombinant nomenclature refers to recombinated subtypes also. Non-CRF recombinant subtype A1GH, means a complex recombination phenomena between subtype A1, G and H [15].

This tricky HIV-1 subtype classification makes necessary to review nomenclatures regularly by expert committees. For instance, firstly named E subtype sequences, were finally renamed as CRF01_AE (recombinant between A and E subtypes) [4]. And formerly called F3 subtype was reissued as K subtype [8]. Some authors consider that CRF02_AG should be a non-recombinant subtype and G subtype as a recombinant of A, J and 02_AG [11].

Some fragments and complete sequences do not match with any subtype. These are U subtype (unclassified). If a recombinant contains unclassified (or unclassifiable) regions, U is included in the name. There are two general circumstances for the assignment of "U". One is when the sequence is so divergent from previously-classified sequences that a subtype cannot be assigned based on current reference sequences. The second situation is when the sequence is of such poor quality that the subtype cannot be determined with any reasonable certainty. There are 23 U subtype complete sequences. And about ten 10 CRFs subtypes contain U fragments. Further knowledge of these sequences and fragments may provide insight into the phylogeny of HIV-1 and its viral evolution.

This article presents the results of an in-depth analysis of the previous work conducted by Guerrero-Tamayo et al. [10], on the most mysterious HIV-1 U sequences, using a methodology based on genomic spectrograms, Deep Learning technologies, and interpretability with the aim not only of demonstrating the robustness and biological coherence of the obtained interpretability patterns, but also of contributing knowledge about the origin and phylogeny of these U sequences based on the interpretability results.

2 Materials and Methods

All experiments were ran in this equipment: Processing Unit: Intel(R) Xeon(R) Gold 5220R CPU @ 2.20 GHz. Installed RAM: 256 GB usable. Operative System: Windows 10 Education. Version: 21H1. GPU: NVIDIA RTX A6000. Total memory: 179451 MB. VRAM: 48571 MB.

Our methodology consisted of three phases described in Figure 1.

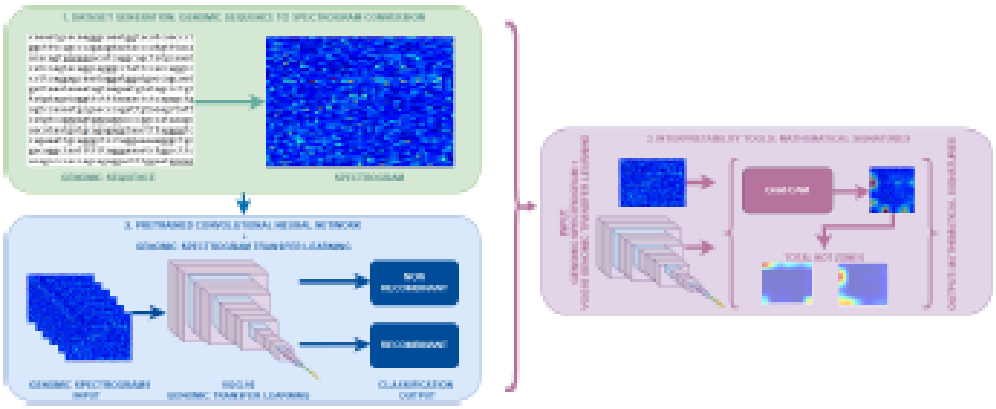


Fig. 1: Methodology

The first phase is a conversion of genomic sequences to biologically-coherent images, this is, genomic spectrograms. The x-axis corresponds to the concept of nucleotide position, so the total length of the x-axis is the length of the genomic sequence. A complete HIV-1 sequence contains approximately 10,000 nucleotides, making this the length of the x-axis. The y-axis represents the frequency range, in this case, from 0 to 0.5 Hz. The z-axis is the sum of the Fast Fourier Transform (FFT) values for each of the four types of nucleotides (A, C, G, and T). This means:

$$S = S_a + S_g + S_c + S_t \tag{1}$$

In order to work with two-dimensional spectrograms, the z-axis is assimilated to a "jet" color palette [16].

HIV-1 genome sequences were downloaded from Los Alamos National Laboratory HIV Database (all subtypes - complete genome) [15]. We generated the spectrogram dataset using Python, Scipy library, Spectrogram function [23]. In order to avoid biases and bad practices during the CNN training, we omitted axes, margins and any other elements that may distort the operation of the network. So, the image is limited to the spectrogram [5, 6].

To build the dataset, we marked all CRF sequences as recombinant and non-CRF sequences as non-recombinant. We then randomly assigned these sequences

to the training, validation, and test folders based on the guidelines specified in Table 1.

Table 1: Dataset Structure

Dataset	Recombinant	Non-Recombinant	TOTAL
Training	1,903 (16.89%)	9,364 (83.11%)	11,267
Validation	338 (17.01%)	1,649 (82.99%)	1,987
Test	150 (50.00%)	150 (50.00%)	300
TOTAL	**2,391**	**11,163**	**13,554**

As illustrated in Table 1, the test set distribution (50% / 50%) is intentionally different from that of the training and validation sets (17% / 83%) to identify underfitting biases during the training phases [9].

Once the spectrogram dataset was generated, we applied transfer learning to a pre-trained VGG-16 using MATLAB 2021b, which we named HIV-1_VGG-16. As performance measures for HIV-1_VGG-16, we used the metrics Training Time, AUC, Validation and Test Accuracy, and Confusion Matrix, not only as overall performance measures but also by category.

Then, we applied Grad-CAM to assess the most relevant areas in recombinant featuring of an HIV-1 sequence. Interpretability tools allow us to identify where the CNN looks to decide the recombinant feature of a sequence (hot zones). Grad-CAM is our tool of choice for visual comparison of interpretability patterns due to its easiest adaptation to human eye.

We calculated global interpretability patterns of non-recombinant and recombinant sequences as an arithmetic summation of the score maps (numerical vectors) using MATLAB 2021b.

To generate the phylogenetic tree, we downloaded the multiple alignment of all HIV-1 sequences up to 2020 from the Los Alamos Database. This multiple alignment includes a single sequence per patient and eliminates very similar sequences and problematic ones [15]. We generated the phylogenetic tree from this multiple alignment using MATLAB Bioinformatics Toolbox. On the basis of the phylogenetic position of all U subtype sequences in the HIV-1 complete sequences tree, we studied their closeness relationships and their interpretability patterns. This analysis was complemented with data from samples registered in the Los Alamos Sequence Database [15].

Our methodology enables deepen knowledge of HIV-1 genetic recombination phenomena. It also provides a way to characterize the virus through its interpretability patterns. By analyzing the results obtained within the U subtype, we were able to extract additional information from the perspective of recombinant featuring. Understanding HIV-1 genetic recombination can contribute to understanding HIV-1 viral evolution itself.

A better knowledge of both complete U sequences and U fragments detected in multiple CRFs will allow us to establish more precisely how the viral evolution

of HIV-1 has been and how it is. The use of Deep Learning tools for this purpose may provide new approaches to elucidate knowledge.

3 Results and Discussion

3.1 Global Results of the Experiment

After testing multiple hyperparameter configurations on a pre-trained VGG-16, Table 2 shows the hyperparameters for the optimal configuration, as well as their results across all performance metrics. The top row in Confusion Matrix Results corresponds to the non-recombinant category (hit - error) and the bottom row to the recombinant category (error - hit).

Table 2: Performance Metrics Results

CNN	Hyperparameters	Validation Accuracy	Training Time	Training Time per Epoch	Test Accuracy	Confusion Matrix Results	AUC
VGG-16	Learning Rate: 0.0001 Batchsize: 32 Epochs: 10	92.90%	15 min 89 sec	1 min 38.90 sec	94.00%	143 – 7 11 – 139	0.9888

Multiple alignment techniques and algorithms are statistical approximations that require complex expert manual adjustments [7]. The execution time of these algorithms increases quadratically depending on the number of sequences and their average length.

For comparative purposes, we conducted multiple tests using traditional multiple sequence alignment methods, employing the most common algorithms (ClustalW and MUSCLE) on the same equipment. Due to computational limitations, it was not possible to complete any of the tests conducted, highlighting the effectiveness of our methodology in terms of computing efficiency and resource management for handling the 13,554 complete genomes of HIV-1 available.

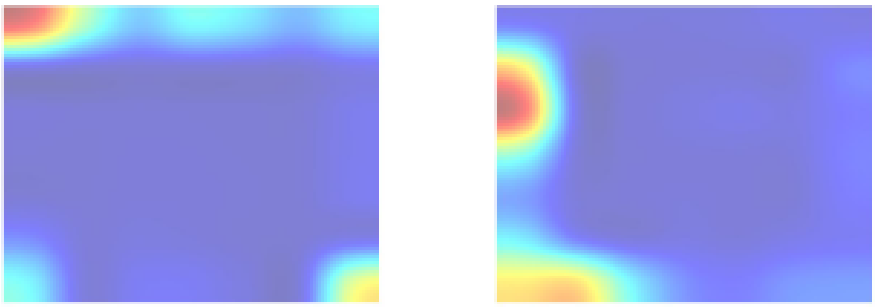
HIV-1_VGG-16 hit rate detecting recombinant feature reached 98.27% on the total of the 13,554 complete HIV-1 sequences (Table 3).

This CNN successfully classified 10,993 HIV-1 complete sequences of the total of 11,163 non-recombinant sequences (hit rate of 98.48%). Of the total of 2,391 complete recombinant sequences, HIV-1_VGG-16 correctly identified 2,326 as recombinants (hit rate of 97.28%). We must take into account that it previously learned with sequences of training and validation sets, so these have higher hit rate than test set.

Table 3: HIV-1_VGG-16 Error Rates

	Non-recombinant		Recombinant	
	No.	%	No.	%
Errors	170	1.52	65	2.72%
Successes	10,993	98.48	2,326	97.28%
TOTAL	**11,163**	**100.00%**	**2,391**	**100.00%**

Figure 2 shows interpretability patterns for the two categories. Figure 2a is the result of the mathematical superposition of the 11,163 non-recombinant patterns and Figure 2b is the recombinant global pattern of the 2,391 recombinant sequences. As shown in Figure 2, the total hot zones are perfectly defined and distinguishable.



(a) Non-recombinant Total Hot Zones (b) Recombinant Total Hot Zones

Fig. 2: Total Interpretability Patterns by Category

Considering subtypes with over a 15% error rate and containing more than 15 complete sequences, we found that "U sequences" has the highest error rate (39.13%), compared to the second subtype with the highest error rate at 25.25%. The high error rate of U sequences suggests that there may be recombination phenomena that can help, not only to propose a review of this subtype sequences, but also to extract similarities through their interpretability patterns by low-computing algorithms. We analyzed in depth interpretability patterns related with recombinant feature of each U subtype sequence.

Our results were biologically consistent and robust. We obtained not only valuable information about U sequences, but we acquired knowledge of inter-subtype parentage and their common and distinct genomic components. All this with a fast, efficient and low computing methodology. Following subsections

describe each level of homology between interpretability patterns and biological evidences.

3.2 Conclusive Biological Coherence

The cases of high homology among the classification results of HIV-1_VGG-16, the interpretability-driven hot zones, and the phylogeny, along with expert literature in the microbiological and virological field, correspond to EF029066, EF029067, EF029068, EF029069, FJ388921, HM215249, HM215251, JF683772, JX140679, KY392775, MN153489, MN153492, MN736701 and MN736702.

Below, we provide a more detailed presentation of the most representative cases.

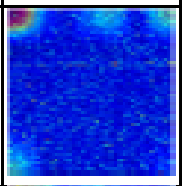
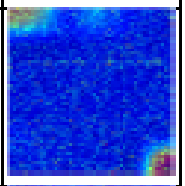
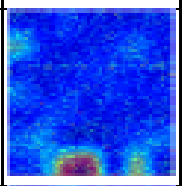
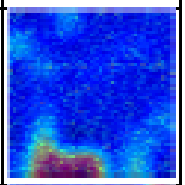
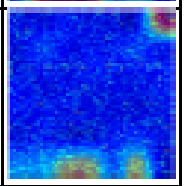
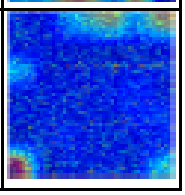
- **HM215249 and HM215251**. HIV-1_VGG-16 classified these sequences as non-recombinant with confidence scores above 99%. Their scores were the maximum of all U sequences. Grad-CAM patterns were almost identical. These U sequences, HM215249 and HM215251, were the unique components of their clade. Quesnel-Vallières et al. considered that HIV-1 isolates infecting patients TV721 (HM215249) and TV749 (HM215251) cannot be considered CRFs. Our results confirmed this point [19]. Therefore, there is a clear logic in our results and the virological evidence.

- **JF683772**. Subtype 01G MN486045 was the only sister of U sequence JF683772 (Patient Code: CY223 - unique sequence). JF683772 was classified as recombinant by HIV-1_VGG-16 (trust level over 75%). JF683772 patterns were described similar to subtypes CRF02_AG and CRF63_02A6 (recombination between CRF02_AG and A6) [1]. This fact endorsed JF683772 recombinant feature. Some studies pointed the recombinant feature of this sequence [13], suggesting a recombination between A1, G, and H [22]. Interpretability patterns were consistent with this.

- **EF029066, EF029067, EF029068 and EF029069 and their related sequences FJ388921 and KY392775**. This case was particularly interesting because we detected a possible phenomenon of hidden recombination in the intrapatient sequences.

FJ388921 (Patient Code: CY090) [14], EF029066 (Patient Code: H10986) and KY392775 (patient code = NGSID12, intrapatient with partial sequences KY365200 and KY365214) [22] were three U direct relatives. Table 4 shows classification results. According to HIV-1_VGG-16, FJ388921 had a strong non-recombinant component, with standard interpretability patterns. EF029066 (classified as non-recombinant) had the lowest confidence score of all U subtype (under 52%). And HIV-1_VGG-16 classified its intrapatient sequences (EF029067, EF029068, EF029069) as recombinant, with medium confidence levels, around 70-75%. EF029066 interpretability patterns were coherent with global non - recombinant hot zones. EF029067, EF029068 and EF029069 had close determinant areas, fairly diffuse in EF029067 and EF029068. Their patterns did not particularly match global recombinant hot zones. KY392775 was classified as non-recombinant just under 90%. However, interpretability patterns were slightly controversial with non-recombinant ones.

Table 4: U Sequences FJ388921, EF029066, EF029067, EF029068, EF029069 and KY392775 Results

Sequence ID	Classification	Confidence Score	Grad-CAM
FJ388921	Non-recombinant	91.853%	
EF029066	Non-recombinant	51.424%	
EF029067	Recombinant	76.708%	
EF029068	Recombinant	69.344%	
EF029069	Recombinant	73.163%	
KY392775	Non-recombinant	89.344%	

FJ388921, EF029066 and KY392775 interpretability patterns were similar. This correlates with their phylogenetic positions. These three sequences shared clade with subtype FK MH705156. AJ249235 and AJ249239 are subtype K.

EF029066, EF029067, EF029068 and EF029069 are intrapatient sequences (Patient ID H10986). HIV-1_VGG-16 classified EF029066 as non-recombinant, but its reliability level was minimum (near 51.50%). EF029067, EF029068 and EF029069 were classified as recombinant. Their interpretability patterns were highly similar and divergent from EF029066. The intrapatient feature of these four sequences unabled to analyze deeply their phylogenetic relationship. But

according to the results obtained, there was a clear recombination phenomenon. EF029067, EF029068 and EF029069 decision areas were similar to several sequences of different recombinant subtypes. This is coherent with HIV-1_VGG-16 classification. Most similar hot areas corresponded to those of CRF26_A5U FM877782, and to a lesser extent to CRF02_AG AB286862, subtype CRF11_cpx AJ291718 and CRF13_cpx DQ845386. Their confidence scores were higher than 90%, except for CRF11_cpx AJ291718 (around 52%). The striking similarity level between these different subtypes suggests the existence of genomic components common to several different recombinant lineages, related with these three U subtype sequences.

Another point that requires further research is the possible relationship between low confidence scores and similarity between interpretability patterns (case of AJ291718).

Van Der Hoek et al. cataloged EF029066 and its intrapatient U sequences EF029067, EF029068 and EF029069 as a new subtype, or a new branch of subtype K [25]. This supports EF029066 classification as non-recombinant. EF029066 low reliability level and the similar recombinant hot zones of EF029067, EF029068 and EF029069 make it necessary to search for recombination events in these sequences.

A sister clade including subtype F2 sequences and the resemblance between EF029066 and these F2 sequences interpretability patterns suggest certain common features between them. Subtype F2 sequences KU749420, and especially AJ249237 and AY371158 had strongly similar hot zones to EF029066. This endorses the non-recombinant classification of EF029066 by HIV-1_VGG-16. However, its low confidence-score and recombinant sister sequences made us consider the presence of recombinant factors embedded.

3.3 Non-Conclusive Biological Coherence

Analyzing the case of U sequences AB485664, AB485665, KP718920, KY392773, KY392774, KY392780, MH078559, and MK095236, homology among the classification results of HIV-1_VGG-16, the interpretability-driven hot zones, and the phylogeny, along with expert literature in the microbiological and virological fields, was rather inconclusive. Sequence AY046058 was the most controversial case.

Next, we provide a detailed analysis of the most significant cases.

- **AY046058, KY392780, MH078559, and MK095236**. MK095236 was classified as possibly recombinant with some uncertainty, while AY046058 and MH078559 were confidently non-recombinant. KY392780 showed similarity to both non - recombinant and recombinant sequences but with low confidence. These U sequences exhibited unique patterns, potentially indicating a shared element responsible for their non-recombinant status despite resemblances to recombinant sequences. Further research is needed to elucidate this common factor. Additionally, there are intriguing relationships between certain CRF sequences and U sequences, suggesting potential ancestral connections. The classification of AY046058 remains contentious due to its documented mixed classification [17].

- **KP718920**. Sequence KP718920 (patient code 469_66 - unique sequence) was labeled non-recombinant due to matching interpretability patterns with global non-recombinant sequences, particularly subtype F2 and A1. Most related sequences are subtype CRF11_cpx, except one (KR017778, subtype 11A1). KP718920's structure indicates a mix of CRF36_cpx, CRF11_cpx, CRF02_AG, and U. Its association with CRF11_cpx and CRF13_cpx suggests a potential for non-recombinant classification, as observed with similar sequences like CRF11_cpx KP718929, JF683802, and CRF13_cpx AY371154, hinting at hidden non-recombinant-like traits in their genomes. Further research is required to validate this hypothesis.

4 Conclusions

HIV-1_VGG-16 classifications as recombinant or non-recombinant and their corresponding patterns were phylogenetically coherent and consistent with sample data. And similarity level correlated with phylogenetic closeness.

Our results were coincident with multiple microbiological evidences, such is the case of EF029066, EF029067, EF029068, EF029069, FJ388921, HM215249, HM215251, JF683772, JX140679, KY392775, MN153489, MN153492, MN736701 and MN736702. Their interpretability patterns were consistent with phylogeny and multiple biological evidences (60.870% of the total U sequences). The 34.783% was rather inconclusive (AB485664, AB485665, KP718920, KY392773, KY392774, KY392780, MH078559 and MK095236). Sequence AY046058 was the most controversial case (4.348%).

We also were able to establish similarities between interpretability patterns of different subtypes whose sequences are phylogenetically close. This can help us to search for common genomic elements in the hot areas pointed by the CNN. Our results also reflected the leading role of genetic recombination in HIV-1 phylogeny.

Many of CRF cpx involved in this research contained subtype G sections. Engagement between subtypes U and G may denote some parentage level between them. And the possible recombinant origin of subtype G could influence U interpretability patterns.

Now we know where does the CNN look at to decide. Our next step is searching what exactly is seeing. MN736702 can significantly clarify this point. This sequence gave us the opportunity to make a direct comparison of both categories in a very specific genomic area (low frequencies of 5'LTR). Perhaps we can unravel those mathematical signatures that define a sequence as recombinant and better understand HIV-1 genetic recombination.

A deeper understanding of both U complete sequences and U fragments detected in multiple CRFs will allow to establish more precisely how HIV-1 viral evolution is. The use of Deep Learning tools for this purpose may provide new approaches to elucidate knowledge.

Our research is a demonstration of this point. Our methodology was alignment-free, much faster and low computing. And it did not need manual adjustments and multiple alignment.

We not only succeeded in demonstrating the robustness of our results in detecting the recombinant feature, but also in establishing homology between the CNN results and interpretability patterns with phylogeny and expert literature, even contributing interesting insights into the phylogenetic relationships of the U subtype, as well as potential hidden recombination events in these sequences.

By applying an alternative methodology, we deepen in the knowledge of U sequences, which are more problematic for classification. And more unknown.

5 Data Availability

The complete HIV-1 phylogenetic tree and the dataset are available at DOI: 10.21227/2917-1s55.

References

1. Baryshev, P.B., Bogachev, V.V., Gashnikova, N.M.: HIV-1 genetic diversity in Russia: CRF63_02A1, a new HIV Type 1 genetic variant spreading in Siberia. AIDS research and human retroviruses **30**(6), 592–597 (2014)
2. Bbosa, N., Kaleebu, P., Ssemwanga, D.: HIV subtype diversity worldwide. Current Opinion in HIV and AIDS **14**(3), 153–160 (2019)
3. Camacho, R.: The significance of subtype-related genetic variability: controversies and unanswered questions. Antiretroviral resistance in clinical practice (2006)
4. Carr, J.K., Salminen, M.O., Koch, C., Gotte, D., Artenstein, A.W., Hegerich, P.A., St Louis, D., Burke, D.S., McCutchan, F.E.: Full-length sequence and mosaic structure of a Human Immunodeficiency Virus Type 1 isolate from Thailand. Journal of Virology **70**(9), 5935–5943 (1996)
5. Catalá, O.D.T., Igual, I.S., Pérez-Benito, F.J., Escrivá, D.M., Castelló, V.O., Llobet, R., Perés-Cortés, J.C.: Bias analysis on public X-ray image datasets of pneumonia and COVID-19 patients. IEEE Access **9**, 42370–42383 (2021)
6. Chen, L., Cruz, A., Ramsey, S., Dickson, C.J., Duca, J.S., Hornak, V., Koes, D.R., Kurtzman, T.: Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. PloS one **14**(8), e0220113 (2019)
7. Gaschen, B., Taylor, J., Yusim, K., Foley, B., Gao, F., Lang, D., Novitsky, V., Haynes, B., Hahn, B.H., Bhattacharya, T., et al.: Diversity considerations in HIV-1 vaccine selection. Science **296**(5577), 2354–2360 (2002)
8. Geretti, A.M.: HIV-1 subtypes: epidemiology and significance for HIV management. Current opinion in infectious diseases **19**(1), 1–7 (2006)
9. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016)
10. Guerrero-Tamayo, A., Urquijo, B.S., Casado, C., Moragues-Tosantos, M.D., Olivares, I., Pastor-López, I.: Discovering mathematical patterns behind HIV-1 genetic recombination: a new methodology to identify viral features. IEEE Access (2023)
11. Hemelaar, J.: The origin and diversity of the HIV-1 pandemic. Trends in molecular medicine **18**(3), 182–192 (2012)

12. Howard, T.M., Rasheed, S.: Genomic structure and nucleotide sequence analysis of a new HIV Type 1 subtype A strain from Nigeria. AIDS research and human retroviruses **12**(15), 1413–1425 (1996)
13. Kousiappa, I., Achilleos, C., Hezka, J., Lazarou, Y., Othonos, K., Demetriades, I., Kostrikis, L.G.: Molecular characterization of HIV type 1 strains from newly diagnosed patients in Cyprus (2007–2009) recovers multiple clades including unique recombinant strains and lack of transmitted drug resistance. AIDS research and human retroviruses **27**(11), 1183–1199 (2011)
14. Kousiappa, I., Van De Vijver, D.A., Kostrikis, L.G.: Near full-length genetic analysis of HIV sequences derived from Cyprus: evidence of a highly polyphyletic and evolving infection. AIDS research and human retroviruses **25**(8), 727–740 (2009)
15. Laboratory, L.A.N.: The HIV database. https://www.hiv.lanl.gov/ (2023), accessed: 2023-06-09
16. Matlab: Matlab colormaps tutorial. https://matplotlib.org/stable/tutorials/colors/colormaps.html (2023), accessed: 2023-06-09
17. Paraskevis, D., Magiorkinis, E., Magiorkinis, G., Anastassopoulou, C., Lazanas, M., Chrysos, G., Vandamme, A.M., Hatzakis, A.: Molecular characterization of a complex, recombinant Human Immunodeficiency Virus Type 1 (HIV-1) isolate (A/G/J/K/?): evidence to support the existence of a novel HIV-1 subtype. Journal of General Virology **82**(10), 2509–2514 (2001)
18. Paraskevis, D., Magiorkinis, M., Vandamme, A.M., Kostrikis, L.G., Hatzakis, A.: Re-analysis of Human Immunodeficiency Virus Type 1 isolates from Cyprus and Greece, initially designated 'subtype I', reveals a unique complex A/C/H/K/? mosaic pattern. Journal of General Virology **82**(3), 575–580 (2001)
19. Quesnel-Vallières, M., Kouzayha, L., Tran, E., Barry, I., Lasgi, C., Merindol, N., Monteil, V., Ransy, D.G., Boucher, M., Lapointe, N., et al.: Novel HIV-1 recombinant forms in antenatal cohort, Montreal, Quebec, Canada. Emerging Infectious Diseases **17**(2), 271 (2011)
20. Rivas, P., Holguín, Á., Ramírez de Arellano, E., Muñoz-Almagro, C., Delgado, R., Ortiz de Lejarazu, R., Soriano, V.: Tratamiento antirretroviral según tipos y subtipos del virus de la inmunodeficiencia humana. Enferm. infecc. microbiol. clín.(Ed. impr.) pp. 29–33 (2006)
21. Robertson, D.L., Anderson, J., Bradac, J., Carr, J., Foley, B., Funkhouser, R., Gao, F., Hahn, B., Kalish, M., Kuiken, C., et al.: HIV-1 nomenclature proposal. Science **288**(5463), 55–55 (2000)
22. Rodgers, M.A., Wilkinson, E., Vallari, A., McArthur, C., Sthreshley, L., Brennan, C.A., Cloherty, G., de Oliveira, T.: Sensitive next-generation sequencing method reveals deep genetic diversity of HIV-1 in the Democratic Republic of the Congo. Journal of virology **91**(6), e01841–16 (2017)
23. SciPy: Scipy v1.11.4 manual. https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.spectrogram.html (2023), accessed: 2023-06-09
24. Smyth, R.P., Davenport, M.P., Mak, J.: The origin of genetic diversity in HIV-1. Virus research **169**(2), 415–429 (2012)
25. Van Der Hoek, L., Pollakis, G., Lukashov, V.V., Jebbink, M.F., Jeeninga, R.E., Bakker, M., Dukers, N., Jurriaans, S., Paxton, W.A., Back, N.K., et al.: Characterization of an HIV-1 group M variant that is distinct from the known subtypes. AIDS research and human retroviruses **23**(3), 466–470 (2007)
26. Williamson, C., Martin, D.P.: HIV-1 genetic diversity. HIV/AIDS in South Africa **2**, 117–126 (2010)



Deusto

Ana Guerrero Tamayo <ana.guerrero@deusto.es>

HAIS 2024: Submit source files by next 22nd July 2024

HAIS 2024 <hais2024@easychair.org> 19 de julio de 2024, 16:37
Para: Ana Guerrero-Tamayo <ana.guerrero@deusto.es>

Dear Ana Guerrero-Tamayo,

You submitted the following paper to be considered for publication at the HAIS 2024 Conference:

Paper ID: 1188
Title: The Third Codon Nucleotide's Role in Genetic Recombination Within SARS-CoV-2 Spike Protein: A Pilot Study

We are pleased to inform you that your paper has been finally accepted, as you successfully addressed the reviewers' comments.

The next step is to upload the source files to Easychair so that the paper can be included in the conference Proceedings.
To do that, please STRICTLY FOLLOW THESE INSTRUCTIONS by July 22nd:

https://haisconference.eu/source-files

The source files submission deadline is 22nd July 2024.
The payment deadline is 25th July 2024.

Best regards,
HAIS 2024 Scientific Programme Committee

The Third Codon Nucleotide's Role in Genetic Recombination Within SARS-CoV-2 Spike Protein: A Pilot Study*

Ana Guerrero-Tamayo[1][0000-0002-1828-350X], Borja Sanz Urquijo[1][0000-0003-2039-7773], Isabel Olivares[2][0000-0002-2347-5717], María-Dolores Moragues Tosantos[3][0000-0002-7272-8810], Concepción Casado[2][0000-0003-3412-2877], and Iker Pastor-López[1][0000-0002-3068-6248]

[1] Faculty of Engineering - University of Deusto, Avenida de las Universidades, 24, Bilbao, 48007 Biscay, Spain
[2] National Microbiology Center (NMC) - Instituto de Salud Carlos III (ISCIII), Ctra. de Pozuelo, 28, Majadahonda, 28222, Madrid, Spain
[3] Faculty of Medicine and Nursing - University of the Basque Country UPV/EHU, Barrio Sarriena s/n, Leioa, 48940, Biscay, Spain

Abstract. Genetic recombination is a key mechanism of viral evolution in coronaviruses. In the case of SARS-CoV-2, the Spike protein (S) is a crucial genomic region where multiple recombination events have occurred. Previous research identified mathematical signatures associated with the recombinant feature in HIV-1. Based on evidence from experiments on SARS-CoV-2 unrelated to the present study, we hypothesized a relationship between a mathematical signature associated with the recombinant feature and arithmetic series with a ratio of 6 (nucleotide periodicity) in the S protein. We conducted a pilot study analyzing arithmetic series with a ratio of 6 in recombinant S proteins, finding a high presence of Thymine in series with an initial position multiple of 3. This pattern was not observed in other proteins like M and N. Repeating the process with non-recombinant sequences, we detected a similar pattern, although common nucleotides differed significantly between both sets. In this pilot study, we identified mathematical patterns associated with genetic recombination, specifically related to the third codon nucleotides. Our methodology was alignment-free and with minimal computational consumption. Our findings suggested the existence of mathematical signatures in the genome, especially in the encoded information of third codon nucleotides. Considering the diversity of perspectives and levels of information when reading and interpreting the genome may be necessary, acknowledging the existence of multiple layers of data and their interpretation.

Keywords: SARS-CoV-2 · Genetic Recombination · Third Codon Nucleotide · Third Nucleotide · Triplet · Mathematical Pattern · Mathematical Signature

* University of Deusto - Research Training Grants Program.

1 Introduction

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) emerged as a novel coronavirus in late 2019, swiftly escalating into a global pandemic [1]. Its remarkable contagiousness significantly fueled the tragic impact of Coronavirus Disease 2019 (COVID-19). The spectrum of SARS-CoV-2 symptoms ranges from mild respiratory ailments to severe pneumonia, and in some instances, fatality [11]. Since the onset of the pandemic, numerous variants have emerged, each possessing distinct evolutionary traits that contribute to heightened transmissibility or increased resistance to vaccines.

A critical mechanism driving viral evolution in coronaviruses is genetic recombination, where genetic material from different strains or viruses is combined, providing the resulting virus with competitive advantages such as adaptation to new hosts, evasion of host immunity, or resistance to antiviral treatments. This process can occasionally result in the emergence of new strains or even entirely new viruses. Hence, it is imperative to further explore the mechanisms underlying genetic recombination [17]. In the case of SARS-CoV-2, nearly all circulating recombinant lineages of SARS-CoV-2 are derived from Omicron [13]. Since 2023, the recombinant lineage XBB emerged dominant in the epidemiological landscape, owing to its increased transmissibility [14]. The Spike (S) protein stands as a pivotal genomic region in viral evolution, with multiple genetic recombination events detected within its genome. Playing a fundamental role in the infection process, the S protein binds to the ACE2 receptor (Angiotensin-converting Enzyme 2) on human cell surfaces, facilitating viral entry [18].

In previous research, we identified a mathematical signature embedded in complete genomic sequences of HIV-1 that determined the recombinant feature [8]. Therefore, the existence of mathematical patterns related to the recombinant feature in SARS-CoV-2 is also plausible. Based on this hypothesis, we conducted a pilot study to explore the existence of the mathematical signature that characterizes a SARS-CoV-2 sequence as recombinant.

This article presents the obtained results, as well as an analysis of their biological significance.

2 Materials

All experiments were ran in this equipment: Processing Unit: Intel(R) Core(TM) i7-4770K CPU. 3.5 GHz. Installed RAM: 32 GB usable. Operative System: Windows 10 Education. Version: 22H2. GPU: NVIDIA GeForce RTX 3090. Total memory: 40 GB.

3 Methods

All analyses were developed in a Python environment, using (among others): Pandas for data manipulation in the form of DataFrames [10]. Numpy, for implementing numerical operations [2]. Math, for applying certain mathematical

functions [4]. Os, for operations related to the operating system [5]. Re, for regular expression operations [6]. The Selenium WebDriver library (Python) for automating the web download of EMBL files using scripts [12].

In experiments from another field unrelated to the research objective of this article, we detected a relationship between the recombinant feature and arithmetic sequences with a span of 6 nucleotides in the S protein within specific recombinant sequences. Out of a total of 1,565 recombinant sequences, we identified a total of 39 sequences where this relationship seemed most evident, and we selected them as the pilot study group.

Subsequently, we selected the same number of non-recombinant sequences (39) from a total of 1,539,728 through an automated random subsampling process. This was done to approximate a search for differential elements in an environment where structural differences appeared to be more evident. In this manner, locating these differences between one study group and the other a priori could be more straightforward.

Subsequently, we extracted all arithmetic series with a common difference of 6 based on their initial position in the S, M, and N proteins in the recombinant sequences, aiming to detect the presence of any pattern exclusively inherent to S. From the selected non-recombinant sequences, we extracted such series in S to establish comparisons between the proteins of recombinant and non-recombinant sequences.

In each of the arithmetic series extracted in the previous step, we calculated the average occurrence of A, T, C, and G in each of the proteins analyzed, as well as the Standard Deviation (SD). The aim was to obtain information regarding the level of abundance of each nucleotide in each of these proteins, as well as the variability found in the sequences under analysis. These data may enable us to identify particularities among the arithmetic series in each analyzed protein. Furthermore, these data may provide clues about the mathematical signature embedded in the genome.

Then, we identified the common nucleotides within the S sequences among sequences belonging to the same group (recombinants and non-recombinants) and conducted a comparative analysis between the results of both groups. The differences between both nucleotide groups may provide information about the formulation of that mathematical signature.

Finally, we made an approach with existing literature in order to approximate the bio-centric discussion of the results.

4 Results

For each of the 39 selected recombinant sequences, we verified whether the position corresponding to the nucleotide periodicity of 6 aligns with the third letter of the second amino acid in S. Upon thorough examination, we confirmed that the sixth letter indeed matches in most cases. See Table 1. The "Position" column indicates the position of this sixth nucleotide from the start of S (not from the start of the entire sequence). $Position = 6$ means that the first sixth nucleotide

of S (counting from the start of the sequence) aligns with the third nucleotide of the second amino acid in S protein.

Table 1: Sixth Nucleotide Position

ID	Variant	Confidence Score	S		First 6th nt	Position
MZ558275	XB	99.99%	21,512	25,333	21,516	5
MZ280957	XB	69.42%	21,519	25,340	21,522	4
MZ817897	XC	68.77%	21,533	25,348	21,534	2
OK039706	XC	80.39%	21,532	25,353	21,534	3
OL791595	XC	99.99%	21,542	25,363	21,546	4
OL920359	XC	99.99%	21,533	25,354	21,534	2
OM059434	XC	99.99%	21,533	25,354	21,534	2
ON144454	XN	98.70%	21,529	25,341	21,534	6
ON330593	XAA	99.98%	21,530	25,342	21,534	5
ON594981	XAC	99.98%	21,529	25,341	21,534	6
ON650495	XAC	99.99%	21,529	25,341	21,534	6
ON800622	XAJ	65.51%	21,529	25,350	21,534	6
ON980049	XE	99.94%	21,526	25,338	21,528	3
OP129015	XAK	99.99%	21,529	25,341	21,534	6
OP294029	XAJ	99.88%	21,529	25,341	21,534	6
OP311263	XAG	74.12%	21,526	25,338	21,528	3
OP313193	XAJ	99.86%	21,529	25,341	21,534	6
OP313256	XAJ	99.99%	21,529	25,341	21,534	6
OP313383	XAJ	96.92%	21,526	25,338	21,528	3
OP359817	XAJ	99.84%	21,529	25,341	21,534	6
OP768851	XE	99.91%	21,526	25,338	21,528	3
OP934590	XBD	99.99%	21,529	25,341	21,534	6
OQ094825	XBB	99.97%	21,529	25,338	21,534	6
OQ111935	XBB	99.99%	21,529	25,338	21,534	6
OQ112611	XBB	99.99%	21,529	25,338	21,534	6
OQ112616	XBB	99.99%	21,542	25,351	21,546	4
OQ112619	XBB	99.99%	21,529	25,338	21,534	6
OQ112624	XBB	99.99%	21,529	25,338	21,534	6
OQ112639	XBB	99.99%	21,529	25,338	21,534	6
OQ112647	XBB	99.99%	21,529	25,338	21,534	6
OQ112649	XBB	99.99%	21,529	25,338	21,534	6
OQ112673	XBB	99.99%	21,529	25,338	21,534	6
OQ112686	XBB	99.99%	21,529	25,338	21,534	6
OQ112691	XBB	99.99%	21,529	25,338	21,534	6
OQ112730	XBB	99.99%	21,529	25,338	21,534	6
OQ112737	XBB	99.99%	21,529	25,338	21,534	6
OQ112740	XBB	99.99%	21,529	25,338	21,534	6

Table 1 – Continuation from the previous page

ID	Variant	Confidence Score	S		First 6th nt	Position
OQ112771	XBB	99.99%	21,529	25,338	21,534	6
OQ112893	XBB	99.99%	21,545	25,354	21,546	2

Table 2 presents a statistical analysis of the occurrence of each nucleotide within the frames exhibiting a ratio of 6 in the S protein (recombinant sequences).

In all tables displaying similar outcomes, the columns labeled "No. Avg." denote the average occurrence of each nucleotide type in the corresponding protein within each frame. "Avg. ± SD (%)" indicates "No. Avg", expressed as a percentage of the total number of nucleotides composing the S protein ± the standard deviation, also in %.

Table 2: SARS-CoV-2 Recombinant Sequences S Protein

	6, 12, ...		**9, 15, ...**	
	No. Avg.	**Avg. ± SD (%)**	**No. Avg.**	**Avg. ± SD (%)**
A	169.54	26.62 ± 1.75	170.92	26.87 ± 1.96
T	293.87	46.13± 1.69	288.90	45.42 ± 1.51
C	99.95	15.69 ± 0.71	99.46	15.64 ± 0.74
G	66.28	10.41 ± 1.21	70.13	11.03 ± 1.14
	7, 13, ...		**10, 16, ...**	
	No. Avg.	**Avg. ± SD (%)**	**No. Avg.**	**Avg. ± SD (%)**
A	188.59	29.65 ± 1.05	195.18	30.69 ± 0.63
T	155.72	24.48 ± 0.53	150.18	23.61 ± 1.02
C	106.54	16.75 ± 1.37	100.41	15.79 ± 1.51
G	178.15	28.01 ± 1.82	183.28	28.82 ± 1.24
	8, 14, ...		**11, 17, ...**	
	No. Avg.	**Avg. ± SD (%)**	**No. Avg.**	**Avg. ± SD (%)**
A	193.79	30.47 ± 0.69	196.21	30.85 ± 0.25
T	182.69	28.73 ± 1.75	183.79	28.90 ± 1.45
C	150.33	23.64 ± 0.99	155.03	24.38 ± 0.77
G	102.59	16.13 ± 1.05	93.62	14.72± 1.42

The amount of Thymine in multiples of 3 arithmetic series rises above 45%. The remaining 55% is distributed among the other three types of nucleotides, with a slight predominance of adenine. This phenomenon does not occur in sequences that are not multiples of 3. In non-multiples of 3 series, the distribution percentages among the 4 types of nucleotides are more balanced. In all cases, the standard deviation (SD) of each nucleotide was less than 2%. This means that in the 39 analyzed sequences, nucleotides at each position are generally similar with minor variations.

The results for the M proteins of these recombinant sequences are shown in Table 3.

Table 3: SARS-CoV-2 Recombinant Sequences M Protein

	6, 12, ...		**9, 15, ...**	
	No. Avg.	**Avg. ± SD (%)**	**No. Avg.**	**Avg. ± SD (%)**
A	22.97	20.70 ± 0.14	31.00	27.93 ± 0.00
T	36.56	32.94± 0.45	46.03	41.46 ± 0.14
C	28.44	25.62± 0.45	21.97	19.80 ± 0.14
G	23.03	20.74 ± 0.14	12.00	10.81 ± 0.00
	7, 13, ...		**10, 16, ...**	
	No. Avg.	**Avg. ± SD (%)**	**No. Avg.**	**Avg. ± SD (%)**
A	31.67	28.53 ± 0.47	35.00	31.82 ± 0.00
T	28.00	25.23 ± 0.00	21.00	19.09 ± 0.00
C	21.33	19.22 ± 0.42	27.00	24.55 ± 0.00
G	30.00	27.03 ± 0.29	27.00	24.55 ± 0.00
	8, 14, ...		**11, 17, ...**	
	No. Avg.	**Avg. ± SD (%)**	**No. Avg.**	**Avg. ± SD (%)**
A	30.95	27.88 ± 0.20	19.00	17.27 ± 0.00
T	37.00	33.33 ± 0.00	43.92	39.93 ± 0.25
C	24.00	21.62 ± 0.00	21.08	19.16 ± 0.25
G	19.05	17.16 ± 0.20	26.00	23.64 ± 0.00

Table 4 displays the results for the N protein.

Table 4: SARS-CoV-2 Recombinant Sequences N Protein

	6, 12, ...		**9, 15, ...**	
	No. Avg.	**Avg. ± SD (%)**	**No. Avg.**	**Avg. ± SD (%)**
A	61.54	29.53 ± 0.45	65.51	31.47 ± 0.49
T	71.69	34.42 ± 2.55	61.33	29.46 ± 2.37
C	40.74	19.54 ± 3.83	53.41	25.66 ± 3.77
G	34.38	16.51 ± 1.73	27.92	13.41 ± 1.88
	7, 13, ...		**10, 16, ...**	
	No. Avg.	**Avg. ± SD (%)**	**No. Avg.**	**Avg. ± SD (%)**
A	60.46	29.04 ± 2.50	68.85	33.21 ± 2.36
T	33.62	16.15 ± 0.49	28.54	13.76 ± 0.51
C	49.82	23.93 ± 0.23	52.00	25.08 ± 0.09
G	64.28	30.88 ± 1.81	57.97	27.95 ± 1.95
	8, 14, ...		**11, 17, ...**	
	No. Avg.	**Avg. ± SD (%)**	**No. Avg.**	**Avg. ± SD (%)**
A	70.26	33.75 ± 0.23	69.64	33.59 ± 0.50
T	36.49	17.53 ± 0.57	32.46	15.65 ± 0.41
C	57.18	27.47 ± 0.14	59.92	28.90 ± 0.20
G	44.26	21.26 ± 0.27	45.33	21.86 ± 0.27

In the comparisons made with the M and N proteins of these 39 recombinant sequences, we did not detect these patterns. The nucleotide quantities appeared more balanced than in the case of the S protein, especially regarding T, where the incidence levels were not reached in any of the arithmetic series in these proteins.

The next step was to compare the results of the S protein in the group of recombinant sequences with the results of the S protein in the non-recombinant sequences. See Table 5.

Table 5: SARS-CoV-2 Non-Recombinant Sequences S Protein

	6, 12, ...		**9, 15, ...**	
	No. Avg.	**Avg. ± SD (%)**	**No. Avg.**	**Avg. ± SD (%)**
A	167,05	26.59 ± 1.91	173.59	27.66 ± 1.96
T	290.74	46.29± 1.50	290.00	46.18 ± 1.69
C	99.15	15.78 ± 0.87	100.49	16.02 ± 0.88
G	71.21	11.35 ± 1.33	63.62	10.14 ± 1.17
	7, 13, ...		**10, 16, ...**	
	No. Avg.	**Avg. ± SD (%)**	**No. Avg.**	**Avg. ± SD (%)**
A	196.69	31.33 ± 1.03	185.74	29.62 ± 0.93
T	146.26	23.30 ± 0.91	157.11	25.05 ± 0.93
C	100.74	16.04 ± 1.30	105.85	16.89 ± 1.38
G	184.05	29.33 ± 1.37	178.38	28.44 ± 1.36
	8, 14, ...		**11, 17, ...**	
	No. Avg.	**Avg. ± SD (%)**	**No. Avg.**	**Avg. ± SD (%)**
A	196.67	31.32 ± 0.70	191.44	30.52 ± 0.70
T	182.00	29.00 ± 1.67	184.36	29.39 ± 1.49
C	151.97	24.20 ± 0.97	151.59	24.18 ± 0.80
G	97.21	15.48 ± 1.41	99.72	15.91 ± 1.29

In the group of non-recombinant sequences, the presence of Thymine in the third codon nucleotide maintained the high percentages of occurrence detected in the recombinant sequences. Therefore, this detected pattern was common between both study groups. It was be necessary to determine the exact common nucleotides shared in both groups to establish a detailed comparison.

5 Common Nucleotides

The common nucleotides in the case of the 39 recombinant sequences were those identified in Table 6. The "Position" column indicates the position of the common nucleotide in the 39 analyzed recombinant sequences, starting from the first nucleotide of S. The "Nucleotide" column specifies the type of nucleotide contained at that position, which is common in these 39 sequences.

Table 6: Common Nucleotides S Protein in Recombinant Sequences

Position	Nucleotide	Position	Nucleotide	Position	Nucleotide
3	G	48	T	123	A
6	T	51	T	165	C
9	T	54	T	177	T
12	T	57	A	189	T
15	T	63	A	198	T
18	T	66	T	318	T
21	A	69	A	363	T
24	G	72	A	369	T
27	A	78	T	399	T
30	A	93	T	405	T
36	T	102	T	582	T
39	T	108	T	1329	T
42	G	117	T	2382	T
45	T	120	C		

Surprisingly, all the common nucleotides in those 39 recombinant sequences were located at positions that were multiples of 3. That is, all the common elements were precisely the third nucleotide of the triplet. There was not a single common nucleotide at positions 1 and 2 of the triplet.

Table 7 compiles the common nucleotides in the 39 non-recombinant sequences under study.

Table 7: Common Nucleotides S Protein in Non-Recombinant Sequences

Position	Nucleotide	Position	Nucleotide	Position	Nucleotide
3	G	18	T	48	T
6	T	24	G	57	A
9	T	27	A	189	T
12	T	42	G	198	T
15	T	45	T	2388	T

Similar to the scenario observed in the 39 recombinant sequences, it is striking that once again, all the common elements in the S proteins of the 39 non-recombinant sequences analyzed were multiples of 3, the third nucleotides of the triplet.

The common nucleotides shared between the 39 recombinant and non-recombinant sequences were those displayed in Table 8.

Table 8: Common Nucleotides in the S Protein Sequences, Shared between Recombinants and Non-Recombinants

Position	Nucleotide	Position	Nucleotide	Position	Nucleotide
3	G	18	T	48	T
6	T	24	G	57	A
9	T	27	A	189	T
12	T	42	G	198	T
15	T	45	T		

The exclusive common elements of the recombinant sequences were positions 21 (A), 30 (A), 36 (T), 39 (T), 51 (T), 54 (T), 63 (A), 66 (T), 69 (A), 72 (A), 78 (T), 93 (T), 102 (T), 108 (T), 117 (T), 120 (C), 123 (A), 165 (C), 177 (T), 318 (T), 363 (T), 369 (T), 399 (T), 405 (T), 582 (T), 1329 (T), 2382 (T).

These exclusive common elements found within recombinant sequences could potentially bear relevance to a mathematically encrypted signature embedded within the third codon nucleotides of S. The pilot study offered valuable insights to guide future research endeavors focused on elucidating this mathematical signature.

6 Bio-Centric Discussion

In genetics and molecular biology, the third letter of a codon is frequently considered less crucial in determining the encoded amino acid compared to the first two letters. This concept was highlighted by Francis Crick in 1966 with his "wobble hypothesis" [3]. This fact provides some flexibility in the DNA sequence and allows mutations at this position to be more tolerated without affecting the final protein produced.

Therefore, we could infer that this third nucleotide is more prone to variability. However, in this pilot study, all the common nucleotides turned out to be precisely third nucleotides, not the first or second ones, which seemed to be more stable.

The limitation of the human ribosome to recognize the necessary information for generating transfer RNA can affect the synthesis of the corresponding amino acid. For instance, the CGC codon (arginine) is not efficiently recognized by the human ribosome, but the CGG, CGA, and CGU codons can encode this amino acid correctly. At times, the ribosome may opt for an alternative codon that fulfills the required function in mRNA translation [20]. The analysis of these factors in transcription must be studied thoroughly.

In arithmetic series multiples of 3, Thymine was significantly more abundant than the other three nucleotide types combined, accounting for nearly half of the entire set. However, this pattern was not observed in sequences not multiples of 3.

Hassanin (2022) discovered that genomes of coronaviruses related to SARS-CoV-2 contain fewer cytosines and more uracils at the third codon position compared to other sarbecoviruses [9], in alignment with findings by Rice et al. [15]

and Tort et al. [19]. These studies indicate a bias in SARS-CoV-2 codon usage resulting from mutations that lead to an enrichment of $A \pm U$ or U, among other factors. Saldivar-Espinoza et al. [16] also documented that mutations where cytosine is replaced by uracil (C>U) are the most common in the SARS-CoV-2 genome. They proposed that host deaminases, enzymes that catalyze the deamination of unpaired nitrogenous bases, were accountable for a significant portion of these mutations.

In accordance with the human genetic code, as outlined in the Codon Usage Frequency Table [7], the third nucleotide of a codon can be either C or U without affecting the encoded amino acid. Hence, if the elevated quantity of U in the third codon position arises from a C>U transition, it would not modify the amino acid sequence of the resultant Spike protein.

However, it is noteworthy to emphasize that in our analysis, other structural proteins such as M and N exhibited distinct quantity distribution patterns compared to S. Consequently, we acknowledge that the elevated amount of U at a specific position may possess biological significance warranting further investigation for interpretation.

While these phenomena do not directly elucidate the connection between the identified mathematical pattern categorizing a SARS-CoV-2 sequence as recombinant and the frequencies of Uracil (Thymine) appearance in the S protein, we cannot dismiss the possibility of a correlation between them.

7 Conclusion

In this pilot study focusing on a group of recombinant SARS-CoV-2 sequences, we identified a pattern characterized by a high Thymine content in the third codon positions of the S protein, which was not observed in other structural proteins like M or N. Our research results, depicted in the genomic spectrogram, revealed this mathematical pattern associated with sequence recombination. By comparing the patterns detected in the recombinant group with equivalent patterns in a non-recombinant group of the same size, we observed a similar characteristic pattern specific to the S protein.

We extracted the common nucleotides in each group of sequences, recombinant and non-recombinant. All these common nucleotides were multiples of 3, that is, the third nucleotide of the triplet.

This search for common nucleotides was conducted without applying any type of multiple alignment; that is, in sequences of varying lengths in S, we found that precisely the common elements were third nucleotides of the codon without performing any prior "artificial" adjustment before the comparison.

Finally, we established which common elements were shared by both recombinants and non-recombinants to try to shed some light on that mathematical signature encrypted in the genome.

The evidence found suggests that this encrypted mathematical signature is located in the third nucleotide of the triplets in S (in the case of SARS-CoV-2).

This pilot study strengthens the hypothesis that the third nucleotide of the triplet, less relevant in terms of amino acid coding, may contain additional encrypted information that characterizes (in this study's case) a SARS-CoV-2 sequence as recombinant.

Using methodologies that do not require multiple alignments and very low-computation tools, we obtained sufficient evidence in this pilot study to firmly hypothesize the possibility that the third nucleotide of the triplet contains encrypted information related to viral characteristics.

Although much remains to be discovered about the intricate rules of the genome, this document may mark the beginning of a new perspective for understanding and analyzing the genome. The results obtained reveal the identification of mathematical signatures in the virus genome through genomic spectrogram analysis, opening up new possibilities for investigating potential functions associated with these mathematical patterns.

References

1. Atzrodt, C.L., Maknojia, I., McCarthy, R.D., Oldfield, T.M., Po, J., Ta, K.T., Stepp, H.E., Clements, T.P.: A guide to COVID-19: a global pandemic caused by the novel coronavirus SARS-CoV-2. The FEBS journal **287**(17), 3633–3650 (2020)
2. Council, N.S.: Numpy library (2024), https://numpy.org/, accessed: 2024-04-30
3. Crick, F.: Codon-anticodon pairing: the Wobble hypothesis. Molecular Biology: A Selection of Papers **19**, 548–555 (1966)
4. Foundation, P.S.: Math Python library (2024), https://docs.python.org/3/library/math.html/, accessed: 2024-04-30
5. Foundation, P.S.: Os Python library (2024), https://docs.python.org/es/3.10/library/os.html/, accessed: 2024-04-30
6. Foundation, P.S.: Re Python library (2024), https://docs.python.org/es/3/library/re.html/, accessed: 2024-04-30
7. GenScript: Genscript Codon Usage Frequency Table (chart) Tool (Human Genetic Code). https://www.genscript.com/tools/codon-frequency-table/ (2024), accessed on March 21, 2024
8. Guerrero-Tamayo, A., Urquijo, B.S., Casado, C., Tosantos, M.D.M., Olivares, I., Pastor-López, I.: Discovering mathematical patterns behind HIV-1 genetic recombination: a new methodology to identify viral features. IEEE Access (2023)
9. Hassanin, A.: Variation in synonymous nucleotide composition among genomes of sarbecoviruses and consequences for the origin of COVID-19. Gene **835**, 146641 (2022)
10. Management, A.C., community of contributors: Pandas library (2024), https://pandas.pydata.org/, accessed: 2024-04-30
11. Marks, K.M., Gulick, R.M.: COVID-19. Annals of Internal Medicine **176**(10), ITC145–ITC160 (2023)
12. Muthukadan, B.: Selenium with Python. https://selenium-python.readthedocs.io/ (2024), accessed on May 03, 2024
13. Ou, J., Lan, W., Wu, X., Zhao, T., Duan, B., Yang, P., Ren, Y., Quan, L., Zhao, W., Seto, D., et al.: Tracking SARS-CoV-2 Omicron diverse spike gene mutations identifies multiple inter-variant recombination events. Signal Transduction and Targeted Therapy **7**(1), 138 (2022)

14. Parums, D.V.: The XBB. 1.5 ('Kraken') subvariant of Omicron SARS-CoV-2 and its rapid global spread. Medical science monitor: international medical journal of experimental and clinical research **29**, e939580–1 (2023)
15. Rice, A.M., Castillo Morales, A., Ho, A.T., Mordstein, C., Mühlhausen, S., Watson, S., Cano, L., Young, B., Kudla, G., Hurst, L.D.: Evidence for strong mutation bias toward, and selection against, U content in SARS-CoV-2: implications for vaccine design. Molecular biology and evolution **38**(1), 67–83 (2021)
16. Saldívar-Espinoza, B., Garcia-Segura, P., Novau-Ferré, N., Macip, G., Martínez, R., Puigbò, P., Cereto-Massagué, A., Pujadas, G., Garcia-Vallve, S.: The mutational landscape of SARS-CoV-2. International Journal of Molecular Sciences **24**(10), 9072 (2023)
17. Su, S., Wong, G., Shi, W., Liu, J., Lai, A.C., Zhou, J., Liu, W., Bi, Y., Gao, G.F.: Epidemiology, genetic recombination, and pathogenesis of coronaviruses. Trends in microbiology **24**(6), 490–502 (2016)
18. Tagliamonte, M.S., Abid, N., Borocci, S., Sangiovanni, E., Ostrov, D.A., Kosakovsky Pond, S.L., Salemi, M., Chillemi, G., Mavian, C.: Multiple recombination events and strong purifying selection at the origin of SARS-CoV-2 spike glycoprotein increased correlated dynamic movements. International Journal of Molecular Sciences **22**(1), 80 (2020)
19. Tort, F.L., Castells, M., Cristina, J.: A comprehensive analysis of genome composition and codon usage patterns of emerging coronaviruses. Virus research **283**, 197976 (2020)
20. Zabolotskii, A.I., Kozlovskiy, S.V., Katrukha, A.G.: The influence of the nucleotide composition of genes and gene regulatory elements on the efficiency of protein expression in Escherichia coli. Biochemistry (Moscow) **88**(Suppl 1), S176–S191 (2023)

Lee y conducirás, no leas y serás conducido. ...

— Santa Teresa de Jesús, *una mujer espectacular, una de las mejores mentes de la historia de la humanidad. Y además, tenía un gran sentido del humor. Cuentan que un día se quejó de lo mal que le trataba la vida y el Señor le dijo "Teresa, así trato yo a mis amigos". Teresa le respondió "Ah, Señor, por eso tienes tan pocos".*

References

[1] Jane E Sykes. "Feline coronavirus infection". In: *Canine and feline infectious diseases* (2014), p. 195.

[2] Istvan Kiss, AM Poland, and Niels C Pedersen. "Disease outcome and cytokine responses in cats immunized with an avirulent feline infectious peritonitis virus (FIPV)-UCD1 and challenge-exposed with virulent FIPV-UCD8". In: *Journal of Feline Medicine and Surgery* 6.2 (2004), pp. 89–97.

[3] Kjavs Hök. "Morbidity, mortality and coronavirus antigen in previously coronavirus free kittens placed in two catteries with feline infectious peritonitis". In: *Acta Veterinaria Scandinavica* 34 (1993), pp. 203–210.

[4] Niels C Pedersen et al. "Efficacy of a 3C-like protease inhibitor in treating various forms of acquired feline infectious peritonitis". In: *Journal of feline medicine and surgery* 20.4 (2018), pp. 378–392.

[5] Arnold APM Herrewegh et al. "Persistence and evolution of feline coronavirus in a closed cat-breeding colony". In: *Virology* 234.2 (1997), pp. 349–363.

[6] Ute Klein-Richers et al. "Prevalence of feline coronavirus shedding in German catteries and associated risk factors". In: *Viruses* 12.9 (2020), p. 1000.

[7] Yvonne Drechsler et al. "Feline coronavirus in multicat environments". In: *Veterinary Clinics: Small Animal Practice* 41.6 (2011), pp. 1133–1169.

[8] Katrin Hartmann. "Feline infectious peritonitis". In: *Veterinary Clinics: Small Animal Practice* 35.1 (2005), pp. 39–79.

[9] Arnold APM Herrewegh et al. "Feline coronavirus type II strains 79-1683 and 79-1146 originate from a double recombination between feline coronavirus type I and canine coronavirus". In: *Journal of virology* 72.5 (1998), pp. 4508–4514.

[10] Yutaka Terada et al. "Emergence of pathogenic coronaviruses in cats by homologous recombination between feline and canine coronaviruses". In: *PloS one* 9.9 (2014), e106534.

[11] Mart M Lamers and Bart L Haagmans. "SARS-CoV-2 pathogenesis". In: *Nature reviews microbiology* 20.5 (2022), pp. 270–284.

[12] Di Wu et al. "The SARS-CoV-2 outbreak: what we know". In: *International journal of infectious diseases* 94 (2020), pp. 44–48.

[13] Jatin Machhi et al. “The natural history, pathobiology, and clinical manifestations of SARS-CoV-2 infections”. In: *Journal of Neuroimmune Pharmacology* 15 (2020), pp. 359–386.

[14] Wendy K Jo et al. “Potential zoonotic sources of SARS-CoV-2 infections”. In: *Transboundary and emerging diseases* 68.4 (2021), pp. 1824–1834.

[15] Spyros Lytras et al. “The animal origin of SARS-CoV-2”. In: *Science* 373.6558 (2021), pp. 968–970.

[16] James C Alwine et al. *A critical analysis of the evidence for the SARS-CoV-2 origin hypotheses*. 2023.

[17] National Center for Biotechnology Information National Library of Medicine (US). *NCBI SARS-CoV-2 Data Hub*. `https://www.ncbi.nlm.nih.gov/sars-cov-2/`. Accessed on March 20, 2023. 2024.

[18] Global Change Data Lab. *Our World in Data*. `https://ourworldindata.org/`. Accessed on March 20, 2023. 2024.

[19] Santiago Emanuel Colina et al. “Clinical and molecular aspects of veterinary coronaviruses”. In: *Virus research* 297 (2021), p. 198382.

[20] Javier A Jaimes and Gary R Whittaker. “Feline coronavirus: Insights into viral pathogenesis based on the spike protein structure and function”. In: *Virology* 517 (2018), pp. 108–121.

[21] Giorgio Palù et al. “Biology of Coronaviruses and Predicted Origin of SARS-CoV-2”. In: *COVID-19*. CRC Press, 2022, pp. 19–26.

[22] Zi-Wei Ye et al. “Zoonotic origins of human coronaviruses”. In: *International journal of biological sciences* 16.10 (2020), p. 1686.

[23] Anne Sofie Hammer et al. “SARS-CoV-2 transmission between mink (Neovison vison) and humans, Denmark”. In: *Emerging infectious diseases* 27.2 (2021), p. 547.

[24] Jianzhong Shi et al. “Susceptibility of ferrets, cats, dogs, and other domesticated animals to SARS–coronavirus 2”. In: *Science* 368.6494 (2020), pp. 1016–1020.

[25] Jingsong Zhang et al. “Potential transmission chains of variant B. 1.1. 7 and co-mutations of SARS-CoV-2”. In: *Cell Discovery* 7.1 (2021), p. 44.

[26] Amalio Telenti, Emma B Hodcroft, and David L Robertson. “The evolution and biology of SARS-CoV-2 variants”. In: *Cold Spring Harbor perspectives in medicine* 12.5 (2022), a041390.

[27] EFSA Panel on Animal Health et al. “SARS-CoV-2 in animals: susceptibility of animal species, risk for animal and public health, monitoring, prevention and control”. In: *EFSA Journal* 21.2 (2023), e07822.

[28] Saverio Paltrinieri et al. “Feline infectious peritonitis (FIP) and coronavirus disease 19 (COVID-19): Are they similar?” In: *Transboundary and Emerging Diseases* 68.4 (2021), pp. 1786–1799.

[29] Shuai Lu et al. "Discovery of a novel canine respiratory coronavirus support genetic recombination among Betacoronavirus1". In: *Virus research* 237 (2017), pp. 7–13.

[30] Ben Jackson et al. "Generation and transmission of interlineage recombinants in the SARS-CoV-2 pandemic". In: *Cell* 184.20 (2021), pp. 5179–5188.

[31] Filippo Scialo et al. "ACE2: the major cell entry receptor for SARS-CoV-2". In: *Lung* 198 (2020), pp. 867–877.

[32] Alison E Stout et al. "Coronaviruses in cats and other companion animals: where does SARS-CoV-2/COVID-19 fit?" In: *Veterinary microbiology* 247 (2020), p. 108777.

[33] Ceren Guney and Fatma Akar. "Epithelial and endothelial expressions of ACE2: SARS-CoV-2 entry routes." In: *Journal of pharmacy & pharmaceutical sciences: a publication of the Canadian Society for Pharmaceutical Sciences, Societe canadienne des sciences pharmaceutiques* 24 (2021).

[34] Lowiese MB Desmarets et al. "Establishment of feline intestinal epithelial cell cultures for the propagation and study of feline enteric coronaviruses". In: *Veterinary research* 44 (2013), pp. 1–13.

[35] Michail Tsimpidis et al. "T-RECs: rapid and large-scale detection of recombination events among different evolutionary lineages of viral genomes". In: *BMC bioinformatics* 18 (2017), pp. 1–8.

[36] Ales Varabyou et al. "Rapid detection of inter-clade recombination in SARS-CoV-2 with Bolotie". In: *Genetics* 218.3 (2021), iyab074.

[37] Laure Wynants et al. "Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal". In: *bmj* 369 (2020).

[38] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[39] ENA European Nucleotide Archive. *ENA*. Accessed: 2023-05-12. 2023. URL: `https://www.ebi.ac.uk/ena/browser/`.

[40] National Center for Biotechnology Information National Library of Medicine (US). *NCBI Virus*. `https://www.ncbi.nlm.nih.gov/labs/virus/vssi/`. Accessed on September 11, 2023. 2023.

[41] Los Alamos National Laboratory. *The HIV Database*. Accessed: 2023-06-09. 2023. URL: `https://www.hiv.lanl.gov/`.

[42] Anthony S Fauci and H Clifford Lane. "Four decades of HIV/AIDSmuch accomplished, much to do". In: *New England Journal of Medicine* 383.1 (2020), pp. 1–4.

[43] Frederick Sanger, Steven Nicklen, and Alan R Coulson. "DNA sequencing with chain-terminating inhibitors". In: *Proceedings of the national academy of sciences* 74.12 (1977), pp. 5463–5467.

[44] Michael S Gottlieb. "Pneumocystis PneumoniaLos Angeles". In: *American journal of public health* 96.6 (2006), pp. 980–981.

[45] Michael H Malim and Michael Emerman. "HIV-1 sequence variation: drift, shift, and attenuation". In: *Cell* 104.4 (2001), pp. 469–472.

[46] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep Learning". In: *Nature* 521.7553 (2015), pp. 436–444.

[47] Matthew D Zeiler and Rob Fergus. "Visualizing and understanding convolutional networks". In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer. 2014, pp. 818–833.

[48] Lisa Torrey and Jude Shavlik. "Transfer learning". In: *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI global, 2010, pp. 242–264.

[49] Thomas D Pollard et al. *Cell Biology E-book*. Elsevier Health Sciences, 2022.

[50] Robert Nussbaum, Roderick R McInnes, and Huntington F Willard. *Thompson & Thompson genetics in medicine e-book*. Elsevier Health Sciences, 2015.

[51] John B Jenkins. *Genética: Principios de la genética*. Reverté, 2014.

[52] Laura Patricia Alejos Velázquez, María del Consuelo Aragón Martínez, and Amelia Cornejo Romero. "Extracción y purificación de ADN". In: *Herramientas moleculares aplicadas en ecología: aspectos teóricos y prácticos* 1 (2014).

[53] Daniel Yero Corona. *Brock. Biología de los microorganismos*. Pearson, 2015.

[54] Nomenclature Committee of the International Union of Biochemistry (NC-IUB). "Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984". In: *Proceedings of the National Academy of Sciences of the United States of America* (1986), pp. 4–8.

[55] Aaron David Goldman and Laura F Landweber. "What is a genome?" In: *PLoS genetics* 12.7 (2016), e1006181.

[56] James W Fickett. "Recognition of protein coding regions in DNA sequences". In: *Nucleic acids research* 10.17 (1982), pp. 5303–5318.

[57] Jan-Willem Taanman. "The mitochondrial genome: Structure, transcription, translation and replication". In: *Biochimica et Biophysica Acta (BBA)-Bioenergetics* 1410.2 (1999), pp. 103–123.

[58] Severo Ochoa. *Escritos*. Editorial CSIC-CSIC Press, 1999.

[59] Jacques Demongeot and Hervé Seligmann. "Why is AUG the start codon? Theoretical minimal RNA rings: Maximizing coded information biases 1st codon for the universal initiation codon AUG". In: *Bioessays* 42.6 (2020), p. 1900201.

[60] Brian Søgaard Laursen et al. "Initiation of protein synthesis in bacteria". In: *Microbiology and molecular biology reviews* 69.1 (2005), pp. 101–123.

[61] National Human Genome Research Institute. *Genetic Code*. Accessed: 2023-10-09. 2023. URL: https://www.genome.gov/es/genetics-glossary/Codigo-genetico.

[62] Helga Ochoterena. "Homology in coding and non-coding DNA sequences: a parsimony perspective". In: *Plant Systematics and Evolution* 282 (2009), pp. 151–168.

[63] James W Fickett and Wyeth W Wasserman. "Discovery and modeling of transcriptional regulatory regions". In: *Current Opinion in Biotechnology* 11.1 (2000), pp. 19–24.

[64] Makoto Hijikata et al. "Gene mapping of the putative structural region of the Hepatitis C virus genome by in vitro processing analysis." In: *Proceedings of the National Academy of Sciences* 88.13 (1991), pp. 5547–5551.

[65] Hui Ling et al. "Junk DNA and the long non-coding RNA twist in cancer genetics". In: *Oncogene* 34.39 (2015), pp. 5003–5011.

[66] Karen Adelman and Emily Egan. "More uses for genomic junk". In: *Nature* 543.7644 (2017), pp. 183–185.

[67] Tetsuro Hirose, Yuichiro Mishima, and Yukihide Tomari. "Elements and machinery of non-coding RNA s: toward their taxonomy". In: *EMBO reports* 15.5 (2014), pp. 489–507.

[68] Joaquín Martínez Martínez, Francisco Martinez-Hernandez, and Manuel Martinez-Garcia. "Single-virus genomics and beyond". In: *Nature Reviews Microbiology* 18.12 (2020), pp. 705–716.

[69] Viswanatha Chaitanya. "Structure and organization of virus genomes". In: *Genome and Genomics: From Archaea to Eukaryotes* (2019), pp. 1–30.

[70] Frederick Sanger et al. "Nucleotide sequence of bacteriophage φX174 DNA". In: *nature* 265.5596 (1977), pp. 687–695.

[71] Amir Saberi et al. "A planarian nidovirus expands the limits of RNA genome size". In: *PLoS pathogens* 14.11 (2018), e1007314.

[72] David P Maison et al. "Genetic characteristics and phylogeny of 969-bp S gene sequence of SARS-CoV-2 from Hawaii reveals the worldwide emerging P681H mutation". In: *Hawaii journal of health & social welfare* 80.3 (2021), p. 52.

[73] Alexander E Gorbalenya et al. "Nidovirales: Evolving the largest RNA virus genome". In: *Virus research* 117.1 (2006), pp. 17–37.

[74] David M Knipe et al. *Fields Virology, Volumes 1 and 2*. Lippincott Williams & Wilkins: Philadelphia, PA, USA, 2013.

[75] Susan F Cotmore et al. "Structure and organization of the viral genome". In: *Parvoviruses. Hodder Arnold, London, United Kingdom* (2005), pp. 73–94.

[76] Ángel San Miguel-Hernández et al. "Variantes o mutaciones del virus de la Hepatitis B". In: *Gaceta Médica de Bilbao* 115.2 (2018), pp. 96–103.

[77] Jason D Fernandes et al. “Functional segregation of overlapping genes in HIV”. In: *Cell* 167.7 (2016), pp. 1762–1773.

[78] Siyang Sun, Venigalla B Rao, and Michael G Rossmann. “Genome packaging in viruses”. In: *Current opinion in structural biology* 20.1 (2010), pp. 114–120.

[79] A Kramvis and MC Kew. “The core promoter of Hepatitis B virus”. In: *Journal of viral hepatitis* 6.6 (1999), pp. 415–427.

[80] Enrico Ne, Robert-Jan Palstra, and Tokameh Mahmoudi. “Transcription: Insights from the HIV-1 promoter”. In: *International review of cell and molecular biology* 335 (2018), pp. 191–243.

[81] George Khoury and Peter Gruss. “Enhancer elements”. In: *Cell* 33.2 (1983), pp. 313–314.

[82] Lili Kuo et al. “Effects of mutations in the gene-start and gene-end sequence motifs on transcription of monocistronic and dicistronic minigenomes of respiratory syncytial virus”. In: *Journal of virology* 70.10 (1996), pp. 6892–6901.

[83] C Martin Stoltzfus and Joshua M Madsen. “Role of viral splicing elements and cellular RNA binding proteins in regulation of HIV-1 alternative RNA splicing”. In: *Current HIV research* 4.1 (2006), pp. 43–55.

[84] E Soeda et al. “Sequence from early region of polyoma virus DNA containing viral replication origin and encoding small, middle and (part of) large T antigens”. In: *Cell* 17.2 (1979), pp. 357–370.

[85] Catherine A Ramsey and Antonito T Panganiban. “Replication of the retroviral terminal repeat sequence during in vivo reverse transcription”. In: *Journal of virology* 67.7 (1993), pp. 4114–4121.

[86] Mark A Samols et al. “Cloning and identification of a microRNA cluster within the latency-associated region of Kaposi’s sarcoma-associated herpesvirus”. In: *Journal of virology* 79.14 (2005), pp. 9301–9305.

[87] Yik Lim Kok, Angela Ciuffi, and Karin J Metzner. “Unravelling HIV-1 latency, one cell at a time”. In: *Trends in microbiology* 25.11 (2017), pp. 932–941.

[88] Leyi Wang and Chang-Won Lee. “Sequencing and mutational analysis of the non-coding regions of Influenza A virus”. In: *Veterinary microbiology* 135.3-4 (2009), pp. 239–247.

[89] Claudia V Filomatori et al. “Dengue virus genomic variation associated with mosquito adaptation defines the pattern of viral non-coding RNAs and fitness in human cells”. In: *PLoS pathogens* 13.3 (2017), e1006265.

[90] Donald B Smith et al. “Discrimination of Hepatitis G virus/GBV-C geographical variants by analysis of the 5’non-coding region”. In: *Journal of General Virology* 78.7 (1997), pp. 1533–1542.

[91] Ken Fujii et al. "Importance of both the coding and the segment-specific noncoding regions of the Influenza A virus NS segment for its efficient incorporation into virions". In: *Journal of virology* 79.6 (2005), pp. 3766–3774.

[92] SIB Swiss Institute of Bioinformatics and Biozentrum. *SWISS-MODEL*. Accessed: 2024-05-28. 2024. URL: `https://swissmodel.expasy.org/`.

[93] Susanna C Manrubia and Ester Lázaro. "Viral evolution". In: *Physics of Life Reviews* 3.2 (2006), pp. 65–92.

[94] Markus G Weinbauer and Fereidoun Rassoulzadegan. "Are viruses driving microbial diversification and diversity?" In: *Environmental microbiology* 6.1 (2004), pp. 1–11.

[95] T Hoenen et al. "Mutation rate and genotype variation of Ebola virus from Mali case sequences". In: *Science* 348.6230 (2015), pp. 117–119.

[96] Siobain Duffy. "Why are RNA virus mutation rates so damn high?" In: *PLoS Biology* 16.8 (2018), e3000003.

[97] Philip W Hedrick. "Adaptive introgression in animals: Examples and comparison to new mutation and standing variation as sources of adaptive variation". In: *Molecular ecology* 22.18 (2013), pp. 4606–4618.

[98] Misha Soskine and Dan S Tawfik. "Mutational effects and the evolution of new protein functions". In: *Nature Reviews Genetics* 11.8 (2010), pp. 572–582.

[99] Marc Jacobsen et al. "A point mutation in PTPRC is associated with the development of multiple sclerosis". In: *Nature genetics* 26.4 (2000), pp. 495–499.

[100] Hidefumi Sasaki et al. "EGFR exon 20 insertion mutation in Japanese lung cancer". In: *Lung Cancer* 58.3 (2007), pp. 324–328.

[101] Gurmeet K Dhaliwal and Raji P Grewal. "Mitochondrial DNA deletion mutation levels are elevated in ALS brains". In: *Neuroreport* 11.11 (2000), pp. 2507–2509.

[102] Susumu Ohno. *Evolution by gene duplication*. Springer Science & Business Media, 2013.

[103] Sharon Windsor, Sherryl AM Taylor, and David Lillicrap. "Direct detection of a common inversion mutation in the genetic diagnosis of severe hemophilia A". In: *Blood* 84.7 (1994), pp. 2202–2205.

[104] Jae-kyung Kyoung Won et al. "Concomitant ALK translocation and EGFR mutation in lung cancer: A comparison of direct sequencing and sensitive assays and the impact on responsiveness to tyrosine kinase inhibitor". In: *Annals of oncology* 26.2 (2015), pp. 348–354.

[105] Michael J Imperiale and Mengxi Jiang. "What DNA viral genomic rearrangements tell us about persistence". In: *Journal of virology* 89.4 (2015), pp. 1948–1950.

[106] Heinrich G Göttlinger. "The HIV-1 assembly machine". In: *Aids* 15 (2001), S13–S20.

[107] Michele W Tang and Robert W Shafer. "HIV-1 antiretroviral resistance: scientific principles and clinical applications". In: *Drugs* 72 (2012), e1–e25.

[108] Jennifer Louten. "Virus replication". In: *Essential Human Virology* (2016), p. 49.

[109] Michael M Lai. "Genetic recombination in RNA viruses". In: *Genetic Diversity of RNA Viruses* (1992), pp. 21–32.

[110] Robert G Webster and W Graeme Laver. "The origin of pandemic Influenza". In: *Bulletin of the World Health Organization* 47.4 (1972), p. 449.

[111] Marcia L Kalish et al. "Recombinant viruses and early global HIV-1 epidemic". In: *Emerging Infectious Diseases* 10.7 (2004), p. 1227.

[112] Peter Spreeuwenberg, Madelon Kroneman, and John Paget. "Reassessing the global mortality burden of the 1918 Influenza pandemic". In: *American Journal of Epidemiology* 187.12 (2018), pp. 2561–2567.

[113] Amanda Mocroft et al. "Changing patterns of mortality across Europe in patients infected with HIV-1". In: *The Lancet* 352.9142 (1998), pp. 1725–1730.

[114] Andrew MQ King. "Genetic recombination in positive strand RNA viruses". In: *RNA Genetics*. CRC Press, 2018, pp. 149–165.

[115] Michael M Lai. "RNA recombination in animal and plant viruses". In: *Microbiological Reviews* 56.1 (1992), pp. 61–79.

[116] Claire Muslin et al. "Recombination in enteroviruses, a multi-step modular evolutionary process". In: *Viruses* 11.9 (2019), p. 859.

[117] Shuo Su et al. "Epidemiology, genetic recombination, and pathogenesis of coronaviruses". In: *Trends in Microbiology* 24.6 (2016), pp. 490–502.

[118] Lei Jia et al. "Recombination pattern reanalysis of some HIV-1 circulating recombination forms suggest the necessity and difficulty of revision". In: *PLoS One* 9.9 (2014), e107349.

[119] Gema Casado et al. "Identification of a novel HIV-1 circulating ADG intersubtype recombinant form (CRF19_cpx) in Cuba". In: *JAIDS Journal of Acquired Immune Deficiency Syndromes* 40.5 (2005), pp. 532–537.

[120] David L Robertson et al. "HIV-1 nomenclature proposal". In: *Science* 288.5463 (2000), pp. 55–55.

[121] Vladimir Likic. "The Needleman-Wunsch algorithm for sequence alignment". In: *Lecture given at the 7th Melbourne Bioinformatics Course, Bi021 Molecular Science and Biotechnology Institute, University of Melbourne* (2008), pp. 1–46.

[122] Zeyu Xia et al. "A review of parallel implementations for the Smith–Waterman algorithm". In: *Interdisciplinary Sciences: Computational Life Sciences* (2021), pp. 1–14.

[123] Julie D Thompson, Toby J Gibson, and Des G Higgins. "Multiple sequence alignment using ClustalW and ClustalX". In: *Current protocols in bioinformatics* 1 (2003), pp. 2–3.

[124] Kazutaka Katoh, George Asimenos, and Hiroyuki Toh. "Multiple alignment of DNA sequences with MAFFT". In: *Bioinformatics for DNA Sequence Analysis*. Springer, 2009, pp. 39–64.

[125] Robert C Edgar and Serafim Batzoglou. "Multiple sequence alignment". In: *Current opinion in structural biology* 16.3 (2006), pp. 368–373.

[126] Fabian Sievers and Desmond G Higgins. "Clustal Omega, accurate alignment of very large numbers of sequences". In: *Multiple sequence alignment methods* (2014), pp. 105–116.

[127] Desmond G Higgins and Paul M Sharp. "CLUSTAL: a package for performing multiple sequence alignment on a microcomputer". In: *Gene* 73.1 (1988), pp. 237–244.

[128] Robert C Edgar. "MUSCLE: multiple sequence alignment with high accuracy and high throughput". In: *Nucleic Acids Research* 32.5 (2004), pp. 1792–1797.

[129] Cédric Notredame, Desmond G Higgins, and Jaap Heringa. "T-Coffee: A novel method for fast and accurate multiple sequence alignment". In: *Journal of Molecular Biology* 302.1 (2000), pp. 205–217.

[130] Chuong B Do et al. "ProbCons: Probabilistic consistency-based multiple sequence alignment". In: *Genome Research* 15.2 (2005), pp. 330–340.

[131] Yongchao Liu, Bertil Schmidt, and Douglas L Maskell. "MSAProbs: multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities". In: *Bioinformatics* 26.16 (2010), pp. 1958–1964.

[132] Yongtao Ye et al. "GLProbs: Aligning multiple sequences adaptively". In: *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*. 2013, pp. 152–160.

[133] Joseph Felsenstein. "Inferring phylogenies". In: *Inferring phylogenies*. 2004, pp. 664–664.

[134] Andrew MQ King et al. *Virus taxonomy: ninth report of the International Committee on Taxonomy of Viruses*. Vol. 9. Elsevier, 2011.

[135] Rafael Sanjuán et al. "Viral mutation rates". In: *Journal of virology* 84.19 (2010), pp. 9733–9748.

[136] MMC Lai. "Genetic recombination in RNA viruses". In: *Genetic diversity of RNA viruses* (1992), pp. 21–32.

[137] David A Steinhauer and JJ Holland. "Rapid evolution of RNA viruses". In: *Annual Reviews in Microbiology* 41.1 (1987), pp. 409–431.

[138] Kyriakos Tsangaras and Alex D Greenwood. “Paleovirology: Viral sequences from historical and ancient DNA”. In: *Paleogenomics: Genome-Scale Analysis of Ancient DNA* (2019), pp. 139–162.

[139] Naruya Saitou and Masatoshi Nei. “The neighbor-joining method: a new method for reconstructing phylogenetic trees.” In: *Molecular biology and evolution* 4.4 (1987), pp. 406–425.

[140] Sebastien Roch. “A short proof that phylogenetic tree reconstruction by maximum likelihood is hard”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 3.1 (2006), pp. 92–94.

[141] Mike Steel. “The maximum likelihood point for a phylogenetic tree is not unique”. In: *Systematic Biology* 43.4 (1994), pp. 560–564.

[142] Lavanya Kannan and Ward C Wheeler. “Maximum parsimony on phylogenetic networks”. In: *Algorithms for molecular biology* 7 (2012), pp. 1–10.

[143] David L Swofford, Jack Sullivan, et al. “Phylogeny inference based on parsimony and other methods using PAUP*”. In: *The phylogenetic handbook: a practical approach to DNA and protein phylogeny, cáp* 7 (2003), pp. 160–206.

[144] M Azim Ansari and Xavier Didelot. “Bayesian inference of the evolution of a phenotype distribution on a phylogenetic tree”. In: *Genetics* 204.1 (2016), pp. 89–98.

[145] Alexei J Drummond and Andrew Rambaut. “BEAST: Bayesian evolutionary analysis by sampling trees”. In: *BMC evolutionary biology* 7 (2007), pp. 1–8.

[146] Peter Dawyndt, Hans De Meyer, and Bernard De Baets. “UPGMA clustering revisited: A weight-driven approach to transitive approximation”. In: *International Journal of Approximate Reasoning* 42.3 (2006), pp. 174–191.

[147] Morgan N Price, Paramvir S Dehal, and Adam P Arkin. “FastTree: computing large minimum evolution trees with profiles instead of a distance matrix”. In: *Molecular biology and evolution* 26.7 (2009), pp. 1641–1650.

[148] Stéphane Guindon et al. “Estimating maximum likelihood phylogenies with PhyML”. In: *Bioinformatics for DNA sequence analysis* (2009), pp. 113–137.

[149] John P Huelsenbeck and Fredrik Ronquist. “MRBAYES: Bayesian inference of phylogenetic trees”. In: *Bioinformatics* 17.8 (2001), pp. 754–755.

[150] Koichiro Tamura, Masatoshi Nei, and Sudhir Kumar. “Prospects for inferring very large phylogenies by using the neighbor-joining method”. In: *Proceedings of the National Academy of Sciences* 101.30 (2004), pp. 11030–11035.

[151] T Stefan Van Dongen and Birgitta Winnepenninckx. “Multiple UPGMA and neighbor-joining trees and the performance of some computer packages”. In: *Mol. Biol. Evol* 13.2 (1996), pp. 309–313.

[152] Ziheng Yang. *Phylogenetic analysis by maximum likelihood (PAML)*. 2000.

[153] Ziheng Yang and Tianqi Zhu. "Bayesian selection of misspecified models is overconfident and may cause spurious posterior probabilities for phylogenetic trees". In: *Proceedings of the National Academy of Sciences* 115.8 (2018), pp. 1854–1859.

[154] Guohua Jin et al. "Maximum likelihood of phylogenetic networks". In: *Bioinformatics* 22.21 (2006), pp. 2604–2611.

[155] John P Huelsenbeck and Bruce Rannala. "Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models". In: *Systematic biology* 53.6 (2004), pp. 904–913.

[156] F James Rohlf et al. "Accuracy of estimated phylogenies: effects of tree topology and evolutionary model". In: *Evolution* 44.6 (1990), pp. 1671–1684.

[157] Fabrícia F Nascimento, Mario dos Reis, and Ziheng Yang. "A biologists guide to Bayesian phylogenetic analysis". In: *Nature ecology & evolution* 1.10 (2017), pp. 1446–1454.

[158] William R Pearson, Gabriel Robins, and Tongtong Zhang. "Generalized neighbor-joining: more reliable phylogenetic tree reconstruction." In: *Molecular Biology and Evolution* 16.6 (1999), pp. 806–816.

[159] John J Wiens and Matthew C Morrill. "Missing data in phylogenetic analysis: reconciling results from simulations and empirical data". In: *Systematic biology* 60.5 (2011), pp. 719–731.

[160] John J Wiens. "Missing data, incomplete taxa, and phylogenetic accuracy". In: *Systematic biology* 52.4 (2003), pp. 528–538.

[161] John J Wiens. "Missing data and the accuracy of Bayesian phylogenetics". In: *Journal of Systematics and Evolution* 46.3 (2008), p. 307.

[162] John J Wiens. "Missing data and the design of phylogenetic analyses". In: *Journal of biomedical informatics* 39.1 (2006), pp. 34–42.

[163] Guilherme P Coelho, Ana Estela A da Silva, and Fernando J Von Zuben. "Evolving phylogenetic trees: A multiobjective approach". In: *Brazilian Symposium on Bioinformatics*. Springer. 2007, pp. 113–125.

[164] Michael J Sanderson and Amy C Driskell. "The challenge of constructing large phylogenetic trees". In: *Trends in plant science* 8.8 (2003), pp. 374–379.

[165] Deepack Jakhar and Ishmeet Kaur. "Artificial intelligence, machine learning and deep learning: definitions and differences". In: *Clinical and experimental dermatology* 45.1 (2020), pp. 131–132.

[166] Prerna Singh. "Systematic review of data-centric approaches in artificial intelligence and machine learning". In: *Data Science and Management* 6.3 (2023), pp. 144–157.

[167] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. "Understanding of a convolutional neural network". In: *2017 international conference on engineering and technology (ICET)*. IEEE. 2017, pp. 1–6.

[168] Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. "Multi-dimensional recurrent neural networks". In: *International conference on artificial neural networks*. Springer. 2007, pp. 549–558.

[169] KR Chowdhary. "Natural language processing". In: *Fundamentals of artificial intelligence* (2020), pp. 603–649.

[170] Xu Wang, Chunyang Chen, and Zhenchang Xing. "Domain-specific machine translation with recurrent neural network for software localization". In: *Empirical Software Engineering* 24.6 (2019), pp. 3514–3545.

[171] Yong Yu et al. "A review of recurrent neural networks: LSTM cells and network architectures". In: *Neural computation* 31.7 (2019), pp. 1235–1270.

[172] Junyoung Chung et al. "Empirical evaluation of gated recurrent neural networks on sequence modeling". In: *arXiv preprint arXiv:1412.3555* (2014).

[173] Alex Sherstinsky. "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network". In: *Physica D: Nonlinear Phenomena* 404 (2020), p. 132306.

[174] Dor Bank, Noam Koenigstein, and Raja Giryes. "Autoencoders". In: *Machine learning for data science handbook: data mining and knowledge discovery handbook* (2023), pp. 353–374.

[175] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[176] M Durgadevi et al. "Generative Adversarial Network (GAN): A general review on different variants of GAN and applications". In: *2021 6th International Conference on Communication and Electronics Systems (ICCES)*. IEEE. 2021, pp. 1–8.

[177] Alankrita Aggarwal, Mamta Mittal, and Gopi Battineni. "Generative adversarial network: An overview of theory and applications". In: *International Journal of Information Management Data Insights* 1.1 (2021), p. 100004.

[178] Lucas Pinheiro Cinelli et al. "Variational autoencoder". In: *Variational Methods for Machine Learning with Applications to Deep Networks*. Springer, 2021, pp. 111–149.

[179] Yu Lu et al. "Attention analysis and calibration for transformer in natural language generation". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), pp. 1927–1938.

[180] Laith Alzubaidi et al. "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions". In: *Journal of big Data* 8 (2021), pp. 1–74.

[181] Heehoon Kim et al. "Performance analysis of CNN frameworks for GPUs". In: *2017 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*. IEEE. 2017, pp. 55–64.

[182] Adam Stani, Vedran Vyroubal, and Vedran Slijepevi. "Classification efficiency of pre-trained deep CNN models on camera trap images". In: *Journal of imaging* 8.2 (2022), p. 20.

[183] Jaya Gupta, Sunil Pathak, and Gireesh Kumar. "Deep learning (CNN) and transfer learning: A review". In: *Journal of Physics: Conference Series*. Vol. 2273. 1. IOP Publishing. 2022, p. 012029.

[184] Quan-shi Zhang and Song-Chun Zhu. "Visual interpretability for deep learning: a survey". In: *Frontiers of Information Technology & Electronic Engineering* 19.1 (2018), pp. 27–39.

[185] Hamed Habibi Aghdam, Elnaz Jahani Heravi, et al. "Guide to convolutional neural networks". In: *New York, NY: Springer* 10.978-973 (2017), p. 51.

[186] Walter Hugo Lopez Pinaya et al. "Convolutional neural networks". In: *Machine learning*. Elsevier, 2020, pp. 173–191.

[187] Rajalingappaa Shanmugamani. *Deep Learning for Computer Vision: Expert techniques to train advanced neural networks using TensorFlow and Keras*. Packt Publishing Ltd, 2018.

[188] Iffat Zafar et al. *Hands-on convolutional neural networks with TensorFlow: Solve computer vision problems with modeling in TensorFlow and Python*. Packt Publishing Ltd, 2018.

[189] Princeton University Stanford Vision Lab Stanford University. *Imagenet*. Accessed: 2024-04-02. 2024. URL: `https://www.image-net.org/`.

[190] Palash Ghosal et al. "Brain tumor classification using ResNet-101 based squeeze and excitation deep neural network". In: *2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP)*. IEEE. 2019, pp. 1–6.

[191] Kaiming He et al. "Identity mappings in deep residual networks". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer. 2016, pp. 630–645.

[192] Kaiming He et al. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[193] Xiaoling Xia, Cui Xu, and Bing Nan. "Inception-v3 for flower classification". In: *2017 2nd international conference on image, vision and computing (ICIVC)*. IEEE. 2017, pp. 783–787.

[194] Cheng Wang et al. "Pulmonary image classification based on inception-v3 transfer learning model". In: *IEEE Access* 7 (2019), pp. 146533–146541.

[195] Christian Szegedy et al. "Rethinking the inception architecture for computer vision". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826.

[196] Google. *Inception V3 Advanced Guide*. Accessed: 2023-06-05. 2023. URL: https://cloud.google.com/tpu/docs/inception-v3-advanced?hl=es-419.

[197] Srikanth Tammina. "Transfer learning using VGG-16 with deep convolutional neural network for classifying images". In: *International Journal of Scientific and Research Publications (IJSRP)* 9.10 (2019), pp. 143–150.

[198] Qing Guan et al. "Deep convolutional neural network VGG-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: a pilot study". In: *Journal of Cancer* 10.20 (2019), p. 4876.

[199] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[200] Sinno Jialin Pan and Qiang Yang. "A survey on transfer learning". In: *IEEE Transactions on knowledge and data engineering* 22.10 (2009), pp. 1345–1359.

[201] Riaan Zoetmulder et al. "Domain-and task-specific transfer learning for medical segmentation tasks". In: *Computer Methods and Programs in Biomedicine* 214 (2022), p. 106539.

[202] Sinno Jialin Pan. "Transfer learning". In: *Learning* 21 (2020), pp. 1–2.

[203] Jochen Garcke and Thomas Vanck. "Importance weighted inductive transfer learning for regression". In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I 14*. Springer. 2014, pp. 466–481.

[204] Andrew Arnold, Ramesh Nallapati, and William W Cohen. "A comparative study of methods for transductive transfer learning". In: *Seventh IEEE international conference on data mining workshops (ICDMW 2007)*. IEEE. 2007, pp. 77–82.

[205] Hang Chang et al. "Unsupervised transfer learning via multi-scale convolutional sparse coding for biomedical applications". In: *IEEE transactions on pattern analysis and machine intelligence* 40.5 (2017), pp. 1182–1194.

[206] Yunhun Jang et al. "Learning what and where to transfer". In: *International conference on machine learning*. PMLR. 2019, pp. 3030–3039.

[207] Mohammadreza Iman, Hamid Reza Arabnia, and Khaled Rasheed. "A review of deep transfer learning and recent advancements". In: *Technologies* 11.2 (2023), p. 40.

[208] Cristian David Rios-Urrego et al. "Transfer learning to detect parkinsons disease from speech in different languages using convolutional neural networks with layer freezing". In: *International Conference on Text, Speech, and Dialogue*. Springer. 2020, pp. 331–339.

[209] Pooyan Jamshidi et al. "Transfer learning for improving model predictions in highly configurable software". In: *2017 IEEE/ACM 12th International Symposium on Software Engineering for Adaptive and Self-Managing Systems (SEAMS)*. IEEE. 2017, pp. 31–41.

[210] Weihua Li et al. "A perspective survey on deep transfer learning for fault diagnosis in industrial scenarios: Theories, applications and challenges". In: *Mechanical Systems and Signal Processing* 167 (2022), p. 108487.

[211] Yang Tan et al. "Transferability-guided cross-domain cross-task transfer learning". In: *IEEE Transactions on Neural Networks and Learning Systems* (2024).

[212] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. "A survey of transfer learning". In: *Journal of Big data* 3 (2016), pp. 1–40.

[213] Thomas Mensink et al. "Factors of influence for transfer learning across diverse appearance domains and task types". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.12 (2021), pp. 9298–9314.

[214] Hee E Kim et al. "Transfer learning for medical image classification: a literature review". In: *BMC medical imaging* 22.1 (2022), p. 69.

[215] Manali Shaha and Meenakshi Pawar. "Transfer learning for image classification". In: *2018 second international conference on electronics, communication and aerospace technology (ICECA)*. IEEE. 2018, pp. 656–660.

[216] Rajdeep Kaur, Rakesh Kumar, and Meenu Gupta. "Review on Transfer Learning for Convolutional Neural Network". In: *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*. IEEE. 2021, pp. 922–926.

[217] Sebastian Ruder et al. "Transfer learning in natural language processing". In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Tutorials*. 2019, pp. 15–18.

[218] Jasmin Bharadiya. "Transfer Learning in Natural Language Processing (NLP)". In: *European Journal of Technology* 7.2 (2023), pp. 26–35.

[219] Dong Wang and Thomas Fang Zheng. "Transfer learning for speech and language processing". In: *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE. 2015, pp. 1225–1237.

[220] Maurice Gerczuk et al. "Emonet: A transfer learning framework for multi-corpus speech emotion recognition". In: *IEEE Transactions on Affective Computing* 14.2 (2021), pp. 1472–1487.

[221] Ning Liu et al. "A transfer learning method for detecting Alzheimer's disease based on speech and natural language processing". In: *Frontiers in Public Health* 10 (2022), p. 772592.

[222] Cynthia Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature machine intelligence* 1.5 (2019), pp. 206–215.

[223] Quan-shi Zhang and Song-Chun Zhu. "Visual interpretability for deep learning: a survey". In: *Frontiers of Information Technology & Electronic Engineering* 19.1 (2018), pp. 27–39.

[224] Chris Olah et al. "The building blocks of interpretability". In: *Distill* 3.3 (2018), e10.

[225] Ramprasaath R Selvaraju et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.

[226] Vicnesh Jahmunah et al. "Explainable detection of myocardial infarction using deep learning models with Grad-CAM technique on ECG signals". In: *Computers in Biology and Medicine* 146 (2022), p. 105550.

[227] Mehmet Akif Gulum, Christopher M Trombley, and Mehmed Kantardzic. "Improved deep learning explanations for prostate lesion classification through grad-CAM and saliency map fusion". In: *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE. 2021, pp. 498–502.

[228] Daniel Dworak and Jerzy Baranowski. "Adaptation of Grad-CAM method to neural network architecture for LiDAR pointcloud object detection". In: *Energies* 15.13 (2022), p. 4681.

[229] Shuihua Wang and Yudong Zhang. "Grad-CAM: understanding AI models". In: *Comput. Mater. Contin* 76.2 (2023), pp. 1321–1324.

[230] Aditya Bhattacharya. *Applied Machine Learning Explainability Techniques: Make ML models explainable and trustworthy for practical applications using LIME, SHAP, and more*. Packt Publishing Ltd, 2022.

[231] Nicholas Hamilton et al. "Enhancing Visualization and Explainability of Computer Vision Models with Local Interpretable Model-Agnostic Explanations (LIME)". In: *2022 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE. 2022, pp. 604–611.

[232] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[233] Damien Garreau and Ulrike Luxburg. "Explaining the explainer: A first theoretical analysis of LIME". In: *International conference on artificial intelligence and statistics*. PMLR. 2020, pp. 1287–1296.

[234] Jaeyoung Shin. "Feasibility of local interpretable model-agnostic explanations (LIME) algorithm as an effective and interpretable feature selection method: comparative fNIRS study". In: *Biomedical Engineering Letters* 13.4 (2023), pp. 689–703.

[235] Shir Gur, Ameen Ali, and Lior Wolf. "Visualization of supervised and self-supervised neural networks via attribution guided factorization". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. 13. 2021, pp. 11545–11554.

[236] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps". In: *arXiv preprint arXiv:1312.6034* (2013).

[237] Jost Tobias Springenberg et al. "Striving for simplicity: The all convolutional net". In: *arXiv preprint arXiv:1412.6806* (2014).

[238] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks". In: *International conference on machine learning*. PMLR. 2017, pp. 3319–3328.

[239] Julius Adebayo et al. "Sanity checks for saliency maps". In: *Advances in neural information processing systems* 31 (2018).

[240] Alessandro Liberati et al. "The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration". In: *Annals of internal medicine* 151.4 (2009), W–65.

[241] Daniele Ravì et al. "Deep learning for health informatics". In: *IEEE Journal of Biomedical and Health Informatics* 21.1 (2016), pp. 4–21.

[242] Longhao Yuan and Jianting Cao. "Patients EEG data analysis via spectrogram image with a Convolution Neural Network". In: *International Conference on Intelligent Decision Technologies*. Springer. 2017, pp. 13–21.

[243] Kaleem Nawaz Khan et al. "Deep Learning based classification of unsegmented phonocardiogram spectrograms leveraging transfer learning". In: *Physiological Measurement* 42.9 (2021), p. 095003.

[244] Giulio Ruffini et al. "Deep Learning with EEG spectrograms in rapid eye movement behavior disorder". In: *Frontiers in neurology* 10 (2019), p. 806.

[245] Wenjie Zhang, Jiqing Han, and Shiwen Deng. "Heart sound classification based on scaled spectrogram and tensor decomposition". In: *Expert Systems with Applications* 84 (2017), pp. 220–231.

[246] Menyun Qiao et al. "Breast calcification detection based on multichannel radiofrequency signals via a unified Deep Learning framework". In: *Expert Systems with Applications* 168 (2021), p. 114218.

[247] Maryam Etemadi et al. "A systematic review of healthcare recommender systems: Open issues, challenges, and techniques". In: *Expert Systems with Applications* (2022), p. 118823.

[248] JS Sahambi, SN Tandon, and RKP Bhatt. "Using wavelet transforms for ECG characterization. An on-line digital signal processing system". In: *IEEE Engineering in Medicine and Biology Magazine* 16.1 (1997), pp. 77–83.

[249] Jingshan Huang et al. “ECG arrhythmia classification using STFT-based spectrogram and convolutional neural network”. In: *IEEE access* 7 (2019), pp. 92871–92880.

[250] Laurence Keselbrener and Solange Akselrod. “Selective discrete Fourier transform algorithm for time-frequency analysis: method and application on simulated and cardiovascular signals”. In: *IEEE Transactions on Biomedical Engineering* 43.8 (1996), pp. 789–802.

[251] Amit Singhal et al. “An efficient removal of power-line interference and baseline wander from ECG signals by employing Fourier decomposition technique”. In: *Biomedical Signal Processing and Control* 57 (2020), p. 101741.

[252] Binish Fatimah et al. “Detection of apnea events from ECG segments using Fourier decomposition method”. In: *Biomedical Signal Processing and Control* 61 (2020), p. 102005.

[253] Zheng Chen et al. “IVAE: An improved deep learning structure for eeg signal characterization and reconstruction”. In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2020, pp. 1909–1913.

[254] Ahmed Fares, Shenghua Zhong, and Jianmin Jiang. “Region level bi-directional deep learning framework for eeg-based image classification”. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 368–373.

[255] Markos G Tsipouras. “Spectral information of EEG signals with respect to epilepsy classification”. In: *EURASIP Journal on Advances in Signal Processing* 2019.1 (2019), pp. 1–17.

[256] Virender Kumar Mehla et al. “An efficient method for identification of epileptic seizures from EEG signals using Fourier analysis”. In: *Physical and Engineering Sciences in Medicine* 44 (2021), pp. 443–456.

[257] Kaveh Samiee, Peter Kovacs, and Moncef Gabbouj. “Epileptic seizure classification of EEG time-series using rational discrete short-time Fourier transform”. In: *IEEE transactions on Biomedical Engineering* 62.2 (2014), pp. 541–552.

[258] Róbert Bódizs et al. “A set of composite, non-redundant EEG measures of NREM sleep based on the power law scaling of the Fourier spectrum”. In: *Scientific reports* 11.1 (2021), p. 2041.

[259] Fuxu Wang et al. “Analysis for early seizure detection system based on deep learning algorithm”. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 2382–2389.

[260] Eugeny Denisovich Adamovich and Oleg Valer’evich Gradov. “Joint Fourier and non-Fourier spectral/pseudo-spectral approach to the lung bioacoustics and biomedical signal fingerprinting as a way to increase the quality of the lung diagnostics using supercomplex hybridization of different DSP methods”. In: *Journal of Biomedical Technologies* 1 (2015), pp. 34–37.

[261] Zeenat Tariq, Sayed Khushal Shah, and Yugyung Lee. "Multimodal lung disease classification using deep convolutional neural network". In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2020, pp. 2530–2537.

[262] Weijie Wang et al. "Revisiting signal processing with spectrogram analysis on EEG, ECG and speech signals". In: *Future Generation Computer Systems* 98 (2019), pp. 227–232.

[263] Mawia Ahmed Hassan and Yasser Mostafa Kadah. "Digital signal processing methodologies for conventional digital medical ultrasound imaging system". In: *American journal of biomedical engineering* 3.1 (2013), pp. 14–30.

[264] Chaoyu Chen et al. "FFPN: Fourier Feature Pyramid Network for Ultrasound Image Segmentation". In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2023, pp. 166–175.

[265] Mengqiang Zhou et al. "The segmentation effect of style transfer on fetal head ultrasound image: a study of multi-source data". In: *Medical & Biological Engineering & Computing* 61.5 (2023), pp. 1017–1031.

[266] Stefano Ricci et al. "Embedded system for real-time digital processing of medical ultrasound Doppler signals". In: *EURASIP Journal on Advances in Signal Processing* 2008 (2008), pp. 1–7.

[267] Louis-Gilles Durand and Philippe Pibarot. "Digital signal processing of the phonocardiogram: review of the most recent advancements". In: *Critical Reviews in biomedical engineering* 23.3-4 (1995).

[268] Shivam Varshney and Satyendra Singh. "Computation of Biological Murmurs in Phonocardiogram Signals using Fast Fourier & Discrete Wavelet Transform". In: *2020 International Conference on Computation, Automation and Knowledge Management (ICCAKM)*. IEEE. 2020, pp. 234–240.

[269] Mahtab Mehrabbeik et al. "Phonocardiography-based mitral valve prolapse detection with using fractional fourier transform". In: *Biomedical Physics & Engineering Express* 7.1 (2020), p. 015003.

[270] Mariko Izumo et al. "Quantitative analyses of circadian gene expression in mammalian cell cultures". In: *PLoS computational biology* 2.10 (2006), e136.

[271] Makoto Akashi et al. "Noninvasive method for assessing the human circadian clock using hair follicle cells". In: *Proceedings of the National Academy of Sciences* 107.35 (2010), pp. 15643–15648.

[272] Nabil Alshurafa et al. "Non-invasive monitoring of eating behavior using spectrogram analysis in a wearable necklace". In: *2014 IEEE healthcare innovation conference (HIC)*. IEEE. 2014, pp. 71–74.

[273] Andrei A Bunaciu, Serban Fleschin, and Hassan Y Aboul-Enein. "Biomedical investigations using Fourier transform-infrared microspectroscopy". In: *Critical Reviews in Analytical Chemistry* 44.3 (2014), pp. 270–276.

[274] R Michael Gendreau. "Biomedical Fourier transform infrared spectroscopy: applications to proteins". In: *Spectroscopy in the Biomedical Sciences*. CRC Press, 2018, pp. 21–52.

[275] EJ Castillo et al. "Surface analysis of biomedical polymers by attenuated total reflectance-Fourier transform infra-red". In: *Biomaterials* 5.4 (1984), pp. 186–193.

[276] Andrew M Tudor et al. "The analysis of biomedical hydroxyapatite powders and hydroxyapatite coatings on metallic medical implants by near-IR Fourier transform Raman spectroscopy". In: *Spectrochimica Acta Part A: Molecular Spectroscopy* 49.5-6 (1993), pp. 675–680.

[277] B Lacroix, Jean Pierre Huvenne, and Marc Deveaux. "Gas chromatography with Fourier transform infrared spectrometry for biomedical applications". In: *Journal of Chromatography B: Biomedical Sciences and Applications* 492 (1989), pp. 109–136.

[278] Shaun Pacheco, Guoan Zheng, and Rongguang Liang. "Reflective Fourier ptychography". In: *Journal of Biomedical Optics* 21.2 (2016), pp. 026010–026010.

[279] KL Andrew Chan and Sergei G Kazarian. "Attenuated total reflection Fourier-transform infrared (ATR-FTIR) imaging of tissues and live cells". In: *Chemical Society Reviews* 45.7 (2016), pp. 1850–1864.

[280] An Pan, Chao Zuo, and Baoli Yao. "High-resolution and large field-of-view Fourier ptychographic microscopy and its applications in biomedicine". In: *Reports on Progress in Physics* 83.9 (2020), p. 096101.

[281] Kar-Yan Su and Wai-Leng Lee. "Fourier transform infrared spectroscopy as a cancer screening and diagnostic tool: a review and prospects". In: *Cancers* 12.1 (2020), p. 115.

[282] Johannes F De Boer, Rainer Leitgeb, and Maciej Wojtkowski. "Twenty-five years of optical coherence tomography: the paradigm shift in sensitivity and speed provided by Fourier domain OCT". In: *Biomedical optics express* 8.7 (2017), pp. 3248–3280.

[283] Maciej Wojtkowski et al. "In vivo human retinal imaging by Fourier domain optical coherence tomography". In: *Journal of biomedical optics* 7.3 (2002), pp. 457–463.

[284] Bernhard Baumann et al. "Total retinal blood flow measurement with ultrahigh speed swept source/Fourier domain OCT". In: *Biomedical optics express* 2.6 (2011), pp. 1539–1552.

[285] Yimin Wang et al. "In vivo total retinal blood flow measurement by Fourier domain Doppler optical coherence tomography". In: *Journal of biomedical optics* 12.4 (2007), pp. 041215–041215.

[286] Donald P Land et al. "Laser-induced thermal desorption with Fourier transform mass spectrometry for the time resolved analysis of catalyst and biomedical implant model surface molecular composition". In: *Analytica chimica acta* 307.2-3 (1995), pp. 321–331.

[287] TV Lakshmi et al. "Design of a state-machine based genomic simulator and development of a system for prediction of Rheumatoid Arthritis (RA) using signal processing techniques". In: *2017 International Conference on Circuits, Controls, and Communications (CCUBE)*. IEEE. 2017, pp. 125–130.

[288] Paul Dan Cristea and Rodica Tuduce. "Signal processing of genomic information: Mitochondrial genomic signals of hominidae". In: *Proceedings EC-VIP-MC 2003. 4th EURASIP Conference focused on Video/Image Processing and Multimedia Communications (IEEE Cat. No. 03EX667)*. Vol. 1. IEEE. 2003, pp. 209–214.

[289] Mahmood Akhtar, Eliathamby Ambikairajah, and Julien Epps. "Detection of period-3 behavior in genomic sequences using singular value decomposition". In: *Proceedings of the IEEE Symposium on Emerging Technologies, 2005.* IEEE. 2005, pp. 13–17.

[290] Juan V Lorenzo-Ginori et al. "Digital signal processing in the analysis of genomic sequences". In: *Current Bioinformatics* 4.1 (2009), pp. 28–40.

[291] Edward R Dougherty. *Genomic signal processing and statistics*. Vol. 2. Hindawi Publishing Corporation, 2005.

[292] Palghat Vaidyanathan. "Genomics and proteomics: A signal processor's tour". In: *IEEE circuits and systems magazine* 4.4 (2004), pp. 6–29.

[293] Ernesto Borrayo et al. "Genomic signal processing methods for computation of alignment-free distances from DNA sequences". In: *PloS one* 9.11 (2014), e110954.

[294] Rabeb Touati et al. "Comparative genomic signature representations of the emerging COVID-19 coronavirus and other coronaviruses: High identity and possible recombination between Bat and Pangolin coronaviruses". In: *Genomics* 112.6 (2020), pp. 4189–4202.

[295] Tung Hoang, Changchuan Yin, and Stephen S-T Yau. "Numerical encoding of DNA sequences by chaos game representation with application in similarity comparison". In: *Genomics* 108.3-4 (2016), pp. 134–142.

[296] Swarna Bai Arniker and Hon Keung Kwan. "Graphical representation of DNA sequences". In: *2009 IEEE International Conference on Electro/Information Technology*. IEEE. 2009, pp. 311–314.

[297] Changchuan Yin and Stephen S-T Yau. "Numerical representation of DNA sequences based on genetic code context and its applications in periodicity analysis of genomes". In: *2008 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*. IEEE. 2008, pp. 223–227.

[298] Paul Dan Cristea, Rodica Tuduce, and Jan Cornelis. "Signal Analysis of Pathogens' Genomic Sequences". In: *2007 Frontiers in the Convergence of Bioscience and Information Technologies*. IEEE. 2007, pp. 245–250.

[299] Dimitris Grigoriadis et al. "DiS-TSS: An annotation agnostic algorithm for TSS identification". In: *Bioinformatics and Biomedical Engineering: 8th International Work-Conference, IWBBIO 2020, Granada, Spain, May 6–8, 2020, Proceedings 8*. Springer. 2020, pp. 613–623.

[300] Matthew J Berryman et al. "Signal processing and statistical methods in analysis of text and DNA". In: *Biomedical Applications of Micro-and Nanoengineering*. Vol. 4937. SPIE. 2002, pp. 231–240.

[301] Imen Messaoudi, Afef Elloumi Oueslati, and Zied Lachiri. "2D DNA representations generated using a new coding and the time-frequency analysis". In: *Journal of Medical Imaging and Health Informatics* 5.5 (2015), pp. 1035–1044.

[302] Muthulakshmi Murugaiah and Murugeswari Ganesan. "A novel frequency based feature extraction technique for classification of corona virus genome and discovery of COVID-19 repeat pattern". In: *Brazilian Archives of Biology and Technology* 64 (2022).

[303] Ning Yu et al. "Signalign: An ontology of DNA as signal for comparative gene structure prediction using information-coding-and-processing techniques". In: *IEEE transactions on nanobioscience* 15.2 (2016), pp. 119–130.

[304] PD Cristea and R Tuduce. "Nucleotide Genomic Signal DNA Analysis". In: *5th European Conference of the International Federation for Medical and Biological Engineering: 14–18 September 2011, Budapest, Hungary*. Springer. 2012, pp. 1374–1377.

[305] Paul Dan Cristea. "Genetic signal representation and analysis". In: *Functional Monitoring and Drug-Tissue Interaction*. Vol. 4623. SPIE. 2002, pp. 77–84.

[306] John A Berger et al. "Visualization and analysis of DNA sequences using DNA walks". In: *Journal of the Franklin Institute* 341.1-2 (2004), pp. 37–53.

[307] Xuan Xiao et al. "Digital coding for amino acid based on cellular automata". In: *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*. Vol. 5. IEEE. 2004, pp. 4593–4598.

[308] Paul Dan Cristea, Dorina Banica, and Rodica Tuduce. "Genomic signal analysis of Mycobacterium tuberculosis". In: *Nanoscale Imaging, Spectroscopy, Sensing, and Actuation for Biomedical Applications IV*. Vol. 6447. SPIE. 2007, pp. 51–59.

[309] Robin Jugas et al. "Cross-correlation based detection of contigs overlaps". In: *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE. 2018, pp. 0155–0158.

[310] Dennis J Hazelett, Daniel L Lakeland, and Joseph B Weiss. "Affinity Density: a novel genomic approach to the identification of transcription factor regulatory targets". In: *Bioinformatics* 25.13 (2009), pp. 1617–1624.

[311] Peijie Zheng et al. "A capsule network-based method for identifying transcription factors". In: *Frontiers in Microbiology* 13 (2022), p. 1048478.

[312] Robert SH Istepanian, Ala Sungoor, and Jean-Christophe Nebel. "Comparative analysis of genomic signal processing for microarray data clustering". In: *IEEE transactions on nanobioscience* 10.4 (2011), pp. 225–238.

[313] Hong Yan and Tuan D Pham. "Spectral estimation techniques for DNA sequence and microarray data analysis". In: *Current Bioinformatics* 2.2 (2007), pp. 145–156.

[314] Yu-Xi Pan et al. "Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach". In: *Journal of protein chemistry* 22 (2003), pp. 395–402.

[315] Bonnie Ngai-Fong Law. "Application of signal processing for DNA sequence compression". In: *IET Signal Processing* 13.6 (2019), pp. 569–580.

[316] Paul Cristea. "Real and complex genomic signals". In: *2002 14th International Conference on Digital Signal Processing Proceedings. DSP 2002 (Cat. No. 02TH8628)*. Vol. 2. IEEE. 2002, pp. 543–546.

[317] Paul Dan Cristea. "Large scale features in DNA genomic signals". In: *Signal Processing* 83.4 (2003), pp. 871–888.

[318] Dawit Nigatu et al. "Relating digital information, thermodynamic stability, and classes of functional genes in E. coli". In: *2014 IEEE global conference on signal and information processing (GlobalSIP)*. IEEE. 2014, pp. 1338–1341.

[319] Dimitris Grigoriadis et al. "DeepTSS: multi-branch convolutional neural network for transcription start site identification from CAGE data". In: *BMC bioinformatics* 23.2 (2022), pp. 1–17.

[320] Imen Messaoudi, Afef Elloumi, and Zied Lachiri. "A Coding Technique Based on the Frequency Evolution Creates with a Time Frequency Analysis a New Genome's Landscape". In: *International Journal of Advanced Computer Science and Applications* 5.4 (2014).

[321] Yiru Zhao et al. "Uncovering the prognostic gene signatures for the improvement of risk stratification in cancers by using deep learning algorithm coupled with wavelet transform". In: *BMC bioinformatics* 21 (2020), pp. 1–24.

[322] Qian Zheng et al. "Gene Prediction by the Scale-limited Gabor Wavelet Transform for Identifying the Protein Coding Regions". In: *2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV)*. IEEE. 2020, pp. 1112–1116.

[323] Earl F Glynn, Jie Chen, and Arcady R Mushegian. "Detecting periodic patterns in unevenly spaced gene expression time series using Lomb–Scargle periodograms". In: *Bioinformatics* 22.3 (2006), pp. 310–316.

[324] Tao Meng et al. “Wavelet analysis in current cancer genome research: a survey”. In: *IEEE/ACM transactions on computational biology and bioinformatics* 10.6 (2013), pp. 1442–14359.

[325] John A Berger et al. “New approaches to genome sequence analysis based on digital signal processing”. In: *University of California* (2002).

[326] Xin-Yun Zhang et al. “Signal processing techniques in genomic engineering”. In: *Proceedings of the IEEE* 90.12 (2002), pp. 1822–1833.

[327] PP Vaidyanathan and Byung Jun Yoon. “The role of signal-processing concepts in genomics and proteomics”. In: *Journal of the Franklin Institute* 341.1-2 (2004), pp. 111–135.

[328] Rabia Aslam Khan, Muhammad Bilal Butt, and Sabreena Nawaz. “Genomic Signal Processing Methods in DNA Mapping Schemes for Prediction of Exon in a Gene Using Digital Filters”. In: *International Journal for Electronic Crime Investigation* 7.3 (2023), pp. 61–68.

[329] Emmanuel Adetiba, Oludayo O Olugbara, and Tunmike B Taiwo. “Identification of pathogenic viruses using genomic cepstral coefficients with radial basis function neural network”. In: *Advances in Nature and Biologically Inspired Computing: Proceedings of the 7th World Congress on Nature and Biologically Inspired Computing (NaBIC2015) in Pietermaritzburg, South Africa, held December 01-03, 2015*. Springer. 2016, pp. 281–291.

[330] Rabeb Touati et al. “Helitrons periodicities identification in C. Elegans based on the smoothed spectral analysis and the frequency Chaos game signal coding”. In: *International Journal of Advanced Computer Science and Applications* 9.4 (2018).

[331] Yashpal Yadav, Sanjeev Narayan Sharma, and Devendra Kumar Shakya. “Detection of tandem repeats in DNA sequences using short-time Ramanujan Fourier transform”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19.3 (2021), pp. 1583–1591.

[332] Changchuan Yin and Jiasong Wang. “Periodic power spectrum with applications in detection of latent periodicities in DNA sequences”. In: *Journal of mathematical biology* 73 (2016), pp. 1053–1079.

[333] Benjamin Simeon Harvey and Soo-Yeon Ji. “Cloud-scale genomic signals processing for robust large-scale cancer genomic microarray data analysis”. In: *IEEE journal of biomedical and health informatics* 21.1 (2015), pp. 238–245.

[334] Micah Thornton and Monnie McGee. “Non-parametric genomic fourier power spectra filter designs”. In: *2021 IEEE Biomedical Circuits and Systems Conference (BioCAS)*. IEEE. 2021, pp. 1–6.

[335] Amit Kumar Singh and Vinay Kumar Srivastava. “Bidirectional filtering approach for the improved protein coding region identification in eukaryotes”. In: *Network Modeling Analysis in Health Informatics and Bioinformatics* 11.1 (2022), p. 13.

[336] Helena Skutkova et al. "Classification of genomic signals using dynamic time warping". In: *BMC bioinformatics* 14.10 (2013), pp. 1–7.

[337] Karel Sedlar et al. "Prokaryotic DNA signal downsampling for fast whole genome comparison". In: *Information Technologies in Biomedicine, Volume 3*. Springer. 2014, pp. 373–383.

[338] Helena Skutkova et al. "Progressive alignment of genomic signals by multiple dynamic time warping". In: *Journal of theoretical biology* 385 (2015), pp. 20–30.

[339] Shreepriya Das and Haris Vikalo. "Optimal haplotype assembly with statistical pruning". In: *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE. 2014, pp. 1330–1333.

[340] Nini Rao and Simon Shepherd. "Detection of 3-periodicity for small genomic sequences based on AR technique". In: *2004 International Conference on Communications, Circuits and Systems (IEEE Cat. No. 04EX914)*. Vol. 2. IEEE. 2004, pp. 1032–1036.

[341] Dimitris Anastassiou. "Frequency-domain analysis of biomolecular sequences". In: *Bioinformatics* 16.12 (2000), pp. 1073–1081.

[342] Dimitris Anastassiou. "Genomic signal processing". In: *IEEE signal processing magazine* 18.4 (2001), pp. 8–20.

[343] Nevenka Dimitrova, Yee Him Cheung, and Michael Zhang. "Analysis and visualization of DNA spectrograms: open possibilities for the genome research". In: *Proceedings of the 14th ACM International Conference on Multimedia*. 2006, pp. 1017–1024.

[344] Dimitris Anastassiou. "DSP in genomics: processing and frequency-domain analysis of character strings". In: *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*. Vol. 2. IEEE. 2001, pp. 1053–1056.

[345] Alin Voina and Petre G Pop. "DNA PATTERNS LOCALIZATION USING SPECTROGRAMS". In: *Acta Technica Napocensis* 51.1 (2010), p. 27.

[346] Liping Du, Hongxia Zhou, and Hong Yan. "OMWSA: detection of DNA repeats using moving window spectral analysis". In: *Bioinformatics* 23.5 (2007), pp. 631–633.

[347] Imen Messaoudi, Afef Elloumi-Oueslati, and Zied Lachiri. "Building specific signals from frequency chaos game and revealing periodicities using a smoothed Fourier analysis". In: *IEEE/ACM transactions on computational biology and bioinformatics* 11.5 (2014), pp. 863–877.

[348] Xingyi Li et al. "Filter shaping for convolutional neural networks". In: *International Conference on Learning Representations*. 2016.

[349] Ren Zhang and Chun-Ting Zhang. "A brief review: The z-curve theory and its application in genome analysis". In: *Current genomics* 15.2 (2014), pp. 78–94.

[350] Hsin-Hsiung Huang and Senthil B Girimurugan. "A novel real-time genome comparison method using discrete wavelet transform". In: *Journal of computational biology* 25.4 (2018), pp. 405–416.

[351] Erik A Cheever, G Christian Overton, and David B Searls. "Fast Fourier transform-based correlation of DNA sequences using complex plane encoding". In: *Bioinformatics* 7.2 (1991), pp. 143–154.

[352] Sajid A Marhon and Stefan C Kremer. "Gene prediction based on DNA spectral analysis: a literature review". In: *Journal of computational biology* 18.4 (2011), pp. 639–676.

[353] Mohammad Al Bataineh and Zouhair Al-qudah. "A novel gene identification algorithm with Bayesian classification". In: *Biomedical Signal Processing and Control* 31 (2017), pp. 6–15.

[354] Petre G Pop and Eugen Lupu. "DNA repeats detection using BW spectrograms". In: *2008 IEEE International Conference on Automation, Quality and Testing, Robotics*. Vol. 3. IEEE. 2008, pp. 408–412.

[355] David Sussillo, Anshul Kundaje, and Dimitris Anastassiou. "Spectrogram analysis of genomes". In: *EURASIP Journal on Advances in Signal Processing* 2004.1 (2004), pp. 1–14.

[356] Giuseppe D'Avenio et al. "SWIFT (sequence-wide investigation with Fourier transform): a software tool for identifying proteins of a given class from the unannotated genome sequence". In: *Bioinformatics* 21.13 (2005), pp. 2943–2949.

[357] Subhajit Kar, Madhabi Ganguly, and Saptarshi Das. "Using DIT-FFT algorithm for identification of protein coding region in eukaryotic gene". In: *Biomedical Engineering: Applications, Basis and Communications* 31.01 (2019), p. 1950002.

[358] Mahendra Roy, Shyam Biswas, and Soma Barman. "Identification and analysis of coding and non-coding regions of a DNA sequence by positional frequency distribution of nucleotides (PFDN) algorithm". In: *2009 4th International Conference on Computers and Devices for Communication (CODEC)*. IEEE. 2009, pp. 1–4.

[359] Biju Issac et al. "Locating probable genes using Fourier transform approach". In: *Bioinformatics* 18.1 (2002), pp. 196–197.

[360] Jayanta Pal et al. "Use of 2D FFT and DTW in Protein Sequence Comparison". In: *The Protein Journal* (2023), pp. 1–11.

[361] Trevor W Fox and Alex Carreira. "A digital signal processing method for gene prediction with improved noise suppression". In: *EURASIP Journal on Advances in Signal Processing* 2004 (2004), pp. 1–7.

[362] S Barman et al. "Prediction of protein coding regions of a DNA sequence through spectral analysis". In: *2012 international conference on informatics, electronics & vision (ICIEV)*. IEEE. 2012, pp. 12–16.

[363] Vikrant Tomar, Dipesh Gandhi, and C Vijaykumar. "Digital signal processing for gene prediction". In: *TENCON 2008-2008 IEEE Region 10 Conference*. IEEE. 2008, pp. 1–5.

[364] Mahmood Akhtar, Julien Epps, and Eliathamby Ambikairajah. "Signal processing in sequence analysis: advances in eukaryotic gene prediction". In: *IEEE journal of selected topics in signal processing* 2.3 (2008), pp. 310–321.

[365] Afef Elloumi Oueslati, Zied Lachiri, and Noureddine Ellouze. "Spectral analysis of DNA sequence: The exon's location method". In: *2007 15th International Conference on Digital Signal Processing*. IEEE. 2007, pp. 115–118.

[366] Benjamin YM Kwan, Jennifer YY Kwan, and Hon Keung Kwan. "Spectral techniques for classifying short exon and intron sequences". In: *2012 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE. 2012, pp. 568–571.

[367] Wei-Feng Zhang and Hong Yan. "Exon prediction using empirical mode decomposition and Fourier transform of structural profiles of DNA sequences". In: *Pattern Recognition* 45.3 (2012), pp. 947–955.

[368] Mahmood Akhtar, Julien Epps, and Eliathamby Ambikairajah. "On DNA numerical representations for period-3 based exon prediction". In: *2007 IEEE international workshop on genomic signal processing and statistics*. IEEE. 2007, pp. 1–4.

[369] Hon Keung Kwan, Benjamin YM Kwan, and Jennifer YY Kwan. "Novel methodologies for spectral classification of exon and intron sequences". In: *EURASIP Journal on Advances in Signal Processing* 2012.1 (2012), pp. 1–14.

[370] Changchuan Yin. "Representation of DNA sequences in genetic codon context with applications in exon and intron prediction". In: *Journal of bioinformatics and computational biology* 13.02 (2015), p. 1550004.

[371] Rajasekhar Kakumani, Vijay Devabhaktuni, and M Omair Ahmad. "Prediction of protein-coding regions in DNA sequences using a model-based approach". In: *2008 IEEE International Symposium on Circuits and Systems*. IEEE. 2008, pp. 1918–1921.

[372] Mahmood Akhtar, Julien Epps, and Eliathamby Ambikairajah. "Time and frequency domain methods for gene and exon prediction in eukaryotes". In: *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*. Vol. 2. IEEE. 2007, pp. II–573.

[373] Parameswaran Ramachandran, Wu-Sheng Lu, and Andreas Antoniou. "Location of exons in DNA sequences using digital filters". In: *2009 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE. 2009, pp. 2337–2340.

[374] Amit Kumar Singh and Vinay Kumar Srivastava. "The three base periodicity of protein coding sequences and its application in exon prediction". In: *2020 7th international conference on signal processing and integrated networks (spin)*. IEEE. 2020, pp. 1089–1094.

[375] Mohammed Abo-Zahhai, Sabah M Ahmed, and Shimaa A Abd-Elrahman. "K11. A New Numerical Mapping Technique for Recognition of Exons and Introns in DNA Sequences". In: *2013 30th National Radio Science Conference (NRSC)*. IEEE. 2013, pp. 573–580.

[376] Fatma Akalin and Nejat Yumusak. "Classification of exon and intron regions obtained using digital signal processing techniques on the DNA genome sequencing with EfficientNetB7 architecture". In: *GUMMFD* 37.3 (2022), pp. 1355–1371.

[377] Alan L Rockwood et al. "Sequence alignment by cross-correlation". In: *Journal of biomolecular techniques: JBT* 16.4 (2005), p. 453.

[378] Mohamed Ameen et al. "Integrative single-cell analysis of cardiogenesis identifies developmental trajectories and non-coding mutations in congenital heart disease". In: *Cell* 185.26 (2022), pp. 4937–4953.

[379] Dong-Wei Liu et al. "Automated detection of cancerous genomic sequences using genomic signal processing and machine learning". In: *Future Generation Computer Systems* 98 (2019), pp. 233–237.

[380] Antara Ghosh and Soma Barman. "Application of BT and PC-BT in Homo sapiens gene prediction". In: *Microsystem Technologies* 22 (2016), pp. 2691–2705.

[381] Inbamalar Tharcis Mariapushpam and Sivakumar Rajagopal. "Improved algorithm for the location of CpG Islands in genomic sequences using discrete Wavelet transforms". In: *Current Bioinformatics* 12.1 (2017), pp. 57–65.

[382] Rida Assaf, Fangfang Xia, and Rick Stevens. "Identifying genomic islands with deep neural networks". In: *BMC genomics* 22.3 (2021), pp. 1–11.

[383] Hongxia Zhou, Liping Du, and Hong Yan. "Detection of tandem repeats in DNA sequences based on parametric spectral estimation". In: *IEEE transactions on information technology in biomedicine* 13.5 (2008), pp. 747–755.

[384] Petre G Pop and Alin Voina. "Numerical representations involved in DNA repeats detection using spectral analysis". In: *Studies in Informatics and Control* 20.2 (2011), pp. 163–180.

[385] Hongxia Zhou and Hong Yan. "Autoregressive models for spectral analysis of short tandem repeats in DNA sequences". In: *2006 IEEE International Conference on Systems, Man and Cybernetics*. Vol. 2. IEEE. 2006, pp. 1286–1290.

[386] Petre G Pop. "New representations in DNA repeats detection". In: *Proceedings of the 8th conference on Signal processing, computational geometry and artificial vision*. 2008, pp. 172–176.

[387] Brian Iaffaldano, Yingxiao Zhang, and Katrina Cornish. "CRISPR/Cas9 genome editing of rubber producing dandelion Taraxacum kok-saghyz using Agrobacterium rhizogenes without selection". In: *Industrial Crops and Products* 89 (2016), pp. 356–362.

[388] Victoria R Li, Zijun Zhang, and Olga G Troyanskaya. "CROTON: an automated and variant-aware deep learning framework for predicting CRISPR/-Cas9 editing outcomes". In: *Bioinformatics* 37.Supplement_1 (2021), pp. i342–i348.

[389] John HC Fong and Alan SL Wong. "Advancing CRISPR/Cas gene editing with machine learning". In: *Current Opinion in Biomedical Engineering* (2023), p. 100477.

[390] Petre G Pop. "Spectral analysis of alpha satellite DNA". In: *2009 International Symposium on Signals, Circuits and Systems*. IEEE. 2009, pp. 1–4.

[391] Petre G Pop. "Alpha satellite DNA analysis using spectrograms". In: *Proceedings of the 10th WSEAS international conference on Mathematics and computers in biology and chemistry*. 2009, pp. 98–101.

[392] Kazuharu Arakawa and Masaru Tomita. "The GC skew index: a measure of genomic compositional asymmetry and the degree of replicational selection". In: *Evolutionary Bioinformatics* 3 (2007), p. 117693430700300006.

[393] Dan Zhao et al. "Potential and whole-genome sequence-based mechanism of elongated-prismatic magnetite magnetosome formation in Acidithiobacillus ferrooxidans BYM". In: *World Journal of Microbiology and Biotechnology* 38.7 (2022), p. 121.

[394] Kazuharu Arakawa, Rintaro Saito, and Masaru Tomita. "Noise-reduction filtering for accurate detection of replication termini in bacterial genomes". In: *FEBS letters* 581.2 (2007), pp. 253–258.

[395] Sophie H Adjalley et al. "Genome-wide transcriptome profiling reveals functional networks involving the Plasmodium falciparum drug resistance transporters PfCRT and PfMDR1". In: *BMC genomics* 16 (2015), pp. 1–23.

[396] Kazuharu Arakawa, Haruo Suzuki, and Masaru Tomita. "Quantitative analysis of replication-related mutation and selection pressures in bacterial chromosomes and plasmids using generalised GC skew index". In: *BMC genomics* 10.1 (2009), pp. 1–10.

[397] Mahmood Akhtar, Eliathamby Ambikairajah, and Julien Epps. "Digital signal processing techniques for gene finding in eukaryotes". In: *Image and Signal Processing: 3rd International Conference, ICISP 2008. Cherbourg-Octeville, France, July 1-3, 2008. Proceedings 3*. Springer. 2008, pp. 144–152.

[398] Lee Sing Cheong, Feng Lin, and Hock Soon Seah. "Frequency-domain algorithms for visual analysis on genomic structures in prokaryotes". In: *International Conference on Computer Graphics, Imaging and Visualisation (CGIV'06)*. IEEE. 2006, pp. 96–103.

[399] Afef Elloumi Oueslati, Noureddine Ellouze, and Zied Lachiri. "3D spectrum analysis of DNA sequence: Application to Caenorhabditis elegans genome". In: *2007 IEEE 7th International Symposium on BioInformatics and BioEngineering*. IEEE. 2007, pp. 864–871.

[400] Manuel Llinas et al. "Comparative whole genome transcriptome analysis of three Plasmodium falciparum strains". In: *Nucleic acids research* 34.4 (2006), pp. 1166–1173.

[401] Zbynek Bozdech et al. "The transcriptome of the intraerythrocytic developmental cycle of Plasmodium falciparum". In: *PLoS biology* 1.1 (2003), e5.

[402] Xuewu Liu et al. "In-depth comparative analysis of malaria parasite genomes reveals protein-coding genes linked to human disease in Plasmodium falciparum genome". In: *BMC genomics* 19 (2018), pp. 1–15.

[403] Michael V D'Ambrosio and Ronald D Vale. "A whole genome RNAi screen of Drosophila S2 cell spreading performed using automated computational image analysis". In: *Journal of Cell Biology* 191.3 (2010), pp. 471–478.

[404] Shaozhen Xing et al. "Complete genome sequence of a novel, virulent Ahjdlikevirus bacteriophage that infects Enterococcus faecium". In: *Archives of virology* 162.12 (2017), pp. 3843–3847.

[405] Rui Dong et al. "A new method to cluster genomes based on cumulative Fourier power spectrum". In: *Gene* 673 (2018), pp. 239–250.

[406] Enrique Canessa. "Physics-Based Signal Analysis of Genome Sequences: An Overview of GenomeBits". In: *Microorganisms* 11.11 (2023), p. 2733.

[407] Kedar Patel et al. "An Non-Invasive Approach of Corona Genome Detection". In: *2020 International Conference on Advances in Computing, Communication & Materials (ICACCM)*. IEEE. 2020, pp. 154–157.

[408] Rabeb Touati et al. "Different genomic representations of novel pathogens base on signal processing algorithms: COVID-19 case study". In: (2022).

[409] Paul Dan Cristea. "Genomic signal analysis of pathogen variability". In: *Imaging, Manipulation, and Analysis of Biomolecules, Cells, and Tissues IV*. Vol. 6088. SPIE. 2006, pp. 191–202.

[410] Ryan F McCormick et al. "The Sorghum bicolor reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization". In: *The Plant Journal* 93.2 (2018), pp. 338–354.

[411] Seid Miad Zandavi et al. "Fotomics: fourier transform-based omics imagification for deep learning-based cell-identity mapping using single-cell omics profiles". In: *Artificial Intelligence Review* 56.7 (2023), pp. 7263–7278.

[412] Vladimira Kubicova and Ivo Provaznik. "Use of whole genome DNA spectrograms in bacterial classification". In: *Computers in Biology and Medicine* 69 (2016), pp. 298–307.

[413] Pardeep Garg and Sunil Datt Sharma. "Optimum Window based Modified Periodicity Spectrum Method for the Detection of Protein Coding Regions in DNA Sequences". In: *Digital Signal Processing* (2023), p. 104137.

[414] Evan Santo and Nevenka Dimitrova. "Improvement of spectral analysis as a genomic analysis tool". In: *2007 IEEE International Workshop on Genomic Signal Processing and Statistics*. IEEE. 2007, pp. 1–4.

[415] Hamidreza Saberkari et al. "Prediction of protein coding regions in DNA sequences using signal processing methods". In: *2012 IEEE Symposium on Industrial Electronics and Applications*. IEEE. 2012, pp. 355–360.

[416] Amit Kumar Singh and Vinay Kumar Srivastava. "Improved filtering approach for identification of protein-coding regions in eukaryotes by background noise reduction using S–G filter". In: *Network Modeling Analysis in Health Informatics and Bioinformatics* 10.1 (2021), p. 19.

[417] Saikat Singha Roy and Soma Barman. "Polyphase filtering with variable mapping rule in protein coding region prediction". In: *Microsystem Technologies* 23 (2017), pp. 4111–4121.

[418] Anca Bucur et al. "Alignment method for spectrograms of DNA sequences". In: *IEEE Transactions on Information Technology in Biomedicine* 14.1 (2009), pp. 3–9.

[419] A Bucur et al. "Frequency sorting method for spectral analysis of DNA sequences". In: *2008 IEEE International Conference on Bioinformatics and Biomedicine*. IEEE. 2008, pp. 43–50.

[420] Gerardo Mendizabal-Ruiz et al. "Genomic signal processing for DNA sequence clustering". In: *PeerJ* 6 (2018), e4264.

[421] Ning Yu, Zhihua Li, and Zeng Yu. "Survey on encoding schemes for genomic data representation and feature learningfrom signal processing to machine learning". In: *Big Data Mining and Analytics* 1.3 (2018), pp. 191–210.

[422] Gurjit S Randhawa et al. "Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study". In: *Plos one* 15.4 (2020), e0232391.

[423] Gurjit S Randhawa, Kathleen A Hill, and Lila Kari. "ML-DSP: Machine Learning with Digital Signal Processing for ultrafast, accurate, and scalable genome classification at all taxonomic levels". In: *BMC genomics* 20 (2019), pp. 1–21.

[424] Luísa C de Souza et al. "New proposal of viral genome representation applied in the classification of SARS-CoV-2 with deep learning". In: *BMC bioinformatics* 24.1 (2023), pp. 1–19.

[425] J Alejandro Morales et al. "Deep learning for the classification of genomic signals". In: *Mathematical Problems in Engineering* 2020 (2020).

[426] Bihter Das. "A deep learning model for identification of diabetes type 2 based on nucleotide signals". In: *Neural Computing and Applications* 34.15 (2022), pp. 12587–12599.

[427] Bihter Da, Suat Toraman, and brahim Türkolu. "A novel genome analysis method with the entropy-based numerical techniqueusing pretrained convolutional neural networks". In: *Turkish Journal of Electrical Engineering and Computer Sciences* 28.4 (2020), pp. 1932–1948.

[428] K Aditya Shastry and HA Sanjay. "Cancer diagnosis using artificial intelligence: A review". In: *Artificial Intelligence Review* (2022), pp. 1–33.

[429] Jia Xu et al. "Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives". In: *Human genetics* 138.2 (2019), pp. 109–124.

[430] Navodini Wijethilake et al. "Glioma survival analysis empowered with data engineeringa survey". In: *Ieee Access* 9 (2021), pp. 43168–43191.

[431] Xiao Tan et al. "Applying machine learning for integration of multi-modal genomics data and imaging data to quantify heterogeneity in tumour tissues". In: *Artificial neural networks* (2021), pp. 209–228.

[432] Alfonso Monaco et al. "A primer on machine learning techniques for genomic applications". In: *Computational and Structural Biotechnology Journal* 19 (2021), pp. 4345–4359.

[433] Md Ashrafuzzaman. "Artificial intelligence, machine learning and deep learning in ion channel bioinformatics". In: *Membranes* 11.9 (2021), p. 672.

[434] Jose Cleydson F Silva et al. "Machine learning approaches and their current application in plant molecular biology: A systematic review". In: *Plant Science* 284 (2019), pp. 37–47.

[435] Jing Lin and Kee Yuan Ngiam. "How data science and AI-based technologies impact genomics". In: *Singapore Medical Journal* 64.1 (2023), p. 59.

[436] Sema Atasever et al. "A comprehensive survey of deep learning research on medical image analysis with focus on transfer learning". In: *Clinical Imaging* 94 (2023), pp. 18–41.

[437] Andrew B Sellergren et al. "Simplified transfer learning for chest radiography models using less data". In: *Radiology* 305.2 (2022), pp. 454–465.

[438] Yuan-Cheng Wang et al. "Dynamic evolution of COVID-19 on chest computed tomography: experience from Jiangsu Province of China". In: *European radiology* 30 (2020), pp. 6194–6203.

[439] Tarek Khorshed, Mohamed N Moustafa, and Ahmed Rafea. "Learning & visualizing genomic signatures of cancer tumors using deep neural networks". In: *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2020, pp. 1–8.

[440] Lei Cui, Jun Feng, and Lin Yang. "Semi-supervised deep linear discriminant analysis for histopathology image classification". In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 2333–2337.

[441] Zhuang Ai et al. "FN-OCT: Disease detection algorithm for retinal optical coherence tomography based on a fusion network". In: *Frontiers in Neuroinformatics* 16 (2022), p. 876927.

[442] Mohammed Zidane et al. "A review on deep learning applications in highly multiplexed tissue imaging data analysis". In: *Frontiers in Bioinformatics* 3 (2023).

[443] Shahzad Ahmad Qureshi et al. "Intelligent ultra-light deep learning model for multi-class brain tumor detection". In: *Applied Sciences* 12.8 (2022), p. 3715.

[444] Gopal S Tandel, Ashish Tiwari, and OG Kakde. "Performance optimisation of deep learning models using majority voting algorithm for brain tumour classification". In: *Computers in Biology and Medicine* 135 (2021), p. 104564.

[445] Masoumeh Siar and Mohammad Teshnehlab. "A combination of feature extraction methods and deep learning for brain tumour classification". In: *IET Image Processing* 16.2 (2022), pp. 416–441.

[446] Dillip Ranjan Nayak et al. "Brain tumour classification using noble deep learning approach with parametric optimization through metaheuristics approaches". In: *Computers* 11.1 (2022), p. 10.

[447] Ramdas Vankdothu, Mohd Abdul Hameed, and Husnah Fatima. "A brain tumor identification and classification using deep learning based on CNN-LSTM method". In: *Computers and Electrical Engineering* 101 (2022), p. 107960.

[448] Sandra Steyaert et al. "Multimodal deep learning to predict prognosis in adult and pediatric brain tumors". In: *Communications Medicine* 3.1 (2023), p. 44.

[449] Liping Huang et al. "Rapid, label-free histopathological diagnosis of liver cancer based on Raman spectroscopy and deep learning". In: *Nature Communications* 14.1 (2023), p. 48.

[450] Simranjeet Randhawa et al. "Deep learning for liver tumour classification: enhanced loss function". In: *Multimedia Tools and Applications* 80 (2021), pp. 4729–4750.

[451] Sayan Saha Roy et al. "An automated liver tumour segmentation and classification model by deep learning based approaches". In: *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 11.3 (2023), pp. 638–650.

[452] Aya Hage Chehade et al. "Lung and colon cancer classification using medical imaging: A feature engineering approach". In: *Physical and Engineering Sciences in Medicine* 45.3 (2022), pp. 729–746.

[453] Yi Niu et al. "Predicting tumor mutational burden from lung adenocarcinoma histopathological images using deep learning". In: *Frontiers in Oncology* 12 (2022), p. 927426.

[454] Selene Tomassini et al. “On-cloud decision-support system for non-small cell lung cancer histology characterization from thorax computed tomography scans”. In: *Computerized Medical Imaging and Graphics* 110 (2023), p. 102310.

[455] Y Fujisawa et al. “Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis”. In: *British Journal of Dermatology* 180.2 (2019), pp. 373–381.

[456] Osamu Iizuka et al. “Deep learning models for histopathological classification of gastric and colonic epithelial tumours”. In: *Scientific reports* 10.1 (2020), p. 1504.

[457] Amina Ben Hamida et al. “Deep learning for colon cancer histopathological images analysis”. In: *Computers in Biology and Medicine* 136 (2021), p. 104730.

[458] Ole-Johan Skrede et al. “Deep learning for prediction of colorectal cancer outcome: a discovery and validation study”. In: *The Lancet* 395.10221 (2020), pp. 350–360.

[459] Hatim Amiji et al. *Deep learning tumor heterogeneity metric from histopathology images vs next generation sequencing-derived scores for colon cancer prognostication.* 2023.

[460] Ashwini Kodipalli et al. “An inception-ResNet deep learning approach to classify tumours in the ovary as benign and malignant”. In: *Expert Systems* (2022), e13215.

[461] Wessam M Salama, Azza M Elbagoury, and Moustafa H Aly. “Novel breast cancer classification framework based on deep learning”. In: *IET Image Processing* 14.13 (2020), pp. 3254–3259.

[462] Hanan Aljuaid et al. “Computer-aided diagnosis for breast cancer classification using deep neural networks and transfer learning”. In: *Computer Methods and Programs in Biomedicine* 223 (2022), p. 106951.

[463] Eren Tekin et al. “Tubule-U-Net: a novel dataset and deep learning-based tubule segmentation framework in whole slide images of breast cancer”. In: *Scientific Reports* 13.1 (2023), p. 128.

[464] Yu He et al. “Deep learning-based classification of primary bone tumors on radiographs: A preliminary study”. In: *EBioMedicine* 62 (2020).

[465] Fahdi Kanavati et al. “A deep learning model for cervical cancer screening on liquid-based cytology specimens in whole slide images”. In: *Cancers* 14.5 (2022), p. 1159.

[466] R Kavitha et al. “Ant colony optimization-enabled CNN deep learning technique for accurate detection of cervical cancer”. In: *BioMed Research International* 2023 (2023).

[467] Noor Badshah and Asif Ahmad. "ResBCU-Net: Deep learning approach for segmentation of skin images". In: *Biomedical Signal Processing and Control* 71 (2022), p. 103137.

[468] Amina Bibi et al. "Skin lesion segmentation and classification using conventional and deep learning based framework". In: *Comput. Mater. Contin* 71 (2022), pp. 2477–2495.

[469] Sarmad Maqsood and Robertas Damaeviius. "Multiclass skin lesion localization and classification using deep learning based features fusion and selection framework for smart healthcare". In: *Neural Networks* 160 (2023), pp. 238–258.

[470] Jitendra V Tembhurne et al. "Skin cancer detection using ensemble of machine learning and deep learning techniques". In: *Multimedia Tools and Applications* (2023), pp. 1–24.

[471] Wenhao Zhang et al. "A benchmark for automatic acral melanoma preliminary screening". In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 2829–2831.

[472] Constantine S Velmahos, Marcus Badgeley, and Ying-Chun Lo. "Using deep learning to identify bladder cancers with FGFR-activating mutations from histology images". In: *Cancer Medicine* 10.14 (2021), pp. 4805–4813.

[473] Sebastian Klein and Dan G Duda. "Machine learning for future subtyping of the tumor microenvironment of gastro-esophageal adenocarcinomas". In: *Cancers* 13.19 (2021), p. 4919.

[474] Kyubum Lee et al. "Deep learning of histopathology images at the single cell level". In: *Frontiers in artificial intelligence* 4 (2021), p. 754641.

[475] Dehua Chen, Guangjun Qian, and Qiao Pan. "Breast cancer classification with electronic medical records using hierarchical attention bidirectional networks". In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 983–988.

[476] Ming Y Lu et al. "AI-based pathology predicts origins for cancers of unknown primary". In: *Nature* 594.7861 (2021), pp. 106–110.

[477] Joung Min Choi and Heejoon Chae. "moBRCA-net: a breast cancer subtype classification framework based on multi-omics attention neural networks". In: *BMC bioinformatics* 24.1 (2023), pp. 1–15.

[478] Seyedeh Maryam Zekavat et al. "Deep learning of the retina enables phenome-and genome-wide analyses of the microvasculature". In: *Circulation* 145.2 (2022), pp. 134–150.

[479] Niousha Bagheri Khoulenjani et al. "Cancer miRNA biomarkers classification using a new representation algorithm and evolutionary deep learning". In: *Soft Computing* 25 (2021), pp. 3113–3129.

[480] Jun Wang et al. "Denoising autoencoder, a deep learning algorithm, aids the identification of a novel molecular signature of lung adenocarcinoma". In: *Genomics, proteomics & bioinformatics* 18.4 (2020), pp. 468–480.

[481] Jayeon Lim et al. "Integrative deep learning for identifying differentially expressed (DE) biomarkers". In: *Computational and mathematical methods in medicine* 2019 (2019).

[482] Li Zhou, Maria Rueda, and Abedalrhman Alkhateeb. "Classification of breast cancer nottingham prognostic index using high-dimensional embedding and residual neural network". In: *Cancers* 14.4 (2022), p. 934.

[483] Kountay Dwivedi et al. "XL1R-Net: Explainable AI-driven improved L1-regularized deep neural architecture for NSCLC biomarker identification". In: *Computational Biology and Chemistry* 108 (2024), p. 107990.

[484] Rui Cao et al. "Development and interpretation of a pathomics-based model for the prediction of microsatellite instability in colorectal cancer". In: *Theranostics* 10.24 (2020), p. 11080.

[485] Karthik Prathaban et al. "Detecting Tumor Infiltration in Diffuse Gliomas with Deep Learning". In: *Advanced Intelligent Systems* 5.12 (2023), p. 2300397.

[486] Shahab Ahmad et al. "A novel hybrid deep learning model for metastatic cancer detection". In: *Computational Intelligence and Neuroscience* 2022 (2022).

[487] Xiaoyuan Li et al. "Exploring and validating the prognostic value of pathomics signatures and genomics in patients with cutaneous melanoma based on bioinformatics and deep learning". In: *Medical Physics* 50.11 (2023), pp. 7049–7059.

[488] Yuan Sh et al. "CaSee: A lightning transfer-learning model directly used to discriminate cancer/normal cells from scRNA-seq". In: *Oncogene* 41.44 (2022), pp. 4866–4876.

[489] Wook Kim et al. "Development of deep learning model for prediction of chemotherapy response using PET images and radiomics features". In: *2018 IEEE Nuclear Science Symposium and Medical Imaging Conference Proceedings (NSS/MIC)*. IEEE. 2018, pp. 1–3.

[490] Feyisope R Eweje et al. "Deep learning for classification of bone lesions on routine MRI". In: *EBioMedicine* 68 (2021).

[491] Leonardo Tanzi et al. "X-ray bone fracture classification using deep learning: a baseline for designing a reliable approach". In: *Applied Sciences* 10.4 (2020), p. 1507.

[492] M Khojaste-Sarakhsi et al. "Deep learning for Alzheimer's disease diagnosis: A survey". In: *Artificial Intelligence in Medicine* 130 (2022), p. 102332.

[493] Gaurav Srivastava et al. "CoviXNet: A novel and efficient deep learning model for detection of COVID-19 using chest X-Ray images". In: *Biomedical Signal Processing and Control* 78 (2022), p. 103848.

[494] Haihua Zhu et al. "CariesNet: a deep learning approach for segmentation of multi-stage caries lesion from oral panoramic X-ray image". In: *Neural Computing and Applications* (2022), pp. 1–9.

[495] Pawan Kumar Mall and Pradeep Kumar Singh. "BoostNet: a method to enhance the performance of deep learning model on musculoskeletal radiographs X-ray images". In: *International Journal of System Assurance Engineering and Management* 13.Suppl 1 (2022), pp. 658–672.

[496] Serkan Sava. "Detecting the stages of Alzheimers disease with pre-trained deep learning architectures". In: *Arabian Journal for Science and Engineering* 47.2 (2022), pp. 2201–2218.

[497] Kazuhiro Suzuki et al. "Development and validation of a modified three-dimensional U-Net deep-learning model for automated detection of lung nodules on chest CT images from the lung image database consortium and Japanese datasets". In: *Academic Radiology* 29 (2022), S11–S17.

[498] Zhiyuan Gao et al. "End-to-end diabetic retinopathy grading based on fundus fluorescein angiography images using deep learning". In: *Graefe's Archive for Clinical and Experimental Ophthalmology* 260.5 (2022), pp. 1663–1673.

[499] Andrew Lin et al. "Deep learning-enabled coronary CT angiography for plaque and stenosis quantification and cardiac risk prediction: an international multicentre study". In: *The Lancet Digital Health* 4.4 (2022), e256–e265.

[500] Chiranjibi Sitaula and Tej Bahadur Shahi. "Monkeypox virus detection using pre-trained deep learning-based approaches". In: *Journal of Medical Systems* 46.11 (2022), p. 78.

[501] Michel Dupuis et al. "External validation of a commercially available deep learning algorithm for fracture detection in children". In: *Diagnostic and Interventional Imaging* 103.3 (2022), pp. 151–159.

[502] Kimberley M Timmins et al. "Geometric deep learning using vascular surface meshes for modality-independent unruptured intracranial aneurysm detection". In: *IEEE Transactions on Medical Imaging* (2023).

[503] Sacide Duman et al. "Detecting the presence of taurodont teeth on panoramic radiographs using a deep learning-based convolutional neural network algorithm". In: *Oral Radiology* 39.1 (2023), pp. 207–214.

[504] Zhaotong Li et al. "Computer-aided diagnosis of spinal tuberculosis from CT images based on deep learning with multimodal feature fusion". In: *Frontiers in Microbiology* 13 (2022), p. 823324.

[505] Madhumita Pal et al. "COVID-19 Prognosis from Chest X-ray Images by using Deep Learning Approaches: A Next Generation Diagnostic Tool." In: *Journal of Pure & Applied Microbiology* 17.2 (2023).

[506] Huanchun Yao and Xinglong Zhang. "A comprehensive review for machine learning based human papillomavirus detection in forensic identification with multiple medical samples". In: *Frontiers in Microbiology* 14 (2023).

[507] Hong-Tao Zhang et al. "Computed tomography–based COVID–19 triage through a deep neural network using mask–weighted global average pooling". In: *Frontiers in cellular and infection microbiology* 13 (2023), p. 187.

[508] Lihong Peng et al. "Analysis of CT scan images for COVID-19 pneumonia based on a deep ensemble framework with DenseNet, Swin transformer, and RegNet". In: *Frontiers in microbiology* 13 (2022), p. 995323.

[509] Yu Zhao et al. "Statistical analysis and machine learning prediction of disease outcomes for COVID-19 and pneumonia patients". In: *Frontiers in cellular and infection microbiology* 12 (2022), p. 838749.

[510] Huan Hu et al. "Identifying SARS-CoV-2 infected cells with scVDN". In: *Frontiers in Microbiology* 14 (2023), p. 1236653.

[511] Heewon Chung et al. "Development and validation of a deep learning model to diagnose COVID-19 using time-series heart rate values before the onset of symptoms". In: *Journal of Medical Virology* 95.2 (2023), e28462.

[512] Erdal In et al. "Using artificial intelligence to improve the diagnostic efficiency of pulmonologists in differentiating COVID-19 pneumonia from community-acquired pneumonia". In: *Journal of Medical Virology* 94.8 (2022), pp. 3698–3705.

[513] Nillmani et al. "Four types of multiclass frameworks for pneumonia classification and its validation in X-ray scans using seven types of deep learning artificial intelligence models". In: *Diagnostics* 12.3 (2022), p. 652.

[514] Anthony Ortiz et al. "Effective deep learning approaches for predicting COVID-19 outcomes from chest computed tomography volumes". In: *Scientific reports* 12.1 (2022), p. 1716.

[515] NB Prakash et al. "Deep transfer learning for COVID-19 detection and infection localization with superpixel based segmentation". In: *Sustainable cities and society* 75 (2021), p. 103252.

[516] Lealem Gedefaw et al. "Artificial intelligence-assisted diagnostic cytology and genomic testing for hematologic disorders". In: *Cells* 12.13 (2023), p. 1755.

[517] Joseph Park et al. "Exome-wide association analysis of CT imaging-derived hepatic fat in a medical biobank". In: *Cell Reports Medicine* 3.12 (2022).

[518] Ojasvi Yadav, Kalpdrum Passi, and Chakresh Kumar Jain. "Using deep learning to classify X-ray images of potential tuberculosis patients". In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 2368–2375.

[519] Zhuang Ai et al. "DR-IIXRN: detection algorithm of diabetic retinopathy based on deep ensemble learning and attention mechanism". In: *Frontiers in Neuroinformatics* 15 (2021), p. 778552.

[520] Mijung Kim et al. "Medinoid: computer-aided diagnosis and localization of glaucoma using deep learning". In: *Applied Sciences* 9.15 (2019), p. 3064.

[521] Taeho Jo et al. "Deep learning detection of informative features in tau PET for Alzheimers disease classification". In: *BMC bioinformatics* 21 (2020), pp. 1–13.

[522] Irene Brusini et al. "Shape information improves the cross-cohort performance of deep learning-based segmentation of the hippocampus". In: *Frontiers in neuroscience* 14 (2020), p. 469755.

[523] Chengsheng Mao et al. "Deep generative classifiers for thoracic disease diagnosis with chest x-ray images". In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 1209–1214.

[524] Olle G Holmberg et al. "Self-supervised retinal thickness prediction enables deep learning from unlabelled data to boost classification of diabetic retinopathy". In: *Nature Machine Intelligence* 2.11 (2020), pp. 719–726.

[525] Fan Gan, Fei-Peng Wu, and Yu-Lin Zhong. "Artificial intelligence method based on multi-feature fusion for automatic macular edema (ME) classification on spectral-domain optical coherence tomography (SD-OCT) images". In: *Frontiers in Neuroscience* 17 (2023), p. 1097291.

[526] Shuo Wang et al. "Deep learning based fetal middle cerebral artery segmentation in large-scale ultrasound images". In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 532–539.

[527] Li Lian Foo et al. "Artificial intelligence in myopia: current and future trends". In: *Current Opinion in Ophthalmology* 32.5 (2021), pp. 413–424.

[528] Umaer Hanif et al. "Estimation of apnea-hypopnea index using deep learning on 3-D craniofacial scans". In: *IEEE Journal of Biomedical and Health Informatics* 25.11 (2021), pp. 4185–4194.

[529] Qicheng Lao, Thomas Fevens, and Boyu Wang. "Leveraging disease progression learning for medical image recognition". In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 671–675.

[530] Jiancheng Ye et al. "Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes". In: *BMC medical informatics and decision making* 20.11 (2020), pp. 1–7.

[531] Jialiang Yang et al. "Prediction of HER2-positive breast cancer recurrence and metastasis risk from histopathological images and clinical information via multimodal deep learning". In: *Computational and structural biotechnology journal* 20 (2022), pp. 333–342.

[532] Mengxin Sun et al. "Early prediction of acute kidney injury in critical care setting using clinical notes and structured multivariate physiological measurements." In: *MedInfo* 264 (2019), pp. 368–372.

[533] Tuoyu Liu et al. "Anticipating the transmissibility of the 2022 mpox outbreak". In: *Journal of Medical Virology* 95.3 (2023), e28683.

[534] Tim Dong et al. “Deep recurrent reinforced learning model to compare the efficacy of targeted local versus national measures on the spread of COVID-19 in the UK”. In: *BMJ open* 12.2 (2022).

[535] Varun N Shenoy et al. “Deepwound: Automated postoperative wound assessment and surgical site surveillance through convolutional neural networks”. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 1017–1021.

[536] Sundreen Asad Kamal et al. “An interpretable risk prediction model for healthcare with pattern attention”. In: *BMC Medical Informatics and Decision Making* 20 (2020), pp. 1–10.

[537] Diego Perna. “Convolutional neural networks learning from respiratory data”. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 2109–2113.

[538] Chunlei Tang et al. “A deep learning approach to handling temporal variation in chronic obstructive pulmonary disease progression”. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 502–509.

[539] Russell Li and Zhandong Liu. “Stress detection using deep neural networks”. In: *BMC Medical Informatics and Decision Making* 20 (2020), pp. 1–10.

[540] Cong Fang et al. “Long-term SARS-CoV-2 neutralizing antibody level prediction using multimodal deep learning: A prospective cohort study on longitudinal data in Wuhan, China”. In: *Journal of Medical Virology* 95.8 (2023), e29036.

[541] Andreas Brink-Kjaer et al. “Age estimation from sleep studies using deep learning predicts life expectancy”. In: *NPJ digital medicine* 5.1 (2022), p. 103.

[542] Wudi Wei et al. “Time series prediction for the epidemic trends of monkeypox using the ARIMA, exponential smoothing, GM (1, 1) and LSTM deep learning methods”. In: *Journal of General Virology* 104.4 (2023), p. 001839.

[543] Pierre Besson et al. “Geometric deep learning on brain shape predicts sex and age”. In: *Computerized Medical Imaging and Graphics* 91 (2021), p. 101939.

[544] Alex Zhavoronkov et al. “PsychoAge and SubjAge: development of deep markers of psychological and subjective age using artificial intelligence”. In: *Aging (Albany NY)* 12.23 (2020), p. 23548.

[545] Orli Bahcall. “Precision medicine”. In: *Nature* 526.7573 (2015), p. 335.

[546] Deep K Hathi et al. *Electronic health record (EHR) and genomics-based machine learning (ML) to predict therapeutic effectiveness among patients with hormone-receptor positive (HR)+/HER2-advanced breast cancer (aBC)*. 2023.

[547] Michael Q Ding et al. "Precision oncology beyond targeted therapy: combining omics data with machine learning matches the majority of cancer cells to effective therapeutics". In: *Molecular cancer research* 16.2 (2018), pp. 269–278.

[548] Jiangsheng Pi et al. "MDGNN: Microbial Drug Prediction Based on Heterogeneous Multi-Attention Graph Neural Network". In: *Frontiers in Microbiology* 13 (2022), p. 819046.

[549] Faheem Ahmed et al. "Drug repurposing for viral cancers: A paradigm of machine learning, deep learning, and virtual screening-based approaches". In: *Journal of Medical Virology* 95.4 (2023), e28693.

[550] Yang Zhang et al. "Deep learning driven drug discovery: tackling severe acute respiratory syndrome coronavirus 2". In: *Frontiers in Microbiology* 12 (2021), p. 739684.

[551] Meriem Bahi and Mohamed Batouche. "Deep semi-supervised learning for DTI prediction using large datasets and H2O-spark platform". In: *2018 international conference on intelligent systems and computer vision (ISCV)*. IEEE. 2018, pp. 1–7.

[552] Hai Yang et al. "From multi-omics data to the cancer druggable gene discovery: a novel machine learning-based approach". In: *Briefings in Bioinformatics* 24.1 (2023), bbac528.

[553] Yongchao Dou et al. "Proteogenomic insights suggest druggable pathways in endometrial carcinoma". In: *Cancer cell* 41.9 (2023), pp. 1586–1605.

[554] Mukunthan Tharmakulasingam et al. "Explainable deep learning approach for multilabel classification of antimicrobial resistance with missing labels". In: *IEEE Access* 10 (2022), pp. 113073–113085.

[555] Jinhong Shi et al. "Antimicrobial resistance genetic factor identification from whole-genome sequence data using deep feature selection". In: *BMC bioinformatics* 20 (2019), pp. 1–14.

[556] Yunxiao Ren et al. "Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning". In: *Bioinformatics* 38.2 (2022), pp. 325–334.

[557] Haoyi Fu et al. "ACEP: improving antimicrobial peptides recognition through automatic feature fusion and amino acid embedding". In: *BMC genomics* 21.1 (2020), pp. 1–14.

[558] Ziye Wang et al. "ARG-SHINE: improve antibiotic resistance class prediction by integrating sequence homology, functional information and deep convolutional neural network". In: *NAR Genomics and Bioinformatics* 3.3 (2021), lqab066.

[559] Matthias I Gröschel et al. "GenTB: A user-friendly genome-based predictor for tuberculosis resistance powered by machine learning". In: *Genome medicine* 13.1 (2021), pp. 1–14.

[560] Alfred Lok Hang Lee et al. "Deep learning model for prediction of extended-spectrum beta-lactamase (ESBL) production in community-onset Enterobacteriaceae bacteraemia from a high ESBL prevalence multi-centre cohort". In: *European Journal of Clinical Microbiology & Infectious Diseases* 40 (2021), pp. 1049–1061.

[561] Sophie J Weiss et al. "Parallel mapping of antibiotic resistance alleles in Escherichia coli". In: *PLoS One* 11.1 (2016), e0146916.

[562] Michael T Paris. "Body composition analysis of computed tomography scans in clinical populations: the role of deep learning". In: *Lifestyle Genomics* 13.1 (2020), pp. 28–31.

[563] Fangping Wan et al. "DeepCPI: a deep learning-based framework for large-scale in silico drug screening". In: *Genomics, proteomics & bioinformatics* 17.5 (2019), pp. 478–495.

[564] Byungjo Lee et al. "Identifying novel antimicrobial peptides from venom gland of spider Pardosa astrigera by deep multi-task learning". In: *Frontiers in Microbiology* 13 (2022), p. 971503.

[565] Manpriya Dua et al. "Guiding Exploration of Antimicrobial Peptide Space with a Deep Neural Network". In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 2082–2087.

[566] Ritesh Sharma et al. "AniAMPpred: artificial intelligence guided discovery of novel antimicrobial peptides in animal kingdom". In: *Briefings in Bioinformatics* 22.6 (2021), bbab242.

[567] K Kennedy et al. "The anti-ageing effects of a natural peptide discovered by artificial intelligence". In: *International Journal of Cosmetic Science* 42.4 (2020), pp. 388–398.

[568] Guadalupe Gonzalez et al. "Predicting anticancer hyperfoods with graph convolutional networks". In: *Human Genomics* 15.1 (2021), pp. 1–12.

[569] Nataliya Sokolovska et al. "Using unlabeled data to discover bivariate causality with deep restricted Boltzmann machines". In: *IEEE/ACM transactions on computational biology and bioinformatics* 17.1 (2018), pp. 358–364.

[570] Son Ali Akbar et al. "An enhanced classification of bacteria pathogen on microscopy images using deep learning". In: *2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*. IEEE. 2021, pp. 119–123.

[571] Alessandro Ferrari, Stefano Lombardi, and Alberto Signoroni. "Bacterial colony counting with convolutional neural networks in digital microbiology imaging". In: *Pattern Recognition* 61 (2017), pp. 629–640.

[572] Sreenath P Kyathanahally et al. "Deep learning classification of lake zooplankton". In: *Frontiers in microbiology* (2021), p. 3226.

[573] Yang Zhang et al. "Deep learning for imaging and detection of microorganisms". In: *Trends in Microbiology* 29.7 (2021), pp. 569–572.

[574] Giovanni Turra, Simone Arrigoni, and Alberto Signoroni. "CNN-based identification of hyperspectral bacterial signatures for digital microbiology". In: *Image Analysis and Processing-ICIAP 2017: 19th International Conference, Catania, Italy, September 11-15, 2017, Proceedings, Part II 19*. Springer. 2017, pp. 500–510.

[575] Ruijun Feng, Sen Li, and Yang Zhang. "A knowledge-integrated deep learning framework for cellular image analysis in parasite microbiology". In: *STAR protocols* 4.3 (2023), p. 102452.

[576] Kwang Seob Lee et al. "Rapid Bacterial Detection in Urine Using Laser Scattering and Deep Learning Analysis". In: *Microbiology spectrum* 10.2 (2022), e01769–21.

[577] Eric Jelli et al. "Single-cell segmentation in bacterial biofilms with an optimized deep learning method enables tracking of cell lineages and measurements of growth rates". In: *Molecular microbiology* (2023).

[578] Chi-Ching Lee et al. "Helminth Egg Analysis Platform (HEAP): An opened platform for microscopic helminth egg identification and quantification based on the integration of deep learning architectures". In: *Journal of Microbiology, Immunology and Infection* 55.3 (2022), pp. 395–404.

[579] Muhd Syamir Azhar et al. "Development of life cycle classification system for Plasmodium knowlesi malaria species using deep learning". In: *AIP Conference Proceedings*. Vol. 2562. 1. AIP Publishing. 2023.

[580] Christoph Spahn et al. "DeepBacs for multi-task bacterial image analysis using open-source deep learning approaches". In: *Communications Biology* 5.1 (2022), p. 688.

[581] Liang Wang et al. "Identification of bacterial pathogens at genus and species levels through combination of Raman spectrometry and deep-learning algorithms". In: *Microbiology Spectrum* 10.6 (2022), e02580–22.

[582] Hongda Wang et al. "Early detection and classification of live bacteria using time-lapse coherent imaging and deep learning". In: *Light: Science & Applications* 9.1 (2020), p. 118.

[583] Zhongxiao Wang et al. "Deep neural networks offer morphologic classification and diagnosis of bacterial vaginosis". In: *Journal of Clinical Microbiology* 59.2 (2021), pp. 10–1128.

[584] Connor Robertson et al. "Video frame prediction of microbial growth with a recurrent neural network". In: *Frontiers in Microbiology* 13 (2023), p. 1034586.

[585] Murat Bacolu et al. "Detection and identification of Bacillus cereus, Bacillus cytotoxicus, Bacillus thuringiensis, Bacillus mycoides and Bacillus weihenstephanensis via machine learning based FTIR spectroscopy". In: *Frontiers in microbiology* 10 (2019), p. 902.

[586] Eliran Dafna and Israel Gannot. "Label-free bacteria identification for clinical applications". In: *Journal of biophotonics* 16.1 (2023), e202200184.

[587] Wei Liu et al. “Rapid discrimination of Shigella spp. and Escherichia coli via label-free surface enhanced Raman spectroscopy coupled with machine learning algorithms”. In: *Frontiers in Microbiology* 14 (2023), p. 1101357.

[588] Emilie Guemas et al. “Automatic patient-level recognition of four Plasmodium species on thin blood smear by a real-time detection transformer (RT-DETR) object detection algorithm: a proof-of-concept and evaluation”. In: *Microbiology Spectrum* (2024), e01440–23.

[589] Blaine A Mathison et al. “Detection of intestinal protozoa in trichrome-stained stool specimens by use of a deep convolutional neural network”. In: *Journal of clinical microbiology* 58.6 (2020), pp. 10–1128.

[590] Zishao Zhong et al. “A study on the diagnosis of the Helicobacter pylori coccoid form with artificial intelligence technology”. In: *Frontiers in Microbiology* 13 (2022), p. 1008346.

[591] Damian J Matuszewski and Ida-Maria Sintorn. “TEM virus images: Benchmark dataset and deep learning classification”. In: *Computer Methods and Programs in Biomedicine* 209 (2021), p. 106318.

[592] Chi Xiao et al. “Virus identification in electron microscopy images by residual mixed attention network”. In: *Computer Methods and Programs in Biomedicine* 198 (2021), p. 105766.

[593] Sofie Van Gassen et al. “CytoNorm: a normalization algorithm for cytometry data”. In: *Cytometry Part A* 97.3 (2020), pp. 268–278.

[594] Konstantin S Kozlov et al. “Digital biology approach for macroscale studies of biofilm growth and biocide effects with electron microscopy”. In: *Digital Discovery* 2.5 (2023), pp. 1522–1539.

[595] Keilor Rojas-Jimenez et al. “Variation of bacterial communities along the vertical gradient in Lake Issyk Kul, Kyrgyzstan”. In: *Environmental Microbiology Reports* 13.3 (2021), pp. 337–347.

[596] Lu Zhang et al. “High Precision Phase Recovery for Single Frame Fringe Pattern of Label-free Cells Detection Based on Deep Learning”. In: *2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA)*. IEEE. 2020, pp. 973–975.

[597] Yichen Wu et al. “Deep learning enables high-throughput analysis of particle-aggregation-based biosensors imaged using holography”. In: *Acs Photonics* 6.2 (2018), pp. 294–301.

[598] Kaiyue Li, Guangtai Ding, and Haitao Wang. “L-fcn: A lightweight fully convolutional network for biomedical semantic segmentation”. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 2363–2367.

[599] Pavel Potocek et al. “Sparse scanning electron microscopy for imaging and segmentation in connectomics”. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 2461–2465.

[600] Michael Kloster et al. “Improving deep learning-based segmentation of diatoms in gigapixel-sized virtual slides by object-based tile positioning and object integrity constraint”. In: *Plos one* 18.2 (2023), e0272103.

[601] Carlos HC Pena et al. “An Ensemble Learning Method for Segmentation Fusion”. In: *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2022, pp. 1–6.

[602] Min Liu et al. “Improved V-Net based image segmentation for 3D neuron reconstruction”. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 443–448.

[603] Nicolas Shiaelis et al. “Virus detection and identification in minutes using single-particle imaging and deep learning”. In: *ACS nano* 17.1 (2022), pp. 697–710.

[604] Claire Lifan Chen et al. “Deep learning in label-free cell classification”. In: *Scientific reports* 6.1 (2016), p. 21471.

[605] Jing Xu et al. “CIForm as a Transformer-based model for cell-type annotation of large-scale single-cell RNA-seq data”. In: *Briefings in Bioinformatics* (2023), bbad195.

[606] Andrew Su et al. “A deep learning model for molecular label transfer that enables cancer cell identification from histopathology images”. In: *NPJ precision oncology* 6.1 (2022), p. 14.

[607] Dan Shao et al. “DeepSec: a deep learning framework for secreted protein discovery in human body fluids”. In: *Bioinformatics* 38.1 (2022), pp. 228–235.

[608] Lijun Cheng et al. “DGCyTOF: Deep learning with graphic cluster visualization to predict cell types of single cell mass cytometry data”. In: *PLoS Computational Biology* 18.4 (2022), e1008885.

[609] Gabriel Rodrigues da Silva et al. “Image dataset of urine test results on petri dishes for deep learning classification”. In: *Data in Brief* 47 (2023), p. 109034.

[610] Kenneth P Smith, Anthony D Kang, and James E Kirby. “Automated interpretation of blood culture gram stains by use of a deep convolutional neural network”. In: *Journal of clinical microbiology* 56.3 (2018), pp. 10–1128.

[611] Yevgeni Nogin et al. “DeepOM: single-molecule optical genome mapping via deep learning”. In: *Bioinformatics* 39.3 (2023), btad137.

[612] Sylwia Majchrowska et al. “Deep neural networks approach to microbial colony detectiona comparative analysis”. In: *Conference on Multimedia, Interaction, Design and Innovation*. Springer. 2021, pp. 98–106.

[613] Sannidhan MS et al. “A swarm-optimized microbial colony counter”. In: *Expert Systems* (), e13510.

[614] Maylis Chapelais-Baron et al. “Colony analysis and deep learning uncover 5-hydroxyindole as an inhibitor of gliding motility and iridescence in Cellulophaga lytica”. In: *Microbiology* 164.3 (2018), pp. 308–321.

[615] László Makrai et al. “Annotated dataset for deep-learning-based bacterial colony detection”. In: (2023).

[616] Upender Kalwa et al. “New methods of removing debris and high-throughput counting of cyst nematode eggs extracted from field soil”. In: *PLoS One* 14.10 (2019), e0223386.

[617] Lewis Howell, Vasileios Anagnostidis, and Fabrice Gielen. “Multi-Object detector yolov4-tiny enables high-throughput combinatorial and spatially-resolved sorting of cells in microdroplets”. In: *Advanced Materials Technologies* 7.5 (2022), p. 2101053.

[618] Sayuri Kuroda et al. “Modeling growth limits of Bacillus spp. spores by using deep-learning algorithm”. In: *Food microbiology* 78 (2019), pp. 38–45.

[619] Ross Michael Worth and Laura Espina. “ScanGrow: Deep Learning-Based Live Tracking of Bacterial Growth in Broth”. In: *Frontiers in Microbiology* 13 (2022), p. 900596.

[620] Yukina Kitahara et al. “A real-time monitoring system for automatic morphology analysis of yeast cultivation in a jar fermenter”. In: *Applied Microbiology and Biotechnology* 106.12 (2022), pp. 4683–4693.

[621] Sreejata Dutta et al. “Identifying dynamical persistent biomarker structures for rare events using modern integrative machine learning approach”. In: *Proteomics* (2023), p. 2200290.

[622] Chia-Hung Yeh et al. “Homogeneous Space Construction and Projection for Single-Cell Expression Prediction Based on Deep Learning”. In: *Bioengineering* 10.9 (2023), p. 996.

[623] Tian Gan et al. “Speeding up collective cell migration using deep reinforcement learning”. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 1277–1280.

[624] Shaojing Wang et al. “CnnPOGTP: a novel CNN-based predictor for identifying the optimal growth temperatures of prokaryotes using only genomic k-mers distribution”. In: *Bioinformatics* 38.11 (2022), pp. 3106–3108.

[625] Tanya Strydom et al. “A roadmap towards predicting species interaction networks (across space and time)”. In: *Philosophical Transactions of the Royal Society B* 376.1837 (2021), p. 20210063.

[626] Haolei Xu et al. “GDReBase: A Knowledge Base for Relations between Human Gut Microbes and Diseases Based on Deep Learning”. In: *Applied Sciences* 13.3 (2023), p. 1614.

[627] Michael Kloster et al. “Deep learning-based diatom taxonomy on virtual slides”. In: *Scientific reports* 10.1 (2020), p. 14416.

[628] Juan C Caicedo et al. "Cell Painting predicts impact of lung cancer variants". In: *Molecular biology of the cell* 33.6 (2022), ar49.

[629] Vito Paolo Pastore et al. "Efficient unsupervised learning of biological images with compressed deep features". In: *Image and Vision Computing* 137 (2023), p. 104764.

[630] Yu Hao et al. "Fiducial marker detection via deep learning approach for electron tomography". In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 642–645.

[631] Mattia Savardi, Sergio Benini, and Alberto Signoroni. "-Hemolysis Detection on Cultured Blood Agar Plates by Convolutional Neural Networks". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 30–38.

[632] Simon Orozco-Arias, Gustavo Isaza, and Romain Guyot. "Retrotransposons in plant genomes: structure, identification, and classification through bioinformatics and machine learning". In: *International journal of molecular sciences* 20.15 (2019), p. 3837.

[633] Simon Orozco-Arias et al. "Automatic curation of LTR retrotransposon libraries from plant genomes through machine learning". In: *Journal of Integrative Bioinformatics* 19.3 (2022), p. 20210036.

[634] Simon Orozco-Arias et al. "InpactorDB: a classified lineage-level plant LTR retrotransposon reference library for free-alignment methods based on machine learning". In: *Genes* 12.2 (2021), p. 190.

[635] Xiaowei Zhu et al. "Machine learning reveals bilateral distribution of somatic L1 insertions in human neurons and glia". In: *Nature neuroscience* 24.2 (2021), pp. 186–196.

[636] Sho Ito, Go Ueno, and Masaki Yamamoto. "DeepCentering: fully automated crystal centering using deep learning for macromolecular crystallography". In: *Journal of synchrotron radiation* 26.4 (2019), pp. 1361–1366.

[637] Anupama Jha et al. "Identifying common transcriptome signatures of cancer by interpreting deep learning models". In: *Genome Biology* 23.1 (2022), pp. 1–23.

[638] Yuchen Liu et al. "Identification and validation of a pyroptosis-related signature in identifying active tuberculosis via a deep learning algorithm". In: *Frontiers in Cellular and Infection Microbiology* 13 (2023).

[639] Nabil Ibtehaz et al. "RamanNet: a generalized neural network architecture for Raman spectrum analysis". In: *Neural Computing and Applications* (2023), pp. 1–17.

[640] Feng Gao, Liwei Gao, and JingYang Gao. "Integrated Detection of Copy Number Variation Based on the Assembly of NGS and 3GS Data". In: *Bioinformatics and Biomedical Engineering: 7th International Work-Conference, IWBBIO 2019, Granada, Spain, May 8-10, 2019, Proceedings, Part I 7*. Springer. 2019, pp. 251–260.

[641] Joseph T Glessner et al. "DeepCNV: a deep learning approach for authenticating copy number variations". In: *Briefings in bioinformatics* 22.5 (2021), bbaa381.

[642] Tom Hill and Robert L Unckless. "A deep learning approach for detecting copy number variation in next-generation sequencing data". In: *G3: Genes, Genomes, Genetics* 9.11 (2019), pp. 3575–3582.

[643] Sarah Alhammad et al. "Efficient DNN-Based Classification of Whole Slide Gram Stain Images for Microbiology". In: *2021 Digital Image Computing: Techniques and Applications (DICTA)*. IEEE. 2021, pp. 01–08.

[644] Haoran Zhu et al. "Dynamic characterization and interpretation for protein-RNA interactions across diverse cellular conditions using HDRNet". In: *Nature Communications* 14.1 (2023), p. 6824.

[645] Ming Li. "Deep Learning Deciphers Protein–RNA Interaction". In: *Genomics, Proteomics & Bioinformatics* 17.5 (2019), p. 475.

[646] Xiaoyang Jing et al. "A web-based protocol for interprotein contact prediction by deep learning". In: *Protein-Protein Interaction Networks: Methods and Protocols* (2020), pp. 67–80.

[647] Weixia Xu et al. "Protein–protein interaction prediction based on ordinal regression and recurrent convolutional neural networks". In: *BMC bioinformatics* 22.6 (2021), pp. 1–20.

[648] Da Zhang and Mansur R Kabuka. "Multimodal deep representation learning for protein-protein interaction networks". In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 595–602.

[649] Min Zeng et al. "A deep learning framework for identifying essential proteins based on protein-protein interaction network and gene expression data". In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 583–588.

[650] Sina Abdollahi, Peng-Chan Lin, and Jung-Hsien Chiang. "WinBinVec: Cancer-Associated Protein-Protein Interaction Extraction and Identification of 20 Various Cancer Types and Metastasis Using Different Deep Learning Models". In: *IEEE Journal of Biomedical and Health Informatics* 25.10 (2021), pp. 4052–4063.

[651] Xue Li et al. "SDNN-PPI: self-attention with deep neural network effect on protein-protein interaction prediction". In: *BMC genomics* 23.1 (2022), p. 474.

[652] Guangyu Zhou et al. "Mutation effect estimation on protein–protein interactions using deep contextualized representation learning". In: *NAR genomics and bioinformatics* 2.2 (2020), lqaa015.

[653] Wenqi Chen et al. "DCSE: Double-Channel-Siamese-Ensemble model for protein protein interaction prediction". In: *BMC genomics* 23.1 (2022), pp. 1–14.

[654] Kanchan Jha, Sriparna Saha, and Sourav Karmakar. "Prediction of Protein-Protein Interactions Using Vision Transformer and Language Model". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2023).

[655] Xiaodi Yang et al. "Deep learning-powered prediction of human-virus protein-protein interactions". In: *Frontiers in microbiology* 13 (2022), p. 842976.

[656] Yu-Ting Hu et al. "Modifying the amino acids in conformational motion pathway of the alpha-amylase of Geobacillus stearothermophilus improved its activity and stability". In: *Frontiers in Microbiology* 14 (2023).

[657] T Mukhopadhyay et al. "Probing the stochastic dynamics of coronaviruses: Machine learning assisted deep computational insights with exploitable dimensions". In: *Advanced Theory and Simulations* 4.7 (2021), p. 2000291.

[658] Zhencheng Fang et al. "DeePVP: Identification and classification of phage virion proteins using deep learning". In: *GigaScience* 11 (2022), giac076.

[659] Charlie F Rowlands, Diana Baralle, and Jamie M Ellingford. "Machine learning approaches for the prioritization of genomic variants impacting pre-mRNA splicing". In: *Cells* 8.12 (2019), p. 1513.

[660] Caleb C Reagor, Nicolas Velez-Angel, and AJ Hudspeth. "Depicting pseudotime-lagged causality across single-cell trajectories for accurate gene-regulatory inference". In: *PNAS nexus* 2.4 (2023), pgad113.

[661] William ND Gao et al. "The crystal structure of vaccinia virus protein E2 and perspectives on the prediction of novel viral protein folds". In: *The Journal of General Virology* 103.1 (2022).

[662] Juan S Rey et al. "Deep-learning in situ classification of HIV-1 virion morphology". In: *Computational and Structural Biotechnology Journal* 19 (2021), pp. 5688–5700.

[663] Lezheng Yu et al. "DeepT3_4: A hybrid deep neural network model for the distinction between bacterial type III and IV secreted effectors". In: *Frontiers in microbiology* 12 (2021), p. 605782.

[664] Zewei Chen et al. "T1SEstacker: A tri-layer stacking model effectively predicts bacterial type 1 secreted proteins based on C-terminal non-repeats-in-toxin-motif sequence features". In: *Frontiers in Microbiology* 12 (2022), p. 813094.

[665] David J Alouani et al. "Specificity of SARS-CoV-2 Real-Time PCR Improved by Deep Learning Analysis". In: *Journal of clinical microbiology* 59.6 (2021), pp. 10–1128.

[666] Yuzhou Chang et al. "Define and visualize pathological architectures of human tissues from spatially resolved transcriptomics using deep learning". In: *Computational and Structural Biotechnology Journal* 20 (2022), pp. 4600–4617.

[667] Marcelo Cicconet et al. “Bots for software-assisted analysis of image-based transcriptomics”. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2017, pp. 134–142.

[668] Fatih Tarlak. “The Use of Predictive Microbiology for the Prediction of the Shelf Life of Food Products”. In: *Foods* 12.24 (2023), p. 4461.

[669] Jian-E Dong et al. “The storage period discrimination of Bolete mushrooms based on deep learning methods combined with two-dimensional correlation spectroscopy and integrative two-dimensional correlation spectroscopy”. In: *Frontiers in Microbiology* 12 (2021), p. 771428.

[670] Anuja Bhargava et al. “A review on tea quality and safety using emerging parameters”. In: *Journal of Food Measurement and Characterization* (2022), pp. 1–21.

[671] Lydia Kienbaum et al. “DeepCob: Precise and high-throughput analysis of maize cob geometry using deep learning with an application in genebank phenomics”. In: *Plant Methods* 17.1 (2021), pp. 1–19.

[672] Siddhartha Das, Sudeepta Pattanayak, and Prateek Ranjan Behera. “Application of machine learning: a recent advancement in plant diseases detection”. In: *Journal of Plant Protection Research* (2022), pp. 122–135.

[673] Abigail Dommer et al. “# COVIDisAirborne: AI-enabled multiscale computational microscopy of delta SARS-CoV-2 in a respiratory aerosol”. In: *The international journal of high performance computing applications* 37.1 (2023), pp. 28–44.

[674] Jiahui Chen, Rui Wang, and Guo-Wei Wei. “SARS-CoV-2 becoming more infectious as revealed by algebraic topology and deep learning”. In: *Communications in information and systems* 21.1 (2021), p. 31.

[675] Jinglong Du et al. “Accelerated super-resolution MR image reconstruction via a 3D densely connected deep convolutional neural network”. In: *2018 IEEE international conference on bioinformatics and biomedicine (BIBM)*. IEEE. 2018, pp. 349–355.

[676] Yuxuan Zhao et al. “Isotropic super-resolution light-sheet microscopy of dynamic intracellular structures at subsecond timescales”. In: *Nature Methods* 19.3 (2022), pp. 359–369.

[677] Lingbo Jin et al. “Deep learning extended depth-of-field microscope for fast and slide-free histology”. In: *Proceedings of the National Academy of Sciences* 117.52 (2020), pp. 33051–33060.

[678] Hideharu Mikami et al. “Virtual-freezing fluorescence imaging flow cytometry”. In: *Nature communications* 11.1 (2020), p. 1162.

[679] Yuta Suzuki et al. “Label-free chemical imaging flow cytometry by high-speed multicolor stimulated Raman scattering”. In: *Proceedings of the National Academy of Sciences* 116.32 (2019), pp. 15842–15848.

[680] Xi Cheng et al. "Fast and lightweight network for single frame structured illumination microscopy super-resolution". In: *IEEE Transactions on Instrumentation and Measurement* 71 (2022), pp. 1–11.

[681] Sen Yang et al. "DeepNoise: signal and noise disentanglement based on classifying fluorescent microscopy images via deep learning". In: *Genomics, Proteomics & Bioinformatics* 20.5 (2022), pp. 989–1001.

[682] Shiro Ihara et al. "Deep learning-based noise filtering toward millisecond order imaging by using scanning transmission electron microscopy". In: *Scientific reports* 12.1 (2022), p. 13462.

[683] Niago Moreira Nobre Leite et al. "Deep convolutional autoencoder for EEG noise filtering". In: *2018 IEEE international conference on bioinformatics and biomedicine (BIBM)*. IEEE. 2018, pp. 2605–2612.

[684] Babak Saravi et al. "Performance of artificial intelligence-based algorithms to predict prolonged length of stay after lumbar decompression surgery". In: *Journal of Clinical Medicine* 11.14 (2022), p. 4050.

[685] Aazad Abbas et al. "Machine learning using preoperative patient factors can predict duration of surgery and length of stay for total knee arthroplasty". In: *International journal of medical informatics* 158 (2022), p. 104670.

[686] Haoyan Zhong et al. "Machine learning approaches in predicting ambulatory same day discharge patients after total hip arthroplasty". In: *Regional Anesthesia & Pain Medicine* 46.9 (2021), pp. 779–783.

[687] ZeYu Huang et al. "Predicting postoperative transfusion in elective total HIP and knee arthroplasty: Comparison of different machine learning models of a case-control study". In: *International Journal of Surgery* 96 (2021), p. 106183.

[688] Wessel T Stam et al. "The prediction of surgical complications using artificial intelligence in patients undergoing major abdominal surgery: A systematic review". In: *Surgery* 171.4 (2022), pp. 1014–1021.

[689] Runyu Jing et al. "DeepT3 2.0: improving type III secreted effector predictions by an integrative deep learning framework". In: *NAR Genomics and Bioinformatics* 3.4 (2021), lqab086.

[690] Amarin Jettakul, Duangdao Wichadakul, and Peerapon Vateekul. "Relation extraction between bacteria and biotopes from biomedical texts with attention mechanisms and domain-specific contextual representations". In: *BMC bioinformatics* 20 (2019), pp. 1–17.

[691] Ran Zhong et al. "Microbial Interaction Extraction from Biomedical Literature using Max-Bi-LSTM". In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2019, pp. 723–726.

[692] Sotirios Paraskevopoulos et al. "Using Artificial Intelligence to extract information on pathogen characteristics from scientific publications". In: *International Journal of Hygiene and Environmental Health* 245 (2022), p. 114018.

[693] Kenbun Sone et al. “Evolution of a surgical system using deep learning in minimally invasive surgery”. In: *Biomedical Reports* 19.1 (2023), pp. 1–8.

[694] Bo Yang et al. “Motion prediction for beating heart surgery with GRU”. In: *Biomedical Signal Processing and Control* 83 (2023), p. 104641.

[695] Jong Woo Choi et al. “Surgery-first orthognathic approach to correct facial asymmetry: artificial intelligence–based cephalometric analysis”. In: *Plastic and Reconstructive Surgery* 149.3 (2022), 496e–499e.

[696] Masashi Takeuchi et al. “Automated surgical-phase recognition for robot-assisted minimally invasive esophagectomy using artificial intelligence”. In: *Annals of Surgical Oncology* 29.11 (2022), pp. 6847–6855.

[697] Ke Cheng et al. “Artificial intelligence-based automated laparoscopic cholecystectomy surgical phase recognition and analysis”. In: *Surgical endoscopy* 36.5 (2022), pp. 3160–3168.

[698] Takahiro Igaki et al. “Artificial intelligence-based total mesorectal excision plane navigation in laparoscopic colorectal surgery”. In: *Diseases of the Colon & Rectum* 65.5 (2022), e329–e333.

[699] Minho Hwang et al. “Automating surgical peg transfer: calibration with deep learning can exceed speed, accuracy, and consistency of humans”. In: *IEEE Transactions on Automation Science and Engineering* 20.2 (2022), pp. 909–922.

[700] Babak Namazi, Ganesh Sankaranarayanan, and Venkat Devarajan. “A contextual detector of surgical tools in laparoscopic videos using deep learning”. In: *Surgical endoscopy* (2022), pp. 1–10.

[701] Jani Koskinen et al. “Automated tool detection with deep learning for monitoring kinematics and eye-hand coordination in microsurgery”. In: *Computers in Biology and Medicine* 141 (2022), p. 105121.

[702] Ziyi Jin et al. “Intelligent SERS navigation system guiding brain tumor surgery by intraoperatively delineating the metabolic acidosis”. In: *Advanced Science* 9.7 (2022), p. 2104935.

[703] Daichi Kitaguchi et al. “Limited generalizability of single deep neural network for surgical instrument segmentation in different surgical environments”. In: *Scientific Reports* 12.1 (2022), p. 12575.

[704] Chieh-Tsai Wu, Yao-Hung Yang, and Yau-Zen Chang. “Three-dimensional deep learning to automatically generate cranial implant geometry”. In: *Scientific reports* 12.1 (2022), p. 2683.

[705] Tian-jian Luo, Ya-chao Fan, Ji-tu Lv, et al. “Deep reinforcement learning from error-related potentials via an EEG-based brain-computer interface”. In: *2018 IEEE international conference on bioinformatics and biomedicine (BIBM)*. IEEE. 2018, pp. 697–701.

[706] Prejwal Prabhakaran et al. “Insilico generation of novel ligands for the inhibition of SARS-CoV-2 main protease (3CLpro) using deep learning”. In: *Frontiers in Microbiology* 14 (2023).

[707] Maged Nasser and Umi Kalsom Yusof. "Deep Learning Based Methods for Breast Cancer Diagnosis: A Systematic Review and Future Direction". In: *Diagnostics* 13.1 (2023), p. 161.

[708] Peng Zhao et al. "A comparative study of deep learning classification methods on a small environmental microorganism image dataset (EMDS-6): from convolutional neural networks to visual transformers". In: *Frontiers in Microbiology* 13 (2022), p. 792166.

[709] Geng Tian et al. "A deep ensemble learning-based automated detection of COVID-19 using lung CT images and Vision Transformer and ConvNeXt". In: *Frontiers in Microbiology* 13 (2022), p. 1024104.

[710] Boyuan Wang, Du Zhang, and Zonggui Tian. "CoroTrans-CL: A Novel Transformer-Based Continual Deep Learning Model for Image Recognition of Coronavirus Infections". In: *Electronics* 12.4 (2023), p. 866.

[711] Jim Clauwaert, Gerben Menschaert, and Willem Waegeman. "Explainability in transformer models for functional genomics". In: *Briefings in bioinformatics* 22.5 (2021), bbab060.

[712] Frederik Wessels et al. "A self-supervised vision transformer to predict survival from histopathology in renal cell carcinoma". In: *World Journal of Urology* 41.8 (2023), pp. 2233–2241.

[713] Zhikang Wang et al. "Surformer: An interpretable pattern-perceptive survival transformer for cancer survival prediction from histopathology whole slide images". In: *Computer Methods and Programs in Biomedicine* 241 (2023), p. 107733.

[714] C-C Jay Kuo. "Understanding convolutional neural networks with a mathematical model". In: *Journal of Visual Communication and Image Representation* 41 (2016), pp. 406–413.

[715] Haibo Wang et al. "Mitosis detection in breast cancer pathology images by combining handcrafted and Convolutional Neural Network features". In: *Journal of Medical Imaging* 1.3 (2014), p. 034003.

[716] Houhong Lu et al. "Multimodal fusion convolutional neural network with cross-attention mechanism for internal defect detection of magnetic tile". In: *IEEE Access* 10 (2022), pp. 60876–60886.

[717] Muhammad Rizwan et al. "Brain tumor and glioma grade classification using Gaussian convolutional neural network". In: *IEEE Access* 10 (2022), pp. 29731–29740.

[718] Md Nahiduzzaman, Md Rabiul Islam, and Rakibul Hassan. "ChestX-Ray6: Prediction of multiple diseases including COVID-19 from chest X-ray images using convolutional neural network". In: *Expert Systems with Applications* 211 (2023), p. 118576.

[719] Phillip M Cheng and Harshawn S Malhi. "Transfer learning with convolutional neural networks for classification of abdominal ultrasound images". In: *Journal of digital imaging* 30.2 (2017), pp. 234–243.

[720] Ibrahem Kandel and Mauro Castelli. "Transfer learning with convolutional neural networks for diabetic retinopathy image classification. A review". In: *Applied Sciences* 10.6 (2020), p. 2021.

[721] Alexander Kensert, Philip J Harrison, and Ola Spjuth. "Transfer learning with deep convolutional neural networks for classifying cellular morphological changes". In: *SLAS Discovery: Advancing Life Sciences R&D* 24.4 (2019), pp. 466–475.

[722] Hui Wen Loh et al. "Decision support system for major depression detection using spectrogram and convolution neural network with EEG signals". In: *Expert Systems* 39.3 (2022), e12773.

[723] Indhupriya Subramanian et al. "Multi-omics data integration, interpretation, and its application". In: *Bioinformatics and biology insights* 14 (2020), p. 1177932219899051.

[724] Francisco Cristovao et al. "Investigating deep learning based breast cancer subtyping using pan-cancer and multi-omic data". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19.1 (2020), pp. 121–134.

[725] ZHU Rong et al. "Diagnostic classification of lung cancer using deep transfer learning technology and multi-omics data". In: *Chinese Journal of Electronics* 30.5 (2021), pp. 843–852.

[726] Bo Yang, Yan Yang, and Xueping Su. "Deep structure integrative representation of multi-omics data for cancer subtyping". In: *Bioinformatics* 38.13 (2022), pp. 3337–3342.

[727] Rezaul Karim et al. "Adversary-aware multimodal neural networks for cancer susceptibility prediction from multiomics data". In: *IEEE Access* 10 (2022), pp. 54386–54409.

[728] Tianle Ma and Aidong Zhang. "Multi-view factorization autoencoder with network constraints for multi-omic integrative analysis". In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 702–707.

[729] Kelin Wang et al. "DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants". In: *Molecular Plant* 16.1 (2023), pp. 279–293.

[730] Dong Sui et al. "Imaging Biomarkers and Gene Expression Data Correlation Framework for Lung Cancer Radiogenomics Analysis Based on Deep Learning". In: *IEEE Access* 9 (2021), pp. 125247–125257.

[731] Alok Sharma et al. "DeepFeature: feature selection in nonimage data using convolutional neural network". In: *Briefings in bioinformatics* 22.6 (2021), bbab297.

[732] Zhiqiang Zhang et al. "Deep learning in omics: a survey and guideline". In: *Briefings in functional genomics* 18.1 (2019), pp. 41–57.

[733] Rohan Gupta et al. "Artificial intelligence to deep learning: machine intelligence approach for drug discovery". In: *Molecular diversity* 25 (2021), pp. 1315–1360.

[734] Nhat Tran and Jean Gao. "Network representation of large-scale heterogeneous RNA sequences with integration of diverse multi-omics, interactions, and annotations data". In: *Pacific Symposium on Biocomputing 2020*. World Scientific. 2019, pp. 499–510.

[735] Xiaoxia Wen et al. "Clinlabomics: leveraging clinical laboratory data by data mining strategies". In: *BMC bioinformatics* 23.1 (2022), pp. 1–20.

[736] Danishuddin, Shawez Khan, and Jong Joo Kim. "From cancer big data to treatment: Artificial intelligence in cancer research". In: *The journal of gene medicine* (2023), e3629.

[737] Elisabetta Manduchi, Joseph D Romano, and Jason H Moore. "The promise of automated machine learning for the genetic analysis of complex traits". In: *Human Genetics* 141.9 (2022), pp. 1529–1544.

[738] Luis Torada et al. "ImaGene: a convolutional neural network to quantify natural selection from genomic data". In: *BMC bioinformatics* 20 (2019), pp. 1–12.

[739] Minjae J Kim et al. "Computational methods in glaucoma research: Current status and future outlook". In: *Molecular Aspects of Medicine* 94 (2023), p. 101222.

[740] Martin Treppner, Harald Binder, and Moritz Hess. "Interpretable generative deep learning: an illustration with single cell gene expression data". In: *Human Genetics* 141.9 (2022), pp. 1481–1498.

[741] Chongyang Chen et al. "Applications of multi-omics analysis in human diseases". In: *MedComm* 4.4 (2023), e315.

[742] Haluk Dogan et al. "Elucidation of dynamic microRNA regulations in cancer progression using integrative machine learning". In: *Briefings in bioinformatics* 22.6 (2021), bbab270.

[743] Nicasia Beebe-Wang et al. "Unified AI framework to uncover deep interrelationships between gene expression and Alzheimers disease neuropathologies". In: *Nature Communications* 12.1 (2021), p. 5369.

[744] Mashooq Ahmad Dar et al. "Multiomics technologies: role in disease biomarker discoveries and therapeutics". In: *Briefings in Functional Genomics* 22.2 (2023), pp. 76–96.

[745] Lei Huang et al. "Deep learning methods for omics data imputation". In: *Biology* 12.10 (2023), p. 1313.

[746] Lijia Yu et al. "Ensemble deep learning of embeddings for clustering multimodal single-cell omics data". In: *Bioinformatics* (2023), btad382.

[747] Wan Xiang Shen et al. “AggMapNet: enhanced and explainable low-sample omics deep learning with feature-aggregated multi-channel networks”. In: *Nucleic Acids Research* 50.8 (2022), e45–e45.

[748] Neo Christopher Chung et al. “Unsupervised classification of multi-omics data during cardiac remodeling using deep learning”. In: *Methods* 166 (2019), pp. 66–73.

[749] Mei Sze Tan et al. “A review on omics-based biomarkers discovery for Alzheimer’s disease from the bioinformatics perspectives: statistical approach vs machine learning approach”. In: *Computers in biology and medicine* 139 (2021), p. 104947.

[750] Song Jiang, Ting Wang, and Kun-He Zhang. “Data-driven decision-making for precision diagnosis of digestive diseases”. In: *BioMedical Engineering OnLine* 22.1 (2023), p. 87.

[751] Barbara Lobato-Delgado, Blanca Priego-Torres, and Daniel Sanchez-Morillo. “Combining Molecular, Imaging, and Clinical Data Analysis for Predicting Cancer Prognosis”. In: *Cancers* 14.13 (2022), p. 3215.

[752] Sören Richard Stahlschmidt, Benjamin Ulfenborg, and Jane Synnergren. “Multimodal deep learning for biomedical data fusion: a review”. In: *Briefings in Bioinformatics* 23.2 (2022), bbab569.

[753] Lucas Schneider et al. “Integration of deep learning-based image analysis and genomic data in cancer pathology: A systematic review”. In: *European journal of cancer* 160 (2022), pp. 80–91.

[754] Gang Li et al. “Deep principal correlated auto-encoders with application to imaging and genomics data integration”. In: *IEEE Access* 8 (2020), pp. 20093–20107.

[755] Muhammad Shoaib et al. “Deep learning for plant bioinformatics: an explainable gradient-based approach for disease detection”. In: *Frontiers in Plant Science* 14 (2023).

[756] Jun Shao et al. “Radiogenomic system for non-invasive identification of multiple actionable mutations and PD-L1 expression in non-small cell lung cancer based on CT images”. In: *Cancers* 14.19 (2022), p. 4823.

[757] Liye Jia et al. “DCCAFN: deep convolution cascade attention fusion network based on imaging genomics for prediction survival analysis of lung cancer”. In: *Complex & Intelligent Systems* (2023), pp. 1–16.

[758] Jennifer B Rattray et al. “Machine learning identification of Pseudomonas aeruginosa strains from colony image data”. In: *PLOS Computational Biology* 19.12 (2023), e1011699.

[759] Edgar Olehnovics et al. “The role of hydrophobic nodes in the dynamics of class A beta-lactamases”. In: *Frontiers in Microbiology* 12 (2021), p. 720991.

[760] Manasi A Pethe, Aliza B Rubenstein, and Sagar D Khare. "Data-driven supervised learning of a viral protease specificity landscape from deep sequencing and molecular simulations". In: *Proceedings of the National Academy of Sciences* 116.1 (2019), pp. 168–176.

[761] Wenjing Qiu et al. "Evaluating the microsatellite instability of colorectal cancer based on multimodal deep learning integrating histopathological and molecular data". In: *Frontiers in Oncology* 12 (2022), p. 925079.

[762] Pei-Chen Tsai et al. "Histopathology images predict multi-omics aberrations and prognoses in colorectal cancer patients". In: *Nature communications* 14.1 (2023), p. 2102.

[763] Han-Ru Chen et al. "Prediction of pathway-omics signature of histopathology images via attention-based deep learning in lung adenocarcinoma and squamous cell carcinoma". In: *Cancer Research* 83.7_Supplement (2023), pp. 3177–3177.

[764] Kaimei Huang et al. "Predicting colorectal cancer tumor mutational burden from histopathological images and clinical information using multi-modal deep learning". In: *Bioinformatics* 38.22 (2022), pp. 5108–5115.

[765] Siew-Kee Low, Hitoshi Zembutsu, and Yusuke Nakamura. "Breast cancer: The translation of big genomic data to cancer precision medicine". In: *Cancer science* 109.3 (2018), pp. 497–506.

[766] Vidhi Malik, Yogesh Kalakoti, and Durai Sundar. "Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer". In: *Bmc Genomics* 22 (2021), pp. 1–11.

[767] Siwen Xu et al. "Integrative analysis of histopathological images and chromatin accessibility data for estrogen receptor-positive breast cancer". In: *BMC Medical Genomics* 13.11 (2020), pp. 1–12.

[768] Yasha Ektefaie et al. "Integrative multiomics-histopathology analysis for breast cancer classification". In: *NPJ Breast Cancer* 7.1 (2021), p. 147.

[769] Xiao Xiao et al. "Transformer with convolution and graph-node co-embedding: an accurate and interpretable vision backbone for predicting gene expressions from local histopathological image". In: *Medical Image Analysis* 91 (2024), p. 103040.

[770] Hong-Jun Yoon et al. "Deep radiogenomics for predicting clinical phenotypes in invasive breast cancer". In: *14th International Workshop on Breast Imaging (IWBI 2018)*. Vol. 10718. SPIE. 2018, pp. 391–396.

[771] Giuseppe Cannizzaro et al. "Automated integration of genomic metadata with sequence-to-sequence models". In: *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part V*. Springer. 2021, pp. 187–203.

[772] Rui Qin, Lara K Mahal, and Daniel Bojar. "Deep learning explains the biology of branched glycans from single-cell sequencing data". In: *Iscience* 25.10 (2022).

[773] Chang Xu et al. "DeepST: identifying spatial domains in spatial transcriptomics by deep learning". In: *Nucleic Acids Research* 50.22 (2022), e131–e131.

[774] Lihong Peng et al. "Stgnnks: identifying cell types in spatial transcriptomics data based on graph neural network, denoising auto-encoder, and k-sums clustering". In: *Computers in Biology and Medicine* 166 (2023), p. 107440.

[775] Jesper B Lund et al. "AntiSplodge: a neural-network-based RNA-profile deconvolution pipeline designed for spatial transcriptomics". In: *NAR Genomics and Bioinformatics* 4.4 (2022), lqac073.

[776] Chunman Zuo and Luonan Chen. "Deep-joint-learning analysis model of single cell transcriptome and open chromatin accessibility data". In: *Briefings in Bioinformatics* 22.4 (2021), bbaa287.

[777] Zhenhao Zhang et al. "A generalizable framework to comprehensively predict epigenome, chromatin organization, and transcriptome". In: *Nucleic Acids Research* 51.12 (2023), pp. 5931–5947.

[778] Aikaterini Symeonidi et al. "Recursive convolutional neural networks for epigenomics". In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, pp. 2567–2574.

[779] Kevin E Wu et al. "BABEL enables cross-modality translation between multiomic profiles at single-cell resolution". In: *Proceedings of the National Academy of Sciences* 118.15 (2021), e2023070118.

[780] Yang Xu, Priyojit Das, and Rachel Patton McCord. "SMILE: mutual information learning for integration of single-cell omics data". In: *Bioinformatics* 38.2 (2022), pp. 476–486.

[781] Zhencheng Fang and Hongwei Zhou. "VirionFinder: identification of complete and partial prokaryote virus virion protein from virome data using the sequence and biochemical properties of amino acids". In: *Frontiers in microbiology* 12 (2021), p. 615711.

[782] Pratik Dutta, Aditya Prakash Patra, and Sriparna Saha. "DeePROG: deep attention-based model for diseased gene prognosis by fusing multi-omics data". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19.5 (2021), pp. 2770–2781.

[783] Francisco Carrillo-Perez et al. "Machine-learning-based late fusion on multi-omics and multi-scale data for non-small-cell lung cancer diagnosis". In: *Journal of Personalized Medicine* 12.4 (2022), p. 601.

[784] Guido Zampieri et al. "Machine and deep learning meet genome-scale metabolic modeling". In: *PLoS computational biology* 15.7 (2019), e1007084.

[785] Yichuan Liu et al. "Application of deep learning algorithm on whole genome sequencing data uncovers structural variants associated with multiple mental disorders in African American patients". In: *Molecular psychiatry* 27.3 (2022), pp. 1469–1478.

[786] Yang Liu et al. "Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean". In: *Frontiers in genetics* 10 (2019), p. 1091.

[787] Imane Boudellioua et al. "DeepPVP: phenotype-based prioritization of causative variants using deep learning". In: *BMC bioinformatics* 20 (2019), pp. 1–8.

[788] Parisa Sarzaeim, Francisco Muñoz-Arriola, and Diego Jarquín. "Climate and genetic data enhancement using deep learning analytics to improve maize yield predictability". In: *Journal of experimental botany* 73.15 (2022), pp. 5336–5354.

[789] Yubin Zhang et al. "Machine learning for image-based multi-omics analysis of leaf veins". In: *Journal of Experimental Botany* 74.17 (2023), pp. 4928–4941.

[790] Shuni Wu et al. "Gene–environment interaction analysis via deep learning". In: *Genetic Epidemiology* 47.3 (2023), pp. 261–286.

[791] Jing-Yi Li et al. "Identifying complex motifs in massive omics data with a variable-convolutional layer in deep neural network". In: *Briefings in Bioinformatics* 22.6 (2021), bbab233.

[792] Edian F Franco et al. "Performance comparison of deep learning autoencoders for cancer subtype detection using multi-omics data". In: *Cancers* 13.9 (2021), p. 2013.

[793] Asia Gervits and Roded Sharan. "Predicting genetic interactions, cell line dependencies and drug sensitivities with variational graph auto-encoder". In: *Frontiers in Bioinformatics* 2 (2022), p. 1025783.

[794] Minhyeok Lee. "Deep Learning Techniques with Genomic Data in Cancer Prognosis: A Comprehensive Review of the 2021–2023 Literature". In: *Biology* 12.7 (2023), p. 893.

[795] Muta Tah Hira et al. "Integrated multi-omics analysis of ovarian cancer using variational autoencoders". In: *Scientific reports* 11.1 (2021), p. 6265.

[796] Bo Yang et al. "Integrating multi-omic data with deep subspace fusion clustering for cancer subtype prediction". In: *IEEE/ACM transactions on computational biology and bioinformatics* 18.1 (2019), pp. 216–226.

[797] Tianwei Yu. "AIME: Autoencoder-based integrative multi-omics data embedding that allows for confounder adjustments". In: *PLoS Computational Biology* 18.1 (2022), e1009826.

[798] Luca Alessandri et al. "Sparsely connected autoencoders: a multi-purpose tool for single cell omics analysis". In: *International Journal of Molecular Sciences* 22.23 (2021), p. 12755.

[799] Shunxin Xiao et al. “Graph Neural Networks with Multiple Prior Knowledge for Multi-Omics Data Analysis”. In: *IEEE Journal of Biomedical and Health Informatics* (2023).

[800] Xiao-Meng Zhang et al. “Graph neural networks and their current applications in bioinformatics”. In: *Frontiers in genetics* 12 (2021), p. 690049.

[801] Ian Morilla et al. “Deep models of integrated multiscale molecular data decipher the endothelial cell response to ionizing radiation”. In: *Iscience* 25.1 (2022), p. 103685.

[802] Bartosz Jan Bartmanski, Miguel Rocha, and Maria Zimmermann-Kogadeeva. “Recent advances in data-and knowledge-driven approaches to explore primary microbial metabolism”. In: *Current Opinion in Chemical Biology* 75 (2023), p. 102324.

[803] Neda Tavakoli. “Seq2image: Sequence analysis using visualization and deep convolutional neural network”. In: *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE. 2020, pp. 1332–1337.

[804] Xiuquan Du et al. “Deepss: Exploring splice site motif through convolutional neural network directly from DNA sequence”. In: *IEEE Access* 6 (2018), pp. 32958–32978.

[805] Rui Yin, Zihan Luo, and Chee Keong Kwoh. “Exploring the lethality of human-adapted coronavirus through alignment-free machine learning approaches using genomic sequences”. In: *Current Genomics* 22.8 (2021), p. 583.

[806] Yue Qi, Peijie Zheng, and Guohua Huang. “DeepLBCEPred: A Bi-LSTM and multi-scale CNN-based deep learning method for predicting linear B-cell epitopes”. In: *Frontiers in Microbiology* 14 (2023), p. 1117027.

[807] Tauhidul Islam and Lei Xing. “Cartography of Genomic Interactions Enables Deep Analysis of Single-Cell Expression Data”. In: *Nature Communications* 14.1 (2023), p. 679.

[808] Michael D Kupperman, Thomas Leitner, and Ruian Ke. “A deep learning approach to real-time HIV outbreak detection using genetic data”. In: *PLOS Computational Biology* 18.10 (2022), e1010598.

[809] Gabriel BM Câmara et al. “Convolutional Neural Network Applied to SARS-CoV-2 Sequence Classification”. In: *Sensors* 22.15 (2022), p. 5730.

[810] Weihua Li et al. “Learning spatiotemporal embedding with gated convolutional recurrent networks for translation initiation site prediction”. In: *Pattern Recognition* 136 (2023), p. 109234.

[811] Mahmoud Ghandi et al. “Enhanced regulatory sequence prediction using gapped k-mer features”. In: *PLoS computational biology* 10.7 (2014), e1003711.

[812] Longzhen Tang et al. "Human activity recognition based on mixed CNN with radar multi-spectrogram". In: *IEEE Sensors Journal* 21.22 (2021), pp. 25950–25962.

[813] Wen-li Liu and Qing-biao Wu. "Analysis method and algorithm design of biological sequence problem based on generalized k-mer vector". In: *Applied Mathematics-A Journal of Chinese Universities* 36.1 (2021), pp. 114–127.

[814] Elly C Knight et al. "Pre-processing spectrogram parameters improve the accuracy of bioacoustic classification using convolutional neural networks". In: *Bioacoustics* 29.3 (2020), pp. 337–355.

[815] Antonio Blanca et al. "The statistics of k-mers from a sequence undergoing a simple mutation process without spurious matches". In: *Journal of Computational Biology* 29.2 (2022), pp. 155–168.

[816] Bernd T Meyer and Birger Kollmeier. "Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition". In: *Speech Communication* 53.5 (2011), pp. 753–767.

[817] Simon R Harris. "SKA: split kmer analysis toolkit for bacterial genomic epidemiology". In: *BioRxiv* (2018), p. 453142.

[818] HB Kekre, A Athawale, and Mrunali Desai. "Speaker identification using row mean vector of spectrogram". In: *Proceedings of the international conference & workshop on emerging trends in technology*. 2011, pp. 171–174.

[819] Magali Jaillard et al. "A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events". In: *PLoS genetics* 14.11 (2018), e1007758.

[820] Raquel de M Barbosa and Marcelo AC Fernandes. "Chaos game representation dataset of SARS-CoV-2 genome". In: *Data in brief* 30 (2020), p. 105618.

[821] Kevin Dick and James R Green. "Chaos game representations & deep learning for proteome-wide protein prediction". In: *2020 IEEE 20th international conference on bioinformatics and bioengineering (BIBE)*. IEEE. 2020, pp. 115–121.

[822] Adrian Kania and Krzysztof Sarapata. "Multifarious aspects of the chaos game representation and its applications in biological sequence analysis". In: *Computers in Biology and Medicine* 151 (2022), p. 106243.

[823] Gao Jie and Xu Zhen-Yuan. "Chaos game representation (CGR)-walk model for DNA sequences". In: *Chinese Physics B* 18.1 (2009), p. 370.

[824] Susana Vinga et al. "Pattern matching through Chaos Game Representation: bridging numerical and discrete data structures for biological sequence analysis". In: *Algorithms for Molecular Biology* 7 (2012), pp. 1–12.

[825] Luis Lugo and Emiliano Barreto- Hernández. "A Recurrent Neural Network approach for whole genome bacteria identification". In: *Applied Artificial Intelligence* 35.9 (2021), pp. 642–656.

[826] Chao Wang et al. “Inference of pan-cancer related genes by orthologs matching based on enhanced LSTM model”. In: *Frontiers in Microbiology* 13 (2022), p. 963704.

[827] Prakash Chourasia et al. “Empowering pandemic response with federated learning for protein sequence data analysis”. In: *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2023, pp. 01–08.

[828] Zhe Liu et al. “Emden: A novel method integrating graph and transformer representations for predicting the effect of mutations on clinical drug response”. In: *Computers in Biology and Medicine* 167 (2023), p. 107678.

[829] Jacob Schreiber and Ritambhara Singh. “Machine Learning for profile prediction in genomics”. In: *Current Opinion in Chemical Biology* 65 (2021), pp. 35–41.

[830] Jérôme Tubiana et al. “Reduced B cell antigenicity of Omicron lowers host serologic response”. In: *Cell reports* 41.3 (2022).

[831] Xin Gao, Jie Zhang, and Zhi Wei. “Deep learning for sequence pattern recognition”. In: *2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC)*. IEEE. 2018, pp. 1–6.

[832] Artem Danilevsky, Avital Luba Polsky, and Noam Shomron. “Adaptive sequencing using nanopores and deep learning of mitochondrial DNA”. In: *Briefings in Bioinformatics* 23.4 (2022), bbac251.

[833] Avantika Lal. “Deciphering the regulatory syntax of genomic DNA with deep learning”. In: *Journal of Biosciences* 47.3 (2022), p. 47.

[834] Wenlong Ma et al. “DeepSATA: A Deep Learning-Based Sequence Analyzer Incorporating the Transcription Factor Binding Affinity to Dissect the Effects of Non-Coding Genetic Variants”. In: *International Journal of Molecular Sciences* 24.15 (2023), p. 12023.

[835] Aditi R Durge, Deepti D Shrimankar, and Ankush D Sawarkar. “Heuristic Analysis of Genomic Sequence Processing Models for High Efficiency Prediction: A Statistical Perspective”. In: *Current Genomics* 23.5 (2022), p. 299.

[836] Pablo Millán Arias et al. “DeLUCS: Deep learning for unsupervised clustering of DNA sequences”. In: *Plos one* 17.1 (2022), e0261531.

[837] Xin Gao, Zhi Wei, and Hakon Hakonarson. “tRNA-DL: a deep learning approach to improve tRNAscan-SE prediction results”. In: *Human heredity* 83.3 (2019), pp. 163–172.

[838] Martin A Smith et al. “Molecular barcoding of native RNAs using nanopore sequencing and deep learning”. In: *Genome research* 30.9 (2020), pp. 1345–1353.

[839] Sargun Nagpal et al. “Genomic surveillance of COVID-19 variants with language models and machine learning”. In: *Frontiers in Genetics* 13 (2022), p. 858252.

[840] Simon Orozco-Arias et al. “Genomic object detection: An improved approach for transposable elements detection and classification using convolutional neural networks”. In: *Plos one* 18.9 (2023), e0291925.

[841] Hung-I Harry Chen et al. “GSAE: an autoencoder with embedded gene-set nodes for genomics functional characterization”. In: *BMC systems biology* 12.8 (2018), pp. 45–57.

[842] Fritz Lekschas et al. “Peax: Interactive visual pattern search in sequential data using unsupervised deep representation learning”. In: *Computer Graphics Forum*. Vol. 39. 3. Wiley Online Library. 2020, pp. 167–179.

[843] Shiquan Sun et al. “A fast and efficient count-based matrix factorization method for detecting cell types from single-cell RNAseq data”. In: *BMC systems biology* 13.2 (2019), pp. 1–8.

[844] Hechen Yang et al. “EMDS-7: Environmental microorganism image dataset seventh version for multiple object detection evaluation”. In: *Frontiers in Microbiology* 14 (2023), p. 1084312.

[845] Buddini Iroshika Karawdeniya et al. “Adeno-associated virus characterization for cargo discrimination through nanopore responsiveness”. In: *Nanoscale* 12.46 (2020), pp. 23721–23731.

[846] Viola Ravasio et al. “Garfield-ngs: Genomic variants filtering by deep learning models in NGS”. In: *Bioinformatics* 34.17 (2018), pp. 3038–3040.

[847] Thi Kieu Khanh Ho and Jeonghwan Gwak. “Toward deep learning approaches for learning structure motifs and classifying biological sequences from RNA A-to-I editing events”. In: *IEEE Access* 7 (2019), pp. 127464–127474.

[848] Quanzhong Liu et al. “DeepGenGrep: a general deep learning-based predictor for multiple genomic signals and regions”. In: *Bioinformatics* 38.17 (2022), pp. 4053–4061.

[849] Jim Clauwaert and Willem Waegeman. “Novel transformer networks for improved sequence labeling in genomics”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19.1 (2020), pp. 97–106.

[850] Gaofei Jiang et al. “DCiPatho: deep cross-fusion networks for genome scale identification of pathogens”. In: *Briefings in Bioinformatics* (2023), bbad194.

[851] Waseem Ullah et al. “Multi-Stage Temporal Convolution Network for COVID-19 Variant Classification”. In: *Diagnostics* 12.11 (2022), p. 2736.

[852] Maria GF Coutinho et al. “SARS-CoV-2 virus classification based on stacked sparse autoencoder”. In: *Computational and Structural Biotechnology Journal* 21 (2023), pp. 284–298.

[853] Indrajit Saha et al. “COVID-DeepPredictor: recurrent neural network to predict SARS-CoV-2 and other pathogenic viruses”. In: *Frontiers in genetics* 12 (2021), p. 569120.

[854] James K Senter et al. “Unaligned sequence similarity search using deep learning”. In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2019, pp. 1892–1899.

[855] Hai Yang et al. “Cancer classification based on chromatin accessibility profiles with deep adversarial learning model”. In: *PLoS Computational Biology* 16.11 (2020), e1008405.

[856] Fergus Imrie et al. “Protein family-specific models using deep neural networks and transfer learning improve virtual screening and highlight the need for more data”. In: *Journal of chemical information and modeling* 58.11 (2018), pp. 2319–2330.

[857] Yi-Fei Huang. “Unified inference of missense variant effects and gene constraints in the human genome”. In: *PLoS genetics* 16.7 (2020), e1008922.

[858] Li Shen and Paul M Thompson. “Brain imaging genomics: Integrated analysis and Machine Learning”. In: *Proceedings of the IEEE* 108.1 (2019), pp. 125–162.

[859] Tristan Cordier et al. “Embracing environmental genomics and Machine Learning for routine biomonitoring”. In: *Trends in microbiology* 27.5 (2019), pp. 387–397.

[860] Arno van Hilten et al. “GenNet framework: interpretable deep learning for predicting phenotypes from genetic data”. In: *Communications biology* 4.1 (2021), p. 1094.

[861] Junjie Shen et al. “Efficient feature extraction from highly sparse binary genotype data for cancer prognosis prediction using an auto-encoder”. In: *Frontiers in Oncology* 12 (2023), p. 1091767.

[862] Iain Carmichael et al. “Joint and individual analysis of breast cancer histologic images and genomic covariates”. In: *The annals of applied statistics* 15.4 (2021), p. 1697.

[863] Joel Rozowsky et al. “The EN-TEx resource of multi-tissue personal epigenomes & variant-impact models”. In: *Cell* 186.7 (2023), pp. 1493–1511.

[864] Nam D Nguyen, Jiawei Huang, and Daifeng Wang. “A deep manifold-regularized learning model for improving phenotype prediction from multi-modal data”. In: *Nature computational science* 2.1 (2022), pp. 38–46.

[865] Junjie Chen and Xinghua Shi. “Sparse convolutional denoising autoencoders for genotype imputation”. In: *Genes* 10.9 (2019), p. 652.

[866] Matteo Togninalli et al. “Multi-modal deep learning improves grain yield prediction in wheat breeding by fusing genomics and phenomics”. In: *Bioinformatics* 39.6 (2023), btad336.

[867] Xu Wang et al. “High-throughput phenotyping with deep learning gives insight into the genetic architecture of flowering time in wheat”. In: *GigaScience* 8.11 (2019), giz120.

[868] Arjan Sammani et al. “Diagnosis and risk prediction of dilated cardiomyopathy in the era of big data and genomics”. In: *Journal of clinical medicine* 10.5 (2021), p. 921.

[869] Simon Höllerer et al. “Large-scale DNA-based phenotypic recording and deep learning enable highly accurate sequence-function mapping”. In: *Nature communications* 11.1 (2020), p. 3551.

[870] Chao Wang et al. “Leveraging population genomics for individualized correction of the hallmarks of alpha-1 antitrypsin deficiency”. In: *Chronic Obstructive Pulmonary Diseases: Journal of the COPD Foundation* 7.3 (2020), p. 224.

[871] Wellison JS Diniz et al. “Machine learning-based Co-expression network analysis unravels potential fertility-related genes in beef cows”. In: *Animals* 12.19 (2022), p. 2715.

[872] Bettina Mieth et al. “DeepCOMBI: explainable artificial intelligence for the analysis and discovery in genome-wide association studies”. In: *NAR genomics and bioinformatics* 3.3 (2021), lqab065.

[873] Torsten Pook et al. “Using local convolutional neural networks for genomic prediction”. In: *Frontiers in genetics* 11 (2020), p. 561497.

[874] Heonwoo Lee et al. “RelCurator: a text mining-based curation system for extracting gene–phenotype relationships specific to neurodegenerative disorders”. In: *Genes & Genomics* (2023), pp. 1–12.

[875] Xinpeng Guo et al. “Using expression quantitative trait loci data and graph-embedded neural networks to uncover genotype–phenotype interactions”. In: *Frontiers in Genetics* 13 (2022), p. 921775.

[876] Victoria Bourgeais, Farida Zehraoui, and Blaise Hanczar. “GraphGONet: a self-explaining neural network encapsulating the Gene Ontology graph for phenotype prediction on gene expression”. In: *Bioinformatics* 38.9 (2022), pp. 2504–2511.

[877] Sambriddhi Mainali, Fredy A Colorado, and Max H Garzon. “Foretelling the phenotype of a genomic sequence”. In: *IEEE/ACM transactions on computational biology and bioinformatics* 18.2 (2020), pp. 777–783.

[878] Raquel Dias et al. “Rapid, Reference-Free human genotype imputation with denoising autoencoders”. In: *Elife* 11 (2022), e75600.

[879] Jiawei Li et al. “DeepATT: a hybrid category attention neural network for identifying functional effects of DNA sequences”. In: *Briefings in bioinformatics* 22.3 (2021), bbaa159.

[880] Peyton Greenside et al. “Discovering epistatic feature interactions from neural network models of regulatory DNA sequences”. In: *Bioinformatics* 34.17 (2018), pp. i629–i637.

[881] Jian Zhou et al. “Whole-genome deep-learning analysis identifies contribution of noncoding mutations to autism risk”. In: *Nature genetics* 51.6 (2019), pp. 973–980.

[882] Cheng-Hong Yang et al. “Applications of deep learning and fuzzy systems to detect cancer mortality in next-generation genomic data”. In: *IEEE Transactions on Fuzzy Systems* 29.12 (2020), pp. 3833–3844.

[883] Mohammad Ruhul Amin et al. “Deepannotator: genome annotation with deep learning”. In: *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*. 2018, pp. 254–259.

[884] Felix Stiehler et al. “Helixer: cross-species gene annotation of large eukaryotic genomes using deep learning”. In: *Bioinformatics* 36.22-23 (2020), pp. 5291–5298.

[885] Alexander Kroll et al. “Deep learning allows genome-scale prediction of Michaelis constants from structural features”. In: *PLoS biology* 19.10 (2021), e3001402.

[886] Anjana Senanayake et al. “DeepSelectNet: deep neural network based selective sequencing for oxford nanopore sequencing”. In: *BMC bioinformatics* 24.1 (2023), p. 31.

[887] Reza Ahsan et al. “Image processing unravels the evolutionary pattern of SARS-CoV-2 against SARS and MERS through position-based pattern recognition”. In: *Computers in Biology and Medicine* 134 (2021), p. 104471.

[888] Xusheng Li et al. “A hybrid deep learning framework for bacterial named entity recognition”. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 428–433.

[889] Walaa Alkady et al. “Classifying COVID-19 based on amino acids encoding with machine learning algorithms”. In: *Chemometrics and Intelligent Laboratory Systems* 224 (2022), p. 104535.

[890] Tariq Sadad et al. “Classification of Highly Divergent Viruses from DNA/RNA Sequence Using Transformer-Based Models”. In: *Biomedicines* 11.5 (2023), p. 1323.

[891] Noam Auslander et al. “Seeker: alignment-free identification of bacteriophage genomes by deep learning”. In: *Nucleic acids research* 48.21 (2020), e121–e121.

[892] Qiaoxing Liang et al. “DeepMicrobes: taxonomic classification for metagenomics with deep learning”. In: *NAR Genomics and Bioinformatics* 2.1 (2020), lqaa009.

[893] Helen Park et al. “Investigation of machine learning algorithms for taxonomic classification of marine metagenomes”. In: *Microbiology Spectrum* 11.5 (2023), e05237–22.

[894] Kuo-ching Liang and Yasubumi Sakakibara. “MetaVelvet-DL: a MetaVelvet deep learning extension for de novo metagenome assembly”. In: *BMC bioinformatics* 22.6 (2021), pp. 1–21.

[895] Om Prakash Singh et al. "Classification of SARS-CoV-2 and non-SARS-CoV-2 using machine learning algorithms". In: *Computers in biology and medicine* 136 (2021), p. 104650.

[896] Xiaojing Yang et al. "Meta-Viromic Sequencing Reveals Virome Characteristics of Mosquitoes and Culicoides on Zhoushan Island, China". In: *Microbiology Spectrum* 11.1 (2023), e02688–22.

[897] Jakub M Bartoszewicz, Anja Seidel, and Bernhard Y Renard. "Interpretable detection of novel human viruses from genome sequencing data". In: *NAR genomics and bioinformatics* 3.1 (2021), lqab004.

[898] Florian Mock et al. "Taxonomic classification of DNA sequences beyond sequence similarity using deep neural networks". In: *Proceedings of the National Academy of Sciences* 119.35 (2022), e2122636119.

[899] Brahim Matougui et al. "NLP-MeTaxa: A Natural Language Processing Approach for Metagenomic Taxonomic Binning Based on Deep Learning". In: *Current Bioinformatics* 16.7 (2021), pp. 992–1003.

[900] Ho-Jin Gwak and Mina Rho. "Transformer-based embedding applied to classify bacterial species using sequencing reads". In: *2022 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE. 2022, pp. 374–377.

[901] Ruither AL Gomes and F Murilo Zerbini. "ConCreT, a 2D convolutional neural network for taxonomic classification applied to viruses in the phylum Cressdnaviricota". In: *Journal of Virological Methods* 320 (2023), p. 114789.

[902] Jinghui Li et al. "Learning functional conservation between human and pig to decipher evolutionary mechanisms underlying gene expression and complex traits". In: *Cell Genomics* 3.10 (2023).

[903] Arnaud Nguembang Fadja et al. "Identification of natural selection in genomic data with deep convolutional neural network". In: *BioData Mining* 14.1 (2021), pp. 1–18.

[904] Wouter Deelder et al. "Using deep learning to identify recent positive selection in malaria parasite sequence data". In: *Malaria journal* 20.1 (2021), p. 270.

[905] Pavel Latyshev et al. "Unsupervised domain adaptation methods for cross-species transfer of regulatory code signals". In: *Frontiers in Big Data* 6 (2023), p. 1140663.

[906] Chris CR Smith et al. "Dispersal inference from population genetic variation using a convolutional neural network". In: *Genetics* 224.2 (2023), iyad068.

[907] Vladimira Kubicova and Ivo Provaznik. "Relationship of bacteria using comparison of whole genome sequences in frequency domain". In: *Information Technologies in Biomedicine, Volume 3*. Springer. 2014, pp. 397–408.

[908] Zhicheng Wang et al. "Fusang: a framework for phylogenetic tree inference via deep learning". In: *Nucleic Acids Research* 51.20 (2023), pp. 10909–10923.

[909] Qiang Zhu et al. “The phylogenetic tree based deep forest for metagenomic data classification”. In: *2018 IEEE international conference on bioinformatics and biomedicine (BIBM)*. IEEE. 2018, pp. 279–282.

[910] Yueyu Jiang, Puoya Tabaghi, and Siavash Mirarab. “Phylogenetic Placement Problem: A Hyperbolic Embedding Approach”. In: *RECOMB International Workshop on Comparative Genomics*. Springer. 2022, pp. 68–85.

[911] Maryam Hayati et al. “Deep clustering of bacterial tree images”. In: *Philosophical Transactions of the Royal Society B* 377.1861 (2022), p. 20210231.

[912] Ananya Bhattacharjee and Md Shamsuzzoha Bayzid. “Machine learning based imputation techniques for estimating phylogenetic trees from incomplete distance matrices”. In: *BMC genomics* 21 (2020), pp. 1–14.

[913] Tian Tian et al. “Complex hierarchical structures in single-cell genomics data unveiled by deep hyperbolic manifold learning”. In: *Genome Research* 33.2 (2023), pp. 232–246.

[914] Alisa Prusokiene, Augustinas Prusokas, and Renata Retkute. “Machine learning based lineage tree reconstruction improved with knowledge of higher level relationships between cells and genomic barcodes”. In: *NAR Genomics and Bioinformatics* 5.3 (2023), lqad077.

[915] Lex Flagel, Yaniv Brandvain, and Daniel R Schrider. “The unreasonable effectiveness of convolutional neural networks in population genetic inference”. In: *Molecular biology and evolution* 36.2 (2019), pp. 220–238.

[916] Pradeep Ruperao et al. “Exploring the sorghum race level diversity utilizing 272 sorghum accessions genomic resources”. In: *Frontiers in Plant Science* 14 (2023), p. 1143512.

[917] Danielle Miller, Adi Stern, and David Burstein. “Deciphering microbial gene function using natural language processing”. In: *Nature Communications* 13.1 (2022), p. 5731.

[918] Yuhan Su et al. “Application of bert to enable gene classification based on clinical evidence”. In: *BioMed research international* 2020 (2020).

[919] Troy M LaPolice and Yi-Fei Huang. “An unsupervised deep learning framework for predicting human essential genes from population and functional genomic data”. In: *BMC bioinformatics* 24.1 (2023), p. 347.

[920] Nguyen Quoc Khanh Le et al. “A computational framework based on ensemble deep neural networks for essential genes identification”. In: *International journal of molecular sciences* 21.23 (2020), p. 9070.

[921] Nour Eldeen M Khalifa et al. “Artificial intelligence technique for gene expression by tumor RNA-Seq data: a novel optimized deep learning approach”. In: *IEEE Access* 8 (2020), pp. 22874–22883.

[922] Peilong Li, Junfeng Wei, and Ying Zhu. “CellGO: a novel deep learning-based framework and webserver for cell-type-specific gene function interpretation”. In: *Briefings in Bioinformatics* 25.1 (2024), bbad417.

[923] Marketa Nykrynova et al. "Using deep learning for gene detection and classification in raw nanopore signals". In: *Frontiers in Microbiology* 13 (2022), p. 942179.

[924] Antonino Fiannaca et al. "GOWDL: gene ontology-driven wide and deep learning model for cell typing of scRNA-seq data". In: *Briefings in Bioinformatics* 24.6 (2023), bbad332.

[925] Ahmad Pesaranghader et al. "deepSimDEF: deep neural embeddings of gene products and gene ontology terms for functional analysis of genes". In: *Bioinformatics* 38.11 (2022), pp. 3051–3061.

[926] Alexander T Xue, Daniel R Schrider, and Andrew D Kern. "Discovery of ongoing selective sweeps within anopheles mosquito populations using deep learning". In: *Molecular biology and evolution* 38.3 (2021), pp. 1168–1183.

[927] Albert Whata and Charles Chimedza. "Deep learning for SARS COV-2 genome sequences". In: *Ieee Access* 9 (2021), pp. 59597–59611.

[928] Ke Ding et al. "CRMnet: a deep learning model for predicting gene expression from large regulatory sequence datasets". In: *Frontiers in Big Data* 6 (2023), p. 1113402.

[929] Tarek Khorshed, Mohamed N Moustafa, and Ahmed Rafea. "Multi-tissue cancer classification of gene expressions using deep learning". In: *2020 IEEE Sixth International Conference on Big Data Computing Service and Applications (BigDataService)*. IEEE. 2020, pp. 128–135.

[930] Tarek Khorshed, Mohamed N Moustafa, and Ahmed Rafea. "Deep learning for multi-tissue cancer classification of gene expressions (GeneXNet)". In: *IEEE Access* 8 (2020), pp. 90615–90629.

[931] Ghazaleh Khodabandelou, Etienne Routhier, and Julien Mozziconacci. "Genome annotation across species using deep convolutional neural networks". In: *PeerJ Computer Science* 6 (2020), e278.

[932] Rui Xie et al. "A deep auto-encoder model for gene expression prediction". In: *BMC genomics* 18 (2017), pp. 39–49.

[933] Wanwen Zeng, Yong Wang, and Rui Jiang. "Integrating distal and proximal information to predict gene expression via a densely connected convolutional neural network". In: *Bioinformatics* 36.2 (2020), pp. 496–503.

[934] Rui Xie et al. "A predictive model of gene expression using a deep learning framework". In: *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2016, pp. 676–681.

[935] Sergii Babichev, Igor Liakh, and Irina Kalinina. "Applying a Recurrent Neural Network-Based Deep Learning Model for Gene Expression Data Classification". In: *Applied Sciences* 13.21 (2023), p. 11823.

[936] Qian Liu and Pingzhao Hu. "Association analysis of deep genomic features extracted by denoising autoencoders in breast cancer". In: *Cancers* 11.4 (2019), p. 494.

[937] Genevieve L Stein-OBrien et al. "Decomposing cell identity for transfer learning across cellular measurements, platforms, tissues, and species". In: *Cell systems* 8.5 (2019), pp. 395–411.

[938] Jonathan Rosenski, Sagiv Shifman, and Tommy Kaplan. "Predicting gene knockout effects from expression data". In: *BMC Medical Genomics* 16.1 (2023), p. 26.

[939] Malik Yousef, Ege Ülgen, and Osman Uur Sezerman. "CogNet: classification of gene expression data based on ranked active-subnetwork-oriented KEGG pathway enrichment analysis". In: *PeerJ Computer Science* 7 (2021), e336.

[940] Aaron M Newman and James B Cooper. "AutoSOME: a clustering method for identifying gene expression modules without prior knowledge of cluster number". In: *BMC bioinformatics* 11 (2010), pp. 1–15.

[941] Marco Jost et al. "Titrating gene expression using libraries of systematically attenuated CRISPR guide RNAs". In: *Nature biotechnology* 38.3 (2020), pp. 355–364.

[942] Fortunato Morabito et al. "Genes selection using deep learning and explainable artificial intelligence for chronic lymphocytic leukemia predicting the need and time to therapy". In: *Frontiers in Oncology* 13 (2023).

[943] Chinedu Anthony Anene et al. "ACSNI: An unsupervised machine-learning tool for prediction of tissue-specific pathway components using gene expression profiles". In: *Patterns* 2.6 (2021).

[944] Ilya Vainberg Slutskin, Adina Weinberger, and Eran Segal. "Sequence determinants of polyadenylation-mediated regulation". In: *Genome research* 29.10 (2019), pp. 1635–1647.

[945] Thai-Hoang Pham et al. "A deep learning framework for high-throughput mechanism-driven phenotype compound screening and its application to COVID-19 drug repurposing". In: *Nature machine intelligence* 3.3 (2021), pp. 247–257.

[946] You Wu et al. "Deep learning prediction of chemical-induced dose-dependent and context-specific multiplex phenotype responses and its application to personalized alzheimers disease drug repurposing". In: *PLOS Computational Biology* 18.8 (2022), e1010367.

[947] Ye Yuan and Ziv Bar-Joseph. "Deep learning for inferring gene relationships from single-cell expression data". In: *Proceedings of the National Academy of Sciences* 116.52 (2019), pp. 27151–27158.

[948] Zhongxiao Li et al. "Applications of deep learning in understanding gene regulation". In: *Cell Reports Methods* (2023).

[949] Linlin Xing et al. "An improved Bayesian network method for reconstructing gene regulatory network based on candidate auto selection". In: *BMC genomics* 18 (2017), pp. 17–30.

[950] Minji Kang et al. "Learning cell-type-specific gene regulation mechanisms by multi-attention based deep learning with regulatory latent space". In: *Frontiers in Genetics* 11 (2020), p. 869.

[951] Madison Dautle, Shaoqiang Zhang, and Yong Chen. "scTIGER: A Deep-Learning Method for Inferring Gene Regulatory Networks from Case versus Control scRNA-seq Datasets". In: *International Journal of Molecular Sciences* 24.17 (2023), p. 13339.

[952] Qian Wang et al. "A gene regulatory network inference model based on pseudo-siamese network". In: *BMC bioinformatics* 24.1 (2023), pp. 1–18.

[953] Feng Chen and Chunhe Li. "Inferring structural and dynamical properties of gene networks from data with deep learning". In: *NAR Genomics and Bioinformatics* 4.3 (2022), lqac068.

[954] Musu Yuan, Liang Chen, and Minghua Deng. "Clustering CITE-seq data with a canonical correlation-based deep learning method". In: *Frontiers in Genetics* 13 (2022), p. 977968.

[955] Yifei Chen et al. "Gene expression inference with deep learning". In: *Bioinformatics* 32.12 (2016), pp. 1832–1839.

[956] Boris Jankovic and Takashi Gojobori. "From shallow to deep: some lessons learned from application of machine learning for recognition of functional genomic elements in human genome". In: *Human Genomics* 16.1 (2022), pp. 1–13.

[957] Shengcheng Dong and Alan P Boyle. "Predicting functional variants in enhancer and promoter elements using RegulomeDB". In: *Human mutation* 40.9 (2019), pp. 1292–1298.

[958] Hu Zhao et al. "An inferred functional impact map of genetic variants in rice". In: *Molecular Plant* 14.9 (2021), pp. 1584–1599.

[959] TaeJin Ahn et al. "Deep learning-based identification of cancer or normal tissue using gene expression data". In: *2018 IEEE international conference on bioinformatics and biomedicine (BIBM)*. IEEE. 2018, pp. 1748–1752.

[960] Saud H AlDubayan et al. "Detection of pathogenic variants with germline genetic testing using deep learning vs standard methods in patients with prostate cancer and melanoma". In: *Jama* 324.19 (2020), pp. 1957–1969.

[961] Ryan Poplin et al. "A universal SNP and small-indel variant caller using deep neural networks". In: *Nature biotechnology* 36.10 (2018), pp. 983–987.

[962] Eugénie Dionnet et al. "Splicing impact of deep exonic missense variants in CAPN3 explored systematically by minigene functional assay". In: *Human Mutation* 41.10 (2020), pp. 1797–1810.

[963] Yubo Zhang et al. "Inferring historical introgression with deep learning". In: *Systematic Biology* (2023), syad033.

[964] K Nandhini and G Tamilpavai. "Hybrid CNN-LSTM and modified wild horse herd Model-based prediction of genome sequences for genetic disorders". In: *Biomedical Signal Processing and Control* 78 (2022), p. 103840.

[965] Zhao Li et al. "DeepBSA: A deep-learning algorithm improves bulked segregant analysis for dissecting complex traits". In: *Molecular Plant* 15.9 (2022), pp. 1418–1427.

[966] Muhammad Tahir ul Qamar et al. "Comparative genomics profiling revealed multi-stress responsive roles of the CC-NBS-LRR genes in three mango cultivars". In: *Frontiers in Plant Science* 14 (2023).

[967] Fatemah Almalki, Janak Sunuwar, and Rajeev K Azad. "Using Machine Learning to Predict Genes Underlying Differentiation of Multipartite and Unipartite Traits in Bacteria". In: *Microorganisms* 11.11 (2023), p. 2756.

[968] Chen Li, Jiaxing Chen, and Shuai Cheng Li. "Deep learning for HGT insertion sites recognition". In: *BMC genomics* 21.11 (2020), pp. 1–18.

[969] Zhencheng Fang and Hongwei Zhou. "Identification of the conjugative and mobilizable plasmid fragments in the plasmidome using sequence signatures". In: *Microbial Genomics* 6.11 (2020).

[970] Simon Orozco-Arias et al. "Measuring performance metrics of machine learning algorithms for detecting and classifying transposable elements". In: *Processes* 8.6 (2020), p. 638.

[971] Bo Wang and Eric R Gamazon. "Modeling mutational effects on biochemical phenotypes using convolutional neural networks: application to SARS-CoV-2". In: *Iscience* 25.7 (2022).

[972] Steve Agajanian, Odeyemi Oluyemi, and Gennady M Verkhivker. "Integration of random forest classifiers and deep convolutional neural networks for classification and biomolecular modeling of cancer driver mutations". In: *Frontiers in molecular biosciences* 6 (2019), p. 44.

[973] Guojie Zhong, Yoolim A Choi, and Yufeng Shen. "VBASS enables integration of single cell gene expression data in Bayesian association analysis of rare variants". In: *Communications Biology* 6.1 (2023), p. 774.

[974] Zahra Shamsi, Matthew Chan, and Diwakar Shukla. "TLmutation: predicting the effects of mutations using transfer learning". In: *The Journal of Physical Chemistry B* 124.19 (2020), pp. 3845–3854.

[975] Erica Sawyer et al. "A Neural Network Approach for Anomaly Detection in Genomic Signals". In: *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE. 2020, pp. 968–971.

[976] Mario Banuelos et al. "Related Inference: A Supervised Learning Approach to Detect Signal Variation in Genome Data". In: *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE. 2021, pp. 1215–1219.

[977] Azza Althagafi et al. "DeepSVP: integration of genotype and phenotype for structural variant prioritization using deep learning". In: *Bioinformatics* 38.6 (2022), pp. 1677–1684.

[978] Li Guo, Tingming Liang, and Zuhong Lu. "A comprehensive study of multiple mapping and feature selection for correction strategy in the analysis of small RNAs from SOLiD sequencing". In: *BioSystems* 104.2-3 (2011), pp. 87–93.

[979] Wenchao Ai et al. "Deep Sequencing of HPV16 E6 Region Reveals Unique Mutation Pattern of HPV16 and Predicts Cervical Cancer". In: *Microbiology Spectrum* 10.4 (2022), e01401–22.

[980] Sumin Yang et al. "Harnessing deep learning into hidden mutations of neurological disorders for therapeutic challenges". In: *Archives of Pharmacal Research* (2023), pp. 1–15.

[981] Guillaume P Ramstein and Edward S Buckler. "Prediction of evolutionary constraint by genomic annotations improves functional prioritization of genomic variants in maize". In: *Genome Biology* 23.1 (2022), pp. 1–26.

[982] Zeeshan Abbas, Hilal Tayara, and Kil To Chong. "ZayyuNet–A unified deep learning model for the identification of epigenetic modifications using raw genomic sequences". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19.4 (2021), pp. 2533–2544.

[983] Giosue Lo Bosco et al. "A deep learning model for epigenomic studies". In: *2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. IEEE. 2016, pp. 688–692.

[984] Zenghong Wu et al. "Epigenetic and Tumor Microenvironment for Prognosis of Patients with Gastric Cancer". In: *Biomolecules* 13.5 (2023), p. 736.

[985] Muhammad Nabeel Asim et al. "DNA-MP: a generalized DNA modifications predictor for multiple species based on powerful sequence encoding method". In: *Briefings in Bioinformatics* 24.1 (2023), bbac546.

[986] Jimin Pei et al. "Mutation severity spectrum of rare alleles in the human genome is predictive of disease type". In: *PLoS computational biology* 16.5 (2020), e1007775.

[987] Zhenyu Yue, Xinlu Chu, and Junfeng Xia. "PredCID: prediction of driver frameshift indels in human cancer". In: *Briefings in Bioinformatics* 22.3 (2021), bbaa119.

[988] Ella Shtifman Segal et al. "Gene essentiality analyzed by in vivo transposon mutagenesis and machine learning in a stable haploid isolate of Candida albicans". In: *MBio* 9.5 (2018), pp. 10–1128.

[989] Anton Levitan et al. "Comparing the utility of in vivo transposon mutagenesis approaches in yeast species to infer gene essentiality". In: *Current Genetics* 66 (2020), pp. 1117–1134.

[990] Roberto Vera Alvarez et al. "SNPDelScore: combining multiple methods to score deleterious effects of noncoding mutations in the human genome". In: *Bioinformatics* 34.2 (2018), pp. 289–291.

[991] Zhuohan Yu et al. "Topological identification and interpretation for single-cell gene regulation elucidation across multiple platforms using scMGCA". In: *Nature Communications* 14.1 (2023), p. 400.

[992] Pengyu Ni and Zhengchang Su. "Deciphering epigenomic code for cell differentiation using deep learning". In: *BMC genomics* 20 (2019), pp. 1–18.

[993] Saad Nadeem et al. "Ki67 proliferation index in medullary thyroid carcinoma: a comparative study of multiple counting methods and validation of image analysis and deep learning platforms". In: *Histopathology* 83.6 (2023), pp. 981–988.

[994] Yen-Jung Chiu, Chung-En Ni, and Yen-Hua Huang. "HArmonized single-cell RNA-seq Cell type Assisted Deconvolution (HASCAD)". In: *BMC Medical Genomics* 16.Suppl 2 (2023), p. 272.

[995] Hyun-Hwan Jeong et al. "Investigating cellular trajectories in the severity of COVID-19 and their transcriptional programs using machine learning approaches". In: *Genes* 12.5 (2021), p. 635.

[996] Yiyuan Fang, Shuyi Deng, and Cai Li. "A generalizable deep learning framework for inferring fine-scale germline mutation rate maps". In: *Nature Machine Intelligence* 4.12 (2022), pp. 1209–1223.

[997] Jiaqi Liu et al. "DrABC: deep learning accurately predicts germline pathogenic mutation status in breast cancer patients based on phenotype data". In: *Genome Medicine* 14.1 (2022), pp. 1–15.

[998] Jason G Parker et al. "Targeting intra-tumoral heterogeneity of human brain tumors with in vivo imaging: A roadmap for imaging genomics from multiparametric MR signals". In: *Medical Physics* 50.4 (2023), pp. 2590–2606.

[999] Shahid Ahmad Wani, Sumeer Ahmad Khan, and SMK Quadri. "scJVAE: A novel method for integrative analysis of multimodal single-cell data". In: *Computers in Biology and Medicine* 158 (2023), p. 106865.

[1000] Rebecca J Carlson et al. "A genome-wide optical pooled screen reveals regulators of cellular antiviral responses". In: *Proceedings of the National Academy of Sciences* 120.16 (2023), e2210623120.

[1001] Yan Zhou et al. "scDLC: a deep learning framework to classify large sample single-cell RNA-seq data". In: *BMC genomics* 23.1 (2022), p. 504.

[1002] Xiang Lin et al. "Clustering of single-cell multi-omics data with a multimodal deep learning method". In: *Nature communications* 13.1 (2022), p. 7705.

[1003] Anjali Gopal and Amy E Herr. "Segmentation-based analysis of single-cell immunoblots". In: *Electrophoresis* 42.20 (2021), pp. 2070–2080.

[1004] Steffen Albrecht et al. "Single-cell specific and interpretable machine learning models for sparse scChIP-seq data imputation". In: *Plos one* 17.7 (2022), e0270043.

[1005] Anjun Ma et al. "Single-cell biological network inference using a heterogeneous graph transformer". In: *Nature Communications* 14.1 (2023), p. 964.

[1006] Yang Eric Li et al. "A comparative atlas of single-cell chromatin accessibility in the human brain". In: *Science* 382.6667 (2023), eadf7044.

[1007] Lucas Seninge et al. "VEGA is an interpretable generative model for inferring biological network activity in single-cell transcriptomics". In: *Nature communications* 12.1 (2021), p. 5684.

[1008] John-William Sidhom et al. "DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires". In: *Nature communications* 12.1 (2021), p. 1605.

[1009] Bohui Li, Maarten Altelaar, and Bas van Breukelen. "Identification of Protein Complexes by Integrating Protein Abundance and Interaction Features Using a Deep Learning Strategy". In: *International Journal of Molecular Sciences* 24.9 (2023), p. 7884.

[1010] Yaosen Min et al. "Learning protein structural fingerprints under the label-free supervision of domain knowledge". In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 69–74.

[1011] Rahul Semwal and Pritish K Varadwaj. "HumDLoc: Human Protein Subcellular Localization Prediction Using Deep Neural Network". In: *Current Genomics* 21.7 (2020), pp. 546–557.

[1012] Xin Deng et al. "An overview of practical applications of protein disorder prediction and drive for faster, more accurate predictions". In: *International journal of molecular sciences* 16.7 (2015), pp. 15384–15404.

[1013] Jack Hanson et al. "SPOT-Disorder2: improved protein intrinsic disorder prediction by ensembled deep learning". In: *Genomics, proteomics & bioinformatics* 17.6 (2019), pp. 645–656.

[1014] Nan Du, Jiayu Shang, and Yanni Sun. "Improving protein domain classification for third-generation sequencing reads using deep learning". In: *BMC genomics* 22.1 (2021), pp. 1–13.

[1015] Nguyen Quoc Khanh Le et al. "Classification of adaptor proteins using recurrent neural networks and PSSM profiles". In: *BMC genomics* 20 (2019), pp. 1–9.

[1016] Manaz Kaleel et al. "SCLpred-MEM: Subcellular localization prediction of membrane proteins by deep N-to-1 convolutional neural networks". In: *Proteins: Structure, Function, and Bioinformatics* 89.10 (2021), pp. 1233–1239.

[1017] Ismail Haberal and Hasan Oul. "Prediction of protein metal binding sites using deep neural networks". In: *Molecular informatics* 38.7 (2019), p. 1800169.

[1018] Di Xiao et al. "SnapKin: a snapshot deep learning ensemble for kinase-substrate prediction from phosphoproteomics data". In: *NAR Genomics and Bioinformatics* 5.4 (2023), lqad099.

[1019] Sam Gelman et al. "Neural networks to learn protein sequence–function relationships from deep mutational scanning data". In: *Proceedings of the National Academy of Sciences* 118.48 (2021), e2104878118.

[1020] Jim Clauwaert et al. "TIS Transformer: remapping the human proteome using deep learning". In: *NAR Genomics and Bioinformatics* 5.1 (2023), lqad021.

[1021] Shilo Ohayon et al. "Simulation of single-protein nanopore sensing shows feasibility for whole-proteome identification". In: *PLoS computational biology* 15.5 (2019), e1007067.

[1022] Will Dampier et al. "HIV-Bidirectional Encoder Representations From Transformers: A Set of Pretrained Transformers for Accelerating HIV Deep Learning Tasks". In: *Frontiers in Virology* 2 (2022), p. 880618.

[1023] Collin Tokheim et al. "Systematic characterization of mutations altering protein degradation in human cancers". In: *Molecular cell* 81.6 (2021), pp. 1292–1308.

[1024] Sini S Raj and SS Vinod Chandra. "Significance of Sequence Features in Classification of Protein–Protein Interactions Using Machine Learning". In: *The Protein Journal* (2023), pp. 1–12.

[1025] Rezarta Islamaj Doan et al. "Overview of the BioCreative VI Precision Medicine Track: mining protein interactions and mutations for precision medicine". In: *Database* 2019 (2019), bay147.

[1026] Yogesh Kalakoti et al. "Modulation of DNA-protein Interactions by Proximal Genetic Elements as Uncovered by Interpretable Deep Learning". In: *Journal of Molecular Biology* 435.13 (2023), p. 168121.

[1027] Shuhui Liu, Zhang Yupei, and Xuequn Shang. "Functional Analysis of Molecular Subtypes with Deep Similarity Learning Model Based on Multi-omics Data". In: *International Conference on Intelligent Computing*. Springer. 2022, pp. 126–137.

[1028] Lorenzo Casalino et al. "AI-driven multiscale simulations illuminate mechanisms of SARS-CoV-2 spike dynamics". In: *The International Journal of High Performance Computing Applications* 35.5 (2021), pp. 432–451.

[1029] Fei He et al. "Extracting molecular entities and their interactions from pathway figures based on deep learning". In: *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. 2019, pp. 397–404.

[1030] Lefteris Koumakis et al. "MinePath: mining for phenotype differential sub-paths in molecular pathways". In: *PLoS computational biology* 12.11 (2016), e1005187.

[1031] Haodong Xu and Zhongming Zhao. "NetBCE: an interpretable deep neural network for accurate prediction of linear B-cell epitopes". In: *Genomics, Proteomics & Bioinformatics* 20.5 (2022), pp. 1002–1012.

[1032] Chao Xia et al. "KaryoNet: Chromosome Recognition with End-to-End Combinatorial Optimization Network". In: *IEEE Transactions on Medical Imaging* (2023).

[1033] Kai-Li Chang et al. "Short human eccDNAs are predictable from sequences". In: *Briefings in Bioinformatics* 24.3 (2023), bbad147.

[1034] Arthur Gilly et al. "Cohort-wide deep whole genome sequencing and the allelic architecture of complex traits". In: *Nature communications* 9.1 (2018), p. 4674.

[1035] Tom W Ouellette and Philip Awadalla. "Inferring ongoing cancer evolution from single tumour biopsies using synthetic supervised learning". In: *PLOS Computational Biology* 18.4 (2022), e1010007.

[1036] Shorya Consul, Ziqi Ke, and Haris Vikalo. "XHap: haplotype assembly using long-distance read correlations learned by transformers". In: *Bioinformatics Advances* 3.1 (2023), vbad169.

[1037] Mattia Antonino Di Gangi et al. "A deep learning network for exploiting positional information in nucleosome related sequences". In: *Bioinformatics and Biomedical Engineering: 5th International Work-Conference, IWBBIO 2017, Granada, Spain, April 26–28, 2017, Proceedings, Part II 5*. Springer. 2017, pp. 524–533.

[1038] Mattia Di Gangi, Giosuè Lo Bosco, and Riccardo Rizzo. "Deep learning architectures for prediction of nucleosome positioning from sequences data". In: *BMC bioinformatics* 19 (2018), pp. 127–135.

[1039] Domenico Amato, Giosue Lo Bosco, and Riccardo Rizzo. "CORENup: a combination of convolutional and recurrent deep neural networks for nucleosome positioning identification". In: *BMC bioinformatics* 21 (2020), pp. 1–14.

[1040] Guo-Sheng Han, Qi Li, and Ying Li. "Nucleosome positioning based on DNA sequence embedding and deep learning". In: *BMC genomics* 23.1 (2022), pp. 1–10.

[1041] Sai Zhang et al. "Analysis of ribosome stalling and translation elongation dynamics by deep learning". In: *Cell systems* 5.3 (2017), pp. 212–220.

[1042] Xin Zeng, Sung-Joon Park, and Kenta Nakai. "Characterizing promoter and enhancer sequences by a deep learning method". In: *Frontiers in Genetics* (2021), p. 1000.

[1043] Yiqiong Chen et al. "AthEDL: Identifying Enhancers in Arabidopsis thaliana Using an Attention-based Deep Learning Method". In: *Current Bioinformatics* 17.6 (2022), pp. 531–540.

[1044] Yifeng Li, Chih-Yu Chen, and Wyeth W Wasserman. "Deep feature selection: theory and application to identify enhancers and promoters". In: *Journal of Computational Biology* 23.5 (2016), pp. 322–336.

[1045] Ying Wang et al. "A successful hybrid deep learning model aiming at promoter identification". In: *BMC bioinformatics* 23.1 (2022), pp. 1–20.

[1046] Xu Min et al. "Predicting enhancers with deep convolutional neural networks". In: *BMC bioinformatics* 18 (2017), pp. 35–46.

[1047] Tianjiao Zhang et al. "DeepITEH: a deep learning framework for identifying tissue-specific eRNAs from the human genome". In: *Bioinformatics* 39.6 (2023), btad375.

[1048] Xu Min et al. "DeepEnhancer: Predicting enhancers by convolutional neural networks". In: *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2016, pp. 637–644.

[1049] Dae Yeong Lim et al. "iEnhancer-RF: Identifying enhancers and their strength by enhanced feature representation using random forest". In: *Chemometrics and Intelligent Laboratory Systems* 212 (2021), p. 104284.

[1050] Muhammad Tahir et al. "An intelligent computational model for prediction of promoters and their strength via natural language processing". In: *Chemometrics and Intelligent Laboratory Systems* 202 (2020), p. 104034.

[1051] Callum MacPhillamy et al. "Cross-species enhancer prediction using machine learning". In: *Genomics* 114.5 (2022), p. 110454.

[1052] Stepan Pachganov et al. "TransPrise: a novel machine learning approach for eukaryotic promoter prediction". In: *PeerJ* 7 (2019), e7990.

[1053] Amandeep Kaur, Ajay Pal Singh Chauhan, and Ashwani Kumar Aggarwal. "Prediction of enhancers in dna sequence data using a hybrid cnn-dlstm model". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 20.2 (2022), pp. 1327–1336.

[1054] Nagina Inayat et al. "iEnhancer-DHF: identification of enhancers and their strengths using optimize deep neural network with multiple features extraction methods". In: *Ieee Access* 9 (2021), pp. 40783–40796.

[1055] Lauro Moraes et al. "CapsProm: a capsule network for promoter prediction". In: *Computers in Biology and Medicine* 147 (2022), p. 105627.

[1056] Hilal Tayara, Muhammad Tahir, and Kil To Chong. "Identification of prokaryotic promoters and their strength by integrating heterogeneous features". In: *Genomics* 112.2 (2020), pp. 1396–1403.

[1057] Quang H Nguyen et al. "iEnhancer-ECNN: identifying enhancers and their strength using ensembles of convolutional neural networks". In: *BMC genomics* 20 (2019), pp. 1–10.

[1058] Li Zeng et al. "iEnhancer-DLRA: identification of enhancers and their strengths by a self-attention fusion strategy for local and global features". In: *Briefings in Functional Genomics* 21.5 (2022), pp. 399–407.

[1059] Wenjun Wang, Qingyao Wu, and Chunshan Li. "iEnhancer-DCSA: identifying enhancers via dual-scale convolution and spatial attention". In: *BMC genomics* 24.1 (2023), p. 393.

[1060] Nguyen Quoc Khanh Le et al. "Classifying promoters by interpreting the hidden information of DNA sequences via deep learning and combination of continuous fasttext N-grams". In: *Frontiers in bioengineering and biotechnology* (2019), p. 305.

[1061] Yansong Wang et al. "Genome-wide identification and characterization of DNA enhancers with a stacked multivariate fusion framework". In: *PLoS Computational Biology* 18.12 (2022), e1010779.

[1062] Ramzan Umarov et al. "Promoter analysis and prediction in the human genome using sequence-based deep learning models". In: *Bioinformatics* 35.16 (2019), pp. 2730–2737.

[1063] Nikita Bhandari et al. "Comparison of machine learning and deep learning techniques in promoter prediction across diverse species". In: *PeerJ Computer Science* 7 (2021), e365.

[1064] Chen Wang et al. "DPProm: a two-layer predictor for identifying promoters and their types on phage genome using deep learning". In: *IEEE Journal of Biomedical and Health Informatics* 26.10 (2022), pp. 5258–5266.

[1065] Xiaoping Min, Fengqing Lu, and Chunyan Li. "Sequence-based deep learning frameworks on enhancer-promoter interactions prediction". In: *Current Pharmaceutical Design* 27.15 (2021), pp. 1847–1855.

[1066] Shashank Singh et al. "Predicting enhancer-promoter interaction from genomic sequence with deep neural networks". In: *Quantitative Biology* 7 (2019), pp. 122–137.

[1067] Xiuchun Lin. "Genomic variation prediction: a summary from different views". In: *Frontiers in Cell and Developmental Biology* 9 (2021), p. 795883.

[1068] Shuai Liu et al. "EPIHC: Improving enhancer-promoter interaction prediction by using hybrid features and communicative learning". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19.6 (2021), pp. 3435–3443.

[1069] Zhong Zhuang, Xiaotong Shen, and Wei Pan. "A simple convolutional neural network for prediction of enhancer–promoter interactions with DNA sequence data". In: *Bioinformatics* 35.17 (2019), pp. 2899–2906.

[1070] Wanwen Zeng, Mengmeng Wu, and Rui Jiang. "Prediction of enhancer-promoter interactions via natural language processing". In: *BMC genomics* 19 (2018), pp. 13–22.

[1071] Zengyan Hong et al. “Identifying enhancer–promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism”. In: *Bioinformatics* 36.4 (2020), pp. 1037–1043.

[1072] Zhangyi Ouyang et al. “Accurate identification of RNA editing sites from primitive sequence with deep neural networks”. In: *Scientific Reports* 8.1 (2018), p. 6005.

[1073] Urminder Singh et al. “Ensemble of deep long short term memory networks for labelling origin of replication sequences”. In: *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE. 2015, pp. 1–7.

[1074] Feng Liu et al. “De novo identification of replication-timing domains in the human genome by deep learning”. In: *Bioinformatics* 32.5 (2016), pp. 641–649.

[1075] Michael A Boemo. “DNAscent v2: detecting replication forks in nanopore sequencing data with deep learning”. In: *BMC genomics* 22.1 (2021), pp. 1–8.

[1076] Elie J Ritch et al. “A generalizable machine learning framework for classifying DNA repair defects using ctDNA exomes”. In: *NPJ Precision Oncology* 7.1 (2023), p. 27.

[1077] Sri Nidhi PV et al. “DNA repair mutation detection using Deep learning strategy–A pharmacogenomic perspective”. In: *2019 Innovations in Power and Advanced Computing Technologies (i-PACT)*. Vol. 1. IEEE. 2019, pp. 1–4.

[1078] Zakaria Louadi et al. “Deep splicing code: Classifying alternative splicing events using deep learning”. In: *Genes* 10.8 (2019), p. 587.

[1079] Michael KK Leung et al. “Deep learning of the tissue-regulated splicing code”. In: *Bioinformatics* 30.12 (2014), pp. i121–i129.

[1080] Sanket Rajan Gupte and Rajgopal Srinivasan. “Joint Prediction of Branchpoint and Splice Acceptor Sites in pre-mRNA Transcripts Using Sequence to Sequence Networks”. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 637–641.

[1081] Anupama Jha, Matthew R Gazzara, and Yoseph Barash. “Integrative deep models for alternative splicing”. In: *Bioinformatics* 33.14 (2017), pp. i274–i282.

[1082] Giulia Riolo, Silvia Cantara, and Claudia Ricci. “Whats wrong in a jump? Prediction and validation of splice site variants”. In: *Methods and Protocols* 4.3 (2021), p. 62.

[1083] Xudong Zou, Xin Gao, and Wei Chen. “Deep learning deepens the analysis of alternative splicing”. In: *Genomics, Proteomics & Bioinformatics* 17.2 (2019), p. 219.

[1084] Da-Cheng Hao et al. “A Global Analysis of Alternative Splicing of Dichocarpum Medicinal Plants, Ranunculales”. In: *Current Genomics* 23.3 (2022), p. 207.

[1085] Feiyang Zhao et al. "Splicing complexity as a pivotal feature of alternative exons in mammalian species". In: *BMC genomics* 24.1 (2023), p. 198.

[1086] Philipp Rentzsch et al. "CADD-Spliceimproving genome-wide variant effect prediction using deep learning-derived splice scores". In: *Genome medicine* 13.1 (2021), pp. 1–12.

[1087] Mhaned Oubounyt et al. "Deep learning models based on distributed feature representations for alternative splicing prediction". In: *IEEE Access* 6 (2018), pp. 58826–58834.

[1088] Victor Akpokiro, Trevor Martin, and Oluwatosin Oluwadare. "EnsembleSplice: ensemble deep learning model for splice site prediction". In: *BMC bioinformatics* 23.1 (2022), p. 413.

[1089] Amin Zabardast et al. "An automated framework for evaluation of deep learning models for splice site predictions". In: *Scientific Reports* 13.1 (2023), p. 10221.

[1090] Indrajit Mandal. "A novel approach for predicting DNA splice junctions using hybrid machine learning algorithms". In: *Soft Computing* 19 (2015), pp. 3431–3444.

[1091] Jasper Zuallaert et al. "SpliceRover: interpretable convolutional neural networks for improved splice site prediction". In: *Bioinformatics* 34.24 (2018), pp. 4180–4188.

[1092] Victor Akpokiro et al. "CNNSplice: Robust Models for Splice Site Prediction Using Convolutional Neural Networks." In: *Computational and Structural Biotechnology Journal* (2023).

[1093] Yi Zhang et al. "Discerning novel splice junctions derived from RNA-seq alignment: a deep learning approach". In: *BMC genomics* 19 (2018), pp. 1–13.

[1094] Shabir Moosa, Prof Abbes Amira, and Dr Sabri Boughorbel. "DASSI: differential architecture search for splice identification from DNA sequences". In: *BioData Mining* 14 (2021), pp. 1–17.

[1095] Kevin Regan, Abolfazl Saghafi, and Zhijun Li. "Splice Junction Identification using Long Short-Term Memory Neural Networks". In: *Current Genomics* 22.5 (2021), p. 384.

[1096] Siguo Wang et al. "Predicting transcription factor binding sites using DNA shape features based on shared hybrid deep learning architecture". In: *Molecular Therapy-Nucleic Acids* 24 (2021), pp. 154–163.

[1097] Dipankar Ranjan Baisya and Stefano Lonardi. "Prediction of histone post-translational modifications using deep learning". In: *Bioinformatics* 36.24 (2020), pp. 5610–5617.

[1098] Yonglin Zhang et al. "Evaluation of deep learning approaches for modeling transcription factor sequence specificity". In: *Genomics* 113.6 (2021), pp. 3774–3781.

[1099] Mike Phuycharoen et al. “Uncovering tissue-specific binding features from differential deep learning”. In: *Nucleic acids research* 48.5 (2020), e27–e27.

[1100] Lijun Quan et al. “Learning useful representations of DNA sequences from ChIP-seq datasets for exploring transcription factor binding specificities”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19.2 (2020), pp. 998–1008.

[1101] Wanwan Ge et al. “Bayesian Markov models improve the prediction of binding motifs beyond first order”. In: *NAR Genomics and Bioinformatics* 3.2 (2021), lqab026.

[1102] Youhong Xu et al. “Using Deep Learning to Predict Transcription Factor Binding Sites Based on Multiple-omics Data”. In: *International Conference on Intelligent Computing*. Springer. 2022, pp. 799–810.

[1103] Hao Li et al. “Inferring transcription factor regulatory networks from single-cell ATAC-seq data based on graph neural networks”. In: *Nature Machine Intelligence* 4.4 (2022), pp. 389–400.

[1104] Dalwinder Singh, Akansha Madhawan, and Joy Roy. “Identification of multiple RNAs using feature fusion”. In: *Briefings in Bioinformatics* 22.6 (2021), bbab178.

[1105] Zixuan Wang et al. “Towards a better understanding of TF-DNA binding prediction from genomic features”. In: *Computers in Biology and Medicine* 149 (2022), p. 105993.

[1106] Erin Jimenez et al. “A regulatory network of Sox and Six transcription factors initiate a cell fate transformation during hearing regeneration in adult zebrafish”. In: *Cell genomics* 2.9 (2022).

[1107] Juexiao Zhou et al. “Annotating TSSs in multiple cell types based on DNA sequence and RNA-seq data via DeeReCT-TSS”. In: *Genomics, Proteomics & Bioinformatics* 20.5 (2022), pp. 959–973.

[1108] Mathys Grapotte et al. “Discovery of widespread transcription initiation at microsatellites predictable by sequence-based deep neural network”. In: *Nature communications* 12.1 (2021), p. 3297.

[1109] Colin Targonski et al. “Cellular state transformations using deep learning for precision medicine applications”. In: *Patterns* 1.6 (2020).

[1110] Nils Wagner et al. “Aberrant splicing prediction across human tissues”. In: *Nature Genetics* (2023), pp. 1–10.

[1111] Raphaël Mourad. “Semi-supervised learning improves regulatory sequence prediction with unlabeled sequences”. In: *BMC bioinformatics* 24.1 (2023), pp. 1–15.

[1112] Koh Onimaru, Osamu Nishimura, and Shigehiro Kuraku. “Predicting gene regulatory regions with a convolutional neural network for processing double-strand genome sequence information”. In: *PloS one* 15.7 (2020), e0235748.

[1113] Mhaned Oubounyt et al. "De-novo reconstruction and identification of transcriptional gene regulatory network modules differentiating single-cell clusters". In: *NAR Genomics and Bioinformatics* 5.1 (2023), lqad018.

[1114] Ramzan Umarov et al. "ReFeaFi: genome-wide prediction of regulatory elements driving transcription initiation". In: *PLoS Computational Biology* 17.9 (2021), e1009376.

[1115] Marco Salvatore et al. "Transfer learning identifies sequence determinants of cell-type specific regulatory element accessibility". In: *NAR genomics and bioinformatics* 5.2 (2023), lqad026.

[1116] Zeyu Chen et al. "Integrative dissection of gene regulatory elements at base resolution". In: *Cell Genomics* 3.6 (2023).

[1117] Hu Zhao et al. "PlantDeepSEA, a deep learning-based web service to predict the regulatory effects of genomic variants in plants". In: *Nucleic Acids Research* 49.W1 (2021), W523–W529.

[1118] Yifeng Li, Wenqiang Shi, and Wyeth W Wasserman. "Genome-wide prediction of cis-regulatory regions using supervised deep learning methods". In: *BMC bioinformatics* 19.1 (2018), pp. 1–14.

[1119] Amalio Telenti et al. "Deep learning of genomic variation and regulatory network data". In: *Human molecular genetics* 27.Supplement_R1 (2018), R63–R71.

[1120] Adam Klie et al. "Predictive analyses of regulatory sequences with EUGENe". In: *Nature Computational Science* 3.11 (2023), pp. 946–956.

[1121] Yifeng Li et al. "The identification of cis-regulatory elements: A review from a machine learning perspective". In: *Biosystems* 138 (2015), pp. 6–17.

[1122] Xin Liu et al. "Learning cis-regulatory principles of ADAR-based RNA editing from CRISPR-mediated mutagenesis". In: *Nature Communications* 12.1 (2021), p. 2165.

[1123] Luca Cappelletti et al. "Bayesian optimization improves tissue-specific prediction of active regulatory regions with deep neural networks". In: *Bioinformatics and Biomedical Engineering: 8th International Work-Conference, IWBBIO 2020, Granada, Spain, May 6–8, 2020, Proceedings 8*. Springer. 2020, pp. 600–612.

[1124] Peter K Koo and Sean R Eddy. "Representation learning of genomic sequence motifs with convolutional neural networks". In: *PLoS computational biology* 15.12 (2019), e1007560.

[1125] Yujia Gao et al. "RicENN: prediction of rice enhancers with neural network based on DNA sequences". In: *Interdisciplinary Sciences: Computational Life Sciences* 14.2 (2022), pp. 555–565.

[1126] Mian Umair Ahsan et al. "NanoCaller for accurate detection of SNPs and indels in difficult-to-map regions from long-read sequencing by haplotype-aware deep neural networks". In: *Genome biology* 22 (2021), pp. 1–33.

[1127] Siavash Manzoori et al. “Detecting SNP markers discriminating horse breeds by deep learning”. In: *Scientific reports* 13.1 (2023), p. 11592.

[1128] Saeed Pirmoradi et al. “A self-organizing deep auto-encoder approach for classification of complex diseases using snp genomics data”. In: *Applied Soft Computing* 97 (2020), p. 106718.

[1129] Sean K Wang et al. “Single-cell multiome of the human retina and deep learning nominate causal variants in complex eye diseases”. In: *Cell genomics* 2.8 (2022).

[1130] Manoj K Mishra et al. “Comparative and Functional Genomic Resource for Mechanistic Studies of Human Blood Pressure–Associated Single Nucleotide Polymorphisms”. In: *Hypertension* 75.3 (2020), pp. 859–868.

[1131] Cameron M Nugent et al. “Genomic and machine learning-based screening of aquaculture-associated introgression into at-risk wild North American Atlantic salmon (Salmo salar) populations”. In: *Molecular Ecology Resources* (2022).

[1132] Casimiro A Curbelo Montañez et al. “SAERMA: stacked autoencoder rule mining algorithm for the interpretation of epistatic interactions in GWAS for extreme obesity”. In: *IEEE Access* 8 (2020), pp. 112379–112392.

[1133] Shima Azizzadeh-Roodpish, Max H Garzon, and Sambriddhi Mainali. “Classifying single nucleotide polymorphisms in humans”. In: *Molecular Genetics and Genomics* 296 (2021), pp. 1161–1173.

[1134] Sapir Labes et al. “Machine-learning of complex evolutionary signals improves classification of SNVs”. In: *NAR Genomics and Bioinformatics* 4.2 (2022), lqac025.

[1135] Sam Friedman et al. “Lean and deep models for more accurate filtering of SNP and INDEL variant calls”. In: *Bioinformatics* 36.7 (2020), pp. 2060–2067.

[1136] Hao Huang et al. “Evaluation of deep learning-based feature selection for single-cell RNA sequencing data analysis”. In: *Genome Biology* 24.1 (2023), p. 259.

[1137] Hongning Zhang et al. “SoCube: an innovative end-to-end doublet detection algorithm for analyzing scRNA-seq data”. In: *Briefings in Bioinformatics* 24.3 (2023), bbad104.

[1138] Li Chen, Ye Wang, and Fengdi Zhao. “Exploiting deep transfer learning for the prediction of functional non-coding variants using genomic sequence”. In: *Bioinformatics* 38.12 (2022), pp. 3164–3172.

[1139] Hilal Tayara and Kil To Chong. “Improving the quantification of DNA sequences using evolutionary information based on deep learning”. In: *Cells* 8.12 (2019), p. 1635.

[1140] Hao Jia, Sung-Joon Park, and Kenta Nakai. “A semi-supervised deep learning approach for predicting the functional effects of genomic non-coding variations”. In: *BMC bioinformatics* 22.6 (2021), pp. 1–11.

[1141] Maria Sindeeva et al. "Cell type–specific interpretation of noncoding variants using deep learning–based methods". In: *GigaScience* 12 (2023), giad015.

[1142] Xiang-He Meng, Hong-Mei Xiao, and Hong-Wen Deng. "Combining artificial intelligence: deep learning with Hi-C data to predict the functional effects of non-coding variants". In: *Bioinformatics* 37.10 (2021), pp. 1339–1344.

[1143] Boqiao Lai et al. "Annotating functional effects of non-coding variants in neuropsychiatric cell types by deep transfer learning". In: *PLoS Computational Biology* 18.5 (2022), e1010011.

[1144] Ning Ai et al. "GDCL-NcDA: identifying non-coding RNA-disease associations via contrastive learning between deep graph learning and deep matrix factorization". In: *BMC genomics* 24.1 (2023), p. 424.

[1145] Tuvshinbayar Chantsalnyam et al. "ncRDense: a novel computational approach for classification of non-coding RNA family by deep learning". In: *Genomics* 113.5 (2021), pp. 3030–3038.

[1146] Alessandro Orro and Gabriele A Trombetti. "High-Accuracy ncRNA Function Prediction via Deep Learning Using Global and Local Sequence Information". In: *Biomedicines* 11.6 (2023), p. 1631.

[1147] Guangsheng Pei et al. "DeepFun: a deep learning sequence-based model to decipher non-coding variant effect in a tissue-and cell type-specific manner". In: *Nucleic acids research* 49.W1 (2021), W131–W139.

[1148] Daniel Quang and Xiaohui Xie. "DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences". In: *Nucleic acids research* 44.11 (2016), e107–e107.

[1149] Lucas F DaSilva et al. "Chromatin landscape distinguishes the genomic loci of hundreds of androgen-receptor-associated LincRNAs from the loci of non-associated LincRNAs". In: *Frontiers in genetics* (2018), p. 132.

[1150] Jael Sanyanda Wekesa, Jun Meng, and Yushi Luan. "A deep learning model for plant lncRNA-protein interaction prediction with graph attention". In: *Molecular Genetics and Genomics* 295 (2020), pp. 1091–1102.

[1151] Yanzhen Xu et al. "Predicting Long non-coding RNAs through feature ensemble learning". In: *BMC genomics* 21.13 (2020), pp. 1–12.

[1152] Ziwei Chen et al. "sORFPred: A Method Based on Comprehensive Features and Ensemble Learning to Predict the sORFs in Plant LncRNAs". In: *Interdisciplinary Sciences: Computational Life Sciences* (2023), pp. 1–13.

[1153] Ahsan Ahmad, Hao Lin, and Swakkhar Shatabda. "Locate-R: subcellular localization of long non-coding RNAs using nucleotide compositions". In: *Genomics* 112.3 (2020), pp. 2583–2589.

[1154] Meng Yang et al. "Integrating convolution and self-attention improves language model of human genome for interpreting non-coding regions at base-resolution". In: *Nucleic acids research* 50.14 (2022), e81–e81.

[1155] Dimitrios Vitsios et al. “Prioritizing non-coding regions based on human genomic constraint and sequence context with deep learning”. In: *Nature communications* 12.1 (2021), p. 1504.

[1156] Alexander Y Lan and M Ryan Corces. “Deep learning approaches for noncoding variant prioritization in neurodegenerative diseases”. In: *Frontiers in Aging Neuroscience* 14 (2022), p. 1027224.

[1157] Xiaoyong Pan et al. “IPMiner: hidden ncRNA-protein interaction sequential pattern mining with stacked autoencoder for accurate computational prediction”. In: *BMC genomics* 17 (2016), pp. 1–14.

[1158] Meihong Gao et al. “GAE-LGA: integration of multi-omics data with graph autoencoders to identify lncRNA–PCG associations”. In: *Briefings in Bioinformatics* 23.6 (2022), bbac452.

[1159] Yingjun Ma. “DeepMNE: deep multi-network embedding for lncRNA-disease association prediction”. In: *IEEE Journal of Biomedical and Health Informatics* 26.7 (2022), pp. 3539–3549.

[1160] Hao Lu et al. “RegVar: Tissue-specific Prioritization of Non-coding Regulatory Variants”. In: *Genomics, Proteomics & Bioinformatics* 21.2 (2023), pp. 385–395.

[1161] Tea Ammunét et al. “Deep learning tools are top performers in long non-coding RNA prediction”. In: *Briefings in Functional Genomics* 21.3 (2022), pp. 230–241.

[1162] Sagar Gupta, Nitesh Kumar Sharma, Ravi Shankar, et al. “DeepPlnc: Bi-modal deep learning for highly accurate plant lncRNA discovery”. In: *Genomics* 114.5 (2022), p. 110443.

[1163] Guishan Zhang et al. “Identifying circular RNA and predicting its regulatory interactions by machine learning”. In: *Frontiers in Genetics* 11 (2020), p. 655.

[1164] Yuning Yang et al. “HCRNet: high-throughput circRNA-binding event identification from CLIP-seq data using deep temporal convolutional network”. In: *Briefings in Bioinformatics* 23.2 (2022), bbac027.

[1165] Miroslav Poláek et al. “Automation of tree-ring detection and measurements using deep learning”. In: *Methods in Ecology and Evolution* 14.9 (2023), pp. 2233–2242.

[1166] Gabriela A Merino et al. “Novel SARS-CoV-2 encoded small RNAs in the passage to humans”. In: *Bioinformatics* 36.24 (2020), pp. 5571–5581.

[1167] Rishab Jain et al. “ICOR: improving codon optimization with recurrent neural networks”. In: *BMC bioinformatics* 24.1 (2023), p. 132.

[1168] Yi Li et al. “Combined genome-wide association study of 136 quantitative ear morphology traits in multiple populations reveal 8 novel loci”. In: *PLoS Genetics* 19.7 (2023), e1010786.

[1169] Binh P Nguyen et al. "iProDNA-CapsNet: identifying protein-DNA binding residues using capsule neural networks". In: *BMC bioinformatics* 20 (2019), pp. 1–12.

[1170] Shuzhen Kuang and Liangjiang Wang. "Identification and analysis of consensus RNA motifs binding to the genome regulator CTCF". In: *NAR Genomics and Bioinformatics* 2.2 (2020), lqaa031.

[1171] Ariel Erijman et al. "A high-throughput screen for transcription activation domains reveals their sequence features and permits prediction by deep learning". In: *Molecular cell* 78.5 (2020), pp. 890–902.

[1172] Abdul Wahab et al. "DNA sequences performs as natural language processing by exploiting deep learning algorithm for the identification of N4-methylcytosine". In: *Scientific reports* 11.1 (2021), p. 212.

[1173] Haodong Xu, Peilin Jia, and Zhongming Zhao. "Deep4mC: systematic assessment and computational prediction for DNA N4-methylcytosine sites by deep learning". In: *Briefings in Bioinformatics* 22.3 (2021), bbaa099.

[1174] Thanh-Hoang Nguyen-Vo et al. "i4mC-GRU: Identifying DNA N4-Methylcytosine sites in mouse genomes using bidirectional gated recurrent unit and sequence-embedded features". In: *Computational and Structural Biotechnology Journal* 21 (2023), pp. 3045–3053.

[1175] Jhabindra Khanal et al. "Identifying DNA N4-methylcytosine sites in the rosaceae genome with a deep learning model relying on distributed feature representation". In: *Computational and Structural Biotechnology Journal* 19 (2021), pp. 1612–1619.

[1176] Lezheng Yu et al. "Systematic analysis and accurate identification of DNA N4-methylcytosine sites by deep learning". In: *Frontiers in microbiology* 13 (2022), p. 843425.

[1177] Peng Ni et al. "DNA 5-methylcytosine detection and methylation phasing using PacBio circular consensus sequencing". In: *Nature communications* 14.1 (2023), p. 4054.

[1178] Hong Wang et al. "MTDeepM6A-2S: A two-stage multi-task deep learning method for predicting RNA N6-methyladenosine sites of Saccharomyces cerevisiae". In: *Frontiers in Microbiology* 13 (2022), p. 999506.

[1179] Yiqian Zhang and Michiaki Hamada. "DeepM6ASeq: prediction and characterization of m6A-containing sequences using deep learning". In: *BMC bioinformatics* 19 (2018), pp. 1–11.

[1180] Sehi Park et al. "i6mA-DNC: Prediction of DNA N6-Methyladenosine sites in rice genome based on dinucleotide representation using deep learning". In: *Chemometrics and Intelligent Laboratory Systems* 204 (2020), p. 104102.

[1181] Mobeen Ur Rehman, Hilal Tayara, and Kil To Chong. "DL-m6A: Identification of N6-methyladenosine Sites in Mammals using deep learning based on different encoding schemes". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 20.2 (2022), pp. 904–911.

[1182] Muhammad Tahir et al. "A deep learning-based computational approach for discrimination of dna n6-methyladenosine sites by fusing heterogeneous features". In: *Chemometrics and Intelligent Laboratory Systems* 206 (2020), p. 104151.

[1183] Jun Wang and Liangjiang Wang. "Deep analysis of RNA N6-adenosine methylation (m6A) patterns in human cells". In: *NAR genomics and bioinformatics* 2.1 (2020), lqaa007.

[1184] Lin Zhang et al. "EDLm6APred: ensemble deep learning approach for mRNA m6A site prediction". In: *BMC bioinformatics* 22.1 (2021), p. 288.

[1185] Eva Klimentova et al. "PENGUINN: Precise exploration of nuclear G-quadruplexes using interpretable neural networks". In: *Frontiers in genetics* 11 (2020), p. 568546.

[1186] Mira Barshai, Alice Aubert, and Yaron Orenstein. "G4detector: convolutional neural network to predict DNA G-quadruplexes". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19.4 (2021), pp. 1946–1955.

[1187] Haopeng Yu et al. "G4Atlas: a comprehensive transcriptome-wide G-quadruplex database". In: *Nucleic Acids Research* 51.D1 (2023), pp. D126–D134.

[1188] Zhen Chen et al. "Integration of a deep learning classifier with a random forest approach for predicting malonylation sites". In: *Genomics, proteomics & bioinformatics* 16.6 (2018), pp. 451–459.

[1189] Qi Tian et al. "MRCNN: a deep learning model for regression of genome-wide DNA methylation". In: *BMC genomics* 20.2 (2019), pp. 1–10.

[1190] Peng Ni et al. "Genome-wide detection of cytosine methylations in plant from Nanopore data using deep learning". In: *Nature communications* 12.1 (2021), p. 5976.

[1191] Zutan Li et al. "Deep6mA: a deep learning framework for exploring similar patterns in DNA N6-methyladenine sites across different species". In: *PLoS computational biology* 17.2 (2021), e1008767.

[1192] Lokuthota Hewage Roland and Champi Thusangi Wannige. "A deep learning model for predicting DNA N6-methyladenine (6mA) sites in eukaryotes". In: *IEEE Access* 8 (2020), pp. 175535–175545.

[1193] Ying Zhang et al. "Leveraging the attention mechanism to improve the identification of DNA N6-methyladenine sites". In: *Briefings in Bioinformatics* 22.6 (2021), bbab351.

[1194] Xingyu Tang et al. "LangMoDHS: A deep learning language model for predicting DNase I hypersensitive sites in mouse genome". In: *Mathematical biosciences and engineering: MBE* 20.1 (2023), pp. 1037–1057.

[1195] Chuqiao Lyu, Lei Wang, and Juhua Zhang. "Deep learning for DNase I hypersensitive sites identification". In: *BMC genomics* 19.10 (2018), pp. 155–165.

[1196] Nazar Beknazarov, Seungmin Jin, and Maria Poptsova. "Deep learning approach for predicting functional Z-DNA regions using omics data". In: *Scientific Reports* 10.1 (2020), p. 19134.

[1197] Zhongxiao Li et al. "DeeReCT-APA: prediction of alternative polyadenylation site usage through deep learning". In: *Genomics, Proteomics and Bioinformatics* 20.3 (2022), pp. 483–495.

[1198] Arturo Magana-Mora, Manal Kalkatawi, and Vladimir B Bajic. "OmniPolyA: a method and tool for accurate recognition of Poly (A) signals in human genomic DNA". In: *BMC genomics* 18 (2017), pp. 1–13.

[1199] Yanbu Guo et al. "Deep Effective k-mer representation learning for polyadenylation signal prediction via co-occurrence embedding". In: *Knowledge-Based Systems* 256 (2022), p. 109887.

[1200] Fahad Albalawi et al. "Hybrid model for efficient prediction of poly (A) signals in human genomic DNA". In: *Methods* 166 (2019), pp. 31–39.

[1201] Xin Gao et al. "DeepPolyA: a convolutional neural network approach for polyadenylation site prediction". In: *IEEE Access* 6 (2018), pp. 24340–24349.

[1202] Michael KK Leung, Andrew Delong, and Brendan J Frey. "Inference of the human polyadenylation code". In: *Bioinformatics* 34.17 (2018), pp. 2889–2898.

[1203] Yanbu Guo et al. "Identifying polyadenylation signals with biological embedding via self-attentive gated convolutional highway networks". In: *Applied Soft Computing* 103 (2021), p. 107133.

[1204] Johannes Linder et al. "Deciphering the impact of genetic variation on human polyadenylation using APARENT2". In: *Genome Biology* 23.1 (2022), pp. 1–33.

[1205] Emma Bohn et al. "A curated census of pathogenic and likely pathogenic UTR variants and evaluation of deep learning models for variant effect prediction". In: *Frontiers in Molecular Biosciences* 10 (2023).

[1206] Jian-Min Chen et al. "The experimentally obtained functional impact assessments of 5'splice site GT> GC variants differ markedly from those predicted". In: *Current Genomics* 21.1 (2020), pp. 56–66.

[1207] Blake Pyman et al. "Exploring microRNA regulation of cancer with context-aware deep cancer classifier". In: *BIOCOMPUTING 2019: Proceedings of the Pacific Symposium*. World Scientific. 2018, pp. 160–171.

[1208] Philip H Williams, Rod Eyles, Georg Weiller, et al. "Plant microRNA prediction by supervised machine learning using C5. 0 decision trees". In: *Journal of nucleic acids* 2012 (2012).

[1209] Lujia Chen and Xinghua Lu. "Discovering functional impacts of miRNAs in cancers using a causal deep learning model". In: *BMC medical genomics* 11.6 (2018), pp. 53–63.

[1210] Xubo Tang and Yanni Sun. "Fast and accurate microRNA search using CNN". In: *BMC bioinformatics* 20.23 (2019), pp. 1–14.

[1211] Li Guo and Zuhong Lu. "Global expression analysis of miRNA gene cluster and family based on isomiRs from deep sequencing data". In: *Computational biology and chemistry* 34.3 (2010), pp. 165–171.

[1212] Santos Kumar Baliarsingh et al. "Analysis of high-dimensional genomic data using MapReduce based probabilistic neural network". In: *Computer methods and programs in biomedicine* 195 (2020), p. 105625.

[1213] Sagar Gupta and Ravi Shankar. "miWords: transformer-based composite deep learning for highly accurate discovery of pre-miRNA regions across plant genomes". In: *Briefings in Bioinformatics* 24.2 (2023), bbad088.

[1214] Susan Higashi et al. "Mirinho: an efficient and general plant and animal pre-miRNA predictor for genomic and deep sequencing data". In: *BMC bioinformatics* 16.1 (2015), pp. 1–14.

[1215] Georgina Stegmayer et al. "High class-imbalance in pre-miRNA prediction: a novel approach based on deepSOM". In: *IEEE/ACM transactions on computational biology and bioinformatics* 14.6 (2016), pp. 1316–1326.

[1216] Jonathan Raad, Georgina Stegmayer, and Diego H Milone. "Complexity measures of the mature miRNA for improving pre-miRNAs prediction". In: *Bioinformatics* 36.8 (2020), pp. 2319–2327.

[1217] Shinya Tasaki et al. "Inferring protein expression changes from mRNA in Alzheimers dementia using deep neural networks". In: *Nature Communications* 13.1 (2022), p. 655.

[1218] Juan Gutiérrez-Cárdenas and Zenghui Wang. "Classification of breast cancer and breast neoplasm scenarios based on machine learning and sequence features from lncRNAs–miRNAs-diseases associations". In: *Interdisciplinary Sciences: Computational Life Sciences* 13.4 (2021), pp. 572–581.

[1219] Pengyu Zhang and Hao Wu. "IChrom-deep: an attention-based deep learning model for identifying chromatin interactions". In: *IEEE Journal of Biomedical and Health Informatics* (2023).

[1220] Aman Agarwal and Li Chen. "DeepPHiC: Predicting promoter-centered chromatin interactions using a novel deep learning approach". In: *Bioinformatics* 39.1 (2023), btac801.

[1221] Zhenhao Zhang, Fan Feng, and Jie Liu. "Characterizing collaborative transcription regulation with a graph-based deep learning approach". In: *PLOS Computational Biology* 18.6 (2022), e1010162.

[1222] Haitham Ashoor et al. "Graph embedding and unsupervised learning predict genomic sub-compartments from HiC chromatin interaction data". In: *Nature communications* 11.1 (2020), p. 1173.

[1223] Alireza Karbalayghareh, Merve Sahin, and Christina S Leslie. "Chromatin interaction–aware gene regulatory modeling with graph attention networks". In: *Genome Research* 32.5 (2022), pp. 930–944.

[1224] Mateusz Chiliski, Anup Kumar Halder, and Dariusz Plewczynski. "Prediction of chromatin looping using deep hybrid learning (DHL)". In: *Quantitative Biology* 11.2 (2023), pp. 155–162.

[1225] Fuzhou Wang et al. "A Lightweight Framework For Chromatin Loop Detection at the Single-Cell Level". In: *Advanced Science* 10.33 (2023), p. 2303502.

[1226] Hao Wu et al. "Be-1DCNN: a neural network model for chromatin loop prediction based on bagging ensemble learning". In: *Briefings in Functional Genomics* (2023), elad015.

[1227] Hao Lv et al. "A sequence-based deep learning approach to predict CTCF-mediated chromatin loop". In: *Briefings in bioinformatics* 22.5 (2021), bbab031.

[1228] Lezheng Yu et al. "Evaluation and development of deep neural networks for RNA 5-Methyluridine classifications using autoBioSeqpy". In: *Frontiers in Microbiology* 14 (2023), p. 1175925.

[1229] Marcin Cielik and Stefan Bekiranov. "Combinatorial epigenetic patterns as quantitative predictors of chromatin biology". In: *BMC genomics* 15.1 (2014), pp. 1–18.

[1230] Francesco Gandolfi and Anna Tramontano. "A computational approach for the functional classification of the epigenome". In: *Epigenetics & Chromatin* 10 (2017), pp. 1–24.

[1231] Pavel Avdeyev and Jian Zhou. "Computational Approaches for Understanding Sequence Variation Effects on the 3D Genome Architecture". In: *Annual Review of Biomedical Data Science* 5 (2022), pp. 183–204.

[1232] Jian Zhou. "Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale". In: *Nature genetics* 54.5 (2022), pp. 725–734.

[1233] Ron Schwessinger et al. "DeepC: predicting 3D genome folding using megabase-scale transfer learning". In: *Nature methods* 17.11 (2020), pp. 1118–1124.

[1234] Zhe Cheng et al. "ReHiC: Enhancing Hi-C data resolution via residual convolutional network". In: *Journal of Bioinformatics and Computational Biology* 19.02 (2021), p. 2150001.

[1235] Tianyin Zhou et al. "Quantitative modeling of transcription factor binding specificities using DNA shape". In: *Proceedings of the National Academy of Sciences* 112.15 (2015), pp. 4654–4659.

[1236] Laura M Gunsalus, Michael J Keiser, and Katherine S Pollard. "In silico discovery of repetitive elements as key sequence determinants of 3D genome folding". In: *Cell Genomics* 3.10 (2023).

[1237] Daniele DAgostino et al. "Advantages of using graph databases to explore chromatin conformation capture experiments". In: *BMC bioinformatics* 22.2 (2021), pp. 1–16.

[1238] Kathleen C Keough et al. “Three-dimensional genome rewiring in loci with human accelerated regions”. In: *Science* 380.6643 (2023), eabm1696.

[1239] Siguo Wang et al. “DLoopCaller: A deep learning approach for predicting genome-wide chromatin loops by integrating accessible chromatin landscapes”. In: *PLoS Computational Biology* 18.10 (2022), e1010572.

[1240] Nguyen Phuoc Nguyen et al. “DRPnet: automated particle picking in cryo-electron micrographs using deep regression”. In: *BMC bioinformatics* 22 (2021), pp. 1–28.

[1241] Jan Zrimec and Aleksej Zelezniak. “Learning the regulatory grammar of DNA for gene expression engineering”. In: *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. 2020, pp. 1–1.

[1242] Niloofar Borhani et al. “A deep learning approach to predict inter-omics interactions in multi-layer networks”. In: *BMC bioinformatics* 23.1 (2022), pp. 1–17.

[1243] Eeshit Dhaval Vaishnav et al. “The evolution, evolvability and engineering of gene regulatory DNA”. In: *Nature* 603.7901 (2022), pp. 455–463.

[1244] Yunhe Wang et al. “scBGEDA: deep single-cell clustering analysis via a dual denoising autoencoder with bipartite graph ensemble clustering”. In: *Bioinformatics* 39.2 (2023), btad075.

[1245] Jiawei Mao et al. “A novel method for identifying key genes in macroevolution based on deep learning with attention mechanism”. In: *Scientific Reports* 13.1 (2023), p. 19727.

[1246] Tulio L Campos et al. “Combined use of feature engineering and machine-learning to predict essential genes in Drosophila melanogaster”. In: *NAR Genomics and Bioinformatics* 2.3 (2020), lqaa051.

[1247] Xue Zhang, Wangxin Xiao, and Weijia Xiao. “DeepHE: Accurately predicting human essential genes based on deep learning”. In: *PLOS Computational Biology* 16.9 (2020), e1008229.

[1248] Qian Li. “scTour: a deep learning architecture for robust inference and accurate prediction of cellular dynamics”. In: *Genome Biology* 24.1 (2023), pp. 1–33.

[1249] Zicheng Zhang et al. “Pan-cancer landscape of T-cell exhaustion heterogeneity within the tumor microenvironment revealed a progressive roadmap of hierarchical dysfunction associated with prognosis and therapeutic efficacy”. In: *EBioMedicine* 83 (2022).

[1250] Tianshi Lu et al. “Deep learning-based prediction of the T cell receptor–antigen binding specificity”. In: *Nature machine intelligence* 3.10 (2021), pp. 864–875.

[1251] Fan Xu et al. “Ensemble learning models that predict surface protein abundance from single-cell multimodal omics data”. In: *Methods* 189 (2021), pp. 65–73.

[1252] Xiang Cheng, Xuan Xiao, and Kuo-Chen Chou. "pLoc-mGneg: Predict subcellular localization of Gram-negative bacterial proteins by deep gene ontology learning via general PseAAC". In: *Genomics* 110.4 (2018), pp. 231–239.

[1253] Meet Barot et al. "NetQuilt: deep multispecies network-based protein function prediction using homology-informed network similarity". In: *Bioinformatics* 37.16 (2021), pp. 2414–2422.

[1254] Sarthak Mishra et al. "A deep learning ensemble for function prediction of hypothetical proteins from pathogenic bacterial species". In: *Computational Biology and Chemistry* 83 (2019), p. 107147.

[1255] Vikas Chauhan et al. "Multi-label classifier for protein sequence using heuristic-based deep convolution neural network". In: *Applied Intelligence* 52.3 (2022), pp. 2820–2837.

[1256] Chenguang Zhao, Tong Liu, and Zheng Wang. "PANDA2: protein function prediction using graph neural networks". In: *NAR genomics and bioinformatics* 4.1 (2022), lqac004.

[1257] Liangzhi Li et al. "AI-driven pan-proteome analyses reveal insights into the biohydrometallurgical properties of Acidithiobacillia". In: *Frontiers in Microbiology* 14 (2023), p. 1243987.

[1258] Fares Saidi et al. "Bacterial outer membrane polysaccharide export (OPX) proteins occupy three structural classes with selective beta-barrel porin requirements for polymer secretion". In: *Microbiology Spectrum* 10.5 (2022), e01290–22.

[1259] Jeffrey A Ruffolo and Jeffrey J Gray. "Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies". In: *Biophysical Journal* 121.3 (2022), 155a–156a.

[1260] Wenjing Xuan et al. "Clpred: a sequence-based protein crystallization predictor using blstm neural network". In: *Bioinformatics* 36.Supplement_2 (2020), pp. i709–i717.

[1261] Abdurrahman Elbasir et al. "DeepCrystal: a deep learning framework for sequence-based protein crystallization prediction". In: *Bioinformatics* 35.13 (2019), pp. 2216–2225.

[1262] Yi-Heng Zhu et al. "Accurate multistage prediction of protein crystallization propensity using deep-cascade forest with sequence-based features". In: *Briefings in Bioinformatics* 22.3 (2021), bbaa076.

[1263] Chiranjib Chakraborty et al. "Structural Landscape of nsp Coding Genomic Regions of SARS-CoV-2-ssRNA Genome: A Structural Genomics Approach Toward Identification of Druggable Genome, Ligand-Binding Pockets, and Structure-Based Druggability". In: *Molecular Biotechnology* (2022), pp. 1–22.

[1264] Qian Wang et al. “Prediction of transport proteins from sequence information with the deep learning approach”. In: *Computers in Biology and Medicine* 160 (2023), p. 106974.

[1265] Pier Luigi Martelli et al. “Computer-aided prediction of protein mitochondrial localization”. In: *Mitochondrial Medicine: Volume 1: Targeting Mitochondria*. Springer, 2021, pp. 433–452.

[1266] Fei He et al. “Protein ubiquitylation and sumoylation site prediction based on ensemble and transfer learning”. In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2019, pp. 117–123.

[1267] Niraj Thapa et al. “DeepSuccinylSite: a deep learning based approach for protein succinylation site prediction”. In: *BMC bioinformatics* 21.3 (2020), pp. 1–10.

[1268] Yubin Xie et al. “DeepNitro: prediction of protein nitration and nitrosylation sites by deep learning”. In: *Genomics, proteomics & bioinformatics* 16.4 (2018), pp. 294–306.

[1269] Qijin Yin et al. “DeepHistone: a deep learning approach to predicting histone modifications”. In: *BMC genomics* 20.2 (2019), pp. 11–23.

[1270] Felix Teufel et al. “DeepPeptide predicts cleaved peptides in proteins using conditional random fields”. In: *Bioinformatics* 39.10 (2023), btad616.

[1271] Filippo Grazioli et al. “On TCR binding predictors failing to generalize to unseen peptides”. In: *Frontiers in Immunology* 13 (2022), p. 1014256.

[1272] Ying Xu et al. “AttnTAP: A Dual-input Framework Incorporating the Attention Mechanism for Accurately Predicting TCR-peptide Binding”. In: *Frontiers in Genetics* 13 (2022), p. 942491.

[1273] Gene Sher, Degui Zhi, and Shaojie Zhang. “DRREP: deep ridge regressed epitope predictor”. In: *BMC genomics* 18.6 (2017), pp. 55–65.

[1274] Chris Fotis et al. “DeepSIBA: chemical structure-based inference of biological alterations using deep learning”. In: *Molecular Omics* 17.1 (2021), pp. 108–120.

[1275] Tian Cai et al. “End-to-end sequence-structure-function meta-learning predicts genome-wide chemical-protein interactions for dark proteins”. In: *PLOS Computational Biology* 19.1 (2023), e1010851.

[1276] Tian Cai et al. “Msa-regularized protein sequence transformer toward predicting genome-wide chemical-protein interactions: Application to gpcrome deorphanization”. In: *Journal of chemical information and modeling* 61.4 (2021), pp. 1570–1582.

[1277] John Jumper et al. “Highly accurate protein structure prediction with AlphaFold”. In: *Nature* 596.7873 (2021), pp. 583–589.

[1278] Andrew W Senior et al. “Improved protein structure prediction using potentials from deep learning”. In: *Nature* 577.7792 (2020), pp. 706–710.

[1279] Daria Gutnik et al. "Using AlphaFold Predictions in Viral Research". In: *Current Issues in Molecular Biology* 45.4 (2023), pp. 3705–3732.

[1280] Max Kilian and Ilka B Bischofs. "Co-evolution at protein–protein interfaces guides inference of stoichiometry of oligomeric protein complexes by de novo structure prediction". In: *Molecular Microbiology* 120.5 (2023), pp. 763–782.

[1281] Vanessa Shivnauth et al. "Structural diversity and stress regulation of the plant immunity-associated calmodulin-binding protein 60 (CBP60) family of transcription factors in Solanum lycopersicum (tomato)". In: *Functional & Integrative Genomics* 23.3 (2023), p. 236.

[1282] Christina B Azodi, John P Lloyd, and Shin-Han Shiu. "The cis-regulatory codes of response to combined heat and drought stress in Arabidopsis thaliana". In: *NAR Genomics and Bioinformatics* 2.3 (2020), lqaa049.

[1283] Bader S Alotaibi et al. "New drug target identification in Vibrio vulnificus by subtractive genome analysis and their inhibitors through molecular docking and molecular dynamics simulations". In: *Heliyon* (2023), e17650.

[1284] Oz Kilim et al. "SARS-CoV-2 receptor-binding domain deep mutational AlphaFold2 structures". In: *Scientific Data* 10.1 (2023), p. 134.

[1285] Mu Gao et al. "Proteome-scale deployment of protein structure prediction workflows on the summit supercomputer". In: *2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. IEEE. 2022, pp. 206–215.

[1286] Yosuke Oyama, Akihiro Tabuchi, and Atsushi Tokuhisa. "Accelerating AlphaFold2 Inference of Protein Three-Dimensional Structure on the Supercomputer Fugaku". In: *Proceedings of the 13th Workshop on AI and Scientific Computing at Scale using Flexible Computing*. 2023, pp. 1–9.

[1287] Wei Zheng et al. "Improving deep learning protein monomer and complex structure prediction using DeepMSA2 with huge metagenomics data". In: *Nature Methods* (2024), pp. 1–11.

[1288] Shiran Barber-Zucker et al. "Stable and functionally diverse versatile peroxidases designed directly from sequences". In: *Journal of the American Chemical Society* 144.8 (2022), pp. 3564–3571.

[1289] Haicang Zhang and Yufeng Shen. "Template-based prediction of protein structure with deep learning". In: *BMC genomics* 21.11 (2020), pp. 1–9.

[1290] Laiyi Fu et al. "UFold: fast and accurate RNA secondary structure prediction with deep learning". In: *Nucleic acids research* 50.3 (2022), e14–e14.

[1291] Magnus Haraldson Høie et al. "Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and conservation". In: *Cell reports* 38.2 (2022).

[1292] Jun Cheng et al. "MMSplice: modular modeling improves the predictions of genetic variant effects on splicing". In: *Genome biology* 20 (2019), pp. 1–15.

[1293] Patricia J Sullivan et al. “Introme accurately predicts the impact of coding and noncoding variants on gene splicing, with clinical applications”. In: *Genome Biology* 24.1 (2023), pp. 1–18.

[1294] Woori Jang et al. “Comparison of in silico tools for splice-altering variant prediction using established spliceogenic variants: an end-users point of view”. In: *International Journal of Genomics* 2022 (2022).

[1295] Jean-Madeleine de Sainte Agathe et al. “SpliceAI-visual: a free online tool to improve SpliceAI splicing variant interpretation”. In: *Human Genomics* 17.1 (2023), pp. 1–16.

[1296] Hao Chen et al. “DIFFUSE: predicting isoform functions from sequences and expression profiles via deep learning”. In: *Bioinformatics* 35.14 (2019), pp. i284–i294.

[1297] Bahrad A Sokhansanj, Zhengqiao Zhao, and Gail L Rosen. “Interpretable and predictive deep neural network modeling of the SARS-Cov-2 spike protein sequence to predict COVID-19 disease severity”. In: *Biology* 11.12 (2022), p. 1786.

[1298] Haicang Zhang et al. “Predicting functional effect of missense variants using graph attention neural networks”. In: *Nature Machine Intelligence* 4.11 (2022), pp. 1017–1028.

[1299] Jinghao Peng et al. “Predicting chromosome flexibility from the genomic sequence based on deep learning neural networks”. In: *Current Bioinformatics* 16.10 (2021), pp. 1311–1319.

[1300] Weiqin Yang, Dexin Li, and Ranran Huang. “EVMP: Enhancing machine learning models for synthetic promoter strength prediction by Extended Vision Mutant Priority framework”. In: *Frontiers in Microbiology* 14 (2023).

[1301] Wenxuan Xu, Lin Zhu, and De-Shuang Huang. “DCDE: an efficient deep convolutional divergence encoding method for human promoter recognition”. In: *IEEE transactions on nanobioscience* 18.2 (2019), pp. 136–145.

[1302] Pora Kim et al. “FusionAI: predicting fusion breakpoint from DNA sequence with deep learning”. In: *Iscience* 24.10 (2021).

[1303] Pora Kim et al. “FusionGDB 2.0: fusion gene annotation updates aided by deep learning”. In: *Nucleic Acids Research* 50.D1 (2022), pp. D1221–D1230.

[1304] Gancheng Zhu et al. “GSRNet, an adversarial training-based deep framework with multi-scale CNN and BiGRU for predicting genomic signals and regions”. In: *Expert Systems with Applications* 229 (2023), p. 120439.

[1305] Yanbu Guo et al. “Deep multi-scale Gaussian residual networks for contextual-aware translation initiation site recognition”. In: *Expert Systems with Applications* 207 (2022), p. 118004.

[1306] Utku Ozbulak et al. “Mutate and observe: utilizing deep neural networks to investigate the impact of mutations on translation initiation”. In: *Bioinformatics* 39.6 (2023), btad338.

[1307] Chao Wei, Junying Zhang, and Yuan Xiguo. "DeepTIS: Improved translation initiation site prediction in genomic sequence via a two-stage deep learning model". In: *Digital Signal Processing* 117 (2021), p. 103202.

[1308] Yanbu Guo et al. "Gated residual neural networks with self-normalization for translation initiation site recognition". In: *Knowledge-Based Systems* 237 (2022), p. 107783.

[1309] Sambit K Mishra, Viraj Muthye, and Gaurav Kandoi. "Computational methods for predicting functions at the mRNA isoform level". In: *International Journal of Molecular Sciences* 21.16 (2020), p. 5686.

[1310] Fei Zhu et al. "Leveraging protein dynamics to identify functional phosphorylation sites using deep learning models". In: *Journal of Chemical Information and Modeling* 62.14 (2022), pp. 3331–3345.

[1311] Mingxin Gan, Wenran Li, and Rui Jiang. "EnContact: predicting enhancer-enhancer contacts using sequence-based deep learning model". In: *PeerJ* 7 (2019), e7657.

[1312] Samuel Khodursky et al. "The evolution and mutational robustness of chromatin accessibility in Drosophila". In: *Genome Biology* 24.1 (2023), p. 232.

[1313] Xu Min et al. "Chromatin accessibility prediction via convolutional long short-term memory networks with k-mer embedding". In: *Bioinformatics* 33.14 (2017), pp. i92–i101.

[1314] Yanbu Guo et al. "Attentive gated neural networks for identifying chromatin accessibility". In: *Neural Computing and Applications* 32 (2020), pp. 15557–15571.

[1315] Ha TH Vu et al. "Unsupervised contrastive peak caller for ATAC-seq". In: *Genome Research* (2023), gr–277677.

[1316] Yanbu Guo et al. "DeepANF: A deep attentive neural framework with distributed representation for chromatin accessibility prediction". In: *Neurocomputing* 379 (2020), pp. 305–318.

[1317] Qiao Liu et al. "DeepCAGE: incorporating transcription factors in genome-wide prediction of chromatin accessibility". In: *Genomics, Proteomics and Bioinformatics* 20.3 (2022), pp. 496–507.

[1318] iga Avsec et al. "Base-resolution models of transcription-factor binding reveal soft motif syntax". In: *Nature Genetics* 53.3 (2021), pp. 354–366.

[1319] Canbiao Wu et al. "DeepHBV: a deep learning model to predict hepatitis B virus (HBV) integration sites". In: *BMC Ecology and Evolution* 21.1 (2021), pp. 1–10.

[1320] Hailin Hu et al. "DeepHINT: understanding HIV-1 integration via deep learning with attention". In: *Bioinformatics* 35.10 (2019), pp. 1660–1667.

[1321] Jiuxing Liang et al. "DeepEBV: a deep learning model to predict Epstein–Barr virus (EBV) integration sites". In: *Bioinformatics* 37.20 (2021), pp. 3405–3411.

[1322] Joseph M Taft et al. "Deep mutational learning predicts ACE2 binding and antibody escape to combinatorial mutations in the SARS-CoV-2 receptor-binding domain". In: *Cell* 185.21 (2022), pp. 4008–4022.

[1323] Ying He et al. "Predicting the sequence specificities of DNA-binding proteins by DNA fine-tuned language model with decaying learning rates". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 20.1 (2022), pp. 616–624.

[1324] Qinhu Zhang, Zhen Shen, and De-Shuang Huang. "Modeling in-vivo protein-DNA binding by combining multiple-instance learning with a hybrid deep neural network". In: *Scientific reports* 9.1 (2019), p. 8484.

[1325] Daniel Quang and Xiaohui Xie. "FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data". In: *Methods* 166 (2019), pp. 40–47.

[1326] Tianqi Yang and Ricardo Henao. "TAMC: A deep-learning approach to predict motif-centric transcriptional factor binding activity based on ATAC-seq profile". In: *PLoS Computational Biology* 18.9 (2022), e1009921.

[1327] Chen Chen et al. "DeepGRN: prediction of transcription factor binding site across cell-types using attention-based deep neural networks". In: *BMC bioinformatics* 22.1 (2021), pp. 1–18.

[1328] Alexander Gulliver Bjørnholt Grønning et al. "DeepCLIP: predicting the effect of mutations on protein–RNA binding with deep learning". In: *Nucleic acids research* 48.13 (2020), pp. 7099–7118.

[1329] Chandra Mohan Dasari, Santhosh Amilpur, and Raju Bhukya. "Exploring variable-length features (motifs) for predicting binding sites through interpretable deep neural networks". In: *Engineering Applications of Artificial Intelligence* 106 (2021), p. 104485.

[1330] Vincentius Martin et al. "QBiC-Pred: quantitative predictions of transcription factor binding changes due to sequence variants". In: *Nucleic acids research* 47.W1 (2019), W127–W135.

[1331] Meghana Kshirsagar et al. "BindVAE: Dirichlet variational autoencoders for de novo motif discovery from accessible chromatin". In: *Genome Biology* 23.1 (2022), p. 174.

[1332] Shabir Moosa, Sabri Boughorbel, and Abbes Amira. "Differential Architecture Search in Deep Learning for Genomic Applications". In: *2019 IEEE 10th GCC Conference & Exhibition (GCC)*. IEEE. 2019, pp. 1–6.

[1333] Jeremy WK Ng et al. "Deep learning for de-convolution of Smad2 versus Smad3 binding sites". In: *BMC genomics* 23.1 (2022), pp. 1–11.

[1334] Kanu Geete and Manish Pandey. "Robust transcription factor binding site prediction using deep neural networks". In: *Current Bioinformatics* 15.10 (2020), pp. 1137–1152.

[1335] Xiao Luo et al. "Expectation pooling: an effective and interpretable pooling method for predicting DNA–protein binding". In: *Bioinformatics* 36.5 (2020), pp. 1405–1412.

[1336] Jon Lundstrøm et al. "LectinOracle: A Generalizable Deep Learning Model for Lectin–Glycan Binding Prediction". In: *Advanced Science* 9.1 (2022), p. 2103807.

[1337] Elisa Fernandez-Castillo et al. "Deep splicer: A CNN model for splice site prediction in genetic sequences". In: *Genes* 13.5 (2022), p. 907.

[1338] Xiaoyong Pan et al. "Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks". In: *BMC genomics* 19.1 (2018), pp. 1–11.

[1339] Xiuquan Du et al. "DeepMVF-RBP: Deep multi-view fusion representation learning for RNA-binding proteins prediction". In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 65–68.

[1340] Lei Deng et al. "Deep neural networks for inferring binding sites of RNA-binding proteins by using distributed representations of RNA primary sequence and secondary structure". In: *BMC genomics* 21.13 (2020), pp. 1–10.

[1341] Upendra K Pradhan et al. "DBPMod: a supervised learning model for computational recognition of DNA-binding proteins in model organisms". In: *Briefings in Functional Genomics* (2023), elad039.

[1342] Upendra K Pradhan et al. "RBPLight: a computational tool for discovery of plant-specific RNA-binding proteins using light gradient boosting machine and ensemble of evolutionary features". In: *Briefings in Functional Genomics* (2023), elad016.

[1343] Jialong Chen and Lei Deng. "Deeparc: an attention-based hybrid model for predicting transcription factor binding sites from positional embedded dna sequence". In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2020, pp. 180–185.

[1344] Junru Jin et al. "iDNA-ABF: multi-scale deep biological language learning model for the interpretable prediction of DNA methylations". In: *Genome biology* 23.1 (2022), pp. 1–23.

[1345] Fei Tan et al. "Elucidation of DNA methylation on N 6-adenine with deep learning". In: *Nature Machine Intelligence* 2.8 (2020), pp. 466–475.

[1346] Wenxing Hu, Lixin Guan, and Mengshan Li. "Prediction of DNA Methylation based on Multi-dimensional feature encoding and double convolutional fully connected convolutional neural network". In: *PLOS Computational Biology* 19.8 (2023), e1011370.

[1347] Hyuk-Jung Kwon et al. "Advances in methylation analysis of liquid biopsy in early cancer detection of colorectal and lung cancer". In: *Scientific Reports* 13.1 (2023), p. 13502.

[1348] Laiyi Fu, Qinke Peng, and Ling Chai. "Predicting dna methylation states with hybrid information based deep-learning model". In: *IEEE/ACM transactions on computational biology and bioinformatics* 17.5 (2019), pp. 1721–1728.

[1349] Christof Angermueller et al. "DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning". In: *Genome biology* 18.1 (2017), pp. 1–13.

[1350] Linlin Zhuo et al. "StableDNAm: towards a stable and efficient model for predicting DNA methylation based on adaptive feature correction learning". In: *BMC genomics* 24.1 (2023), p. 742.

[1351] Amidou N'Diaye et al. "Machine learning analyses of methylation profiles uncovers tissue-specific gene expression patterns in wheat". In: *The Plant Genome* 13.2 (2020), e20027.

[1352] Zhenxing Wang and Yadong Wang. "Exploring dna methylation data of lung cancer samples with variational autoencoders". In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 1286–1289.

[1353] Wei Ning et al. "Accurate prediction of pan-cancer types using machine learning with minimal number of DNA methylation sites". In: *Journal of Molecular Cell Biology* (2023), mjad023.

[1354] Ximei Luo et al. "Recall DNA methylation levels at low coverage sites using a CNN model in WGBS". In: *PLOS Computational Biology* 19.6 (2023), e1011205.

[1355] Jingxian Zhou et al. "4acCPred: Weakly supervised prediction of N4-acetyldeoxycytosine DNA modification from sequences". In: *Molecular Therapy-Nucleic Acids* 30 (2022), pp. 337–345.

[1356] Zutan Li, Bingbing Jin, and Jingya Fang. "MetaAc4C: A multi-module deep learning framework for accurate prediction of N4-acetylcytidine sites based on pre-trained bidirectional encoder representation and generative adversarial networks". In: *Genomics* 116.1 (2024), p. 110749.

[1357] Jianhua Jia, Xiaojing Cao, and Zhangying Wei. "DLC-ac4C: A Prediction Model for N4-acetylcytidine Sites in Human mRNA Based on DenseNet and Bidirectional LSTM Methods". In: *Current Genomics* 24.3 (2023), pp. 171–186.

[1358] Zhen Shen et al. "Prediction of Back-splicing sites for CircRNA formation based on convolutional neural networks". In: *BMC genomics* 23.1 (2022), p. 581.

[1359] Qianqian Yang et al. "EpiCas-DL: Predicting sgRNA activity for CRISPR-mediated epigenome editing by deep learning". In: *Computational and Structural Biotechnology Journal* 21 (2023), pp. 202–211.

[1360] Ye Han et al. "SiRNA silencing efficacy prediction based on a deep architecture". In: *BMC genomics* 19 (2018), pp. 59–65.

[1361] Ye Han et al. "Effective small interfering RNA design based on convolutional neural network". In: *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2017, pp. 16–21.

[1362] Jael Sanyanda Wekesa, Jun Meng, and Yushi Luan. "Multi-feature fusion for deep learning to predict plant lncRNA-protein interaction". In: *Genomics* 112.5 (2020), pp. 2928–2936.

[1363] Chengshuai Zhao et al. "Graph embedding ensemble methods based on the heterogeneous network for lncRNA-miRNA interaction prediction". In: *BMC genomics* 21 (2020), pp. 1–12.

[1364] Mira Barshai et al. "G4mismatch: Deep neural networks to predict G-quadruplex propensity based on G4-seq data". In: *PLOS Computational Biology* 19.3 (2023), e1010948.

[1365] Dong Chen, Wenjie Shu, and Shaoliang Peng. "Predicting crispr-cas9 off-target with self-supervised neural networks". In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2020, pp. 245–250.

[1366] Jeffrey K Mak, Florian Störtz, and Peter Minary. "Comprehensive computational analysis of epigenetic descriptors affecting CRISPR-Cas9 off-target activity". In: *BMC genomics* 23.1 (2022), p. 805.

[1367] Vuong Le et al. "Deep in the bowel: highly interpretable neural encoder-decoder networks predict gut metabolites from gut microbiome". In: *BMC genomics* 21.4 (2020), pp. 1–15.

[1368] Shang Gao, Jalees Rehman, and Yang Dai. "Assessing comparative importance of DNA sequence and epigenetic modifications on gene expression using a deep convolutional neural network". In: *Computational and Structural Biotechnology Journal* 20 (2022), pp. 3814–3823.

[1369] Jing Li et al. "Genomic representation predicts an asymptotic host adaptation of bat coronaviruses using deep learning". In: *Frontiers in Microbiology* 14 (2023), p. 1157608.

[1370] Lucas Paulo de Lima Camillo, Louis R Lapierre, and Ritambhara Singh. "A pan-tissue DNA-methylation epigenetic clock based on deep learning". In: *npj Aging* 8.1 (2022), p. 4.

[1371] Meng Lan, Shixiong Zhang, and Lin Gao. "Efficient Generation of Paired Single-Cell Multiomics Profiles by Deep Learning". In: *Advanced Science* (2023), p. 2301169.

[1372] Jeffrey R Adrion, Jared G Galloway, and Andrew D Kern. "Predicting the landscape of recombination using deep learning". In: *Molecular biology and evolution* 37.6 (2020), pp. 1790–1808.

[1373] Fatima Khan et al. "Prediction of recombination spots using novel hybrid feature extraction method via deep learning approach". In: *Frontiers in Genetics* 11 (2020), p. 539227.

[1374] Zaheer Ullah Khan et al. "iRSpot-SPI: Deep learning-based recombination spots prediction by incorporating secondary sequence information coupled with physio-chemical properties via Chou's 5-step rule and pseudo components". In: *Chemometrics and Intelligent Laboratory Systems* 189 (2019), pp. 169–180.

[1375] Yu Li et al. "Deep learning identifies and quantifies recombination hotspot determinants". In: *Bioinformatics* 38.10 (2022), pp. 2683–2691.

[1376] Abhigyan Nath and S Karthikeyan. "Enhanced prediction of recombination hotspots using input features extracted by class specific autoencoders". In: *Journal of Theoretical Biology* 444 (2018), pp. 73–82.

[1377] Meijun Gao and Kevin J Liu. "Statistical analysis of GC-biased gene conversion and recombination hotspots in eukaryotic genomes: a phylogenetic hidden Markov model-based approach". In: *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. 2021, pp. 1–24.

[1378] John C Wooley, Adam Godzik, and Iddo Friedberg. "A primer on metagenomics". In: *PLoS computational biology* 6.2 (2010), e1000667.

[1379] Jinny X Zhang et al. "A deep learning model for predicting next-generation sequencing depth from DNA sequence". In: *Nature communications* 12.1 (2021), p. 4387.

[1380] Atif Khurshid Wani et al. "Discovering untapped microbial communities through metagenomics for microplastic remediation: Recent advances, challenges, and way forward". In: *Environmental Science and Pollution Research* (2023), pp. 1–24.

[1381] Lily Tong et al. "Discovery of RNA and DNA viruses using next-generation sequencing: Metagenomics". In: *CVR Genomics* (2023).

[1382] Zhao-xiang Lu et al. "Advances in the Discovery of Novel Antibiotic-resistant Genes Based on Functional Metagenomics". In: *Biotechnology Bulletin* 38.9 (2022), p. 17.

[1383] Yuntao Lu, Qi Li, and Tao Li. "PPA-GCN: A Efficient GCN Framework for Prokaryotic Pathways Assignment". In: *Frontiers in Genetics* 13 (2022), p. 839453.

[1384] Alban Mathieu et al. "Machine Learning and Deep Learning Applications in Metagenomic Taxonomy and Functional Annotation". In: *Frontiers in Microbiology* 13 (2022), p. 811495.

[1385] Laura Judith Marcos-Zambrano et al. "A toolbox of machine learning software to support microbiome analysis". In: *Frontiers in Microbiology* 14 (2023).

[1386] Pu Liu et al. "Towards Strain-Level Complexity: Sequencing Depth Required for Comprehensive Single-Nucleotide Polymorphism Analysis of the Human Gut Microbiome". In: *Frontiers in Microbiology* 13 (2022), p. 828254.

[1387] Yuguo Zha, Hui Chong, and Kang Ning. "Microbiome sample comparison and search: from pair-wise calculations to model-based matching". In: *Frontiers in Microbiology* 12 (2021), p. 642439.

[1388] Alexander Wichmann et al. "MetaTransformer: deep metagenomic sequencing read classification using self-attention models". In: *NAR Genomics and Bioinformatics* 5.3 (2023), lqad082.

[1389] Lindsay M Hopson et al. "Bioinformatics and machine learning in gastrointestinal microbiome research and clinical application". In: *Progress in Molecular Biology and Translational Science* 176 (2020), pp. 141–178.

[1390] Qiang Zhu et al. "An ensemble feature selection method based on deep forest for microbiome-wide association studies". In: *2018 IEEE international conference on bioinformatics and biomedicine (BIBM)*. IEEE. 2018, pp. 248–253.

[1391] Rajnish Kumar et al. "Unlocking the microbial studies through computational approaches: how far have we reached?" In: *Environmental Science and Pollution Research* 30.17 (2023), pp. 48929–48947.

[1392] James MWR McElhinney et al. "Interfacing Machine Learning and Microbial Omics: A Promising Means to Address Environmental Challenges". In: *Frontiers in Microbiology* 13 (2022), p. 851450.

[1393] Huynh Quang Bao, Tran Van Hoai, et al. "A deep embedded clustering algorithm for the binning of metagenomic sequences". In: *IEEE Access* 10 (2022), pp. 54348–54357.

[1394] Sharanbasappa D Madival et al. "A Deep Clustering-based Novel Approach for Binning of Metagenomics Data". In: *Current Genomics* 23.5 (2022), p. 353.

[1395] Timmy Manning et al. "Phylogeny-aware deep 1-dimensional convolutional neural network for the classification of metagenomes". In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 1826–1831.

[1396] Robin L Miller-Coleman et al. "Korarchaeota diversity, biogeography, and abundance in Yellowstone and Great Basin hot springs and ecological niche modeling based on machine learning". In: *PLoS One* 7.5 (2012), e35964.

[1397] Gregory Ditzler, Robi Polikar, and Gail Rosen. "Multi-layer and recursive neural networks for metagenomic classification". In: *IEEE transactions on nanobioscience* 14.6 (2015), pp. 608–616.

[1398] Ying-Feng Hsu et al. “Deep Learning Approach for Pathogen Detection Through Shotgun Metagenomics Sequence Classification”. In: *Artificial Intelligence in Medicine: 17th Conference on Artificial Intelligence in Medicine, AIME 2019, Poznan, Poland, June 26–29, 2019, Proceedings 17*. Springer. 2019, pp. 24–30.

[1399] Yassin Mreyoud and Tae-Hyuk Ahn. “Deep Neural Network Modeling for Phenotypic Prediction of Metagenomic Samples”. In: *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. 2020, pp. 1–1.

[1400] Micha Karlicki, Stanisaw Antonowicz, and Anna Karnkowska. “Tiara: deep learning-based classification system for eukaryotic sequences”. In: *Bioinformatics* 38.2 (2022), pp. 344–350.

[1401] Jie Ren et al. “Identifying viruses from metagenomic data using deep learning”. In: *Quantitative Biology* 8 (2020), pp. 64–77.

[1402] Grigorii Sukhorukov et al. “VirHunter: A deep learning-based method for detection of novel RNA viruses in plant sequencing data”. In: *Frontiers in Bioinformatics* 2 (2022), p. 867111.

[1403] Yan Miao et al. “DETIRE: a hybrid deep learning model for identifying viral sequences from metagenomes”. In: *Frontiers in Microbiology* 14 (2023), p. 1169791.

[1404] Colin Price and Joseph A Russell. “AMAnD: an automated metagenome anomaly detection methodology utilizing DeepSVDD neural networks”. In: *Frontiers in Public Health* 11 (2023).

[1405] Fu Liu et al. “RNN-VirSeeker: a deep learning method for identification of short viral sequences from metagenomes”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19.3 (2020), pp. 1840–1849.

[1406] Yunmeng Chu et al. “DeephageTP: a convolutional neural network framework for identifying phage-specific proteins from metagenomic sequencing data”. In: *PeerJ* 10 (2022), e13404.

[1407] Tor Einar Møller et al. “Mapping microbial abundance and prevalence to changing oxygen concentration in deep-sea sediments using machine learning and differential abundance”. In: *Frontiers in Microbiology* 13 (2022), p. 804575.

[1408] Mary Maranga et al. “Comprehensive functional annotation of metagenomes and microbial genomes using a deep learning-based method”. In: *Msystems* 8.2 (2023), e01178–22.

[1409] Rundong Tan et al. “Prediction of minimal inhibitory concentration of meropenem against klebsiella pneumoniae using metagenomic data”. In: *Frontiers in Microbiology* 12 (2021), p. 712886.

[1410] Zhencheng Fang et al. “PPR-Meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning”. In: *GigaScience* 8.6 (2019), giz066.

[1411] Theodoros Koutsandreas et al. "ANASTASIA: an automated metagenomic analysis pipeline for novel enzyme discovery exploiting next generation sequencing data". In: *Frontiers in Genetics* (2019), p. 469.

[1412] Amani Al-Ajlan and Achraf El Allali. "CNN-MGP: convolutional neural networks for metagenomics gene prediction". In: *Interdisciplinary Sciences: Computational Life Sciences* 11 (2019), pp. 628–635.

[1413] Yan Wang et al. "Fueling ab initio folding with marine metagenomics enables structure and function predictions of new protein families". In: *Genome biology* 20.1 (2019), pp. 1–14.

[1414] Muhammad Umar Nasir et al. "Advance genome disorder prediction model empowered with deep learning". In: *IEEE Access* 10 (2022), pp. 70317–70328.

[1415] Meryem Altn Karagöz and O Ufuk Nalbantoglu. "Taxonomic classification of metagenomic sequences from Relative Abundance Index profiles using deep learning". In: *Biomedical Signal Processing and Control* 67 (2021), p. 102539.

[1416] Yuguo Zha et al. "Ontology-aware deep learning enables ultrafast and interpretable source tracking among sub-million microbial community samples from hundreds of niches". In: *Genome medicine* 14.1 (2022), p. 43.

[1417] Yueyu Jiang et al. "DEPP: deep learning enables extending species trees using single genes". In: *Systematic Biology* 72.1 (2023), pp. 17–34.

[1418] Shiva Abdollahi Aghdam and Amanda May Vivian Brown. "Deep learning approaches for natural product discovery from plant endophytic microbiomes". In: *Environmental microbiome* 16.1 (2021), pp. 1–20.

[1419] Venkata Suhas Maringanti, Vanni Bucci, and Georg K Gerber. "MDITRE: scalable and interpretable machine learning for predicting host status from temporal microbiome dynamics". In: *Msystems* 7.5 (2022), e00132–22.

[1420] Yang Shen et al. "EnsDeepDP: An Ensemble Deep Learning Approach for Disease Prediction Through Metagenomics". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 20.2 (2022), pp. 986–998.

[1421] Yassin Mreyoud et al. "MegaD: deep learning for rapid and accurate disease status prediction of metagenomic samples". In: *Life* 12.5 (2022), p. 669.

[1422] Xingjian Chen et al. "Human disease prediction from microbiome data by multiple feature fusion and deep learning". In: *Iscience* 25.4 (2022).

[1423] Mohammad Arifur Rahman and Huzefa Rangwala. "IDMIL: an alignment-free Interpretable Deep Multiple Instance Learning (MIL) for predicting disease from whole-metagenomic data". In: *Bioinformatics* 36.Supplement_1 (2020), pp. i39–i47.

[1424] Jinglie Zhou et al. "Diversity of virophages in metagenomic data sets". In: *Journal of virology* 87.8 (2013), pp. 4225–4236.

[1425] Kalimuthu Jawaharraj et al. "Transcriptome-wide marker gene expression analysis of stress-responsive sulfate-reducing bacteria". In: *Scientific Reports* 13.1 (2023), p. 16181.

[1426] Wenhuan Zeng, Anupam Gautam, and Daniel H Huson. "DeepToA: an ensemble deep-learning approach to predicting the theater of activity of a microbiome". In: *Bioinformatics* 38.20 (2022), pp. 4670–4676.

[1427] Zhiyu Deng et al. "Application of deep learning in plant–microbiota association analysis". In: *Frontiers in Genetics* 12 (2021), p. 697090.

[1428] Jiaquan Huang et al. "A deep learning model for type II polyketide natural product prediction without sequence alignment". In: *Digital Discovery* 2.5 (2023), pp. 1484–1493.

[1429] Joachim Johansen et al. "Genome binning of viral entities from bulk metagenomics data". In: *Nature Communications* 13.1 (2022), p. 965.

[1430] Cédric G Arisdakessian et al. "CoCoNet: an efficient deep learning tool for viral metagenome binning". In: *Bioinformatics* 37.18 (2021), pp. 2803–2810.

[1431] Naveen Duhan, Jeanette M Norton, and Rakesh Kaundal. "deepNEC: a novel alignment-free tool for the identification and classification of nitrogen biochemical network-related enzymes using deep learning". In: *Briefings in Bioinformatics* 23.3 (2022), bbac071.

[1432] Yun Wu et al. "Transcriptomics based network analyses and molecular docking highlighted potentially therapeutic biomarkers for colon cancer". In: *Biochemical Genetics* (2023), pp. 1–19.

[1433] Jimmy Bell and David A Hendrix. "Predicting Drosha and Dicer Cleavage Sites with DeepMirCut". In: *Frontiers in Molecular Biosciences* 8 (2022), p. 799056.

[1434] Veronika Petrova et al. "Increased chromatin accessibility facilitates intron retention in specific cell differentiation states". In: *Nucleic Acids Research* 50.20 (2022), pp. 11563–11579.

[1435] Yi Li, Daniel Quang, and Xiaohui Xie. "Understanding sequence conservation with deep learning". In: *Proceedings of the 8th ACM international conference on bioinformatics, computational biology, and health informatics*. 2017, pp. 400–406.

[1436] Wanshan Ning et al. "HybridSucc: a hybrid-learning architecture for general and species-specific succinylation site prediction". In: *Genomics, proteomics & bioinformatics* 18.2 (2020), pp. 194–207.

[1437] Anuradha Wickramarachchi and Yu Lin. "Binning long reads in metagenomics datasets using composition and coverage information". In: *Algorithms for Molecular Biology* 17.1 (2022), p. 14.

[1438] Xudong Yu et al. "Single-cell omics traces the heterogeneity of prostate cancer cells and the tumor microenvironment". In: *Cellular & Molecular Biology Letters* 28.1 (2023), pp. 1–16.

[1439] Pedro Madrigal et al. “Machine learning algorithm to characterize antimicrobial resistance associated with the International Space Station surface microbiome”. In: *Microbiome* 10.1 (2022), p. 134.

[1440] Gustavo Arango-Argoty et al. “DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data”. In: *Microbiome* 6 (2018), pp. 1–15.

[1441] Seung Jae Lee and Mina Rho. “Multimodal deep learning applied to classify healthy and disease states of human microbiome”. In: *Scientific Reports* 12.1 (2022), p. 824.

[1442] Alberto Catanese et al. “Multiomics and machine-learning identify novel transcriptional and mutational signatures in amyotrophic lateral sclerosis”. In: *Brain* (2023), awad075.

[1443] Parvathaneni Naga Srinivasu et al. “Using recurrent neural networks for predicting type-2 diabetes from genomic and tabular data”. In: *Diagnostics* 12.12 (2022), p. 3067.

[1444] Gangcai Xie et al. “Group lasso regularized deep learning for cancer prognosis from multi-omics and clinical features”. In: *Genes* 10.3 (2019), p. 240.

[1445] Zhi Huang et al. “Deep learning-based cancer survival prognosis from RNA-seq data: approaches and evaluations”. In: *BMC medical genomics* 13 (2020), pp. 1–12.

[1446] Kumardeep Chaudhary et al. “Deep learning–based multi-omics integration robustly predicts survival in liver cancer”. In: *Clinical Cancer Research* 24.6 (2018), pp. 1248–1259.

[1447] Duc-Hau Le. “Machine learning-based approaches for disease gene prediction”. In: *Briefings in functional genomics* 19.5-6 (2020), pp. 350–363.

[1448] Qingyan Yin et al. “A convolutional neural network model for survival prediction based on prognosis-related cascaded Wx feature selection”. In: *Laboratory Investigation* 102.10 (2022), pp. 1064–1074.

[1449] Hua Chai et al. “Integrating multi-omics data through deep learning for accurate cancer prognosis prediction”. In: *Computers in biology and medicine* 134 (2021), p. 104481.

[1450] Jie-Yi Shi et al. “Exploring prognostic indicators in the pathological images of hepatocellular carcinoma based on deep learning”. In: *Gut* 70.5 (2021), pp. 951–961.

[1451] Chihyun Park et al. “Improved prediction of cancer outcome using graph-embedded generative adversarial networks”. In: *IEEE Access* 9 (2021), pp. 20076–20088.

[1452] Abdur Rahaman et al. “Deep multimodal predictome for studying mental disorders”. In: *Human Brain Mapping* 44.2 (2023), pp. 509–522.

[1453] Cheng-Hong Yang et al. "An Information Fusion System-Driven Deep Neural Networks With Application to Cancer Mortality Risk Estimate". In: *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[1454] Yichuan Liu et al. "Identification of risk variants related to malignant tumors in children with birth defects by whole genome sequencing". In: *Biomarker Research* 10.1 (2022), pp. 1–9.

[1455] Nadia G Elseddeq et al. "A selected deep learning cancer prediction framework". In: *IEEE Access* 9 (2021), pp. 151476–151492.

[1456] Olga Mineeva et al. "ResMiCo: Increasing the quality of metagenome-assembled genomes with deep learning". In: *PLOS Computational Biology* 19.5 (2023), e1011001.

[1457] Kalana Wijegunarathna et al. "MetaG: a comprehensive visualization tool to explore metagenomes". In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2020, pp. 460–463.

[1458] Kevin B Dsouza et al. "Latent representation of the human pan-celltype epigenome through a deep recurrent neural network". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19.4 (2021), pp. 2313–2323.

[1459] Uditha Maduranga et al. "Dimensionality reduction for cluster identification in metagenomics using autoencoders". In: *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*. IEEE. 2020, pp. 113–118.

[1460] Liang Chen et al. "Deep soft K-means clustering with self-training for single-cell RNA sequence data". In: *NAR genomics and bioinformatics* 2.2 (2020), lqaa039.

[1461] Kristiina Ausmees and Carl Nettelblad. "A deep learning framework for characterization of genotype data". In: *G3* 12.3 (2022), jkac020.

[1462] Jiarui Ding, Anne Condon, and Sohrab P Shah. "Interpretable dimensionality reduction of single cell transcriptome data with deep generative models". In: *Nature communications* 9.1 (2018), p. 2002.

[1463] Kaikun Xie et al. "scAIDE: clustering of large-scale single-cell RNA-seq data reveals putative and rare cell types". In: *NAR genomics and bioinformatics* 2.4 (2020), lqaa082.

[1464] Xiaoyu Zhang et al. "Integrated multi-omics analysis using variational autoencoders: application to pan-cancer classification". In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2019, pp. 765–769.

[1465] Yanyu Hu et al. "Classifying the multi-omics data of gastric cancer using a deep feature selection method". In: *Expert Systems with Applications* 200 (2022), p. 116813.

[1466] Abeer Alzubaidi, Jonathan Tepper, and Ahmad Lotfi. "A novel deep mining model for effective knowledge discovery from omics data". In: *Artificial Intelligence in Medicine* 104 (2020), p. 101821.

[1467] Imene Zenbout et al. "Efficient Bioinspired Feature Selection and Machine Learning Based Framework Using Omics Data and Biological Knowledge Data Bases in Cancer Clinical Endpoint Prediction". In: *IEEE Access* 11 (2023), pp. 2674–2699.

[1468] Marcelo AC Fernandes and HT Kung. "A novel training strategy for deep learning model compression applied to viral classifications". In: *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2021, pp. 1–9.

[1469] Mindia Adinda Sari Haryono et al. "Recovery of high quality metagenome-assembled genomes from full-scale activated sludge microbial communities in a tropical climate using longitudinal metagenome sampling". In: *Frontiers in Microbiology* 13 (2022), p. 869135.

[1470] Andrei Dmitrenko, Michelle Reid, and Nicola Zamboni. "Regularized adversarial learning for normalization of multi-batch untargeted metabolomics data". In: *Bioinformatics* 39.3 (2023), btad096.

[1471] Thanh Hai Nguyen. "Metagenome-based disease classification with deep learning and visualizations based on self-organizing maps". In: *Future Data and Security Engineering: 6th International Conference, FDSE 2019, Nha Trang City, Vietnam, November 27–29, 2019, Proceedings 6*. Springer. 2019, pp. 307–319.

[1472] Xuping Xie et al. "Predicting miRNA-disease associations based on PPMI and attention network". In: *BMC bioinformatics* 24.1 (2023), pp. 1–19.

[1473] Qingquan Liao et al. "An efficient model for predicting human diseases through miRNA based on multiple-types of contrastive learning". In: *Frontiers in Microbiology* 14 (2023).

[1474] Xuan Jiang et al. "Human cytomegalovirus infection perturbs neural progenitor cell fate via the expression of viral microRNAs". In: *Journal of Medical Virology* 95.2 (2023), e28574.

[1475] Hua Hu et al. "Adaptive deep propagation graph neural network for predicting miRNA–disease associations". In: *Briefings in Functional Genomics* (2023), elad010.

[1476] Jiashu Li et al. "FCGCNMDA: predicting miRNA-disease associations by applying fully connected graph convolutional networks". In: *Molecular Genetics and Genomics* 295.5 (2020), pp. 1197–1209.

[1477] Ping Luo et al. "Enhancing the prediction of disease–gene associations with multimodal deep learning". In: *Bioinformatics* 35.19 (2019), pp. 3735–3742.

[1478] Nam D Nguyen, Ting Jin, and Daifeng Wang. "Varmole: a biologically drop-connect deep neural network model for prioritizing disease risk variants and genes". In: *Bioinformatics* 37.12 (2021), pp. 1772–1775.

[1479] Junhao Su et al. "RENET2: high-performance full-text gene–disease relation extraction with iterative training data expansion". In: *NAR Genomics and Bioinformatics* 3.3 (2021), lqab062.

[1480] Prashnna K Gyawali et al. "Improving genetic risk prediction across diverse population by disentangling ancestry representations". In: *Communications Biology* 6.1 (2023), p. 964.

[1481] Andrew G Sharo et al. "StrVCTVRE: a supervised learning method to predict the pathogenicity of human genome structural variants". In: *The American Journal of Human Genetics* 109.2 (2022), pp. 195–209.

[1482] Victoria Popic et al. "Cue: a deep-learning framework for structural variant discovery and genotyping". In: *Nature Methods* 20.4 (2023), pp. 559–568.

[1483] Hongjian Qi et al. "MVP predicts the pathogenicity of missense variants by deep learning". In: *Nature communications* 12.1 (2021), p. 510.

[1484] Da Xu et al. "Multi-scale supervised clustering-based feature selection for tumor classification and identification of biomarkers and targets on genomic data". In: *BMC genomics* 21.1 (2020), pp. 1–17.

[1485] Ruchi Chauhan, PK Vinod, and CV Jawahar. "Exploring genetic-histologic relationships in breast cancer". In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2021, pp. 1187–1190.

[1486] Prasoon Joshi, Seokho Jeong, and Taesung Park. "Cancer subtype classification based on superlayered neural network". In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2019, pp. 1988–1992.

[1487] Ana Paula Alarcón-Zendejas et al. "The promising role of new molecular biomarkers in prostate cancer: From coding and non-coding genes to artificial intelligence approaches". In: *Prostate cancer and prostatic diseases* 25.3 (2022), pp. 431–443.

[1488] Max Garzon et al. "A computational approach to biological pathogenicity". In: *Molecular Genetics and Genomics* 297.6 (2022), pp. 1741–1754.

[1489] Kountay Dwivedi et al. "An explainable AI-driven biomarker discovery framework for Non-Small Cell Lung Cancer classification". In: *Computers in Biology and Medicine* 153 (2023), p. 106544.

[1490] Arwinder Dhillon, Ashima Singh, and Vinod Kumar Bhalla. "Biomarker identification and cancer survival prediction using random spatial local best cat swarm and Bayesian optimized DNN". In: *Applied Soft Computing* 146 (2023), p. 110649.

[1491] Ray Bahado-Singh et al. "Precision oncology: Artificial intelligence and DNA methylation analysis of circulating cell-free DNA for lung cancer detection". In: *Frontiers in oncology* 12 (2022), p. 790645.

[1492] Vivek Bhakta Mathema et al. "Deep learning facilitates multi-data type analysis and predictive biomarker discovery in cancer precision medicine". In: *Computational and Structural Biotechnology Journal* (2023).

[1493] Bastian Pfeifer, Anna Saranti, and Andreas Holzinger. "GNN-SubNet: disease subnetwork detection with explainable graph neural networks". In: *Bioinformatics* 38.Supplement_2 (2022), pp. ii120–ii126.

[1494] Yulin Dai et al. "Association of CXCR6 with COVID-19 severity: delineating the host genetic factors in transcriptomic regulation". In: *Human genetics* 140.9 (2021), pp. 1313–1328.

[1495] Xiaolong Zhang et al. "Robust prognostic subtyping of muscle-invasive bladder cancer revealed by deep learning-based multi-omics data integration". In: *Frontiers in Oncology* 11 (2021), p. 689626.

[1496] Jili Hu et al. "A Deep Neural Network for Gastric Cancer Prognosis Prediction Based on Biological Information Pathways". In: *Journal of Oncology* 2022 (2022).

[1497] Jiening Zhu et al. "Geometric graph neural networks on multi-omics data to predict cancer survival outcomes". In: *Computers in Biology and Medicine* (2023), p. 107117.

[1498] Hamid Reza Hassanzadeh, John H Phan, and May D Wang. "A multi-modal graph-based semi-supervised pipeline for predicting cancer survival". In: *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2016, pp. 184–189.

[1499] Akshat Gupta et al. "Advancement in Deep Learning Methods for Diagnosis and Prognosis of Cervical Cancer". In: *Current Genomics* 23.4 (2022), p. 234.

[1500] Leili Tapak et al. "Identification of gene profiles related to the development of oral cancer using a deep learning technique". In: *BMC Medical Genomics* 16.1 (2023), p. 35.

[1501] Lu Qiu et al. "Hierarchical multimodal fusion framework based on noisy label learning and attention mechanism for cancer classification with pathology and genomic features". In: *Computerized Medical Imaging and Graphics* 104 (2023), p. 102176.

[1502] Arjun Magge et al. "GeoBoost2: a natural languageprocessing pipeline for GenBank metadata enrichment for virus phylogeography". In: *Bioinformatics* 36.20 (2020), pp. 5120–5121.

[1503] Pegah Mavaie, Lawrence Holder, and Michael Skinner. "Identifying unique exposure-specific transgenerational differentially DNA methylated region epimutations in the genome using hybrid deep learning prediction models". In: *Environmental Epigenetics* 9.1 (2023), dvad007.

[1504] Jiarun Cao et al. "A Novel Automated Approach to Mutation-Cancer Relation Extraction by Incorporating Heterogeneous Knowledge". In: *IEEE Journal of Biomedical and Health Informatics* 27.2 (2022), pp. 1096–1105.

[1505] Carlos A Garcia-Prieto et al. "Detection of oncogenic and clinically actionable mutations in cancer genomes critically depends on variant calling tools". In: *Bioinformatics* 38.12 (2022), pp. 3181–3191.

[1506] Kanggeun Lee et al. “CPEM: Accurate cancer type classification based on somatic alterations using an ensemble of a random forest and a deep neural network”. In: *Scientific reports* 9.1 (2019), p. 16927.

[1507] Suixue Wang, Shuling Wang, and Zhengxia Wang. “A survey on multi-omics-based cancer diagnosis using machine learning with the potential application in gastrointestinal cancer”. In: *Frontiers in Medicine* 9 (2023), p. 1109365.

[1508] Kseniia Cheloshkina and Maria Poptsova. “Comprehensive analysis of cancer breakpoints reveals signatures of genetic and epigenetic contribution to cancer genome rearrangements”. In: *PLOS Computational Biology* 17.3 (2021), e1008749.

[1509] A Sexton Oates et al. “MA01. 07 From Molecular to Histological Characterization of Lung Carcinoids via Computer Vision and Spatial Genomics”. In: *Journal of Thoracic Oncology* 17.9 (2022), S44–S45.

[1510] Jie Xia et al. “RDAClone: deciphering tumor heterozygosity through single-cell genomics data analysis with robust deep autoencoder”. In: *Genes* 12.12 (2021), p. 1847.

[1511] Yasir Ali et al. “IDriveGenes: Cancer Driver Genes Prediction Using Machine Learning”. In: *IEEE Access* 11 (2023), pp. 28439–28453.

[1512] Nitay Itzhacky and Roded Sharan. “Prediction of cancer dependencies from expression data using deep learning”. In: *Molecular Omics* 17.1 (2021), pp. 66–71.

[1513] Jianing Xi et al. “An omics-to-omics joint knowledge association subtensor model for radiogenomics cross-modal modules from genomics and ultrasonic images of breast cancers”. In: *Computers in Biology and Medicine* 155 (2023), p. 106672.

[1514] Bryan He et al. “Integrating spatial gene expression and breast tumour morphology via deep learning”. In: *Nature biomedical engineering* 4.8 (2020), pp. 827–834.

[1515] Yingshuai Sun et al. “Identification of 12 cancer types through genome deep learning”. In: *Scientific reports* 9.1 (2019), p. 17256.

[1516] Shahzad Ahmad Qureshi et al. “Radiogenomic classification for MGMT promoter methylation status using multi-omics fused feature space for least invasive diagnosis through mpMRI scans”. In: *Scientific reports* 13.1 (2023), p. 3291.

[1517] Jan-Michaél Hirsch et al. “A paradigm shift in the prevention and diagnosis of oral squamous cell carcinoma”. In: *Journal of Oral Pathology & Medicine* 52.9 (2023), pp. 826–833.

[1518] Zhen Zhao et al. “Deep learning-based model for predicting progression in patients with head and neck squamous cell carcinoma”. In: *Cancer Biomarkers* 27.1 (2020), pp. 19–28.

[1519] Jiapan Gu et al. "Multi-phase cross-modal learning for noninvasive gene mutation prediction in hepatocellular carcinoma". In: *2020 42nd annual international conference of the IEEE engineering in medicine & biology society (EMBC)*. IEEE. 2020, pp. 5814–5817.

[1520] Chanwoo Park, Boram Kim, and Taesung Park. "DeepHisCoM: deep learning pathway analysis using hierarchical structural component models". In: *Briefings in Bioinformatics* 23.5 (2022), bbac171.

[1521] Hua Chai et al. "An adaptive transfer-learning-based deep Cox neural network for hepatocellular carcinoma prognosis prediction". In: *Frontiers in oncology* 11 (2021), p. 692774.

[1522] Jun Yu et al. "A model for predicting prognosis in patients with esophageal squamous cell carcinoma based on joint representation learning". In: *Oncology Letters* 20.6 (2020), pp. 1–1.

[1523] M DOrazio et al. "Machine learning phenomics (MLP) combining deep learning with time-lapse-microscopy for monitoring colorectal adenocarcinoma cells gene expression and drug-response". In: *Scientific Reports* 12.1 (2022), p. 8545.

[1524] Xin Lai et al. "A disease network-based deep learning approach for characterizing melanoma". In: *International Journal of Cancer* 150.6 (2022), pp. 1029–1044.

[1525] Jingting Zhao et al. "A multi-omics deep learning model for hypoxia phenotype to predict tumor aggressiveness and prognosis in uveal melanoma for rationalized hypoxia-targeted therapy". In: *Computational and Structural Biotechnology Journal* 20 (2022), pp. 3182–3194.

[1526] Mohaiminul Islam et al. "An integrative deep learning framework for classifying molecular subtypes of breast cancer". In: *Computational and structural biotechnology journal* 18 (2020), pp. 2185–2199.

[1527] Alaa El-Nabawy, Nahla A Belal, and Nashwa El-Bendary. "A cascade deep forest model for breast cancer subtype classification using multi-omics data". In: *Mathematics* 9.13 (2021), p. 1574.

[1528] Nermin Abdelhakim Othman, Manal A Abdel-Fattah, and Ahlam Talaat Ali. "A Hybrid Deep Learning Framework with Decision-Level Fusion for Breast Cancer Survival Prediction". In: *Big Data and Cognitive Computing* 7.1 (2023), p. 50.

[1529] Chen Peng, Yang Zheng, and De-Shuang Huang. "Capsule network based modeling of multi-omics data for discovery of breast cancer-related genes". In: *IEEE/ACM transactions on computational biology and bioinformatics* 17.5 (2019), pp. 1605–1612.

[1530] Cristiano Guttà, Christoph Morhard, and Markus Rehm. "Applying a GAN-based classifier to improve transcriptome-based prognostication in breast cancer". In: *PLOS Computational Biology* 19.4 (2023), e1011035.

[1531] Lu Ye et al. "An ovarian cancer susceptible gene prediction method based on deep learning methods". In: *Frontiers in Cell and Developmental Biology* 9 (2021), p. 730475.

[1532] Zhe Zhang et al. "Assessing the clinical utility of multi-omics data for predicting serous ovarian cancer prognosis". In: *Journal of Obstetrics and Gynaecology* 43.1 (2023), p. 2171778.

[1533] Ke Feng Sun et al. "XGBG: A Novel Method for Identifying Ovarian Carcinoma Susceptible Genes Based on Deep Learning". In: *Frontiers in Oncology* 12 (2022), p. 897503.

[1534] Min-Koo Park et al. "Deep-Learning Algorithm and Concomitant Biomarker Identification for NSCLC Prediction Using Multi-Omics Data Integration". In: *Biomolecules* 12.12 (2022), p. 1839.

[1535] Kimberly Skead et al. "Interacting evolutionary pressures drive mutation dynamics and health outcomes in aging blood". In: *Nature communications* 12.1 (2021), p. 4921.

[1536] Kountay Dwivedi et al. "Enlightening the path to NSCLC biomarkers: Utilizing the power of XAI-guided deep learning". In: *Computer Methods and Programs in Biomedicine* 243 (2024), p. 107864.

[1537] Vivek Ruhela et al. "BDL-SP: A Bio-inspired DL model for the identification of altered Signaling Pathways in Multiple Myeloma using WES data". In: *American Journal of Cancer Research* 13.4 (2023), p. 1155.

[1538] Francisco Azuaje et al. "Connecting histopathology imaging and proteomics in kidney cancer through machine learning". In: *Journal of clinical medicine* 8.10 (2019), p. 1535.

[1539] Yanan Wang et al. "Genomics and pathology based deep learning to predict cancers of unknown primary". In: *Cancer Research* 82.12_Supplement (2022), pp. 5045–5045.

[1540] Yue Zhao et al. "CUP-AI-Dx: A tool for inferring cancer tissue of origin and molecular subtype using RNA gene-expression data and artificial intelligence". In: *EBioMedicine* 61 (2020).

[1541] Tejaswini Mallavarapu et al. "PASCL: Pathway-based Sparse Deep Clustering for Identifying Unknown Cancer Subtypes". In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 470–475.

[1542] Sayed Mohammad Ebrahim Sahraeian et al. "Achieving robust somatic mutation detection with deep learning models derived from reference data sets of a cancer sample". In: *Genome Biology* 23.1 (2022), p. 12.

[1543] Jing Meng et al. "DeepSSV: detecting somatic small variants in paired tumor and normal sequencing data with convolutional neural network". In: *Briefings in Bioinformatics* 22.4 (2021), bbaa272.

[1544] Cheng Wang and Jingjing Li. "A deep learning framework identifies pathogenic noncoding somatic mutations from personal prostate cancer genomes". In: *Cancer Research* 80.21 (2020), pp. 4644–4654.

[1545] Jordan Anaya et al. "Multiple-instance learning of somatic mutations for the classification of tumour type and the prediction of microsatellite status". In: *Nature biomedical engineering* 8.1 (2024), pp. 57–67.

[1546] Ander Díaz-Navarro et al. "RFcaller: a machine learning approach combined with read-level features to detect somatic mutations". In: *NAR Genomics and Bioinformatics* 5.2 (2023), lqad056.

[1547] Eric Pellegrino et al. "Machine learning random forest for predicting oncosomatic variant NGS analysis". In: *Scientific reports* 11.1 (2021), p. 21820.

[1548] Prima Sanjaya et al. "Mutation-Attention (MuAt): deep representation learning of somatic mutations for tumour typing and subtyping". In: *Genome medicine* 15.1 (2023), p. 47.

[1549] Kiran Krishnamachari et al. "Accurate somatic variant detection using weakly supervised deep learning". In: *Nature Communications* 13.1 (2022), p. 4248.

[1550] Firda Aminy Maruf, Rian Pratama, and Giltae Song. "DNN-Boost: Somatic mutation identification of tumor-only whole-exome sequencing data using deep neural network and XGBoost". In: *Journal of Bioinformatics and Computational Biology* 19.06 (2021), p. 2140017.

[1551] Benjamin J Ainscough et al. "A deep learning approach to automate refinement of somatic variant calling from cancer sequencing data". In: *Nature genetics* 50.12 (2018), pp. 1735–1743.

[1552] Martin Palazzo, Pierre Beauseroy, and Patricio Yankilevich. "A pan-cancer somatic mutation embedding using autoencoders". In: *BMC bioinformatics* 20 (2019), pp. 1–10.

[1553] Yuchen Yuan et al. "Cancer type prediction based on copy number aberration and chromatin 3D structure with convolutional neural networks". In: *BMC genomics* 19 (2018), pp. 1–8.

[1554] Sunkyu Kim et al. "Mut2Vec: distributed representation of cancerous mutations". In: *BMC medical genomics* 11 (2018), pp. 57–69.

[1555] Sina Abdollahi, Peng-Chan Lin, and Jung-Hsien Chiang. "DiaDeL: An accurate deep learning-based model with mutational signatures for predicting metastasis stage and cancer types". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19.3 (2021), pp. 1336–1343.

[1556] Z Wang et al. "7P Predicting lymph node metastasis and molecular subtype from pathology united with genomics in breast cancer with multitask deep learning". In: *Annals of Oncology* 33 (2022), S1433.

[1557] Daria Lazic et al. “Landscape of bone marrow metastasis in human neuroblastoma unraveled by transcriptomics and deep multiplex imaging”. In: *Cancers* 13.17 (2021), p. 4311.

[1558] Dan Liu et al. “Multi-omics immune regulatory mechanisms in lung adenocarcinoma metastasis and survival time”. In: *Computers in Biology and Medicine* 164 (2023), p. 107333.

[1559] Wei Jiao et al. “A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns”. In: *Nature communications* 11.1 (2020), p. 728.

[1560] Weisheng Zheng et al. “Deep learning model accurately classifies metastatic tumors from primary tumors based on mutational signatures”. In: *Scientific Reports* 13.1 (2023), p. 8752.

[1561] Milad Mostavi et al. “CancerSiamese: one-shot learning for predicting primary and metastatic tumor types unseen during model training”. In: *BMC bioinformatics* 22.1 (2021), pp. 1–17.

[1562] Guojie Hou et al. “Deep learning approach for predicting lymph node metastasis in non-small cell lung cancer by fusing image–gene data”. In: *Engineering Applications of Artificial Intelligence* 122 (2023), p. 106140.

[1563] Saeed Pirmoradi et al. “A self-organizing deep neuro-fuzzy system approach for classification of kidney cancer subtypes using miRNA genomics data”. In: *Computer Methods and Programs in Biomedicine* 206 (2021), p. 106132.

[1564] Chun-Yu Lin et al. “Deep learning with evolutionary and genomic profiles for identifying cancer subtypes”. In: *Journal of Bioinformatics and Computational Biology* 17.03 (2019), p. 1940005.

[1565] Senthil Kumar K ME et al. “Machine Learning Algorithms for Breast Cancer Prediction”. In: *Journal of Population Therapeutics and Clinical Pharmacology* 30.7 (2023), pp. 245–250.

[1566] Aswathi Viswanathan et al. “Deep learning-based classifier of diffuse large B-cell lymphoma cell-of-origin with clinical outcome”. In: *Briefings in Functional Genomics* 22.1 (2023), pp. 42–48.

[1567] Richard J Chen et al. “Pan-cancer integrative histology-genomic analysis via multimodal deep learning”. In: *Cancer Cell* 40.8 (2022), pp. 865–878.

[1568] Raktim Kumar Mondol et al. “AFExNet: An adversarial autoencoder for differentiating breast cancer sub-types and extracting biologically relevant genes”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19.4 (2021), pp. 2060–2070.

[1569] Adham Beykikhoshk et al. “DeepTRIAGE: interpretable and individualised biomarker scores using attention mechanism for the classification of breast cancer sub-types”. In: *BMC medical genomics* 13 (2020), pp. 1–10.

[1570] Richard Lupat et al. “Moanna: Multi-Omics Autoencoder-Based Neural Network Algorithm for Predicting Breast Cancer Subtypes”. In: *IEEE Access* 11 (2023), pp. 10912–10924.

[1571] Milad Mostavi et al. "Convolutional neural network models for cancer type prediction based on gene expression". In: *BMC medical genomics* 13 (2020), pp. 1–13.

[1572] Smriti Chawla et al. "Gene expression based inference of cancer drug sensitivity". In: *Nature communications* 13.1 (2022), p. 5680.

[1573] Long-Yi Guo et al. "Deep learning-based ovarian cancer subtypes identification using multi-omics data". In: *BioData Mining* 13.1 (2020), pp. 1–12.

[1574] Qiuwen Sun et al. "SADLN: Self-attention based deep learning network of integrating multi-omics data for cancer subtype recognition". In: *Frontiers in Genetics* 13 (2023), p. 1032768.

[1575] Liu Shuangshuang et al. "A Deep Learning Fusion Clustering framework for breast cancer subtypes identification by integrating multi-omics data". In: *2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*. IEEE. 2020, pp. 1710–1714.

[1576] Bingjun Li, Tianyu Wang, and Sheida Nabavi. "Cancer molecular subtype classification by graph convolutional networks on multi-omics data". In: *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. 2021, pp. 1–9.

[1577] Hai Yang et al. "Subtype-GAN: a deep learning approach for integrative cancer subtyping of multi-omics data". In: *Bioinformatics* 37.16 (2021), pp. 2231–2237.

[1578] Ge Zhang et al. "MultiGATAE: a novel cancer subtype identification method based on multi-omics and attention mechanism". In: *Frontiers in Genetics* 13 (2022), p. 855629.

[1579] Bo Yang et al. "Deep subspace mutual learning for cancer subtypes prediction". In: *Bioinformatics* 37.21 (2021), pp. 3715–3722.

[1580] Yanting Zhang and Hisanori Kiryu. "MODEC: an unsupervised clustering method integrating omics data for identifying cancer subtypes". In: *Briefings in Bioinformatics* 23.6 (2022), bbac372.

[1581] Jasper Zhongyuan Zhang, Wei Xu, and Pingzhao Hu. "Tightly integrated multiomics-based deep tensor survival model for time-to-event prediction". In: *Bioinformatics* 38.12 (2022), pp. 3259–3266.

[1582] Zheng Chen et al. "Learning vector quantized representation for cancer subtypes identification". In: *Computer Methods and Programs in Biomedicine* 236 (2023), p. 107543.

[1583] Jing Zhao et al. "Subtype-DCC: decoupled contrastive clustering method for cancer subtype identification based on multi-omics data". In: *Briefings in Bioinformatics* 24.2 (2023), bbad025.

[1584] Sana Munquad and Asim Bikas Das. "DeepAutoGlioma: a deep learning autoencoder-based multi-omics data integration and classification tools for glioma subtyping". In: *BioData Mining* 16.1 (2023), p. 32.

[1585] Xingze Wang et al. “Lung cancer subtype diagnosis by fusing image-genomics data and hybrid deep networks”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2021).

[1586] Yang Guo, Xuequn Shang, and Zhanhuai Li. “Identification of cancer subtypes by integrating multiple types of transcriptomics data with deep learning in breast cancer”. In: *Neurocomputing* 324 (2019), pp. 20–30.

[1587] Qingchun Liu and Kai Song. “ProgCAE: a deep learning-based method that integrates multi-omics data to predict cancer subtypes”. In: *Briefings in Bioinformatics* (2023), bbad196.

[1588] Eleftherios Trivizakis et al. “Radiotranscriptomics of non-small cell lung carcinoma for assessing high-level clinical outcomes using a machine learning-derived multi-modal signature”. In: *BioMedical Engineering OnLine* 22.1 (2023), p. 125.

[1589] Thi-Oanh Tran, Thanh Hoa Vo, and Nguyen Quoc Khanh Le. “Omics-based deep learning approaches for lung cancer decision-making and therapeutics development”. In: *Briefings in Functional Genomics* (2023), elad031.

[1590] Jiarui Feng, Heming Zhang, and Fuhai Li. “Investigating the relevance of major signaling pathways in cancer survival using a biologically meaningful deep learning model”. In: *BMC bioinformatics* 22.1 (2021), pp. 1–13.

[1591] Luís A Vale-Silva and Karl Rohr. “Long-term cancer survival prediction using multimodal deep learning”. In: *Scientific Reports* 11.1 (2021), p. 13505.

[1592] Jing Tian et al. “Deep learning algorithm reveals two prognostic subtypes in patients with gliomas”. In: *BMC bioinformatics* 23.1 (2022), p. 417.

[1593] Diksha Sharma et al. “A deep learning-based integrative model for survival time prediction of head and neck squamous cell carcinoma patients”. In: *Neural Computing and Applications* 34.23 (2022), pp. 21353–21365.

[1594] Jiao Liu et al. “Improved ReliefF-based feature selection algorithm for cancer histology”. In: *Biomedical Signal Processing and Control* (2023), p. 104980.

[1595] Ariana Rahman, Yining Zhang, and Jin G Park. “A Breast Cancer Recurrence Risk Predictor Using a Deep Learning Multi-omics Data Integration Framework”. In: *2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE. 2023, pp. 921–922.

[1596] Shiqi Li et al. “Multi-omics integration analysis of GPCRs in pan-cancer to uncover inter-omics relationships and potential driver genes”. In: *Computers in Biology and Medicine* 161 (2023), p. 106988.

[1597] Haiyao Dong et al. “Prediction of distinct populations of innate lymphoid cells by transcriptional profiles”. In: *Frontiers in Genetics* 14 (2023).

[1598] G Tamilmani et al. “Cancer MiRNA biomarker classification based on improved generative adversarial network optimized with Mayfly optimization algorithm”. In: *Biomedical Signal Processing and Control* 75 (2022), p. 103545.

[1599] Ningyi Zhang et al. "DeepGP: An Integrated Deep Learning Method for Endocrine Disease Gene Prediction Using Omics Data". In: *Frontiers in Cell and Developmental Biology* 9 (2021), p. 700061.

[1600] Andrew J Sedgewick et al. "Learning subgroup-specific regulatory interactions and regulator independence with PARADIGM". In: *Bioinformatics* 29.13 (2013), pp. i62–i70.

[1601] Abedalrhman Alkhateeb et al. "Multi-omics Data Integration Model based on Isomap and Convolutional Neural Network". In: *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE. 2022, pp. 1381–1385.

[1602] Hui-Qing Wang et al. "MMDAE-HGSOC: A novel method for high-grade serous ovarian cancer molecular subtypes classification based on multimodal deep autoencoder". In: *Computational Biology and Chemistry* 105 (2023), p. 107906.

[1603] Yuan Luo, Faraz S Ahmad, and Sanjiv J Shah. "Tensor factorization for precision medicine in heart failure with preserved ejection fraction". In: *Journal of cardiovascular translational research* 10 (2017), pp. 305–312.

[1604] Vignesh Venkat et al. "Investigating genes associated with heart failure, atrial fibrillation, and other cardiovascular diseases, and predicting disease using machine learning techniques for translational research and precision medicine". In: *Genomics* 115.2 (2023), p. 110584.

[1605] Shrey Gandhi et al. "Evolutionarily conserved transcriptional landscape of the heart defining the chamber specific physiology". In: *Genomics* 113.6 (2021), pp. 3782–3792.

[1606] Wei Xu et al. "Rapid computer-aided diagnosis of stroke by serum metabolic fingerprint based multi-modal recognition". In: *Advanced Science* 7.21 (2020), p. 2002021.

[1607] Fabien Lareyre et al. "Machine Learning and Omics Analysis in Aortic Aneurysm". In: *Angiology* (2023), p. 00033197231206427.

[1608] Annie M Westerlund et al. "Risk prediction of cardiovascular events by exploration of molecular data with explainable artificial intelligence". In: *International Journal of Molecular Sciences* 22.19 (2021), p. 10291.

[1609] Mirzat Turhon et al. "Development and validation of a deep learning model for prediction of intracranial aneurysm rupture risk based on multi-omics factor". In: *European Radiology* (2023), pp. 1–12.

[1610] Dabin Jeong et al. "GOAT: Gene-level biomarker discovery from multi-Omics data using graph ATtention neural network for eosinophilic asthma subtype". In: *Bioinformatics* 39.10 (2023), btad582.

[1611] Bing Li et al. "Function-Wise Dual-Omics analysis for radiation pneumonitis prediction in lung cancer patients". In: *Frontiers in Pharmacology* 13 (2022), p. 971849.

[1612] Lanlan Li et al. “Use of deep learning genomics to discriminate Alzheimer’s disease and healthy controls”. In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2021, pp. 5788–5791.

[1613] Giorgio Dolci et al. “Deep Generative Transfer Learning Predicts Conversion To AlzheimerS Disease From Neuroimaging Genomics Data”. In: *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE. 2023, pp. 1–5.

[1614] Seungeun Lee et al. “Integrating Genetic Information for Early Alzheimers Diagnosis through MRI Interpretation”. In: *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE. 2023, pp. 1–6.

[1615] Carlo Maj et al. “Integration of machine learning methods to dissect genetically imputed transcriptomic profiles in Alzheimers disease”. In: *Frontiers in genetics* 10 (2019), p. 726.

[1616] Yuting Zhang et al. “Predicting AT (N) pathologies in Alzheimers disease from blood-based proteomic data using neural networks”. In: *Frontiers in Aging Neuroscience* 14 (2022), p. 1040001.

[1617] Yu Jin et al. “Classification of Alzheimer’s disease using robust TabNet neural networks on genetic data”. In: *Mathematical biosciences and engineering: MBE* 20.5 (2023), pp. 8358–8374.

[1618] Naveen Sundar Gnanadesigan et al. “An integrated network topology and deep learning model for prediction of Alzheimer disease candidate genes”. In: *Soft Computing* (2023), pp. 1–15.

[1619] Giorgio Dolci et al. “A deep generative multimodal imaging genomics framework for Alzheimer’s disease prediction”. In: *2022 IEEE 22nd International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE. 2022, pp. 41–44.

[1620] Atlas Khan and Kai Wang. “A deep learning based scoring system for prioritizing susceptibility variants for mental disorders”. In: *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2017, pp. 1698–1705.

[1621] Atlas Khan, Qian Liu, and Kai Wang. “iMEGES: integrated mental-disorder GEnome score by deep neural network for prioritizing the susceptibility genes for mental disorders in personal genomes”. In: *BMC bioinformatics* 19 (2018), pp. 95–107.

[1622] Sundaram Laksshman et al. “DeepBipolar: Identifying genomic mutations for bipolar disorder via deep learning”. In: *Human mutation* 38.9 (2017), pp. 1217–1224.

[1623] S Karthik and M Sudha. “Predicting bipolar disorder and schizophrenia based on non-overlapping genetic phenotypes using deep neural network”. In: *Evolutionary Intelligence* 14 (2021), pp. 619–634.

[1624] Haishuai Wang, Paul Avillach, et al. "Diagnostic classification and prognostic prediction using common genetic variants in autism spectrum disorder: Genotype-based deep learning". In: *JMIR medical informatics* 9.4 (2021), e24754.

[1625] Ray O Bahado-Singh et al. "Artificial intelligence analysis of newborn leucocyte epigenomic markers for the prediction of autism". In: *Brain research* 1724 (2019), p. 146457.

[1626] Yousang Jo et al. "Interpretation of SNP combination effects on schizophrenia etiology based on stepwise deep learning with multi-precision data". In: *Briefings in Functional Genomics* (2023), elad041.

[1627] Martina Tarozzi et al. "Identification of recurrent genetic patterns from targeted sequencing panels with advanced data science: a case-study on sporadic and genetic neurodegenerative diseases". In: *BMC Medical Genomics* 15.1 (2022), p. 26.

[1628] Uppala Radhakrishna et al. "Placental DNA methylation profiles in opioid-exposed pregnancies and associations with the neonatal opioid withdrawal syndrome". In: *Genomics* 113.3 (2021), pp. 1127–1135.

[1629] Bojian Yin et al. "Using the structure of genome data in the design of deep neural networks for predicting amyotrophic lateral sclerosis from genotype". In: *Bioinformatics* 35.14 (2019), pp. i538–i547.

[1630] Jiyun Zhou et al. "Deep learning predicts DNA methylation regulatory variants in the human brain and elucidates the genetics of psychiatric disorders". In: *Proceedings of the National Academy of Sciences* 119.34 (2022), e2206069119.

[1631] Paula Esteller-Cucala et al. "Genomic analysis of the natural history of attention-deficit/hyperactivity disorder using Neanderthal and ancient Homo sapiens samples". In: *Scientific Reports* 10.1 (2020), p. 8622.

[1632] Sarah A Graham et al. "Artificial intelligence approaches to predicting and detecting cognitive decline in older adults: A conceptual review". In: *Psychiatry research* 284 (2020), p. 112732.

[1633] Michael Elgart, Susan Redline, and Tamar Sofer. "Machine and deep learning in molecular and genetic aspects of sleep research". In: *Neurotherapeutics* 18 (2021), pp. 228–243.

[1634] Noha E El-Attar, Bossy M Moustafa, and Wael A Awad. "Deep Learning Model to Detect Diabetes Mellitus Based on DNA Sequence." In: *Intelligent Automation & Soft Computing* 31.1 (2022).

[1635] Agata Wesolowska-Andersen et al. "Deep learning models predict regulatory variants in pancreatic islets and refine type 2 diabetes association signals". In: *Elife* 9 (2020), e51503.

[1636] Fang Xu et al. "Protein deep profile and model predictions for identifying the causal genes of male infertility based on deep learning". In: *Information Fusion* 75 (2021), pp. 70–89.

[1637] Simmi Kharb and Anagha Joshi. "Multi-omics and machine learning for the prevention and management of female reproductive health". In: *Frontiers in Endocrinology* 14 (2023), p. 358.

[1638] Charalampos Siristatidis et al. "Omics and artificial intelligence to improve in vitro fertilization (IVF) success: a proposed protocol". In: *Diagnostics* 11.5 (2021), p. 743.

[1639] Hao Wang et al. "Machine learning of plasma metabolome identifies biomarker panels for metabolic syndrome: findings from the China Suboptimal Health Cohort". In: *Cardiovascular Diabetology* 21.1 (2022), pp. 1–12.

[1640] Helena Castañé et al. "Coupling machine learning and lipidomics as a tool to investigate metabolic dysfunction-associated fatty liver disease. A general overview". In: *Biomolecules* 11.3 (2021), p. 473.

[1641] Hamid R Marateb et al. "Prediction of dyslipidemia using gene mutations, family history of diseases and anthropometric indicators in children and adolescents: the CASPIAN-III study". In: *Computational and structural biotechnology journal* 16 (2018), pp. 121–130.

[1642] Wanchen Ning et al. "Deep learning reveals key immunosuppression genes and distinct immunotypes in periodontitis". In: *Frontiers in Genetics* 12 (2021), p. 648329.

[1643] Sophia Y Wang, Benjamin Tseng, and Tina Hernandez-Boussard. "Deep learning approaches for predicting glaucoma progression using electronic health records and natural language processing". In: *Ophthalmology Science* 2.2 (2022), p. 100127.

[1644] Jakob Weiss et al. "Deep learning to estimate lung disease mortality from chest radiographs". In: *Nature Communications* 14.1 (2023), p. 2797.

[1645] Jesús M Antoñanzas et al. "Symptom-based predictive model of COVID-19 disease in children". In: *Viruses* 14.1 (2021), p. 63.

[1646] Cheng Song et al. "Artificial intelligence for HPV status prediction based on disease-specific images in head and neck cancer: A systematic review and meta-analysis". In: *Journal of Medical Virology* 95.9 (2023), e29080.

[1647] SN Manoharan, KMV Madan Kumar, and N Vadivelan. "A Novel CNN-TLSTM Approach for Dengue Disease Identification and Prevention using IoT-Fog Cloud Architecture". In: *Neural Processing Letters* 55.2 (2023), pp. 1951–1973.

[1648] Hong Zhang et al. "Predicting tumor mutational burden from liver cancer pathological images using convolutional neural network". In: *2019 IEEE international conference on bioinformatics and biomedicine (BIBM)*. IEEE. 2019, pp. 920–925.

[1649] Léon-Charles Tranchevent, Francisco Azuaje, and Jagath C Rajapakse. "A deep neural network approach to predicting clinical outcomes of neuroblastoma patients". In: *BMC medical genomics* 12 (2019), pp. 1–11.

[1650] Qasem Al-Tashi et al. "SwarmDeepSurv: swarm intelligence advances deep survival network for prognostic radiomics signatures in four solid cancers". In: *Patterns* 4.8 (2023).

[1651] Sijie Yao et al. "Prediction-oriented prognostic biomarker discovery with survival machine learning methods". In: *NAR Genomics and Bioinformatics* 5.2 (2023), lqad055.

[1652] Olivier B Poirion et al. "DeepProg: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data". In: *Genome medicine* 13.1 (2021), pp. 1–15.

[1653] Hua Chai et al. "Predicting bladder cancer prognosis by integrating multi-omics data through a transfer learning-based Cox proportional hazards network". In: *CCF Transactions on High Performance Computing* 3.3 (2021), pp. 311–319.

[1654] Long Liu et al. "Explainable deep transfer learning model for disease risk prediction using high-dimensional genomic data". In: *PLOS Computational Biology* 18.7 (2022), e1010328.

[1655] Wei Shao et al. "Characterizing the Survival-Associated Interactions between Tumor-infiltrating Lymphocytes and Tumors from Pathological Images and Multi-omics Data". In: *IEEE Transactions on Medical Imaging* (2023).

[1656] Shuo Wang et al. "A novel deep learning method to predict lung cancer long-term survival with biological knowledge incorporated gene expression images and clinical data". In: *Frontiers in Genetics* 13 (2022), p. 800853.

[1657] Hanyu Zhang et al. "Chronic kidney disease survival prediction with artificial neural networks". In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 1351–1356.

[1658] Li Tong et al. "Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis". In: *BMC medical informatics and decision making* 20 (2020), pp. 1–12.

[1659] Arnór I Sigurdsson et al. "Deep integrative models for large-scale human genomics". In: *Nucleic Acids Research* 51.12 (2023), e67–e67.

[1660] Wenkai Han et al. "Predicting the antigenic evolution of SARS-COV-2 with deep learning". In: *Nature Communications* 14.1 (2023), p. 3478.

[1661] Guangyu Wang et al. "Deep-learning-enabled protein–protein interaction analysis for prediction of SARS-CoV-2 infectivity and variant evolution". In: *Nature Medicine* 29.8 (2023), pp. 2007–2018.

[1662] Jiahui Chen et al. "Persistent Laplacian projected Omicron BA. 4 and BA. 5 to become new dominating variants". In: *Computers in Biology and Medicine* 151 (2022), p. 106262.

[1663] Drew H Bryant et al. "Deep diversification of an AAV capsid protein by machine learning". In: *Nature Biotechnology* 39.6 (2021), pp. 691–696.

[1664] Thanh Thi Nguyen et al. "Genomic mutations and changes in protein secondary structure and solvent accessibility of SARS-CoV-2 (COVID-19 virus)". In: *Scientific Reports* 11.1 (2021), p. 3487.

[1665] Yanhua Xu and Dominik Wojtczak. "Predicting influenza A viral host using PSSM and word embeddings". In: *2021 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE. 2021, pp. 1–10.

[1666] Zhaoyan Ming et al. "HostNet: improved sequence representation in deep neural networks for virus-host prediction". In: *BMC bioinformatics* 24.1 (2023), p. 455.

[1667] Qian Guo et al. "Predicting hosts based on early SARS-CoV-2 samples and analyzing the 2020 pandemic". In: *Scientific Reports* 11.1 (2021), p. 17422.

[1668] Haodong Xu, Peilin Jia, and Zhongming Zhao. "DeepVISP: deep learning for virus site integration prediction and motif discovery". In: *Advanced Science* 8.9 (2021), p. 2004958.

[1669] Haodong Xu et al. "Deep learning for detecting and elucidating human T-cell leukemia virus type 1 integration in the human genome". In: *Patterns* 4.2 (2023).

[1670] Geeta Rani et al. "Applying deep learning-based multi-modal for detection of coronavirus". In: *Multimedia Systems* (2021), pp. 1–12.

[1671] Yuanyuan Bu et al. "COPPER: an ensemble deep-learning approach for identifying exclusive virus-derived small interfering RNAs in plants". In: *Briefings in Functional Genomics* 22.3 (2023), pp. 274–280.

[1672] Shabnam Ataee, Xavier Brochet, and Carlos Andrés Peña-Reyes. "Bacteriophage Genetic Edition using LSTM". In: *Frontiers in Bioinformatics* 2 (2022), p. 932319.

[1673] Lefteris Koumakis. "Deep learning models in genomics; are we there yet?" In: *Computational and Structural Biotechnology Journal* 18 (2020), pp. 1466–1473.

[1674] Eugene Lin, Chieh-Hsin Lin, and Hsien-Yuan Lane. "Deep learning with neuroimaging and genomics in Alzheimers disease". In: *International Journal of Molecular Sciences* 22.15 (2021), p. 7911.

[1675] Dongming Liu et al. "Imaging-genomics in glioblastoma: Combining molecular and imaging signatures". In: *Frontiers in Oncology* 11 (2021), p. 699265.

[1676] Yu-Chiao Chiu et al. "Deep learning of pharmacogenomics resources: moving towards precision oncology". In: *Briefings in bioinformatics* 21.6 (2020), pp. 2066–2083.

[1677] Matteo Bastico et al. "DrOGA: an artificial intelligence solution for driver-status prediction of genomics mutations in precision cancer medicine". In: *IEEE Access* (2023).

[1678] Sarah J MacEachern and Nils D Forkert. "Machine learning for precision medicine". In: *Genome* 64.4 (2021), pp. 416–425.

[1679] Godwin Woo et al. "DeepCOP: deep learning-based approach to predict gene regulating effects of small molecules". In: *Bioinformatics* 36.3 (2020), pp. 813–818.

[1680] Glaucia C Pereira. "Genomics and artificial intelligence working together in drug discovery and repositioning: the advent of adaptive pharmacogenomics in glioblastoma and chronic arterial inflammation therapies". In: *Biotechnology and production of anti-cancer compounds* (2017), pp. 253–281.

[1681] Mustafa Al-Maini et al. "Artificial intelligence-based preventive, personalized and precision medicine for cardiovascular disease/stroke risk assessment in rheumatoid arthritis patients: a narrative review". In: *Rheumatology International* 43.11 (2023), pp. 1965–1982.

[1682] Dmitry Grapov et al. "Rise of deep learning for genomic, proteomic, and metabolomic data integration in precision medicine". In: *Omics: a journal of integrative biology* 22.10 (2018), pp. 630–636.

[1683] Dibyendu Dana et al. "Deep learning in drug discovery and medicine; scratching the surface". In: *Molecules* 23.9 (2018), p. 2384.

[1684] Jinzhuang Liao et al. "Artificial intelligence assists precision medicine in cancer treatment". In: *Frontiers in Oncology* 12 (2023), p. 998222.

[1685] Fatima Noor et al. "Machine learning for synergistic network pharmacology: a comprehensive overview". In: *Briefings in Bioinformatics* (2023), bbad120.

[1686] Sara Momtazmanesh, Ali Nowroozi, and Nima Rezaei. "Artificial intelligence in rheumatoid arthritis: current status and future perspectives: a state-of-the-art review". In: *Rheumatology and Therapy* 9.5 (2022), pp. 1249–1304.

[1687] Minsik Oh et al. "DRIM: a web-based system for investigating drug response at the molecular level by condition-specific multi-omics data integration". In: *Frontiers in Genetics* 11 (2020), p. 564792.

[1688] Michael R Kosorok and Eric B Laber. "Precision medicine". In: *Annual review of statistics and its application* 6 (2019), pp. 263–286.

[1689] Hamidreza Taleghamar et al. "Deep learning of quantitative ultrasound multi-parametric images at pre-treatment to predict breast cancer response to chemotherapy". In: *Scientific reports* 12.1 (2022), p. 2244.

[1690] Shelli F Farhadian et al. "HIV viral transcription and immune perturbations in the CNS of people with HIV despite ART". In: *JCI insight* 7.13 (2022).

[1691] Siqi Chen et al. "DNN-PNN: A parallel deep neural network model to improve anticancer drug sensitivity". In: *Methods* 209 (2023), pp. 1–9.

[1692] Min Li et al. "DeepDSC: a deep learning method to predict drug sensitivity of cancer cell lines". In: *IEEE/ACM transactions on computational biology and bioinformatics* 18.2 (2019), pp. 575–582.

[1693] Xiao-Ying Liu and Xin-Yue Mei. "Prediction of drug sensitivity based on multi-omics data using deep learning and similarity network fusion approaches". In: *Frontiers in Bioengineering and Biotechnology* 11 (2023), p. 1156372.

[1694] Conghao Wang et al. "Deep learning and multi-omics approach to predict drug responses in cancer". In: *BMC bioinformatics* 22.10 (2021), pp. 1–15.

[1695] Yimeng Wang et al. "XGraphCDS: An explainable deep learning model for predicting drug sensitivity from gene pathways and chemical structures". In: *Computers in Biology and Medicine* 168 (2024), p. 107746.

[1696] Yongcui Wang et al. "DeepDRK: a deep learning framework for drug repurposing through kernel-based multi-omics integration". In: *Briefings in Bioinformatics* 22.5 (2021), bbab048.

[1697] Davinder Paul Singh and Baijnath Kaushik. "CTDN (Convolutional Temporal Based Deep-Neural Network): An Improvised Stacked Hybrid Computational Approach for Anticancer Drug Response Prediction". In: *Computational Biology and Chemistry* 105 (2023), p. 107868.

[1698] Sejin Park, Jihee Soh, and Hyunju Lee. "Super. FELT: supervised feature extraction learning using triplet loss for drug response prediction with multi-omics data". In: *BMC bioinformatics* 22.1 (2021), pp. 1–23.

[1699] Yi-Ching Tang and Assaf Gottlieb. "Explainable drug sensitivity prediction through cancer pathway enrichment". In: *Scientific reports* 11.1 (2021), p. 3128.

[1700] Qi Wei and Stephen A Ramsey. "Predicting chemotherapy response using a variational autoencoder approach". In: *Bmc Bioinformatics* 22.1 (2021), pp. 1–22.

[1701] Deeba Khan and Seema Shedole. "Leveraging deep learning techniques and integrated omics data for tailored treatment of breast cancer". In: *Journal of Personalized Medicine* 12.5 (2022), p. 674.

[1702] Yi Yang et al. "A multi-omics-based serial deep learning approach to predict clinical outcomes of single-agent anti-PD-1/PD-L1 immunotherapy in advanced stage non-small-cell lung cancer". In: *American journal of translational research* 13.2 (2021), p. 743.

[1703] Hong Wang et al. "GADRP: graph convolutional networks and autoencoders for cancer drug response prediction". In: *Briefings in Bioinformatics* 24.1 (2023), bbac501.

[1704] Abhishek Majumdar et al. "kESVR: An ensemble model for drug response prediction in precision medicine using cancer cell lines gene expression". In: *Genes* 12.6 (2021), p. 844.

[1705] Likun Jiang et al. "DeepTTA: a transformer-based model for predicting cancer drug response". In: *Briefings in bioinformatics* 23.3 (2022), bbac100.

[1706] Prateek Singh et al. "A machine learning-based approach to determine infection status in recipients of BBV152 (Covaxin) whole-virion inactivated SARS-CoV-2 vaccine for serological surveys". In: *Computers in Biology and Medicine* 146 (2022), p. 105419.

[1707] Rui Wang et al. "Emerging vaccine-breakthrough SARS-CoV-2 variants". In: *ACS infectious diseases* 8.3 (2022), pp. 546–556.

[1708] Rui Wang et al. "Vaccine-escape and fast-growing mutations in the United Kingdom, the United States, Singapore, Spain, India, and other COVID-19-devastated countries". In: *Genomics* 113.4 (2021), pp. 2158–2170.

[1709] Margo VanOeffelen et al. "A genomic data resource for predicting antimicrobial resistance from laboratory-derived antimicrobial susceptibility phenotypes". In: *Briefings in Bioinformatics* 22.6 (2021), bbab313.

[1710] Ran Su et al. "Deep-Resp-Forest: a deep forest model to predict anti-cancer drug response". In: *Methods* 166 (2019), pp. 91–102.

[1711] Tamer N Jarada, Jon G Rokne, and Reda Alhajj. "SNF–CVAE: computational method to predict drug–disease interactions using similarity network fusion and collective variational autoencoder". In: *Knowledge-Based Systems* 212 (2021), p. 106585.

[1712] Thang Chu et al. "Graph Transformer for drug response prediction". In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 20.2 (2022), pp. 1065–1072.

[1713] Aron Park, Yeeun Lee, and Seungyoon Nam. "A performance evaluation of drug response prediction models for individual drugs". In: *Scientific Reports* 13.1 (2023), p. 11911.

[1714] Feng Xie et al. "Multifactorial deep learning reveals pan-cancer genomic tumor clusters with distinct immunogenomic landscape and response to immunotherapy". In: *Clinical Cancer Research* 26.12 (2020), pp. 2908–2920.

[1715] Tingting Zhao et al. "Identification of significant gene expression changes in multiple perturbation experiments using knockoffs". In: *Briefings in Bioinformatics* 24.2 (2023), bbad084.

[1716] Jia Hu et al. "Prediction of clinicopathological features, multi-omics events and prognosis based on digital pathology and deep learning in HR+/HER2- breast cancer". In: *Journal of Thoracic Disease* 15.5 (2023), p. 2528.

[1717] Sotiris Ouzounis et al. "A Robust Machine Learning Framework Built Upon Molecular Representations Predicts CYP450 Inhibition: Toward Precision in Drug Repurposing". In: *OMICS: A Journal of Integrative Biology* 27.7 (2023), pp. 305–314.

[1718] Eunchong Huang, Sarah Kim, and TaeJin Ahn. "Deep learning for integrated analysis of insulin resistance with multi-omics data". In: *Journal of Personalized Medicine* 11.2 (2021), p. 128.

[1719] Maria Giovanna Danieli et al. "A machine learning analysis to predict the response to intravenous and subcutaneous immunoglobulin in inflammatory myopathies. A proposal for a future multi-omics approach in autoimmune diseases". In: *Autoimmunity Reviews* 21.6 (2022), p. 103105.

[1720] Hossein Sharifi-Noghabi et al. "MOLI: multi-omics late integration with deep neural networks for drug response prediction". In: *Bioinformatics* 35.14 (2019), pp. i501–i509.

[1721] Jielin Xu et al. "Interpretable deep learning translation of GWAS and multi-omics findings to identify pathobiology and drug repurposing in Alzheimers disease". In: *Cell reports* 41.9 (2022), p. 111717.

[1722] Xia-an Bi et al. "A novel generation adversarial network framework with characteristics aggregation and diffusion for brain disease classification and feature selection". In: *Briefings in Bioinformatics* 23.6 (2022), bbac454.

[1723] Guanghui Li et al. "DeepWalk-aware graph attention networks with CNN for circRNA–drug sensitivity association identification". In: *Briefings in Functional Genomics* (2023), elad053.

[1724] Alireza Tafazoli et al. "Pharmacovariome scanning using whole pharmacogene resequencing coupled with deep computational analysis and machine learning for clinical pharmacogenomics". In: *Human Genomics* 17.1 (2023), p. 62.

[1725] Hiroyuki Kurata and Sho Tsukiyama. "ICAN: Interpretable cross-attention network for identifying drug and target protein interactions". In: *PLoS One* 17.10 (2022), e0276609.

[1726] Yu Peng et al. "LGBMDF: A cascade forest framework with LightGBM for predicting drug-target interactions". In: *Frontiers in Microbiology* 13 (2023), p. 1092467.

[1727] Mohammad Mustafa Kazan, Misgana Mengistu Asmare, and Rajani Kanta Mahapatra. "Identification of Potential Drug Targets in Erythrocyte Invasion Pathway of Plasmodium falciparum". In: *Current Microbiology* 80.5 (2023), p. 165.

[1728] Jiaying You, Robert D McLeod, and Pingzhao Hu. "Predicting drug-target interaction network using deep learning model". In: *Computational biology and chemistry* 80 (2019), pp. 90–101.

[1729] Zhong-Hao Ren et al. "DeepMPF: deep learning framework for predicting drug–target interactions based on multi-modal representation with meta-path semantic analysis". In: *Journal of Translational Medicine* 21.1 (2023), pp. 1–18.

[1730] Nansu Zong et al. "Deep mining heterogeneous networks of biomedical linked data to predict novel drug–target associations". In: *Bioinformatics* 33.15 (2017), pp. 2337–2344.

[1731] Jie Zheng, Xuan Xiao, and Wang-Ren Qiu. "DTI-BERT: identifying drug-target interactions in cellular networking based on BERT and deep learning method". In: *Frontiers in Genetics* 13 (2022), p. 859188.

[1732] Masatoshi Hamanaka et al. "CGBVS-DNN: Prediction of Compound-protein Interactions Based on Deep Learning". In: *Molecular informatics* 36.1-2 (2017), p. 1600045.

[1733] Lingwei Xie et al. "Deep learning-based transcriptome data classification for drug-target interaction prediction". In: *BMC genomics* 19 (2018), pp. 93–102.

[1734] Nobuaki Yasuo, Yusuke Nakashima, and Masakazu Sekijima. "Code-dti: Collaborative deep learning-based drug-target interaction prediction". In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 792–797.

[1735] Leiming Xia et al. "Drug-target binding affinity prediction using message passing neural network and self supervised learning". In: *BMC genomics* 24.1 (2023), p. 557.

[1736] Ying Zheng and Zheng Wu. "Cascade Deep Forest With Heterogeneous Similarity Measures for Drug–Target Interaction Prediction". In: *Frontiers in Genetics* 12 (2021), p. 702259.

[1737] Kejie Fang et al. "ColdDTA: utilizing data augmentation and attention-based feature fusion for drug-target binding affinity prediction". In: *Computers in Biology and Medicine* 164 (2023), p. 107372.

[1738] Zhenyu Wu et al. "Single-cell techniques and deep learning in predicting drug response". In: *Trends in pharmacological sciences* 41.12 (2020), pp. 1050–1065.

[1739] Zhaorui Zuo et al. "SWnet: a deep learning model for drug response prediction from cancer genomic signatures and compound chemical structures". In: *BMC bioinformatics* 22.1 (2021), pp. 1–16.

[1740] Heming Zhang, Yixin Chen, and Fuhai Li. "Predicting anticancer drug response with deep learning constrained by signaling pathways". In: *Frontiers in Bioinformatics* 1 (2021), p. 639349.

[1741] Minjae Joo et al. "A deep learning model for cell growth inhibition IC50 prediction and its application for gastric cancer patients". In: *International journal of molecular sciences* 20.24 (2019), p. 6276.

[1742] Giang TT Nguyen, Hoa D Vu, and Duc-Hau Le. "Integrating molecular graph data of drugs and multiple-omic data of cell lines for drug response prediction". In: *IEEE/ACM transactions on computational biology and bioinformatics* 19.2 (2021), pp. 710–717.

[1743] Xiaoqing Huang et al. "ParsVNN: parsimony visible neural networks for uncovering cancer-specific and drug-sensitive genes and pathways". In: *NAR genomics and bioinformatics* 3.4 (2021), lqab097.

[1744] Xiaoxiao Cheng et al. "NeRD: a multichannel neural network to predict cellular response of drugs by integrating multidimensional data". In: *BMC medicine* 20.1 (2022), pp. 1–16.

[1745] Aron Park et al. "A comprehensive evaluation of regression-based drug responsiveness prediction models, using cell viability inhibitory concentrations (IC50 values)". In: *Bioinformatics* 38.10 (2022), pp. 2810–2817.

[1746] Bo Xu et al. "Full-attention based drug drug interaction extraction exploiting user-generated content". In: *2018 IEEE international conference on bioinformatics and biomedicine (BIBM)*. IEEE. 2018, pp. 560–565.

[1747] Shengnan She et al. "Deep learning-based multi-drug synergy prediction model for individually tailored anti-cancer therapies". In: *Frontiers in Pharmacology* 13 (2022), p. 1032875.

[1748] Yanglan Gan et al. "Predicting synergistic anticancer drug combination based on low-rank global attention mechanism and bilinear predictor". In: *Bioinformatics* 39.10 (2023), btad607.

[1749] Anna Torkamannia, Yadollah Omidi, and Reza Ferdousi. "SYNDEEP: a deep learning approach for the prediction of cancer drugs synergy". In: *Scientific Reports* 13.1 (2023), p. 6184.

[1750] António J Preto et al. "SYNPRED: prediction of drug combination effects in cancer using different synergy metrics and ensemble learning". In: *GigaScience* 11 (2022), giac087.

[1751] Hanan Ahmed et al. "Enhanced deep learning model for personalized cancer treatment". In: *IEEE Access* 10 (2022), pp. 106050–106058.

[1752] Benedek Rozemberczki et al. "MOOMIN: Deep Molecular Omics Network for Anti-Cancer Drug Combination Therapy". In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2022, pp. 3472–3483.

[1753] Xiaowen Wang et al. "PRODeepSyn: predicting anticancer synergistic drug combinations by embedding cell lines with protein–protein interaction network". In: *Briefings in Bioinformatics* 23.2 (2022), bbab587.

[1754] Qiao Liu and Lei Xie. "TranSynergy: Mechanism-driven interpretable deep neural network for the synergistic prediction and pathway deconvolution of drug combinations". In: *PLoS computational biology* 17.2 (2021), e1008653.

[1755] Luc Blassel et al. "Drug resistance mutations in HIV: new bioinformatics approaches and challenges". In: *Current opinion in virology* 51 (2021), pp. 56–64.

[1756] Sam Benkwitz-Bedford et al. "Machine learning prediction of resistance to subinhibitory antimicrobial concentrations from Escherichia coli genomes". In: *Msystems* 6.4 (2021), pp. 10–1128.

[1757] Xingyan Kuang et al. "Accurate and rapid prediction of tuberculosis drug resistance from genome sequence data using traditional machine learning algorithms and CNN". In: *Scientific reports* 12.1 (2022), p. 2427.

[1758] Andrew Zhang, Ling Teng, and Gil Alterovitz. "An explainable machine learning platform for pyrazinamide resistance prediction and genetic feature identification of Mycobacterium tuberculosis". In: *Journal of the American Medical Informatics Association* 28.3 (2021), pp. 533–540.

[1759] Anna G Green et al. "A convolutional neural network highlights mutations relevant to antimicrobial resistance in Mycobacterium tuberculosis". In: *Nature communications* 13.1 (2022), p. 3817.

[1760] Yang Yang et al. "DeepAMR for predicting co-occurrent resistance of Mycobacterium tuberculosis". In: *Bioinformatics* 35.18 (2019), pp. 3240–3249.

[1761] Boya Ji et al. "HyperVR: a hybrid deep ensemble learning approach for simultaneously predicting virulence factors and antibiotic resistance genes". In: *NAR Genomics and Bioinformatics* 5.1 (2023), lqad012.

[1762] Yang Yang et al. "An end-to-end heterogeneous graph attention network for Mycobacterium tuberculosis drug-resistance prediction". In: *Briefings in Bioinformatics* 22.6 (2021), bbab299.

[1763] Margaret C Steiner, Keylie M Gibson, and Keith A Crandall. "Drug resistance prediction using deep learning techniques on HIV-1 sequence data". In: *Viruses* 12.5 (2020), p. 560.

[1764] Mykola Galushka et al. "Toxicity prediction using pre-trained autoencoder". In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 299–304.

[1765] Badireddi Subathra Lakshmi, Mohan Latha Abillasha, and Pandjassarame Kangueane. "Emerging Technologies for Antiviral Drug Discovery". In: *Global Virology III: Virology in the 21st Century* (2019), pp. 59–71.

[1766] Lei Deng et al. "Graph2MDA: a multi-modal variational graph embedding model for predicting microbe–drug associations". In: *Bioinformatics* 38.4 (2022), pp. 1118–1125.

[1767] Jannis Born et al. "PaccMannRL: De novo generation of hit-like anticancer molecules from transcriptomic data via reinforcement learning". In: *Iscience* 24.4 (2021).

[1768] Alokkumar Jha et al. "Deep convolution neural network model to predict relapse in breast cancer". In: *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE. 2018, pp. 351–358.

[1769] Pengfei Liu et al. "Improving prediction of phenotypic drug response on cancer cell lines using deep convolutional network". In: *BMC bioinformatics* 20.1 (2019), pp. 1–14.

[1770] Hui Liu et al. "NeuMF: Predicting Anti-cancer Drug Response Through a Neural Matrix Factorization Model". In: *Current Bioinformatics* 17.9 (2022), pp. 835–847.

[1771] Alok Sharma et al. “DeepInsight-3D architecture for anti-cancer drug response prediction with deep-learning on multi-omics”. In: *Scientific Reports* 13.1 (2023), p. 2483.

[1772] Wei Gan et al. “A computational method to predict topologically associating domain boundaries combining histone marks and sequence information”. In: *BMC genomics* 20.13 (2019), pp. 1–12.

[1773] Atsushi Yoshimori et al. “Strategies for design of molecular structures with a desired pharmacophore using deep reinforcement learning”. In: *Chemical and Pharmaceutical Bulletin* 68.3 (2020), pp. 227–233.

[1774] Xuhan Liu et al. “An exploration strategy improves the diversity of de novo ligands using deep reinforcement learning: a case for the adenosine A2A receptor”. In: *Journal of cheminformatics* 11.1 (2019), p. 35.

[1775] Bo Wen et al. “Cancer neoantigen prioritization through sensitive and reliable proteogenomics analysis”. In: *Nature communications* 11.1 (2020), p. 1759.

[1776] Chenkai Li et al. “AMPlify: attentive deep learning model for discovery of novel antimicrobial peptides effective against WHO priority pathogens”. In: *BMC genomics* 23.1 (2022), p. 77.

[1777] Zichang Xu et al. “Advances in antibody discovery from human BCR repertoires”. In: *Frontiers in Bioinformatics* 2 (2022), p. 1044975.

[1778] Derek M Mason et al. “Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning”. In: *Nature Biomedical Engineering* 5.6 (2021), pp. 600–612.

[1779] Emily K Makowski et al. “Co-optimization of therapeutic antibody affinity and specificity using machine learning models that generalize to novel mutational space”. In: *Nature communications* 13.1 (2022), p. 3788.

[1780] Igor Kozlovskii and Petr Popov. “Structure-based deep learning for binding site detection in nucleic acid macromolecules”. In: *NAR Genomics and Bioinformatics* 3.4 (2021), lqab111.

[1781] Li Xue et al. “A deep learning framework for sequence-based bacteria type IV secreted effectors prediction”. In: *Chemometrics and Intelligent Laboratory Systems* 183 (2018), pp. 134–139.

[1782] Youngju Kim et al. “Characterization of Klebsiella pneumoniae bacteriophages, KP1 and KP12, with deep learning-based structure prediction”. In: *Frontiers in Microbiology* 13 (2023), p. 5289.

[1783] Lucia Grenga et al. “Analyzing the complex regulatory landscape of Hfq–an integrative, multi-omics approach”. In: *Frontiers in microbiology* 8 (2017), p. 1784.

[1784] Sajad Ali et al. “Revisiting Alternaria-host interactions: New insights on its pathogenesis, defense mechanisms and control strategies”. In: *Scientia Horticulturae* 322 (2023), p. 112424.

[1785] Marcus Nguyen et al. "Using machine learning to predict antimicrobial MICs and associated genomic features for nontyphoidal Salmonella". In: *Journal of clinical microbiology* 57.2 (2019), pp. 10–1128.

[1786] Scott N Dean et al. "PepVAE: variational autoencoder framework for antimicrobial peptide generation and activity prediction". In: *Frontiers in microbiology* 12 (2021), p. 725727.

[1787] Yungang Xu et al. "Deep learning of the splicing (epi) genetic code reveals a novel candidate mechanism linking histone modifications to ESC fate decision". In: *Nucleic acids research* 45.21 (2017), pp. 12100–12112.

[1788] David Martínez-Enguita et al. "NCAE: data-driven representations using a deep network-coherent DNA methylation autoencoder identify robust disease and risk factor signatures". In: *Briefings in Bioinformatics* 24.5 (2023), bbad293.

[1789] Chaitanya Srinivasan et al. "Addiction-associated genetic variants implicate brain cell type-and region-specific cis-regulatory elements in addiction neurobiology". In: *Journal of Neuroscience* 41.43 (2021), pp. 9008–9030.

[1790] Hossein Sharifi-Noghabi et al. "AITL: adversarial inductive transfer learning with input and output space adaptation for pharmacogenomics". In: *Bioinformatics* 36.Supplement_1 (2020), pp. i380–i388.

[1791] Haitham A Elmarakeby et al. "Biologically informed deep neural network for prostate cancer discovery". In: *Nature* 598.7880 (2021), pp. 348–352.

[1792] Chao Fang et al. "DeePaN: dee p pa tient graph convolutional n etwork integrating clinico-genomic evidence to stratify lung cancers for immunotherapy". In: *NPJ digital medicine* 4.1 (2021), p. 14.

[1793] Li Zhang et al. "Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma". In: *Frontiers in genetics* 9 (2018), p. 477.

[1794] Satoshi Takahashi et al. "Predicting deep learning based multi-omics parallel integration survival subtypes in lung cancer using reverse phase protein array data". In: *Biomolecules* 10.10 (2020), p. 1460.

[1795] Weijia Li et al. "A deep learning approach based on multi-omics data integration to construct a risk stratification prediction model for skin cutaneous melanoma". In: *Journal of Cancer Research and Clinical Oncology* 149.17 (2023), pp. 15923–15938.

[1796] Ann-Kristin Becker et al. "Analysis of epidemiological association patterns of serum thyrotropin by combining random forests and Bayesian networks". In: *Plos one* 17.7 (2022), e0271610.

[1797] Tao Sun et al. "Genome-wide association study-based deep learning for survival prediction". In: *Statistics in medicine* 39.30 (2020), pp. 4605–4620.

[1798] Young-Ji Lee, Jun-Hyung Park, and Seung-Ho Lee. "A study on the prediction of cancer using whole-genome data and deep learning". In: *International Journal of Molecular Sciences* 23.18 (2022), p. 10396.

[1799] Lisa Barros de Andrade e Sousa et al. “Kinetics of Xist-induced gene silencing can be predicted from combinations of epigenetic and genomic features”. In: *Genome research* 29.7 (2019), pp. 1087–1099.

[1800] Hai Wang et al. “Deep learning for plant genomics and crop improvement”. In: *Current opinion in plant biology* 54 (2020), pp. 34–41.

[1801] Yifan Wang et al. “A deep learning approach to automate whole-genome prediction of diverse epigenomic modifications in plants”. In: *New Phytologist* 232.2 (2021), pp. 880–897.

[1802] Ashlyn Rairdin et al. “Deep learning-based phenotyping for genome wide association studies of sudden death syndrome in soybean”. In: *Frontiers in plant science* 13 (2022), p. 966244.

[1803] Seongmun Jeong, Jae-Yoon Kim, and Namshin Kim. “GMStool: GWAS-based marker selection tool for genomic prediction from genomic data”. In: *Scientific reports* 10.1 (2020), p. 19653.

[1804] Upendra Kumar Pradhan et al. “ASmiR: a machine learning framework for prediction of abiotic stress–specific miRNAs in plants”. In: *Functional & Integrative Genomics* 23.2 (2023), p. 92.

[1805] Zahra Agharbaoui et al. “An integrative approach to identify hexaploid wheat miRNAome associated with development and tolerance to abiotic stress”. In: *BMC genomics* 16 (2015), pp. 1–17.

[1806] Yating Zhong et al. “MODILM: towards better complex diseases classification using a novel multi-omics data integration learning model”. In: *BMC Medical Informatics and Decision Making* 23.1 (2023), pp. 1–18.

[1807] Gautier Viaud, Prasanna Mayilvahanan, and Paul-Henry Cournède. “Representation learning for the clustering of multi-omics data”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19.1 (2021), pp. 135–145.

[1808] Hakim Benkirane et al. “CustOmics: A versatile deep-learning based strategy for multi-omics integration”. In: *PLoS Computational Biology* 19.3 (2023), e1010921.

[1809] Munhwan Lee et al. “Gene-centric multi-omics integration with convolutional encoders for cancer drug response prediction”. In: *Computers in Biology and Medicine* 151 (2022), p. 106192.

[1810] Sehwan Moon and Hyunju Lee. “JDSNMF: Joint deep semi-non-negative matrix factorization for learning integrative representation of molecular signals in Alzheimers disease”. In: *Journal of Personalized Medicine* 11.8 (2021), p. 686.

[1811] Tongxin Wang et al. “MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification”. In: *Nature Communications* 12.1 (2021), p. 3445.

[1812] Muhammad Luthfi et al. "Comparative omics analysis for novel target discovery in plant pathogens: A case study for Magnaporthe oryzae". In: *Plant Pathology* (2023).

[1813] Chengming Zhang et al. "Contrastively generative self-expression model for single-cell and spatial multimodal data". In: *Briefings in Bioinformatics* 24.5 (2023), bbad265.

[1814] Muhammad Usman Tariq and Fahad Saeed. "SpeCollate: Deep cross-modal similarity network for mass spectrometry data based peptide deductions". In: *PloS one* 16.10 (2021), e0259349.

[1815] Chayan Maitra et al. "Unsupervised neural network for single cell Multi-omics INTegration (UMINT): an application to health and disease". In: *Frontiers in Molecular Biosciences* 10 (2023), p. 1184748.

[1816] Chaozhong Liu, Linhua Wang, and Zhandong Liu. "Single-cell multi-omics integration for unpaired data by a siamese network with graph-based contrastive loss". In: *BMC bioinformatics* 24.1 (2023), p. 5.

[1817] Ling Du et al. "Uncertainty-aware dynamic integration for multi-omics classification of tumors". In: *Journal of Cancer Research and Clinical Oncology* 149.7 (2023), pp. 3301–3312.

[1818] Nikola Simidjievski et al. "Variational autoencoders for cancer data integration: design principles and computational practice". In: *Frontiers in genetics* 10 (2019), p. 1205.

[1819] Yingxin Cao et al. "Integrated analysis of multimodal single-cell data with structural similarity". In: *Nucleic acids research* 50.21 (2022), e121–e121.

[1820] Jan Zrimec et al. "Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure". In: *Nature communications* 11.1 (2020), p. 6141.

[1821] Xiaoyu Zhang et al. "OmiEmbed: a unified multi-task deep learning framework for multi-omics data". In: *Cancers* 13.12 (2021), p. 3047.

[1822] Mingon Kang, Euiseong Ko, and Tesfaye B Mersha. "A roadmap for multi-omics data integration using deep learning". In: *Briefings in Bioinformatics* 23.1 (2022), bbab454.

[1823] Zhiwei Rong et al. "MCluster-VAEs: an end-to-end variational deep learning-based clustering method for subtype discovery using multi-omics data". In: *Computers in Biology and Medicine* 150 (2022), p. 106085.

[1824] Bashier ElKarami et al. "Multi-omics data integration model based on UMAP embedding and convolutional neural network". In: *Cancer Informatics* 21 (2022), p. 11769351221124205.

[1825] Tasbiraha Athaya et al. "Multimodal deep learning approaches for single-cell multi-omics data integration". In: *Briefings in Bioinformatics* 24.5 (2023), bbad313.

[1826] Wenwen Cui et al. “Compressing genomic sequences by using deep learning”. In: *International Conference on Artificial Neural Networks*. Springer. 2020, pp. 92–104.

[1827] Jacob Schreiber et al. “Avocado: a multi-scale deep tensor factorization method learns a latent representation of the human epigenome”. In: *Genome biology* 21.1 (2020), pp. 1–18.

[1828] Roman Kudrin and Andrey Mironov. “Inferring chromatin states with stochastic autoencoder”. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 2477–2478.

[1829] Nicolas Mathis et al. “Predicting prime editing efficiency and product purity by deep learning”. In: *Nature Biotechnology* (2023), pp. 1–9.

[1830] Tao Cui and Tingting Wang. “JOINT for large-scale single-cell RNA-sequencing analysis via soft-clustering and parallel computing”. In: *BMC genomics* 22 (2021), pp. 1–16.

[1831] Haoyue Fu, Lianping Yang, and Xiangde Zhang. “Noncoding variants functional prioritization methods based on predicted regulatory factor binding sites”. In: *Current Genomics* 18.4 (2017), pp. 322–331.

[1832] Tanzila Islam et al. “DeepCGP: A Deep Learning Method to Compress Genome-wide Polymorphisms for Predicting Phenotype of Rice”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* (2023).

[1833] Haojiang Tan et al. “Weighted deep factorizing heterogeneous molecular network for genome-phenome association prediction”. In: *Methods* 205 (2022), pp. 18–28.

[1834] Ehsaneddin Asgari and Mohammad RK Mofrad. “Continuous distributed representation of biological sequences for deep proteomics and genomics”. In: *PloS one* 10.11 (2015), e0141287.

[1835] Somnath Rakshit et al. “Deep learning for integrated analysis of breast cancer subtype specific multi-omics data”. In: *TENCON 2018-2018 IEEE region 10 conference*. IEEE. 2018, pp. 1917–1922.

[1836] Chengming Zhang et al. “Deep latent space fusion for adaptive representation of heterogeneous multi-omics data”. In: *Briefings in Bioinformatics* 23.2 (2022), bbab600.

[1837] María Martínez-García and Pablo M Olmos. “Handling ill-conditioned omics data with deep probabilistic models”. In: *IEEE Journal of Biomedical and Health Informatics* (2023).

[1838] Tanmay Nath et al. “A machine learning based approach towards high-dimensional mediation analysis”. In: *NeuroImage* 268 (2023), p. 119843.

[1839] Yunchuan Kong and Tianwei Yu. “forgeNet: a graph deep neural network model using tree-based ensemble classifiers for feature graph construction”. In: *Bioinformatics* 36.11 (2020), pp. 3507–3515.

[1840] Madhumita and Sushmita Paul. "Capturing the latent space of an Autoencoder for multi-omics integration and cancer subtyping". In: *Computers in Biology and Medicine* 148 (2022), p. 105832.

[1841] Kim F Marquart et al. "Predicting base editing outcomes with an attention-based deep learning algorithm trained on high-throughput target library screens". In: *Nature Communications* 12.1 (2021), p. 5114.

[1842] Antonio Majdandzic, Chandana Rajesh, and Peter K Koo. "Correcting gradient-based interpretations of deep neural networks for genomics". In: *Genome Biology* 24.1 (2023), pp. 1–13.

[1843] Lingxi Chen, Jiao Xu, and Shuai Cheng Li. "DeepMF: Deciphering the latent patterns in omics profiles with a deep learning method". In: *BMC bioinformatics* 20.23 (2019), pp. 1–13.

[1844] Stephen J Fleming et al. "Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender". In: *Nature methods* 20.9 (2023), pp. 1323–1335.

[1845] Soroosh Gholamizoj and Bin Ma. "SPEQ: quality assessment of peptide tandem mass spectra with deep learning". In: *Bioinformatics* 38.6 (2022), pp. 1568–1574.

[1846] Robert W Link et al. "HIV-Quasipore: A Suite of HIV-1-Specific Nanopore Basecallers Designed to Enhance Viral Quasispecies Detection". In: *Frontiers in Virology* 2 (2022), p. 858375.

[1847] Luotong Wang et al. "NanoReviser: an error-correction tool for nanopore sequencing based on a deep learning algorithm". In: *Frontiers in Genetics* 11 (2020), p. 900.

[1848] Luotong Wang et al. "An Error Correction Method of Nanopore Sequencing Data Using Deep Learning". In: *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE. 2020, pp. 23–28.

[1849] Paul Anderson et al. "Reducing variability of breast cancer subtype predictors by grounding deep learning models in prior knowledge". In: *Computers in Biology and Medicine* 138 (2021), p. 104850.

[1850] Fengyao Yan et al. "Deep neural network based tissue deconvolution of circulating tumor cell RNA". In: *Journal of Translational Medicine* 21.1 (2023), p. 783.

[1851] Sophie Peacock, Etai Jacob, and Nikolay Burlutskiy. "Coupling deep imputation with multitask learning for downstream tasks on omics data". In: *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2022, pp. 1–10.

[1852] Jin-Hong Du, Zhanrui Cai, and Kathryn Roeder. "Robust probabilistic modeling for single-cell multimodal mosaic integration and imputation via scVAEIT". In: *Proceedings of the National Academy of Sciences* 119.49 (2022), e2214414119.

[1853] Tian Tian, Martin Renqiang Min, and Zhi Wei. "Model-based autoencoders for imputing discrete single-cell RNA-seq data". In: *Methods* 192 (2021), pp. 112–119.

[1854] Mary Qu Yang et al. "MISC: missing imputation for single-cell RNA sequencing data". In: *BMC systems biology* 12 (2018), pp. 55–63.

[1855] Zimo Huang et al. "scGGAN: single-cell RNA-seq imputation by graph-based generative adversarial network". In: *Briefings in bioinformatics* 24.2 (2023), bbad040.

[1856] Eric Chen et al. "GapPredict–A Language Model for Resolving Gaps in Draft Genome Assemblies". In: *IEEE/ACM transactions on computational biology and bioinformatics* 18.6 (2021), pp. 2802–2808.

[1857] Binhai Zhu et al. "Deep Learning Approaches for the Protein Scaffold Filling Problem". In: *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE. 2022, pp. 1055–1061.

[1858] Yuge Wang, Tianyu Liu, and Hongyu Zhao. "ResPAN: a powerful batch correction model for scRNA-seq data through residual adversarial networks". In: *Bioinformatics* 38.16 (2022), pp. 3942–3949.

[1859] Jingyang Niu et al. "Joint deep learning for batch effect removal and classification toward MALDI MS based metabolomics". In: *BMC bioinformatics* 23.1 (2022), pp. 1–19.

[1860] Yao-zhong Zhang et al. "Sequence-specific bias correction for RNA-seq data using recurrent neural networks". In: *BMC genomics* 18 (2017), pp. 1–6.

[1861] Steffen Albrecht et al. "seqQscorer: automated quality control of next-generation sequencing data using machine learning". In: *Genome biology* 22.1 (2021), pp. 1–20.

[1862] Xin Bai, Jie Ren, and Fengzhu Sun. "MLR-OOD: A Markov Chain Based Likelihood Ratio Method for Out-Of-Distribution Detection of Genomic Sequences". In: *Journal of Molecular Biology* 434.15 (2022), p. 167586.

[1863] Andrea Riba et al. "Cell cycle gene regulation dynamics revealed by RNA velocity and deep-learning". In: *Nature Communications* 13.1 (2022), p. 2865.

[1864] Tanglong Yuan et al. "Optimization of C-to-G base editors with sequence context preference predictable by machine learning methods". In: *Nature Communications* 12.1 (2021), p. 4902.

[1865] Myungjae Song et al. "Sequence-specific prediction of the efficiencies of adenine and cytosine base editors". In: *Nature biotechnology* 38.9 (2020), pp. 1037–1043.

[1866] Yang-Ming Lin, Ching-Tai Chen, and Jia-Ming Chang. "MS2CNN: predicting MS/MS spectrum based on protein sequence using deep convolutional neural networks". In: *BMC genomics* 20.9 (2019), pp. 1–10.

[1867] Huan Hu et al. “Modeling and analyzing single-cell multimodal data with deep parametric inference”. In: *Briefings in Bioinformatics* 24.1 (2023), bbad005.

[1868] You Wu, Qiao Liu, and Lei Xie. “Hierarchical multi-omics data integration and modeling predict cell-specific chemical proteomics and drug responses”. In: *Cell Reports Methods* 3.4 (2023).

[1869] Xi Xiang et al. “Enhancing CRISPR-Cas9 gRNA efficiency prediction by data integration and deep learning”. In: *Nature communications* 12.1 (2021), p. 3238.

[1870] Hari Shrawgi and Dilip Singh Sisodia. “Convolution neural network model for predicting single guide rna efficiency in crispr/cas9 system”. In: *Chemometrics and Intelligent Laboratory Systems* 189 (2019), pp. 149–154.

[1871] Samual MacDonald et al. “Generalising uncertainty improves accuracy and safety of deep learning analytics applied to oncology”. In: *Scientific Reports* 13.1 (2023), p. 7395.

[1872] Fabrizio Kuruc, Harald Binder, and Moritz Hess. “Stratified neural networks in a time-to-event setting”. In: *Briefings in Bioinformatics* 23.1 (2022), bbab392.

[1873] Conghao Wang et al. “Network-based integration of multi-omics data for clinical outcome prediction in neuroblastoma”. In: *Scientific Reports* 12.1 (2022), p. 15425.

[1874] Elika Chalupová et al. “ENNGene: an Easy Neural Network model building tool for Genomics”. In: *BMC genomics* 23.1 (2022), pp. 1–11.

[1875] Hüseyin Anil Gündüz et al. “A self-supervised deep learning method for data-efficient training in genomics”. In: *Communications Biology* 6.1 (2023), p. 928.

[1876] Lorna Richardson et al. “MGnify: the microbiome sequence data analysis resource in 2023”. In: *Nucleic Acids Research* 51.D1 (2023), pp. D753–D759.

[1877] Wolfgang Kopp et al. “Deep learning for genomics using Janggu”. In: *Nature communications* 11.1 (2020), p. 3488.

[1878] Etienne Routhier, Ayman Bin Kamruddin, and Julien Mozziconacci. “keras_dna: a wrapper for fast implementation of deep learning models in genomics”. In: *Bioinformatics* 37.11 (2021), pp. 1593–1594.

[1879] Tianze Cao et al. “plotnineSeqSuite: a Python package for visualizing sequence data using ggplot2 style”. In: *BMC genomics* 24.1 (2023), p. 585.

[1880] Daniel Quang, Yuanfang Guan, and Stephen CJ Parker. “YAMDA: thousandfold speedup of EM-based motif discovery using deep learning libraries and GPU”. In: *Bioinformatics* 34.20 (2018), pp. 3578–3580.

[1881] Magnus Haraldson Høie et al. “NetSurfP-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning”. In: *Nucleic acids research* 50.W1 (2022), W510–W515.

[1882] Tudor-Stefan Cotet et al. "ePlatypus: an ecosystem for computational analysis of immunogenomics data". In: *Bioinformatics* 39.9 (2023), btad553.

[1883] Chunyuan Qin et al. "Deep learning embedder method and tool for mass spectra similarity search". In: *Journal of proteomics* 232 (2021), p. 104070.

[1884] Dongfang Wang et al. "iMAP: integration of multiple single-cell datasets by adversarial paired transfer networks". In: *Genome biology* 22.1 (2021), pp. 1–24.

[1885] Tongjun Gu and Xiwu Zhao. "Integrating multi-platform genomic datasets for kidney renal clear cell carcinoma subtyping using stacked denoising autoencoders". In: *Scientific reports* 9.1 (2019), p. 16668.

[1886] Elena Sügis et al. "HENA, heterogeneous network-based data set for Alzheimers disease". In: *Scientific data* 6.1 (2019), p. 151.

[1887] Barbara R Terlouw et al. "MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters". In: *Nucleic acids research* 51.D1 (2023), pp. D603–D610.

[1888] Katarína Greová et al. "Genomic benchmarks: a collection of datasets for genomic sequence classification". In: *BMC Genomic Data* 24.1 (2023), p. 25.

[1889] William ZD Zeng et al. "Selecting precise reference normal tissue samples for cancer research using a deep learning approach". In: *BMC Medical Genomics* 12 (2019), pp. 179–189.

[1890] Amandeep Kaur, Ajay Pal Singh Chauhan, and Ashwani Kumar Aggarwal. "Dynamic deep genomics sequence encoder for managed file transfer". In: *IETE Journal of Research* (2022), pp. 1–13.

[1891] Mohammed Aledhari et al. "A deep learning-based data minimization algorithm for fast and secure transfer of big genomic datasets". In: *IEEE transactions on big data* 7.2 (2018), pp. 271–284.

[1892] Dohoon Lee and Sun Kim. "Knowledge-guided artificial intelligence technologies for decoding complex multiomics interactions in cells". In: *Clinical and Experimental Pediatrics* 65.5 (2022), p. 239.

[1893] Roman D Bülow et al. "Extending the landscape of omics technologies by pathomics". In: *npj Systems Biology and Applications* 9.1 (2023), p. 38.

[1894] Nida Nasir et al. "Review of Machine Learning Advancements for Single-Cell Analysis". In: *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE. 2023, pp. 1383–1387.

[1895] Pratik Dutta, Piyush Mishra, and Sriparna Saha. "Incomplete multi-view gene clustering with data regeneration using Shape Boltzmann Machine". In: *Computers in Biology and Medicine* 125 (2020), p. 103965.

[1896] Tribikram Dhar et al. "Challenges of Deep Learning in Medical Image AnalysisImproving Explainability and Trust". In: *IEEE Transactions on Technology and Society* 4.1 (2023), pp. 68–75.

[1897] Supriyo Chakraborty et al. "Interpretability of deep learning models: A survey of results". In: *2017 IEEE smartworld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, Internet of people and smart city innovation (smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI)*. IEEE. 2017, pp. 1–6.

[1898] Jian-Xun Mi, An-Di Li, and Li-Fang Zhou. "Review study of interpretation methods for future interpretable machine learning". In: *IEEE Access* 8 (2020), pp. 191969–191985.

[1899] Md Rezaul Karim et al. "Explainable ai for bioinformatics: Methods, tools and applications". In: *Briefings in bioinformatics* 24.5 (2023), bbad236.

[1900] Kevin Chew Figueroa et al. "Interpretable deep learning approach for oral cancer classification using guided attention inference network". In: *Journal of biomedical optics* 27.1 (2022), pp. 015001–015001.

[1901] Xia-an Bi et al. "Explainable and programmable hypergraph convolutional network for imaging genetics data fusion". In: *Information Fusion* 100 (2023), p. 101950.

[1902] Håvard Horgen Thunold et al. "A Deep Diagnostic Framework Using Explainable Artificial Intelligence and Clustering". In: *Diagnostics* 13.22 (2023), p. 3413.

[1903] Alicja Rczkowska et al. "Deep learning-based tumor microenvironment segmentation is predictive of tumor mutations and patient survival in non-small-cell lung cancer". In: *BMC cancer* 22.1 (2022), p. 1001.

[1904] Rezaul Karim et al. "Deepcovidexplainer: explainable COVID-19 diagnosis from chest X-ray images". In: *2020 IEEE international conference on bioinformatics and biomedicine (BIBM)*. IEEE. 2020, pp. 1034–1037.

[1905] João Moranguinho et al. "Attention based deep multiple instance learning approach for lung cancer prediction using histopathological images". In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE. 2021, pp. 2852–2855.

[1906] Sarah M Ryan et al. "Cluster activation mapping with application to computed tomography scans of the lung". In: *Journal of Medical Imaging* 9.2 (2022), pp. 026001–026001.

[1907] Gherman Novakovsky et al. "Obtaining genetics insights from Deep Learning via explainable Artificial Intelligence". In: *Nature Reviews Genetics* 24.2 (2023), pp. 125–137.

[1908] Sanghyuk Roy Choi and Minhyeok Lee. "Transformer architecture and attention mechanisms in genome data analysis: a comprehensive review". In: *Biology* 12.7 (2023), p. 1033.

[1909] Shuya Abe et al. "Explainable AI for Estimating Pathogenicity of Genetic Variants Using Large-Scale Knowledge Graphs". In: *Cancers* 15.4 (2023), p. 1118.

[1910] Adrien Badré and Chongle Pan. “LINA: A linearizing neural network architecture for accurate first-order and second-order interpretations”. In: *IEEE Access* 10 (2022), pp. 36166–36176.

[1911] Chandra Mohan Dasari and Raju Bhukya. “Explainable deep neural networks for novel viral genome prediction”. In: *Applied Intelligence* 52.3 (2022), pp. 3002–3017.

[1912] Jin Zhu et al. “Computational analysis of pathological image enables interpretable prediction for microsatellite instability”. In: *Frontiers in Oncology* 12 (2022), p. 825353.

[1913] Huan Yang et al. “Deep learning-based six-type classifier for lung cancer and mimics from histopathological whole slide images: a retrospective study”. In: *BMC medicine* 19 (2021), pp. 1–14.

[1914] Daiqiao Ai et al. “A multi-task FP-GNN framework enables accurate prediction of selective PARP inhibitors”. In: *Frontiers in Pharmacology* 13 (2022), p. 971369.

[1915] Matthias Mann et al. “Artificial intelligence for proteomics and biomarker discovery”. In: *Cell systems* 12.8 (2021), pp. 759–770.

[1916] Hryhorii Chereda et al. “Explaining decisions of graph convolutional neural networks: patient-specific molecular subnetworks responsible for metastasis prediction in breast cancer”. In: *Genome medicine* 13 (2021), pp. 1–16.

[1917] Ge Liu, Haoyang Zeng, and David K Gifford. “Visualizing complex feature interactions and feature sharing in genomic deep neural networks”. In: *BMC bioinformatics* 20.1 (2019), pp. 1–14.

[1918] Peter K Koo et al. “Global importance analysis: An interpretability method to quantify importance of genomic features in deep neural networks”. In: *PLoS computational biology* 17.5 (2021), e1008925.

[1919] Yi Liu, Kenneth Barr, and John Reinitz. “Fully interpretable deep learning model of transcriptional control”. In: *Bioinformatics* 36.Supplement_1 (2020), pp. i499–i507.

[1920] Gherman Novakovsky et al. “ExplaiNN: interpretable and transparent neural networks for genomics”. In: *Genome Biology* 24.1 (2023), p. 154.

[1921] Melvyn Yap et al. “Verifying explainability of a deep learning tissue classifier trained on RNA-seq data”. In: *Scientific reports* 11.1 (2021), p. 2641.

[1922] Ali Foroughi Pour et al. “Deep learning features encode interpretable morphologies within histological images”. In: *Scientific Reports* 12.1 (2022), p. 9428.

[1923] Katyna Sada Del Real and Angel Rubio. “Discovering the mechanism of action of drugs with a sparse explainable network”. In: *Ebiomedicine* 95 (2023).

[1924] Liviu Badea and Emil Stnescu. “Identifying transcriptomic correlates of histology using deep learning”. In: *PloS one* 15.11 (2020), e0242858.

[1925] Oscar A Zarate et al. "Boostmec: predicting crispr-cas9 cleavage efficiency through boosting models". In: *BMC bioinformatics* 23.1 (2022), pp. 1–14.

[1926] Anupama Jha et al. "Enhanced integrated gradients: improving interpretability of deep learning models using splicing codes as a case study". In: *Genome biology* 21 (2020), pp. 1–22.

[1927] Miguel Graça et al. "Interpreting High Order Epistasis Using Sparse Transformers". In: *Proceedings of the 8th ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies*. 2023, pp. 114–125.

[1928] Abdul Karim et al. "Molecular classification and interpretation of amyotrophic lateral sclerosis using deep convolution neural networks and shapley values". In: *Genes* 12.11 (2021), p. 1754.

[1929] Leonardo R Souza et al. "Using amino acids co-occurrence matrices and explainability model to investigate patterns in dengue virus proteins". In: *BMC bioinformatics* 23.1 (2022), pp. 1–19.

[1930] Sherlyn Jemimah, Aamna AlShehhi, and Alzheimers Disease Neuroimaging Initiative. "c-Diadem: a constrained dual-input deep learning model to identify novel biomarkers in Alzheimers disease". In: *BMC Medical Genomics* 16.Suppl 2 (2023), p. 244.

[1931] Santhosh Amilpur and Raju Bhukya. "EDeepSSP: Explainable deep neural networks for exact splice sites prediction". In: *Journal of Bioinformatics and Computational Biology* 18.04 (2020), p. 2050024.

[1932] Meiqin Gong et al. "Interpretable single-cell transcription factor prediction based on deep learning with attention mechanism". In: *Computational Biology and Chemistry* 106 (2023), p. 107923.

[1933] Li-Ming Xiao, Yun-Qi Wan, and Zhen-Ran Jiang. "AttCRISPR: a spacetime interpretable model for prediction of sgRNA on-target activity". In: *BMC bioinformatics* 22 (2021), pp. 1–17.

[1934] Guishan Zhang et al. "Prediction of CRISPR/Cas9 single guide RNA cleavage efficiency and specificity by attention-based convolutional neural networks". In: *Computational and structural biotechnology journal* 19 (2021), pp. 1445–1457.

[1935] Qingyuan Zheng et al. "Accurate diagnosis and survival prediction of bladder cancer using deep learning on histological slides". In: *Cancers* 14.23 (2022), p. 5807.

[1936] Zhenliang Li et al. "Deep multi-instance learning with induced self-attention for medical image classification". In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2020, pp. 446–450.

[1937] Takuma Kikutsuji et al. "Explaining reaction coordinates of alanine dipeptide isomerization obtained from deep neural networks using Explainable Artificial Intelligence (XAI)". In: *The journal of chemical physics* 156.15 (2022).

[1938] Nikolaus Fortelny and Christoph Bock. "Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data". In: *Genome biology* 21.1 (2020), pp. 1–36.

[1939] Kodai Minoura et al. "A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data". In: *Cell reports methods* 1.5 (2021).

[1940] Wenxing Hu et al. "Interpretable multimodal fusion networks reveal mechanisms of brain cognition". In: *IEEE transactions on medical imaging* 40.5 (2021), pp. 1474–1483.

[1941] Eloise Withnell et al. "XOmiVAE: an interpretable deep learning model for cancer classification using high-dimensional omics data". In: *Briefings in Bioinformatics* 22.6 (2021), bbab315.

[1942] Yuanyuan Liu et al. "Deep feature selection and causal analysis of Alzheimers disease". In: *Frontiers in neuroscience* 13 (2019), p. 1198.

[1943] Yujia Zhang et al. "Why Should You Trust My Explanation? Understanding Uncertainty in LIME Explanations". In: *arXiv preprint arXiv:1904.12991* (2019).

[1944] Krishna Mridha et al. "An Interpretable Skin Cancer Classification using Optimized Convolutional Neural Network for a smart Healthcare System". In: *IEEE Access* (2023).

[1945] Ian E Nielsen et al. "Robust Explainability: A Tutorial on Gradient-Based Attribution Methods for Deep Neural Networks". In: *arXiv preprint arXiv: 2107.11400* (2021).

[1946] Zohaib Salahuddin et al. "Transparency of deep neural networks for medical image analysis: A review of interpretability methods". In: *Computers in Biology and Medicine* 140 (2022), p. 105111.

[1947] Bas HM Van der Velden et al. "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis". In: *Medical Image Analysis* 79 (2022), p. 102470.

[1948] Yue Yang Teo, Artem Danilevsky, and Noam Shomron. "Overcoming interpretability in deep learning cancer classification". In: *Deep Sequencing Data Analysis* (2021), pp. 297–309.

[1949] W Richard McCombie, John D McPherson, and Elaine R Mardis. "Next-generation sequencing technologies". In: *Cold Spring Harbor perspectives in medicine* 9.11 (2019), a036798.

[1950] Taishan Hu et al. "Next-generation sequencing technologies: An overview". In: *Human Immunology* 82.11 (2021), pp. 801–811.

[1951] Charles Y Chiu and Steven A Miller. "Clinical metagenomics". In: *Nature Reviews Genetics* 20.6 (2019), pp. 341–355.

[1952] Iain M Wallace and Desmond G Higgins. "Evaluation of iterative alignment algorithms for multiple alignment". In: *Bioinformatics* 21.8 (2005), pp. 1408–1414.

[1953] Brian Gaschen et al. "Diversity considerations in HIV-1 vaccine selection". In: *Science* 296.5577 (2002), pp. 2354–2360.

[1954] Tanya Golubchik et al. "Mind the gaps: evidence of bias in estimates of multiple sequence alignments". In: *Molecular biology and evolution* 24.11 (2007), pp. 2433–2442.

[1955] Fatimah Noni Muhamad et al. "Performance Analysis of Needleman-Wunsch algorithm (global) and Smith-Waterman algorithm (local) in reducing search space and time for DNA sequence alignment". In: *Journal of Physics: Conference Series*. Vol. 1019. 1. IOP Publishing. 2018, p. 012085.

[1956] Isidro Cortés-Ciriano et al. "Computational analysis of cancer genome sequencing data". In: *Nature Reviews Genetics* 23.5 (2022), pp. 298–314.

[1957] INSDC International Nucleotide Sequence Database Collaboration. *SARS-CoV2 Database*. Accessed: 2023-05-23. 2023. URL: `https://covid19.sfb.uit.no/`.

[1958] Ting Wang et al. "The Human Pangenome Project: A global resource to map genomic diversity". In: *Nature* 604.7906 (2022), pp. 437–446.

[1959] Felipe Llinares-López et al. "Deep embedding and alignment of protein sequences". In: *Nature Methods* 20.1 (2023), pp. 104–111.

[1960] Hannah L Nicholls et al. "Reaching the end-game for GWAS: machine learning approaches for the prioritization of complex disease loci". In: *Frontiers in genetics* 11 (2020), p. 350.

[1961] Ana Guerrero-Tamayo and Fernando José Sadio-Ramos. *£Es posible saber Deep Learning sin ser ingeniero? Fundamentos y reflexiones sobre modelos generativos*. 2024.

[1962] Sean Whalen et al. "Navigating the pitfalls of applying machine learning in genomics". In: *Nature Reviews Genetics* 23.3 (2022), pp. 169–181.

[1963] Muhammad Arif et al. "StackACPred: Prediction of anticancer peptides by integrating optimized multiple feature descriptors with stacked ensemble approach". In: *Chemometrics and Intelligent Laboratory Systems* 220 (2022), p. 104458.

[1964] Xiaoyuan Li et al. "Exploring and validating the prognostic value of pathomics signatures and genomics in patients with cutaneous melanoma based on bioinformatics and deep learning". In: *Medical Physics* 50.11 (2023), pp. 7049–7059.

[1965] Ruifeng Hu et al. "Decoding regulatory structures and features from epigenomics profiles: A Roadmap-ENCODE Variational Auto-Encoder (RE-VAE) model". In: *Methods* 189 (2021), pp. 44–53.

[1966] S Remya and T Anjali. "An intelligent and optimal deep learning approach in sensor based networks for detecting microbes". In: *IEEE Sensors Journal* (2023).

[1967] Jie Peng et al. "Deep learning to estimate durable clinical benefit and prognosis from patients with non-small cell lung cancer treated with pd-1/Pd-L1 blockade". In: *Frontiers in Immunology* 13 (2022), p. 960459.

[1968] Eliana Marostica et al. "Development of a histopathology informatics pipeline for classification and prediction of clinical outcomes in subtypes of renal cell carcinoma". In: *Clinical Cancer Research* 27.10 (2021), pp. 2868–2878.

[1969] Pasquale Rescigno et al. "Characterizing CDK12-mutated prostate cancers". In: *Clinical Cancer Research* 27.2 (2021), pp. 566–574.

[1970] Wenxi Yu et al. "MSGA+ RF: A two-stage deep learning-based multi-scale guided attention models to diagnose idiopathic pulmonary fibrosis from CT images." In: *Medical Physics* (2022).

[1971] Wenyao Li et al. "Multi-Omics Data Integration Patient Classification Method Based on Deep Dense Residual Shrinkage Network". In: *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE. 2022, pp. 1098–1104.

[1972] Yaoliang Liu et al. "Machine learning framework for gut microbiome biomarkers discovery and modulation analysis in large-scale obese population". In: *BMC genomics* 23.1 (2022), pp. 1–14.

[1973] Yuhua Fu et al. "A gene prioritization method based on a swine multi-omics knowledgebase and a deep learning model". In: *Communications Biology* 3.1 (2020), p. 502.

[1974] Xin Niu et al. "Improved prediction of brain age using multimodal neuroimaging data". In: *Human brain mapping* 41.6 (2020), pp. 1626–1643.

[1975] Alexander Sasse et al. "Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings". In: *Nature Genetics* (2023), pp. 1–5.

[1976] Hamid Reza Hassanzadeh and May D Wang. "An integrated deep network for cancer survival prediction using omics data". In: *Frontiers in big Data* 4 (2021), p. 568352.

[1977] Ying Feng et al. "Pseudotargeted Metabolomic Fingerprinting and Deep Learning for Identification and Visualization of Common Pathogens". In: *Frontiers in Microbiology* 13 (2022), p. 830832.

[1978] Letizia Squarcina et al. "Deep learning for the prediction of treatment response in depression". In: *Journal of affective disorders* 281 (2021), pp. 618–622.

[1979] Pooya Mobadersany et al. "Predicting cancer outcomes from histology and genomics using convolutional networks". In: *Proceedings of the National Academy of Sciences* 115.13 (2018), E2970–E2979.

[1980] Xiaohan Xing et al. "Multi-level attention graph neural network based on co-expression gene modules for disease diagnosis and prognosis". In: *Bioinformatics* 38.8 (2022), pp. 2178–2186.

[1981] Huma Naz and Sachin Ahuja. "Deep learning approach for diabetes prediction using PIMA Indian dataset". In: *Journal of Diabetes & Metabolic Disorders* 19 (2020), pp. 391–403.

[1982] Tianle Ma and Aidong Zhang. "Integrate multi-omics data with biological interaction networks using Multi-view Factorization AutoEncoder (MAE)". In: *BMC genomics* 20 (2019), pp. 1–11.

[1983] Liping Song et al. "A deep fuzzy model for diagnosis of COVID-19 from CT images". In: *Applied Soft Computing* 122 (2022), p. 108883.

[1984] Shuangquan Zhang et al. "Assessing deep learning methods in cis-regulatory motif finding based on genomic sequencing data". In: *Briefings in Bioinformatics* 23.1 (2022), bbab374.

[1985] Amirreza Mahbod et al. "A dual decoder U-Net-based model for nuclei instance segmentation in hematoxylin and eosin-stained histological images". In: *Frontiers in Medicine* 9 (2022), p. 978146.

[1986] Derek Van Booven et al. "MP33-13 advancement of Automated Gleason Grading using Deep Learning integrated with genomics shows promise inprostate cancer". In: *Journal of Urology* 206.Supplement 3 (2021), e611.

[1987] Carlos Garcia-Perez et al. "Efficient detection of longitudinal bacteria fission using transfer learning in deep neural networks". In: *frontiers in microbiology* 12 (2021), p. 645972.

[1988] Jing Zhou et al. "Qualification of soybean responses to flooding stress using UAV-based imagery and deep learning". In: *Plant Phenomics* (2021).

[1989] Wei Liu et al. "Development and Validation of Multi-Omics Thymoma Risk Classification Model Based on Transfer Learning". In: *Journal of Digital Imaging* (2023), pp. 1–10.

[1990] Wenxi Yu et al. "An automatic diagnosis of idiopathic pulmonary fibrosis (IPF) using domain knowledge-guided attention models in HRCT images". In: *Medical Imaging 2021: Computer-Aided Diagnosis*. Vol. 11597. SPIE. 2021, pp. 458–463.

[1991] Wenxi Yu et al. "Mga-Net: Multi-Scale Guided Attention Models for an Automated Diagnosis of Idiopathic Pulmonary Fibrosis (IPF)". In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2021, pp. 1777–1780.

[1992] Chihyun Park, Jihwan Ha, and Sanghyun Park. "Prediction of Alzheimer's disease based on deep neural network by integrating gene expression and DNA methylation dataset". In: *Expert Systems with Applications* 140 (2020), p. 112873.

[1993] Nivedhitha Mahendran and Durai Raj Vincent PM. "Deep belief network-based approach for detecting Alzheimer's disease using the multi-omics data". In: *Computational and Structural Biotechnology Journal* 21 (2023), pp. 1651–1660.

[1994] Geoffrey D Hannigan et al. "A deep learning genome-mining strategy for biosynthetic gene cluster prediction". In: *Nucleic acids research* 47.18 (2019), e110–e110.

[1995] Tzung-Chien Hsieh et al. "PEDIA: prioritization of exome data by image analysis". In: *Genetics in Medicine* 21.12 (2019), pp. 2807–2814.

[1996] Camilla Nero et al. "Deep-learning to predict BRCA mutation and survival from digital h&e slides of epithelial ovarian cancer". In: *International Journal of Molecular Sciences* 23.19 (2022), p. 11326.

[1997] Bartosz Zieliski et al. "Deep learning approach to bacterial colony classification". In: *PloS one* 12.9 (2017), e0184554.

[1998] Okyaz Eminaga et al. "Artificial Intelligence Reveals Distinct Prognostic Subgroups of Muscle-Invasive Bladder Cancer on Histology Images". In: *Cancers* 15.20 (2023), p. 4998.

[1999] MB Sudhan et al. "Segmentation and classification of glaucoma using U-net with deep learning model". In: *Journal of Healthcare Engineering* 2022 (2022).

[2000] Qi Wang et al. "Deep learning-based brain transcriptomic signatures associated with the neuropathological and clinical severity of Alzheimers disease". In: *Brain Communications* 4.1 (2022), fcab293.

[2001] Donglin Wang et al. "Identifying ARG-carrying bacteriophages in a lake replenished by reclaimed water using deep learning techniques". In: *Water Research* 248 (2024), p. 120859.

[2002] Marco Camardo Leggieri, Marco Mazzoni, and Paola Battilani. "Machine learning for predicting mycotoxin occurrence in maize". In: *Frontiers in Microbiology* 12 (2021), p. 661132.

[2003] Siamak Salimy et al. "A deep learning-based framework for predicting survival-associated groups in colon cancer by integrating multi-omics and clinical data". In: *Heliyon* 9.7 (2023).

[2004] Zhi Huang et al. "SALMON: survival analysis learning with multi-omics neural networks on breast cancer". In: *Frontiers in genetics* 10 (2019), p. 166.

[2005] Omar Maddouri, Xiaoning Qian, and Byung-Jun Yoon. "Deep graph representations embed network information for robust disease marker identification". In: *Bioinformatics* 38.4 (2022), pp. 1075–1086.

[2006] Erkan Bostanci et al. "Machine learning analysis of RNA-seq data for diagnostic and prognostic prediction of colon cancer". In: *Sensors* 23.6 (2023), p. 3080.

[2007] Linmin Pei et al. “Deep neural network analysis of pathology images with integrated molecular data for enhanced glioma classification and grading”. In: *Frontiers in oncology* 11 (2021), p. 668694.

[2008] Rafael Vieira et al. “CHEIC: Chemical Image Classificator. An intelligent system for identification of volatiles compounds with potential for respiratory diseases using Deep Learning”. In: *Expert Systems with Applications* 234 (2023), p. 121178.

[2009] Hasan Zulfiqar et al. “Deep-4mCW2V: A sequence-based predictor to identify N4-methylcytosine sites in Escherichia coli”. In: *Methods* 203 (2022), pp. 558–563.

[2010] Zhen Cao and Shihua Zhang. “Probe efficient feature representation of gapped k-mer frequency vectors from sequences using deep neural networks”. In: *IEEE/ACM transactions on computational biology and bioinformatics* 17.2 (2018), pp. 657–667.

[2011] Alejandro Lopez-Rincon et al. “Classification and specific primer design for accurate detection of SARS-CoV-2 using deep learning”. In: *Scientific reports* 11.1 (2021), p. 947.

[2012] Ines P Nearchou et al. “Automated detection and classification of desmoplastic reaction at the colorectal tumour front using deep learning”. In: *Cancers* 13.7 (2021), p. 1615.

[2013] Yichuan Liu et al. “Deep learning prediction of attention-deficit hyperactivity disorder in African Americans by copy number variation”. In: *Experimental Biology and Medicine* 246.21 (2021), pp. 2317–2323.

[2014] Basel Elsayed et al. “Applications of Artificial Intelligence in Philadelphia-Negative Myeloproliferative Neoplasms”. In: *Diagnostics* 13.6 (2023), p. 1123.

[2015] Anahita Fathi Kazerooni et al. “Clinical measures, radiomics, and genomics offer synergistic value in AI-based prediction of overall survival in patients with glioblastoma”. In: *Scientific Reports* 12.1 (2022), p. 8784.

[2016] Diego Machado Reyes et al. “Genomics transformer for diagnosing Parkinson’s disease”. In: *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE. 2022, pp. 01–04.

[2017] Silvia Seidlitz et al. “Robust deep learning-based semantic organ segmentation in hyperspectral images”. In: *Medical Image Analysis* 80 (2022), p. 102488.

[2018] Abelardo Montesinos-López et al. “Multimodal deep learning methods enhance genomic prediction of wheat breeding”. In: *G3: Genes, Genomes, Genetics* 13.5 (2023), jkad045.

[2019] Mohsin Bilal et al. “Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study”. In: *The Lancet Digital Health* 3.12 (2021), e763–e772.

[2020] Van Bettauer et al. “A deep learning approach to capture the essence of Candida albicans morphologies”. In: *Microbiology Spectrum* 10.5 (2022), e01472–22.

[2021] Xiuying Liang et al. “StomataScorer: a portable and high-throughput leaf stomata trait scorer combined with deep learning and an improved CV model”. In: *Plant biotechnology journal* 20.3 (2022), pp. 577–591.

[2022] D Yang et al. “HISMD: A Novel Immune Subtyping System for HNSCC”. In: *Journal of Dental Research* 102.3 (2023), pp. 270–279.

[2023] Tzu-Hao Wang et al. “Biomarker identification through multiomics data analysis of prostate cancer prognostication using a deep learning model and similarity network fusion”. In: *Cancers* 13.11 (2021), p. 2528.

[2024] Lin Wang et al. “Integrative Serum Metabolic Fingerprints Based Multi-Modal Platforms for Lung Adenocarcinoma Early Detection and Pulmonary Nodule Classification”. In: *Advanced Science* 9.34 (2022), p. 2203786.

[2025] Sunan Cui, Randall K Ten Haken, and Issam El Naqa. “Integrating multiomics information in deep learning architectures for joint actuarial outcome prediction in non-small cell lung cancer patients after radiation therapy”. In: *International Journal of Radiation Oncology* Biology* Physics* 110.3 (2021), pp. 893–904.

[2026] Lihong Peng et al. “Predicting potential microbe-disease associations with graph attention autoencoder, positive-unlabeled learning, and deep neural network”. In: *Frontiers in Microbiology* 14 (2023).

[2027] Ying Chen et al. “Revealing novel pyroptosis-related therapeutic targets for sepsis based on machine learning”. In: *BMC Medical Genomics* 16.1 (2023), p. 23.

[2028] Md Abdur Rahaman et al. “Two-Dimensional Attentive Fusion for Multi-Modal Learning of Neuroimaging and Genomics Data”. In: *2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE. 2022, pp. 1–6.

[2029] Haoran Ma, Tin Wee Tan, and Kenneth Hon Kim Ban. “A multi-task CNN learning model for taxonomic assignment of human viruses”. In: *BMC bioinformatics* 22.6 (2021), pp. 1–15.

[2030] Castrense Savojardo et al. “DeepMito: accurate prediction of protein sub-mitochondrial localization using convolutional neural networks”. In: *Bioinformatics* 36.1 (2020), pp. 56–64.

[2031] Jiaojiao Zhao et al. “Deep learning radiomics model related with genomics phenotypes for lymph node metastasis prediction in colorectal cancer”. In: *Radiotherapy and Oncology* 167 (2022), pp. 195–202.

[2032] Dmitrii Bychkov et al. “Deep learning based tissue analysis predicts outcome in colorectal cancer”. In: *Scientific reports* 8.1 (2018), p. 3395.

[2033] Somayah Albaradei et al. "MetaCancer: A deep learning-based pan-cancer metastasis prediction model developed using multi-omics data". In: *Computational and Structural Biotechnology Journal* 19 (2021), pp. 4404–4411.

[2034] Ziwei Wei et al. "Deep learning-based multi-omics integration robustly predicts relapse in prostate cancer". In: *Frontiers in oncology* 12 (2022), p. 893424.

[2035] Titus J Brinker et al. "Deep learning approach to predict sentinel lymph node status directly from routine histology of primary melanoma tumours". In: *European Journal of Cancer* 154 (2021), pp. 227–234.

[2036] Jing Xu et al. "A hierarchical integration deep flexible neural forest framework for cancer subtype classification by integrating multi-omics data". In: *BMC bioinformatics* 20.1 (2019), pp. 1–11.

[2037] Huan Zhao et al. "Deep learning-based multi-omics study reveals the polymolecular phenotypic of diabetic kidney disease". In: *Clinical and Translational Medicine* 13.6 (2023).

[2038] Dibyendu Bikash Seal et al. "Estimating gene expression from DNA methylation and copy number variation: a deep learning regression model for multi-omics integration". In: *Genomics* 112.4 (2020), pp. 2833–2841.

[2039] Ghazal Mirabnahrazam et al. "Predicting time-to-conversion for dementia of Alzheimer's type using multi-modal deep survival analysis". In: *Neurobiology of Aging* 121 (2023), pp. 139–156.

[2040] Tzong-Yi Lee et al. "Incorporating deep learning and multi-omics autoencoding for analysis of lung adenocarcinoma prognostication". In: *Computational Biology and Chemistry* 87 (2020), p. 107277.

[2041] Simone Rampelli et al. "G2S: a new deep learning tool for predicting stool microbiome structure from oral microbiome data". In: *Frontiers in Genetics* 12 (2021), p. 644516.

[2042] Bing Feng et al. "Bi-stream CNN Down Syndrome screening model based on genotyping array". In: *BMC medical genomics* 11 (2018), pp. 25–33.

[2043] Naryttza N Diaz et al. "TACOA–Taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach". In: *BMC bioinformatics* 10 (2009), pp. 1–16.

[2044] Xiaoping Li et al. "Development and validation of a prognostic model for HER2-low breast cancer to evaluate neoadjuvant therapy". In: *Gland Surgery* 12.2 (2023), p. 183.

[2045] Alice R Owens et al. "Novel deep learning-based solution for identification of prognostic subgroups in liver cancer (Hepatocellular carcinoma)". In: *BMC bioinformatics* 22 (2021), pp. 1–22.

[2046] Ping Gong et al. "Multi-omics integration method based on attention deep learning network for biomedical data classification". In: *Computer Methods and Programs in Biomedicine* 231 (2023), p. 107377.

[2047] Hui Ying et al. “Using multiple indicators to predict the risk of surgical site infection after ORIF of tibia fractures: a machine learning based study”. In: *Frontiers in Cellular and Infection Microbiology* 13 (2023).

[2048] Jiaxin Hou et al. “Integrative Histology-Genomic Analysis Predicts Hepatocellular Carcinoma Prognosis Using Deep Learning”. In: *Genes* 13.10 (2022), p. 1770.

[2049] Narges Aghakhan Olia et al. “Deep learning–based denoising of low-dose SPECT myocardial perfusion images: quantitative assessment and clinical performance”. In: *European journal of nuclear medicine and molecular imaging* (2022), pp. 1–15.

[2050] DN Wickramaratne, CR Wijesinghe, and AR Weerasinghe. “A Deep Learning Approach to Predict Health Status Using Microbiome Profiling”. In: *2022 22nd International Conference on Advances in ICT for Emerging Regions (ICTer)*. IEEE. 2022, pp. 075–079.

[2051] Jing Li et al. “Predicting gastric cancer tumor mutational burden from histopathological images using multimodal deep learning”. In: *Briefings in Functional Genomics* (2023), elad032.

[2052] Saraswathi Duraisamy and Srinivasan Emperumal. “Computer-aided mammogram diagnosis system using deep learning convolutional fully complex-valued relaxation neural network classifier”. In: *IET Computer Vision* 11.8 (2017), pp. 656–662.

[2053] Mael Zuercher et al. “Retraining an artificial intelligence algorithm to calculate left ventricular ejection fraction in pediatrics”. In: *Journal of Cardiothoracic and Vascular Anesthesia* 36.9 (2022), pp. 3610–3616.

[2054] Fuguang Ji, Shuai Zhou, and Zhangshuan Bi. “Computed tomography angiography under deep learning in the treatment of atherosclerosis with rapamycin”. In: *Journal of Healthcare Engineering* 2021 (2021), pp. 1–8.

[2055] Emmanuel Adetiba et al. “DeepCOVID-19: A model for identification of COVID-19 virus sequences with genomic signal processing and deep learning”. In: *Cogent Engineering* 9.1 (2022), p. 2017580.

[2056] Shuai Li et al. “A novel radiogenomics framework for genomic and image feature correlation using deep learning”. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 899–906.

[2057] Xiangyu Liu et al. “Development of Prognostic Biomarkers by TMB-Guided WSI Analysis: A Two-Step Approach”. In: *IEEE Journal of Biomedical and Health Informatics* 27.4 (2023), pp. 1780–1789.

[2058] Cowan Ho et al. “A promising deep learning-assistive algorithm for histopathological screening of colorectal cancer”. In: *Scientific Reports* 12.1 (2022), p. 2222.

[2059] Gi Jeong Kim et al. "Efficient diagnosis of IDH-mutant gliomas: 1p/19qNET assesses 1p/19q codeletion status using weakly-supervised learning". In: *NPJ Precision Oncology* 7.1 (2023), p. 94.

[2060] Johannes Smolander, Matthias Dehmer, and Frank Emmert-Streib. "Comparing deep belief networks with support vector machines for classifying gene expression data from complex disorders". In: *FEBS Open Bio* 9.7 (2019), pp. 1232–1248.

[2061] Arthur O Frankel et al. "Machine learning for rhabdomyosarcoma histopathology". In: *Modern Pathology* 35.9 (2022), pp. 1193–1203.

[2062] Can Liu et al. "A classification method of gastric cancer subtype based on residual graph convolution network". In: *Frontiers in Genetics* 13 (2023), p. 1090394.

[2063] Diego Fioravanti et al. "Phylogenetic convolutional neural networks in metagenomics". In: *BMC bioinformatics* 19 (2018), pp. 1–13.

[2064] Disha Sushant Wankhede and R Selvarani. "Dynamic architecture based deep learning approach for glioblastoma brain tumor survival prediction". In: *Neuroscience Informatics* 2.4 (2022), p. 100062.

[2065] Cecile M Cres et al. "DL-TODA: A Deep Learning Tool for Omics Data Analysis". In: *Biomolecules* 13.4 (2023), p. 585.

[2066] Stefan Schulz et al. "Multimodal deep learning for prognosis prediction in renal cancer". In: *Frontiers in oncology* 11 (2021), p. 788740.

[2067] Shaohua Xu and Huicheng Yuan. "A three-methylation-driven gene–based deep learning model for tuberculosis diagnosis in patients with and without human immunodeficiency virus co-infection". In: *Microbiology and Immunology* 66.6 (2022), pp. 317–323.

[2068] Amin ul Haq et al. "DACBT: Deep learning approach for classification of brain tumors using MRI data in IoT healthcare environment". In: *Scientific Reports* 12.1 (2022), p. 15331.

[2069] Daniele Raimondi et al. "Genome interpretation in a federated learning context allows the multi-center exome-based risk prediction of Crohns disease patients". In: *Scientific Reports* 13.1 (2023), p. 19449.

[2070] Naofumi Tomita et al. "Predicting oncogene mutations of lung cancer using deep learning and histopathologic features on whole-slide images". In: *Translational Oncology* 24 (2022), p. 101494.

[2071] Reetesh K Pai et al. "Development and initial validation of a deep learning algorithm to quantify histological features in colorectal carcinoma including tumour budding/poorly differentiated clusters". In: *Histopathology* 79.3 (2021), pp. 391–405.

[2072] Safaa M Naeem et al. "A diagnostic genomic signal processing (GSP)-based system for automatic feature analysis and detection of COVID-19". In: *Briefings in Bioinformatics* 22.2 (2021), pp. 1197–1205.

[2073] Jonathan Thompson et al. “Diagnostic Screening of Bovine Mastitis Using MALDI-TOF MS Direct-Spotting of Milk and Machine Learning”. In: *Veterinary Sciences* 10.2 (2023), p. 101.

[2074] Chao Lv et al. “Prediction of 3-year recurrence rate of hepatocellular carcinoma after resection based on contrast-enhanced CT: A single-centre study”. In: *The British Journal of Radiology* 96.1145 (2023), p. 20220702.

[2075] Turki Turki and YH Taguchi. “Discriminating the single-cell gene regulatory networks of human pancreatic islets: A novel deep learning application”. In: *Computers in Biology and Medicine* 132 (2021), p. 104257.

[2076] Karthik Devarajan. “A statistical framework for non-negative matrix factorization based on generalized dual divergence”. In: *Neural Networks* 140 (2021), pp. 309–324.

[2077] Daksh Chauhan et al. “Comparison of machine learning and deep learning for view identification from cardiac magnetic resonance images”. In: *Clinical Imaging* 82 (2022), pp. 121–126.

[2078] Abdullah Alsulaiman et al. “Profiling of pharmacogenomic variants in CYP2D6 and DPYD in indigenous Arab breast cancer patients”. In: *Pharmacogenomics* 24.7 (2023), pp. 411–423.

[2079] Frederik Wessels et al. “Deep learning approach to predict lymph node metastasis directly from primary tumour histology in prostate cancer”. In: *BJU international* 128.3 (2021), pp. 352–360.

[2080] Liang Bin et al. “A deep learning-based dual-omics prediction model for radiation pneumonitis”. In: *Medical physics* 48.10 (2021), pp. 6247–6256.

[2081] Amir Baghdadi et al. “Automated differentiation of benign renal oncocytoma and chromophobe renal cell carcinoma on computed tomography using deep learning”. In: *BJU international* 125.4 (2020), pp. 553–560.

[2082] Eleftherios Trivizakis et al. “Deep radiotranscriptomics of non-small cell lung carcinoma for assessing molecular and histology subtypes with a data-driven analysis”. In: *Diagnostics* 11.12 (2021), p. 2383.

[2083] Edward B James et al. “SymbiQuant: a machine learning object detection tool for polyploid independent estimates of endosymbiont population size”. In: *Frontiers in microbiology* 13 (2022), p. 816608.

[2084] Javier Juan-Albarracín et al. “ONCOhabitats: A system for glioblastoma heterogeneity assessment through MRI”. In: *International journal of medical informatics* 128 (2019), pp. 53–61.

[2085] David Unnersjö-Jess et al. “Deep learning–based segmentation and quantification of podocyte foot process morphology suggests differential patterns of foot process effacement across kidney pathologies”. In: *Kidney International* 103.6 (2023), pp. 1120–1130.

[2086] Chen Qiao et al. “The effective diagnosis of schizophrenia by using multilayer RBMs deep networks”. In: *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2015, pp. 603–606.

[2087] Gun Woo Lee, Hyunkwang Shin, and Min Cheol Chang. "Deep learning algorithm to evaluate cervical spondylotic myelopathy using lateral cervical spine radiograph". In: *BMC neurology* 22.1 (2022), p. 147.

[2088] Mohamed Yacin Sikkandar. "Automatic detection of genetics and genomics of eye disease using deep assimilation learning algorithm". In: *Interdisciplinary Sciences: Computational Life Sciences* 13 (2021), pp. 286–298.

[2089] Sara Sheehan and Yun S Song. "Deep learning for population genetic inference". In: *PLoS computational biology* 12.3 (2016), e1004845.

[2090] Jielu Yan et al. "Deep-AmPEP30: improve short antimicrobial peptides prediction with deep learning". In: *Molecular Therapy-Nucleic Acids* 20 (2020), pp. 882–894.

[2091] Ayush Kanyal et al. "Multi-modal deep learning on imaging genetics for schizophrenia classification". In: *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*. IEEE. 2023, pp. 1–5.

[2092] Susann Stephan-Falkenau et al. "Integrated Clinical, Molecular and Immunological Characterization of Pulmonary Sarcomatoid Carcinomas Reveals an Immune Escape Mechanism That May Influence Therapeutic Strategies". In: *International Journal of Molecular Sciences* 24.13 (2023), p. 10558.

[2093] Jan Oldenburg et al. "Methodology for comprehensive cell-level analysis of wound healing experiments using deep learning in MATLAB". In: *BMC Molecular and Cell Biology* 22.1 (2021), p. 32.

[2094] Seokmin Han, Sung Il Hwang, and Hak Jong Lee. "The classification of renal cancer in 3-phase CT images using a deep learning method". In: *Journal of digital imaging* 32 (2019), pp. 638–643.

[2095] Vidit Kumar et al. "Hybrid spatiotemporal contrastive representation learning for content-based surgical video retrieval". In: *Electronics* 11.9 (2022), p. 1353.

[2096] Ling Jia et al. "AI-Blue-Carba: a rapid and improved carbapenemase producer detection assay using Blue-Carba with deep learning". In: *Frontiers in Microbiology* 11 (2020), p. 585417.

[2097] Zixiao Lu et al. "BrcaSeg: a deep learning approach for tissue quantification and genomic correlations of histopathological images". In: *Genomics, proteomics & bioinformatics* 19.6 (2021), pp. 1032–1042.

[2098] Xiaoyong Pan et al. "RBPsuite: RNA-protein binding sites prediction suite based on deep learning". In: *BMC genomics* 21 (2020), pp. 1–8.

[2099] Taro Sakamoto et al. "A collaborative workflow between pathologists and deep learning for the evaluation of tumour cellularity in lung adenocarcinoma". In: *Histopathology* 81.6 (2022), pp. 758–769.

[2100] Jie Ju et al. "Robust deep learning model for prognostic stratification of pancreatic ductal adenocarcinoma patients". In: *Iscience* 24.12 (2021).

[2101] Yiming Song. "CT Radio Genomics of Non-Small Cell Lung Cancer Using Machine and Deep Learning". In: *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)*. IEEE. 2021, pp. 128–139.

[2102] Wei Wang et al. "A novel deep learning model for glioma epilepsy associated with the identification of human cytomegalovirus infection injuries based on head MR". In: *Frontiers in Microbiology* 14 (2023).

[2103] Stefano Trebeschi et al. "Deep learning for fully-automated localization and segmentation of rectal cancer on multiparametric MR". In: *Scientific reports* 7.1 (2017), p. 5301.

[2104] Chi Wang et al. "Automatic liver segmentation using multi-plane integrated fully convolutional neural networks". In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 1–6.

[2105] iga Avsec et al. "Modeling positional effects of regulatory sequences with spline transformations increases prediction accuracy of deep neural networks". In: *Bioinformatics* 34.8 (2018), pp. 1261–1269.

[2106] Ritaban Kirtania et al. "Deepsgp: Deep learning for gene selection and survival group prediction in glioblastoma". In: *Electronics* 10.12 (2021), p. 1463.

[2107] Mahendra Khened et al. "3D convolution neural networks for molecular subtype prediction in glioblastoma multiforme". In: *Medical Imaging 2019: Imaging Informatics for Healthcare, Research, and Applications*. Vol. 10954. SPIE. 2019, pp. 256–262.

[2108] Ruchika Verma et al. "Deep learning-based integration of esophageal cancer morphology with genomics". In: *Cancer Research* 81.13_Supplement (2021), pp. 3–3.

[2109] Avnish Panwar et al. "Stratification of the lesions in color fundus images of diabetic retinopathy patients using deep learning models and machine learning classifiers". In: *Edge Analytics: Select Proceedings of 26th International ConferenceADCOM 2020*. Springer. 2022, pp. 653–666.

[2110] Jing-Wen Lyu et al. "Rapid prediction of multidrug-resistant klebsiella pneumoniae through deep learning analysis of sers spectra". In: *Microbiology Spectrum* 11.2 (2023), e04126–22.

[2111] Nam Nhut Phan et al. "High-performance deep learning pipeline predicts individuals in mixtures of DNA using sequencing data". In: *Briefings in Bioinformatics* 22.6 (2021), bbab283.

[2112] Shao-Wu Zhang, Xiang-Yang Jin, Teng Zhang, et al. "Gene prediction in metagenomic fragments with deep learning". In: *BioMed research international* 2017 (2017).

[2113] Hiroki Kiyokawa et al. “Deep learning analysis of histologic images from intestinal specimen reveals adipocyte shrinkage and mast cell infiltration to predict postoperative Crohn disease”. In: *The American journal of pathology* 192.6 (2022), pp. 904–916.

[2114] Wei Chen et al. “Integrating radiomics with genomics for non-small cell lung cancer survival analysis”. In: *Journal of Oncology* 2022 (2022).

[2115] Ahmad Algohary et al. “Uncovering prostate cancer aggressiveness signal in T2-weighted MRI through a three-reference tissues normalization technique”. In: *NMR in Biomedicine* (2023), e5069.

[2116] Daniel Gonzalez, Robin L Dietz, and Liron Pantanowitz. “Feasibility of a deep learning algorithm to distinguish large cell neuroendocrine from small cell lung carcinoma in cytology specimens”. In: *Cytopathology* 31.5 (2020), pp. 426–431.

[2117] Martin Sturm et al. “TargetSpy: a supervised machine learning approach for microRNA target prediction”. In: *BMC bioinformatics* 11 (2010), pp. 1–17.

[2118] Peilin Jia, Ruifeng Hu, and Zhongming Zhao. “Benchmark of embedding-based methods for accurate and transferable prediction of drug response”. In: *Briefings in Bioinformatics* 24.3 (2023), bbad098.

[2119] Gil Pinheiro et al. “Deep homography based localization on videos of endoscopic capsules”. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 724–727.

[2120] Ke Li et al. “Association of radiomic imaging features and gene expression profile as prognostic factors in pancreatic ductal adenocarcinoma”. In: *American journal of translational research* 11.7 (2019), p. 4491.

[2121] Mateusz Buda et al. “Deep radiogenomics of lower-grade gliomas: Convolutional neural networks predict tumor genomic subtypes using mr images”. In: *Radiology: Artificial Intelligence* 2.1 (2020), e180050.

[2122] Jong Ho Jhee et al. “Transformer-Based Gene Scoring Model for Extracting Representative Characteristic of Central Dogma Process to Prioritize Pathogenic Genes Applying Breast Cancer Multi-omics Data”. In: *2023 IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE. 2023, pp. 149–154.

[2123] M Kalaiyarasi and Harikumar Rajaguru. “Performance Analysis of Ovarian Cancer Detection and Classification for Microarray Gene Data”. In: *BioMed Research International* 2022 (2022).

[2124] Rohita Sinha and Jennifer Clarke. “When technology meets technology: Retrained "Inception V3" classifier for NGS based pathogen detection”. In: *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2017, pp. 1–5.

[2125] Indrani Bhattacharya et al. “Selective identification and localization of indolent and aggressive prostate cancers via CorrSigNIA: an MRI-pathology correlation and deep learning framework”. In: *Medical Image Analysis* 75 (2022), p. 102288.

[2126] Meixuan Wu et al. “Exploring prognostic indicators in the pathological images of ovarian cancer based on a deep survival network”. In: *Frontiers in Genetics* 13 (2023), p. 1069673.

[2127] Kazuma Sakoda et al. “Genetic diversity in stomatal density among soybeans elucidated using high-throughput technique based on an algorithm for object detection”. In: *Scientific reports* 9.1 (2019), p. 7610.

[2128] Zhaohui Liang et al. “DL-ADR: a novel deep learning model for classifying genomic variants into adverse drug reactions”. In: *BMC medical genomics* 9.2 (2016), pp. 195–204.

[2129] Ala Saleh Alluhaidan. “DNA Sequence Analysis for Brain Disorder Using Deep Learning and Secure Storage.” In: *Computers, Materials & Continua* 71.3 (2022).

[2130] Moozhan Nikpanah et al. “A deep-learning based artificial intelligence (AI) approach for differentiation of clear cell renal cell carcinoma from oncocytoma on multi-phasic MRI”. In: *Clinical imaging* 77 (2021), pp. 291–298.

[2131] Yanmin Chen et al. “Instance Segmentation and Number Counting of Grape Berry Images Based on Deep Learning”. In: *Applied Sciences* 13.11 (2023), p. 6751.

[2132] Carmelo Pino et al. “Interpretable deep model for predicting gene-addicted non-small-cell lung cancer in CT scans”. In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2021, pp. 891–894.

[2133] Pradeep Kumar Mallick et al. “Convergent learning–based model for leukemia classification from gene expression”. In: *Personal and Ubiquitous Computing* 27.3 (2023), pp. 1103–1110.

[2134] Raphael Carapito et al. “Identification of driver genes for critical forms of COVID-19 in a deeply phenotyped young patient cohort”. In: *Science translational medicine* 14.628 (2022), eabj7521.

[2135] Su-Jhen Hung et al. “Assessment of the risk of severe dengue using intrahost viral population in dengue virus serotype 2 patients via machine learning”. In: *Frontiers in Cellular and Infection Microbiology* 12 (2022), p. 831281.

[2136] Xiaoxiao Wang et al. “Prediction of BRCA gene mutation in breast cancer based on deep learning and histopathology images”. In: *Frontiers in Genetics* 12 (2021), p. 661109.

[2137] Amina Lemsara, Salima Ouadfel, and Holger Fröhlich. “PathME: pathway based multi-modal sparse autoencoders for clustering of patient-level multi-omics data”. In: *BMC bioinformatics* 21 (2020), pp. 1–20.

[2138] Takahiro Ueda et al. "Deep learning reconstruction of diffusion-weighted MRI improves image quality for prostatic imaging". In: *Radiology* 303.2 (2022), pp. 373–381.

[2139] Dan Wu et al. "Panicle-3D: A low-cost 3D-modeling method for rice panicles based on deep learning, shape from silhouette, and supervoxel clustering". In: *The Crop Journal* 10.5 (2022), pp. 1386–1398.

[2140] Ding Li et al. "Relationship between the deep features of the full-scan pathological map of mucinous gastric carcinoma and related genes based on deep learning". In: *Heliyon* 9.3 (2023).

[2141] Tahsin Kurc et al. "Segmentation and classification in digital pathology for glioma research: challenges and deep learning approaches". In: *Frontiers in neuroscience* 14 (2020), p. 27.

[2142] Jianpeng Yuan et al. "Predictive effect of computed tomography imaging omics features under deep learning on metastatic lymph nodes of nasopharyngeal carcinoma". In: *Expert Systems* 39.3 (2022), e12860.

[2143] Pratik Jagtap et al. "Deep metaproteomic analysis of human salivary supernatant". In: *Proteomics* 12.7 (2012), pp. 992–1001.

[2144] Yue Zhang et al. "Application of deep learning method on ischemic stroke lesion segmentation". In: *Journal of Shanghai Jiaotong University (Science)* (2022), pp. 1–13.

[2145] Francesco Mercaldo et al. "Coronavirus COVID-19 detection by means of explainable deep learning". In: *Scientific Reports* 13.1 (2023), p. 462.

[2146] Carlos Fernandes et al. "Artificial neural networks classification of patients with Parkinsonism based on gait". In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 2024–2030.

[2147] Daichi Kitaguchi et al. "Deep learning-based automatic surgical step recognition in intraoperative videos for transanal total mesorectal excision". In: *Surgical Endoscopy* (2022), pp. 1–9.

[2148] Siqi Gao et al. "TRAmHap: accurate prediction of transcriptional activity from DNA methylation haplotypes in bisulfite-sequencing data". In: *Briefings in Bioinformatics* (2023), bbad214.

[2149] Hanyu Luo et al. "SENet: a deep learning framework for discriminating super-and typical enhancers by sequence information". In: *Computational Biology and Chemistry* (2023), p. 107905.

[2150] Ray O Bahado-Singh et al. "Artificial intelligence and amniotic fluid multiomics: prediction of perinatal outcome in asymptomatic women with short cervix". In: *Ultrasound in Obstetrics & Gynecology* 54.1 (2019), pp. 110–118.

[2151] Alberto Signoroni et al. "Hierarchical AI enables global interpretation of culture plates in the era of digital microbiology". In: *Nature Communications* 14.1 (2023), p. 6874.

[2152] Maria Carlota Dao et al. “A data integration multi-omics approach to study calorie restriction-induced changes in insulin sensitivity”. In: *Frontiers in physiology* 9 (2019), p. 1958.

[2153] Ghazala Khan and Naimul Mefraz Khan. “Brain MRI Segmentation using efficient 3D Fully Convolutional Neural Networks”. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 2351–2356.

[2154] Azeem Ahmad et al. “Highly sensitive quantitative phase microscopy and deep learning aided with whole genome sequencing for rapid detection of infection and antimicrobial resistance”. In: *Frontiers in Microbiology* 14 (2023), p. 1154620.

[2155] Lei Cai, Yufeng Wu, and Jingyang Gao. “DeepSV: accurate calling of genomic deletions from high-throughput sequencing data using deep convolutional neural network”. In: *BMC bioinformatics* 20.1 (2019), pp. 1–17.

[2156] Ardi Tampuu et al. “ViraMiner: Deep learning on raw DNA sequences for identifying viral genomes in human samples”. In: *PloS one* 14.9 (2019), e0222271.

[2157] Xue-han Wang et al. “EEG emotion recognition using dynamical graph convolutional neural networks and broad learning system”. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. 2018, pp. 1240–1244.

[2158] Sanchita Sil et al. “Raman spectroscopy and artificial intelligence open up accurate detection of pathogens from DNA-based sub-species level classification”. In: *Journal of Raman Spectroscopy* 52.12 (2021), pp. 2648–2659.

[2159] Kaya Turgut, Helin Dutagaci, and David Rousseau. “RoseSegNet: An attention-based deep learning architecture for organ segmentation of plants”. In: *Biosystems Engineering* 221 (2022), pp. 138–153.

[2160] Georgios Asimomitis et al. “Predicting single cell genotypes from single cell expression profiles in AML using deep learning”. In: *Proceedings of the 2023 13th International Conference on Bioscience, Biochemistry and Bioinformatics*. 2023, pp. 1–9.

[2161] Saumya Singh et al. “Culture-Independent Raman Spectroscopic Identification of Bacterial Pathogens from Clinical Samples Using Deep Transfer Learning”. In: *Analytical Chemistry* 94.42 (2022), pp. 14745–14754.

[2162] Bihter Das and Suat Toraman. “Deep transfer learning for automated liver cancer gene recognition using spectrogram images of digitized DNA sequences”. In: *Biomedical Signal Processing and Control* 72 (2022), p. 103317.

[2163] Ricardo Lemes Gonçalves et al. “Integrative transcriptome analysis of SARS-CoV-2 human-infected cells combined with deep learning algorithms identifies two potential cellular targets for the treatment of coronavirus disease”. In: *Brazilian Journal of Microbiology* 54.1 (2023), pp. 53–68.

[2164] Yonghua Zhuang et al. “Deep learning on graphs for multi-omics classification of COPD”. In: *Plos one* 18.4 (2023), e0284563.

[2165] Meihong Gao and Xuequn Shang. “Identification of associations between lncRNA and drug resistance based on deep learning and attention mechanism”. In: *Frontiers in Microbiology* 14 (2023), p. 1147778.

[2166] Tatsuhiko Naito. “Predicting the impact of single nucleotide variants on splicing via sequence-based deep neural networks and genomic features”. In: *Human mutation* 40.9 (2019), pp. 1261–1269.

[2167] Ejay Nsugbe and Stephanie Connelly. “Multiscale depth of anaesthesia prediction for surgery using frontal cortex electroencephalography”. In: *Healthcare Technology Letters* 9.3 (2022), pp. 43–53.

[2168] Zixu Yuan et al. “Development and validation of an image-based deep learning algorithm for detection of synchronous peritoneal carcinomatosis in colorectal cancer”. In: *Annals of surgery* 275.4 (2022), e645–e651.

[2169] Bifang He et al. “PDL1Binder: Identifying programmed cell death ligand 1 binding peptides by incorporating next-generation phage display data and different peptide descriptors”. In: *Frontiers in Microbiology* 13 (2022), p. 928774.

[2170] Tafadzwa L Chaunzwa et al. “Deep learning classification of lung cancer histology using CT images”. In: *Scientific reports* 11.1 (2021), p. 5471.

[2171] Phillip Chlap et al. “A review of medical image data augmentation techniques for deep learning applications”. In: *Journal of Medical Imaging and Radiation Oncology* 65.5 (2021), pp. 545–563.

[2172] Kavitha Shaga Devan et al. “Improved automatic detection of herpesvirus secondary envelopment stages in electron microscopy by augmenting training data with synthetic labelled images generated by a generative adversarial network”. In: *Cellular Microbiology* 23.2 (2021), e13280.

[2173] Yuanwei Xu et al. “CACONET: a novel classification framework for microbial correlation networks”. In: *Bioinformatics* 38.6 (2022), pp. 1639–1647.

[2174] Yu Zhang et al. “DL-CRISPR: a deep learning method for off-target activity prediction in CRISPR/Cas9 with data augmentation”. In: *IEEE Access* 8 (2020), pp. 76610–76617.

[2175] Alexandra Danyi, Myrthe Jager, and Jeroen de Ridder. “Cancer type classification in liquid biopsies based on sparse mutational profiles enabled through data augmentation and integration”. In: *Life* 12.1 (2021), p. 1.

[2176] Supratim Das and Xinghua Shi. “Offspring GAN augments biased human genomic data”. In: *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. 2022, pp. 1–10.

[2177] Stefan Lenz, Moritz Hess, and Harald Binder. “Deep generative models in DataSHIELD”. In: *BMC Medical Research Methodology* 21.1 (2021), pp. 1–16.

[2178] Jens NuSSberger et al. “Synthetic observations from deep generative models and binary omics data with limited sample size”. In: *Briefings in Bioinformatics* 22.4 (2021), bbaa226.

[2179] Shinji Kanazawa et al. “Fake metabolomics chromatogram generation for facilitating deep learning of peak-picking neural networks”. In: *Journal of bioscience and bioengineering* 131.2 (2021), pp. 207–212.

[2180] Bahareh Jahanyar, Hamid Tabatabaee, and Alireza Rowhanimanesh. “MS-ACGAN: A modified auxiliary classifier generative adversarial network for schizophrenia’s samples augmentation based on microarray gene expression data”. In: *Computers in Biology and Medicine* 162 (2023), p. 107024.

[2181] Weifeng Zhang et al. “Compound Raman microscopy for rapid diagnosis and antimicrobial susceptibility testing of pathogenic bacteria in urine”. In: *Frontiers in Microbiology* 13 (2022), p. 874966.

[2182] Fan Wang et al. “Deep learning identifies genome-wide DNA binding sites of long noncoding RNAs”. In: *RNA biology* 15.12 (2018), pp. 1468–1476.

[2183] Hongda Bu et al. “A new method for enhancer prediction based on deep belief network”. In: *BMC bioinformatics* 18 (2017), pp. 99–105.

[2184] Martin Treppner et al. “Synthetic single cell rna sequencing data from small pilot studies using deep generative models”. In: *Scientific reports* 11.1 (2021), p. 9403.

[2185] Francisco Carrillo-Perez et al. “Synthetic whole-slide image tile generation with gene expression profile-infused deep generative models”. In: *Cell Reports Methods* 3.8 (2023).

[2186] Junjie Chen, Mohammad Erfan Mowlaei, and Xinghua Shi. “Population-scale Genomic Data Augmentation Based on Conditional Generative Adversarial Networks”. In: *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. 2020, pp. 1–6.

[2187] Xue Jiang et al. “A generative adversarial network model for disease gene prediction with RNA-seq data”. In: *IEEE Access* 8 (2020), pp. 37352–37360.

[2188] Nicholas Keone Lee et al. “EvoAug: improving generalization and interpretability of genomic deep neural networks with evolution-inspired data augmentations”. In: *Genome Biology* 24.1 (2023), p. 105.

[2189] Jan Zrimec et al. “Controlling gene expression with deep generative design of regulatory DNA”. In: *Nature communications* 13.1 (2022), p. 5099.

[2190] Chi Xiao et al. “Automatic SARS-CoV-2 segmentation in electron microscopy based on few-shot learning”. In: *International Journal of Wavelets, Multiresolution and Information Processing* 22.02 (2024), p. 2350047.

[2191] William Jones et al. “Computational biology: deep learning”. In: *Emerging Topics in Life Sciences* 1.3 (2017), pp. 257–274.

[2192] Muhammad Toseef et al. "Deep transfer learning for clinical decision-making based on high-throughput data: comprehensive survey with benchmark results". In: *Briefings in Bioinformatics* 24.4 (2023), bbad254.

[2193] Walia Farzana et al. "Radiogenomic Prediction of MGMT Using Deep Learning with Bayesian Optimized Hyperparameters". In: *International MICCAI Brainlesion Workshop*. Springer. 2021, pp. 357–366.

[2194] Meenal Chaudhari et al. "DeepRMethylSite: a deep learning based approach for prediction of arginine methylation sites in proteins". In: *Molecular omics* 16.5 (2020), pp. 448–454.

[2195] Victoria Bourgeais et al. "Deep GONet: self-explainable deep neural network based on Gene Ontology for phenotype prediction from gene expression data". In: *BMC bioinformatics* 22.10 (2021), pp. 1–25.

[2196] Bertil Schmidt and Andreas Hildebrandt. "Deep learning in next-generation sequencing". In: *Drug discovery today* 26.1 (2021), pp. 173–180.

[2197] Chayakrit Krittanawong et al. "Artificial intelligence and cardiovascular genetics". In: *Life* 12.2 (2022), p. 279.

[2198] Eugene Lin, Chieh-Hsin Lin, and Hsien-Yuan Lane. "Machine learning and deep learning for the pharmacogenomics of antidepressant treatments". In: *Clinical Psychopharmacology and Neuroscience* 19.4 (2021), p. 557.

[2199] Sheng Wang et al. "The application of deep learning in biomedical informatics". In: *2018 International Conference on Robots & Intelligent System (ICRIS)*. IEEE. 2018, pp. 391–394.

[2200] Sijie Yang et al. "Intelligent health care: Applications of deep learning in computational medicine". In: *Frontiers in Genetics* 12 (2021), p. 607471.

[2201] Noam Auslander, Ayal B Gussow, and Eugene V Koonin. "Incorporating machine learning into established bioinformatics frameworks". In: *International journal of molecular sciences* 22.6 (2021), p. 2903.

[2202] Binhua Tang et al. "Recent advances of deep learning in bioinformatics and computational biology". In: *Frontiers in genetics* 10 (2019), p. 214.

[2203] Connie Huang et al. "Personal transcriptome variation is poorly explained by current genomic deep learning models". In: *Nature Genetics* (2023), pp. 1–4.

[2204] Tianwei Yue et al. "Deep Learning for Genomics: From Early Neural Nets to Modern Large Language Models". In: *International Journal of Molecular Sciences* 24.21 (2023), p. 15858.

[2205] Robert Zeibich et al. "Applications for Deep Learning in Epilepsy Genetic Research". In: *International journal of molecular sciences* 24.19 (2023), p. 14645.

[2206] Ziqi Tang, Shushan Toneyan, and Peter K Koo. "Current approaches to genomic deep learning struggle to fully capture human genetic variation". In: *Nature Genetics* (2023), pp. 1–2.

[2207] Raul I Perez Martell et al. “Supervised promoter recognition: a benchmark framework”. In: *BMC bioinformatics* 23.1 (2022), p. 118.

[2208] Georgina Stegmayer et al. “Predicting novel microRNA: a comprehensive comparison of machine learning approaches”. In: *Briefings in bioinformatics* 20.5 (2019), pp. 1607–1620.

[2209] Amlan Talukder et al. “Interpretation of deep learning in genomics and epigenomics”. In: *Briefings in Bioinformatics* 22.3 (2021), bbaa177.

[2210] Manuel Martínez-Sellés and Manuel Marina-Breysse. “Current and future use of artificial intelligence in electrocardiography”. In: *Journal of Cardiovascular Development and Disease* 10.4 (2023), p. 175.

[2211] Philipp Lohmann et al. “Radiomics in neuro-oncology: Basics, workflow, and applications”. In: *Methods* 188 (2021), pp. 112–121.

[2212] Nafiseh Erfanian et al. “Deep learning applications in single-cell genomics and transcriptomics data analysis”. In: *Biomedicine & Pharmacotherapy* 165 (2023), p. 115077.

[2213] Simone Marini et al. “Towards routine employment of computational tools for antimicrobial resistance determination via high-throughput sequencing”. In: *Briefings in bioinformatics* 23.2 (2022), bbac020.

[2214] Etienne Routhier and Julien Mozziconacci. “Genomics enters the deep learning era”. In: *PeerJ* 10 (2022), e13613.

[2215] Polina Mamoshina et al. “Applications of deep learning in biomedicine”. In: *Molecular pharmaceutics* 13.5 (2016), pp. 1445–1454.

[2216] Travers Ching et al. “Opportunities and obstacles for deep learning in biology and medicine”. In: *Journal of The Royal Society Interface* 15.141 (2018), p. 20170387.

[2217] Divneet Mandair, Jorge S Reis-Filho, and Alan Ashworth. “Biological insights and novel biomarker discovery through deep learning approaches in breast cancer histopathology”. In: *NPJ Breast Cancer* 9.1 (2023), p. 21.

[2218] Octavian Sabin Ttaru et al. “Artificial intelligence and machine learning in prostate cancer patient managementcurrent trends and future perspectives”. In: *Diagnostics* 11.2 (2021), p. 354.

[2219] Elena Ivanova et al. “Empowering Renal Cancer Management with AI and Digital Pathology: Pathology, Diagnostics and Prognosis”. In: *Biomedicines* 11.11 (2023), p. 2875.

[2220] Kenbun Sone et al. “Application of artificial intelligence in gynecologic malignancies: A review”. In: *Journal of Obstetrics and Gynaecology Research* 47.8 (2021), pp. 2577–2585.

[2221] Fabio Boniolo et al. “Artificial intelligence in early drug discovery enabling precision medicine”. In: *Expert Opinion on Drug Discovery* 16.9 (2021), pp. 991–1007.

[2222] Raihan Rafique, SM Riazul Islam, and Julhash U Kazi. "Machine learning in the prediction of cancer therapy". In: *Computational and Structural Biotechnology Journal* 19 (2021), pp. 4003–4017.

[2223] Chaoyi Zhang et al. "Novel research and future prospects of artificial intelligence in cancer diagnosis and treatment". In: *Journal of Hematology & Oncology* 16.1 (2023), p. 114.

[2224] Roman David Bülow et al. "How will artificial intelligence and bioinformatics change our understanding of IgA Nephropathy in the next decade?" In: *Seminars in Immunopathology*. Vol. 43. 5. Springer. 2021, pp. 739–752.

[2225] Tony Hauptmann and Stefan Kramer. "A fair experimental comparison of neural network architectures for latent representations of multi-omics for drug response prediction". In: *BMC bioinformatics* 24.1 (2023), p. 45.

[2226] Yanan Qin et al. "Deep learning identifies erroneous microarray-based, gene-level conclusions in literature". In: *NAR Genomics and Bioinformatics* 3.4 (2021), lqab089.

[2227] Sanjay Saxena et al. "Role of artificial intelligence in radiogenomics for cancers in the era of precision medicine". In: *Cancers* 14.12 (2022), p. 2860.

[2228] Dolly Das, Saroj Kr Biswas, and Sivaji Bandyopadhyay. "A critical review on diagnosis of diabetic retinopathy using machine learning and deep learning". In: *Multimedia Tools and Applications* 81.18 (2022), pp. 25613–25655.

[2229] Bo Lin et al. "Intelligent oncology: the convergence of artificial intelligence and oncology". In: *Journal of the National Cancer Center* 3.1 (2023), pp. 83–91.

[2230] Hang Qiu et al. "Applications of artificial intelligence in screening, diagnosis, treatment, and prognosis of colorectal cancer". In: *Current Oncology* 29.3 (2022), pp. 1773–1795.

[2231] Simin Zhang et al. "Artificial Intelligence Applications in Glioma With 1p/19q Co-Deletion: A Systematic Review". In: *Journal of Magnetic Resonance Imaging* (2023).

[2232] Wardah S Alharbi and Mamoon Rashid. "A review of deep learning applications in human genomics using next-generation sequencing data". In: *Human Genomics* 16.1 (2022), pp. 1–20.

[2233] Ninon Burgos et al. "Deep learning for brain disorders: from data processing to disease treatment". In: *Briefings in Bioinformatics* 22.2 (2021), pp. 1560–1576.

[2234] Kaitlyn Alleman et al. "Multimodal Deep Learning-Based Prognostication in Glioma Patients: A Systematic Review". In: *Cancers* 15.2 (2023), p. 545.

[2235] Pierre Murchan et al. "Deep learning of histopathological features for the prediction of tumour molecular genetics". In: *Diagnostics* 11.8 (2021), p. 1406.

[2236] Matthew Brendel et al. "Application of deep learning on single-cell RNA sequencing data analysis: a review". In: *Genomics, Proteomics & Bioinformatics* 20.5 (2022), pp. 814–835.

[2237] Alessio Gerussi et al. "Artificial intelligence for precision medicine in autoimmune liver disease". In: *Frontiers in Immunology* 13 (2022), p. 966329.

[2238] Shang-Fu Chen et al. "Artificial Intelligence for Risk Assessment on Primary Prevention of Coronary Artery Disease". In: *Current Cardiovascular Risk Reports* 17.12 (2023), pp. 215–231.

[2239] Christian F Luz et al. "Machine learning in infection management using routine electronic health records: tools, techniques, and reporting of future technologies". In: *Clinical Microbiology and Infection* 26.10 (2020), pp. 1291–1299.

[2240] Mehnaz Sarwar, Hassaan Malik, and Insia Zahra. "Innovative Computational Moulding Approach for Genomics". In: *2021 International Conference on Innovative Computing (ICIC)*. IEEE. 2021, pp. 1–7.

[2241] Diksha Pandey and P Onkara Perumal. "A scoping review on deep learning for next-generation RNA-Seq. data analysis". In: *Functional & Integrative Genomics* 23.2 (2023), p. 134.

[2242] Tanvir Alam, Hamada RH Al-Absi, and Sebastian Schmeier. "Deep learning in LncRNAome: contribution, challenges, and perspectives". In: *Non-coding RNA* 6.4 (2020), p. 47.

[2243] Xinran Xu et al. "A systematic review of computational methods for predicting long noncoding RNAs". In: *Briefings in functional genomics* 20.3 (2021), pp. 162–173.

[2244] Shuai Zeng et al. "G2PDeep: a web-based deep-learning framework for quantitative phenotype prediction and discovery of genomic markers". In: *Nucleic acids research* 49.W1 (2021), W228–W236.

[2245] Chensi Cao et al. "Deep learning and its applications in biomedicine". In: *Genomics, proteomics & bioinformatics* 16.1 (2018), pp. 17–32.

[2246] Yuanqi Zeng et al. "A review about transcription factor binding sites prediction based on deep learning". In: *Ieee Access* 8 (2020), pp. 219256–219274.

[2247] Louis-Oscar Morel et al. "Preliminary evaluation of deep learning for first-line diagnostic prediction of tumor mutational status". In: *Scientific Reports* 13.1 (2023), p. 6927.

[2248] Burak Yelmen and Flora Jay. "An Overview of Deep Generative Models in Functional and Evolutionary Genomics". In: *Annual Review of Biomedical Data Science* 6 (2023).

[2249] Mira Barshai, Eitamar Tripto, and Yaron Orenstein. "Identifying regulatory elements via deep learning". In: *Annual Review of Biomedical Data Science* 3 (2020), pp. 315–338.

[2250] Muhammad Shahid Iqbal et al. "Breast Cancer Dataset, Classification and Detection Using Deep Learning". In: *Healthcare*. Vol. 10. 12. MDPI. 2022, p. 2395.

[2251] Wan Zhu et al. "The application of deep learning in cancer prognosis prediction". In: *Cancers* 12.3 (2020), p. 603.

[2252] Aly Al-Amyn Valliani, Daniel Ranti, and Eric Karl Oermann. "Deep learning and neurology: a systematic review". In: *Neurology and therapy* 8 (2019), pp. 351–365.

[2253] Sushmita Mitra. "Deep Learning with Radiogenomics towards Personalized Management of Gliomas". In: *IEEE Reviews in Biomedical Engineering* 16 (2021), pp. 579–593.

[2254] Osval Antonio Montesinos-López et al. "A review of deep learning applications for genomic selection". In: *BMC genomics* 22 (2021), pp. 1–23.

[2255] Mazin Abed Mohammed et al. "Rise of Deep Learning Clinical Applications and Challenges in Omics Data: A Systematic Review". In: *Diagnostics* 13.4 (2023), p. 664.

[2256] Fangyoumin Feng et al. "Large-scale pharmacogenomic studies and drug response prediction for personalized cancer medicine". In: *Journal of Genetics and Genomics* 48.7 (2021), pp. 540–551.

[2257] Luis Marti-Bonmati et al. "Pancreatic cancer, radiomics and artificial intelligence". In: *The British Journal of Radiology* 95.1137 (2022), p. 20220072.

[2258] Michael W Mullowney et al. "Artificial intelligence for natural product drug discovery". In: *Nature Reviews Drug Discovery* 22.11 (2023), pp. 895–916.

[2259] Seonwoo Min, Byunghan Lee, and Sungroh Yoon. "Deep learning in bioinformatics". In: *Briefings in bioinformatics* 18.5 (2017), pp. 851–869.

[2260] Nadia Terranova, Karthik Venkatakrishnan, and Lisa J Benincosa. "Application of machine learning in translational medicine: current status and future opportunities". In: *The AAPS Journal* 23.4 (2021), p. 74.

[2261] Zakir Hossain et al. "The role of machine learning in developing non-magnetic resonance imaging based biomarkers for multiple sclerosis: a systematic review". In: *BMC Medical Informatics and Decision Making* 22.1 (2022), p. 242.

[2262] Stephanie Chan et al. "Machine learning in dermatology: current applications, opportunities, and limitations". In: *Dermatology and therapy* 10 (2020), pp. 365–386.

[2263] Alfredo Cesario et al. "The role of artificial intelligence in managing multimorbidity and cancer". In: *Journal of Personalized Medicine* 11.4 (2021), p. 314.

[2264] Chiara Corti et al. "Artificial intelligence in cancer research and precision medicine: Applications, limitations and priorities to drive transformation in the delivery of equitable and unbiased care". In: *Cancer Treatment Reviews* 112 (2023), p. 102498.

[2265] Khoa A Tran et al. "Deep learning in cancer diagnosis, prognosis and treatment selection". In: *Genome Medicine* 13.1 (2021), pp. 1–17.

[2266] Swati B Bhonde and Jayashree R Prasad. "Deep Learning Techniques in Cancer Prediction Using Genomic Profiles". In: *2021 6th International Conference for Convergence in Technology (I2CT)*. IEEE. 2021, pp. 1–9.

[2267] Swaroop Aradhya et al. "Applications of artificial intelligence in clinical laboratory genomics". In: *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*. Vol. 193. 3. Wiley Online Library. 2023, e32057.

[2268] Dipankar Baisya et al. "Genome-wide functional screens enable the prediction of high activity CRISPR-Cas9 and-Cas12a guides in Yarrowia lipolytica". In: *Nature communications* 13.1 (2022), p. 922.

[2269] Md Rezaul Karim et al. "Deep learning-based clustering approaches for bioinformatics". In: *Briefings in bioinformatics* 22.1 (2021), pp. 393–415.

[2270] Rufeng Li et al. "Machine learning meets omics: applications and perspectives". In: *Briefings in Bioinformatics* 23.1 (2022), bbab460.

[2271] Somayah Albaradei et al. "Machine learning and deep learning methods that use omics data for metastasis prediction". In: *Computational and structural biotechnology journal* 19 (2021), pp. 5008–5018.

[2272] Naiem T Issa et al. "Machine and deep learning approaches for cancer drug repurposing". In: *Seminars in cancer biology*. Vol. 68. Elsevier. 2021, pp. 132–142.

[2273] Nathan LaPierre et al. "MetaPheno: a critical evaluation of deep learning and machine learning in metagenome-based disease prediction". In: *Methods* 166 (2019), pp. 74–82.

[2274] Koji Arikawa and Masahito Hosokawa. "Uncultured prokaryotic genomes in the spotlight: An examination of publicly available data from metagenomics and single-cell genomics". In: *Computational and Structural Biotechnology Journal* (2023).

[2275] Sandra Brasil et al. "Artificial intelligence in epigenetic studies: Shedding light on rare diseases". In: *Frontiers in Molecular Biosciences* 8 (2021), p. 648012.

[2276] Alan J Mueller-Breckenridge et al. "Machine-learning based patient classification using Hepatitis B virus full-length genome quasispecies from Asian and European cohorts". In: *Scientific Reports* 9.1 (2019), p. 18892.

[2277] Michal Krassowski et al. "State of the field in multi-omics research: from computational needs to data mining and sharing". In: *Frontiers in Genetics* 11 (2020), p. 610798.

[2278] Yihao Liu and Minghua Wu. "Deep learning in precision medicine and focus on glioma". In: *Bioengineering & Translational Medicine* (2023), e10553.

[2279] Conceicao Bettencourt et al. "Artificial intelligence for dementia genetics and omics". In: *Alzheimer's & Dementia* 19.12 (2023), pp. 5905–5921.

[2280] Yurui Chen and Louxin Zhang. "How much can deep learning improve prediction of the responses to drugs in cancer cell lines?" In: *Briefings in bioinformatics* 23.1 (2022), bbab378.

[2281] Caijie Qin et al. "Application of artificial intelligence in diagnosis of craniopharyngioma". In: *Frontiers in Neurology* 12 (2022), p. 752119.

[2282] Tung-Hung Su, Chih-Horng Wu, and Jia-Horng Kao. "Artificial intelligence in precision medicine in hepatology". In: *Journal of Gastroenterology and Hepatology* 36.3 (2021), pp. 569–580.

[2283] Valeria Visco et al. "Artificial Intelligence in Hypertension Management: An Ace up Your Sleeve". In: *Journal of Cardiovascular Development and Disease* 10.2 (2023), p. 74.

[2284] Yulian Ding et al. "Machine learning approaches for predicting biomolecule–disease associations". In: *Briefings in Functional Genomics* 20.4 (2021), pp. 273–287.

[2285] Milan Picard et al. "Integration strategies of multi-omics data for machine learning analysis". In: *Computational and Structural Biotechnology Journal* 19 (2021), pp. 3735–3746.

[2286] Matteo Tortora et al. "RadioPathomics: Multimodal Learning in Non-Small Cell Lung Cancer for Adaptive Radiotherapy". In: *IEEE Access* (2023).

[2287] Liangrui Ren et al. "Single-cell RNA-seq data clustering by deep information fusion". In: *Briefings in Functional Genomics* (2023), elad017.

[2288] Yi Pan, Xiujuan Lei, and Yuchen Zhang. "Association predictions of genomics, proteinomics, transcriptomics, microbiome, metabolomics, pathomics, radiomics, drug, symptoms, environment factor, and disease networks: A comprehensive approach". In: *Medicinal Research Reviews* 42.1 (2022), pp. 441–461.

[2289] Richard J Chen et al. "Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis". In: *IEEE Transactions on Medical Imaging* 41.4 (2020), pp. 757–770.

[2290] Yi Hao Chan et al. "Combining neuroimaging and omics datasets for disease classification using graph neural networks". In: *Frontiers in Neuroscience* 16 (2022), p. 866666.

[2291] Xingxin Pan et al. "i-Modern: Integrated multi-omics network model identifies potential therapeutic targets in glioma by deep learning with interpretability". In: *Computational and Structural Biotechnology Journal* 20 (2022), pp. 3511–3521.

[2292] Yotsawat Pomyen et al. “Deep metabolome: Applications of deep learning in metabolomics”. In: *Computational and Structural Biotechnology Journal* 18 (2020), pp. 2818–2825.

[2293] Xi Wang and Bin-bin Li. “Deep learning in head and neck tumor multiomics diagnosis and analysis: review of the literature”. In: *Frontiers in Genetics* 12 (2021), p. 624820.

[2294] Yifeng Li, Fang-Xiang Wu, and Alioune Ngom. “A review on machine learning principles for multi-view biological data integration”. In: *Briefings in bioinformatics* 19.2 (2018), pp. 325–340.

[2295] Ren Yuan Lee et al. “The promise and challenge of spatial omics in dissecting tumour microenvironment and the role of AI”. In: *Frontiers in Oncology* 13 (2023), p. 1172314.

[2296] Ge Zhang et al. “A novel biomarker identification approach for gastric cancer using gene expression and DNA methylation dataset”. In: *Frontiers in Genetics* 12 (2021), p. 644378.

[2297] Souham Meshoul et al. “Explainable Multi-Class Classification Based on Integrative Feature Selection for Breast Cancer Subtyping”. In: *Mathematics* 10.22 (2022), p. 4271.

[2298] Dinesh K Saini et al. “Comprehensive evaluation of mapping complex traits in wheat using genome-wide association studies”. In: *Molecular Breeding* 42 (2022), pp. 1–52.

[2299] Burcu Darst et al. “Data mining and machine learning approaches for the integration of genome-wide association and methylation data: methodology and main conclusions from GAW20”. In: *BMC genetics* 19.1 (2018), pp. 81–88.

[2300] Eva Brombacher et al. “The performance of deep generative models for learning joint embeddings of single-cell multi-omics data”. In: *Frontiers in Molecular Biosciences* 9 (2022), p. 962644.

[2301] Lianhe Zhao et al. “DeepOmix: A scalable and interpretable multi-omics deep learning framework and application in cancer survival analysis”. In: *Computational and structural biotechnology journal* 19 (2021), pp. 2719–2725.

[2302] Jael Sanyanda Wekesa and Michael Kimwele. “A review of multi-omics data integration through deep learning approaches for disease diagnosis, prognosis, and treatment”. In: *Frontiers in Genetics* 14 (2023).

[2303] Xiao-Xia Yin et al. “MRI radiogenomics for intelligent diagnosis of breast tumors and accurate prediction of neoadjuvant chemotherapy responses-a review”. In: *Computer methods and programs in biomedicine* 214 (2022), p. 106510.

[2304] Chang Su et al. “Deep learning in mental health outcome research: a scoping review”. In: *Translational Psychiatry* 10.1 (2020), p. 116.

[2305] Suzanna E Ledgister Hanchard et al. "Scoping review and classification of deep learning in medical genetics". In: *Genetics in Medicine* 24.8 (2022), pp. 1593–1603.

[2306] Guangyao Wu et al. "Structural and functional radiomics for lung cancer". In: *European Journal of Nuclear Medicine and Molecular Imaging* 48 (2021), pp. 3961–3974.

[2307] David Wissel, Daniel Rowson, and Valentina Boeva. "Systematic comparison of multi-omics survival models reveals a widespread lack of noise resistance". In: *Cell Reports Methods* 3.4 (2023).

[2308] Amrita Chattopadhyay and Tzu-Pin Lu. "Gene-gene interaction: the curse of dimensionality". In: *Annals of translational medicine* 7.24 (2019).

[2309] Tyrone Chen and Sonika Tyagi. "Integrative computational epigenomics to build data-driven gene regulation hypotheses". In: *GigaScience* 9.6 (2020), giaa064.

[2310] Roberto Gatta et al. "Integrating radiomics into holomics for personalised oncology: from algorithms to bedside". In: *European radiology experimental* 4.1 (2020), pp. 1–9.

[2311] Teppei Matsubara et al. "Convolutional neural network approach to lung cancer classification integrating protein interaction network and gene expression profiles". In: *Journal of bioinformatics and computational biology* 17.03 (2019), p. 1940007.

[2312] Giovanna Nicora et al. "Integrated multi-omics analyses in oncology: a review of machine learning methods and tools". In: *Frontiers in oncology* 10 (2020), p. 1030.

[2313] Fangfang Xia et al. "A cross-study analysis of drug response prediction in cancer cell lines". In: *Briefings in bioinformatics* 23.1 (2022), bbab356.

[2314] Sukanya Panja et al. "Big data to knowledge: application of machine learning to predictive modeling of therapeutic response in cancer". In: *Current genomics* 22.4 (2021), p. 244.

[2315] Sandra Brasil et al. "Artificial intelligence (AI) in rare diseases: is the future brighter?" In: *Genes* 10.12 (2019), p. 978.

[2316] Yogesh Kumar et al. "Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda". In: *Journal of ambient intelligence and humanized computing* 14.7 (2023), pp. 8459–8486.

[2317] Bahrad A Sokhansanj and Gail L Rosen. "Mapping data to deep understanding: Making the most of the deluge of SARS-CoV-2 genome sequences". In: *Msystems* 7.2 (2022), e00035–22.

[2318] Xuehai Hu, Alisdair R Fernie, and Jianbing Yan. "Deep learning in regulatory genomics: from identification to design". In: *Current Opinion in Biotechnology* 79 (2023), p. 102887.

[2319] Claudia Caudai et al. “AI applications in functional genomics”. In: *Computational and Structural Biotechnology Journal* 19 (2021), pp. 5762–5790.

[2320] Mariangela Santorsola and Francesco Lescai. “The Promise of Explainable Deep Learning for Omics Data Analysis: Adding New Discovery Tools to AI”. In: *New Biotechnology* (2023).

[2321] Riccardo Miotto et al. “Deep learning for healthcare: review, opportunities and challenges”. In: *Briefings in bioinformatics* 19.6 (2018), pp. 1236–1246.

[2322] Zhen-Jie Yao, Jie Bi, and Yi-Xin Chen. “Applying deep learning to individual and community health monitoring data: A survey”. In: *International Journal of Automation and Computing* 15 (2018), pp. 643–655.

[2323] Xiaoyong Pan et al. “Recent methodology progress of deep learning for RNA–protein interaction prediction”. In: *Wiley Interdisciplinary Reviews: RNA* 10.6 (2019), e1544.

[2324] James A Diao, Isaac S Kohane, and Arjun K Manrai. “Biomedical informatics and machine learning for clinical genomics”. In: *Human molecular genetics* 27.R1 (2018), R29–R34.

[2325] Chun Yen Lee and Yi-Ping Phoebe Chen. “New insights into drug repurposing for COVID-19 using deep learning”. In: *IEEE Transactions on Neural Networks and Learning Systems* 32.11 (2021), pp. 4770–4780.

[2326] Animesh Ray. “Machine learning in postgenomic biology and personalized medicine”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 12.2 (2022), e1451.

[2327] Nikita Bhandari et al. “A comprehensive survey on computational learning methods for analysis of gene expression data”. In: *Frontiers in Molecular Biosciences* 9 (2022), p. 907150.

[2328] Yu Zhang et al. “A survey on neural network interpretability”. In: *IEEE Transactions on Emerging Topics in Computational Intelligence* 5.5 (2021), pp. 726–742.

[2329] Raquel Dias and Ali Torkamani. “Artificial intelligence in clinical and genomic diagnostics”. In: *Genome medicine* 11.1 (2019), pp. 1–12.

[2330] Xindi Yu et al. “Survey of deep learning techniques for disease prediction based on omics data”. In: *Human Gene* 35 (2023), p. 201140.

[2331] Yunfeng Wu, Sridhar Krishnan, and Behnaz Ghoraani. “Computational Methods for Physiological Signal Processing and Data Analysis”. In: *Computational and Mathematical Methods in Medicine* 2022 (2022).

[2332] Bikash Ranjan Samal et al. “Opportunities and challenges in interpretable deep learning for drug sensitivity prediction of cancer cells”. In: *Frontiers in Bioinformatics* (2022), p. 104.

[2333] Yuqi Wen et al. “Deep Learning-Based Multiomics Data Integration Methods for Biomedical Application”. In: *Advanced Intelligent Systems* (2023), p. 2200247.

[2334] Andreas Holzinger et al. "Machine learning and knowledge extraction in digital pathology needs an integrative approach". In: *Towards Integrative Machine Learning and Knowledge Extraction: BIRS Workshop, Banff, AB, Canada, July 24-26, 2015, Revised Selected Papers*. Springer. 2017, pp. 13–50.

[2335] Ruoqi Wei and Ausif Mahmood. "Recent advances in variational autoencoders with representation learning for biomedical informatics: A survey". In: *Ieee Access* 9 (2020), pp. 4939–4956.

[2336] Paul Villoutreix. "What machine learning can do for developmental biology". In: *Development* 148.1 (2021), dev188474.

[2337] Sunil Kapila et al. "Connecting the dots towards precision orthodontics". In: *Orthodontics & craniofacial research* 26 (2023), pp. 8–19.

[2338] Tamizhini Loganathan and George Priya Doss C. "The influence of machine learning technologies in gut microbiome research and cancer studies-A review". In: *Life Sciences* (2022), p. 121118.

[2339] Youngjun Park, Dominik Heider, and Anne-Christin Hauschild. "Integrative analysis of next-generation sequencing for next-generation cancer research toward artificial intelligence". In: *Cancers* 13.13 (2021), p. 3148.

[2340] Kevin M Mendez, David I Broadhurst, and Stacey N Reinke. "The application of artificial neural networks in metabolomics: a historical perspective". In: *Metabolomics* 15 (2019), pp. 1–14.

[2341] Tehseen Mazhar et al. "The role of machine learning and deep learning approaches for the detection of skin cancer". In: *Healthcare*. Vol. 11. 3. MDPI. 2023, p. 415.

[2342] Peter K Koo and Matt Ploenzke. "Deep learning for inferring transcription factor binding sites". In: *Current opinion in systems biology* 19 (2020), pp. 16–23.

[2343] Kexin Huang et al. "Machine learning applications for therapeutic tasks with genomics data". In: *Patterns* 2.10 (2021).

[2344] Qian Liu and Pingzhao Hu. "Extendable and explainable deep learning for pan-cancer radiogenomics research". In: *Current Opinion in Chemical Biology* 66 (2022), p. 102111.

[2345] Óscar Álvarez-Machancoses et al. "On the role of artificial intelligence in genomics to enhance precision medicine". In: *Pharmacogenomics and personalized medicine* (2020), pp. 105–119.

[2346] Mehmet A Gulum, Christopher M Trombley, and Mehmed Kantardzic. "A review of explainable deep learning cancer detection models in medical imaging". In: *Applied Sciences* 11.10 (2021), p. 4573.

[2347] Magdalena Wysocka et al. "A systematic review of biologically-informed deep learning models for cancer: fundamental trends for encoding and interpreting oncology data". In: *BMC bioinformatics* 24.1 (2023), pp. 1–31.

[2348] Mamatha Bhat et al. "Artificial intelligence, machine learning, and deep learning in liver transplantation". In: *Journal of hepatology* 78.6 (2023), pp. 1216–1233.

[2349] Wensi Zhu et al. "Evaluation of AlphaFold-Multimer prediction on multi-chain protein complexes". In: *Bioinformatics* 39.7 (2023), btad424.

[2350] Rui Yin and Brian G Pierce. "Evaluation of AlphaFold antibody–antigen modeling with implications for improving predictive accuracy". In: *Protein Science* 33.1 (2024), e4865.

[2351] Rui Yin et al. "Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants". In: *Protein Science* 31.8 (2022), e4379.

[2352] Seifeddine Messaoud et al. "Virtual healthcare center for COVID-19 patient detection based on artificial intelligence approaches". In: *Canadian Journal of Infectious Diseases and Medical Microbiology* 2022 (2022).

[2353] Ramón Alvarado. "Should we replace radiologists with deep learning? Pigeons, error and trust in medical AI". In: *Bioethics* 36.2 (2022), pp. 121–133.

[2354] Christopher E Lawson et al. "Machine learning for metabolic engineering: A review". In: *Metabolic Engineering* 63 (2021), pp. 34–60.

[2355] Ryuji Hamamoto et al. "Epigenetics analysis and integrated analysis of multiomics data, including epigenetic data, using artificial intelligence in the era of precision medicine". In: *Biomolecules* 10.1 (2019), p. 62.

[2356] David Sidak et al. "Interpretable machine learning methods for predictions in systems biology from omics data". In: *Frontiers in Molecular Biosciences* 9 (2022), p. 926623.

[2357] Stanislav Y Shvartsman and Ruth E Baker. "Mathematical models of morphogen gradients and their effects on gene expression". In: *Wiley Interdisciplinary Reviews: Developmental Biology* 1.5 (2012), pp. 715–730.

[2358] Jörg Stelling. "Mathematical models in microbial systems biology". In: *Current opinion in microbiology* 7.5 (2004), pp. 513–518.

[2359] David Sankoff. "Gene and genome duplication". In: *Current opinion in genetics & development* 11.6 (2001), pp. 681–684.

[2360] Peter Schuster. "Mathematical modeling of evolution. Solved and open problems". In: *Theory in Biosciences* 130 (2011), pp. 71–89.

[2361] Pietro Liò and Nick Goldman. "Models of molecular evolution and phylogeny". In: *Genome research* 8.12 (1998), pp. 1233–1244.

[2362] David Moreira and Hervé Philippe. "Molecular phylogeny: pitfalls and progress". In: *International Microbiology* 3.1 (2000), pp. 9–16.

[2363] Olivier Gascuel. *Mathematics of evolution and phylogeny*. OUP Oxford, 2005.

[2364] Ahmet Ay and David N Arnosti. "Mathematical modeling of gene expression: a guide for the perplexed biologist". In: *Critical reviews in biochemistry and molecular biology* 46.2 (2011), pp. 137–151.

[2365] Deepak K Agrawal et al. "Mathematical modeling of RNA-based architectures for closed loop control of gene expression". In: *ACS synthetic biology* 7.5 (2018), pp. 1219–1228.

[2366] Alireza Sepehri. "A mathematical model for DNA". In: *International Journal of Geometric Methods in Modern Physics* 14.11 (2017), p. 1750152.

[2367] MI Drobotenko et al. "A mathematical model for basepair opening in a DNA double helix". In: *Biophysics* 63 (2018), pp. 177–182.

[2368] Puneet Rawat et al. "AggreRATE-Pred: a mathematical model for the prediction of change in aggregation rate upon point mutation". In: *Bioinformatics* 36.5 (2020), pp. 1439–1444.

[2369] Kouki Uchinomiya and Masanori Tomita. "A mathematical model for cancer risk and accumulation of mutations caused by replication errors and external factors". In: *Plos one* 18.6 (2023), e0286499.

[2370] Xiaoqiang Sun and Bin Hu. "Mathematical modeling and computational prediction of cancer drug resistance". In: *Briefings in bioinformatics* 19.6 (2018), pp. 1382–1399.

[2371] Eleanor Chu and Alan George. *Inside the FFT black box: Serial and parallel Fast Fourier Transform algorithms*. CRC press, 1999.

[2372] Sou Miyamoto et al. "Two-stage sequence-to-sequence neural voice conversion with low-to-high definition spectrogram mapping". In: *Recent Advances in Intelligent Information Hiding and Multimedia Signal Processing: Proceeding of the Fourteenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, November, 26-28, 2018, Sendai, Japan, Volume 2 14*. Springer. 2019, pp. 132–139.

[2373] S Jothimani and Kandhasamy Premalatha. "MFF-SAug: Multi feature fusion with spectrogram augmentation of speech emotion recognition using Convolution Neural Network". In: *Chaos, Solitons & Fractals* 162 (2022), p. 112512.

[2374] Coen PH Elemans, Kier Heeck, and Mees Muller. "Spectrogram analysis of animal sound production". In: *Bioacoustics* 18.2 (2008), pp. 183–212.

[2375] Francois Leonard et al. "Free-vibration behaviour of a cracked cantilever beam and crack detection". In: *Mechanical systems and signal processing* 15.3 (2001), pp. 529–548.

[2376] Xiaopeng Dong et al. "Detection and identification of vehicles based on their unintended electromagnetic emissions". In: *IEEE Transactions on Electromagnetic Compatibility* 48.4 (2006), pp. 752–759.

[2377] RIA Harmanny, Jacco JM De Wit, and G Prémel Cabic. "Radar micro-Doppler feature extraction using the spectrogram and the cepstrogram". In: *2014 11th European Radar Conference*. IEEE. 2014, pp. 165–168.

[2378] Vincent Roma et al. “Astrophysics with core-collapse supernova gravitational wave signals in the next generation of gravitational wave detectors”. In: *Physical Review D* 99.6 (2019), p. 063018.

[2379] Emma Suriñach et al. “Estimation of avalanche development and frontal velocities based on the spectrogram of the seismic signals generated at the Vallée de la Sionne Test Site”. In: *Geosciences* 10.3 (2020), p. 113.

[2380] Millaray Curilem et al. “Discriminating seismic events of the Llaima volcano (Chile) based on spectrogram cross-correlations”. In: *Journal of Volcanology and Geothermal Research* 367 (2018), pp. 63–78.

[2381] Hongzu Li and Pierre Boulanger. “Structural Anomalies Detection from Electrocardiogram (ECG) with Spectrogram and Handcrafted Features”. In: *Sensors* 22.7 (2022), p. 2467.

[2382] James McNames and Andrew M Fraser. “Obstructive sleep apnea classification based on spectrogram patterns in the electrocardiogram”. In: *Computers in Cardiology 2000. vol. 27 (Cat. 00CH37163)*. IEEE. 2000, pp. 749–752.

[2383] Md Nurul Ahad Tawhid et al. “A spectrogram image based intelligent technique for automatic detection of autism spectrum disorder from EEG”. In: *Plos one* 16.6 (2021), e0253094.

[2384] Robson P Bonidia et al. “Feature extraction approaches for biological sequences: A comparative study of mathematical features”. In: *Briefings in Bioinformatics* 22.5 (2021), bbab011.

[2385] Alan Wee Chung Liew, Hong Yan, and Mengsu Yang. “Pattern recognition techniques for the emerging field of bioinformatics: A review”. In: *Pattern Recognition* 38.11 (2005), pp. 2055–2073.

[2386] Shelly Soffer et al. “Convolutional neural networks for radiologic images: a radiologists guide”. In: *Radiology* 290.3 (2019), pp. 590–606.

[2387] Shrish Tiwari et al. “Prediction of probable genes by Fourier analysis of genomic sequences”. In: *Bioinformatics* 13.3 (1997), pp. 263–270.

[2388] Henri J Nussbaumer. “The Fast Fourier Transform”. In: *Fast Fourier Transform and Convolution Algorithms*. Springer, 1981, pp. 80–111.

[2389] John G Proakis and Dimitris G Manolakis. *Introduction to Digital Signal Processing*. Prentice Hall Professional Technical Reference, 1988.

[2390] Andy Extance. “How DNA could store all the world’s data”. In: *Nature* 537.7618 (2016).

[2391] Jessica Alföldi et al. “The genome of the green anole lizard and a comparative analysis with birds and mammals”. In: *Nature* 477.7366 (2011), pp. 587–591.

[2392] J Michael Cherry. “The Saccharomyces genome database: a tool for discovery”. In: *Cold Spring Harbor Protocols* 2015.12 (2015), pdb–top083840.

[2393] Michael McClelland et al. “Complete genome sequence of Salmonella enterica serovar Typhimurium LT2”. In: *Nature* 413.6858 (2001), pp. 852–856.

[2394] Sherwood Casjens. "Prophages and bacterial genomics: what have we learned so far?" In: *Molecular microbiology* 49.2 (2003), pp. 277–300.

[2395] Matplotlib Contributors. *Matplotlib: Visualization with Python.* Accessed: 2024-03-22. 2024. URL: `https://matplotlib.org`.

[2396] Omar Del Tejo Catalá et al. "Bias analysis on public X-ray image datasets of pneumonia and COVID-19 patients". In: *IEEE Access* 9 (2021), pp. 42370–42383.

[2397] Lieyang Chen et al. "Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening". In: *PloS One* 14.8 (2019), e0220113.

[2398] SciPy Contributors. *Documentation of Spectrogram function of the Scipy library in Python environment.* Accessed: 2024-03-22. 2024. URL: `https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.spectrogram.html`.

[2399] Orhan Gazi. "Understanding digital signal processing". In: (2018).

[2400] Henry Pfister. "Discrete-Time Signal Processing". In: *Lecture Note, pfister. ee. duke. edu/courses/ece485/dtsp. pdf* (2017).

[2401] Ronald E Crochiere and Lawrence R Rabiner. *Multirate digital signal processing*. Vol. 18. Prentice-hall Englewood Cliffs, NJ, 1983.

[2402] Sanjit K Mitra and James F Kaiser. *Handbook for digital signal processing*. John Wiley & Sons, Inc., 1993.

[2403] Monson H Hayes. *Statistical digital signal processing and modeling*. John Wiley & Sons, 1996.

[2404] Donald B Percival and Andrew T Walden. *Spectral analysis for physical applications*. cambridge university press, 1993.

[2405] Francis Castanié. *Spectral analysis: parametric and non-parametric digital methods*. John Wiley & Sons, 2013.

[2406] Lloyd N Trefethen. *Spectral methods in MATLAB*. SIAM, 2000.

[2407] Paul Dan Cristea. "Conversion of nucleotides sequences into genomic signals". In: *Journal of cellular and molecular medicine* 6.2 (2002), pp. 279–303.

[2408] S Paget-McNicol et al. "The Plasmodium falciparum var gene switching rate, switching mechanism and patterns of parasite recrudescence described by mathematical modelling". In: *Parasitology* 124.3 (2002), pp. 225–235.

[2409] Ludmil B Alexandrov and Michael R Stratton. "Mutational signatures: the patterns of somatic mutations hidden in cancer genomes". In: *Current opinion in genetics & development* 24 (2014), pp. 52–60.

[2410] *UNAIDS - AIDS by the Numbers*. Accessed on 2021/10/22. UNAIDS. URL: `https://www.unaids.org/`.

[2411] Maria Mercedes Santoro and Carlo Federico Perno. "HIV-1 genetic variability and clinical implications". In: *International Scholarly Research Notices* 2013 (2013).

[2412] DA Collier, Christopher Monit, and Ravindra Kumar Gupta. "The impact of HIV-1 drug escape on the global treatment landscape". In: *Cell host & microbe* 26.1 (2019), pp. 48–60.

[2413] Pilar Mendoza et al. "Combination therapy with anti-HIV-1 antibodies maintains viral suppression". In: *Nature* 561.7724 (2018), pp. 479–484.

[2414] Eric J Arts and Daria J Hazuda. "HIV-1 antiretroviral drug therapy". In: *Cold Spring Harbor perspectives in medicine* 2.4 (2012).

[2415] Myron S Cohen et al. "Antiretroviral therapy for the prevention of HIV-1 transmission". In: *New England Journal of Medicine* 375.9 (2016), pp. 830–839.

[2416] Tao Sheng Li et al. "Long-lasting recovery in CD4 T-cell function and viral-load reduction after highly active antiretroviral therapy in advanced HIV-1 disease". In: *The Lancet* 351.9117 (1998), pp. 1682–1686.

[2417] Leonidas Stamatatos. "Immunization during ART and ATI for HIV-1 vaccine discovery/development". In: *Current Opinion in HIV and AIDS* 18.6 (2023), pp. 309–314.

[2418] Emmanuel Ndashimye, Paul S Reyes, and Eric J Arts. "New antiretroviral inhibitors and HIV-1 drug resistance: More focus on 90% HIV-1 isolates?" In: *FEMS Microbiology Reviews* 47.1 (2023), fuac040.

[2419] Rafael Nájera et al. "Genetic recombination and its role in the development of the HIV-1 pandemic". In: *AIDS* 16 (2002), S3–S16.

[2420] Joëlle Eid, Marylène Mougel, and Marius Socol. "Advances in continuous microfluidics-based technologies for the study of HIV infection". In: *Viruses* 12.9 (2020), p. 982.

[2421] Heinrich G Göttlinger. "HIV-1 Gag: A molecular machine driving viral particle assembly and release". In: *HIV sequence Compendium* 2001 (2001), p. 2.

[2422] Yasemin van Heuvel et al. "Infectious RNA: human immunodeficiency virus (HIV) biology, therapeutic intervention, and the quest for a vaccine". In: *Toxins* 14.2 (2022), p. 138.

[2423] Eric O Freed. "HIV-1 replication". In: *Somatic cell and molecular genetics* 26 (2001), pp. 13–33.

[2424] Richard Wyatt et al. "Structure of the core of the HIV-1 gp120 exterior envelope glycoprotein". In: *Nature* (1998), pp. 3–9.

[2425] Jonathan Karn. "Tat, a novel regulator of HIV transcription and latency". In: *HIV Seq Compend* 1990 (2000), pp. 2–18.

[2426] Victoria W Pollard and Michael H Malim. "The HIV-1 rev protein". In: *Annual review of microbiology* 52.1 (1998), pp. 491–532.

[2427] Stefanie Jäger et al. "Vif hijacks CBF-β to degrade APOBEC3G and promote HIV-1 infection". In: *Nature* 481.7381 (2012), pp. 371–375.

[2428] Joshua L Andersen and Vicente Planelles. "The role of Vpr in HIV-1 pathogenesis". In: *Current HIV research* 3.1 (2005), pp. 43–51.

[2429] Stuart J D Neil et al. "HIV-1 Vpu promotes release and prevents endocytosis of nascent retrovirus particles from the plasma membrane". In: *PLoS pathogens* 2.5 (2006), e39.

[2430] Daniel Sauter et al. "Tetherin-driven adaptation of Vpu and Nef function and the evolution of pandemic and nonpandemic HIV-1 strains". In: *Cell host & microbe* 6.5 (2009), pp. 409–421.

[2431] Stéphane Basmaciogullari and Massimo Pizzato. "The activity of Nef on HIV-1 infectivity". In: *Frontiers in microbiology* 5 (2014), p. 232.

[2432] Oliver T Fackler and Andreas S Baur. "Live and let die: Nef functions beyond HIV replication". In: *Immunity* 16.4 (2002), pp. 493–497.

[2433] Etienne Simon-Loriere and Edward C Holmes. "Why do RNA viruses recombine?" In: *Nature Reviews Microbiology* 9.8 (2011), pp. 617–626.

[2434] John M Coffin, Stephen H Hughes, and Harold E Varmus. "Retroviruses". In: *Cold Spring Harbor Laboratory Press.* (1997).

[2435] Fred C Krebs et al. "Lentiviral LTR-directed expression, sequence variation, and disease pathogenesis". In: *HIV sequence compendium* 2001 (2001), pp. 29–70.

[2436] Alfredo Santana et al. "Biología celular y molecular del virus de inmunodeficiencia humana (VIH)". In: *Revista de Diagnóstico Biológico* 52.1 (2003), pp. 07–18.

[2437] Raúl Gómez-Román and Carmen Soler-Claudín. "La importancia de la Secuencia Terminal Repetida Larga (LTR) en la patogenia del virus de la inmunodeficiencia humana". In: *Revista Biomédica* 11.1 (2000), pp. 61–71.

[2438] Piotr Widak and William T Garrard. "Nucleosomes and regulation of gene expression. Structure of the HIV-1 5'LTR." In: *Acta Biochimica Polonica* 45.1 (1998), pp. 209–219.

[2439] Bryan R Gullen and MH Malim. "Regulation of HIV-1 gene expression". In: *Nucleic Acids and Molecular Biology 4* (1990), pp. 176–184.

[2440] Wei Shau Hu and Howard M Temin. "Genetic consequences of packaging two RNA genomes in one retroviral particle: pseudodiploidy and high rate of genetic recombination." In: *Proceedings of the National Academy of Sciences* 87.4 (1990), pp. 1556–1560.

[2441] Gabriela Turk et al. "Higher transactivation activity associated with LTR and Tat elements from HIV-1 BF intersubtype recombinant variants". In: *Retrovirology* 3.1 (2006), pp. 1–12.

[2442] Joris Hemelaar et al. "Global and regional molecular epidemiology of HIV-1, 1990–2015: a systematic review, global survey, and trend analysis". In: *The Lancet Infectious Diseases* 19.2 (2019), pp. 143–155.

[2443] Dennis Maletich Junqueira and Sabrina Esteves de Matos Almeida. "HIV-1 subtype B: Traces of a pandemic". In: *Virology* 495 (2016), pp. 173–184.

[2444] Polly R Walker et al. "Comparative population dynamics of HIV-1 subtypes B and C: Subtype-specific differences in patterns of epidemic growth". In: *Infection, Genetics and Evolution* 5.3 (2005), pp. 199–208.

[2445] Xingguang Li et al. "Tracing the epidemic history of HIV-1 CRF01_AE clusters using near-complete genome sequences". In: *Scientific Reports* 7.1 (2017), p. 4024.

[2446] Daiana Mir et al. "Phylodynamics of the major HIV-1 CRF02_AG African lineages and its global dissemination". In: *Infection, Genetics and Evolution* 46 (2016), pp. 190–199.

[2447] Vaishali Gupta and Ruchi Patel. "Lungs disease classification using VGG-16 architecture with PCA". In: *2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT)*. IEEE. 2023, pp. 495–500.

[2448] Zishun Zhou et al. "Detection of Pneumonia based on ResNet improved by attention mechanism". In: *2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA)*. IEEE. 2023, pp. 859–863.

[2449] Rahul Singh, Neha Sharma, and Rupesh Gupta. "Detection of Alzheimer's risk level using Inception V3 transfer learning model". In: *2023 International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*. IEEE. 2023, pp. 1–6.

[2450] *MathWorks*. Accessed on 2021/11/13. URL: `https://www.mathworks.com`.

[2451] Mohanad Azeez Joodi, Muna Hadi Saleh, and Dheyaa Jasim Kadhim. "Increasing validation accuracy of a face mask detection by new deep learning model-based classification". In: *Indonesian Journal of Electrical Engineering and Computer Science* 29.1 (2023), pp. 304–314.

[2452] Saptadeep Pal et al. "Optimizing multi-GPU parallelization strategies for deep learning training". In: *Ieee Micro* 39.5 (2019), pp. 91–101.

[2453] Sofia Visa et al. "Confusion matrix-based feature selection." In: *Maics* 710.1 (2011), pp. 120–127.

[2454] Tianbao Yang and Yiming Ying. "AUC maximization in the era of Big Data and AI: A survey". In: *ACM Computing Surveys* 55.8 (2022), pp. 1–37.

[2455] *Grad-CAM Explains Why*. Accessed on 2021/10/22. MathWorks. URL: `https://es.mathworks.com/help/deeplearning/ug/gradcam-explains-why.html`.

[2456] *Understand Network Predictions using LIME*. Accessed on 2021/10/22. MathWorks. URL: https://es.mathworks.com/help/deeplearning/ug/lime-understand-network-predictions.html.

[2457] *Investigate Spectrogram Classifications Using LIME*. Accessed on 2021/10/22. MathWorks. URL: https://www.mathworks.com/help/deeplearning/ug/investigate-spectrogram-classifications-using-lime.html.

[2458] *Deep Learning Visualization Methods*. Accessed on 2021/10/22. MathWorks. URL: https://es.mathworks.com/help/deeplearning/ug/deep-learning-visualization-methods.html.

[2459] Kevin D Hyde et al. "Incorporating molecular data in fungal systematics: a guide for aspiring researchers". In: *arXiv preprint arXiv:1302.3244* (2013).

[2460] Eduardo Castro-Nallar et al. "The evolution of HIV: inferences using phylogenetics". In: *Molecular phylogenetics and evolution* 62.2 (2012), pp. 777–792.

[2461] Carla Kuiken, Bette Korber, and Robert W Shafer. "HIV sequence databases". In: *AIDS Reviews* 5.1 (2003), p. 52.

[2462] Sean Eddy and the HMMER development team. *HMMER Software Suite*. Accessed: 2024-06-25. 2024. URL: http://hmmer.org/.

[2463] Tom Hall. *BioEdit Software*. Accessed: 2024-06-25. 2024. URL: https://thalljiscience.github.io/.

[2464] Rachel Colquhoun et al. *Se-Al Software*. Accessed: 2024-06-25. 2024. URL: http://tree.bio.ed.ac.uk/software/seal/.

[2465] Anders Larsson. "AliView: a fast and lightweight alignment viewer and editor for large datasets". In: *Bioinformatics* 30.22 (2014), pp. 3276–3278.

[2466] Genetic Information Research Institute. *MASE Biotool*. Accessed: 2024-06-25. 2024. URL: https://www.girinst.org/downloads/software/mase/.

[2467] Yi Feng et al. "The rapidly expanding CRF01_AE epidemic in China is driven by multiple lineages of HIV-1 viruses introduced in the 1990s". In: *AIDS (London, England)* 27.11 (2013), p. 1793.

[2468] Matthew J Gartner et al. "Understanding the mechanisms driving the spread of subtype C HIV-1". In: *EBioMedicine* 53 (2020).

[2469] Richard A Neher and Thomas Leitner. "Recombination rate and selection strength in HIV intra-patient evolution". In: *PLoS Computational Biology* 6.1 (2010), e1000660.

[2470] Wei Shau Hu and Howard M Temin. "Retroviral recombination and reverse transcription". In: *Science* 250.4985 (1990), pp. 1227–1233.

[2471] Bertha Cecilia Ramirez et al. "Implications of recombination for HIV diversity". In: *Virus Research* 134.1-2 (2008), pp. 64–73.

[2472] Eva Ramírez de Arellano et al. "Drastic decrease of transcription activity due to hypermutated long terminal repeat (LTR) region in different HIV-1 subtypes and recombinants". In: *Antiviral Research* 88.2 (2010), pp. 152–159.

[2473] Jean-Pierre Vartanian et al. "Selection, recombination, and G-A hypermutation of human immunodeficiency virus type 1 genomes". In: *Journal of Virology* 65.4 (1991), pp. 1779–1788.

[2474] Abayomi S Olabode et al. "Revisiting the recombinant history of HIV-1 group M with dynamic network community detection". In: *Proceedings of the National Academy of Sciences* 119.19 (2022), e2108815119.

[2475] Lisa R Banner, James G Keck, and Michael MC Lai. "A clustering of RNA recombination sites adjacent to a hypervariable region of the peplomer gene of murine coronavirus". In: *Virology* 175.2 (1990), pp. 548–555.

[2476] Nathalie Désiré et al. "Characterization update of HIV-1 M subtypes diversity and proposal for subtypes A and D sub-subtypes reclassification". In: *Retrovirology* 15 (2018), pp. 1–7.

[2477] Syed Hani Abidi et al. "Origin and evolution of HIV-1 subtype A6". In: *Plos One* 16.12 (2021), e0260604.

[2478] Madita Schlösser et al. "HIV-1 sub-subtype A6: settings for normalised identification and molecular epidemiology in the Southern Federal District, Russia". In: *Viruses* 12.4 (2020), p. 475.

[2479] Gkikas Magiorkinis et al. "The global spread of HIV-1 subtype B epidemic". In: *Infection, genetics and evolution* 46 (2016), pp. 169–179.

[2480] Gonzalo Bello et al. "Origin and evolutionary history of HIV-1 subtype C in Brazil". In: *Aids* 22.15 (2008), pp. 1993–2000.

[2481] André F Santos et al. "Characterization of a new circulating recombinant form comprising HIV-1 subtypes C and B in southern Brazil". In: *Aids* 20.16 (2006), pp. 2011–2019.

[2482] Stéphanie Raymond et al. "Genotypic prediction of HIV-1 subtype D tropism". In: *Retrovirology* 8 (2011), pp. 1–8.

[2483] Paula C Aulicino et al. "Description of the first full-length HIV type 1 subtype F1 strain in Argentina: implications for the origin and dispersion of this subtype in South America". In: *AIDS research and human retroviruses* 23.10 (2007), pp. 1176–1182.

[2484] LG Gurtler et al. "HIV-1 subtype O: epidemiology, pathogenesis, diagnosis, and perspectives of the evolution of HIV". In: *Archives of Virology-Supplements* 11 (1996), p. 195.

[2485] Shannon Bush and Denis M Tebit. "HIV-1 Group O Origin, Evolution, Pathogenesis, and Treatment: Unraveling the Complexity of an Outlier 25 Years Later." In: *AIDS reviews* 17.3 (2015).

[2486] Frank AJ Konings et al. "Human immunodeficiency virus type 1 (HIV-1) circulating recombinant form 02_AG (CRF02_AG) has a higher in vitro replicative capacity than its parental subtypes A and G". In: *Journal of medical virology* 78.5 (2006), pp. 523–534.

[2487] Min Wei et al. "Simple subtyping assay for human immunodeficiency virus type 1 subtypes B, C, CRF01-AE, CRF07-BC, and CRF08-BC". In: *Journal of clinical microbiology* 42.9 (2004), pp. 4261–4267.

[2488] Yi Feng et al. "Geographic origin and evolutionary history of Chinas two predominant HIV-1 circulating recombinant forms, CRF07_BC and CRF08_BC". In: *Scientific reports* 6.1 (2016), p. 19279.

[2489] Celine Montavon et al. "Identification of a new circulating recombinant form of HIV type 1, CRF11-cpx, involving subtypes A, G, J, and CRF01-AE, in Central Africa". In: *AIDS research and human retroviruses* 18.3 (2002), pp. 231–236.

[2490] Jiangqin Zhao et al. "Identification and genetic characterization of a novel CRF22_01A1 recombinant form of HIV type 1 in Cameroon". In: *AIDS research and human retroviruses* 26.9 (2010), pp. 1033–1045.

[2491] Jiangqin Zhao et al. "CRF22_01A1 is involved in the emergence of new HIV-1 recombinants in Cameroon". In: *JAIDS Journal of Acquired Immune Deficiency Syndromes* 60.4 (2012), pp. 344–350.

[2492] Sayed-Hamidreza Mozhgani et al. "CRF35-AD as the main circulating genotype of human immunodeficiency virus type 1 infection in Iran: a phylogenetic and demographic-based study". In: *Intervirology* 60.4 (2018), pp. 144–148.

[2493] Francesco Saladini et al. "Near full-length sequence analysis of HIV type 1 BF recombinants from Italy". In: *AIDS research and human retroviruses* 28.3 (2012), pp. 299–303.

[2494] Olivier ter Brake, Karin J von Eije, and Ben Berkhout. "Probing the sequence space available for HIV-1 evolution". In: *Aids* 22.14 (2008), pp. 1875–1877.

[2495] Martine Peeters and Paul M Sharp. "Genetic diversity of HIV-1: the moving target". In: *Aids* 14 (2000), S129–S140.

[2496] Ben Berkhout, Atze T Das, and Nancy Beerens. "HIV-1 RNA editing, hypermutation, and error-prone reverse transcription". In: *Science* 292.5514 (2001), pp. 7–7.

[2497] Catherine S Adamson and Eric O Freed. "Novel approaches to inhibiting HIV-1 replication". In: *Antiviral research* 85.1 (2010), pp. 119–141.

[2498] Yuntao Wu. "HIV-1 gene expression: lessons from provirus and non-integrated DNA". In: *Retrovirology* 1 (2004), pp. 1–10.

[2499] David L Robertson, Beatrice H Hahn, and Paul M Sharp. "Recombination in AIDS viruses". In: *Journal of molecular evolution* 40 (1995), pp. 249–259.

[2500] Adewunmi Onafuwa-Nuga and Alice Telesnitsky. "The remarkable frequency of human immunodeficiency virus type 1 genetic recombination". In: *Microbiology and Molecular Biology Reviews* 73.3 (2009), pp. 451–480.

[2501] Ming Zhang et al. "The role of recombination in the emergence of a complex and dynamic HIV epidemic". In: *Retrovirology* 7.1 (2010), pp. 1–15.

[2502] Stacy A Seibert et al. "Natural selection on the gag, pol, and env genes of human immunodeficiency virus 1 (HIV-1)." In: *Molecular biology and evolution* 12.5 (1995), pp. 803–813.

[2503] Andrew J Leigh Brown and Douglas D Richman. "HIV-1: gambling on the evolution of drug resistance?" In: *Nature medicine* 3.3 (1997), pp. 268–271.

[2504] Thomas Unterthiner et al. "Detection of viral sequence fragments of HIV-1 subfamilies yet unknown". In: *BMC bioinformatics* 12 (2011), pp. 1–13.

[2505] Feng Gao et al. "Evidence of two distinct subsubtypes within the HIV-1 subtype A radiation". In: *AIDS research and human retroviruses* 17.8 (2001), pp. 675–688.

[2506] Rob DeSalle, Michael Tessler, and Jeffrey Rosenfeld. *Phylogenomics: a primer*. CRC Press, 2020.

[2507] Alexander E Gorbalenya and Chris Lauber. "Phylogeny of viruses". In: *Reference Module in Biomedical Sciences* (2017).

[2508] Jemma L Geoghegan and Edward C Holmes. "The phylogenomics of evolving virus virulence". In: *Nature Reviews Genetics* 19.12 (2018), pp. 756–769.

[2509] Francine E McCutchan et al. "HIV Type 1 Circulating Recombinant Form CRF09_cpx from West Africa combines subtypes A, F, G, and may share ancestors with CRF02_AG and Z321". In: *AIDS research and human retroviruses* 20.8 (2004), pp. 819–826.

[2510] Tomasz Stadejek et al. "Definition of subtypes in the European genotype of porcine reproductive and respiratory syndrome virus: nucleocapsid characteristics and geographical distribution in Europe". In: *Archives of virology* 153 (2008), pp. 1479–1488.

[2511] Saverio Caini et al. "Clinical characteristics and severity of influenza infections by virus type, subtype, and lineage: a systematic literature review". In: *Influenza and other respiratory viruses* 12.6 (2018), pp. 780–792.

[2512] Mahmoud Reza Pourkarim et al. "Are hepatitis B virus subgenotypes defined accurately?" In: *Journal of clinical virology* 47.4 (2010), pp. 356–360.

[2513] Ioanna Kousiappa, David AMC Van De Vijver, and Leondios G Kostrikis. "Near full-length genetic analysis of HIV sequences derived from Cyprus: evidence of a highly polyphyletic and evolving infection". In: *AIDS research and human retroviruses* 25.8 (2009), pp. 727–740.

[2514] Mary A Rodgers et al. "Sensitive next-generation sequencing method reveals deep genetic diversity of HIV-1 in the Democratic Republic of the Congo". In: *Journal of virology* 91.6 (2017), e01841–16.

[2515] Lia Van Der Hoek et al. "Characterization of an HIV-1 group M variant that is distinct from the known subtypes". In: *AIDS research and human retroviruses* 23.3 (2007), pp. 466–470.

[2516] Mathieu Quesnel-Vallières et al. "Novel HIV-1 recombinant forms in antenatal cohort, Montreal, Quebec, Canada". In: *Emerging Infectious Diseases* 17.2 (2011), p. 271.

[2517] Nicole Vidal et al. "Genetic characterization of eight full-length HIV Type 1 genomes from the Democratic Republic of Congo (DRC) reveal a new subsubtype, A5, in the A radiation that predominates in the recombinant structure of CRF26_A5U". In: *AIDS research and human retroviruses* 25.8 (2009), pp. 823–832.

[2518] Pavel B Baryshev, Vladislav V Bogachev, and Natalya M Gashnikova. "HIV-1 genetic diversity in Russia: CRF63_02A1, a new HIV Type 1 genetic variant spreading in Siberia". In: *AIDS research and human retroviruses* 30.6 (2014), pp. 592–597.

[2519] Ioanna Kousiappa et al. "Molecular characterization of HIV Type 1 strains from newly diagnosed patients in Cyprus (2007–2009) recovers multiple clades including unique recombinant strains and lack of transmitted drug resistance". In: *AIDS research and human retroviruses* 27.11 (2011), pp. 1183–1199.

[2520] Ana M Sanchez et al. "Development of a contemporary globally diverse HIV viral panel by the EQAPOL program". In: *Journal of immunological methods* 409 (2014), pp. 117–130.

[2521] Celine Montavon et al. "CRF06-cpx: a new circulating recombinant form of HIV-1 in West Africa involving subtypes A, G, K, and J." In: *Journal of acquired immune deficiency syndromes (1999)* 29.5 (2002), pp. 522–530.

[2522] Rebecca LR Powell et al. "Circulating recombinant form (CRF) 37_cpx: an old strain in Cameroon composed of diverse, genetically distant lineages of subtypes A and G". In: *AIDS Research and Human retroviruses* 23.7 (2007), pp. 923–933.

[2523] Jiangqin Zhao et al. "Identification, Genetic Characterization and Validation of Highly Diverse HIV-1 Viruses for Reference Panel Development". In: *Viruses* 13.7 (2021), p. 1417.

[2524] Jean K Carr et al. "Diverse BF recombinants have spread widely since the introduction of HIV-1 into South America". In: *Aids* 15.15 (2001), F41–F47.

[2525] Maja M Lunar et al. "HIV-1 unique recombinant forms identified in Slovenia and their characterization by near full-length genome sequencing". In: *Viruses* 12.1 (2020), p. 63.

[2526] Marie Leoz et al. "Characterization of CRF56_cpx, a new circulating B/CRF02/G recombinant form identified in MSM in France". In: *Aids* 27.14 (2013), pp. 2309–2312.

[2527] Jiangqin Zhao et al. "Identification and genetic characterization of a novel CRF22_01A1 recombinant form of HIV type 1 in Cameroon". In: *AIDS research and human retroviruses* 26.9 (2010), pp. 1033–1045.

[2528] Diana D Huang et al. "Complex mosaic composition of near full-length genomes of two NED (NIH-ENVA-DOD) subtype panel HIV Type 1 strains, BCF-Dioum and BCF-Kita, originating from the Democratic Republic of Congo (DRC)". In: *AIDS research and human retroviruses* 25.10 (2009), pp. 1039–1043.

[2529] Marcel Tongo et al. "Near full-length HIV type 1M genomic sequences from Cameroon: Evidence of early diverging under-sampled lineages in the country". In: *Evolution, medicine, and public health* 2015.1 (2015), pp. 254–265.

[2530] Michael M Thomson et al. "Identification of a novel HIV-1 complex circulating recombinant form (CRF18_cpx) of Central African origin in Cuba". In: *Aids* 19.11 (2005), pp. 1155–1163.

[2531] Fabien R Niama et al. "CRF45_AKU, a circulating recombinant from Central Africa, is probably the common ancestor of HIV Type 1 MAL and HIV Type 1 NOGIL". In: *AIDS research and human retroviruses* 25.12 (2009), pp. 1345–1353.

[2532] Manuel Gómez-Carrillo et al. "Drug resistance testing provides evidence of the globalization of HIV type 1: a new circulating recombinant form". In: *AIDS research and human retroviruses* 20.8 (2004), pp. 885–888.

[2533] Unchalee Visawapoka et al. "Circulating and unique recombinant forms of HIV type 1 containing subsubtype A2". In: *AIDS Research & Human Retroviruses* 22.7 (2006), pp. 695–702.

[2534] D Paraskevis et al. "Molecular characterization of a complex, recombinant Human Immunodeficiency Virus Type 1 (HIV-1) isolate (A/G/J/K/?): evidence to support the existence of a novel HIV-1 subtype". In: *Journal of General Virology* 82.10 (2001), pp. 2509–2514.

[2535] Rebecca LR Powell et al. "Identification of a novel circulating recombinant form (CRF) 36_cpx in Cameroon that combines two CRFs (01_AE and 02_AG) with ancestral lineages of subtypes A and G". In: *AIDS Research and Human retroviruses* 23.8 (2007), pp. 1008–1019.

[2536] Ka-Cheung Luk et al. "Utility of metagenomic next-generation sequencing for characterization of HIV and human pegivirus diversity". In: *PloS one* 10.11 (2015), e0141723.

[2537] Nicole Vidal et al. "Characterization of an old complex circulating recombinant form, CRF27_cpx, originating from the Democratic Republic of Congo (DRC) and circulating in France". In: *AIDS research and human retroviruses* 24.2 (2008), pp. 315–321.

[2538] Ming Zhang et al. "HIV type 1 CRF13_cpx revisited: identification of a new sequence from Cameroon and signal for subsubtype J2". In: *AIDS Research & Human Retroviruses* 21.11 (2005), pp. 955–960.

[2539] Karin Wilbe et al. "Identification of two CRF11-cpx genomes and two preliminary representatives of a new circulating recombinant form (CRF13-cpx) of HIV Type 1 in Cameroon". In: *AIDS research and human retroviruses* 18.12 (2002), pp. 849–856.

[2540] Ricardo Camacho. "The significance of subtype-related genetic variability: controversies and unanswered questions". In: *Antiretroviral resistance in clinical practice* (2006).

[2541] Jürgen Weidner et al. "An improved PCR method for detection of HIV-1 proviral DNA of a wide range of subtypes and recombinant forms circulating globally". In: *Journal of virological methods* 172.1-2 (2011), pp. 22–26.

[2542] Mika O Salminen et al. "Full-length sequence of an Ethiopian human immunodeficiency virus type 1 (HIV-1) isolate of genetic subtype C". In: *AIDS research and human retroviruses* 12.14 (1996), pp. 1329–1339.

[2543] Feng Gao et al. "A comprehensive panel of near-full-length clones and reference sequences for non-subtype B isolates of human immunodeficiency virus type 1". In: *Journal of virology* 72.7 (1998), pp. 5680–5698.

[2544] Tiina Laukkanen et al. "Virtually full-length sequences of HIV type 1 subtype J reference strains". In: *AIDS research and human retroviruses* 15.3 (1999), pp. 293–297.

[2545] Karine Triques et al. "Near-full-length genome sequencing of divergent African HIV type 1 subtype F viruses leads to the identification of a new HIV type 1 subtype designated K". In: *AIDS research and human retroviruses* 16.2 (2000), pp. 139–151.

[2546] Jean K Carr et al. "Full genome sequences of human immunodeficiency virus type 1 subtypes G and A/G intersubtype recombinants". In: *Virology* 247.1 (1998), pp. 22–31.

[2547] Ana B Abecasis et al. "Recombination confounds the early evolutionary history of human immunodeficiency virus type 1: subtype G is a circulating recombinant form". In: *Journal of virology* 81.16 (2007), pp. 8543–8551.

[2548] Katherine A Lau and Justin JL Wong. "Current trends of HIV recombination worldwide". In: *Infectious disease reports* 5.Suppl 1 (2013).

[2549] Fabio Zanini et al. "Population genomics of intrapatient HIV-1 evolution". In: *Elife* 4 (2015), e11282.

[2550] Megan C Steain et al. "HIV-1 co-infection, superinfection and recombination". In: *Sexual health* 1.4 (2004), pp. 239–250.

[2551] Todd M Allen and Marcus Altfeld. "HIV-1 superinfection". In: *Journal of allergy and clinical immunology* 112.5 (2003), pp. 829–835.

[2552] Antoinette C Van der Kuyl and Marion Cornelissen. "Identifying HIV-1 dual infections". In: *Retrovirology* 4 (2007), pp. 1–12.

[2553] Ricardo S Diaz et al. "Dual human immunodeficiency virus type 1 infection and recombination in a dually exposed transfusion recipient. The Transfusion Safety Study Group". In: *Journal of Virology* 69.6 (1995), pp. 3273–3281.

[2554] Monica Sala, Eric Pelletier, and Simon Wain-Hobson. "HIV-1 gpl20 sequences from a doubly infected drug user". In: *AIDS research and human retroviruses* 11.5 (1995), pp. 653–655.

[2555] Andrew W Artenstein et al. "Dual infection with human immunodeficiency virus type 1 of distinct envelope subtypes in humans". In: *Journal of Infectious Diseases* 171.4 (1995), pp. 805–810.

[2556] Graziella Becker-Pergola et al. "Identification of diverse HIV type 1 subtypes and dual HIV type 1 infection in pregnant Ugandan women". In: *AIDS research and human retroviruses* 16.12 (2000), pp. 1099–1104.

[2557] JL Mellquist et al. "Dual transmission of subtype A and D HIV type 1 viruses from a Ugandan woman to her infant". In: *AIDS research and human retroviruses* 15.2 (1999), pp. 217–221.

[2558] Luiz M Janini et al. "Identification of single and dual infections with distinct subtypes of human immunodeficiency virus type 1 by using restriction fragment length polymorphism analysis". In: *Virus Genes* 13 (1996), pp. 69–81.

[2559] Luiz M Janini et al. "Horizontal and vertical transmission of human immunodeficiency virus type 1 dual infections caused by viruses of subtypes B and C". In: *Journal of Infectious Diseases* 177.1 (1998), pp. 227–231.

[2560] Artur Ramos et al. "Dual and recombinant infections: an integral part of the HIV-1 epidemic in Brazil." In: *Emerging infectious diseases* 5.1 (1999), p. 65.

[2561] Idhaliz Flores et al. "HIV-1 subtype F in single and dual infections in Puerto Rico: A potential sentinel site for monitoring novel genetic HIV variants in North America." In: *Emerging Infectious Diseases* 5.3 (1999), p. 481.

[2562] S Butto et al. "Dual infection with different strains of the same HIV-1 subtype". In: *Aids* 11.5 (1997), pp. 694–696.

[2563] Adam Trickey et al. "Life expectancy of people with HIV on long-term antiretroviral therapy in Europe and North America: a cohort study". In: *The Lancet Healthy Longevity* 3 (2022), S2.

[2564] Arantxa Rosales Ruiz. "Las enfermedades de transmisión sexual, un problema para la sociedad actual." In: (2020).

[2565] Min Du et al. "Increasing incidence rates of sexually transmitted infections from 2010 to 2019: an analysis of temporal trends by geographical regions and age groups from the 2019 Global Burden of Disease Study". In: *BMC infectious diseases* 22.1 (2022), p. 574.

[2566] Jérémy Zeggagh et al. “Incidence and risk factors for recurrent sexually transmitted infections among MSM on HIV pre-exposure prophylaxis”. In: *Aids* 36.8 (2022), pp. 1129–1134.

[2567] Bluma Brenner et al. “Persistence of multidrug-resistant HIV-1 in primary infection leading to superinfection”. In: *Aids* 18.12 (2004), pp. 1653–1660.

[2568] Anne Piantadosi et al. “Chronic HIV-1 infection frequently fails to protect against superinfection”. In: *PLoS pathogens* 3.11 (2007), e177.

[2569] Concepción Casado et al. “Coinfection and superinfection in patients with long-term, nonprogressive HIV-1 disease”. In: *Journal of Infectious Diseases* 196.6 (2007), pp. 895–899.

[2570] Bhavna H Chohan, Anne Piantadosi, and Julie Overbaugh. “HIV-1 superinfection and its implications for vaccine design”. In: *Current HIV Research* 8.8 (2010), pp. 596–601.

[2571] Marcus Altfeld et al. “HIV-1 superinfection despite broad CD8+ T-cell responses containing replication of the primary virus”. In: *Nature* 420.6914 (2002), pp. 434–439.

[2572] Stephanie Jost et al. “A patient with HIV-1 superinfection”. In: *New England Journal of Medicine* 347.10 (2002), pp. 731–736.

[2573] Kersten K Koelsch et al. “Clade B HIV-1 superinfection with wild-type virus after primary infection with drug-resistant clade B virus”. In: *Aids* 17.7 (2003), F11–F16.

[2574] Artur Ramos et al. “Intersubtype human immunodeficiency virus type 1 superinfection following seroconversion to primary infection in two injection drug users”. In: *Journal of Virology* 76.15 (2002), pp. 7444–7452.

[2575] Kimber L Gross, Travis C Porco, and Robert M Grant. “HIV-1 superinfection and viral diversity”. In: *Aids* 18.11 (2004), pp. 1513–1520.

[2576] Daniel J Sheward et al. “HIV-1 superinfection resembles primary infection”. In: *The Journal of infectious diseases* 212.6 (2015), pp. 904–908.

[2577] Derek J Chan. “HIV-1 superinfection: evidence and impact”. In: *Current HIV Research* 2.3 (2004), pp. 271–274.

[2578] Guowei Fang et al. “Recombination following superinfection by HIV-1”. In: *Aids* 18.2 (2004), pp. 153–159.

[2579] Sandra E Chaudron et al. “A systematic molecular epidemiology screen reveals numerous HIV-1 superinfections in the Swiss HIV Cohort Study.” In: *The journal of infectious diseases* 226.7 (2022), pp. 1256–1266.

[2580] Hui Zhang et al. “HIV-1–Specific Immunodominant T-Cell Responses Drive the Dynamics of HIV-1 Recombination Following Superinfection”. In: *Frontiers in immunology* 12 (2022), p. 820628.

[2581] Maria Pernas et al. “A dual superinfection and recombination within HIV-1 subtype B 12 years after primoinfection”. In: *JAIDS Journal of Acquired Immune Deficiency Syndromes* 42.1 (2006), pp. 12–18.

[2582] Fransje A Koning et al. "Dynamics of HIV type 1 recombination following superinfection". In: *AIDS research and human retroviruses* 29.6 (2013), pp. 963–970.

[2583] Ahmed I Abulsoud et al. "Mutations in SARS-CoV-2: Insights on structure, variants, vaccines, and biomedical interventions". In: *Biomedicine & Pharmacotherapy* 157 (2023), p. 113977.

[2584] David A Jamison Jr et al. "A comprehensive SARS-CoV-2 and COVID-19 review, Part 1: Intracellular overdrive for SARS-CoV-2 infection". In: *European Journal of Human Genetics* 30.8 (2022), pp. 889–898.

[2585] Dong Yang and Julian L Leibowitz. "The structure and functions of coronavirus genomic 3' and 5' ends". In: *Virus research* 206 (2015), pp. 120–133.

[2586] Craig E Nelson, Bradley M Hersh, and Sean B Carroll. "The regulatory content of intergenic DNA shapes genome architecture". In: *Genome biology* 5 (2004), pp. 1–15.

[2587] SM Wilson and JCS Clegg. "Sequence analysis of the S RNA of the African arenavirus Mopeia: an unusual secondary structure feature in the intergenic region". In: *Virology* 180.2 (1991), pp. 543–552.

[2588] Kayvan Etebari et al. "Identification of Aedes aegypti long intergenic non-coding RNAs and their association with Wolbachia and dengue virus infection". In: *PLoS neglected tropical diseases* 10.10 (2016), e0005069.

[2589] Xuming Zhang and MM Lai. "Interactions between the cytoplasmic proteins and the intergenic (promoter) sequence of mouse hepatitis virus RNA: correlation with the amounts of subgenomic mRNA transcribed". In: *Journal of virology* 69.3 (1995), pp. 1637–1644.

[2590] Daniele Mercatelli and Federico M Giorgi. "Geographic and genomic distribution of SARS-CoV-2 mutations". In: *Frontiers in microbiology* 11 (2020), p. 1800.

[2591] Zhichao Miao et al. "Secondary structure of the SARS-CoV-2 5'-UTR". In: *RNA biology* 18.4 (2021), pp. 447–456.

[2592] Rachel O Niederer et al. "Direct analysis of ribosome targeting illuminates thousand-fold regulation of translation initiation". In: *Cell systems* 13.3 (2022), pp. 256–264.

[2593] Shiori Ueno et al. "Eight-amino-acid sequence at the N-terminus of SARS-CoV-2 nsp1 is involved in stabilizing viral genome replication". In: *Virology* (2024), p. 110068.

[2594] Moumita Mukherjee and Srikanta Goswami. "Global cataloguing of variations in untranslated regions of viral genome and prediction of key host RNA binding protein-microRNA interactions modulating genome stability in SARS-CoV-2". In: *PloS one* 15.8 (2020), e0237559.

[2595] Rohit Verma et al. "RNA-Protein interaction analysis of SARS-CoV-2 5' and 3' untranslated regions reveals a role of lysosome-associated membrane protein-2A during viral infection". In: *Msystems* 6.4 (2021), e00643–21.

[2596] Manijeh Mohammadi-Dehcheshmeh et al. "A transcription regulatory sequence in the 5' untranslated region of SARS-CoV-2 is vital for virus replication with an altered evolutionary pattern against human inhibitory microRNAs". In: *Cells* 10.2 (2021), p. 319.

[2597] Yaara Finkel et al. "The coding capacity of SARS-CoV-2". In: *Nature* 589.7840 (2021), pp. 125–130.

[2598] Tal Fisher et al. "Parsing the role of NSP1 in SARS-CoV-2 infection". In: *Cell Reports* 39.11 (2022).

[2599] Resal Raj. "Analysis of non-structural proteins, NSPs of SARS-CoV-2 as targets for computational drug designing". In: *Biochemistry and biophysics reports* 25 (2021), p. 100847.

[2600] Zhang Xu et al. "SARS-CoV-2 impairs interferon production via NSP2-induced repression of mRNA translation". In: *Proceedings of the National Academy of Sciences* 119.32 (2022), e2204539119.

[2601] Lee A Armstrong et al. "Biochemical characterization of protease activity of Nsp3 from SARS-CoV-2 and its inhibition by nanobodies". In: *PloS one* 16.7 (2021), e0253364.

[2602] Chew Theng Lim et al. "Identifying SARS-CoV-2 antiviral compounds by screening for small molecule inhibitors of Nsp3 papain-like protease". In: *Biochemical Journal* 478.13 (2021), pp. 2517–2531.

[2603] Maryline Santerre et al. "Why do SARS-CoV-2 NSPs rush to the ER?" In: *Journal of neurology* 268 (2021), pp. 2013–2022.

[2604] Xiaoyuan Lin et al. "The NSP4 T492I mutation increases SARS-CoV-2 infectivity by altering non-structural protein cleavage". In: *Cell Host & Microbe* 31.7 (2023), pp. 1170–1184.

[2605] Niti Yashvardhini, Amit Kumar, and Deepak Kumar Jha. "Analysis of SARS-CoV-2 mutations in the main viral protease (NSP5) and its implications on the vaccine designing strategies". In: *Vacunas* 23 (2022), S1–S13.

[2606] Yongzhen Liu et al. "SARS-CoV-2 Nsp5 demonstrates two distinct mechanisms targeting RIG-I and MAVS to evade the innate immune response". In: *MBio* 12.5 (2021), pp. 10–1128.

[2607] Simona Ricciardi et al. "The role of NSP6 in the biogenesis of the SARS-CoV-2 replication organelle". In: *Nature* 606.7915 (2022), pp. 761–768.

[2608] Domenico Benvenuto et al. "Evolutionary analysis of SARS-CoV-2: how mutation of Non-Structural Protein 6 (NSP6) could affect viral autophagy". In: *Journal of Infection* 81.1 (2020), e24–e27.

[2609] Mahamaya Biswal et al. "Two conserved oligomer interfaces of NSP7 and NSP8 underpin the dynamic assembly of SARS-CoV-2 RdRP". In: *Nucleic acids research* 49.10 (2021), pp. 5956–5966.

[2610] Quan Wang et al. "Structural basis for RNA replication by the SARS-CoV-2 polymerase". In: *Cell* 182.2 (2020), pp. 417–428.

[2611] Jian Deng et al. “SARS-CoV-2 NSP7 inhibits type I and III IFN production by targeting the RIG-I/MDA5, TRIF, and STING signaling pathways”. In: *Journal of Medical Virology* 95.3 (2023), e28561.

[2612] Qi Peng et al. “Structural and biochemical characterization of the nsp12-nsp7-nsp8 core polymerase complex from ŧSARS-CoV-2”. In: *Cell reports* 31.11 (2020).

[2613] Shamlan MS Reshamwala et al. “Mutations in SARS-CoV-2 nsp7 and nsp8 proteins and their predicted impact on replication/transcription complex structure”. In: *Journal of medical virology* 93.7 (2021), pp. 4616–4619.

[2614] Xiaolin Zhang et al. “SARS-CoV-2 Nsp8 suppresses MDA5 antiviral immune responses by impairing TRIM4-mediated K63-linked polyubiquitination”. In: *Plos Pathogens* 19.11 (2023), e1011792.

[2615] Heiko Slanina et al. “Coronavirus replication–transcription complex: Vital and selective NMPylation of a conserved site in nsp9 by the NiRAN-RdRp subunit”. In: *Proceedings of the National Academy of Sciences* 118.6 (2021), e2022310118.

[2616] Yihua Zhang et al. “SARS-COV-2 protein NSP9 promotes cytokine production by targeting TBK1”. In: *Frontiers in Immunology* 14 (2023).

[2617] Sheng Lin et al. “Crystal structure of SARS-CoV-2 nsp10 bound to nsp14-ExoN domain reveals an exoribonuclease with both structural and functional integrity”. In: *Nucleic acids research* 49.9 (2021), pp. 5382–5392.

[2618] Sheng Lin et al. “Crystal structure of SARS-CoV-2 nsp10/nsp16 2’-O-methylase and its implication on antiviral drug design”. In: *Signal transduction and targeted therapy* 5.1 (2020), p. 131.

[2619] Kundlik Gadhave et al. “Conformational dynamics of 13 amino acids long NSP11 of SARS-CoV-2 under membrane mimetics and different solvent conditions”. In: *Microbial Pathogenesis* 158 (2021), p. 105041.

[2620] Md Jahirul Islam et al. “A review on structural, non-structural, and accessory proteins of SARS-CoV-2: Highlighting drug target sites”. In: *Immunobiology* 228.1 (2023), p. 152302.

[2621] Ashleigh Shannon et al. “Remdesivir and SARS-CoV-2: Structural requirements at both nsp12 RdRp and nsp14 Exonuclease active-sites”. In: *Antiviral research* 178 (2020), p. 104793.

[2622] Joseph A Newman et al. “Structure, mechanism and crystallographic fragment screening of the SARS-CoV-2 NSP13 helicase”. In: *Nature communications* 12.1 (2021), p. 4848.

[2623] Natacha S Ogando et al. “The enzymatic activity of the nsp14 exoribonuclease is critical for replication of MERS-CoV and SARS-CoV-2”. In: *Journal of virology* 94.23 (2020), pp. 10–1128.

[2624] Sheng Lin et al. “Crystal structure of SARS-CoV-2 nsp10 bound to nsp14-ExoN domain reveals an exoribonuclease with both structural and functional integrity”. In: *Nucleic acids research* 49.9 (2021), pp. 5382–5392.

[2625] Chao Liu et al. “Potential treatment with Chinese and Western medicine targeting NSP14 of SARS-CoV-2”. In: *Journal of pharmaceutical analysis* 11.3 (2021), pp. 272–277.

[2626] Monica C Pillon et al. “Cryo-EM structures of the SARS-CoV-2 endoribonuclease Nsp15 reveal insight into nuclease specificity and dynamics”. In: *Nature communications* 12.1 (2021), p. 636.

[2627] Clayton J Otter et al. “SARS-CoV-2 nsp15 endoribonuclease antagonizes dsRNA-induced antiviral signaling”. In: *Proceedings of the National Academy of Sciences* 121.15 (2024), e2320194121.

[2628] Petra Krafcikova et al. “Structural analysis of the SARS-CoV-2 methyltransferase complex involved in RNA cap creation bound to sinefungin”. In: *Nature communications* 11.1 (2020), p. 3717.

[2629] Sheng Lin et al. “Crystal structure of SARS-CoV-2 nsp10/nsp16 2-O-methylase and its implication on antiviral drug design”. In: *Signal transduction and targeted therapy* 5.1 (2020), p. 131.

[2630] Neha Vithani et al. “SARS-CoV-2 Nsp16 activation mechanism and a cryptic pocket with pan-coronavirus antiviral potential”. In: *Biophysical Journal* 120.14 (2021), pp. 2880–2889.

[2631] Jun Zhang et al. “Structure of SARS-CoV-2 spike protein”. In: *Current opinion in virology* 50 (2021), pp. 173–182.

[2632] William T Harvey et al. “SARS-CoV-2 variants, spike mutations and immune escape”. In: *Nature Reviews Microbiology* 19.7 (2021), pp. 409–424.

[2633] Sayantan Laha et al. “Characterizations of SARS-CoV-2 mutational profile, spike protein stability and viral transmission”. In: *Infection, genetics and evolution* 85 (2020), p. 104445.

[2634] Sanu Augustine et al. “Targeting viral ORF3a protein: A new approach to mitigate COVID-19 induced Immune cell apoptosis and associated respiratory complications”. In: *Advanced Pharmaceutical Bulletin* 13.4 (2023), p. 678.

[2635] Yujie Ren et al. “The ORF3a protein of SARS-CoV-2 induces apoptosis in cells”. In: *Cellular & molecular immunology* 17.8 (2020), pp. 881–883.

[2636] Jiantao Zhang et al. “Understanding the role of SARS-CoV-2 ORF3a in viral pathogenesis and COVID-19”. In: *Frontiers in microbiology* 13 (2022), p. 854567.

[2637] Jiantao Zhang et al. “Endoplasmic reticulum-associated SARS-CoV-2 ORF3a elicits heightened cytopathic effects despite robust ER-associated degradation”. In: *Mbio* 15.1 (2024), e03030–23.

[2638] Venkata S Mandala et al. “Structure and drug binding of the SARS-CoV-2 envelope protein transmembrane domain in lipid bilayers”. In: *Nature structural & molecular biology* 27.12 (2020), pp. 1202–1208.

[2639] Sébastien Eymieux et al. "Ultrastructural modifications induced by SARS-CoV-2 in Vero cells: a kinetic analysis of viral factory formation, viral particle morphogenesis and virion release". In: *Cellular and Molecular Life Sciences* 78 (2021), pp. 3565–3576.

[2640] Daniel K Beyer and Adriana Forero. "Mechanisms of antiviral immune evasion of SARS-CoV-2". In: *Journal of molecular biology* 434.6 (2022), p. 167265.

[2641] Ariel Shepley-McTaggart et al. "SARS-CoV-2 Envelope (E) protein interacts with PDZ-domain-2 of host tight junction protein ZO1". In: *PLoS One* 16.6 (2021), e0251955.

[2642] Zhikuan Zhang et al. "Structure of SARS-CoV-2 membrane protein essential for virus assembly". In: *Nature communications* 13.1 (2022), p. 4399.

[2643] Bertrand Boson et al. "The SARS-CoV-2 envelope and membrane proteins modulate maturation and retention of the spike protein, allowing assembly of virus-like particles". In: *Journal of Biological Chemistry* 296 (2021).

[2644] Shan Lu et al. "The SARS-CoV-2 nucleocapsid phosphoprotein forms mutually exclusive condensates with RNA and the membrane-associated M protein". In: *Nature communications* 12.1 (2021), p. 502.

[2645] Yi Zheng et al. "Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) membrane (M) protein inhibits type I and III interferon production by targeting RIG-I/MDA-5 signaling". In: *Signal transduction and targeted therapy* 5.1 (2020), p. 299.

[2646] Chun-Kit Yuen et al. "SARS-CoV-2 nsp13, nsp14, nsp15 and orf6 function as potent interferon antagonists". In: *Emerging microbes & infections* 9.1 (2020), pp. 1418–1428.

[2647] Jin-Gu Lee et al. "Characterization of SARS-CoV-2 proteins reveals Orf6 pathogenicity, subcellular localization, host interactions and attenuation by Selinexor". In: *Cell & bioscience* 11 (2021), pp. 1–12.

[2648] Gianni Gori Savellini et al. "Nucleopore traffic is hindered by SARS-CoV-2 ORF6 protein to efficiently suppress IFN-β and IL-6 secretion". In: *Viruses* 14.6 (2022), p. 1273.

[2649] Amin Addetia et al. "SARS-CoV-2 ORF6 disrupts bidirectional nucleocytoplasmic transport through interactions with Rae1 and Nup98". In: *MBio* 12.2 (2021), pp. 10–1128.

[2650] Zengguo Cao et al. "Ubiquitination of SARS-CoV-2 ORF7a promotes antagonism of interferon response". In: *Cellular & molecular immunology* 18.3 (2021), pp. 746–748.

[2651] Zhixin Liu et al. "Ubiquitination of SARS-CoV-2 ORF7a prevents cell death induced by recruiting BclXL to activate ER stress". In: *Microbiology Spectrum* 10.6 (2022), e01509–22.

[2652] Najla Arshad et al. "SARS-CoV-2 accessory proteins ORF7a and ORF3a use distinct mechanisms to down-regulate MHC-I surface expression". In: *Proceedings of the National Academy of Sciences* 120.1 (2023), e2208525120.

[2653] Peili Hou et al. "The ORF7a protein of SARS-CoV-2 initiates autophagy and limits autophagosome-lysosome fusion via degradation of SNAP29 to promote virus replication". In: *Autophagy* 19.2 (2023), pp. 551–569.

[2654] Natalia Redondo et al. "SARS-CoV-2 accessory proteins in viral pathogenesis: knowns and unknowns". In: *Frontiers in immunology* 12 (2021), p. 708264.

[2655] Maya Shemesh et al. "SARS-CoV-2 suppresses IFNβ production mediated by NSP1, 5, 6, 15, ORF6 and ORF7b but does not suppress the effects of added interferon". In: *PLoS pathogens* 17.8 (2021), e1009800.

[2656] Jing Zhang et al. "A systemic and molecular study of subcellular localization of SARS-CoV-2 proteins". In: *Signal transduction and targeted therapy* 5.1 (2020), p. 269.

[2657] Ruiping Yang et al. "SARS-CoV-2 accessory protein ORF7b mediates tumor necrosis factor-α-induced apoptosis in cells". In: *Frontiers in microbiology* 12 (2021), p. 654709.

[2658] Matthew D Park. "Immune evasion via SARS-CoV-2 ORF8 protein?" In: *Nature Reviews Immunology* 20.7 (2020), pp. 408–408.

[2659] Smita Vinjamuri, Lenong Li, and Marlene Bouvier. "SARS-CoV-2 ORF8: one protein, seemingly one structure, and many functions". In: *Frontiers in immunology* 13 (2022), p. 1035559.

[2660] Rushikesh Deshpande and Chunbin Zou. "Programmed cell death in SARS-CoV-2 infection: A short review". In: *Journal of Respiration* 1.4 (2021), pp. 223–228.

[2661] Sk Sarif Hassan et al. "A unique view of SARS-CoV-2 through the lens of ORF8 protein". In: *Computers in biology and medicine* 133 (2021), p. 104380.

[2662] Tianyi Gao et al. "Identification and functional analysis of the SARS-COV-2 nucleocapsid protein". In: *BMC microbiology* 21 (2021), pp. 1–10.

[2663] Qiaozhen Ye et al. "Architecture and self-assembly of the SARS-CoV-2 nucleocapsid protein". In: *Protein Science* 29.9 (2020), pp. 1890–1901.

[2664] Zhihua Bai et al. "The SARS-CoV-2 nucleocapsid protein and its role in viral structure, biological functions, and a potential target for drug or vaccine mitigation". In: *Viruses* 13.6 (2021), p. 1115.

[2665] Pan Pan et al. "SARS-CoV-2 N protein promotes NLRP3 inflammasome activation to induce hyperinflammation". In: *Nature communications* 12.1 (2021), p. 4664.

[2666] Lulu Han et al. "SARS-CoV-2 ORF10 antagonizes STING-dependent interferon activation and autophagy". In: *Journal of medical virology* 94.11 (2022), pp. 5174–5188.

[2667] Xingyu Li et al. "SARS-CoV-2 ORF10 suppresses the antiviral innate immune response by degrading MAVS through mitophagy". In: *Cellular & molecular immunology* 19.1 (2022), pp. 67–78.

[2668] Liying Wang et al. "SARS-CoV-2 ORF10 impairs cilia by enhancing CUL2ZYG11B activity". In: *Journal of Cell Biology* 221.7 (2022), e202108015.

[2669] Sk Sarif Hassan et al. "Notable sequence homology of the ORF10 protein introspects the architecture of SARS-CoV-2". In: *International Journal of Biological Macromolecules* 181 (2021), pp. 801–809.

[2670] Maria Romano et al. "A structural view of SARS-CoV-2 RNA replication machinery: RNA synthesis, proofreading and final capping". In: *Cells* 9.5 (2020), p. 1267.

[2671] Pinghua Liu et al. "Functional analysis of the stem loop S3 and S4 structures in the coronavirus 3 UTR". In: *Virology* 443.1 (2013), pp. 40–47.

[2672] Peter Simmonds. "Pervasive RNA secondary structure in the genomes of SARS-CoV-2 and other coronaviruses". In: *MBio* 11.6 (2020), pp. 10–1128.

[2673] Valeria Lulla et al. "Targeting the conserved stem loop 2 motif in the SARS-CoV-2 genome". In: *Journal of virology* 95.14 (2021), pp. 10–1128.

[2674] Joseph A Makowski et al. "Stem-loop II motif (s2m) structural and dynamical differences between SARS-CoV-2 and b. 1.617. 2 (Delta) variant". In: *Biophysical Journal* 121.3 (2022), 207a.

[2675] Jing Li and Jonathan T Sczepanski. "Targeting a conserved structural element from the SARS-CoV-2 genome using l-DNA aptamers". In: *RSC Chemical Biology* 3.1 (2022), pp. 79–84.

[2676] AQR Capital Management and community of contributors. *Pandas library*. Accessed: 2024-04-30. 2024. URL: `https://pandas.pydata.org/`.

[2677] NumPy Steering Council. *Numpy library*. Accessed: 2024-04-30. 2024. URL: `https://numpy.org/`.

[2678] Python Software Foundation. *Math Python library*. Accessed: 2024-04-30. 2024. URL: `https://docs.python.org/3/library/math.html/`.

[2679] Python Software Foundation. *Os Python library*. Accessed: 2024-04-30. 2024. URL: `https://docs.python.org/es/3.10/library/os.html/`.

[2680] Python Software Foundation. *Re Python library*. Accessed: 2024-04-30. 2024. URL: `https://docs.python.org/es/3/library/re.html/`.

[2681] Baiju Muthukadan. *Selenium with Python*. `https://selenium-python.readthedocs.io/`. Accessed on May 03, 2024. 2024.

[2682] Guenter Stoesser et al. "The EMBL nucleotide sequence database". In: *Nucleic Acids Research* 29.1 (2001), pp. 17–21.

[2683] Sascha El-Sharkawy, Nozomi Yamagishi-Eichler, and Klaus Schmid. "Metrics for analyzing variability and its implementation in software product lines: A systematic literature review". In: *Information and Software Technology* 106 (2019), pp. 1–30.

[2684] Elizabeth Spencer, Lucy Bryant, and Kim Colyvas. "Minimizing variability in language sampling analysis: A practical way to calculate text length and time variability and measure reliable change when assessing clients". In: *Topics in Language Disorders* 40.2 (2020), pp. 166–181.

[2685] Mollee J Pezold, Caitlin M Imgrund, and Holly L Storkel. "Using computer programs for language sample analysis". In: *Language, Speech, and Hearing Services in Schools* 51.1 (2020), pp. 103–114.

[2686] Yiming Yang. "An evaluation of statistical approaches to text categorization". In: *Information retrieval* 1.1 (1999), pp. 69–90.

[2687] Sergio Jimenez et al. "Generalized mongue-elkan method for approximate text string comparison". In: *Computational Linguistics and Intelligent Text Processing: 10th International Conference, CICLing 2009, Mexico City, Mexico, March 1-7, 2009. Proceedings 10*. Springer. 2009, pp. 559–570.

[2688] Sergio Jimenez, Fabio Gonzalez, and Alexander Gelbukh. "Text comparison using soft cardinality". In: *International symposium on string processing and information retrieval*. Springer. 2010, pp. 297–302.

[2689] Abdullah Ammar Karcioglu and Hasan Bulut. "Improving hash-q exact string matching algorithm with perfect hashing for DNA sequences". In: *Computers in Biology and Medicine* 131 (2021), p. 104292.

[2690] Joshua A Imperatore et al. "Conserved elements in the 3'-UTR of SARS-CoV-2: involvement in genomic dimerization and interactions with cellular microrna". In: *Biophysical Journal* 120.3 (2021), 313a.

[2691] Hongbing Jiang et al. "The highly conserved stem-loop II motif is dispensable for SARS-CoV-2". In: *Journal of virology* 97.6 (2023), e00635–23.

[2692] Xue Liu et al. "A comprehensive evolutionary and epidemiological characterization of insertion and deletion mutations in SARS-CoV-2 genomes". In: *Virus evolution* 7.2 (2021), veab104.

[2693] Alan G Hinnebusch. "Molecular mechanism of scanning and start codon selection in eukaryotes". In: *Microbiology and Molecular Biology Reviews* 75.3 (2011), pp. 434–467.

[2694] Katsura Asano. "Why is start codon selection so precise in eukaryotes?" In: *Translation* 2.1 (2014), e28387.

[2695] Graziano Pesole et al. "Analysis of oligonucleotide AUG start codon context in eukariotic mRNAs". In: *Gene* 261.1 (2000), pp. 85–91.

[2696] Nimisha Ghosh, Suman Nandi, and Indrajit Saha. "A review on evolution of emerging SARS-CoV-2 variants based on spike glycoprotein". In: *International immunopharmacology* 105 (2022), p. 108565.

[2697] Alexandra C Walls et al. "Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein". In: *Cell* 181.2 (2020), pp. 281–292.

[2698] Martina Bianchi et al. “Sars-CoV-2 envelope and membrane proteins: Structural differences linked to virus characteristics?” In: *BioMed Research International* 2020.1 (2020), p. 4389089.

[2699] Alexander Kuzmin et al. “Structure and dynamics of the SARS-CoV-2 envelope protein monomer”. In: *Proteins: Structure, Function, and Bioinformatics* 90.5 (2022), pp. 1102–1114.

[2700] Prateek Kumar et al. “An insight into SARS-CoV-2 membrane protein interaction with spike, envelope, and nucleocapsid proteins”. In: *Journal of Biomolecular Structure and Dynamics* 41.3 (2023), pp. 1062–1071.

[2701] Xueyan Liu et al. “Rampant C-to-U deamination accounts for the intrinsically high mutation rate in SARS-CoV-2 spike gene”. In: *Rna* 28.7 (2022), pp. 917–926.

[2702] Kejie Mou et al. “Emerging mutations in Nsp1 of SARS-CoV-2 and their effect on the structural stability”. In: *Pathogens* 10.10 (2021), p. 1285.

[2703] Francesca Benedetti et al. “Emerging of a SARS-CoV-2 viral strain with a deletion in nsp1”. In: *Journal of Translational Medicine* 18 (2020), pp. 1–6.

[2704] Neha Kaushal et al. “Mutational frequencies of SARS-CoV-2 genome during the beginning months of the outbreak in USA”. In: *Pathogens* 9.7 (2020), p. 565.

[2705] Tetsuya Akaishi, Kei Fujiwara, and Tadashi Ishii. “Insertion/deletion hotspots in the Nsp2, Nsp3, S1, and ORF8 genes of SARS-related coronaviruses”. In: *BMC ecology and evolution* 22.1 (2022), p. 123.

[2706] SM Ali Hosseini Rad et al. “The impact of non-synonymous mutations on miRNA binding sites within the SARS-CoV-2 NSP3 and NSP4 genes”. In: *Scientific Reports* 13.1 (2023), p. 16945.

[2707] M Shaminur Rahman et al. “Evolutionary dynamics of SARS-CoV-2 nucleocapsid protein and its consequences”. In: *Journal of medical virology* 93.4 (2021), pp. 2177–2195.

[2708] Juliana M Serpeloni et al. “Genome interaction of the virus and the host genes and non-coding RNAs in SARS-CoV-2 infection”. In: *Immunobiology* 226.5 (2021), p. 152130.

[2709] Xiaoxing Liu et al. “Non-coding RNAs expression in SARS-CoV-2 infection: pathogenesis, clinical significance, and therapeutic targets”. In: *Signal Transduction and Targeted Therapy* 8.1 (2023), p. 441.

[2710] E Domingo and JJ. Holland. “RNA virus mutations and fitness for survival”. In: *Annual review of microbiology* 51.1 (1997), pp. 151–178.

[2711] Rafael Sanjuán and Pilar Domingo-Calap. “Mechanisms of viral mutation”. In: *Cellular and molecular life sciences* 73 (2016), pp. 4433–4448.

[2712] Grigorios D Amoutzias et al. "The remarkable evolutionary plasticity of coronaviruses by mutation and recombination: insights for the COVID-19 pandemic and the future evolutionary paths of SARS-CoV-2". In: *Viruses* 14.1 (2022), p. 78.

[2713] Harr F Njai et al. "The predominance of Human Immunodeficiency Virus Type 1 (HIV-1) circulating recombinant form 02 (CRF02_AG) in West Central Africa may be related to its replicative fitness". In: *Retrovirology* 3.1 (2006), pp. 1–11.

[2714] Nicolas Combelas et al. "Recombination between poliovirus and coxsackie A viruses of species C: a model of viral genetic plasticity and emergence". In: *Viruses* 3.8 (2011), pp. 1460–1484.

[2715] Panida Poonsin et al. "Canine respiratory coronavirus in Thailand undergoes mutation and evidences a potential putative parent for genetic recombination". In: *Microbiology Spectrum* (2023), e02268–23.

[2716] Gergely Tekes et al. "Chimeric Feline Coronaviruses that encode Type II Spike protein on Type I genetic background display accelerated viral growth and altered receptor usage". In: *Journal of Virology* 84.3 (2010), pp. 1326–1333.

[2717] Kerry Gordon et al. "A Recombinant Variant with Increased Envelope Entry Efficiency Emerged During Early Infection of an HIV-1 Subtype C Dual Infected Rapid Progressor". In: *AIDS research and human retroviruses* 32.3 (2016), pp. 303–310.

[2718] Anastasia N Vlasova et al. "Novel canine coronavirus isolated from a hospitalized patient with pneumonia in East Malaysia". In: *Clinical Infectious Diseases* 74.3 (2022), pp. 446–454.

[2719] Nancy Stella Cuervo et al. "Genomic features of intertypic recombinant Sabin poliovirus strains excreted by primary vaccinees". In: *Journal of Virology* 75.13 (2001), pp. 5740–5751.

[2720] Sophie Jegouic et al. "Recombination between polioviruses and co-circulating Coxsackie A viruses: role in the emergence of pathogenic vaccine-derived polioviruses". In: *PLoS pathogens* 5.5 (2009), e1000412.

[2721] Ben Hu et al. "Bat origin of human coronaviruses". In: *Virology journal* 12.1 (2015), pp. 1–10.

[2722] Chung-Chau Hon et al. "Evidence of the recombinant origin of a bat severe acute respiratory syndrome (SARS)-like coronavirus and its implications on the direct ancestor of SARS coronavirus". In: *Journal of virology* 82.4 (2008), pp. 1819–1826.

[2723] Jamal SM Sabir et al. "Co-circulation of three camel coronavirus species and recombination of MERS-CoVs in Saudi Arabia". In: *Science* 351.6268 (2016), pp. 81–84.

[2724] Cassandra L. Atzrodt et al. "A Guide to COVID-19: a global pandemic caused by the novel coronavirus SARS-CoV-2". In: *The FEBS journal* 287.17 (2020), pp. 3633–3650.

[2725] Kristen M Marks and Roy M Gulick. "COVID-19". In: *Annals of Internal Medicine* 176.10 (2023), ITC145–ITC160.

[2726] World Health Organization. *WHO COVID-19 Dashboard*. `https://covid19.who.int/`. Accessed on September 11, 2023. 2020.

[2727] Xiang Yu and Na Li. "Understanding the beginning of a pandemic: Chinas response to the emergence of COVID-19". In: *Journal of Infection and Public Health* 14.3 (2021), pp. 347–352.

[2728] Nicholas G Davies et al. "Increased mortality in community-tested cases of SARS-CoV-2 lineage B. 1.1. 7". In: *Nature* 593.7858 (2021), pp. 270–274.

[2729] Katherine A Twohig et al. "Hospital admission and emergency care attendance risk for SARS-CoV-2 Delta (B. 1.617. 2) compared with Alpha (B. 1.1. 7) variants of concern: a cohort study". In: *The Lancet Infectious Diseases* 22.1 (2022), pp. 35–42.

[2730] Alireza Zali et al. "Mortality among hospitalized COVID-19 patients during surges of SARS-CoV-2 Alpha (B. 1.1. 7) and Delta (B. 1.617. 2) variants". In: *Scientific Reports* 12.1 (2022), p. 18918.

[2731] Peter V. Markov et al. "The evolution of SARS-CoV-2". In: *Nature Reviews Microbiology* 21.6 (2023), pp. 361–379.

[2732] Yusha Araf et al. "Omicron variant of SARS-CoV-2: genomics, transmissibility, and responses to current COVID-19 vaccines". In: *Journal of medical virology* 94.5 (2022), pp. 1825–1832.

[2733] Nitin K Saksena et al. "SARS-CoV-2 variants, its recombinants and epigenomic exploitation of host defenses". In: *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease* (2023), p. 166836.

[2734] Yuliang Wang et al. "Characterization of SARS-CoV-2 recombinants and emerging Omicron sublineages". In: *International journal of medical sciences* 20.1 (2023), p. 151.

[2735] Liang Wang and George F. Gao. "The Wolf is indeed coming: recombinant Deltacron SARS-CoV-2 detected". In: *China CDC Weekly* 4.14 (2022), p. 285.

[2736] Jocelyne Piret and Guy Boivin. "Pandemics throughout history". In: *Frontiers in microbiology* 11 (2021), p. 631736.

[2737] Sunday Adeola Ajagbe and Matthew O Adigun. "Deep learning techniques for detection and prediction of pandemic diseases: a systematic literature review". In: *Multimedia Tools and Applications* (2023), pp. 1–35.

[2738] International Committee on Taxonomy of Viruses (ICTV). *Official Taxonomic Resources*. Accessed: 2024-05-09. 2024. URL: `https://ictv.global/`.

[2739] Peter J Walker et al. "ICTV virus taxonomy profile: Roniviridae". In: *Journal of General Virology* 102.1 (2021), p. 001514.

[2740] Dave Cavanagh. "Coronaviridae: a review of coronaviruses and toroviruses". In: *Coronaviruses with special emphasis on first insights concerning SARS* (2005), pp. 1–54.

[2741] Stuart G Siddell. *The Coronaviridae: an introduction*. Springer, 1995.

[2742] Harvey F Lodish. *Molecular cell biology*. Macmillan, 2008.

[2743] Golzar Hossain et al. "SARS-CoV-2 host diversity: An update of natural infections and experimental evidence". In: *Journal of Microbiology, Immunology and Infection* 54.2 (2021), pp. 175–181.

[2744] Huiping Shuai et al. "Emerging SARS-CoV-2 variants expand species tropism to murines". In: *EBioMedicine* 73 (2021).

[2745] Emma C Hobbs and Tristan J Reid. "Animals and SARS-CoV-2: Species susceptibility and viral transmission in experimental and natural conditions, and the potential implications for community transmission". In: *Transboundary and Emerging Diseases* 68.4 (2021), pp. 1850–1867.

[2746] Ronald S Murray et al. "Coronavirus infects and causes demyelination in primate central nervous system". In: *Virology* 188.1 (1992), pp. 274–284.

[2747] Gary F Cabirac et al. "Coronavirus JHM OMP1 pathogenesis in owl monkey CNS and coronavirus infection of owl monkey CNS via peripheral routes". In: *Coronaviruses: Molecular Biology and Virus-Host Interactions* (1993), pp. 347–352.

[2748] L Kim et al. "Molecular characterization and pathogenesis of transmissible gastroenteritis coronavirus (TGEV) and porcine respiratory coronavirus (PRCV) field isolates co-circulating in a swine herd". In: *Archives of virology* 145 (2000), pp. 1133–1147.

[2749] C Schwegmann-Wessels and G Herrler. "Transmissible gastroenteritis virus infection: a vanishing specter". In: *DTW. Deutsche tierarztliche Wochenschrift* 113.4 (2006), pp. 157–159.

[2750] Heike Hofmann and Stefan Pöhlmann. "Cellular entry of the SARS coronavirus". In: *Trends in microbiology* 12.10 (2004), pp. 466–472.

[2751] Stanley G Sawicki, Dorothea L Sawicki, and Stuart G Siddell. "A contemporary view of coronavirus transcription". In: *Journal of virology* 81.1 (2007), pp. 20–29.

[2752] Márta Lrincz et al. "Sporadic re-emergence of enzootic porcine transmissible gastroenteritis in Hungary". In: *Acta Veterinaria Hungarica* 62.1 (2014), pp. 125–133.

[2753] Christel Schwegmann-WeSSels and Georg Herrler. "Sialic acids as receptor determinants for coronaviruses". In: *Glycoconjugate journal* 23 (2006), pp. 51–58.

[2754] Yutaka Terada et al. “Feline infectious peritonitis virus with a large deletion in the 5-terminal region of the spike gene retains its virulence for cats”. In: *Journal of general virology* 93.9 (2012), pp. 1930–1934.

[2755] Shinji Makino et al. “High-frequency RNA recombination of murine coronaviruses”. In: *Journal of virology* 57.3 (1986), pp. 729–737.

[2756] Isabel Sola et al. “Continuous and discontinuous RNA synthesis in coronaviruses”. In: *Annual review of virology* 2 (2015), pp. 265–288.

[2757] Jennifer Gribble et al. “The coronavirus proofreading exoribonuclease mediates extensive viral recombination”. In: *PLoS pathogens* 17.1 (2021), e1009226.

[2758] Sanneke A Kottier, David Cavanagh, and Paul Britton. “First experimental evidence of recombination in infectious bronchitis virus: Recombination in IBV”. In: *Corona-and Related Viruses: Current Concepts in Molecular Biology and Pathogenesis* (1995), pp. 551–556.

[2759] Sanneke A Kottier, David Cavanagh, and Paul Britton. “Experimental evidence of recombination in coronavirus infectious bronchitis virus”. In: *Virology* 213.2 (1995), pp. 569–580.

[2760] Steven J van Beurden et al. “A reverse genetics system for avian coronavirus infectious bronchitis virus based on targeted RNA recombination”. In: *Virology journal* 14 (2017), pp. 1–13.

[2761] Qinghong Zeng et al. “Structure of coronavirus hemagglutinin-esterase offers insight into corona and influenza virus evolution”. In: *Proceedings of the National Academy of Sciences* 105.26 (2008), pp. 9065–9069.

[2762] Li Chen and Guang Hao. “The role of angiotensin-converting enzyme 2 in coronaviruses/influenza viruses and cardiovascular disease”. In: *Cardiovascular research* 116.12 (2020), pp. 1932–1936.

[2763] Reinhard Vlasak et al. “Human and bovine coronaviruses recognize sialic acid-containing receptors similar to those of influenza C viruses.” In: *Proceedings of the National Academy of Sciences* 85.12 (1988), pp. 4526–4529.

[2764] Linda J Saif. “Comparative biology of animal coronaviruses: lessons for SARS”. In: *Severe acute respiratory syndrome* (2005), pp. 84–99.

[2765] Richard Molenkamp and Willy JM Spaan. “Identification of a specific interaction between the coronavirus mouse hepatitis virus A59 nucleocapsid protein and packaging signal”. In: *Virology* 239.1 (1997), pp. 78–86.

[2766] SG Sawicki and DL Sawicki. “Coronavirus transcription: a perspective”. In: *Coronavirus replication and reverse genetics* (2005), pp. 31–55.

[2767] Samantha Gildenhuys. “Expanding our understanding of the role polyprotein conformation plays in the coronavirus life cycle”. In: *Biochemical Journal* 477.8 (2020), pp. 1479–1482.

[2768] Volker Thiel et al. "Mechanisms and enzymes involved in SARS coronavirus genome expression". In: *Journal of General Virology* 84.9 (2003), pp. 2305–2315.

[2769] Adriaan H De Wilde et al. "Host factors in coronavirus replication". In: *Roles of host gene and non-coding RNA expression in virus infection* (2018), pp. 1–42.

[2770] Mohammed A Rohaim et al. "Structural and functional insights into non-structural proteins of coronaviruses". In: *Microbial pathogenesis* 150 (2021), p. 104641.

[2771] Marc Desforges et al. "Human coronaviruses: viral and cellular factors involved in neuroinvasiveness and neuropathogenesis". In: *Virus research* 194 (2014), pp. 145–158.

[2772] K Nakagawa, KG Lokugamage, and Shinji Makino. "Viral and cellular mRNA translation in coronavirus-infected cells". In: *Advances in virus research* 96 (2016), pp. 165–192.

[2773] Fernando Almazán et al. "Coronavirus reverse genetic systems: infectious clones and replicons". In: *Virus research* 189 (2014), pp. 262–270.

[2774] Trevor Scobey et al. "Reverse genetics with a full-length infectious cDNA of the Middle East respiratory syndrome coronavirus". In: *Proceedings of the National Academy of Sciences* 110.40 (2013), pp. 16157–16162.

[2775] Paul S Masters. "Reverse genetics of the largest RNA viruses". In: *Advances in Virus Research* 53 (1999), pp. 245–264.

[2776] Boyd Yount, Kristopher M Curtis, and Ralph S Baric. "Strategy for systematic assembly of large RNA and DNA genomes: transmissible gastroenteritis virus model". In: *Journal of virology* 74.22 (2000), pp. 10600–10611.

[2777] Patrick CY Woo et al. "Coronavirus genomics and bioinformatics analysis". In: *viruses* 2.8 (2010), pp. 1804–1820.

[2778] Patrick CY Woo et al. "Discovery of seven novel Mammalian and avian coronaviruses in the genus deltacoronavirus supports bat coronaviruses as the gene source of alphacoronavirus and betacoronavirus and avian coronaviruses as the gene source of gammacoronavirus and deltacoronavirus". In: *Journal of virology* 86.7 (2012), pp. 3995–4008.

[2779] International Committee on Taxonomy of Viruses: ICTV. *Coronaviridae*. Accessed: 2023-11-23. 2023. URL: `https://ictv.global/report_9th/RNApos/Nidovirales/Coronaviridae`.

[2780] Nicola Decaro. "Alphacoronavirus: Coronaviridae". In: *The Springer Index of Viruses* (2011), p. 371.

[2781] Justyna Miek and Katarzyna Blicharz-Domaska. "Coronaviruses in avian species–review with focus on epidemiology and diagnosis in wild birds". In: *Journal of veterinary research* 62.3 (2018), p. 249.

[2782] Yosra A Helmy et al. "The COVID-19 pandemic: a comprehensive review of taxonomy, genetics, epidemiology, diagnosis, treatment, and control". In: *Journal of clinical medicine* 9.4 (2020), p. 1225.

[2783] Swiss-Prot group. *ViralZone*. Accessed: 2023-11-24. 2023. URL: `https://viralzone.expasy.org`.

[2784] Nicola Decaro et al. "Full-length genome analysis of canine coronavirus type I". In: *Virus research* 210 (2015), pp. 100–105.

[2785] Nicola Decaro and Canio Buonavoglia. "An update on canine coronaviruses: viral evolution and pathobiology". In: *Veterinary microbiology* 132.3-4 (2008), pp. 221–234.

[2786] DD Addie et al. "Persistence and transmission of natural type I feline coronavirus infection". In: *Journal of general virology* 84.10 (2003), pp. 2735–2744.

[2787] Niels C Pedersen, Claire E Allen, and Leslie A Lyons. "Pathogenesis of feline enteric coronavirus infection". In: *Journal of Feline Medicine and Surgery* 10.6 (2008), pp. 529–541.

[2788] José V Diaz and Roberto Poma. "Diagnosis and clinical signs of feline infectious peritonitis in the central nervous system". In: *The Canadian Veterinary Journal* 50.10 (2009), p. 1091.

[2789] Harry Vennema et al. "Feline infectious peritonitis viruses arise by mutation from endemic feline enteric coronaviruses". In: *Virology* 243.1 (1998), pp. 150–157.

[2790] Niels C Pedersen. "A review of feline infectious peritonitis virus infection: 1963–2008". In: *Journal of feline medicine and surgery* 11.4 (2009), pp. 225–258.

[2791] Yiwu Chen et al. "Transmissible Gastroenteritis Virus: An Update Review and Perspective". In: *Viruses* 15.2 (2023), p. 359.

[2792] Khulud Bukhari et al. "Description and initial characterization of metatranscriptomic nidovirus-like genomes from the proposed new family Abyssoviridae, and from a sister group to the Coronavirinae, the proposed genus Alphaletovirus". In: *Virology* 524 (2018), pp. 160–171.

[2793] Ju-Yi Peng et al. "The cell tropism of porcine respiratory coronavirus for airway epithelial cells is determined by the expression of porcine aminopeptidase N". In: *Viruses* 12.11 (2020), p. 1211.

[2794] Wen Wang et al. "Discovery, diversity and evolution of novel coronaviruses sampled from rodents in China". In: *Virology* 474 (2015), pp. 19–27.

[2795] Ding X Liu, Jia Q Liang, and To S Fung. "Human coronavirus-229E,-OC43,-NL63, and-HKU1 (Coronaviridae)". In: *Encyclopedia of virology* (2021), p. 428.

[2796] Sahar Abdul-Rasool and Burtram C Fielding. "Understanding human coronavirus HCoV-NL63". In: *The open virology journal* 4 (2010), p. 76.

[2797] Ying Tao et al. "Surveillance of bat coronaviruses in Kenya identifies relatives of human coronaviruses NL63 and 229E and their recombination history". In: *Journal of virology* 91.5 (2017), pp. 10–1128.

[2798] Nicola Decaro and Alessio Lorusso. "Novel human coronavirus (SARS-CoV-2): A lesson from animal coronaviruses". In: *Veterinary microbiology* 244 (2020), p. 108693.

[2799] Susanna KP Lau et al. "Recent transmission of a novel alphacoronavirus, bat coronavirus HKU10, from Leschenault's rousettes to pomona leaf-nosed bats: first evidence of interspecies transmission of coronavirus between bats of different suborders". In: *Journal of virology* 86.21 (2012), pp. 11906–11918.

[2800] DKW Chu et al. "Genomic characterizations of bat coronaviruses (1A, 1B and HKU8) and evidence for co-infections in Miniopterus bats". In: *Journal of General Virology* 89.5 (2008), pp. 1282–1287.

[2801] Yongfei Pan et al. "Discovery of a novel swine enteric alphacoronavirus (SeACoV) in southern China". In: *Veterinary microbiology* 211 (2017), pp. 15–21.

[2802] Yi Fan et al. "Bat coronaviruses in China". In: *Viruses* 11.3 (2019), p. 210.

[2803] Zhijian Zhou, Ye Qiu, and Xingyi Ge. "The taxonomy, host range and pathogenicity of coronaviruses and other viruses in the Nidovirales order". In: *Animal Diseases* 1.1 (2021), pp. 1–28.

[2804] Linda J Saif and Kwonil Jung. "Comparative pathogenesis of bovine and porcine respiratory coronaviruses in the animal host species and SARS-CoV-2 in humans". In: *Journal of Clinical Microbiology* 58.8 (2020), pp. 10–1128.

[2805] Kwonil Jung and Linda J Saif. "Porcine epidemic diarrhea virus infection: Etiology, epidemiology, pathogenesis and immunoprophylaxis". In: *The Veterinary Journal* 204.2 (2015), pp. 134–143.

[2806] Yi-Ning Chen et al. "Entry of Scotophilus bat coronavirus-512 and severe acute respiratory syndrome coronavirus in human and multiple animal cells". In: *Pathogens* 8.4 (2019), p. 259.

[2807] ChuanMei Zhang et al. "Research progress on mink Coronavirus". In: *Chinese Journal of Virology* (2021), pp. 964–968.

[2808] Yifei Xu. "Genetic diversity and potential recombination between ferret coronaviruses from European and American lineages". In: *Journal of Infection* 80.3 (2020), pp. 350–371.

[2809] Lela Urushadze et al. "A cross sectional sampling reveals novel coronaviruses in bat populations of Georgia". In: *Viruses* 14.1 (2021), p. 72.

[2810] Harindranath Cholleti et al. "Characterization of Pipistrellus pygmaeus Bat Virome from Sweden". In: *Viruses* 14.8 (2022), p. 1654.

[2811] Anastasia N Vlasova and Linda J Saif. "Bovine coronavirus and the associated diseases". In: *Frontiers in Veterinary Science* 8 (2021), p. 643220.

[2812] N Pusterla et al. "Equine coronavirus: An emerging enteric virus of adult horses". In: *Equine veterinary education* 28.4 (2016), pp. 216–223.

[2813] Su-fen Zhang et al. "Epidemiology characteristics of human coronaviruses in patients with respiratory infection symptoms and phylogenetic analysis of HCoV-OC43 during 2010-2015 in Guangzhou". In: *Plos one* 13.1 (2018), e0191789.

[2814] Lin Xu et al. "Identification of bovine coronavirus in a Daurian ground squirrel expands the host range of Betacoronavirus 1". In: *Virologica Sinica* 38.2 (2023), p. 321.

[2815] LeAnn Luna et al. "Betacoronavirus 1 in alpacas (Vicugna pacos) in the High Peruvian Andes". In: *Small Ruminant Research* 133 (2015), pp. 7–9.

[2816] Juan Carlos Mora-Díaz et al. "Porcine hemagglutinating encephalomyelitis virus: a review". In: *Frontiers in veterinary science* 6 (2019), p. 53.

[2817] Susanna KP Lau et al. "Discovery of a novel coronavirus, China Rattus coronavirus HKU24, from Norway rats supports the murine origin of Betacoronavirus 1 and has implications for the ancestor of Betacoronavirus lineage A". In: *Journal of virology* 89.6 (2015), pp. 3076–3092.

[2818] Susan J Bender and Susan R Weiss. "Pathogenesis of murine coronavirus in the central nervous system". In: *Journal of Neuroimmune Pharmacology* 5.3 (2010), pp. 336–354.

[2819] Rachael L Hardison et al. "Evaluation of surface disinfection methods to inactivate the beta coronavirus Murine Hepatitis Virus". In: *Journal of Occupational and Environmental Hygiene* 19.8 (2022), pp. 455–468.

[2820] Susanna KP Lau et al. "Isolation and characterization of a novel Betacoronavirus subgroup A coronavirus, rabbit coronavirus HKU14, from domestic rabbits". In: *Journal of virology* 86.10 (2012), pp. 5481–5496.

[2821] Patrick CY Woo et al. "Genetic relatedness of the novel human group C betacoronavirus to Tylonycteris bat coronavirus HKU4 and Pipistrellus bat coronavirus HKU5". In: *Emerging microbes & infections* 1.1 (2012), pp. 1–5.

[2822] Susanna KP Lau et al. "Genetic characterization of Betacoronavirus lineage C viruses in bats reveals marked sequence divergence in the spike protein of pipistrellus bat coronavirus HKU5 in Japanese pipistrelle: implications for the origin of the novel Middle East respiratory syndrome coronavirus". In: *Journal of virology* 87.15 (2013), pp. 8638–8650.

[2823] Joseph O Obameso et al. "The persistent prevalence and evolution of cross-family recombinant coronavirus GCCDC1 among a bat population: a two-year follow-up". In: *Science China Life Sciences* 60 (2017), pp. 1357–1363.

[2824] Susanna KP Lau et al. "Coexistence of different genotypes in the same bat and serological characterization of Rousettus bat coronavirus HKU9 belonging to a novel Betacoronavirus subgroup". In: *Journal of virology* 84.21 (2010), pp. 11385–11394.

[2825] William J Liu et al. "T-cell immunity of SARS-CoV: Implications for vaccine development against MERS-CoV". In: *Antiviral research* 137 (2017), pp. 82–92.

[2826] Nigeer Te et al. "Middle East respiratory syndrome coronavirus infection in camelids". In: *Veterinary pathology* 59.4 (2022), pp. 546–555.

[2827] Ji-Eun Park et al. "MERS transmission and risk factors: a systematic review". In: *BMC public health* 18 (2018), pp. 1–15.

[2828] Magorzata Pomorska-Mól et al. "First detection of Hedgehog coronavirus 1 in Poland". In: *Scientific Reports* 12.1 (2022), p. 2386.

[2829] Simon L Priestnall et al. "Serological prevalence of canine respiratory coronavirus". In: *Veterinary Microbiology* 115.1-3 (2006), pp. 43–53.

[2830] Simon L Priestnall. "Canine respiratory coronavirus: A naturally occurring model of COVID-19?" In: *Veterinary Pathology* 57.4 (2020), pp. 467–471.

[2831] Wentao Zhu et al. "A novel bat coronavirus with a polybasic furin-like cleavage site". In: *Virologica Sinica* (2023).

[2832] Antoni G Wrobel et al. "Structure and binding properties of Pangolin-CoV spike glycoprotein inform the evolution of SARS-CoV-2". In: *Nature communications* 12.1 (2021), p. 837.

[2833] Susanna KP Lau et al. "Discovery and sequence analysis of four deltacoronaviruses from birds in the Middle East reveal interspecies jumping with recombination as a potential mechanism for avian-to-avian and avian-to-mammalian transmission". In: *Journal of Virology* 92.15 (2018), pp. 10–1128.

[2834] Leyi Wang, Beverly Byrum, and Yan Zhang. "Porcine coronavirus HKU15 detected in 9 US states, 2014". In: *Emerging infectious diseases* 20.9 (2014), p. 1594.

[2835] Qi-Zhang Liang et al. "Chicken or porcine aminopeptidase n mediates cellular entry of pseudoviruses carrying spike glycoprotein from the avian deltacoronaviruses HKU11, HKU13, and HKU17". In: *Journal of Virology* 97.2 (2023), e01947–22.

[2836] Vasily Y Marchenko et al. "Diversity of coronaviruses in wild and domestic birds in Vietnam". In: *Asian Pacific Journal of Tropical Medicine* 15.10 (2022), pp. 442–450.

[2837] Qiong Wang et al. "Epidemiology and evolution of novel deltacoronaviruses in birds in central China". In: *Transboundary and Emerging Diseases* 69.2 (2022), pp. 632–644.

[2838] Francine C Paim et al. "Epidemiology of deltacoronaviruses (δ-CoV) and gammacoronaviruses (γ-CoV) in wild birds in the United States". In: *Viruses* 11.10 (2019), p. 897.

[2839] Patrick CY Woo et al. "Discovery of a novel bottlenose dolphin coronavirus reveals a distinct species of marine mammal coronavirus in Gammacoronavirus". In: *Journal of Virology* 88.2 (2014), pp. 1318–1331.

[2840] Jasper FW Chan et al. "Is the discovery of the novel human betacoronavirus 2c EMC/2012 (HCoV-EMC) the beginning of another SARS-like pandemic?" In: *Journal of Infection* 65.6 (2012), pp. 477–489.

[2841] Jasper FW Chan et al. "Middle East respiratory syndrome coronavirus: another zoonotic betacoronavirus causing SARS-like disease". In: *Clinical microbiology reviews* 28.2 (2015), pp. 465–522.

[2842] Mélanie J Boileau and Sanjay Kapil. "Bovine coronavirus associated syndromes". In: *Veterinary Clinics: Food Animal Practice* 26.1 (2010), pp. 123–146.

[2843] Diego Forni, Rachele Cagliani, and Manuela Sironi. "Recombination and positive selection differentially shaped the diversity of betacoronavirus subgenera". In: *Viruses* 12.11 (2020), p. 1313.

[2844] Carla Miranda et al. "Genomic evolution of the human and animal coronavirus diseases". In: *Molecular Biology Reports* 48.9 (2021), pp. 6645–6653.

[2845] Jun Yong Choi and Davey M Smith. "SARS-CoV-2 variants of concern". In: *Yonsei medical journal* 62.11 (2021), p. 961.

[2846] GISAID Initiative. *GISAID-Tracking of hCoV-19 Variants*. `https://gisaid.org/hcov19-variants/`. Accessed on November 21, 2023. 2023.

[2847] Andrew Rambaut et al. "A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology". In: *Nature microbiology* 5.11 (2020), pp. 1403–1407.

[2848] Alessandro M Carabelli et al. "SARS-CoV-2 variant biology: immune escape, transmission and fitness". In: *Nature Reviews Microbiology* 21.3 (2023), pp. 162–177.

[2849] Jun Yu, Mingyao Ai, and Zhiqiang Ye. "A review on design inspired subsampling for big data". In: *Statistical Papers* (2023), pp. 1–44.

[2850] Ana Guerrero-Tamayo et al. "Discovering Mathematical Patterns Behind HIV-1 Genetic Recombination: a new methodology to identify viral features". In: *IEEE Access* (2023).

[2851] Simon Kornblith, Jonathon Shlens, and Quoc V Le. "Do better imagenet models transfer better?" In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 2661–2671.

[2852] Peter M Howley et al. *Fields Virology: RNA Viruses*. Lippincott Williams & Wilkins, 2022.

[2853] Jin Huang and Charles X Ling. "Using AUC and accuracy in evaluating learning algorithms". In: *IEEE Transactions on knowledge and Data Engineering* 17.3 (2005), pp. 299–310.

[2854] Jose Antonio Marmolejo-Saucedo and Utku Kose. "Numerical grad-CAM based explainable convolutional neural network for brain tumor diagnosis". In: *Mobile Networks and Applications* (2022), pp. 1–10.

[2855] Valentine Wargnier-Dauchelle et al. "A Weakly Supervised Gradient Attribution Constraint for Interpretable Classification and Anomaly Detection". In: *IEEE Transactions on Medical Imaging* (2023).

[2856] Karl J Friston et al. "Spatial registration and normalization of images". In: *Human brain mapping* 3.3 (1995), pp. 165–189.

[2857] Jie Tan et al. "Image-Contrast Technology Based on the Electrochemiluminescence of Porous Silicon and Its Application in Fingerprint Visualization". In: *Angewandte Chemie* 126.37 (2014), pp. 9980–9984.

[2858] G Bradski. "The OpenCV Library". In: *Dr. Dobb's Journal of Software Tools* (2000).

[2859] Marios Nikolaidis et al. "The neighborhood of the Spike gene is a hotspot for modular intertypic homologous and nonhomologous recombination in coronavirus genomes". In: *Molecular biology and evolution* 39.1 (2022), msab292.

[2860] Yatish Turakhia et al. "Pandemic-scale phylogenomics reveals the SARS-CoV-2 recombination landscape". In: *Nature* 609.7929 (2022), pp. 994–997.

[2861] Chaoyang Li et al. "Deep Learning and Image Recognition". In: *2023 IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT)*. IEEE. 2023, pp. 557–562.

[2862] Lei Cai, Jingyang Gao, and Di Zhao. "A review of the application of deep learning in medical image classification and segmentation". In: *Annals of translational medicine* 8.11 (2020).

[2863] Sazan Qadir Maulud et al. "Deltacron: Apprehending a new phase of the COVID-19 pandemic". In: *International Journal of Surgery* 102 (2022), p. 106654.

[2864] Ranjan K Mohapatra et al. "The recombinant variants of SARS-CoV-2: concerns continues amid COVID-19 pandemic". In: *Journal of Medical Virology* 94.8 (2022), p. 3506.

[2865] Lauren J Beesley et al. "SARS-CoV-2 variant transition dynamics are associated with vaccination rates, number of co-circulating variants, and convalescent immunity". In: *EBioMedicine* 91 (2023).

[2866] Rishad Shiraz and Shashank Tripathi. "Enhanced recombination among Omicron subvariants of SARS-CoV-2 contributes to viral immune escape". In: *Journal of Medical Virology* 95.2 (2023), e28519.

[2867] Tomokazu Tamura et al. "Virological characteristics of the SARS-CoV-2 XBB variant derived from recombination of two Omicron subvariants". In: *Nature communications* 14.1 (2023), p. 2800.

[2868] Fabio Scarpa et al. "Genome-based comparison between the recombinant SARS-CoV-2 XBB and its parental lineages". In: *Journal of Medical Virology* 95.3 (2023), e28625.

[2869] Antonio Vitiello et al. “Advances in the Omicron variant development”. In: *Journal of internal medicine* 292.1 (2022), pp. 81–90.

[2870] Isabel Pagani et al. “Origin and evolution of SARS-CoV-2”. In: *The European Physical Journal Plus* 138.2 (2023), p. 157.

[2871] Pei Du, George Fu Gao, and Qihui Wang. “The mysterious origins of the Omicron variant of SARS-CoV-2”. In: *The Innovation* 3.2 (2022).

[2872] Changshuo Wei et al. “Evidence for a mouse origin of the SARS-CoV-2 Omicron variant”. In: *Journal of genetics and genomics* 48.12 (2021), pp. 1111–1121.

[2873] Yamin Sun et al. “Origin and evolutionary analysis of the SARS-CoV-2 Omicron variant”. In: *Journal of biosafety and biosecurity* 4.1 (2022), pp. 33–37.

[2874] Pushpendra Singh et al. “Mosaic Recombination Inflicted Various SARS-CoV-2 Lineages to Emerge into Novel Virus Variants: a Review Update”. In: *Indian J Clin Biochem* (2022), pp. 1–8.

[2875] Ewan P Plant and Zhiping Ye. “Bias at the third nucleotide of codon pairs in virus and host genomes”. In: *Scientific Reports* 12.1 (2022), p. 4522.

[2876] Kathleen P Talaro et al. *Foundations in Microbiology*. Vol. 622. Wm. C. Brown, 1996.

[2877] FH Crick. “Codon-anticodon pairing: the Wobble hypothesis.” In: *Molecular Biology: A Selection of Papers* 19 (1966), pp. 548–555.

[2878] Alfred L Goldberg and Robert E Wittes. “Genetic code: aspects of organization”. In: *Science* 153.3734 (1966), pp. 420–424.

[2879] Haixia Zhou et al. “Structural definition of a neutralization epitope on the N-terminal domain of MERS-CoV spike glycoprotein”. In: *Nature communications* 10.1 (2019), p. 3068.

[2880] Guiqing Peng et al. “Crystal structure of bovine coronavirus spike protein lectin domain”. In: *Journal of Biological Chemistry* 287.50 (2012), pp. 41931–41938.

[2881] Xiangyang Chi et al. “A neutralizing human antibody binds to the N-terminal domain of the Spike protein of SARS-CoV-2”. In: *Science* 369.6504 (2020), pp. 650–655.

[2882] Gabriele Cerutti et al. “Potent SARS-CoV-2 neutralizing antibodies directed against spike N-terminal domain target a single supersite”. In: *Cell host & microbe* 29.5 (2021), pp. 819–833.

[2883] William N Voss et al. “Prevalent, protective, and convergent IgG recognition of SARS-CoV-2 non-RBD spike epitopes”. In: *Science* 372.6546 (2021), pp. 1108–1112.

[2884] Pengfei Wang et al. “Antibody resistance of SARS-CoV-2 variants B. 1.351 and B. 1.1. 7”. In: *Nature* 593.7857 (2021), pp. 130–135.

[2885] Kevin R McCarthy et al. "Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape". In: *Science* 371.6534 (2021), pp. 1139–1142.

[2886] Alexandre Hassanin. "Variation in synonymous nucleotide composition among genomes of sarbecoviruses and consequences for the origin of COVID-19". In: *Gene* 835 (2022), p. 146641.

[2887] Alan M Rice et al. "Evidence for strong mutation bias toward, and selection against, U content in SARS-CoV-2: implications for vaccine design". In: *Molecular biology and evolution* 38.1 (2021), pp. 67–83.

[2888] Fernando L Tort, Matías Castells, and Juan Cristina. "A comprehensive analysis of genome composition and codon usage patterns of emerging coronaviruses". In: *Virus research* 283 (2020), p. 197976.

[2889] Bryan Saldivar-Espinoza et al. "The mutational landscape of SARS-CoV-2". In: *International Journal of Molecular Sciences* 24.10 (2023), p. 9072.

[2890] GenSscript. *Human Genetic Code*. `https://www.genscript.com/tools/codon-frequency-table/`. Accessed on March 21, 2024. 2024.

[2891] Artur I Zabolotskii, Stanislav V Kozlovskiy, and Alexey G Katrukha. "The Influence of the Nucleotide Composition of Genes and Gene Regulatory Elements on the Efficiency of Protein Expression in Escherichia coli". In: *Biochemistry (Moscow)* 88.Suppl 1 (2023), S176–S191.

[2892] Jean-Luc Jestin and Achim Kempf. "Degeneracy in the genetic code: how and why". In: *Genes Genomes Genomics* 1.1 (2007), pp. 100–103.

[2893] Eugene V Koonin and Valerian V Dolja. "A virocentric perspective on the evolution of life". In: *Current opinion in virology* 3.5 (2013), pp. 546–557.

[2894] Edward C Holmes. "The evolution of viral emergence". In: *Novel and Re-Emerging Respiratory Viral Diseases: Novartis Foundation Symposium 290*. Vol. 290. Wiley Online Library. 2008, pp. 17–31.

[2895] Joel O Wertheim et al. "A case for the ancient origin of coronaviruses". In: *Journal of Virology* 87.12 (2013), pp. 7039–7045.

[2896] Beatrice H Hahn et al. "AIDS as a zoonosis: scientific and public health implications". In: *Science* 287.5453 (2000), pp. 607–614.

[2897] Alexander Hayward. "Origin of the retroviruses: when, where, and how?" In: *Current opinion in virology* 25 (2017), pp. 23–27.

[2898] Michael J Benton. *Vertebrate palaeontology*. John Wiley & Sons, 2014.

[2899] Kholhring Lalchhandama. "A history of coronaviruses". In: *WikiJournal of Medicine* 9.1 (2022), pp. 1–19.

[2900] Gregory J Retallack et al. "Middle-Late Permian mass extinction on land". In: *Geological Society of America Bulletin* 118.11-12 (2006), pp. 1398–1411.

[2901] John Ruben. "The evolution of endothermy in mammals and birds: from physiology to fossils". In: *Annual review of physiology* 57.1 (1995), pp. 69–95.

[2902] Thomas Stainforth Kemp. *The origin and evolution of mammals*. Oxford University Press, 2005.

[2903] Monique R Ferguson et al. "HIV-1 replication cycle". In: *Clinics in laboratory medicine* 22.3 (2002), pp. 611–635.

[2904] Christian M Zmasek et al. "Genomic evolution of the Coronaviridae family". In: *Virology* 570 (2022), pp. 123–133.

[2905] Stephen P Goff. "Host factors exploited by retroviruses". In: *Nature Reviews Microbiology* 5.4 (2007), pp. 253–263.

[2906] Nathan J Hardenbrook and Peijun Zhang. "A structural view of the SARS-CoV-2 virus and its assembly". In: *Current opinion in virology* 52 (2022), pp. 123–134.

[2907] Ann Emery, Sarah B Joseph, Ronald Swanstrom, et al. "Nonsuppressible viremia during HIV-1 therapy meets molecular virology". In: *The Journal of Clinical Investigation* 133.6 (2023).

[2908] Bing Chen. "Molecular mechanism of HIV-1 entry". In: *Trends in microbiology* 27.10 (2019), pp. 878–891.

[2909] Marina Lusic and Robert F Siliciano. "Nuclear landscape of HIV-1 infection and integration". In: *Nature Reviews Microbiology* 15.2 (2017), pp. 69–82.

[2910] Alan N Engelman and Parmit K Singh. "Cellular and molecular mechanisms of HIV-1 integration targeting". In: *Cellular and Molecular Life Sciences* 75 (2018), pp. 2491–2507.

[2911] Stefan G Sarafianos et al. "Structure and function of HIV-1 reverse transcriptase: molecular mechanisms of polymerization and inhibition". In: *Journal of molecular biology* 385.3 (2009), pp. 693–713.

[2912] Wei-Shau Hu and Stephen H Hughes. "HIV-1 reverse transcription". In: *Cold Spring Harbor perspectives in medicine* 2.10 (2012), a006882.

[2913] Yang-Hui Jimmy Yeh et al. "The clonal expansion dynamics of the HIV-1 reservoir: mechanisms of integration site-dependent proliferation and HIV-1 persistence". In: *Viruses* 13.9 (2021), p. 1858.

[2914] Robert Craigie and Frederic D Bushman. "HIV DNA integration". In: *Cold Spring Harbor perspectives in medicine* 2.7 (2012), a006890.

[2915] Thang K Chiu and David R Davies. "Structure and function of HIV-1 integrase". In: *Current topics in medicinal chemistry* 4.9 (2004), pp. 965–977.

[2916] Olivier Delelis et al. "Integrase and integration: biochemical activities of HIV-1 integrase". In: *Retrovirology* 5 (2008), pp. 1–13.

[2917] David D Chang and Phillip A Sharp. "Messenger RNA transport and HIV rev regulation". In: *Science* 249.4969 (1990), pp. 614–615.

[2918] Carine Van Lint, Sophie Bouchat, and Alessandro Marcello. "HIV-1 transcription and latency: an update". In: *Retrovirology* 10 (2013), pp. 1–38.

[2919] Daniela Toro-Ascuy et al. "Interactions between the HIV-1 unspliced mRNA and host mRNA decay machineries". In: *Viruses* 8.11 (2016), p. 320.

[2920] Jean-Marc Jacque, Karine Triques, and Mario Stevenson. "Modulation of HIV-1 replication by RNA interference". In: *Nature* 418.6896 (2002), pp. 435–438.

[2921] Luc Geeraert, Günter Kraus, and Roger J Pomerantz. "Hide-and-seek: the challenge of viral persistence in HIV-1 infection". In: *Annu. Rev. Med.* 59.1 (2008), pp. 487–501.

[2922] Nancie M Archin et al. "Eradicating HIV-1 infection: seeking to clear a persistent pathogen". In: *Nature Reviews Microbiology* 12.11 (2014), pp. 750–764.

[2923] Behnaz Bakhshandeh et al. "Mutations in SARS-CoV-2; Consequences in structure, function, and pathogenicity of the virus". In: *Microbial Pathogenesis* 154 (2021), p. 104831.

[2924] Gilbert Eriani and Franck Martin. "Viral and cellular translation during SARS-CoV-2 infection". In: *FEBS Open Bio* 12.9 (2022), pp. 1584–1601.

[2925] Massimo Pizzato et al. "SARS-CoV-2 and the host cell: a tale of interactions". In: *Frontiers in Virology* 1 (2022), p. 815388.

[2926] Philip Vkovski et al. "Coronavirus biology and replication: implications for SARS-CoV-2". In: *Nature Reviews Microbiology* 19.3 (2021), pp. 155–170.

[2927] Ayslan Castro Brant et al. "SARS-CoV-2: from its discovery to genome structure, transcription, and replication". In: *Cell & bioscience* 11 (2021), pp. 1–17.

[2928] Zhiyong Lou and Zihe Rao. "The life of SARS-CoV-2 inside cells: Replication - transcription complex assembly and function". In: *Annual Review of Biochemistry* 91.1 (2022), pp. 381–401.

[2929] Frank J Palella Jr et al. "Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection". In: *New England Journal of Medicine* 338.13 (1998), pp. 853–860.

[2930] Chaolin Huang et al. "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China". In: *The lancet* 395.10223 (2020), pp. 497–506.

[2931] Rebecca F Baggaley, Richard G White, and Marie-Claude Boily. "Systematic review of orogenital HIV-1 transmission probabilities". In: *International journal of epidemiology* 37.6 (2008), pp. 1255–1265.

[2932] Andrew G Harrison, Tao Lin, and Penghua Wang. "Mechanisms of SARS-CoV-2 transmission and pathogenesis". In: *Trends in immunology* 41.12 (2020), pp. 1100–1115.

[2933] Daniel C Douek, Mario Roederer, and Richard A Koup. "Emerging concepts in the immunopathogenesis of AIDS". In: *Annual review of medicine* 60 (2009), pp. 471–484.

[2934] Markus Hoffmann et al. "SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor". In: *cell* 181.2 (2020), pp. 271–280.

[2935] Andrew S Edelman and Susan Zolla-Pazner. "AIDS: a syndrome of immune dysregulation, dysfunction, and deficiency". In: *The FASEB journal* 3.1 (1989), pp. 22–30.

[2936] Puja Mehta et al. "COVID-19: consider cytokine storm syndromes and immunosuppression". In: *The lancet* 395.10229 (2020), pp. 1033–1034.

[2937] Steven G Deeks, Sharon R Lewin, and Diane V Havlir. "The end of AIDS: HIV infection as a chronic disease". In: *The lancet* 382.9903 (2013), pp. 1525–1533.

[2938] Suresh Thakur et al. "SARS-CoV-2 mutations and their impact on diagnostics, therapeutics and vaccines". In: *Frontiers in medicine* 9 (2022), p. 815389.

[2939] Mark P Ashe et al. "Poly (A) site selection in the HIV-1 provirus: inhibition of promoter-proximal polyadenylation by the downstream major splice donor site." In: *Genes & development* 9.23 (1995), pp. 3008–3025.

[2940] Sarah C Keane et al. "NMR detection of intermolecular interaction sites in the dimeric 5-leader of the HIV-1 genome". In: *Proceedings of the National Academy of Sciences* 113.46 (2016), pp. 13033–13038.

[2941] Marc S Weinberg et al. "An RNA targeted to the HIV-1 LTR promoter modulates indiscriminate off-target gene activation". In: *Nucleic acids research* 35.21 (2007), pp. 7303–7312.

[2942] Bradley D Preston, Bernard J Poiesz, and Lawrence A Loeb. "Fidelity of HIV-1 reverse transcriptase". In: *Science* 242.4882 (1988), pp. 1168–1171.

[2943] George J Klarmann, Cherylene A Schauber, and Bradley D Preston. "Template-directed pausing of DNA synthesis by HIV-1 reverse transcriptase during polymerization of HIV-1 sequences in vitro". In: *Journal of Biological Chemistry* 268.13 (1993), pp. 9793–9802.

[2944] Viktor Müller and Sebastian Bonhoeffer. "Guanine–adenine bias: a general property of retroid viruses that is unrelated to host-induced hypermutation". In: *Trends in Genetics* 21.5 (2005), pp. 264–268.

[2945] Antoinette C van der Kuyl and Ben Berkhout. "The biased nucleotide composition of the HIV genome: a constant factor in a highly variable virus". In: *Retrovirology* 9 (2012), pp. 1–14.

[2946] KK Vidyavijayan et al. "Biased nucleotide composition and differential codon usage pattern in HIV-1 and HIV-2". In: *AIDS Research and Human Retroviruses* 33.3 (2017), pp. 298–307.

[2947] Luca Bellucci et al. "Unconventional plasticity of HIV-1 reverse transcriptase: how inhibitors could open a connection gate between allosteric and catalytic sites". In: *Journal of chemical information and modeling* 53.12 (2013), pp. 3117–3122.

[2948] Olga A Nikolaitchik et al. "Selective packaging of HIV-1 RNA genome is guided by the stability of 5 untranslated region polyA stem". In: *Proceedings of the National Academy of Sciences* 118.50 (2021), e2114494118.

[2949] Jonathan MO Rawson et al. "Recombination is required for efficient HIV-1 replication and the maintenance of viral genome integrity". In: *Nucleic acids research* 46.20 (2018), pp. 10535–10545.

[2950] David Posada. *Bioinformatics for DNA sequence analysis*. Springer, 2009.

[2951] Peter Simmonds. "Rampant C–>U hypermutation in the genomes of SARS-CoV-2 and other coronaviruses: causes and consequences for their short-and long-term evolutionary trajectories". In: *Msphere* 5.3 (2020), pp. 10–1128.

[2952] Salvatore Di Giorgio et al. "Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2". In: *Science advances* 6.25 (2020), eabb5813.

[2953] Dae-Kyum Kim et al. "A proteome-scale map of the SARS-CoV-2–human contactome". In: *Nature Biotechnology* 41.1 (2023), pp. 140–149.

[2954] Yoshihiro Nakata et al. "Cellular APOBEC3A deaminase drives mutations in the SARS-CoV-2 genome". In: *Nucleic Acids Research* 51.2 (2023), pp. 783–795.

[2955] Takaomi Ishida et al. "5'long terminal repeat (LTR)-selective methylation of latently infected HIV-1 provirus that is demethylated by reactivation signals". In: *Retrovirology* 3 (2006), pp. 1–7.

[2956] Kateina Trejbalová et al. "Development of 5' LTR DNA methylation of latent HIV-1 provirus in cell line models and in long-term-infected individuals". In: *Clinical epigenetics* 8 (2016), pp. 1–20.

[2957] Lindsey Reynolds et al. "Repression of the HIV-1 5 LTR promoter and inhibition of HIV-1 replication by using engineered zinc-finger transcription factors". In: *Proceedings of the National Academy of Sciences* 100.4 (2003), pp. 1615–1620.

[2958] Andrew P Rice. "Roles of CDKs in RNA polymerase II transcription of the HIV-1 genome". In: *Transcription* 10.2 (2019), pp. 111–117.

[2959] Sylvain Thierry et al. "Integrase inhibitor reversal dynamics indicate unintegrated HIV-1 dna initiate de novo integration". In: *Retrovirology* 12 (2015), pp. 1–12.

[2960] Zeger Debyser et al. "Assays for the evaluation of HIV-1 integrase inhibitors". In: *Nuclease Methods and Protocols* (2001), pp. 139–155.

[2961] Edward Reed-Inderbitzin and Wendy Maury. "Cellular specificity of HIV-1 replication can be controlled by LTR sequences". In: *Virology* 314.2 (2003), pp. 680–695.

[2962] Jori Symons et al. "HIV integration sites in latently infected cell lines: evidence of ongoing replication". In: *Retrovirology* 14 (2017), pp. 1–11.

[2963] Atze T Das, Bep Klaver, and Ben Berkhout. "A hairpin structure in the R region of the human immunodeficiency virus type 1 RNA genome is instrumental in polyadenylation site selection". In: *Journal of virology* 73.1 (1999), pp. 81–91.

[2964] Xing Guo et al. "A unified classification system for HIV-1 5'long terminal repeats". In: *Plos one* 19.5 (2024), e0301809.

[2965] Antonella Baldassarre et al. "Potential use of noncoding RNAs and innovative therapeutic strategies to target the 5'UTR of SARS-CoV-2". In: *Epigenomics* 12.15 (2020), pp. 1349–1361.

[2966] María Teresa Hernández-Huerta et al. "3'-UTR of the SARS-CoV-2 genome as a possible source of piRNAs". In: *Genes & Diseases* 10.3 (2023), p. 668.

[2967] Shih-Cheng Chen et al. "Optimization of 5' UTR to evade SARS-CoV-2 Nonstructural protein 1-directed inhibition of protein synthesis in cells". In: *Applied Microbiology and Biotechnology* 107.7 (2023), pp. 2451–2468.

[2968] Lin Zhu et al. "Evidence for selection on SARS-CoV-2 RNA translation revealed by the evolutionary dynamics of mutations in UTRs and CDSs". In: *RNA biology* 19.1 (2022), pp. 866–876.

[2969] Setu M Vora et al. "Targeting stem-loop 1 of the SARS-CoV-2 5' UTR to suppress viral translation and Nsp1 evasion". In: *Proceedings of the National Academy of Sciences* 119.9 (2022), e2117198119.

[2970] Ramya Rangan et al. "RNA genome conservation and secondary structure in SARS-CoV-2 and SARS-related viruses: a first look". In: *Rna* 26.8 (2020), pp. 937–959.

[2971] Imre Berger and Christiane Schaffitzel. "The SARS-CoV-2 spike protein: balancing stability and infectivity". In: *Cell research* 30.12 (2020), pp. 1059–1060.

[2972] Naveenchandra Suryadevara et al. "Neutralizing and protective human monoclonal antibodies recognizing the N-terminal domain of the SARS-CoV-2 spike protein". In: *Cell* 184.9 (2021), pp. 2316–2331.

[2973] Houcemeddine Othman et al. "Interaction of the spike protein RBD from SARS-CoV-2 with ACE2: Similarity with SARS-CoV, hot-spot analysis and effect of the receptor polymorphism". In: *Biochemical and biophysical research communications* 527.3 (2020), pp. 702–708.

[2974] Shuai Xia et al. "The role of furin cleavage site in SARS-CoV-2 spike protein-mediated membrane fusion in the presence or absence of trypsin". In: *Signal transduction and targeted therapy* 5.1 (2020), p. 92.

[2975] Jun Siong Low et al. "ACE2-binding exposes the SARS-CoV-2 fusion peptide to broadly neutralizing coronavirus antibodies". In: *Science* 377.6607 (2022), pp. 735–742.

[2976] Lixiao Xing et al. "A five-helix-based SARS-CoV-2 fusion inhibitor targeting heptad repeat 2 domain against SARS-CoV-2 and its variants of concern". In: *Viruses* 14.3 (2022), p. 597.

[2977] Shi Yu et al. “SARS-CoV-2 spike engagement of ACE2 primes S2 site cleavage and fusion initiation”. In: *Proceedings of the National Academy of Sciences* 119.1 (2022), e2111199119.

[2978] Daniel Birtles and Jinwoo Lee. “Identifying distinct structural features of the SARS-CoV-2 spike protein fusion domain essential for membrane interaction”. In: *Biochemistry* 60.40 (2021), pp. 2978–2986.

[2979] Yang Yang and Lanying Du. “SARS-CoV-2 spike protein: a key target for eliciting persistent neutralizing antibodies”. In: *Signal transduction and targeted therapy* 6.1 (2021), p. 95.

[2980] Joseph Thomas Ortega et al. “Role of changes in SARS-CoV-2 spike protein in the interaction with the human ACE2 receptor: An in silico analysis”. In: *EXCLI journal* 19 (2020), p. 410.

[2981] Makoto Takeda. “Proteolytic activation of SARS-CoV-2 spike protein”. In: *Microbiology and immunology* 66.1 (2022), pp. 15–23.

[2982] Xinyu Li et al. “Spike protein mediated membrane fusion during SARS-CoV-2 infection”. In: *Journal of medical virology* 95.1 (2023), e28212.

[2983] Subhomoi Borkotoky, Debajit Dey, and Manidipa Banerjee. “Computational insight into the mechanism of SARS-CoV-2 membrane fusion”. In: *Journal of Chemical Information and Modeling* 61.1 (2021), pp. 423–431.

[2984] Naveen Vankadari et al. “Structure of human TMPRSS2 in complex with SARS-CoV-2 spike glycoprotein and implications for potential therapeutics”. In: *The Journal of Physical Chemistry Letters* 13.23 (2022), pp. 5324–5333.

[2985] Chan Yang et al. “Salvianolic acid C potently inhibits SARS-CoV-2 infection by blocking the formation of six-helix bundle core of spike protein”. In: *Signal transduction and targeted therapy* 5.1 (2020), p. 220.

[2986] Jun Zhang et al. “Structural impact on SARS-CoV-2 spike protein by D614G substitution”. In: *Science* 372.6541 (2021), pp. 525–530.

[2987] Lizhou Zhang et al. “SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity”. In: *Nature communications* 11.1 (2020), p. 6013.

[2988] Yang Liu et al. “The N501Y spike substitution enhances SARS-CoV-2 infection and transmission”. In: *Nature* 602.7896 (2022), pp. 294–299.

[2989] Bo Meng et al. “Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the Alpha variant B. 1.1. 7”. In: *Cell reports* 35.13 (2021).

[2990] Abbas Khan et al. “Higher infectivity of the SARS-CoV-2 new variants is associated with K417N/T, E484K, and N501Y mutants: an insight from structural data”. In: *Journal of cellular physiology* 236.10 (2021), pp. 7045–7057.

[2991] Akira Aoki et al. “A rapid screening assay for L452R and T478K spike mutations in SARS-CoV-2 Delta variant using high-resolution melting analysis”. In: *The Journal of Toxicological Sciences* 46.10 (2021), pp. 471–476.

[2992] Jun Zhang et al. “Structural and functional impact by SARS-CoV-2 Omicron spike mutations”. In: *Cell reports* 39.4 (2022).

[2993] Sangeeta Khare et al. “Conformational changes of the receptor binding domain of SARS-CoV-2 spike protein and prediction of a B-cell antigenic epitope using structural data”. In: *Frontiers in Artificial Intelligence* 4 (2021), p. 630955.

[2994] Yu D Nechipurenko, AA Anashkina, and OV Matveeva. “Change of antigenic determinants of SARS-CoV-2 virus S-protein as a possible cause of antibody-dependent enhancement of virus infection and cytokine storm”. In: *Biophysics* 65 (2020), pp. 703–709.

[2995] Sandipan Chakraborty. “E484K and N501Y SARS-CoV 2 spike mutants Increase ACE2 recognition but reduce affinity for neutralizing antibody”. In: *International immunopharmacology* 102 (2022), p. 108424.

[2996] Miloslav Sanda, Lindsay Morrison, and Radoslav Goldman. “N-and O-glycosylation of the SARS-CoV-2 spike protein”. In: *Analytical chemistry* 93.4 (2021), pp. 2003–2009.

[2997] Aitor Casas-Sanchez et al. “Inhibition of protein N-glycosylation blocks SARS-CoV-2 infection”. In: *MBio* 13.1 (2022), e03718–21.

[2998] Alessandra Lo Presti, Giovanni Rezza, and Paola Stefanelli. “Selective pressure on SARS-CoV-2 protein coding genes and glycosylation site prediction”. In: *Heliyon* 6.9 (2020).

[2999] Massimiliano S Tagliamonte et al. “Multiple recombination events and strong purifying selection at the origin of SARS-CoV-2 spike glycoprotein increased correlated dynamic movements”. In: *International Journal of Molecular Sciences* 22.1 (2020), p. 80.

[3000] Louis-Marie Bobay, Angela C ODonnell, and Howard Ochman. “Recombination events are concentrated in the spike protein region of Betacoronaviruses”. In: *PLoS Genetics* 16.12 (2020), e1009272.

[3001] Ahmed M Almehdi et al. “SARS-CoV-2 spike protein: pathogenesis, vaccines, and potential therapies”. In: *Infection* 49.5 (2021), pp. 855–876.

[3002] Juan Shi et al. “Effective vaccination strategy using SARS-CoV-2 spike cocktail against Omicron and other variants of concern”. In: *npj Vaccines* 7.1 (2022), p. 169.

[3003] César MJA Metzger et al. “PCR performance in the SARS-CoV-2 Omicron variant of concern?” In: *Swiss medical weekly* 151.4950 (2021), w30120–w30120.

[3004] Branislav Kovacech et al. "Monoclonal antibodies targeting two immunodominant epitopes on the Spike protein neutralize emerging SARS-CoV-2 variants of concern". In: *EBioMedicine* 76 (2022).

[3005] Ana Guerrero-Tamayo et al. "Validating by Deep Learning an Efficient Method for Genomic Sequence Analysis: Genomic Spectrograms". In: *International Conference on Hybrid Artificial Intelligence Systems*. Springer. 2023, pp. 281–292.

[3006] Ana Guerrero-Tamayo et al. "Classification of SARS-CoV-2 sequences as recombinants via a pre-trained CNN and identification of a mathematical signature relative to recombinant feature at Spike, via interpretability". In: *PLoS ONE* (2024).

[3007] Ana Guerrero-Tamayo et al. *Discovering Mathematical Patterns Behind HIV-1 Genetic Recombination: a new methodology to identify viral features - Supplementary Information*. 2023. URL: `https://dx.doi.org/10.21227/2917-1s55`.

[3008] Ian Goodfellow et al. "Generative adversarial networks". In: *Communications of the ACM* 63.11 (2020), pp. 139–144.

[3009] Minsuk Kahng et al. "Gan lab: Understanding complex deep generative models using interactive visual experimentation". In: *IEEE transactions on visualization and computer graphics* 25.1 (2018), pp. 310–320.

[3010] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[3011] Matt J Kusner, Brooks Paige, and José Miguel Hernández-Lobato. "Grammar variational autoencoder". In: *International conference on machine learning*. PMLR. 2017, pp. 1945–1954.

[3012] Tianyang Lin et al. "A survey of transformers". In: *AI open* 3 (2022), pp. 111–132.

[3013] Abel Chandra et al. "Transformer-based deep learning for predicting protein properties in the life sciences". In: *Elife* 12 (2023), e82819.

[3014] Alex Graves. "Generating sequences with recurrent neural networks". In: *arXiv preprint arXiv:1308.0850* (2013).

[3015] Anvita Gupta et al. "Generative recurrent networks for de novo drug design". In: *Molecular informatics* 37.1-2 (2018), p. 1700111.

[3016] William Bort et al. "Discovery of novel chemical reactions by deep generative recurrent neural network". In: *Scientific reports* 11.1 (2021), p. 3178.

[3017] Danilo Rezende and Shakir Mohamed. "Variational inference with normalizing flows". In: *International conference on machine learning*. PMLR. 2015, pp. 1530–1538.

[3018] Emmanuel Bengio et al. "Flow network based generative models for non-iterative diverse candidate generation". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 27381–27394.

[3019] Heiko Zimmermann et al. “A variational perspective on generative flow networks”. In: *arXiv preprint arXiv:2210.07992* (2022).

[3020] Kamran Ghasedi Dizaji, Xiaoqian Wang, and Heng Huang. “Semi-supervised generative adversarial network for gene expression inference”. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018, pp. 1435–1444.

[3021] Hugo Dalla-Torre et al. “The nucleotide transformer: Building and evaluating robust foundation models for human genomics”. In: *BioRxiv* (2023), pp. 2023–01.

[3022] Konstantinos Vrettos et al. “Generative adversarial networks for spine imaging: A critical review of current applications”. In: *European Journal of Radiology* (2024), p. 111313.

[3023] Ruopeng Gao and Limin Wang. “MeMOTR: Long-term memory-augmented transformer for multi-object tracking”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 9901–9910.

[3024] Muqiao Yang et al. “Complex transformer: A framework for modeling complex-valued sequence”. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 4232–4236.

[3025] Willone Lim et al. “Future of generative adversarial networks (GAN) for anomaly detection in network security: A review”. In: *Computers & Security* (2024), p. 103733.

[3026] Xiaohan Ma et al. “An adaptive control algorithm for stable training of generative adversarial networks”. In: *IEEE Access* 7 (2019), pp. 184103–184114.

[3027] Prabu Selvam et al. “A transformer-based framework for scene text recognition”. In: *IEEE Access* 10 (2022), pp. 100895–100910.

[3028] Ming-Hsiang Su, Chung-Hsien Wu, and Hao-Tse Cheng. “A two-stage transformer-based approach for variable-length abstractive summarization”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), pp. 2061–2072.

[3029] Zhiwei Yang et al. “HiTRANS: a hierarchical transformer network for nested named entity recognition”. In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. 2021, pp. 124–132.

[3030] Yongqing Zhang et al. “Uncovering the relationship between tissue-specific TF-DNA binding and chromatin features through a transformer-based model”. In: *Genes* 13.11 (2022), p. 1952.

[3031] Tom Sexton, Frédéric Bantignies, and Giacomo Cavalli. “Genomic interactions: chromatin loops and gene meeting points in transcriptional regulation”. In: *Seminars in cell & developmental biology*. Vol. 20. 7. Elsevier. 2009, pp. 849–855.

[3032] Erin Pleasance et al. "Pan-cancer analysis of advanced patient tumors reveals interactions between therapy and genomic landscapes". In: *Nature Cancer* 1.4 (2020), pp. 452–468.

[3033] Di Yao et al. "TrajGAT: A graph-based long-term dependency modeling approach for trajectory similarity computation". In: *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*. 2022, pp. 2275–2285.

[3034] Dohoon Lee, Jeewon Yang, and Sun Kim. "Learning the histone codes with large genomic windows and three-dimensional chromatin interactions using transformer". In: *Nature Communications* 13.1 (2022), p. 6678.

[3035] Yang Li et al. "Heterogeneous transformer: A scale adaptable neural network architecture for device activity detection". In: *IEEE Transactions on Wireless Communications* 22.5 (2022), pp. 3432–3446.

[3036] Jiayu Shang et al. "Accurate identification of bacteriophages from metagenomic data using Transformer". In: *Briefings in Bioinformatics* 23.4 (2022), bbac258.

[3037] Amoon Jamzad et al. "Graph transformers for characterization and interpretation of surgical margins". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24*. Springer. 2021, pp. 88–97.

[3038] Theodore T Jiang, Li Fang, and Kai Wang. "Deciphering the language of nature: A transformer-based language model for deleterious mutations in proteins". In: *The Innovation* 4.5 (2023).

[3039] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. "Waveglow: A flow-based generative network for speech synthesis". In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 3617–3621.

[3040] Xuezhe Ma et al. "Macow: Masked convolutional generative flow". In: *Advances in Neural Information Processing Systems* 32 (2019).

[3041] Xingjian Zhen et al. "Flow-based generative models for learning manifold to manifold mappings". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 12. 2021, pp. 11042–11052.

[3042] Anja Butter et al. "Generative networks for precision enthusiasts". In: *SciPost Physics* 14.4 (2023), p. 078.

[3043] Xinling Wang, Lu Lu, and Shibo Jiang. "SARS-CoV-2 evolution from the BA. 2.86 to JN. 1 variants: unexpected consequences". In: *Trends in Immunology* (2024).

[3044] Ashley C Hoyte et al. "Resistance to pyridine-based inhibitor KF116 reveals an unexpected role of integrase in HIV-1 Gag-Pol polyprotein proteolytic processing". In: *Journal of Biological Chemistry* 292.48 (2017), pp. 19814–19825.

[3045] Katrina A Lythgoe and Christophe Fraser. “New insights into the evolutionary rate of HIV-1 at the within-host and epidemiological levels”. In: *Proceedings of the Royal Society B: Biological Sciences* 279.1741 (2012), pp. 3367–3375.

[3046] Anna Garbelli et al. “How to win the HIV-1 drug resistance hurdle race: running faster or jumping higher?” In: *Biochemical Journal* 474.10 (2017), pp. 1559–1577.

[3047] Eric Nguyen et al. “Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution”. In: *Advances in neural information processing systems* 36 (2024).

[3048] Minghua Deng et al. “Mapping gene ontology to proteins based on protein–protein interaction data”. In: *Bioinformatics* 20.6 (2004), pp. 895–902.

[3049] Shujun He et al. “Nucleic Transformer: Classifying DNA Sequences with Self-Attention and Convolutions”. In: *ACS Synthetic Biology* 12.11 (2023), pp. 3205–3214.

[3050] Zuher Jahshan, Esteban Garzón, and Leonid Yavits. “ViRAL: Vision transformer based accelerator for ReAL time lineage assignment of viral pathogens”. In: *IEEE Access* (2024).

[3051] Yijia Zheng et al. “Learning manifold dimensions with conditional variational autoencoders”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 34709–34721.

[3052] Antoine Plumerault, Hervé Le Borgne, and Céline Hudelot. “AVAE: adversarial variational auto encoder”. In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, pp. 8687–8694.

[3053] Te-Cheng Hsu and Che Lin. “Learning from small medical datarobust semi-supervised cancer prognosis classifier with Bayesian variational autoencoder”. In: *Bioinformatics Advances* 3.1 (2023), vbac100.

[3054] Oskar Wysocki et al. “Transformers and the representation of biomedical background knowledge”. In: *Computational Linguistics* 49.1 (2023), pp. 73–115.

[3055] Tianle Ma and Aidong Zhang. “Incorporating biological knowledge with factor graph neural network for interpretable deep learning”. In: *arXiv preprint arXiv:1906.00537* (2019).

[3056] Zhengliang Liu et al. “Deid-gpt: Zero-shot medical text de-identification by gpt-4”. In: *arXiv preprint arXiv:2303.11032* (2023).

[3057] Ioan Ieremie, Rob M Ewing, and Mahesan Niranjan. “TransformerGO: predicting protein–protein interactions by modelling the attention between sets of gene ontology terms”. In: *Bioinformatics* 38.8 (2022), pp. 2269–2277.

[3058] Ting-He Zhang et al. “Transformer for Gene Expression Modeling (T-GEM): An Interpretable Deep Learning Model for Gene Expression-Based Phenotype Predictions”. In: *Cancers* 14.19 (2022), p. 4763.

[3059] Fathi M Salem and Fathi M Salem. "Recurrent Neural Networks (RNN)". In: *Recurrent Neural Networks: From Simple to Gated Architectures* (2022), pp. 43–67.

[3060] Antônio H Ribeiro et al. "Beyond exploding and vanishing gradients: analysing RNN training using attractors and smoothness". In: *International conference on artificial intelligence and statistics*. PMLR. 2020, pp. 2370–2380.

[3061] Nguyen Quoc Khanh Le. "Leveraging transformers-based language models in proteome bioinformatics". In: *Proteomics* 23.23-24 (2023), p. 2300011.

[3062] Shigeki Karita et al. "A comparative study on transformer vs rnn in speech applications". In: *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*. IEEE. 2019, pp. 449–456.

[3063] Ilya Sutskever, James Martens, and Geoffrey E Hinton. "Generating text with recurrent neural networks". In: *Proceedings of the 28th international conference on machine learning (ICML-11)*. 2011, pp. 1017–1024.

[3064] Miguel Graça et al. "Distributed transformer for high order epistasis detection in large-scale datasets". In: *Scientific Reports* 14.1 (2024), p. 14579.

[3065] Emma Yann Zhang et al. "From Turing to transformers: A comprehensive review and tutorial on the evolution and applications of generative transformer models". In: *Sci* 5.4 (2023), p. 46.

[3066] Yafei Xiang et al. "Text Understanding and Generation Using Transformer Models for Intelligent E-commerce Recommendations". In: *arXiv preprint arXiv:2402.16035* (2024).

[3067] Ruijin Liu et al. "PBFormer: Capturing Complex Scene Text Shape with Polynomial Band Transformer". In: *Proceedings of the 31st ACM International Conference on Multimedia*. 2023, pp. 2112–2120.

[3068] Zhufeng Li et al. "Whole Genome Transformer for Gene Interaction Effects in Microbiome Habitat Specificity". In: *arXiv preprint arXiv:2405.05998* (2024).

[3069] Atul Sharma et al. "Lerna: transformer architectures for configuring error correction tools for short-and long-read genome sequencing". In: *BMC bioinformatics* 23.1 (2022), p. 25.

[3070] Wei Li et al. "Prediction of SARS-CoV-2 Infection Phosphorylation Sites and Associations of these Modifications with Lung Cancer Development". In: *Current Gene Therapy* 24.3 (2024), pp. 239–248.

[3071] Nguyen Quoc Khanh Le and Quang-Thai Ho. "Deep transformers and convolutional neural network in identifying DNA N6-methyladenine sites in cross-species genomes". In: *Methods* 204 (2022), pp. 199–206.

[3072] Binbin Zhou et al. "TEMPO: A transformer-based mutation prediction framework for SARS-CoV-2 evolution". In: *Computers in Biology and Medicine* 152 (2023), p. 106264.

[3073] Ladislav Rampáek et al. “Recipe for a general, powerful, scalable graph transformer”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 14501–14515.

[3074] Emily Kaczmarek et al. “Multi-omic graph transformers for cancer classification and interpretation”. In: *Pacific Symposium on Biocomputing 2022*. World Scientific. 2021, pp. 373–384.

[3075] Anthony Gillioz et al. “Overview of the Transformer-based Models for NLP Tasks”. In: *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*. IEEE. 2020, pp. 179–183.

[3076] Yang Zhao, Jiajun Zhang, and Chengqing Zong. “Transformer: A general framework from machine translation to others”. In: *Machine Intelligence Research* 20.4 (2023), pp. 514–538.

[3077] Hanqing Zhang et al. “A survey of controllable text generation using transformer-based pre-trained language models”. In: *ACM Computing Surveys* 56.3 (2023), pp. 1–37.

[3078] Soonbeom Choi and Juhan Nam. “A melody-unsupervision model for singing voice synthesis”. In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 7242–7246.

[3079] R Ranjith and A Chandrasekar. “GTSO: Gradient tangent search optimization enabled voice transformer with speech intelligibility for aphasia”. In: *Computer Speech & Language* 84 (2024), p. 101568.

[3080] Arshi Parvaiz et al. “Vision transformers in medical computer visionA contemplative retrospection”. In: *Engineering Applications of Artificial Intelligence* 122 (2023), p. 106126.

[3081] Xin Deng et al. “Classifying ASD based on time-series fMRI using spatial–temporal transformer”. In: *Computers in biology and medicine* 151 (2022), p. 106320.

[3082] Xin Xie et al. “From discrimination to generation: Knowledge graph completion with generative transformer”. In: *Companion Proceedings of the Web Conference 2022*. 2022, pp. 162–165.

[3083] Zihang Dai et al. “Transformer-xl: Language modeling with longer-term dependency”. In: (2018).

[3084] Martin R Chavez et al. “Chat Generative Pre-trained Transformer: why we should embrace this technology”. In: *American Journal of Obstetrics and Gynecology* 228.6 (2023), pp. 706–711.

[3085] Kyle Hamilton, Luca Longo, and Bojan Bozic. “GPT Assisted Annotation of Rhetorical and Linguistic Features for Interpretable Propaganda Technique Detection in News Text.” In: *Companion Proceedings of the ACM on Web Conference 2024*. 2024, pp. 1431–1440.

[3086] Gokul Yenduri et al. "GPT (Generative Pre-trained Transformer)–A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions". In: *IEEE Access* (2024).

[3087] R Thomas McCoy et al. "How much do language models copy from their training data? evaluating linguistic novelty in text generation using raven". In: *Transactions of the Association for Computational Linguistics* 11 (2023), pp. 652–670.

[3088] Hanqing Zhang et al. "A survey of controllable text generation using transformer-based pre-trained language models". In: *ACM Computing Surveys* 56.3 (2023), pp. 1–37.

[3089] Anwar Khan and Boreom Lee. "DeepGene transformer: Transformer for the gene expression-based classification of cancer subtypes". In: *Expert Systems with Applications* 226 (2023), p. 120047.

[3090] Xiaodan Zhang et al. "A deep learning approach for orphan gene identification in moso bamboo (Phyllostachys edulis) based on the CNN+ Transformer model". In: *BMC bioinformatics* 23.1 (2022), p. 162.

[3091] Gi Bae Kim et al. "Functional annotation of enzyme-encoding genes using deep learning with transformer layers". In: *Nature Communications* 14.1 (2023), p. 7370.

[3092] Jonathan Raad et al. "miRe2e: a full end-to-end deep model based on transformers for prediction of pre-miRNAs". In: *Bioinformatics* 38.5 (2022), pp. 1191–1197.

[3093] Zhonghui Gu et al. "Hierarchical graph transformer with contrastive learning for protein function prediction". In: *Bioinformatics* 39.7 (2023), btad410.

[3094] Wenhuan Zeng, Anupam Gautam, and Daniel H Huson. "MuLan-Methylmultiple transformer-based language models for accurate DNA methylation prediction". In: *GigaScience* 12 (2023), giad054.

[3095] Joshua A Hendrickson et al. "Antifungal resistance: a concerning trend for the present and future". In: *Current infectious disease reports* 21 (2019), pp. 1–8.

[3096] Matthew C Fisher et al. "Tackling the emerging threat of antifungal resistance to human health". In: *Nature reviews microbiology* 20.9 (2022), pp. 557–571.

[3097] David S Perlin, Riina Rautemaa-Richardson, and Ana Alastruey-Izquierdo. "The global problem of antifungal resistance: prevalence, mechanisms, and management". In: *The Lancet infectious diseases* 17.12 (2017), e383–e392.

[3098] Yunjin Lee, Nicole Robbins, and Leah E Cowen. "Molecular mechanisms governing antifungal drug resistance". In: *npj Antimicrobials and Resistance* 1.1 (2023), p. 5.

[3099] Theodore C White, Kieren A Marr, and Raleigh A Bowden. "Clinical, cellular, and molecular factors that contribute to antifungal drug resistance". In: *Clinical microbiology reviews* 11.2 (1998), pp. 382–402.

[3100] Zeina A Kanafani and John R Perfect. "Resistance to antifungal agents: mechanisms and clinical impact". In: *Clinical infectious diseases* 46.1 (2008), pp. 120–128.

[3101] Ana Espinel-Ingroff. "Clinical relevance of antifungal resistance". In: *Infectious disease clinics of North America* 11.4 (1997), pp. 929–944.

[3102] Alexandre M Fuentefria et al. "Antifungals discovery: an insight into new strategies to combat antifungal resistance". In: *Letters in applied microbiology* 66.1 (2018), pp. 2–13.

[3103] Hsiang-Kuang Tseng and John R Perfect. "Strategies to manage antifungal drug resistance". In: *Expert opinion on pharmacotherapy* 12.2 (2011), pp. 241–256.

[3104] B Anne Neville, Christophe d'Enfert, and Marie-Elisabeth Bougnoux. "Candida albicans commensalism in the gastrointestinal tract". In: *FEMS yeast research* 15.7 (2015), fov081.

[3105] Heather R Conti and Sarah L Gaffen. "IL-17–Mediated immunity to the opportunistic fungal pathogen Candida albicans". In: *The Journal of Immunology* 195.3 (2015), pp. 780–788.

[3106] Markus Ruhnke. "Epidemiology of Candida albicans infections and role of non-Candidaalbicans yeasts". In: *Current drug targets* 7.4 (2006), pp. 495–504.

[3107] Jose A Vazquez et al. "Nosocomial acquisition of Candida albicans: an epidemiologic study". In: *Journal of Infectious Diseases* 168.1 (1993), pp. 195–201.

[3108] Somanon Bhattacharya, Sutthichai Sae-Tia, and Bettina C Fries. "Candidiasis and mechanisms of antifungal resistance". In: *Antibiotics* 9.6 (2020), p. 312.

[3109] Stephanie A Flowers et al. "Contribution of clinically derived mutations in ERG11 to azole resistance in Candida albicans". In: *Antimicrobial agents and chemotherapy* 59.1 (2015), pp. 450–460.

[3110] Andrea Patricia Cortés Hidalgo et al. "Opciones terapéuticas frente a especies de Candida resistentes a las equinocandinas". In: *Universitas Medica* 59.2 (2018), pp. 37–51.

[3111] Amparo Valentín et al. "Actividad in vitro de la anfotericina B y la anidulafungina sobre biopelículas de Candida albicans y Candida tropicalis". In: *Revista Iberoamericana de Micología* 24.4 (2007), pp. 272–277.

[3112] Sarah G Whaley et al. "Azole antifungal resistance in Candida albicans and emerging non-albicans Candida species". In: *Frontiers in microbiology* 7 (2017), p. 231851.

[3113] Andrew T Nishimoto, Cheshta Sharma, and P David Rogers. "Molecular and genetic basis of azole antifungal resistance in the opportunistic pathogenic fungus Candida albicans". In: *Journal of Antimicrobial Chemotherapy* 75.2 (2020), pp. 257–270.

[3114] Manuela Schuetzer-Muehlbauer et al. "Reversal of antifungal resistance mediated by ABC efflux pumps from Candida albicans functionally expressed in yeast". In: *International journal of antimicrobial agents* 22.3 (2003), pp. 291–300.

[3115] Babak Pourakbari et al. "Expression of major efflux pumps in fluconazole-resistant Candida albicans". In: *Infectious Disorders-Drug Targets (Formerly Current Drug Targets-Infectious Disorders)* 17.3 (2017), pp. 178–184.

[3116] Jyotsna Chandra et al. "Biofilm formation by the fungal pathogen Candida albicans: development, architecture, and drug resistance". In: *Journal of bacteriology* 183.18 (2001), pp. 5385–5394.

[3117] R Pereira et al. "Biofilm of Candida albicans: formation, regulation and resistance". In: *Journal of Applied Microbiology* 131.1 (2021), pp. 11–22.

[3118] Cécile Gautier, Eli I Maciel, and Iuliana V Ene. "Approaches for identifying and measuring heteroresistance in azole-susceptible Candida isolates". In: *Microbiology Spectrum* 12.4 (2024), e04041–23.

[3119] Ronen Ben-Ami et al. "Heteroresistance to fluconazole is a continuously distributed phenotype among Candida glabrata clinical strains associated with in vivo persistence". In: *MBio* 7.4 (2016), pp. 10–1128.

[3120] Daniel Poulain. "Candida albicans, plasticity and pathogenesis". In: *Critical reviews in microbiology* 41.2 (2015), pp. 208–217.

[3121] Marianne Frieri, Krishan Kumar, and Anthony Boutin. "Antibiotic resistance". In: *Journal of infection and public health* 10.4 (2017), pp. 369–378.

[3122] Niel Woodford and Matthew J Ellington. "The emergence of antibiotic resistance by mutation". In: *Clinical Microbiology and Infection* 13.1 (2007), pp. 5–18.

[3123] Philip S Stewart and J William Costerton. "Antibiotic resistance of bacteria in biofilms". In: *The lancet* 358.9276 (2001), pp. 135–138.

[3124] Thien-Fah Mah. "Biofilm-specific antibiotic resistance". In: *Future microbiology* 7.9 (2012), pp. 1061–1072.

[3125] Dongchang Sun et al. "Horizontal gene transfer mediated bacterial antibiotic resistance". In: *Frontiers in microbiology* 10 (2019), p. 478460.

[3126] Christian JH Von Wintersdorff et al. "Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer". In: *Frontiers in microbiology* 7 (2016), p. 173.

[3127] Johan Bengtsson-Palme, Erik Kristiansson, and DG Joakim Larsson. "Environmental factors influencing the development and spread of antibiotic resistance". In: *FEMS microbiology reviews* 42.1 (2018), fux053.

[3128] B Keith English and Aditya H Gaur. "The use and abuse of antibiotics and the development of antibiotic resistance". In: *Hot topics in infection and immunity in children VI* (2010), pp. 73–82.

[3129] Vangelis Economou and Panagiota Gousia. "Agriculture and food animals as a source of antimicrobial-resistant bacteria". In: *Infection and drug resistance* (2015), pp. 49–61.

[3130] Marc Lipsitch, Carl T Bergstrom, and Bruce R Levin. "The epidemiology of antibiotic resistance in hospitals: paradoxes and prescriptions". In: *Proceedings of the National Academy of Sciences* 97.4 (2000), pp. 1938–1943.

[3131] Maurizio Ferri et al. "Antimicrobial resistance: A global emerging threat to public health systems". In: *Critical reviews in food science and nutrition* 57.13 (2017), pp. 2857–2876.

[3132] N Deborah Friedman, Elizabeth Temkin, and Yehuda Carmeli. "The negative impact of antibiotic resistance". In: *Clinical Microbiology and Infection* 22.5 (2016), pp. 416–422.

[3133] Chang-Ro Lee et al. "Strategies to minimize antibiotic resistance". In: *International journal of environmental research and public health* 10.9 (2013), pp. 4274–4305.

[3134] Benjamin D Brooks and Amanda E Brooks. "Therapeutic strategies to combat antibiotic resistance". In: *Advanced drug delivery reviews* 78 (2014), pp. 14–27.

[3135] Evelina Tacconelli et al. "Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis". In: *The Lancet infectious diseases* 18.3 (2018), pp. 318–327.

[3136] Isabelle Vergne et al. "Cell biology of Mycobacterium tuberculosis phagosome". In: *Annu. Rev. Cell Dev. Biol.* 20 (2004), pp. 367–394.

[3137] Liem Nguyen. "Antibiotic resistance mechanisms in M. tuberculosis: an update". In: *Archives of toxicology* 90 (2016), pp. 1585–1604.

[3138] Lawrence Peter Ormerod. "Multidrug-resistant tuberculosis (MDR-TB): epidemiology, prevention and treatment". In: *British medical bulletin* 73.1 (2005), pp. 17–24.

[3139] AI De La Iglesia and Y HR Morbidoni. "Mecanismos de acción y de resistencia a rifampicina e isoniacida en Mycobacterium tuberculosis: nueva información sobre viejos conocidos". In: *Revista Argentina de microbiologia* 38.2 (2006), pp. 97–109.

[3140] Hiroki Ando et al. "Identification of katG mutations associated with high-level isoniazid resistance in Mycobacterium tuberculosis". In: *Antimicrobial agents and chemotherapy* 54.5 (2010), pp. 1793–1799.

[3141] Todd J Sullivan et al. "High affinity InhA inhibitors with activity against drug-resistant strains of Mycobacterium tuberculosis". In: *ACS chemical biology* 1.1 (2006), pp. 43–53.

[3142] Hatsumi Taniguchi et al. "Rifampicin resistance and mutation of the rpoB gene in Mycobacterium tuberculosis". In: *FEMS microbiology letters* 144.1 (1996), pp. 103–108.

[3143] Robert C Goldman, Kevin V Plumley, and Barbara E Laughon. "The evolution of extensively drug resistant tuberculosis (XDR-TB): history, status and issues for global control". In: *Infectious Disorders-Drug Targets (Formerly Current Drug Targets-Infectious Disorders)* 7.2 (2007), pp. 73–91.

[3144] José A Caminero et al. "Diagnóstico y tratamiento de la tuberculosis con resistencia a fármacos". In: *Archivos de bronconeumología* 53.9 (2017), pp. 501–509.

[3145] Elisea Avalos et al. "Frequency and geographic distribution of gyrA and gyrB mutations associated with fluoroquinolone resistance in clinical Mycobacterium tuberculosis isolates: a systematic review". In: *PloS one* 10.3 (2015), e0120470.

[3146] Priti Kambli et al. "Correlating rrs and eis promoter mutations in clinical isolates of Mycobacterium tuberculosis with phenotypic susceptibility levels to the second-line injectables". In: *The International Journal of Mycobacteriology* 5.1 (2016), pp. 1–6.

[3147] Parissa Farnia et al. "Growth and cell-division in extensive (XDR) and extremely drug resistant (XXDR) tuberculosis strains: transmission and atomic force observation". In: *International journal of clinical and experimental medicine* 3.4 (2010), p. 308.

[3148] Luisa Maria Nieto Ramirez, Karina Quintero Vargas, and Gustavo Diaz. "Whole genome sequencing for the analysis of drug resistant strains of Mycobacterium tuberculosis: a systematic review for bedaquiline and delamanid". In: *Antibiotics* 9.3 (2020), p. 133.

[3149] Thomas R Ioerger et al. "Identification of new drug targets and resistance mechanisms in Mycobacterium tuberculosis". In: *PloS one* 8.9 (2013), e75245.

[3150] Pedro Eduardo Almeida da Silva et al. "Efflux as a mechanism for drug resistance in Mycobacterium tuberculosis". In: *FEMS Immunology & Medical Microbiology* 63.1 (2011), pp. 1–9.

[3151] Neelima Gorityala, Anthony Samit Baidya, and Someswar R Sagurthi. "Genome mining of Mycobacterium tuberculosis: targeting SufD as a novel drug candidate through in silico characterization and inhibitor screening". In: *Frontiers in Microbiology* 15 (2024), p. 1369645.

[3152] Aanchal Kashyap, Pankaj Kumar Singh, and Om Silakari. "Mechanistic investigation of resistance via drug-inactivating enzymes in Mycobacterium tuberculosis". In: *Drug metabolism reviews* 50.4 (2018), pp. 448–465.

[3153] Tamanna Tasnim et al. "Pre-extensively drug resistant tuberculosis (Pre-XDR-TB) among pulmonary multidrug resistant tuberculosis (MDR-TB) patients in Bangladesh". In: *Journal of Tuberculosis Research* 6.3 (2018), pp. 199–206.

[3154] Ethel Leonor Noia Maciel et al. "Adverse effects of the new tuberculosis treatment regimen recommended by the Brazilian Ministry of Health". In: *Jornal Brasileiro de Pneumologia* 36 (2010), pp. 232–238.

[3155] Marlo Möller et al. "Genetic resistance to Mycobacterium tuberculosis infection and disease". In: *Frontiers in immunology* 9 (2018), p. 410221.

[3156] Amanda Mabhula and Vinayak Singh. "Drug-resistance in Mycobacterium tuberculosis: where we stand". In: *Medchemcomm* 10.8 (2019), pp. 1342–1360.

[3157] Vinayak Singh and Kelly Chibale. "Strategies to combat multi-drug resistance in tuberculosis". In: *Accounts of chemical research* 54.10 (2021), pp. 2361–2376.

[3158] Wiep Klaas Smits et al. "Clostridium difficile infection". In: *Nature reviews Disease primers* 2.1 (2016), pp. 1–20.

[3159] Daniel Paredes-Sabja, Aimee Shen, and Joseph A Sorg. "Clostridium difficile spore biology: sporulation, germination, and spore structural proteins". In: *Trends in microbiology* 22.7 (2014), pp. 406–416.

[3160] Daniel E Voth and Jimmy D Ballard. "Clostridium difficile toxins: mechanism of action and role in disease". In: *Clinical microbiology reviews* 18.2 (2005), pp. 247–263.

[3161] Jhon Walter Zea and Clara Lina Salazar. "Enfermedad asociada a Clostridium difficile: prevalencia y diagnóstico por laboratorio". In: *Infectio* 16.4 (2012), pp. 211–222.

[3162] Torbjorn Norén et al. "Clindamycin resistant strains of Clostridium difficile isolated from cases of C. difficile associated diarrhea (CDAD) in a hospital in Sweden". In: *Diagnostic microbiology and infectious disease* 42.2 (2002), pp. 149–151.

[3163] Patrizia Spigaglia et al. "Fluoroquinolone resistance in Clostridium difficile isolates from a prospective study of C. difficile infections in Europe". In: *Journal of medical microbiology* 57.6 (2008), pp. 784–789.

[3164] T Pelaez et al. "Reassessment of Clostridium difficile susceptibility to metronidazole and vancomycin". In: *Antimicrobial agents and chemotherapy* 46.6 (2002), pp. 1647–1650.

[3165] Rawish Fatima and Muhammad Aziz. "The hypervirulent strain of Clostridium difficile: NAP1/B1/027-a brief overview". In: *Cureus* 11.1 (2019).

[3166] Barry Cookson. "Hypervirulent strains of Clostridium difficile". In: *Postgraduate medical journal* 83.979 (2007), pp. 291–295.

[3167] Patrizia Spigaglia. "Recent advances in the understanding of antibiotic resistance in Clostridium difficile infection". In: *Therapeutic advances in infectious disease* 3.1 (2016), pp. 23–42.

[3168] L Dridi, J Tankovic, and J-C Petit. "CdeA of Clostridium difficile, a new multidrug efflux transporter of the MATE family". In: *Microbial drug resistance* 10.3 (2004), pp. 191–196.

[3169] Marta Toth et al. "Intrinsic class D β-lactamases of Clostridium difficile". In: *MBio* 9.6 (2018), pp. 10–1128.

[3170] Tanya M Monaghan et al. "Clostridioides difficile: innovations in target discovery and potential for therapeutic success". In: *Expert Opinion on Therapeutic Targets* 25.11 (2021), pp. 949–963.

[3171] Matthew Samore. "Epidemiology of nosocomial Clostridium difficile diarrhoea". In: *Journal of hospital infection* 43 (1999), S183–S190.

[3172] Brett G Mitchell and Anne Gardner. "Mortality and Clostridium difficile infection: a review". In: *Antimicrobial resistance and infection control* 1 (2012), pp. 1–6.

[3173] Jennie H Kwon, Margaret A Olsen, and Erik R Dubberke. "The morbidity, mortality, and costs associated with Clostridium difficile infection". In: *Infectious Disease Clinics* 29.1 (2015), pp. 123–134.

[3174] Louis B Rice. "The Maxwell Finland Lecture: for the durationrational antibiotic administration in an era of antimicrobial resistance and Clostridium difficile". In: *Clinical infectious diseases* 46.4 (2008), pp. 491–496.

[3175] George G Zhanel, Andrew J Walkty, James A Karlowsky, et al. "Fidaxomicin: a novel agent for the treatment of Clostridium difficile infection". In: *Canadian Journal of Infectious Diseases and Medical Microbiology* 26 (2015), pp. 305–312.

[3176] Matthew Austin, Mark Mellow, and William M Tierney. "Fecal microbiota transplantation in the treatment of Clostridium difficile infections". In: *The American journal of medicine* 127.6 (2014), pp. 479–483.

[3177] William A Rutala and David J Weber. "Role of the hospital environment in disease transmission, with a focus on Clostridium difficile". In: *Healthcare infection* 18.1 (2013), pp. 14–22.

[3178] R-P Vonberg et al. "Infection control measures to limit the spread of Clostridium difficile". In: *Clinical Microbiology and Infection* 14 (2008), pp. 2–20.

[3179] Dhara Shah et al. "Clostridium difficile infection: update on emerging antibiotic treatment options and antibiotic resistance". In: *Expert review of anti-infective therapy* 8.5 (2010), pp. 555–564.

[3180] Chandrabali Ghose and Ciarán P Kelly. "The prospect for vaccines to prevent Clostridium difficile infection". In: *Infectious Disease Clinics* 29.1 (2015), pp. 145–162.

[3181] Séverine Péchiné et al. "Targeting Clostridium difficile surface components to develop immunotherapeutic strategies against Clostridium difficile infection". In: *Frontiers in microbiology* 9 (2018), p. 1009.

[3182] Louis B Rice. "Emergence of vancomycin-resistant enterococci." In: *Emerging infectious diseases* 7.2 (2001), p. 183.

[3183] Yesim Cetinkaya, Pamela Falk, and C Glen Mayhall. "Vancomycin-resistant enterococci". In: *Clinical microbiology reviews* 13.4 (2000), pp. 686–707.

[3184] Krista Dubin and Eric G Pamer. "Enterococci and their interactions with the intestinal microbiome". In: *Microbiology spectrum* 5.6 (2017), pp. 10–1128.

[3185] Gary A Noskin, Lance R Peterson, and John R Warren. "Enterococcus faecium and Enterococcus faecalis bacteremia: acquisition and outcome". In: *Clinical infectious diseases* 20.2 (1995), pp. 296–301.

[3186] Patricia Cornejo-Juárez et al. "Trend of antimicrobial drug-susceptibility of blood isolates at an oncological center (1998-2003)." In: *Salud Pública de México* 47.4 (2005), pp. 288–293.

[3187] Kelli L Palmer, Veronica N Kos, and Michael S Gilmore. "Horizontal gene transfer and the genomics of enterococcal antibiotic resistance". In: *Current opinion in microbiology* 13.5 (2010), pp. 632–639.

[3188] Taskeen Raza et al. "Vancomycin resistant Enterococci: A brief review". In: *J Pak Med Assoc* 68.5 (2018), pp. 768–772.

[3189] Stefan Evers, Peter E Reynolds, and Patrice Courvalin. "Sequence of the vanB and ddl genes encoding D-alanine: D-lactate and D-alanine: D-alanine ligases in vancomycin-resistant Enterococcus faecalis V583". In: *Gene* 140.1 (1994), pp. 97–102.

[3190] Li-Hwei Sng et al. "Antimicrobial susceptibility testing of a clinical isolate of vancomycin-dependent enterococcus using D-alanine-D-alanine as a growth supplement". In: *American journal of clinical pathology* 109.4 (1998), pp. 399–403.

[3191] Peter E Reynolds and Patrice Courvalin. "Vancomycin resistance in enterococci due to synthesis of precursors terminating in D-alanyl-D-serine". In: *Antimicrobial agents and chemotherapy* 49.1 (2005), pp. 21–25.

[3192] Stephan Harbarth, Sara Cosgrove, and Yehuda Carmeli. "Effects of antibiotics on nosocomial epidemiology of vancomycin-resistant enterococci". In: *Antimicrobial agents and chemotherapy* 46.6 (2002), pp. 1619–1628.

[3193] Emilia Cercenado. "Enterococcus: resistencias fenotípicas y genotípicas y epidemiología en España". In: *Enfermedades Infecciosas y Microbiología Clínica* 29 (2011), pp. 59–65.

[3194] John M Boyce. "Vancomycin-resistant enterococcus: detection, epidemiology, and control measures". In: *Infectious disease clinics of North America* 11.2 (1997), pp. 367–384.

[3195] Simona de Niederhäusern et al. "Vancomycin-resistance transferability from VanA enterococci to Staphylococcus aureus". In: *Current microbiology* 62 (2011), pp. 1363–1367.

[3196] Katherine Shea et al. "Successful treatment of vancomycin-resistant Enterococcus faecium pyelonephritis with daptomycin during pregnancy". In: *Annals of Pharmacotherapy* 42.5 (2008), pp. 722–725.

[3197] Jatapat Hemapanpairoa et al. "Vancomycin-resistant enterococcal infection in a Thai university hospital: clinical characteristics, treatment outcomes, and synergistic effect". In: *Infection and Drug Resistance* (2019), pp. 2049–2057.

[3198] Nico T Mutters et al. "Control of the spread of vancomycin-resistant enterococci in hospitals: epidemiology and clinical relevance". In: *Deutsches Ärzteblatt International* 110.43 (2013), p. 725.

[3199] Peter K Linden. "Treatment options for vancomycin-resistant enterococcal infections". In: *Drugs* 62 (2002), pp. 425–441.

[3200] Ermioni Kalfopoulou and Johannes Huebner. "Advances and prospects in vaccine development against enterococci". In: *Cells* 9.11 (2020), p. 2397.

[3201] Jatinder Kaur et al. "Optimization of acetazolamide-based scaffold as potent inhibitors of vancomycin-resistant Enterococcus". In: *Journal of medicinal chemistry* 63.17 (2020), pp. 9540–9562.

[3202] Yehuda Carmeli et al. "Health and economic outcomes of vancomycin-resistant enterococci". In: *Archives of internal medicine* 162.19 (2002), pp. 2223–2228.

[3203] Keiichi Hiramatsu et al. "The emergence and evolution of methicillin-resistant Staphylococcus aureus". In: *Trends in microbiology* 9.10 (2001), pp. 486–493.

[3204] Andie S Lee et al. "Methicillin-resistant Staphylococcus aureus". In: *Nature reviews Disease primers* 4.1 (2018), pp. 1–23.

[3205] G Guillén Fiel et al. "Bullous impetigo caused by methicillin-resistant Staphylococcus aureus". In: *Acta Pediatrica Espanola* 66.8 (2008), p. 415.

[3206] M Yilmaz et al. "First report of catalase-negative methicillin-resistant Staphylococcus aureus sepsis". In: *Journal of hospital infection* 60.2 (2005), pp. 188–189.

[3207] Inaam A Al Obaid et al. "Isolation and characterization of coagulase-negative methicillin-resistant Staphylococcus aureus from patients in an intensive care unit". In: *Medical Principles and Practice* 8.3 (1999), pp. 230–236.

[3208] LA Mandell and R Wunderink. *Methicillin-resistant Staphylococcus aureus and community-acquired pneumonia: an evolving relationship*. 2012.

[3209] Maribel J Castellano González and Armindo J Perozo-Mena. "Mecanismos de resistencia a antibióticos β-lactámicos en Staphylococcus aureus". In: *Kasmera* 38.1 (2010), pp. 18–35.

[3210] Daniel Lim and Natalie CJ Strynadka. "Structural basis for the β lactam resistance of PBP2a from methicillin-resistant Staphylococcus aureus". In: *Nature structural biology* 9.11 (2002), pp. 870–876.

[3211] H Hanaki et al. "Activated cell-wall synthesis is associated with vancomycin resistance in methicillin-resistant Staphylococcus aureus clinical strains Mu3 and Mu50". In: *The Journal of antimicrobial chemotherapy* 42.2 (1998), pp. 199–209.

[3212] Junyan Liu et al. "Staphylococcal chromosomal cassettes mec (SCCmec): A mobile genetic element in methicillin-resistant Staphylococcus aureus". In: *Microbial pathogenesis* 101 (2016), pp. 56–67.

[3213] Raimundo Castro-Orozco et al. "Methicillin-resistant Staphylococcus aureus in children attending school in Cartagena, Colombia". In: *Revista de Salud Pública* 12.3 (2010), pp. 454–463.

[3214] Noelle I Samia et al. "Methicillin-resistant staphylococcus aureus nosocomial infection has a distinct epidemiological position and acts as a marker for overall hospital-acquired infection trends". In: *Scientific reports* 12.1 (2022), p. 17007.

[3215] José R Mediavilla et al. "Global epidemiology of community-associated methicillin resistant Staphylococcus aureus (CA-MRSA)". In: *Current opinion in microbiology* 15.5 (2012), pp. 588–595.

[3216] Michelle L Campbell et al. "Treatment of methicillin-resistant Staphylococcus aureus infections with a minimal inhibitory concentration of 2 μg/mL to vancomycin: old (trimethoprim/sulfamethoxazole) versus new (daptomycin or linezolid) agents". In: *Annals of Pharmacotherapy* 46.12 (2012), pp. 1587–1597.

[3217] Loreen A Herwaldt. "Control of methicillin-resistant Staphylococcus aureus in the hospital setting". In: *The American journal of medicine* 106.5 (1999), pp. 11–18.

[3218] Michael W Ellis et al. "Hygiene strategies to prevent methicillin-resistant Staphylococcus aureus skin and soft tissue infections: a cluster-randomized controlled trial among high-risk military trainees". In: *Clinical Infectious Diseases* 58.11 (2014), pp. 1540–1548.

[3219] Simona Purrello et al. "Methicillin-resistant Staphylococcus aureus infections: A review of the currently available treatment options". In: *Journal of Global Antimicrobial Resistance* 7 (2016), pp. 178–186.

[3220] Palanichamy Nandhini et al. "Recent developments in methicillin-resistant Staphylococcus aureus (MRSA) treatment: a review". In: *Antibiotics* 11.5 (2022), p. 606.

[3221] David P Calfee et al. "Strategies to prevent transmission of methicillin-resistant Staphylococcus aureus in acute care hospitals". In: *Infection Control & Hospital Epidemiology* 29.S1 (2008), S62–S80.

[3222] JV Robotham, David Richard Jenkins, and Graham F Medley. "Screening strategies in surveillance and control of methicillin-resistant Staphylococcus aureus (MRSA)". In: *Epidemiology & Infection* 135.2 (2007), pp. 328–342.

[3223] Matteo Bassetti et al. "Important new therapies for methicillin-resistant Staphylococcus aureus". In: *Expert Opinion on Pharmacotherapy* 20.18 (2019), pp. 2317–2334.

[3224] Alex Soriano et al. "Influence of vancomycin minimum inhibitory concentration on the treatment of methicillin-resistant Staphylococcus aureus bacteremia". In: *Clinical Infectious Diseases* 46.2 (2008), pp. 193–200.

[3225] Mireille Goetghebeur et al. "Methicillin-resistant Staphylococcus aureus: a public health issue with economic consequences". In: *Canadian Journal of Infectious Diseases and Medical Microbiology* 18 (2007), pp. 27–34.

[3226] Magnus Unemo and William M Shafer. "Antimicrobial resistance in Neisseria gonorrhoeae in the 21st century: past, evolution, and future". In: *Clinical microbiology reviews* 27.3 (2014), pp. 587–613.

[3227] Deborah M Tobiason and H Steven Seifert. "The obligate human pathogen, Neisseria gonorrhoeae, is polyploid". In: *PLoS biology* 4.6 (2006), e185.

[3228] Frederik S Archibald and Minh-ngoc Duong. "Superoxide dismutase and oxygen toxicity defenses in the genus Neisseria". In: *Infection and immunity* 51.2 (1986), pp. 631–641.

[3229] Sarah Jane Quillin and H Steven Seifert. "Neisseria gonorrhoeae host adaptation and pathogenesis". In: *Nature Reviews Microbiology* 16.4 (2018), pp. 226–240.

[3230] Magnus Unemo and William M Shafer. "Antibiotic resistance in Neisseria gonorrhoeae: origin, evolution, and lessons learned for the future". In: *Annals of the New York Academy of Sciences* 1230.1 (2011), E19–E28.

[3231] John W Tapsall. "Antibiotic resistance in Neisseria gonorrhoeae". In: *Clinical Infectious Diseases* 41.Supplement_4 (2005), S263–S268.

[3232] Makoto Ohnishi et al. "Spread of a chromosomal cefixime-resistant penA gene among different Neisseria gonorrhoeae lineages". In: *Antimicrobial agents and chemotherapy* 54.3 (2010), pp. 1060–1067.

[3233] Xiaomian Lin et al. "Dissemination and genome analysis of high-level ceftriaxone-resistant penA 60.001 Neisseria gonorrhoeae strains from the Guangdong Gonococcal antibiotics susceptibility Programme (GD-GASP), 2016–2019". In: *Emerging Microbes & Infections* 11.1 (2022), pp. 344–350.

[3234] Jason P Folster et al. "MtrR modulates rpoH expression and levels of antimicrobial resistance in Neisseria gonorrhoeae". In: *Journal of bacteriology* 191.1 (2009), pp. 287–297.

[3235] Kornelius Zeth et al. "Structure and function of the PorB porin from disseminating Neisseria gonorrhoeae". In: *Biochemical Journal* 449.3 (2013), pp. 631–642.

[3236] Melanie Olesky et al. "Porin-mediated antibiotic resistance in Neisseria gonorrhoeae: ion, solute, and antibiotic permeation through PIB proteins with penB mutations". In: *Journal of bacteriology* 188.7 (2006), pp. 2300–2308.

[3237] Sheeba Santhini Manoharan-Basil et al. "Evidence of horizontal gene transfer of 50S ribosomal genes rplB, rplD, and rplY in Neisseria gonorrhoeae". In: *Frontiers in Microbiology* 12 (2021), p. 683901.

[3238] Sheeba Santhini Manoharan-Basil et al. "Horizontal gene transfer of fluoroquinolone resistance-conferring genes from commensal Neisseria to Neisseria gonorrhoeae: a global phylogenetic analysis of 20,047 isolates". In: *Frontiers in Microbiology* 13 (2022), p. 793612.

[3239] Lai-King Ng et al. "Mutation in 23S rRNA associated with macrolide resistance in Neisseria gonorrhoeae". In: *Antimicrobial agents and chemotherapy* 46.9 (2002), pp. 3020–3025.

[3240] Mei Hu et al. "High-level chromosomally mediated tetracycline resistance in Neisseria gonorrhoeae results from a point mutation in the rpsJ gene encoding ribosomal protein S10 in combination with the mtrR and penB resistance determinants". In: *Antimicrobial agents and chemotherapy* 49.10 (2005), pp. 4327–4334.

[3241] Magnus Unemo, Carlos Del Rio, and William M Shafer. "Antimicrobial resistance expressed by Neisseria gonorrhoeae: a major global public health problem in the 21st century". In: *Emerging Infections 10* (2016), pp. 213–237.

[3242] John W Tapsall et al. "Meeting the public health challenge of multidrug- and extensively drug-resistant Neisseria gonorrhoeae". In: *Expert review of anti-infective therapy* 7.7 (2009), pp. 821–834.

[3243] David M Whiley et al. "The ticking time bomb: escalating antibiotic resistance in Neisseria gonorrhoeae is a public health disaster in waiting". In: *Journal of Antimicrobial Chemotherapy* 67.9 (2012), pp. 2059–2061.

[3244] World Health Organization et al. *Global action plan to control the spread and impact of antimicrobial resistance in Neisseria gonorrhoeae*. World Health Organization, 2012.

[3245] Fan Yang and Jing Yan. "Antibiotic resistance and treatment options for multidrug-resistant gonorrhea". In: *Infectious Microbes & Diseases* 2.2 (2020), pp. 67–76.

[3246] David Terkelsen et al. "Multidrug-resistant Neisseria gonorrhoeae infection with ceftriaxone resistance and intermediate resistance to azithromycin, Denmark, 2017". In: *Eurosurveillance* 22.42 (2017), pp. 17–00659.

[3247] David A Lewis. "New treatment options for Neisseria gonorrhoeae in the era of emerging antimicrobial resistance". In: *Sexual Health* 16.5 (2019), pp. 449–456.

[3248] Jutamas Shaughnessy et al. "A novel factor H–Fc chimeric immunotherapeutic molecule against Neisseria gonorrhoeae". In: *The Journal of Immunology* 196.4 (2016), pp. 1732–1740.

[3249] Gunther Kern et al. "Inhibition of Neisseria gonorrhoeae type II topoisomerases by the novel spiropyrimidinetrione AZD0914". In: *Journal of Biological Chemistry* 290.34 (2015), pp. 20984–20994.

[3250] Eric Y Lin, Paul C Adamson, and Jeffrey D Klausner. "Epidemiology, treatments, and vaccine development for antimicrobial-resistant Neisseria gonorrhoeae: current strategies and future directions". In: *Drugs* 81.10 (2021), pp. 1153–1169.

[3251] Julia Colón Pérez et al. "Addressing Sexually Transmitted Infections Due to Neisseria gonorrhoeae in the Present and Future". In: *Microorganisms* 12.5 (2024), p. 884.

[3252] Neil Gupta et al. "Carbapenem-resistant Enterobacteriaceae: epidemiology and prevention". In: *Clinical infectious diseases* 53.1 (2011), pp. 60–67.

[3253] Gian Carlo González-Carballo, Cristina García-Marín, et al. "Mecanismos de resistencia a la colistina, la nitrofurantoina y la fosfomicina en Enterobacterias". In: *Acta Médica Costarricense* 65.2 (2023), pp. 1–10.

[3254] Jorge Alberto Guevara Díaz et al. "Bacterial resistance: organisms of the ESKAPE group". In: *Enfermedades Infecciosas y Microbiología* 41.3 (2021), pp. 111–117.

[3255] Alexander YG Yip et al. "Antibiotics promote intestinal growth of carbapenem-resistant Enterobacteriaceae by enriching nutrients and depleting microbial metabolites". In: *Nature Communications* 14.1 (2023), p. 5094.

[3256] Maryn McKenna. "The last resort: health officials are watching in horror as bacteria become resistant to powerful carbapenem antibiotics–one of the last drugs on the shelf". In: *Nature* 499.7459 (2013), pp. 394–397.

[3257] Patrice Nordmann, Thierry Naas, and Laurent Poirel. "Global spread of carbapenemase-producing Enterobacteriaceae". In: *Emerging infectious diseases* 17.10 (2011), p. 1791.

[3258] Liang Chen et al. "Carbapenemase-producing Klebsiella pneumoniae: molecular and genetic decoding". In: *Trends in microbiology* 22.12 (2014), pp. 686–696.

[3259] Asad U Khan, Lubna Maryam, and Raffaele Zarrilli. "Structure, genetics and worldwide spread of New Delhi metallo-β-lactamase (NDM): a threat to public health". In: *BMC microbiology* 17 (2017), pp. 1–12.

[3260] Laurent Poirel, Anas Potron, and Patrice Nordmann. "OXA-48-like carbapenemases: the phantom menace". In: *Journal of Antimicrobial Chemotherapy* 67.7 (2012), pp. 1597–1606.

[3261] Philipp Kohler et al. "Dissemination of Verona Integron-encoded Metallo-β-lactamase among clinical and environmental Enterobacteriaceae isolates in Ontario, Canada". In: *Scientific Reports* 10.1 (2020), p. 18580.

[3262] Pisut Pongchaikul and Paninee Mongkolsuk. "Comprehensive analysis of imipenemase (IMP)-type metallo-β-lactamase: a global distribution threatening asia". In: *Antibiotics* 11.2 (2022), p. 236.

[3263] Min Hao et al. "Porin deficiency in carbapenem-resistant Enterobacter aerogenes strains". In: *Microbial Drug Resistance* 24.9 (2018), pp. 1277–1283.

[3264] Robert F Potter, Alaric W DSouza, and Gautam Dantas. "The rapid spread of carbapenem-resistant Enterobacteriaceae". In: *Drug Resistance Updates* 29 (2016), pp. 30–46.

[3265] Marta Hernández-García et al. "Deciphering mechanisms affecting cefepime-taniborbactam in vitro activity in carbapenemase-producing Enterobacterales and carbapenem-resistant Pseudomonas spp. isolates recovered during a surveillance study in Spain". In: *European Journal of Clinical Microbiology & Infectious Diseases* 43.2 (2024), pp. 279–296.

[3266] Alice Y Guh, L Clifford McDonald, and Ronda Sinkowitz-Cochran. "Assessment of Public Health Perspectives on Responding to an Emerging Pathogen: Carbapenem-Resistant: Enterobacteriaceae". In: *Journal of Public Health Management and Practice* 19.4 (2013), E27–E32.

[3267] N Deborah Friedman et al. "Carbapenem-resistant Enterobacteriaceae: a strategic roadmap for infection control". In: *infection control & hospital epidemiology* 38.5 (2017), pp. 580–594.

[3268] Haiyang Yu et al. "Multicenter Study of Carbapenemase-Producing Enterobacterales in Havana, Cuba, 2016–2021". In: *Antibiotics* 11.4 (2022), p. 514.

[3269] Chau-Chyun Sheu et al. "Infections caused by carbapenem-resistant Enterobacteriaceae: an update on therapeutic options". In: *Frontiers in microbiology* 10 (2019), p. 80.

[3270] Sandra S Richter and Dror Marchaim. "Screening for carbapenem-resistant Enterobacteriaceae: who, when, and how?" In: *Virulence* 8.4 (2017), pp. 417–426.

[3271] Nicole Bradley and Yuman Lee. "Practical implications of new antibiotic agents for the treatment of carbapenem-resistant Enterobacteriaceae". In: *Microbiology insights* 12 (2019), p. 1178636119840367.

[3272] Martin Everett et al. "Discovery of a novel metallo-β-lactamase inhibitor that potentiates meropenem activity against carbapenem-resistant Enterobacteriaceae". In: *Antimicrobial Agents and Chemotherapy* 62.5 (2018), pp. 10–1128.

[3273] Hyeonji Seo et al. "Clinical and microbiological analysis of risk factors for mortality in patients with carbapenem-resistant Enterobacteriaceae bacteremia". In: *International journal of antimicrobial agents* 56.4 (2020), p. 106126.

[3274] SM Bartsch et al. "Potential economic burden of carbapenem-resistant Enterobacteriaceae (CRE) in the United States". In: *Clinical Microbiology and Infection* 23.1 (2017), 48–e9.

[3275] Jesús Troya and José Bascuñana. "Safety and Tolerability: Current Challenges to Antiretroviral Therapy for the Long-Term Management of HIV Infection." In: *AIDS reviews* 18.3 (2016).

[3276] Anita Chawla et al. "A review of long-term toxicity of antiretroviral treatment regimens and implications for an aging population". In: *Infectious diseases and therapy* 7 (2018), pp. 183–195.

[3277] Harrison Rudd and Michal Toborek. "Pitfalls of antiretroviral therapy: current status and long-term CNS toxicity". In: *Biomolecules* 12.7 (2022), p. 894.

[3278] Huseyin Ekin Ergin et al. "HIV, antiretroviral therapy and metabolic alterations: a review". In: *Cureus* 12.5 (2020).

[3279] Claire Lagathu et al. "Metabolic complications affecting adipose tissue, lipid and glucose metabolism associated with HIV antiretroviral treatment". In: *Expert opinion on drug safety* 18.9 (2019), pp. 829–840.

[3280] Clay Bavinger et al. "Risk of cardiovascular disease from antiretroviral therapy for HIV: a systematic review". In: *PloS one* 8.3 (2013), e59551.

[3281] Kevin J Liner, Michelle J Ro, and Kevin R Robertson. "HIV, antiretroviral therapies, and the brain". In: *Current HIV/AIDS Reports* 7 (2010), pp. 85–91.

[3282] Lucette A Cysique and Bruce J Brew. "Neuropsychological functioning and antiretroviral treatment in HIV/AIDS: a review". In: *Neuropsychology review* 19 (2009), pp. 169–185.

[3283] Jeffrey M Jacobson and Kamel Khalili. "Toward the cure of HIV-1 infection: Lessons learned and yet to be learned as new strategies are developed". In: *AIDS reviews* 20.4 (2018), p. 220.

[3284] Jennifer M Zerbato and Sharon R Lewin. "A cure for HIV: how would we know?" In: *The Lancet HIV* 7.5 (2020), e304–e306.

[3285] Fanke Gong. "Case Study of Co-morbidities—HIV/AIDS and Blood Cancer". In: *SHS Web of Conferences*. Vol. 158. EDP Sciences. 2023, p. 02004.

[3286] Chayarani Kelgeri et al. "Clinical spectrum of children with acute hepatitis of unknown cause". In: *New England Journal of Medicine* 387.7 (2022), pp. 611–619.

[3287] Kai Gong et al. "Acute hepatitis of unknown origin in children: A combination of factors". In: *Frontiers in Pharmacology* 13 (2022), p. 1056385.

[3288] Jason G Smith et al. "Adenovirus". In: *Cell entry by non-enveloped viruses* (2010), pp. 195–224.

[3289] Malik Sallam et al. "Hepatitis of unknown origin and etiology (acute non HepA-E hepatitis) among children in 2021/2022: review of the current findings". In: *Healthcare*. Vol. 10. 6. MDPI. 2022, p. 973.

[3290] Muge Cevik et al. *Acute hepatitis of unknown origin in children*. 2022.

[3291] María Teresa Pérez-Gracia, Antonio Tarín-Pelló, and Beatriz Suay-García. "Severe acute hepatitis of unknown origin in children: what do we know today?" In: *Journal of Clinical and Translational Hepatology* 10.4 (2022), p. 711.

[3292] Barath Jagadisan and Anil Dhawan. "Outbreak of indeterminate acute hepatitis in children, not a new disease but an epidemiological phenomenon". In: *Hepatology* (2024), pp. 10–1097.

[3293] Olivier Uwishema et al. "A review on acute, severe hepatitis of unknown origin in children: A call for concern". In: *Annals of Medicine and Surgery* 81 (2022).

[3294] Jodie Haring and Stanley Perlman. "Mouse hepatitis virus". In: *Current opinion in microbiology* 4.4 (2001), pp. 462–466.

[3295] Graziele Pereira Oliveira and Erna Geessien Kroon. "Mouse hepatitis virus: A betacoronavirus model to study the virucidal activity of air disinfection equipment on surface contamination". In: *Journal of Virological Methods* 297 (2021), p. 114274.

[3296] Susan R Weiss and Julian L Leibowitz. "Coronavirus pathogenesis". In: *Advances in virus research* 81 (2011), pp. 85–164.

[3297] Ehud Lavi et al. "MHV-A59 pathogenesis in mice". In: *Molecular Biology and Pathogenesis of Coronaviruses* (1984), pp. 237–245.

[3298] AE Matthews, Susan R Weiss, and Yvonne Paterson. "Murine hepatitis virus–a model for virus-induced CNS demyelination". In: *Journal of neurovirology* 8.2 (2002), pp. 76–85.

[3299] Robert W Koerner et al. "Of mice and men: the coronavirus MHV and mouse models as a translational approach to understand SARS-CoV-2". In: *Viruses* 12.8 (2020), p. 880.

[3300] Kurt Bedell, Adam H Buchaklian, and Stanley Perlman. "Efficacy of an automated multiple emitter whole-room ultraviolet-C disinfection system against coronaviruses MHV and MERS-CoV". In: *infection control & hospital epidemiology* 37.5 (2016), pp. 598–599.

[3301] Susan R Compton et al. "Pathogenesis of enterotropic mouse hepatitis virus in immunocompetent and immunodeficient mice". In: *Comparative medicine* 54.6 (2004), pp. 681–689.

[3302] Susan R Compton et al. "Pathogenesis of mouse hepatitis virus infection in gamma interferon-deficient mice is modulated by co-infection with Helicobacter hepaticus". In: *Comparative medicine* 53.2 (2003), pp. 197–206.

[3303] Wallace A Jones and Richard B Cohen. "The effect of a murine hepatitis virus on the liver: An anatomic and histochemical study". In: *The American Journal of Pathology* 41.3 (1962), p. 329.

[3304] Nadine De Albuquerque et al. "Murine Hepatitis virus strain 1 produces a clinically relevant model of severe acute respiratory syndrome in A/J mice". In: *Journal of virology* 80.21 (2006), pp. 10382–10394.

[3305] Julian L Leibowitz et al. "Genetic determinants of mouse hepatitis virus strain 1 pneumovirulence". In: *Journal of virology* 84.18 (2010), pp. 9278–9291.

[3306] Felix R Homberger. "Enterotropic mouse hepatitis virus". In: *Laboratory animals* 31.2 (1997), pp. 97–115.

[3307] John C Hierholzer, J ROGER Broderson, and Frederik A Murphy. "New strain of mouse hepatitis virus as the cause of lethal enteritis in infant mice". In: *Infection and Immunity* 24.2 (1979), pp. 508–522.

[3308] Charles H Calisher and Wallace P Rowe. "Mouse Hepatitis Viruses". In: *Viruses of Laboratory Rodents* 20 (1966), p. 67.

[3309] SW Barthold and Abigail L Smith. "Mouse hepatitis virus strainrelated patterns of tissue tropism in suckling mice". In: *Archives of virology* 81 (1984), pp. 103–112.

[3310] Bruce D Zelus et al. "Purified, soluble recombinant mouse hepatitis virus receptor, Bgp1b, and Bgp2 murine coronavirus receptors differ in mouse hepatitis virus binding and neutralizing activities". In: *Journal of virology* 72.9 (1998), pp. 7237–7244.

[3311] Margaret G Kidwell. "Horizontal transfer". In: *Current opinion in genetics & development* 2.6 (1992), pp. 868–873.

[3312] Christine L Schneider. "Bacteriophage-mediated horizontal gene transfer: transduction". In: *Bacteriophages: biology, technology, therapy* (2021), pp. 151–192.

[3313] Nicholas AT Irwin et al. "Systematic evaluation of horizontal gene transfer between eukaryotes and viruses". In: *Nature microbiology* 7.2 (2022), pp. 327–336.

[3314] Ilana Lauren Brito. "Examining horizontal gene transfer in microbial communities". In: *Nature Reviews Microbiology* 19.7 (2021), pp. 442–453.

[3315] Holger Heuer and Kornelia Smalla. "Horizontal gene transfer between bacteria". In: *Environmental biosafety research* 6.1-2 (2007), pp. 3–13.

[3316] Huiquan Liu et al. "Widespread horizontal gene transfer from circular single-stranded DNA viruses to eukaryotic genomes". In: *BMC evolutionary biology* 11 (2011), pp. 1–15.

[3317] Huiquan Liu et al. "Widespread horizontal gene transfer from double-stranded RNA viruses to eukaryotic nuclear genomes". In: *Journal of virology* 84.22 (2010), pp. 11876–11887.

[3318] Shannon M Soucy, Jinling Huang, and Johann Peter Gogarten. "Horizontal gene transfer: building the web of life". In: *Nature Reviews Genetics* 16.8 (2015), pp. 472–482.

[3319] Valerian V Dolja and Eugene V Koonin. "Metagenomics reshapes the concepts of RNA virus evolution by revealing extensive horizontal virus transfer". In: *Virus research* 244 (2018), pp. 36–52.

[3320] Luis Boto. "Horizontal gene transfer in evolution: facts and challenges". In: *Proceedings of the Royal Society B: Biological Sciences* 277.1683 (2010), pp. 819–827.

[3321] Gerald Schönknecht, Andreas PM Weber, and Martin J Lercher. "Horizontal gene acquisitions by eukaryotes as drivers of adaptive evolution". In: *Bioessays* 36.1 (2014), pp. 9–20.

[3322] Mya Breitbart and Forest Rohwer. "Here a virus, there a virus, everywhere the same virus?" In: *Trends in microbiology* 13.6 (2005), pp. 278–284.

[3323] Jan F Finke et al. "Variation in the genetic repertoire of viruses infecting Micromonas pusilla reflects horizontal gene transfer and links to their environmental distribution". In: *Viruses* 9.5 (2017), p. 116.