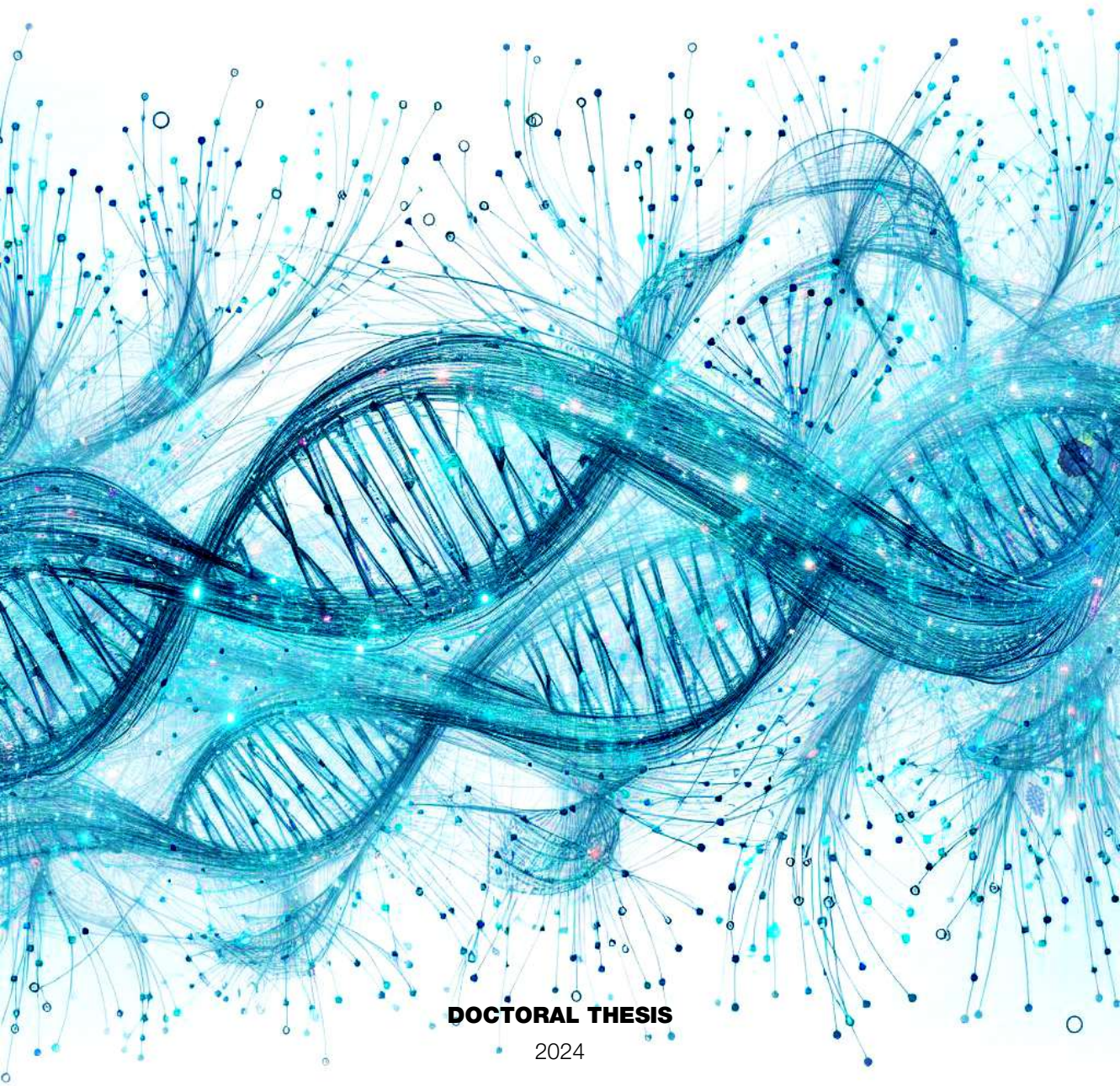
APPLICATION OF MACHINE LEARNING METHODS ON SARS-COV-2: MORTALITY PREDICTION BY USING HEALTH AND NUTRITIONAL FACTORS AND PREDICTION OF RECURRENT MUTATIONS

Bryan Percy Saldivar Espinoza

UNIVERSITAT ROVIRA i VIRGILI

Application of machine learning methods on SARS-Cov-2: Mortality prediction by using health and nutritional factors and prediction of recurrent mutations

Bryan Percy Saldivar Espinoza



DOCTORAL THESIS

2024

UNIVERSITAT ROVIRA I VIRGILI
APPLICATION OF MACHINE LEARNING METHODS ON SARS-COV-2: MORTALITY PREDICTION BY USING HEALTH AND NUTRITIONAL FACTORS AND PREDICTION OF RECURRENT MUTATIONS
Bryan Percy Saldivar Espinoza

Bryan Percy Saldivar Espinoza

Application of machine learning methods on SARS-CoV-2: Mortality prediction by using health and nutritional factors and prediction of recurrent mutations

Doctoral Thesis

Supervised by
Dr. Gerard Pujadas Anguiano, Dr. Santiago Garcia Vallvé and Dr. Adrià Cereto-Massagué

Cheminformatics & Nutrition Research Group
Biochemistry & Biotechnology Department

UNIVERSITAT ROVIRA i VIRGILI

Tarragona 2024

UNIVERSITAT ROVIRA I VIRGILI
APPLICATION OF MACHINE LEARNING METHODS ON SARS-COV-2: MORTALITY PREDICTION BY USING HEALTH AND NUTRITIONAL FACTORS AND PREDICTION OF RECURRENT MUTATIONS
Bryan Percy Saldivar Espinoza

UNIVERSITAT ROVIRA I VIRGILI

FAIG CONSTAR que aquest treball, titulat **"Application of machine learning methods on SARS-Cov-2: Mortality prediction by using health and nutritional factors and prediction of recurrent mutations"**, que presenta **Bryan Percy Saldivar Espinoza** per a l'obtenció del títol de Doctor, ha estat realitzat sota la meva direcció a Departament de Bioquímica i Biotecnologia d'aquesta universitat.

HAGO CONSTAR que el presente trabajo, titulado **"Application of machine learning methods on SARS-Cov-2: Mortality prediction by using health and nutritional factors and prediction of recurrent mutations"**, que presenta **Bryan Percy Saldivar Espinoza** para la obtención del título de Doctor, ha sido realizado bajo mi dirección en el Departamento de Bioquímica y Biotecnología de esta universidad.

I STATE that the present study, entitled **"Application of machine learning methods on SARS-Cov-2: Mortality prediction by using health and nutritional factors and prediction of recurrent mutations"**, presented by **Bryan Percy Saldivar Espinoza** for the award of the degree of Doctor, has been carried out under my supervision at the Department of Biochemistry and Biotechnology of this university.

Tarragona, 4 d'abril de 2024

El/s director/s de la tesi doctoral
El/los director/es de la tesis doctoral
Doctoral Thesis Supervisor/s

Santiago Garcia Valvé

Gerard Pujadas Anguiano

Adrià Cereto Massagué

UNIVERSITAT ROVIRA I VIRGILI
APPLICATION OF MACHINE LEARNING METHODS ON SARS-COV-2: MORTALITY PREDICTION BY USING HEALTH AND NUTRITIONAL FACTORS AND PREDICTION OF RECURRENT MUTATIONS
Bryan Percy Saldivar Espinoza

The research presented in this thesis was carried out with funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 713679 and from the Universitat Rovira i Virgili (URV).

UNIVERSITAT ROVIRA I VIRGILI
APPLICATION OF MACHINE LEARNING METHODS ON SARS-COV-2: MORTALITY PREDICTION BY USING HEALTH AND NUTRITIONAL FACTORS AND PREDICTION OF RECURRENT MUTATIONS
Bryan Percy Saldivar Espinoza

Acknowledgments

I appreciate the help, assistance, mentoring, tolerance, guidance and many other qualities and ways of support from Dr. Gerard Pujadas Anguiano, Dr. Santiago Garcia Vallvé, Dr. Adrià Cereto-Massagué and Dr Pere Puigbò. In addition, I would like to thank Dr Aleix Gimeno Vives for also helping me with the thesis guidance. In addition, I appreciate Dr Patrick Aloy for granting me some time off work for completing my thesis. I dedicate my work to my family: Rocio Espinoza Rojas, William Saldivar Mora, Ray Enrique Saldivar Espinoza and William Cesar Saldivar Espinoza, who were there for me at all times and allowed me follow my doctoral studies.

I extend my gratitude towards all my friends that were encouraging me to complete the thesis, to do not give up and gave me enthusiasm so I keep the pace and progress. Specially towards Arnau Comajuncosa, Dylan Dalton, Elena Pareja, Gema Rojas and Guillem Jorba (alphabetically to not disclose my favorite).

I would like to also thank Michael Bamiloshin for helping me with additional guidance and support with the thesis and to his great tolerance for keeping our friendship. I extend also my gratitude towards Reyda Akdemir, who helped me through tough times.

Last but not least, I am in debt with Marco Trombetti and Sébastien Bratières who were there for me through challenging times. Without their help I might not have completed this work.

UNIVERSITAT ROVIRA I VIRGILI
APPLICATION OF MACHINE LEARNING METHODS ON SARS-COV-2: MORTALITY PREDICTION BY USING HEALTH AND NUTRITIONAL FACTORS AND PREDICTION OF RECURRENT MUTATIONS
Bryan Percy Saldivar Espinoza

Table of Contents

Summary

In 2019 we witnessed the emergence of a new pandemic that has made society, healthcare systems and economy to tremble worldwide, unveiling how unprepared we were in terms of readiness, knowledge and protocols to minimize its negative impact. The pandemic was caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), a positive single stranded RNA betacoronavirus. During the pandemic, it was seen that the disease COVID-19 affected people differently, depending on health, socioeconomic, nutritional and other factors. In addition, the virus showed an elevated propensity to mutate, generating uncertainty about the efficacy of treatments to fight it. In this regard, this thesis uses Machine Learning models to analyze the main factors affecting COVID-19 mortality and predict SARS-CoV-2 recurrent mutations. The predictive model developed for COVID-19 mortality at the US county level integrates health, socioeconomic, and nutritional data, achieving a notable correlation of 0.715. The analysis of influential variables revealed that the proportion of primary care physicians and other health providers relative to the population, along with socioeconomic indicators such as median household income and rates of physical inactivity and poverty, significantly impact COVID-19 mortality rates. Surprisingly, metabolism or nutrition-related variables show little importance in predictive power, while hypertension-related deaths in specific age and gender groups emerge as significant predictors. In addition to predicting COVID-19 mortality, this thesis explores SARS-CoV-2 mutations. Artificial Neural Network (ANN) models were developed to predict SARS-CoV-2 recurrent mutations. These models effectively predicted positions with recurrent mutations and their recurrence levels. Notably, model robustness was demonstrated by false positive mutations becoming true positives in an updated test set. Moreover, the exploration of the most influencial variables revealed that nucleotides in the central position within each evaluated window, as well as surrounding nucleotides and RNA reactivity data, play significant roles. These findings agrees with the prevalence in SARS-CoV-2 of C>U mutations originated by host deaminases.

Resumen

En 2019 fuimos testigos del surgimiento de una nueva pandemia que ha hecho temblar a la sociedad, a los sistemas de salud y a la economía en todo el mundo, revelando cuán poco competentes estábamos en términos de preparación, conocimiento y protocolos para minimizar su impacto negativo. La pandemia fue causada por el coronavirus 2 del síndrome respiratorio agudo severo (SARS-CoV-2), un betacoronavirus de ARN monocatenario positivo. Durante la pandemia, se vio que la enfermedad COVID-19 afectaba a las personas de manera diferente, dependiendo de factores de salud, socioeconómicos, nutricionales y otros. Además, el virus mostró una elevada propensión a mutar, generando incertidumbre sobre la eficacia de los tratamientos para combatirlo. En este sentido, esta tesis utiliza modelos de Machine Learning para analizar los principales factores que afectan la mortalidad por COVID-19 y predecir mutaciones recurrentes del SARS-CoV-2. El modelo predictivo desarrollado para la mortalidad por COVID-19 a nivel de condados de EE. UU. integra datos de salud, socioeconómicos y nutricionales, logrando una notable correlación de 0,715. El análisis de variables más influyentes reveló que la proporción de médicos de atención primaria y otros proveedores de salud en relación con la población, junto con indicadores socioeconómicos como el ingreso familiar medio y las tasas de inactividad física y pobreza, impactan significativamente en las tasas de mortalidad por COVID-19. Sorprendentemente, las variables relacionadas con el metabolismo o la nutrición muestran poca importancia en el poder predictivo, mientras que las muertes relacionadas con la hipertensión en grupos específicos de edad y género emergen como predictores significativos. Además de predecir la mortalidad por COVID-19, esta tesis explora las mutaciones del SARS-CoV-2. Se desarrollaron modelos de redes neuronales artificiales (ANN) para predecir mutaciones recurrentes del SARS-CoV-2. Estos modelos predijeron eficazmente posiciones del genoma del virus con mutaciones recurrentes y sus niveles de recurrencia. En particular, la fortaleza del modelo se demostró mediante mutaciones falsas positivas que se convirtieron en verdaderas positivas en un conjunto de datos actualizado. Además, la exploración de las variables más influyentes reveló que los nucleótidos en la posición central dentro de cada ventana evaluada, así como los nucleótidos circundantes y los datos de reactividad del ARN, desempeñan papeles importantes. Estos hallazgos concuerdan con la prevalencia en el SARS-CoV-2 de mutaciones C>U originadas por las desaminasas del huésped.

Resum

El 2019 vam assistir a l'aparició d'una nova pandèmia que ha fet tremolar la societat, els sistemes sanitaris i l'economia a tot el món, revelant com de poc preparats estàvem en termes de coneixements i protocols per minimitzar el seu impacte negatiu. La pandèmia va ser causada pel coronavirus 2 de la síndrome respiratòria aguda severa (SARS-CoV-2), un betacoronavirus d'ARN monocatenari positiu. Durant la pandèmia, es va veure que la malaltia COVID-19 afectava a les persones de manera diferent, segons factors de salut, socioeconòmics, nutricionals i altres. A més, el virus va mostrar una elevada propensió a mutar, generant incertesa sobre l'eficàcia dels tractaments per combatre'l. En aquest sentit, aquesta tesi utilitza models d'aprenentatge automàtic per analitzar els principals factors que afecten la mortalitat per COVID-19 i predir mutacions recurrents del SARS-CoV-2. El model predictiu desenvolupat per a la mortalitat per COVID-19 a nivell de comtats dels EUA integra dades de salut, socioeconòmiques i nutricionals, aconseguint una correlació notable de 0,715. L'anàlisi de variables influents va revelar que la proporció de metges d'atenció primària i altres proveïdors de salut en relació a la població, juntament amb indicadors socioeconòmics com la renda mitjana de les llars i les taxes d'inactivitat física i pobresa, afecten significativament les taxes de mortalitat per COVID-19. Sorprenentment, les variables relacionades amb el metabolisme o la nutrició mostren poca importància en el poder predictiu, mentre que les morts relacionades amb la hipertensió en grups d'edat i gènere específics emergeixen com a predictors significatius. A més de predir la mortalitat per COVID-19, aquesta tesi explora les mutacions del SARS-CoV-2. Es van desenvolupar models de xarxa neuronal artificial (ANN) per predir mutacions recurrents del SARS-CoV-2. Aquests models van predir eficaçment les posicions amb mutacions recurrents i els seus nivells de recurrència. En particular, la robustesa del model es va demostrar quan alguns falsos positius es van convertir en veritables positius en un conjunt de proves més actualitzat. A més, l'exploració de les variables més influents va revelar que els nucleòtids en la posició central dins de cada finestra avaluada, així com els nucleòtids circumdants i les dades de reactivitat de l'ARN, tenen un paper important. Aquestes troballes coincideixen amb la prevalença en SARS-CoV-2 de mutacions C>U originades per desaminases de l'hoste.

Introduction

1. Introduction to The SARS-CoV-2 and COVID-19 Pandemic:

1.1 Background and Global Impact

The emergence of SARS-CoV-2, the virus responsible for the COVID-19 pandemic [1], disrupted global public health due to the incomplete understanding of its pathogenicity [2]. Identified in late 2019 in Wuhan, China [3], the virus rapidly crossed borders, reaching 43 thousand confirmed cases in 28 countries/regions by February 2020 [4]. The scale and severity of the pandemic have posed immense challenges to healthcare systems, economies, and societies worldwide [5,6], reaching by October 2023, 6.8 million fatalities and 676 million cases worldwide [7]. The virus, belonging to the coronavirus family, exhibited high transmissibility and the ability to cause severe respiratory illness [8,9]. It affected with higher intensity to vulnerable populations: people older than 65 and with immunocompromising conditions [10] and people with cancer [11,12]. It has also highly affected to those with nutritional related diseases, such as obesity [13,14], diabetes [15], hypertension [16] and other comorbidities [17]. The non-stop spread of SARS-CoV-2 underline the need for heterogeneous approaches, including predictive modeling [18] and mutational analysis [19], to understand, mitigate, and manage the evolving challenges created by SARS-CoV-2.

1.2. SARS-CoV-2

SARS-CoV-2 is a beta-coronavirus that belongs to the Coronaviridae family [3]. This virus family infects avian and mammal species [3]. Among those that infect are HcoV-229E, HcoV-OC43, HCoV-NL63 and HcoV-HKU1, which produce “common cold”-like symptoms through seasonal and mild respiratory infections [20]. However, other coronaviruses such as Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV), Middle East Respiratory Syndrome Coronavirus (MERS-CoV) and SARS-CoV-2 are more aggressive and produce life-threatening, severe lung injuries and respiratory pathologies [20]. SARS-CoV, in 2002, resulted in 8 thousand cases with a 10% mortality rate. Meanwhile, MERS-CoV, in 2012, generated 2.5 thousand cases and a 36% mortality rate [20].

SARS-CoV-2 genome is a positive single-stranded RNA (+ssRNA) [21]. This means that it can be transcribed directly into proteins by human cells’ ribosomes. SARS-CoV-2 genome is about 29.9 kilobases long [3], making it one of the largest RNA viruses [22]. This genome encodes for 28 proteins [23], of which 4 are structural, 16 are nonstructural and 8 are accessory proteins [24] (the proteins and genes encoding for them are shown in **Table 1**). At the 5’ of the genome are present the Open Reading Frames (ORFs) 1a and 1b, occupying the first two thirds of the genome. They together encode the 16 nonstructural proteins (nsp), from which 15 (nsp2 to nsp16) form the replication and transcription complex (RTC), which is in charge of RNA processing, modification and proof-reading [20]. Meanwhile, structural and accessory proteins are located in the last third part of the genome. After ORF1a and ORF1b, the S gene is found, followed by the ORF3a, ORF3b, E, M, ORF6, ORF7a, ORF7b, ORF8, N (with ORF9b inside of it) and ORF10 genes [25]. The accessory proteins (ORF 3a, 3b, 6, 7a, 7b, 8, 9b and 10) have high variability, limited conservation and are thought to contribute to module the host

response to infection [26–29]. Even though they were not seen required for replication in cell cultures, ORF8 has been seen in cell cultures to bind to the Major Histocompatibility Complex (MHC) mediating its degradation [20]. Thus, suggesting a mechanism of immune evasion. In addition ORF3b, which was suggested as an effective interferon antagonist [20].

Table 1. SARS-CoV-2 Genes, proteins and functions summary

Gene	Protein	Start Position [25]	End Position [25]	Function
ORF1a	NSP1	266	805	Reduces Interferon I expression and binds to ribosome so only viral mRNA is translated [20,30,31].
	NSP2	806	2719	Disrupts the cell environment by binding to prohibitin 1 and 2 [26,31].
	NSP3 (PLpro)	2720	8554	Cleaves polyprotein 1a and 1b and component of the Replication Transcription Complex (RTC) [20,26,31].
	NSP4	8555	10054	Component of the RTC with NSP3 and NSP6 [20,26,31].
	NSP5 (3CLpro/Mpro)	10055	10972	Cleaves polyprotein 1a and 1b [20,26,31].
	NSP6	10973	11842	Component of the RTC with NSP3 and NSP4 [20,26,31].
	NSP7	11843	12091	Cofactor of NSP12 (RdRp) with NSP8 [20,26,31].
	NSP8	12092	12685	Cofactor of NSP12 (RdRp) with NSP7 [20,26,31] and adenylyltransferase RNA 3′-terminal activity [31].
	NSP9	12686	13024	Binds viral RNA for replication [26,31–33].

	NSP10	13025	13441	Cofactor with NSP14 and NSP16 [26,31] and is part of the capping machinery [20].
	NSP11	13442	13480	Unknown [26,31].
ORF1b	NSP12 (RdRp)	13442	16236	Core component of the RTC for RNA replication [20,26,31].
	NSP13 (Helicase)	16237	18039	Unwinds duplex RNA [26,31] and has RNA 5′-triphosphatase activity [20,26].
	NSP14	18040	19620	Proofreading, 5′ exoribonuclease and a guanosine-N7 methyltransferase for RNA cap formation [20,26,31].
	NSP15 (endoRNAse)	19621	20658	Endoribonuclease [20,26,31], degrades viral RNA to avoid immune system recognition [26].
	NSP16	20659	21552	2'-O-methyltransferase [20], methylates viral RNA for cap formation [26,31] and promoting translation [26].
S	Spike glycoprotein	21563	25384	Binds to the ACE2 cell receptor for viral entry [20,26,31].
ORF3a	Accessory protein	25393	26220	Activates the NLRP3 inflammasome [26,27].
E	Envelope protein	26245	26472	Viral assembly [20,26], virion release [26] and promotes pathogenecity [26,34].
M	Membrane protein	26523	27191	Viral assembly [20,26,31] and induce apoptosis [26].
ORF6	Accessory protein	27202	27387	Interacts with NSP8 contributing to the RNA polymerase [26] and antagonizes the interferon signalling pathway [20].
ORF7a	Non-structural protein 7a	27394	27759	Downregulates Major Histocompatibility Complex class I (MHC-I) expression [28]. Can deregulate interferon class 1 (IFN-I) [29].
ORF7b	Non-structural protein 7b	27756	27887	Probable inhibition of the activation of the Signal transducer and activator of transcription (STAT2) through phosphorylation [29].
ORF8	Non-structural protein 8	27894	28259	Downregulates MHC-I [35].
N	Nucleocapsid protein	28274	29533	Provides stability to the viral RNA through binding with it and packaging [20,26,31]. Also obstructs interferon signalling pathway [20].
ORF10	Accessory protein	29558	29674	Downregulates IFN-I [29,36].

SARS-CoV-2 infects bronchial epithelial cells and pneumocytes mainly in the upper respiratory tract, where the Angiotensin-Converting Enzyme 2 (ACE2) receptor, on the surface of human cells, is abundantly expressed [20]. For viral entry into the cell, the spike (S) protein, which is expressed on the surface of SARS-CoV-2, binds with ACE2. Also expressed on the surface of the virus are the envelope (E) and membrane (M) proteins which conform the virion envelope [24]. These three are SARS-CoV-2 structural proteins, in addition to the nucleocapsid (N) protein, which resides inside the virus, keeping the genetic material together [24]. The S protein, a homotrimer, is glycosylated on its surface, granting it evasion from the host immune system by protecting epitopes from neutralizing antibodies [20]. In addition, the S protein has two parts, S1 and S2. S1 is exposed on the surface of the S protein, where the Receptor Binding Domain (RBD) is present and makes contact with ACE2 [24]. On the other hand, S2 is the part closer to the virus envelope, it has the trans-membrane domain, which mediates fusion with the human cell membrane [20]. S1 and S2 are split apart at the boundary between them by the human Transmembrane Protease Serine 2 (TMPRSS2), which is present on the surface of the cell [20]. Nevertheless, the ease of separation at this boundary is because of a previously cleavage done by host cell-derived proteases, like the prototype proprotein convertase furin [24], which is located inside the cell, mainly in the Golgi apparatus. Once the S protein is freed from S1, S2 joins the host cell membrane and SARS-CoV-2's for fusion and entrance of the N protein, carrying the viral genome.

Once the viral genome is in the cell, it is translated into proteins using the cell ribosomes [20]. The first gene to be translated is the ORF1a, which encodes for the polyprotein 1a (pp1a) [20]. Pp1a contains a chain of 11 proteins, from nsp1 to nsp11, that will be later separated. At the end of the translation process of ORF1a, the ribosome makes a -1 programmed frameshit (A change in the reading frame codon returning one position) and starts translating the pp1b, which contains another chain of proteins, from nsp12 to nsp16 [20]. Once translated the pp1a, nsp3 (papain–like protease, PLpro) cleaves and releases nsp1 [26]. Nsp1 immediately leads towards ribosomes and binds, so they only translate viral RNA instead of the host's [26]. This action also impacts the reduction of the expression of interferons type I and III and of other proteins in the innate immune system [20]. Later, PLpro cuts himself and other nsps including another protease, nsp5 (chymotrypsin-like protease, 3C-like protease, 3CLpro, Mpro). Once Mpro is released from the poly-protein chain, it takes the lead in cleaving the majority of other nsps leveraged by a broader substrate specificity than PLpro [26]. Overall, nsps 2 to 11 accommodate the RTC, modulate intracellular membranes, provide evasion against the immune system and participate in providing cofactors for replication [20]. Meanwhile, nsps 12 to 16, contain the main enzymatic functions required for RNA synthesis, proofreading and modification [20]. More specifically, RNA synthesis is performed by nsp12 (RNA-dependent RNA polymerase, RdRp) and two co-factors, nsp7 and nsp8. In addition, nsp14 (ExoN) provides proof-reading for genome stability and cuts out erroneous mutagenic nucleotides [20,26]. Other examples are nsp10, nsp13, nsp14 and nsp16, which provide assistance in the capping process, catalize the removal of the 5' phosphate group, methylate the N7 position of the guanine that is added at the 5' end during capping and methylate the 2'-OH group of the ribose sugar in the first nucleotide of the virus RNA, respectively [20,26].

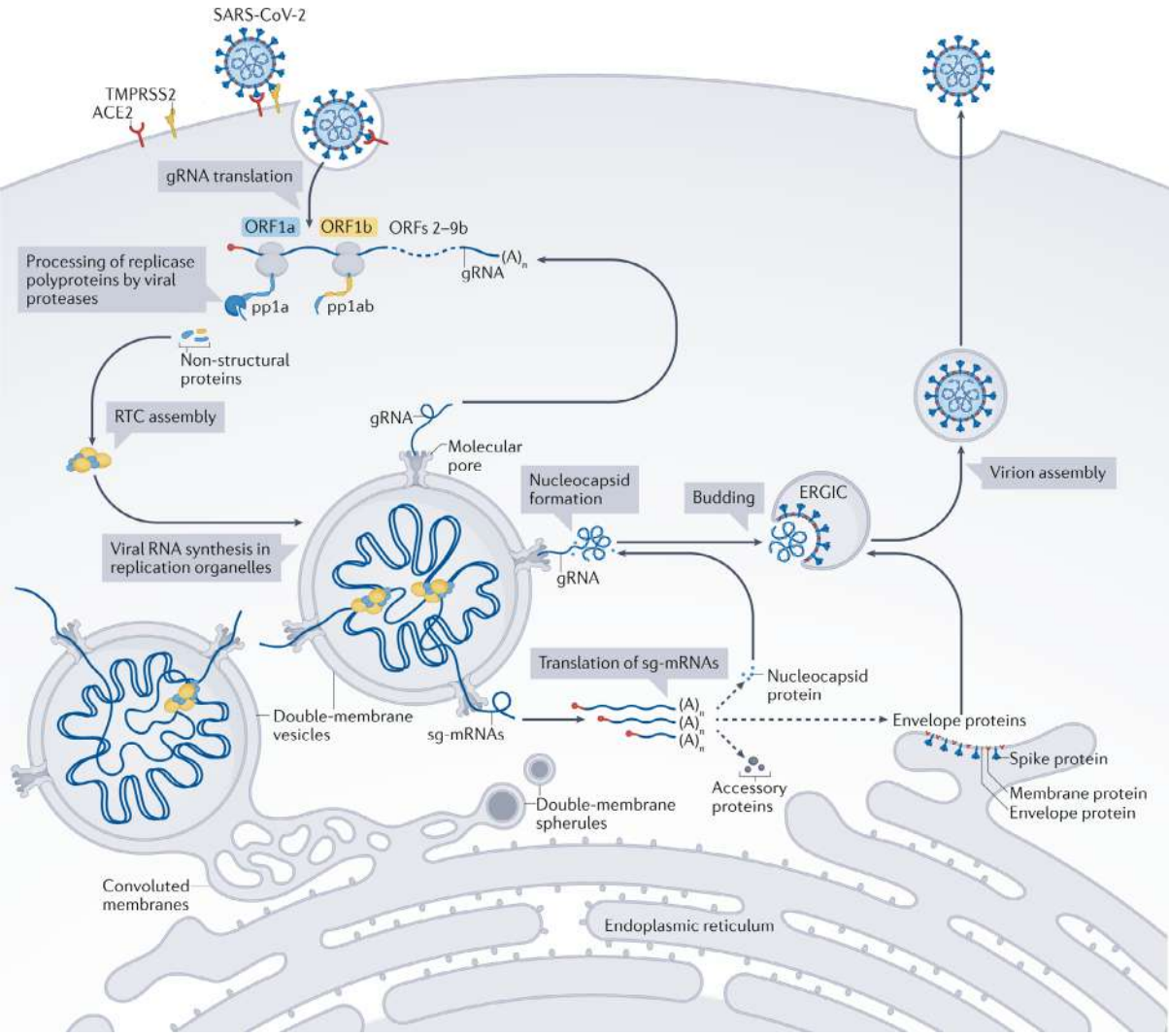


Figure 1. *SARS-CoV-2 Entry, translation, replication, assembly and exit. Source: (Malone et al. 2022) [37].*

For replication, special compartments are created during the first replication cycles of the virus (Figure 1). This as consequence of interactions between cell factors and the nsps 3, 4, 12, 13 and 16 [20,26]. These compartments derive from the Endoplasmic Reticulum (ER) and were hypothesized to be double-membrane vesicles (DMVs) with the possibility of having nsp3 and nsp4 as pores for the passage of viral components [20]. These compartments provide a favorable environment with the proper concentrations of macromolecules required for RNA synthesis. They also protect virus replication intermediate products from exposure to cytosolic innate immune sensors [20]. Within these compartments, RNA synthesis and replication start with the synthesis of full-length negative sense RNAs, which behave as templates for the creation of complimentary positive-sense RNA [20]. The positive-sense RNAs are used later for translation to obtain more nsps or for getting packed in new virions. During the synthesis of the negative strand RNA, the RTC interrupts the transcription when encountering Transcription Regulatory Sequences (TRSs) located at different parts of the RNA [20]. At any of these TRS, the synthesis of the negative strand stops and is re-started at the TRS close to the leader sequence (TRS-L). At the re-initiation of RNA synthesis in the TRS-L region, a negative strand copy of the leader sequence is added to the nascent RNA to complete the generation of the negative strand sub genomic RNA (sgRNA). This discontinuous generation of sgRNAs produces the characteristic nested set of positive sense sgmRNAs that are translated into structural and accessory proteins [20].

A change in the structure of the resulted proteins, as an effect of mutations [38], impact us negatively because they could provide the virus higher infectivity [39] and decrease the efficacy of developed medicine and vaccines [40] against SARS-CoV-2. Until June 2022, there were more than 10 million SARS-CoV-2 genomes collected in GISAID [41] accounting for up to 108 thousand mutations, from which 84 thousand were substitutions, 22 thousand were deletions and 2 thousand were insertions [42]. In addition, mutations can also occur at the same genome gene and position in different hosts independently, therefore becoming recurrent mutations [43]. Moreover, when mutations make a virus vary from the others, it becomes a variant [44]. These variants are classified as Variants of Concern (VoC) if they show increased rates of transmissibility, virulence, change in disease symptoms and a decrease in the effectiveness of treatments, medicine, vaccines, and measurements against it [45].

2. Machine Learning:

2.1 Algorithms

Machine learning (ML) is a subset of computational algorithms aimed at approximating human intelligence [46]. The key to their application lies in the improvement of these algorithms through experience/data [47]. ML algorithms can learn through tabular data [48], images/videos [49], text [50], interaction with an environment [51], among others [52,53]. The terminology of Artificial Intelligence (AI) is usually employed to address ML and the other way around [54]. Nevertheless, ML belongs in the broader category

of AI [54]. These ML algorithms can be classified into supervised, unsupervised and reinforcement learning [55]. The key difference between supervised and unsupervised methods is the presence of a label or a target in the data that is to be used [55]. An example of supervised learning is predicting if a patient will need an ICU [56]. This is possible if there is clinical data of patients, including variables (age, comorbidities, etc.) and a record whether they required an ICU (the label). On the other hand, unsupervised learning can be used for clustering data, such as computerized tomography scans [57], and to extract important features for later supervised applications [58]. This last method does not require a label/category for the clustering, just the data itself. If the goal is to predict a particular outcome, such as a mutation or number of fatalities, it is required to focus on supervised learning. Among the supervised algorithms, we can find Linear regression [59], K-nearest Neighbor [60], Naive Bayes [61], Support vector machines [62], Decision Trees [63], Random Forest [64], XGBoost [65], Artificial Neural Networks [66] and a extension of the last one, deep learning [67]. All these algorithms have different possible configurations, which are called hyperparameters, specific values for them can transform the performance power for each. **Table 2** shows some hyperparameters for these models.

Table 2. Some Machine learning models and possible hyperparameters

Model Type	Task	Some hyperparameters
Linear Regression	Regression	Learning rate, regularization penalty, number of iterations
Logistic Regression	Classification	Learning rate, regularization penalty, solver used for optimization
K-Nearest Neighbors (KNN)	Classification	Number of neighbors (k), distance metric used
Naive Bayes	Classification	
Support Vector Machine (SVM)	Classification, regression	Kernel type, regularization parameter, cost parameter
Decision Trees	Classification, regression	Maximum depth, splitting criterion, minimum number of samples per leaf
Random Forest	Classification, regression	Number of trees, maximum depth per tree, minimum number of samples per leaf
XGBoost	Classification, regression	Learning rate, number of trees, maximum depth per tree, regularization parameters
Multilayer Perceptron (MLP)	Classification, regression	Number of hidden layers, number of neurons per layer, activation function, learning rate, optimizer, dropout, skip connections
Convolutional Neural Network (CNN)	Image classification, object detection	Kernel size, stride, padding, number of filters per layer, learning rate, optimizer, dropout, skip connections
Recurrent Neural Network (RNN)	Sequence prediction, text generation	Number of hidden units, number of layers, activation function, learning rate, optimizer

Some of these algorithms combine multiple models to create a more robust and stronger model, they are called ensemble models [68]. When heterogeneous models are combined and another model is trained with the predictions of these previous models as input, it is called stacking [68]. On the other hand, when a model is trained weakly, a similar instance of the model is trained, giving more importance to those samples with higher errors in the prediction in a consecutive manner; this is called boosting [68]. An example of this last type of XGBoost [65].

Inspired by how the brain works, the perceptron was introduced, aiming to imitate how neurons communicate for learning. The perceptron [66] placed the base for current Artificial Neural Networks (ANN) since they are the collection, arrangement, and improvement of them. The way these systems learn is by back propagation [69]. Data comes into the network and produces an output. This output is compared to the target (the value we want the network to predict), and the difference is computed as an error. Then the gradient of this error is propagated backwards towards the input of the network, and each part of the ANN (parameters) is adjusted proportionally to correct for their error contribution. There are many ways to optimize the learning process. Among those are the optimization algorithms Stochastic Gradient Descent [70], Adam [71], Ada Delta [72], etc.

The explosive incorporation of ANNs into current applications is owed to the fact that ANNs work with matrix multiplication, which can be done extremely quickly in GPUs, in comparison to CPUs. This is because GPUs, or Graphic Processing Units, were designed for parallel processing [73]. Many types of ANNs were created afterwards, including the present Large Language Models (LLMs), GPT-like systems, Automatic Image Segmentation tools, Diffusion models for image generation. A vast amount of them have incorporated a particular kind of module called transformers [74]. Transformers were introduced to improve learning sequences by varying the attention that each part of the sequence receive during processing. The attention block has three main components: the Key, Query and Value. The three of them are computed as shown in formula **(1)**. When those three items come from the same source it is considered self-attention [75]. On the other hand, if the Query comes from a different source than the other two, it is called cross-attention [75].

$$Attention(Q, K, V) = softmax\left(\frac{Q K^T}{\sqrt{D_k}}\right) V$$

$$D_k = Dimensions\ of\ K;\ K^T = K\ transposed \quad \textbf{(1)}$$

Given the myriad of ML algorithms and the abundance of possible configurations for each one, their applicability becomes an exhausting task. For this matter, another branch of AI/ML, Automatic Machine Learning (AutoML) [76,77], focuses on finding the right algorithm or combinations of them, reducing human workload. Among the frameworks that aim in this direction, we can find MLJAR [78], TabPFN [79], TPOT [80] and AutoGluon [81].

Some ML algorithms are interpretable models (white box models) [82]. They are designed with the objective of providing information of how the decisions for the prediction were made and what the contribution of each variable is [82]. Nevertheless, many of the listed ML algorithms and AutoML frameworks behave like blackbox models [82]. They provide better prediction performance at the expense of a lack of understanding of the importance and behavior of the used variables. Despite it, some of them are explainable through the application of other methods [82] such as LIME [83], Saliency maps [84], SHAP values [85] and sensitivity analysis [86].

2.2 Data management

Since ML methods get their power from data, how data is handled plays a crucial role in obtaining not only good results but also fair, comparable, reproducible, and valid. Before feeding data into a ML algorithm, it is required to split the available data into two sets, one to train the ML system and another set of data that was not used during the training phase for testing the model's ability to generalize to unseen data. It is also often a common practice that this split is not done beforehand, but rather, there are two sources of data. One of them was used for training and the other, from another source, separated by its origin, for testing [87].

Many ML competitions and benchmarks add another layer of evaluation. They provide a dataset for training with labels, another set of data without labels, the testing set, but, for evaluating the real predictive power of the models, they use a subset of this testing set as a public result board that is shown during the validity of the competition. However, another subset of this testing set is used to qualify after all possible improvements are done. This is done to avoid tailoring the generalization ability of the system to the feedback of the scoring board. By doing this, this publicly visible testing set works as a validation set.

The validation set is another data split used to evaluate progressively the generalization capacity of the model being trained. Therefore, getting three splits: training, validation, and testing sets. In addition to this setting, it is also common to use another data splitting technique called n-fold cross validation [87]. The objective is to split the data n-times, each time a different part of the data behaves as the validation set, without replacing. By validating the predictions of the model with different splits of data for training and validation, the probability of the ML model and its configuration to generalize to unseen data increases. A ML model can have multiple configurations and each performing differently.

In the case of an ANN, it can have different number of layers, number of neurons per layer, activation functions, dropout layer probability, batch normalization layers, skip connections, learning rate, data batch size, number of epochs, etc. (**Table 2**). This group of variables used in the design of the training strategy are called hyperparameters. The n-fold cross validation technique, or just using a separate validation set is used to find the right set of hyperparameters for the ML algorithm being used. When there is a high discrepancy in the performance of the ML model between the training-validation and testing sets, and the performance with the training-validation set is higher, it is considered overfitting (**Figure 2 C**) [88]. On the other hand, when the model performance behaves

poorly for both sets, it is called underfitting (**Figure 2 A**) [88]. Therefore, the importance of the validation set to reduce these possible unwanted scenarios.

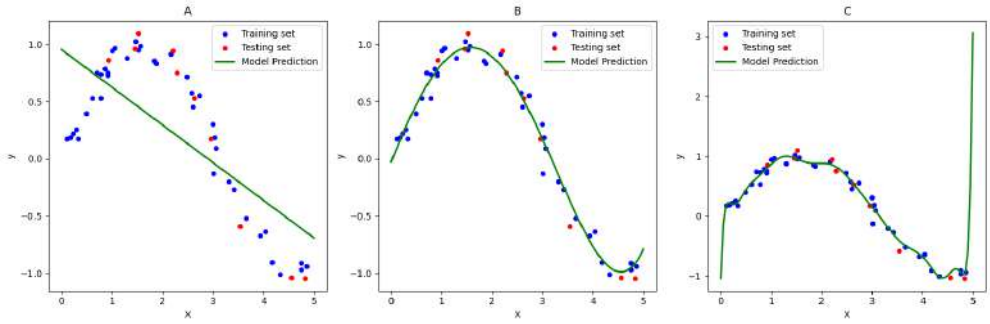


Figure 2. *An example for underfitting, a good fit and overfitting are shown in panels* ***A***, ***B*** *and* ***C*** *respectively.*

Data size plays an important role in the validation strategy. When the amount of data exceeds the computation capacity for a ML training task, it is not possible to perform a n-fold cross validation with high values for n, but rather keep the value of n as 1 and become the first case of having one split for training and one for validation. It is very common to find ML applications where the data is unbalanced. This means, there is more abundance of one type, kind, and range of values for the data, while another is scarce. As an example, in the medical field, if we would like to differentiate healthy patients from those with a disease, the collection of data for people without carrying the disease will provide more samples, and the number of cases with the disease will be lower [89]. When this happens the data split strategy should change by doing an stratified sampling for the definition of the sets and if an n-fold, n>1, method is used, the value of n is higher than normal. A default value of n is 5 and with unbalanced data, the value goes higher, to 7 or 10.

In addition to controlling the data split for unbalanced scenarios, it is also required to control data leakage from one split to another [90]. It is possible that samples that are very similar or identical can be present in the training set and the validation or testing sets. If these similar samples are in the training and validation sets but not in the testing set, the performance evaluation would provide a false sense of generalization, since the evaluation is using data used for training the model. Following this scenario, when the model performance is to be measured on the testing set, the measurement of the capacity will drop, since it was overestimated. On the other hand, if the data leakage goes to the testing set, all reported metrics of performance would show an inaccurate reflection of reality.

Many ML applications also face a general data scarcity because of the cost of acquiring more samples or their nature of being sparse, infrequent, or in a shortage. To deal with this scenario, there are data augmentation techniques that increase the amount of available data for training in a synthetic way [91]. An example in the computer vision realm (ML applied to images) would be to rotate the image, flip it, reflect it horizontally/

vertically, add noise to it, increase or decrease brightness or contrast, crop subsections of the image and use combinations of all of them. Nevertheless, these techniques demand care in how and when they are used, a modification of a medical image for data augmentation could damage what the ML model is learning. For instance, for an X-ray image, it would make no sense if the image were reflected vertically and if done horizontally, it could provide a bad judgment by moving the heart to the other side. Moreover, data augmentation should be done after data splitting, since if it is done before, it could lead to data leakage by having pairs representing the same sample in different data splits.

2.3 Data cleaning and processing

Data rarely comes in the right format to be used directly with a ML model or framework. There could be missing values or outliers; values could be numerical, categorical, or textual. ML models can only be fed with numerical values, thus, any other types of value should be handled and transformed before their utilization. The first step would be to separate variables by type, numerical, categorical, text, or others. Categorical variables are those that do not reflect a magnitude, like sex (male or female). The most common transformation of categorical variables is one-hot encoding [92]. This method transforms categories into binary vectors that have as many columns as categories are present in the data but without redundancy. If we were to categorize if a drug binds to a specific protein, we would mark it with 1 if there is binding or 0 otherwise. In the same way, if we were to move the variable sex to a vector, it could be 1 for females and 0 for males. If our dataset is tabular and one of the columns is to which set of proteins the samples bind, we would create a vector of as many columns as the total amount of registered proteins present in the table. If a sample binds to 3 proteins, from 10 possible proteins, the vector will have 7 zeros and 3 ones. Text values, on the other hand, might require a more advanced transformation. In Natural Language Processing (NLP), sentences are split into tokens, groups of characters, or chunks of the sentence and assigned an integer number. This number will correspond to an index in a dictionary of vectors. These vectors are called embeddings and were optimized to better represent the before-mentioned tokens for later being placed together [93]. Numerical values, on the other hand, already come in a format ready to be used, nonetheless, there could be missing values that affect the direct employment of this set of variables. Missing values can be handled by being replaced with zeros, the mean value for that variable, the most frequent value, or a more sophisticated imputation technique [94,95]. Missing values can also be replaced by predictions of uni-variate and multi-variate models that use the existing values and predict what values should be present in those missing positions. Despite these techniques, depending on the data size and importance of the variables, samples with many missing values can be discarded, or also variables that possess a high abundance of missing values.

Variables can be present in a suboptimal representation; for instance, we could have weight and volume to describe a compound, but a combination of both (density) could not only represent the samples better but also decrease the number of used variables, decreasing computation and the possibility of overfitting. Other variables for which we do not know their relationships to physical properties, might also benefit from such change of representation. Feature engineering [96] deals with these combinations, the

AutoML frameworks described above , incorporate such steps to create new features from combinations of the existing ones.

In addition to these transformations, it is still required to analyze the values of each variable regarding how they are distributed. Some values could be outside normal ranges, outliers, modifying the general perception of how the variable behaves from an analytical point of view and for the processing of some ML algorithms. Nevertheless, when the amount of data is large enough some ML methods are not affected by these outliers. Therefore, it is only important to identify them for interpretation at the end of the training stage. The most common method for outlier detection is Interquartile Range (IQR) [97]. The IQR method considers outliers to those values outside the range between Q1-IQR and Q3+IQR [where Q1 and Q3 are the values of the first and third quartiles, while IQR corresponds to 1.5x(Q3-Q1)] (**Figure 3**). With this method, the identified outliers are removed for training if the ML method could be sensitive to them or after training to evaluate them separately.

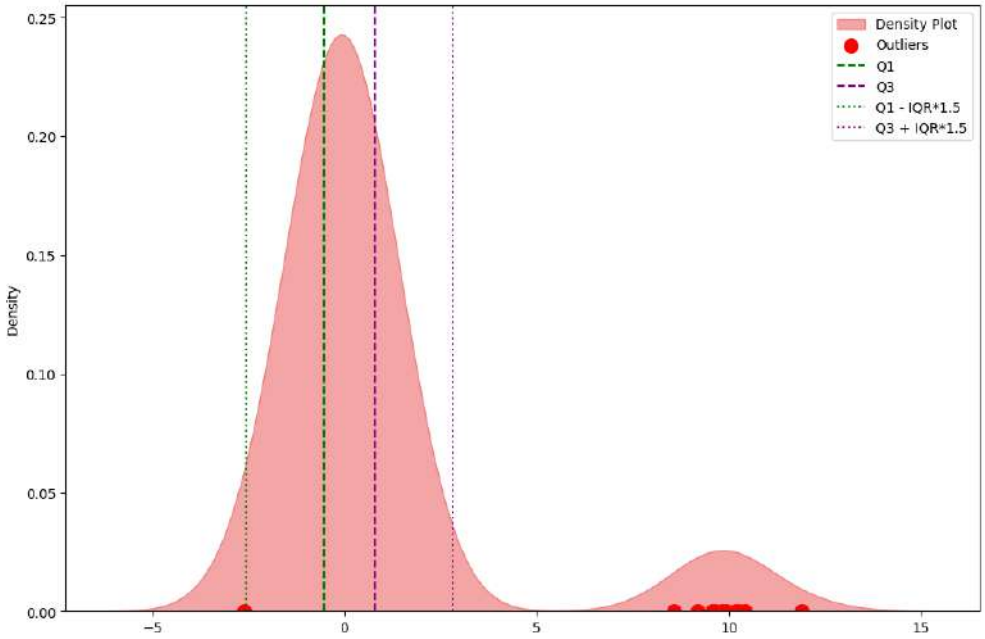


Figure 3. *Outlier detection using the IQR method.*

2.4 Machine Learning Evaluation

ML models require to be evaluated with metrics that could show how they perform and thus reflect their relevance, enabling them to be compared with other models dealing with the same task, problem, or application. The most common types of supervised tasks that a ML model can aim for are classification and regression. In a regression task, the output of the ML model is a continuous variable or variables. In the case of a classification task, the output is trained to predict categories, therefore being discrete. Nevertheless, the output is also continuous, representing the probability of each category being present. For regression, the most common metrics to evaluate the performance of a ML model are Mean Average Error (MAE) [98], Mean Squared Error (MSE) [98], Root Mean Squared Error (RMSE) [98] and Pearson correlation (r), as shown in formulas (2), (3), (4) and (5). Lower values for MAE, MSE and RMSE are indicative of a good performance. Meanwhile, for r bing closer to 1 indicates better performance.

$$MAE=\frac{1}{m}\sum_{i=1}^{m}|X_i-Y_i| \quad \textbf{(2)}$$

$$MSE=\frac{1}{m}\sum_{i=1}^{m}(X_i-Y_i)^2 \quad \textbf{(3)}$$

$$RMSE=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(X_i-Y_i)^2} \quad \textbf{(4)}$$

$$r=\frac{\sum_{i=1}^{m}(x_i-\bar{x})\sum_{i=1}^{m}(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i-\bar{x})^2\sum_{i=1}^{m}(y_i-\bar{y})^2}};\bar{x}=\frac{1}{m}\sum_{i=1}^{m}x_i;\bar{y}=\frac{1}{m}\sum_{i=1}^{m}y_i \quad \textbf{(5)}$$

Meanwhile, for a classification task, such as prediciting if a position in a genome will mutate or not, we can evaluate the performance of a ML model by its accuracy, sensitivity, specificity and Receiver Operating Characteristic – Area Under the Curve (ROC-AUC). The real labels/categories are considered the ground truth. When the real and predicted labels have the value of 1, positive, it is called a true positive (TP). When the real and predicted labels have a value of 0, it is considered a true negative (TN). Nevertheless, if the real value is 1 and the predicted value is 0, it is a false negative (FN). In addition, if the real value is 0 and the predicted value is 1, it is a false positive (FP). With these definitions, we can define accuracy, sensitivity (also called True Positive Rate TPR), specificity, false positive rate (FPR) as shown in formulas (6), (7), (8) and (9).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

$$Sensitivity\,(TPR) = \frac{TP}{TP + FN} \quad (7)$$

$$Specificity = \frac{TN}{TN + FP} \quad (8)$$

$$FPR = \frac{FP}{TN + FP} \quad (9)$$

The ROC curve is obtained by changing the decision threshold to consider a prediction positive (value of 1). Since the predictions are probabilities, the threshold changes and a new value for TPR and FPR is obtained. With these values for TPR and FPR at different thresholds, it is possible to have a plot with one on the horizontal axis and the other on the vertical axis. In addition, the ROC-AUC is the area under this curve.

2.5 Hyperparameter search

As mentioned before, a ML algorithm has hyperparameters, each with a specific resulting performance. AutoML frameworks take care of finding the right set of hyperparameters, but for some tasks, it is required to have a ML model that is not included in the AutoML framework or that the framework does not allow modification. In this scenario, the hyperparameter search should be done manually. Some techniques for reaching these goal are Grid Search, Bayesian Optimization and Genetic-evolutionary algorithms, among others [99]. Grid search consist of trying all combinations of hyperparameters and at the end, keeping the mixture with the best result. An example would be to search among the hyperparameters learning rate and batch size, giving two possible values for each one as shown in **Table 3**.

Table 3. Hyper-parameter grid search example

Learning rate	Batch size	Number of Neurons
1e-2	32	10
1e-2	32	20
1e-2	64	10
1e-2	64	20
1e-4	32	10
1e-4	32	20
1e-4	64	10
1e-4	64	20

2.6 Feature importance with SHAP values

When simple models, like linear models, are used to predict an outcome, the model itself shows the importance given to each variable. Nevertheless, for more complex models, such as ensemble models and deep learning models, this is not possible. Therefore, the way to determine the importance of each variable comes from an interpretable approximation of the original model. In this regard, SHapley Additive exPlanation values [85] obtain this result by taking each prediction of an input as a model to be interpreted, providing a feature importance interpretation with a strong consistency with human understanding, and by including three theoretical properties: Local accuracy, Missingness and Consistency. Local accuracy implies that the output of a model approximation given a simplified input (instead of the original input) equals the output of the original complex model given the original input (Formula 10). The property of missingness requires that variables that are missing in the original input have no impact in the variable interpretation (Formula 11). Finally, Consistency demands that if a modified model makes the simplified input increase its contribution or keep it as it was, then the attribution of that input should not decrease (Formula 12).

$$f(x) = g(x') = \phi_0 + \sum_{i=1}^{M} \phi_i x_i'$$

$$\phi = \text{to be solved}; x' = \text{simplified input}; g(x') = \textit{model approximation} \quad (10)$$

$$x_i' = 0 \rightarrow \phi_i = 0 \quad (11)$$

$$f'_x(z') - f_{z_0}' \geq f_x(z') - f_{z_0}$$

$$f_x(z') = f(h_x(z')); z_0 \textit{ when } z_i' = 0; z' \in \{0,1\}^M$$

$$M = \text{number of simplified inputs}; x = h_x(x'); h_x = \textit{mapping function}; f' = \textit{modified model} \quad (12)$$

In practice, a python library [100] can be used to determine the importance of each feature and obtain a plot, as shown in **Figure 4**. This type of plot places the most important variables at the top. The horizontal axis and the color go together. If the samples are in the negative part of the horizontal axis and they have a red color, it means that when that variable has negative values, the output of the model increases. For instance, in **Figure 4**, the three variables correlate negatively with the target variable (all with random values). This observation comes from seeing red values on the left and blue values on the right. This means, when the value of the variable decreases, so does the target.

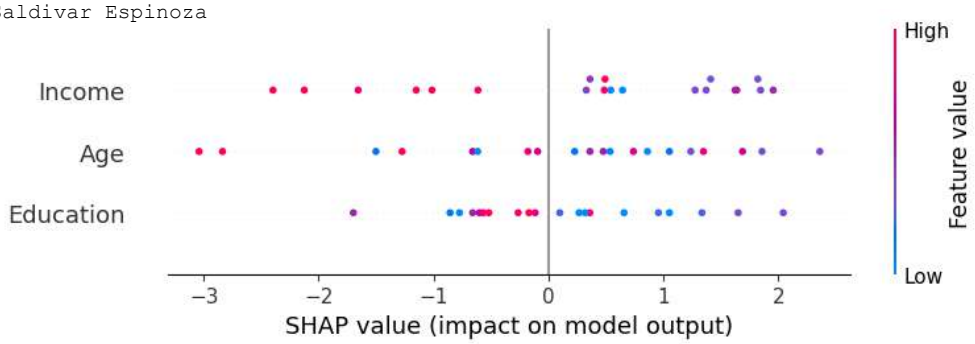


Figure 4. *SHAP values summary plot of three variables with random values.*

3. Machine Learning for SARS-CoV-2

ML has been applied for COVID-19 diagnosis, discriminating positive from negative samples. These negative samples might include other pathogens that produce respiratory infections like SARS-CoV, MERS-Cov or influenza. However, the diagnosis is no novelty, since it is possible to diagnose COVID-19 through Real Time – Polymerase Chain Reaction (RT-PCR). The advantage arrives by overcoming lack of this kind of data (RT-PCR), because it is expensive or because they try to make faster methods more reliable, like standardized blood tests. Thus, these studies leverage on ML for taming this limitations.

An example for applying ML for small amounts of genomic data, is the introduction of Neurochaos Learning for classifying SARS-CoV-2 from other viruses [101]. The application of this ML model comes handy since traditional ML models thrive from abundant data and their proposed approach deals with this scarcity. They [101] also have benchmarked this ML method with other classical models like K-Nearest Neighbors (KNN), Logistic Regression, Random Forest (RF), Support Sector Machines (SVM) and Naive Bayes classifiers.

In another case, ML is used to improve the reliability of faster methods for COVID-19 diagnosis. They aim to diagnose COVID-19 using an ultra-fast COVID-19 diagnostic sensor (UFC-19) [102]. Their negative samples include other virus like SARS-CoV, MERS-CoV, Human CoV and influenza. They [102] also benchmark traditional ML classifiers (AdaBoost, Decision Trees, Multilayer Perceptron and SVM) against Convolutional Neural Networks (CNN) [103]. A CNN is a modern deep learning method that extracts positional invariant patterns through the optimization of filters that utilize the convolution operation against the data. The CNN showed better performance over the other evaluated algorithms using UFC-19 data.

Moreover, in the line of bettering the applicability of standardized methods for COVID-19 diagnosis, ML has been used using common laboratory markers [104]. They evaluated the performance of Categorical gradient Boosting (CatBoost), SVM and LR. The best performing model was CatBoost, with which the rest of the evaluation was performed. These markers include white blood cells count, lymphocytes count, C-reactive protein levels and lactate dehydrogenase (LDH) levels. They [104] also include, for the evaluation of the diagnosis, a chest radiography for a medical doctor interpretation. They compared the results of the ML method alone, results evaluated from a radiologist and a radiologist helped with the ML method. They [104] include in the evaluation of the performance of these approaches, a temporal split. There is a testing set in the sample collection cohort when the model was made. In addition, there are other data collection cohorts after, that are used for testing. Something remarkable in the importance of these following cohorts is the elimination of the seasonal flu as a co-factor influencing the tests.

Furthermore, ML has been foreseen helpful to diagnose COVID-19 using hemograms. They start by the hypothesis that monocytes' and neutrophils' conformation and function will change on presence of SARS-CoV-2 [105]. They aim to capture this patter of changes with SVM as the ML method. They [105] fine tune their model and hyperparameters using a 10-fold cross validation to finally evaluate the generalization capabilities in a separate testing set. They [105] compare the predictions using hemogram data as input to the SVM ML model against RT-PCR.

Sharing this data split style of using a 10-fold cross validation plus an additional testing set for evaluating the model performance, in other study they benchmark Artificial Neural Networks (ANN), KNN plus SVMs to diagnose COVID-19 using laboratory data [106]. The data was collected from 18 laboratories, including variables such as C-reactive protein, lymphocytes and white blood cells as used before [104]. In addition, they have also analyzed hemoglobin, red blood cells, hematocrit, urea, potassium, sodium, age plus other 15 variables.

A similar study also applies ML on routine blood tests for COVID-19 diagnosis [107]. In the blood analysis they include LDH as before [104], in addition to hemoglobin, red blood cells, hematocrit, urea, potassium, sodium and age as the last mentioned study [106] plus other 20 variables. In contrast with the last two studies [105,106], in this study [107] they have used 5-fold cross validation with no mention of a separate testing set. The ML method they [107] used was RF.

In addition to the previous studies that diagnose COVID-19 from blood samples, other study aims to predict the same by using symptomatic information from an early stage instead of using blood samples [108]. They evaluate different ML models and configurations, including bagging and stacking. Among the models they used are XGBoost, RF, AdaBoost, ANN-Multi Layer Perceptron (MLP), KNN, Gradient Boosting Machine, Naive Bayes and LR. For the data split they used 10-fold cross validation, but did not include a separate testing set. Their best performing model was the stacking of other models.

The application of ML for diagnosis goes beyond and focus on predicting the possibility that a patient on hemodialysis (blood filtering for a person whose kidneys are not working properly) has a SARS-CoV-2 infection that was undetected [109]. They use XGBoost as ML model and their data was split into training, validation and testing, what would be equivalent to a 1-fold cross validation plus a separate testing set. They also utilize SHAP values to explain the most important features that the model take into account for its predictions. They found that the most important variable is the interdialytic weight gain comparing with the previous month. At higher values, the probability of having COVID-19 goes lower. The second most important variable they find was the temperature before the hemodialysis. Higher values correlate with higher probability of having COVID-19.

After the diagnosing stage, researchers have also found application for ML. They employed ML for predicting safe discharge from the emergency department, disease severity and mortality during hospitalization [110]. They evaluated the performance of RF, Gradient Boosting Machine and Decision Trees using a separate test split in addition to a 10-fold cross validation strategy. Their data included patients' LDH levels as previous studies [104,107], but they attribute more predictive power to the ROX index, which is the ratio of oxygen saturation (measured with a pulse oximeter) to oxygen fraction (FiO2) and the ratio of Partial Pressure of oxygen (PaO2) to FiO2.

In addition to this example, there is another study that use historical electronic health records from patients, to predict COVID-19 mortality within a 12-week period since the first positive test [111]. They also apply SHAP values to understand what factors increase the mortality probability. They show that at higher values for age and count of medicines that were taken previously, the mortality increases. On the other hand, at higher values of body mass index (BMI), the lower the risk.

Extending prognosis prediction, another group worked on predicting COVID-19 mortality, need for an ICU or ventilator and need for hospitalization [56]. For this, they used data from 91 thousand patients, including, demographic variables (age, gender, location, nationality), previous medical conditions and current COVID-19 diagnosis. The algorithms they used were LR, SVM, RF and gradient boosted decision trees (XGBoost). They took a 30% of the data randomly, for the test set, and for the hyperparameter tuning they used a stratified 10-fold cross-validation including a process of recursive feature elimination to keep only the most important variables. From the used models, they found that SVM and LR performed better than RF and XGBoost. When analyzing the most important variables for predicting hospitalization they found that age (Older than 65), pregnancy, diabetes, chronic renal insufficiency and immunosuppression were the most important variables. Meanwhile, when predicting mortality, age was also the most important variable, followed by immunosuppression, chronic renal insufficiency, obesity and diabetes. Moreover, for the prediction of ICU or ventilator need, if the development of pneumonia was known, this was the most predictive variable, followed by age, obesity, diabetes and hypertension.

Also aiming to predict the need of ICU or mechanical ventilation, mortality and hospitalization, in other study [112] they used variables such as dyspnea, past medical history, SpO2, social determinants of health, respiratory rate, age, BMI, diastolic blood

pressure, among others. Similar to [56], in this study they use LR, SVM, RF, and XGBoost. However, they included LSTM-Transformer neural networks, which performed as well as the others, but outperformed them for longer evaluation time windows (using data from 36 hours ago). The data split strategy included a separate test set, which was picked randomly. They also used a 5-fold cross validation for hyperparameter tuning. They repeated these two steps 5 times, thus, 5 randomly picked test sets, separated from the 5-fold cross validation splits. Similar to [56], in this study they also use a recursive feature elimination method, but besides, they eliminate variables that are not statistically significant and in addition remove variables that are highly correlated (>0.8) with each other. For mortality prediction, vital signs evolution was the most predictive set of factors through the usage of the LSTM-Transformer. The most important variables for predicting ICU need were history of having renal disease, cardiac arrest, C-reactive protein (similar to [105,107]), ferritin and LDH (similar to [105,108,111]). For hospitalization, they report as the top predictive features food insecurity and need for transportation, from a set of social determinants of health. In addition they also show that this kind of model has a racial-bias, having most false positives as a result of predicting that Black people will be hospitalized.

Among other applications during the pandemic, ML was used for predicting high-risk variants analyzing genomes of SARS-CoV-2 from GISAID [113]. They obtained features from a Haplotype network, used for tracing and visualizing SARS-CoV genealogies. They account for node size, network centrality, sequence growth ratio, geographic information entropy and mutation score in the spike protein. They trained the LightGBM ML model and have a testing split to validate their results. Other study aims towards the same objective but use incremental/online ML [114]. Incremental ML is a set of ML that is trained as samples are introduced, in contrast with traditional batch ML where when new samples wants to be included, it is required to include all previous samples plus the new one [115]. They [114] used the amino acid sequences to obtain k-mers (small sequences of amino acids) from the spike protein as features for training the Logistic Regression Incremental Learner. In a similar manner, other methods used for detecting variants of interest and concern, include the usage of Convolutional Neural Networks and Long Short Term Memory with genome sequences of the Spike protein [116].

Moreover, in the realm of drug discovery against SARS-CoV-2, ML has been used to prioritize FDA-approved compounds for in vitro testing [117]. They have trained several ML models (Bernoulli Naive Bayes, Adaboost, RF,SVM, KNN and deep learning) using 5-fold cross validation. They [117] have used as training data in vitro essays that measure the inhibition produced by compounds against SARS-CoV-2. On the other hand, for the same purpose, other study has used about 9.5 thousand compounds' docking scores (in silico) and tested the performance on a separate testing set (ZINC in vivo set) [118]. This later study, is focused on the inhibition of SARS-CoV-2 Main Protease applying XGBoost, decision trees and Artificial Neural Networks. Additionally, ML has been used to predict what medicinal plants from traditional Chinese medicine could be used to treat COVID-19 [119]. They first used bioactivity data from the ChEMBL database obtaining information about compounds that could target SARS-CoV-2 and build two ML models, RF and SVM. Afterwards, they [119] use the trained models to predict what compounds from a traditional Chinese medicine compounds database could be considered active (high affinity) against SARS-CoV-2. They obtain 1011 compounds,

from which 24 are very similar to FDA-approved drugs. Then they identify 74 plants from traditional Chinese medicine that carry these found compounds.

Through all these studies for diagnosing COVID-19, the ML models used were Neurochaos Learning [101], K-Nearest Neighbors [101,106,117], Logistic Regression [56,101,104,112,114], Random Forest [56,101,107,112,117,119], Support Vector Machines [56,101,102,104,105,112,117,119], Naive Bayes Classifier [101,117], Decision Trees [102,118], Multilayer Perceptron [102,106,118], Categorical gradient Boosting [104], XGBoost [56,112,118], LightGBM [113] , Adaboost [102,108,117] and deep learning (LSTM [112,116], Convolutional Neural Networks [102,116], Transformers [112]). These studies focused on improving alternative faster and cheaper diagnosis methods to the Real Time Polymerase Chain reaction in order to facilitate and increment the access to a reliable source for diagnosis. In addition, ML was used for predicting prognosis, variants of interest/variants of concern and identification of potential inhibitors against SARS-CoV-2.

REFERENCES

1. Yamayoshi, S.; Kawaoka, Y. Emergence of SARS-CoV-2 and Its Outlook. Glob. Health Med. 2020, 2, 1–2, doi:10.35772/ghm.2020.01009.
2. Feng, W.; Zong, W.; Wang, F.; Ju, S. Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2): A Review. Mol. Cancer 2020, 19, 100, doi:10.1186/s12943-020-01218-1.
3. Zhu, N.; Zhang, D.; Wang, W.; Li, X.; Yang, B.; Song, J.; Zhao, X.; Huang, B.; Shi, W.; Lu, R.; et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N. Engl. J. Med. 2020, 382, 727–733, doi:10.1056/NEJMoa2001017.
4. Lai, C.-C.; Shih, T.-P.; Ko, W.-C.; Tang, H.-J.; Hsueh, P.-R. Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) and Coronavirus Disease-2019 (COVID-19): The Epidemic and the Challenges. Int. J. Antimicrob. Agents 2020, 55, 105924, doi:10.1016/j.ijantimicag.2020.105924.
5. Shahin, M.; Alabed, H. Healthcare Management Challenges and Opportunities during COVID Pandemic: Management Challenges and Opportunities During COVID Pandemic. Curr. Res. Public Health 2023, 53–59.
6. Filip, R.; Gheorghita Puscaselu, R.; Anchidin-Norocel, L.; Dimian, M.; Savage, W.K. Global Challenges to Public Health Care Systems during the COVID-19 Pandemic: A Review of Pandemic Measures and Problems. J. Pers. Med. 2022, 12, 1295, doi:10.3390/jpm12081295.
7. COVID-19 Map Available online: https://coronavirus.jhu.edu/map.html (accessed on 11 February 2024).
8. Kaul, R.; Devi, S. Coronavirus - A Crippling Affliction to Humans. Recent Pat. Biotechnol. 16, 226–242.
9. gianni Transmission Mode Associated with Coronavirus Disease 2019: A Review. Eur. Rev. 2020.
10. Connors, J.; Bell, M.R.; Marcy, J.; Kutzler, M.; Haddad, E.K. The Impact of Immuno-Aging on SARS-CoV-2 Vaccine Development. GeroScience 2021, 43, 31–51, doi:10.1007/s11357-021-00323-3.
11. Dai, M.; Liu, D.; Liu, M.; Zhou, F.; Li, G.; Chen, Z.; Zhang, Z.; You, H.; Wu, M.; Zheng, Q.; et al. Patients with Cancer Appear More Vulnerable to SARS-CoV-2: A Multicenter Study during the COVID-19 Outbreak. Cancer Discov. 2020, 10, 783–791, doi:10.1158/2159-8290.CD-20-0422.
12. Kaur, H.; Thakur, J.S.; Paika, R.; Advani, S.M. Impact of Underlying Comorbidities on Mortality in SARS-COV-2 Infected Cancer Patients: A Systematic Review and Meta-Analysis. Asian Pac. J. Cancer Prev. 2021, 22, 1333–1349, doi:10.31557/APJCP.2021.22.5.1333.
13. Seidu, S.; Gillies, C.; Zaccardi, F.; Kunutsor, S.K.; Hartmann-Boyce, J.; Yates, T.; Singh, A.K.; Davies, M.J.; Khunti, K. The Impact of Obesity on Severe Disease and Mortality in People with SARS-CoV-2: A Systematic Review and Meta-Analysis. Endocrinol. Diabetes Metab. 2021, 4, e00176, doi:10.1002/edm2.176.
14. Michalakis, K.; Ilias, I. SARS-CoV-2 Infection and Obesity: Common Inflammatory and Metabolic Aspects. Diabetes Metab. Syndr. Clin. Res. Rev. 2020, 14, 469–471, doi:10.1016/j.dsx.2020.04.033.
15. Piracha, Z.Z.; Saeed, U.; Sarfraz, R.; Asif, U.; Waheed, Y.; Raheem, A.; Tahir, M.; Kiran, S.; Pervaiz, I.; Uppal, R. Impact of SARS-CoV-2 on Onset of Diabetes and Associated Complications. Arch. Clin. Biomed. Res. 2022, 6, 217–227.
16. Luque, P.; Castilla-Guerra, L.; Sanchez, N.; Delgado, M.; Alegre, M.; Carmona, E.; Rico, M.A. THE IMPACT OF HYPERTENSION AND ANTIHIPERTENSIVE TREATMENTS ON PATIENTS WITH SARS-COV-2: A RETROSPECTIVE-COHORT STUDY. J. Hypertens. 2022, 40, e29, doi:10.1097/01.hjh.0000835528.11250.2d.
17. Koyyada, R.; Nagalla, B.; Tummala, A.; Singh, A.D.; Patnam, S.; Barigala, R.; Kandala, M.; Krishna, V.; Manda, S.V. Prevalence and Impact of Preexisting Comorbidities on Overall Clinical Outcomes of Hospitalized COVID-19 Patients. BioMed Res. Int. 2022, 2022, e2349890, doi:10.1155/2022/2349890.

18. *gianni Novel Coronavirus 2019 (Sars-CoV2): A Global Emergency That Needs New Approaches? Eur. Rev. 2020.*

19. *Garvin, M.R.; T. Prates, E.; Pavicic, M.; Jones, P.; Amos, B.K.; Geiger, A.; Shah, M.B.; Streich, J.; Felipe Machado Gazolla, J.G.; Kainer, D.; et al. Potentially Adaptive SARS-CoV-2 Mutations Discovered with Novel Spatiotemporal and Explainable AI Models. Genome Biol. 2020, 21, 304, doi:10.1186/s13059-020-02191-0.*

20. *V'kovski, P.; Kratzel, A.; Steiner, S.; Stalder, H.; Thiel, V. Coronavirus Biology and Replication: Implications for SARS-CoV-2. Nat. Rev. Microbiol. 2021, 19, 155–170, doi:10.1038/s41579-020-00468-6.*

21. *Kim, D.; Lee, J.-Y.; Yang, J.-S.; Kim, J.W.; Kim, V.N.; Chang, H. The Architecture of SARS-CoV-2 Transcriptome. Cell 2020, 181, 914-921.e10, doi:10.1016/j.cell.2020.04.011.*

22. *Huston, N.C.; Wan, H.; Strine, M.S.; Tavares, R. de C.A.; Wilen, C.B.; Pyle, A.M. Comprehensive in Vivo Secondary Structure of the SARS-CoV-2 Genome Reveals Novel Regulatory Motifs and Mechanisms. Mol. Cell 2021, 81, 584-598.e5, doi:10.1016/j.molcel.2020.12.041.*

23. *Lubin, J.H.; Zardecki, C.; Dolan, E.M.; Lu, C.; Shen, Z.; Dutta, S.; Westbrook, J.D.; Hudson, B.P.; Goodsell, D.S.; Williams, J.K.; et al. Evolution of the SARS-CoV-2 Proteome in Three Dimensions (3D) during the First 6 Months of the COVID-19 Pandemic. Proteins Struct. Funct. Bioinforma. 2022, 90, 1054–1080, doi:10.1002/prot.26250.*

24. *Bai, C.; Zhong, Q.; Gao, G.F. Overview of SARS-CoV-2 Genome-Encoded Proteins. Sci. China Life Sci. 2022, 65, 280–294, doi:10.1007/s11427-021-1964-4.*

25. *Severe Acute Respiratory Syndrome Coronavirus 2 Isolate Wuhan-Hu-1, Complete Genome 2020.*

26. *Yoshimoto, F.K. The Proteins of Severe Acute Respiratory Syndrome Coronavirus-2 (SARS CoV-2 or n-COV19), the Cause of COVID-19. Protein J. 2020, 39, 198–216, doi:10.1007/s10930-020-09901-4.*

27. *Severe Acute Respiratory Syndrome Coronavirus ORF3a Protein Activates the NLRP3 Inflammasome by Promoting TRAF3□dependent Ubiquitination of ASC., doi:10.1096/fj.201802418R.*

28. *Arshad, N.; Laurent-Rolle, M.; Ahmed, W.S.; Hsu, J.C.-C.; Mitchell, S.M.; Pawlak, J.; Sengupta, D.; Biswas, K.H.; Cresswell, P. SARS-CoV-2 Accessory Proteins ORF7a and ORF3a Use Distinct Mechanisms to down-Regulate MHC-I Surface Expression. Proc. Natl. Acad. Sci. 2023, 120, e2208525120, doi:10.1073/pnas.2208525120.*

29. *Zandi, M.; Shafaati, M.; Kalantar-Neyestanaki, D.; Pourghadamyari, H.; Fani, M.; Soltani, S.; Kaleji, H.; Abbasi, S. The Role of SARS-CoV-2 Accessory Proteins in Immune Evasion. Biomed. Pharmacother. 2022, 156, 113889, doi:10.1016/j.biopha.2022.113889.*

30. *Schubert, K.; Karousis, E.D.; Jomaa, A.; Scaiola, A.; Echeverria, B.; Gurzeler, L.-A.; Leibundgut, M.; Thiel, V.; Mühlemann, O.; Ban, N. SARS-CoV-2 Nsp1 Binds the Ribosomal mRNA Channel to Inhibit Translation. Nat. Struct. Mol. Biol. 2020, 27, 959–966, doi:10.1038/s41594-020-0511-8.*

31. *Brant, A.C.; Tian, W.; Majerciak, V.; Yang, W.; Zheng, Z.-M. SARS-CoV-2: From Its Discovery to Genome Structure, Transcription, and Replication. Cell Biosci. 2021, 11, 136, doi:10.1186/s13578-021-00643-z.*

32. *Littler, D.R.; Gully, B.S.; Colson, R.N.; Rossjohn, J. Crystal Structure of the SARS-CoV-2 Non-Structural Protein 9, Nsp9. iScience 2020, 23, doi:10.1016/j.isci.2020.101258.*

33. *Slanina, H.; Madhugiri, R.; Bylapudi, G.; Schultheiß, K.; Karl, N.; Gulyaeva, A.; Gorbalenya, A.E.; Linne, U.; Ziebuhr, J. Coronavirus Replication–Transcription Complex: Vital and Selective NMPylation of a Conserved Site in Nsp9 by the NiRAN-RdRp Subunit. Proc. Natl. Acad. Sci. 2021, 118, e2022310118, doi:10.1073/pnas.2022310118.*

34. *Mandala, V.S.; McKay, M.J.; Shcherbakov, A.A.; Dregni, A.J.; Kolocouris, A.; Hong, M. Structure and Drug Binding of the SARS-CoV-2 Envelope Protein Transmembrane Domain in Lipid Bilayers. Nat.*

Struct. Mol. Biol. 2020, 27, 1202–1208, doi:10.1038/s41594-020-00536-8.

35. Zhang, Y.; Chen, Y.; Li, Y.; Huang, F.; Luo, B.; Yuan, Y.; Xia, B.; Ma, X.; Yang, T.; Yu, F.; et al. The ORF8 Protein of SARS-CoV-2 Mediates Immune Evasion through down-Regulating MHC-Ⅰ. Proc. Natl. Acad. Sci. 2021, 118, e2024202118, doi:10.1073/pnas.2024202118.

36. Han, L.; Zheng, Y.; Deng, J.; Nan, M.-L.; Xiao, Y.; Zhuang, M.-W.; Zhang, J.; Wang, W.; Gao, C.; Wang, P.-H. SARS-CoV-2 ORF10 Antagonizes STING-Dependent Interferon Activation and Autophagy. J. Med. Virol. 2022, 94, 5174–5188, doi:10.1002/jmv.27965.

37. Malone, B.; Urakova, N.; Snijder, E.J.; Campbell, E.A. Structures and Functions of Coronavirus Replication–Transcription Complexes and Their Relevance for SARS-CoV-2 Drug Design. Nat. Rev. Mol. Cell Biol. 2022, 23, 21–39, doi:10.1038/s41580-021-00432-z.

38. Fitzgerald, D.M.; Rosenberg, S.M. What Is Mutation? A Chapter in the Series: How Microbes "Jeopardize" the Modern Synthesis. PLOS Genet. 2019, 15, e1007995, doi:10.1371/journal.pgen.1007995.

39. Plante, J.A.; Liu, Y.; Liu, J.; Xia, H.; Johnson, B.A.; Lokugamage, K.G.; Zhang, X.; Muruato, A.E.; Zou, J.; Fontes-Garfias, C.R.; et al. Spike Mutation D614G Alters SARS-CoV-2 Fitness. Nature 2021, 592, 116–121, doi:10.1038/s41586-020-2895-3.

40. Murano, K.; Guo, Y.; Siomi, H. The Emergence of SARS-CoV-2 Variants Threatens to Decrease the Efficacy of Neutralizing Antibodies and Vaccines. Biochem. Soc. Trans. 2021, 49, 2879–2890, doi:10.1042/BST20210859.

41. Khare, S.; Gurry, C.; Freitas, L.; Schultz, M.B.; Bach, G.; Diallo, A.; Akite, N.; Ho, J.; Lee, R.T.; Yeo, W.; et al. GISAID's Role in Pandemic Response. China CDC Wkly. 2021, 3, 1049–1051, doi:10.46234/ccdcw2021.255.

42. SARS-CoV-2 Mutation Portal Available online: http://sarscov2-mutation-portal.urv.cat/SARS-CoV-2_mutation-portal/ (accessed on 21 February 2024).

43. Kuipers, J.; Jahn, K.; Raphael, B.J.; Beerenwinkel, N. A Statistical Test on Single-Cell Data Reveals Widespread Recurrent Mutations in Tumor Evolution 2016, 094722.

44. CDC Coronavirus Disease 2019 (COVID-19) Available online: https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html (accessed on 8 November 2021).

45. Scovino, A.M.; Dahab, E.C.; Vieira, G.F.; Freire-de-Lima, L.; Freire-de-Lima, C.G.; Morrot, A. SARS-CoV-2's Variants of Concern: A Brief Characterization. Front. Immunol. 2022, 13.

46. El Naqa, I.; Murphy, M.J. What Is Machine Learning? In Machine Learning in Radiation Oncology: Theory and Applications; El Naqa, I., Li, R., Murphy, M.J., Eds.; Springer International Publishing: Cham, 2015; pp. 3–11 ISBN 978-3-319-18305-3.

47. Jordan, M.I.; Mitchell, T.M. Machine Learning: Trends, Perspectives, and Prospects. Science 2015, 349, 255–260, doi:10.1126/science.aaa8415.

48. Szijártó, Á.; Fábián, A.; Lakatos, B.K.; Tolvaj, M.; Merkely, B.; Kovács, A.; Tokodi, M. A Machine Learning Framework for Performing Binary Classification on Tabular Biomedical Data. Imaging 2023, 15, 1–6, doi:10.1556/1647.2023.00109.

49. Khan, A.I.; Al-Habsi, S. Machine Learning in Computer Vision. Procedia Comput. Sci. 2020, 167, 1444–1451, doi:10.1016/j.procs.2020.03.355.

50. Nagarhalli, T.P.; Mhatre, S.; Patil, S.; Patil, P. The Review of Natural Language Processing Applications with Emphasis on Machine Learning Implementations. In Proceedings of the 2022 International Conference on Electronics and Renewable Systems (ICEARS); March 2022; pp. 1353–1358.

51. Akanksha, E.; Jyoti; Sharma, N.; Gulati, K. Review on Reinforcement Learning, Research Evolution and Scope of Application. In Proceedings of the 2021 5th International Conference on Computing Methodologies and Communication (ICCMC); April 2021; pp. 1416–1423.

52. Dhall, D.; Kaur, R.; Juneja, M. Machine Learning: A Review of the Algorithms and Its Applications. In Proceedings of the Proceedings of ICRIC 2019; Singh, P.K., Kar, A.K., Singh, Y., Kolekar, M.H., Tanwar, S.,

Eds.; Springer International Publishing: Cham, 2020; pp. 47–63.

53. Dimoulas, C.A. Machine Learning. In The SAGE Encyclopedia of Surveillance, Security, and Privacy; SAGE Publications, Inc.: Thousand Oaks, 2018; pp. 591–592.

54. Lyu, S. Artificial Intelligence and Machine Learning. In Practical Rust Projects: Building Game, Physical Computing, and Machine Learning Applications; Lyu, S., Ed.; Apress: Berkeley, CA, 2020; pp. 187–235 ISBN 978-1-4842-5599-5.

55. Ling, Q. Machine Learning Algorithms Review. Appl. Comput. Eng. 2023, 4, 91–98, doi:10.54254/2755-2721/4/20230355.

56. Wollenstein-Betech, S.; Cassandras, C.G.; Paschalidis, I.Ch. Personalized Predictive Models for Symptomatic COVID-19 Patients Using Basic Preconditions: Hospitalizations, Mortality, and the Need for an ICU or Ventilator. Int. J. Med. Inf. 2020, 142, 104258, doi:10.1016/j.ijmedinf.2020.104258.

57. Damilola, S. A Review of Unsupervised Artificial Neural Networks with Applications. Int. J. Comput. Appl. 2019, 181, 22–26, doi:10.5120/ijca2019918425.

58. Unsupervised Class-Expert Learning for Supporting Covid-19 Triage Based on Computed Tomography Data. Learn. NonLinear Models.

59. Wolberg, E.J. The Method of Least Squares. In Designing Quantitative Experiments: Prediction Analysis; Wolberg, J., Ed.; Springer: Berlin, Heidelberg, 2010; pp. 47–89 ISBN 978-3-642-11589-9.

60. Uddin, S.; Haque, I.; Lu, H.; Moni, M.A.; Gide, E. Comparative Performance Analysis of K-Nearest Neighbour (KNN) Algorithm and Its Different Variants for Disease Prediction. Sci. Rep. 2022, 12, 6256, doi:10.1038/s41598-022-10358-x.

61. Chen, H.; Hu, S.; Hua, R.; Zhao, X. Improved Naive Bayes Classification Algorithm for Traffic Risk Management. EURASIP J. Adv. Signal Process. 2021, 2021, 30, doi:10.1186/s13634-021-00742-6.

62. Zhang, Y. Support Vector Machine Classification Algorithm and Its Application. In Proceedings of the Information Computing and Applications; Liu, C., Wang, L., Yang, A., Eds.; Springer: Berlin, Heidelberg, 2012; pp. 179–186.

63. FAST MULTI-CLASS IMAGE ANNOTATION WITH RANDOM SUBWINDOWS AND MULTIPLE OUTPUT RANDOMIZED TREES: In Proceedings of the Proceedings of the Fourth International Conference on Computer Vision Theory and Applications; SciTePress - Science and and Technology Publications: Lisboa, Portugal, 2009; pp. 196–203.

64. Breiman, L. Random Forests. Mach. Learn. 2001, 45, 5–32, doi:10.1023/A:1010933404324.

65. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Association for Computing Machinery: New York, NY, USA, August 13 2016; pp. 785–794.

66. Block, H.D. A Review of "perceptrons: An Introduction to Computational Geometry". Inf. Control 1970, 17, 501–522, doi:10.1016/S0019-9958(70)90409-2.

67. Chen, C.L.P. Deep Learning for Pattern Learning and Recognition. In Proceedings of the 2015 IEEE 10th Jubilee International Symposium on Applied Computational Intelligence and Informatics; May 2015; pp. 17–17.

68. Gohar, U.; Biswas, S.; Rajan, H. Towards Understanding Fairness and Its Composition in Ensemble Machine Learning. In Proceedings of the Proceedings of the 45th International Conference on Software Engineering; IEEE Press: Melbourne, Victoria, Australia, July 26 2023; pp. 1533–1545.

69. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning Representations by Back-Propagating Errors. Nature 1986, 323, 533–536, doi:10.1038/323533a0.

70. Tian, Y.; Zhang, Y.; Zhang, H. Recent Advances in Stochastic Gradient Descent in Deep Learning. Mathematics 2023, 11, 682, doi:10.3390/math11030682.

71. Choi, D.; Shallue, C.J.; Nado, Z.; Lee, J.; Maddison, C.J.; Dahl, G.E. On Empirical Comparisons of Optimizers for Deep Learning. 2019.

72. Zeiler, M.D. ADADELTA: An Adaptive Learning Rate Method 2012.

73. Wynters, L. Parallel Processing on NVIDIA Graphics Processing Units Using CUDA. J. Comput. Sci. Coll. 2011, 26, 58–66.
74. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. ArXiv170603762 Cs 2017.
75. Turner, R.E. An Introduction to Transformers 2023.
76. Eldeeb, H.; Maher, M.; Elshawi, R.; Sakr, S. AutoMLBench: A Comprehensive Experimental Evaluation of Automated Machine Learning Frameworks 2023.
77. Conrad, F.; Mälzer, M.; Schwarzenberger, M.; Wiemer, H.; Ihlenfeldt, S. Benchmarking AutoML for Regression Tasks on Small Tabular Data in Materials Design. Sci. Rep. 2022, 12, 19350, doi:10.1038/s41598-022-23327-1.
78. Płońska, A.; Płoński, P. MLJAR: State-of-the-Art Automated Machine Learning Framework for Tabular Data. Version 0.10.3 2021.
79. Hollmann, N.; Müller, S.; Eggensperger, K.; Hutter, F. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second 2022.
80. Le, T.T.; Fu, W.; Moore, J.H. Scaling Tree-Based Automated Machine Learning to Biomedical Big Data with a Feature Set Selector. Bioinformatics 2020, 36, 250–256, doi:10.1093/bioinformatics/btz470.
81. Erickson, N.; Mueller, J.; Shirkov, A.; Zhang, H.; Larroy, P.; Li, M.; Smola, A. AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data 2020.
82. Marcinkevičs, R.; Vogt, J.E. Interpretable and Explainable Machine Learning: A Methods-Centric Overview with Concrete Examples. WIREs Data Min. Knowl. Discov. 2023, 13, e1493, doi:10.1002/widm.1493.
83. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier 2016.
84. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps 2014.
85. Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. ArXiv170507874 Cs Stat 2017.
86. Pizarroso, J.; Portela, J.; Muñoz, A. NeuralSens: Sensitivity Analysis of Neural Networks. J. Stat. Softw. 2022, 102, 1–36, doi:10.18637/jss.v102.i07.
87. Birba, D.E. A Comparative Study of Data Splitting Algorithms for Machine Learning Model Selection; 2020;
88. Géron, A. Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow : Concepts, Tools, and Techniques to Build Intelligent Systems; Second edition.; O'Reilly Media, Inc.: Sebastopol, CA, 2019; ISBN 978-1-4920-3264-9.
89. Kim, J.; Kim, J. The Impact of Imbalanced Training Data on Machine Learning for Author Name Disambiguation. Scientometrics 2018, 117, 511–526, doi:10.1007/s11192-018-2865-9.
90. Murphy, K.P. Machine Learning: A Probabilistic Perspective; Adaptive computation and machine learning series; 4. print. (fixed many typos).; MIT Press: Cambridge, Mass., 2013; ISBN 978-0-262-01802-9.
91. Chollet, F. Deep Learning with Python; Manning, 2017; ISBN 978-1-61729-443-3.
92. Hancock, J.T.; Khoshgoftaar, T.M. Survey on Categorical Data for Neural Networks. J. Big Data 2020, 7, 28, doi:10.1186/s40537-020-00305-w.
93. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space 2013.
94. Buuren, S. van; Groothuis-Oudshoorn, K. Mice: Multivariate Imputation by Chained Equations in R. J. Stat. Softw. 2011, 45, 1–67, doi:10.18637/jss.v045.i03.
95. Buck, S.F. A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with an Electronic Computer. J. R. Stat. Soc. Ser. B Methodol. 1960, 22, 302–306.

96. Zheng, A.; Casari, A. Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists; O'Reilly, 2018; ISBN 978-1-4919-5324-2.
97. Dash, Ch.S.K.; Behera, A.K.; Dehuri, S.; Ghosh, A. An Outliers Detection and Elimination Framework in Classification Task of Data Mining. Decis. Anal. J. 2023, 6, 100164, doi:10.1016/j.dajour.2023.100164.
98. Chicco, D.; Warrens, M.J.; Jurman, G. The Coefficient of Determination R-Squared Is More Informative than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis Evaluation. PeerJ Comput. Sci. 2021, 7, e623, doi:10.7717/peerj-cs.623.
99. Claesen, M.; De Moor, B. Hyperparameter Search in Machine Learning 2015.
100. Welcome to the SHAP Documentation — SHAP Latest Documentation Available online: https://shap.readthedocs.io/en/latest/index.html (accessed on 27 February 2024).
101. Harikrishnan, N.B.; Pranay, S.Y.; Nagaraj, N. Classification of SARS-CoV-2 Viral Genome Sequences Using Neurochaos Learning. Med. Biol. Eng. Comput. 2022, 60, 2245–2255, doi:10.1007/s11517-022-02591-3.
102. Gecgel, O.; Ramanujam, A.; Botte, G.G. Selective Electrochemical Detection of SARS-CoV-2 Using Deep Learning. Viruses 2022, 14, 1930, doi:10.3390/v14091930.
103. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-Based Learning Applied to Document Recognition. Proc. IEEE 1998, 86, 2278–2324, doi:10.1109/5.726791.
104. Du, R.; Tsougenis, E.D.; Ho, J.W.K.; Chan, J.K.Y.; Chiu, K.W.H.; Fang, B.X.H.; Ng, M.Y.; Leung, S.-T.; Lo, C.S.Y.; Wong, H.-Y.F.; et al. Machine Learning Application for the Prediction of SARS-CoV-2 Infection Using Blood Tests and Chest Radiograph. Sci. Rep. 2021, 11, 14250, doi:10.1038/s41598-021-93719-2.
105. Gómez-Rojas, S.; Segura, G.P.; Ollé, J.; Carreño Gómez-Tarragona, G.; Medina, J.G.; Aguado, J.M.; Guerrero, E.V.; Santaella, M.P.; Martínez-López, J. A Machine Learning Tool for the Diagnosis of SARS-CoV-2 Infection from Hemogram Parameters. J. Cell. Mol. Med. 2023, 27, 3423–3430, doi:10.1111/jcmm.17864.
106. Salman, A.O.; Geman, O. Evaluating Three Machine Learning Classification Methods for Effective COVID-19 Diagnosis. Int. J. Math. Stat. Comput. Sci. 2023, 1, 1–14, doi:10.59543/ijmscs.v1i.7693.
107. Tschoellitsch, T.; Dünser, M.; Böck, C.; Schwarzbauer, K.; Meier, J. Machine Learning Prediction of SARS-CoV-2 Polymerase Chain Reaction Results with Routine Blood Tests. Lab. Med. 2021, 52, 146–149, doi:10.1093/labmed/lmaa111.
108. Dritsas, E.; Trigka, M. Supervised Machine Learning Models to Identify Early-Stage Symptoms of SARS-CoV-2. Sensors 2023, 23, 40, doi:10.3390/s23010040.
109. Monaghan, C.K.; Larkin, J.W.; Chaudhuri, S.; Han, H.; Jiao, Y.; Bermudez, K.M.; Weinhandl, E.D.; Dahne-Steuber, I.A.; Belmonte, K.; Neri, L.; et al. Machine Learning for Prediction of Patients on Hemodialysis with an Undetected SARS-CoV-2 Infection. Kidney360 2021, 2, 456, doi:10.34067/KID.0003802020.
110. Casano, N.; Santini, S.J.; Vittorini, P.; Sinatti, G.; Carducci, P.; Mastroianni, C.M.; Ciardi, M.R.; Pasculli, P.; Petrucci, E.; Marinangeli, F.; et al. Application of Machine Learning Approach in Emergency Department to Support Clinical Decision Making for SARS-CoV-2 Infected Patients. J. Integr. Bioinforma. 2023, 20, doi:10.1515/jib-2022-0047.
111. Zucco, A.G.; Agius, R.; Svanberg, R.; Moestrup, K.S.; Marandi, R.Z.; MacPherson, C.R.; Lundgren, J.; Ostrowski, S.R.; Niemann, C.U. Personalized Survival Probabilities for SARS-CoV-2 Positive Patients by Explainable Machine Learning. Sci. Rep. 2022, 12, 13879, doi:10.1038/s41598-022-17953-y.
112. Rodriguez, V.A.; Bhave, S.; Chen, R.; Pang, C.; Hripcsak, G.; Sengupta, S.; Elhadad, N.; Green, R.; Adelman, J.; Metitiri, K.S.; et al. Development and Validation of Prediction Models for Mechanical Ventilation, Renal Replacement Therapy, and Readmission in COVID-19 Patients. J. Am. Med. Inform. Assoc. 2021, 28, 1480–1488, doi:10.1093/jamia/ocab029.

113. Li, L.; Li, C.; Li, N.; Zou, D.; Zhao, W.; Xue, Y.; Zhang, Z.; Bao, Y.; Song, S. Machine Learning Detection of SARS-CoV-2 High-Risk Variants 2023, 2023.04.19.537460.

114. Nicora, G.; Marini, S.; Salemi, M.; Bellazzi, R. Dynamic Prediction of Non-Neutral SARS-Cov-2 Variants Using Incremental Machine Learning. In Challenges of Trustable AI and Added-Value on Health; IOS Press, 2022; pp. 654–658.

115. Bouchachia, A.; Gabrys, B.; Sahel, Z. Overview of Some Incremental Learning Algorithms. In Proceedings of the 2007 IEEE International Fuzzy Systems Conference; July 2007; pp. 1–6.

116. Mwanga, M.J.; Obura, H.O.; Evans, M.; Awe, O.I. Enhanced Deep Convolutional Neural Network for SARS-CoV-2 Variants Classification 2023, 2023.08.09.552643.

117. Gawriljuk, V.O.; Zin, P.P.K.; Puhl, A.C.; Zorn, K.M.; Foil, D.H.; Lane, T.R.; Hurst, B.; Tavella, T.A.; Costa, F.T.M.; Lakshmanane, P.; et al. Machine Learning Models Identify Inhibitors of SARS-CoV-2. J. Chem. Inf. Model. 2021, 61, 4224–4235, doi:10.1021/acs.jcim.1c00683.

118. Bucinsky, L.; Bortňák, D.; Gall, M.; Matúška, J.; Milata, V.; Pitoňák, M.; Štekláč, M.; Végh, D.; Zajaček, D. Machine Learning Prediction of 3CLpro SARS-CoV-2 Docking Scores. Comput. Biol. Chem. 2022, 98, 107656, doi:10.1016/j.compbiolchem.2022.107656.

119. Liang, J.; Zheng, Y.; Tong, X.; Yang, N.; Dai, S. In Silico Identification of Anti-SARS-CoV-2 Medicinal Plants Using Cheminformatics and Machine Learning. Molecules 2023, 28, 208, doi:10.3390/molecules28010208.

Objectives

Despite the original focus of our research program on Nutrigenomics and personalized nutrition, the emergence of the COVID-19 pandemic prompted a shift in our research priorities. Consequently, our research group collaboratively redirected our efforts towards gaining a better understanding of the SARS-CoV-2 virus and its societal impact. Based on my experience applying machine learning and in alignment with this shift, we outlined the following objectives for this PhD thesis:

To develop a predictive model for COVID-19 mortality. We aim to utilize health, socioeconomic, and nutritional data from United States at a county level from 2019, to predict COVID-19 fatality rates in these counties until September 2023. This objective will be achieved by employing Automatic Machine Learning (AutoML) frameworks to derive the predictive model (Manuscript 1).

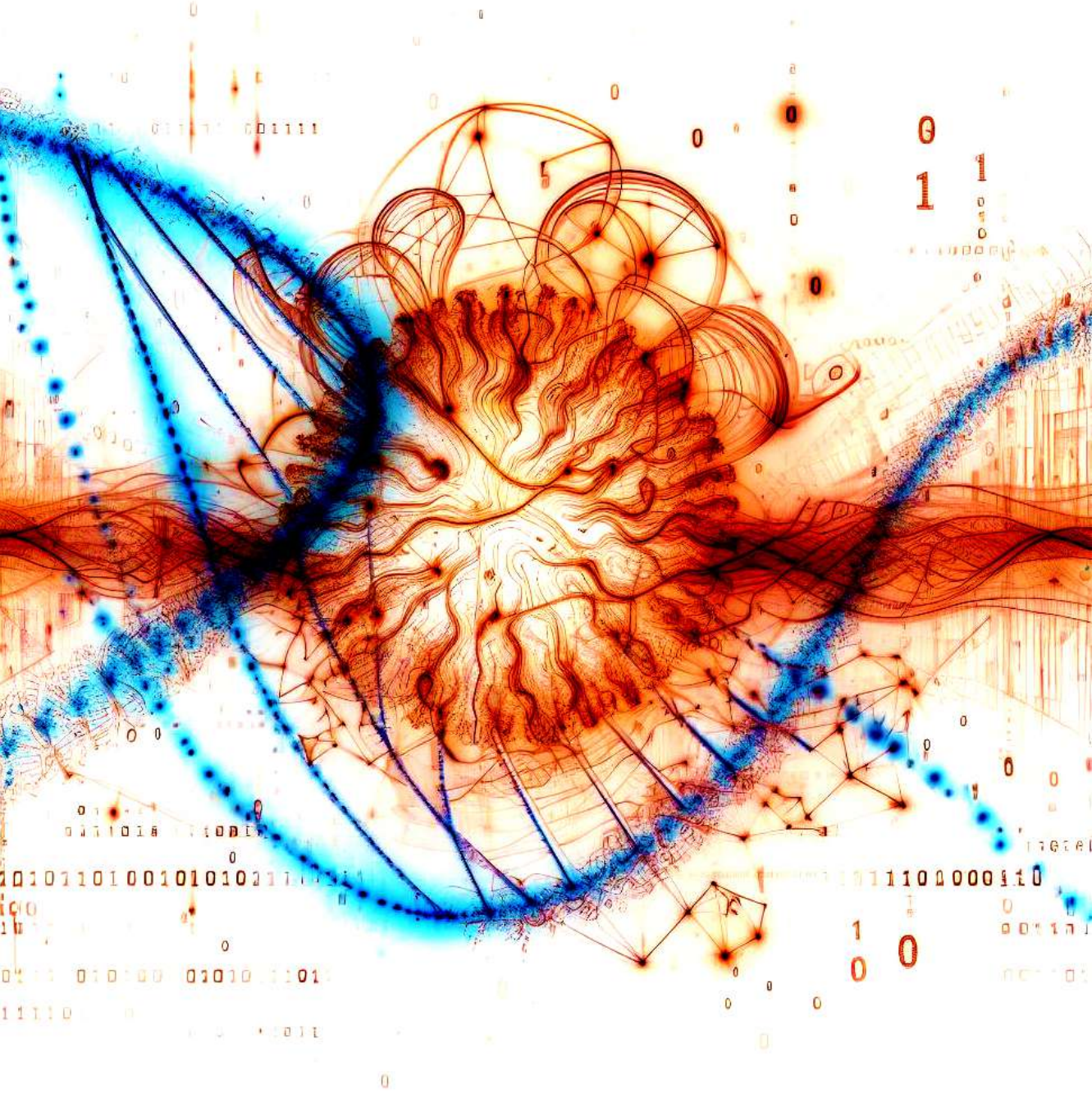
To identify the most influential health, socioeconomic, and nutritional factors that have a significant impact on COVID-19 mortality. Using the machine learning model for predicting COVID-19 fatalities developed in the previous objective, our aim is to assess the significance of each variable used by the model, using SHapley Additive exPlanation (SHAP) values (Manuscript 1).

To characterize SARS-CoV-2 mutations and identify recurrent mutations. For this purpose we will utilize the SARS-CoV-2 genome sequences available in the Global Initiative on Sharing All Influenza Data (GISAID) database to identify and characterize Single Nucleotide Variants (SNV), deletions, and insertions, and analyze their frequency (Manuscript 2). The analysis of SNV present in distantly-related lineages will allow us to identify recurrent mutations, i.e. mutations that occur independently and many times throughout the virus' evolution (Manuscript 3).

To predict recurrent mutations in SARS-CoV-2 using a machine learning model. We will use the characterized SARS-CoV-2 mutations from the previous objective, plus RNA reactivity values and RNA secondary structure information, to train a machine learning model for predicting SARS-CoV-2 recurrent mutations. We will validate the effectiveness of the model using more recent data, extending beyond the dataset utilized for training (Manuscript 3).

To identify the most important factors for predicting recurrent mutations in SARS-CoV-2. We will use the machine learning model developed in the previous objective, to obtain its SHapley Additive exPlanation (SHAP) values to pinpoint the most relevant variables for predicting recurrent mutations in SARS-CoV-2 (Manuscript 3).

Results



Manuscript 1

Prediction of COVID-19 Mortality: Integrating Health, Socioeconomic, and Nutritional Factors at the County Level

Bryan Saldivar-Espinoza (1), Pere Puigbò (2,3,4), Adrià Cereto-Massagué (5),Gerard Pujadas (1) and Santiago Garcia-Vallve (1,*)

1 Research Group in Cheminformatics & Nutrition, Departament de Bioquímica i Biotecnologia, Campus de Sescelades, Universitat Rovira i Virgili, 43007 Tarragona, Spain

2 Department of Biology, University of Turku, 20500 Turku, Finland

3 Department of Biochemistry and Biotechnology, Rovira i Virgili University, 43007 Tarragona, Spain

4 Nutrition and Health Unit, Eurecat Technology Centre of Catalonia, 43204 Reus, Spain

5 EURECAT Centre Tecnològic de Catalunya, Centre for Omic Sciences (COS), Joint Unit Universitat Rovira i Virgili-EURECAT, Unique Scientific and Technical Infrastructures (ICTS), 43204 Reus, Spain

* Author to whom correspondence should be addressed. (santi.garcia-vallve@urv.cat)

ABSTRACT

The COVID-19 pandemic has caused more than 6.8 thousand million deaths worldwide, understanding better the society factors that increase the mortality rate could have reduced the number of fatalities. In this regard, we analyze retrospectively diverse aspects such as demographics, economic factors, healthcare access, behavioral health and nutrition related variables and its relationship with the number of COVID-19 deaths at a US county level. To find how they are linked, we used Automatic the Machine Learning (AutoML) frameworks MLJAR, TPOT, and TabPFN for predicting the number of COVID-19 fatalities registered until November 2022, using county data from 2019. After data cleansing we employed 50 variables to train the models. We obtained a Pearson correlation of 0.715 between the real and predicted number of COVID-19 deaths in the testing set. Through Shapley Additive exPlanation (SHAP) values we find as the most important predictors: the proportion of primary care physicians and providers, the median household income, physical inactivity, children in poverty, long commutes/ driving alone and diabetes. In addition, opposite to previous studies we do not find the proportion of African-Americans as a highly predictive factor. We expect that the provided results will contribute to policy makers by providing insights that were overseen during the pandemic and that could be used for a better pandemic preparedness and management strategies in the future.
Keywords: COVID-19 pandemic, Mortality rate, AutoML, Machine learning, pandemic preparedness.

I. INTRODUCTION

COVID-19, the respiratory disease caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) [1], has caused more than 6.8 thousand million deaths and infected 676.6 thousand individuals around the world, as reported by Johns Hopkins University and the World Health Organization in September 2023 [2,3]. Since its emergence in China in 2019 [4], it has affected all aspects of society [5], global economy [6], mental health [7,8], education [9] and the management of other diseases [10–14]. Despite reaching all the globe, the outcome of the infection, cases and vaccination coverage has varied geographically [15–17], reflecting socio-demographic inequalities and disparities [18–21]. Several studies have aimed at uncovering these socioeconomic factors that influence the incidence of COVID-19 hospitalizations [22,23], death rates [24–29] and cases [26,28,30]. In order to identify the main drivers of the named outcomes at a county level in the United States, these studies have considered factors such as comorbidity (e.g. hypertension, diabetes, heart, kidney and respiratory diseases) [24,25,27,29–34], age [24,25,28–34], ethnicity [24,25,27–29,31,32,34], education level [25,31], household income [22,24,27,30,31], air pollution [24,34] and other variables such as insurance coverage and incarceration rates [24]. Some other studies have included unconventional sources and information, taking it from Facebook and using political party preferences [24] and gender-related factors [33]. Many of these studies coincide in the unavoidable role of demographics in shaping COVID-19 outcomes. Thus, the proportion of African-Americans in a US county's population is a major predictor of both COVID-19 cases and deaths [24,25,27,28,31,32,34]. Nonetheless, the importance

of other variables varies between studies. For instance, during the pandemic it was found that lower education levels, a greater percentage of black residents and a higher population older than 65 years had the greatest association with COVID-19 cases [31]. In terms of socioeconomic evaluation, the proportion of black residents and household income were considered important predictive factors [31,32]. In another study [34], a vulnerability index to categorize each county based on their susceptibility to COVID-19 was introduced. They found population size, proportion of black residents, air quality indicator (fine particulate matter less than or equal to 2.5 micrometers in diameter PM2.5), insurance percentage coverage and proportion of hispanic residents were the most important predictive factors [34]. Furthermore, despite obesity being accounted for more COVID-19 hospitalizations [17] and a higher number of obesity hospitalizations showed more COVID-19 risk [16], it was not found as a predictive factor [24,27]. Therefore, there is no concluding linkage between obesity and COVID-19 outcomes and it requires further study [31].

Most of the studies mentioned above commonly used data from the US Census [24,25,27–29,31,32,34] and the Centers for Disease Control and Prevention (CDC) [24,25,27,29,31,32,34]. Meanwhile, they have used regression analyses with statistical and machine learning models such as Ensemble models [27], Neural Networks [30], Bayesian models [29], transformers (Neural Networks) [22], and other custom algorithms [23,26] to identify the most important variables.

This aforementioned research corpus, highlights the diversity of factors that are linked to COVID-19 and how they influence the outcomes. In this study, we leverage on Automated Machine Learning (AutoML) models [35] to analyze these underlying socio-economical factors influencing COVID-19 deaths in the United States. By taking advantage of AutoML models, we harness their robustness to human error, technical power and ease to use [36]. In addition, by doing it retrospectively, we examine a more abundant coverage of data regarding COVID-19 deaths, from January 2020 to November 2022. This enable us to discard misconceptions originated during the pandemic, such as the importance given to the proportion of black residents in a county and to unveil factors that were overlooked when data was scarce, like the ratio of population to primary care physicians/providers in a county. We expect to contribute to a better understanding of these factors and, as a result, to be better prepared to fight future pandemics.

II. RESULTS AND DISCUSSION

Among the features of our dataset we had county health data from 2019 [37] and hypertension-related cardiovascular disease death rates from 2019 [38], making together 102 features (**Figure 1**).

Our target variable corresponds to the number of COVID-19 Deaths by County, reported until November 13th 2022 [39], normalized by county population. In order to reduce misinterpretation while analyzing the most important variables that the machine learning model captures, we excluded those variables that have a correlation (absolute value) lower than 0.15, against the target. In addition, to avoid multicolinearity we removed those variables with a correlation (absolute value) higher than 0.8 against other variables. We prioritized those variables with higher correlation against the target. After this filtering process we ended up with 50 variables. We have grouped them from most connected to nutrition to least related, starting form the categories present at the

County-health Rankings Model [40]. These variables can be summarized as follows: **Nutrition:** Food environment index (Access to healthy foods), food insecurity (A household-level economic and social condition of limited access to adequate food [41]), % low birthweight, % adult obesity, % diabetic, number of hypertension-related deaths by sex and age. **Healthcare Access and Providers:** Primary care physicians ratio (Population to Primary Care Physicians ratio), primary care physicians rate (Primary Care Physicians per 100,000 population), other primary care providers ratio (primary care providers other than physicians, such as nurse practitioners, physician assistants and clinical nurse specialists [42]), other primary care providers rate, % uninsured children, % uninsured adults. **Behavioral Health:** % Frequent mental distress, % insufficient sleep, % adult smokers, % Alcohol-impaired driving deaths, % excessive drinking. **Physical Health:** % Physically Inactive. **Community and Environment:** % Driving alone to work, % long commute – driving alone, residential segregation – non-white/ white, % with access to exercise opportunities, % flu vaccinated. **Social and Safety Indicators:** Social associations rate (Number of membership associations per 10,000 population [43]), violent crime rate, teen births rate, preventable hospital stays rate, motor vehicle crash deaths, injury deaths rate, age-adjusted mortality, life expectancy. **Economic Factors:** Median household income, % unemployed, income inequality ratio (Household income of the 80th percentile to income at the 20th percentile [44]), % homeowners. **Demographics:** % African American, % female, % 65 years old and over, % native Hawaiian/Other Pacific Islander, % Asian, % children in poverty, % not proficient in English, % single-parent households. From this list, a percentage % indicates the percentage of the county population. In **Figure 1** we can see the variables that correlate the most against COVID-19 deaths by county's population (target). At the top left is the median household income, which has the highest magnitude of negative correlation against the target (-0.444). Meanwhile, at the top right, premature age-adjusted mortality hast the highest positive correlation (0.428). These two corresponding to the variables' groups **Economic Factors** and **Social and Safety Indicators**. From the group **Healthcare Access and Providers,** the variable Other primary care providers rate, has the highest correlation (0.410). One position below (0.407) is the percentage of people who are physically inactive from the **Physical Health** group. Immediately next, it is the percentage of children in poverty from the **Demographics** group (0.404). This last one followed by the percentage of people with diabetes (prevalence), with the highest correlation (0.380) among the **Nutrition variables**. With a negative correlation (-0.367), is the percentage of people that takes a long commute and drive alone with the highest magnitude among the Community and Environment variables. Finally, the variable with the highest magnitude of correlation (0.349) from **Behavioral Health** variables is the percentage of people with frequent mental distress (The value of the correlations is in **Table S1). In Figure 1**, the hypertension (ht) variable that ends with Overall Overall means that accounts for all ethnicities and genders; MV Mortality Rate, stands for the mortality rate caused by Motor Vehicles; % Limited Access, stands for the percentage of a county's population with limited access to healthy foods. All variables used, kept and their correlations are present in **Table S1**.

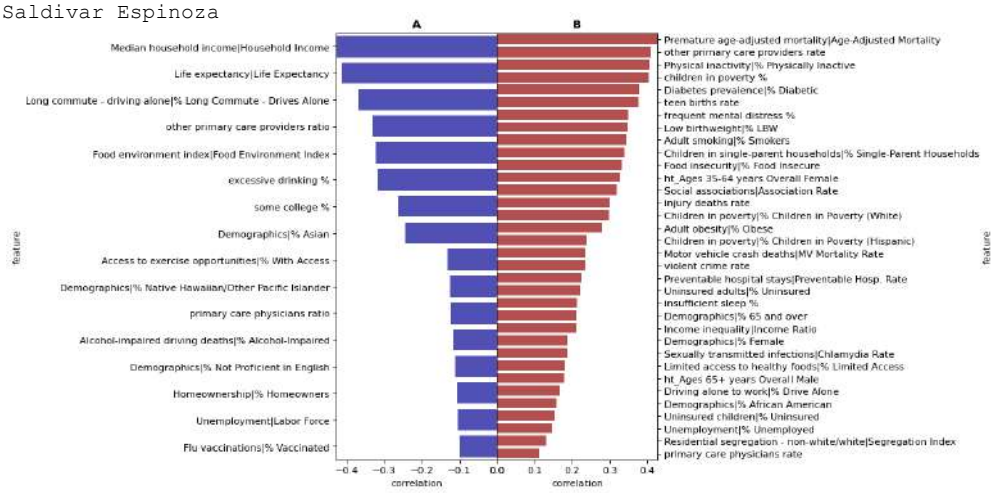


Figure 1. Correlation of all variables against the number of COVID-19 deaths by county's population (target). Only correlations that are statistically significant are shown. On the right (**B**) in red variables that correlate positively with the number of deaths and on the left (**A**) in blue, variables that correlate negatively with it. A vertical line | is placed before some variables' names to indicate the category where they were originally grouped in.

II.1 Prediction of COVID-19 deaths

After examining the correlations between our dataset variables and the COVID-19 deaths per population in US counties, we take a close look at predictive modeling. While correlations provide valuable insights into the relationships among variables, they offer a limited perspective on the predictive capacity of our dataset. The next logical step in our analysis is to explore predictive models, aiming to discern patterns and dependencies that may not be immediately apparent through correlation alone. By employing predictive models, we seek to answer a critical question: Can the COVID-19 deaths per population be reliably predicted based on the information gleaned from the correlated variables? This shift in focus from correlation to prediction promises a richer understanding of our dataset, unraveling nuanced relationships and unveiling hidden dynamics that may elude traditional correlation analysis. In our predictive models, we split the data into a training and testing set. We used the training set to train and a part of it, the validation set, was used by the AutoML frameworks to fine tune their configuration. At the end we evaluated the performance, predictions and feature importance on the testing set, which was not used during training. We tried the AutoML methods TabPFN, TPOT and MLJAR (using explain, compete and optuna modes) obtaining the performance shown in **Table 1**. In this table we show the comparison between the real and predicted values using the metrics: Pearson correlation and its p-value, R2, Mean Absolute Error (MAE) and Mean Squared Error (MSE) [45]. Despite having good values for the coefficient of determination R2 in the training set, some models have a negative value in the testing

set. This indicates that regardless of a high correlation they are not a good fit to the testing data and perform worse than the average line [46]. We also have included if the method allows a straightforward way to see the importance of the used variables. MLJAR in explain mode is the only one from the group of AutoML frameworks that provides an easy way to understand directly the features' importance. Nonetheless, as a result of using TPOT, that searched among different machine learning models, configurations and combinations, the pipeline that provided the best result, lowest MSE, was XGBoost [47] . Since the best configuration found was purely XGBoost and the performance is equivalent to the other best performing methods, we picked XGBoost for its simplicity. With XGBoost we obtained a correlation of 0.715 in the testing set between the real number of COVID-19 caused deaths by county population against the predicted values, as shown in **Figure 2**. TabPFN did not have a better prediction power despite using transformers [48], a modern machine learning architecture with great success in Natural Language Processing and Computer Vision [49,50].

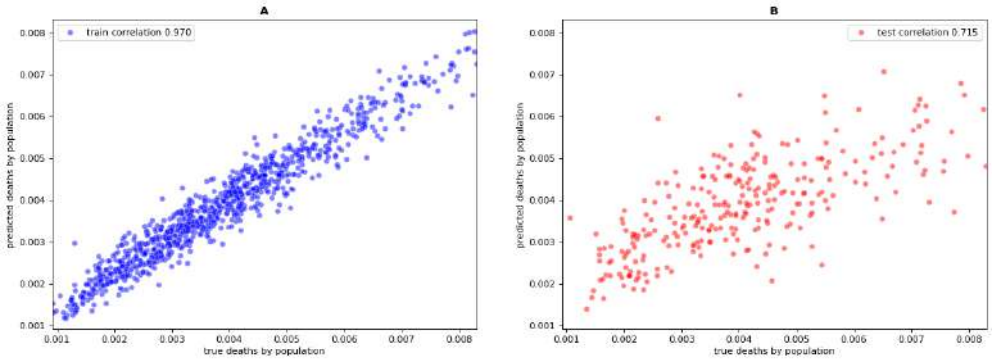


Figure 2. Prediction of XGBoost. Panel **A** shows the prediction on the training set and on panel **B** prediction on the testing set. The horizontal axis corresponds to the real number of COVID-19 caused deaths by county population and the vertical axis to the predicted values.

Table 1. Regression metrics of the used methods and model. The results are sorted from top to bottom based on the method with the highest correlation in the testing set. Natively explainable indicates if the method has a direct way to see how important is each feature without additional processing.

Train							
Correlation	R^2	MAE	MSE	Correlation pvalue	Correlation p<0.05	Method	Natively explainable
0.95	0.81	4.2E-04	3.1E-07	0.0E+00	True	MLJAR Optuna	False
0.970	0.930	3.2E-04	1.7E-07	0.0E+00	True	XGBoost	True
0.970	0.930	3.2E-04	1.7E-07	0.0E+00	True	TPOT	False
0.905	0.798	5.0E-04	5.0E-07	0.0E+00	True	MLJAR explain	True
0.959	0.911	3.2E-04	2.2E-07	0.0E+00	True	MLJAR compete	False
0.941	0.881	1.5E-01	5.6E-01	0.0E+00	True	TabPFN	False

Test							
correlation	R^2	MAE	MSE	correlation pvalue	Correlation p<0.05	method	natively explainable
0.730	-0.20	8.7E-04	1.3E-06	1.3E-47	True	MLJAR Optuna	False
0.715	**0.501**	**8.8E-04**	**1.3E-06**	**5.2E-45**	**True**	**XGBoost**	**True**
0.715	0.501	8.8E-04	1.3E-06	5.2E-45	True	TPOT	False
0.718	0.497	9.0E-04	1.3E-06	1.7E-45	True	MLJAR explain	True
0.700	0.480	8.9E-04	1.4E-06	2.3E-42	True	MLJAR compete	False
0.357	-0.445	1.8E+00	5.5E+00	8.5E-10	True	TabPFN	False

II.2 Feature importance

XGBoost allows to obtain, in a straight forward manner, the importance of the used variables through Shapley Additive exPlanation (SHAP) [51]. We gathered the SHAP values and displayed them in **Figure 3**, where they are listed from most important variable at the top, to least important at the bottom. In addition we applied a Sensitivity Analysis [52,53] to know by how much the number of COVID-19 deaths will change based on changes to values of the variables. These values are present in **Table S3** (More detail in **Materials and methods**. All the variables with their respective feature importance are present in **Table S2**).

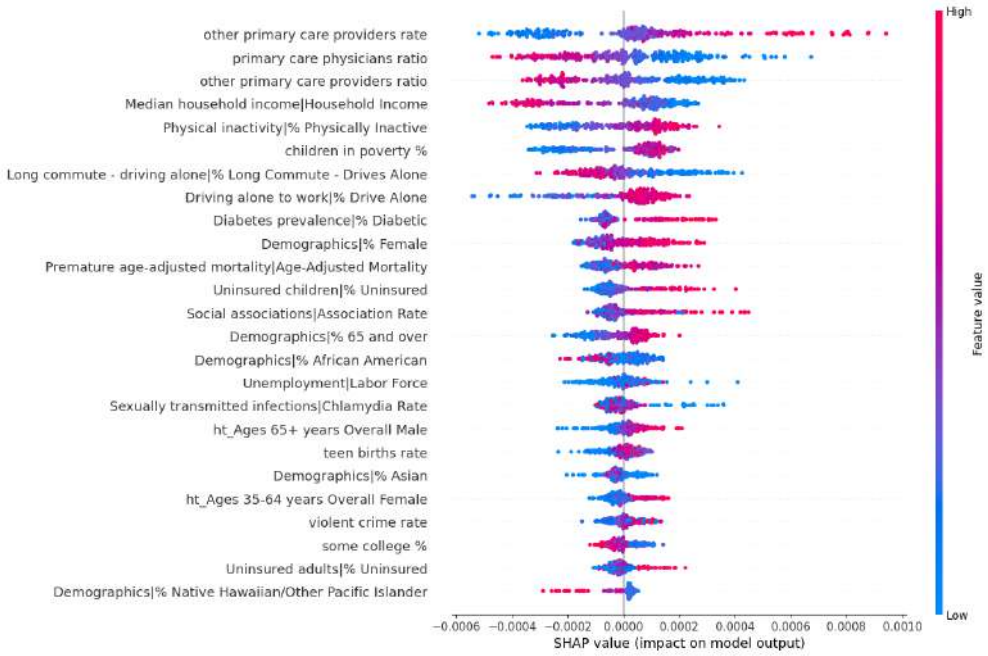


Figure 3. SHAP summary plot of the 25 most important variables sorted from most to least important, top to bottom. If the colors go from blue to red, left to right, it indicates that if the variable increases its value, also the amount of deaths by COVID-19.

II.2.1 Most important variables

In contrast with previous studies [24,25,27,28,31,32,34] which found the proportion of African-American in a county as one of the most important variables to predcit COVID-19 deaths, this variable was accounted as the 15th most important factor in our results, showing a lower contribution to the number of COVID-19 deaths in a county **(Figure 3)**. We identified the proportion of primary care physicians and other primary care providers within a county as a crucial set of variables, as shown in **Figure 3**. These variables have been previously associated with population mortality in the United States [54]. Notably, an increase in the number of primary care providers historically correlated with improved life expectancy between 2005 and 2015 [54]. However, contrary to our expectations, our analysis revealed an inverse correlation (**Figure 3)**. Surprisingly, an elevated rate of other primary care providers (Other primary care providers per 100,000 population) was linked to an increase in deaths caused by COVID-19. Simultaneously, a reduction in the ratio of primary care physicians and providers (Population to Primary Care Physicians ratio) was associated with a rise in the predicted number of COVID-19 deaths, given that the ratio is the inverse of the rate. Importantly, it is worth noting that this observed relationship is not an artifact of the predictive method. The same inverse correlation is evident in the correlation of these variables with the COVID-19 death variable by county, as illustrated in **Figure 1**. This consistency across analyses reinforces the unexpected nature of the findings and suggests a robust association between the primary care provider variables and COVID-19 mortality. It is important to note that while these correlations are observed, they do not imply causation. One possible explanation may lie in the regional variation of healthcare accessibility. In some scenarios, a higher ratio of other primary care providers may be associated with lower population density, particularly in rural areas. Conversely, densely populated urban areas may exhibit a lower ratio due to a higher concentration of people and potentially more healthcare facilities. Notably, our study lacks variables directly related to population density, which could play a significant role in virus transmission and mortality. Further investigation is warranted to fully understand the intricate factors influencing these unexpected correlations. After the amount of primary care providers and physicians, we found that the median household income posses a high predictive power. This magnitude of importance was also found in other studies [31,32]. A decrease in the median household income increases the predicted number of deaths by COVID-19 (**Figure 3**). A possible explanation for this is that high income counties will have more possibilities to have remote jobs [24], therefore less transmission. Nonetheless, lower median household income could also represent lower medical services taken. High median household income values, specially when there is a high income inequalities, have been associated with a higher COVID-19 incidence [30], opposite to our findings. We did not include the variable “Income inequality Ratio” for training our models, however, we show in **Figure 4A**, that without including a predictive model, there is a negative correlation between the Income inequality Ratio (Higher values, more inequalities) and the median household income (-0.33). However, it is seen that higher numbers of COVID-19 deaths are present where the income inequality index (Higher values more inequality) is between 4 and 5 and the median household income is between 40,000 and 60,000. Meanwhile, in other research [31], tthey also find that high income values are associated with a higher COVID-19 incidence. Nevertheless, they specify that this happens when a county has high median income and low education levels. Despite of this report, we observe in **Figure 4B** that the percentage of people

with some college education correlates positively with the median household income (0.65) and both correlate negatively with the number of COVID-19 deaths. Therefore, this scenario of low education and high household income is unlikely. Among other variables related to education, the percentage of the county population that has some college education has the highest magnitude of correlation (-0.26) against the number of COVID-19 deaths.

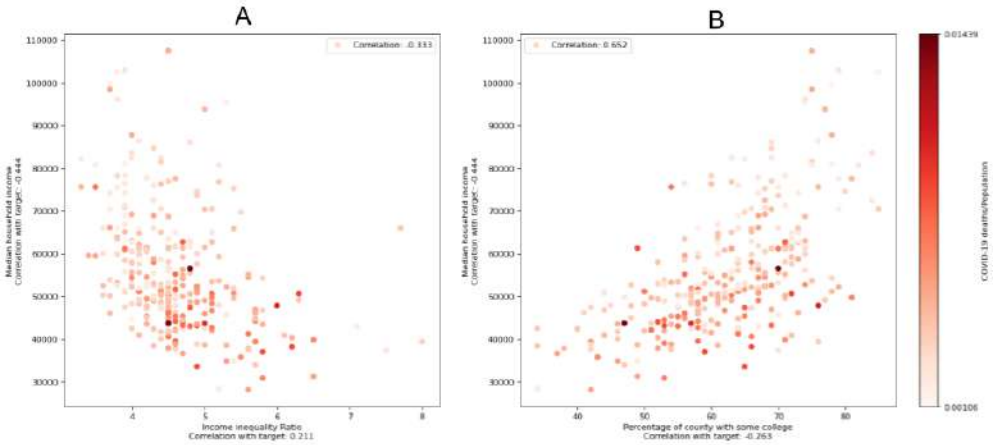


Figure 4. Relationships between median household income, COVID-19 deaths/population, income inequality ratio and percentage of county with some college. Darker colors mean more COVID-19 deaths. The correlation shown on top is between the x and y variables. Target stands for number of COVID-19 deaths/population.

II.2.2 Nutritional variables

The nutritional variables present in the predictive model are: the percentage of people with diabetes, number of deaths caused by hypertension in men over 65 and women between 35 to 64, index of factors that contribute to a healthy food environment (food environment index) [55], percentage of adults with obesity, percentage of population who lack adequate access to food (% food insecurity) [56], percentage of population who are low-income and do not live close to a grocery store (% Limited access) [57] and percentage of live births with low birthweight (< 2,500 grams) [58]. The importance of these listed variables, in respective order are 3%, 1.7%, 1.4%, 1.1%, 1.1%, 0.8%, 0.6% and 0.2% as can be seen in **Table S2**. To put these values in the right scale for comparison, the rate of other primary care providers, the most important variable, has 8.8% of importance, as attributed by the model. The top 10 variables, which include the percentage of people with diabetes (in the 9th position), sum up to 52% of all the importance captured by the predictive model (**Table S2)**. However, the number

of deaths caused by hypertension in men and women, the following most important nutrition related variables, are in the 18th and 21th position of importance. An increase of any of these three variables also increases also the number of COVID-19 deaths (**Figure 3**). The same behavior is shown in **Figure 1** since all of them correlate positively with COVID-19 mortality. In previous studies [24,25,31,33], diabetes and cardiovascular diseases were also found as predictive variables for COVID-19 deaths. The index of factors that contribute to a healthy food environment (food environment index, higher better environment) correlates negatively with the percentage of people with diabetes and deaths caused by hypertension (**Figure S1**). Meanwhile, the limited access to healthy foods, adult obesity and food insecurity correlate positively with them (**Figure S1**). These factors have been reported as connected to diabetes and heart disease from a dietary point of view [59].

III. CONCLUSIONS

In this study, we aimed to predict the number of COVID-19 deaths by county, normalized by population, using a dataset with 102 features, including county health data, socioeconomic data and hypertension-related cardiovascular disease death rates from 2019. The dataset was processed by filtering out variables with low correlation to the target and addressing multicollinearity issues, resulting in a reduced set of 50 variables. We employed three AutoML methods (TabPFN, TPOT, and MLJAR) to develop predictive models and found that XGBoost (result of TPOT), performed the best, achieving a correlation of 0.715 on the testing set between predicted and actual COVID-19 deaths normalized by population size. Through the usage Shapley Additive exPlanation (SHAP) values in the predictive model we found that the most leading important factor was the ratio of primary care physicians and providers, contradicting previous studies that emphasized the proportion of African-Americans as a crucial factor. Additionally, the median household income emerged as a significant predictor, with lower income levels linked to higher predicted COVID-19 deaths. However, it was previously reported that in counties with high median household and high income inequality there were more COVID-19 fatalities. This was not ratified. In addition, it was also previously suggested that high median household income values were associated with a higher COVID-19 incidence when a county had low education levels. We showed that this scenario is unlikely, since education and median household income levels correlated positively and the number of COVID-19 deaths correlated negatively with both. We also highlighted the relevance of nutritional factors, including diabetes prevalence and deaths caused by hypertension, in predicting COVID-19 deaths. We showed that as the amount of people with diabetes and hypertension related deaths increase, also COVID-19 fatalities. Overall, the findings underscore the complex nature of the determinants of COVID-19 outcomes and emphasize the need for a comprehensive understanding of healthcare, socioeconomic, and nutritional factors in predicting and managing pandemic outcomes at the county level.

IV. MATERIALS AND METHODS

IV.1 Primary sources of data acquisition

We used US county health data [37] and hypertension related cardiovascular disease death rates [38] from 2019. The US county health data was obtained from the University of Wisconsin, that joined data from different sources, such as Census Population Estimates [60], USDA Food Environment Atlas [61], National Center For Health Statistics [62], Behavioral Risk Factor Surveillance System [63], Stanford Education data Archive [64] and others. Meanwhile the hypertension related cardiovascular disease death rates were obtained from the Centers for Disease Control And Prevention (CDC) [38]. The number of counties in these data sources were 3142, corresponding to 50 US states plus the federal District of Columbia.

IV.2 Data cleaning and normalization

In the data pre-processing step, we defined the number of deaths caused by COVID-19 at a US county level as our target variable. This data was obtained from the CDC [39]. Our dataset had originally 269 variables, including confidence intervals and quartiles for each variable. After removing the statistical information and those variables with more than 20% missing values we obtained 120 variables. Consequently we also removed variables that were expressed in a quantity when there was also a percentage of the population, keeping the percentage format. This resulting in 103 variables. Afterwards, we used Scikit-learn [65] for multivariate imputation [66,67] to address missing values [68]. To avoid the confounding effect of the number of inhabitants on the results, we divided the target variable by the county population. Without this approach, any trained model's predictions had a correlation larger than 0.9, but this was attributable to the population component. With this normalization we tried to find more meaningful connections with the other county variables. Therefore, we also removed the population variable, obtaining 102 variables. We also removed highly correlated (>0.8) variables to mitigate multicollinearity. Additionally, we removed variables with a low correlation (<0.15) to the target variable to avoid inconsistent interpretations (positive or negative sign) with the models to be trained. Finally ending with 50 variables in total. Then we performed the removal of outliers, using the IQR technique [69] applied on the target variable. This means that we removed those counties with COVID-19 death counts considered outliers by the IQR method. From 1168 US counties, we removed 51 considered. The IQR method considers outliers to those values outside the range between **Q1-IQR** and **Q3+IQR** [where Q1 and Q3 are the values of the first and third quartile, while IQR corresponds to **1.5x(Q3-Q1)**].

IV.3 Automated Machine Learning (AutoML) models generation

For the machine learning training we used three AutoML python packages to find the best model, MLJar [70], TPOT [71] and TabPFN [72]. The last one only supports classification tasks, so it was required to discretize the target into 10 intervals/bins.
The data was divided into a training and testing set, allocating randomly 20% for the testing set and the remaining 80% for the training set. Afterwards, for each machine learning model, the training set was split into training and validation, following a n-fold cross validation process. The MLJar [70] method used a 10-fold cross validation automatically, for TPOT we used 5-fold cross validation and for TabPFN there was no option for specifying it since it does not use any cross validation [72]. We trained MLJAR in the explain, compete and optuna modes to obtain different levels of performance. The configuration for TPOT was set to work for 3 generations and a population size of 50, leaving all other parameters as default. We also trained TabPFN with the default parameters shown in their github respository.

IV.4 Feature Importance

To obtain the importance given to each variable we obtained the Shapley Additive exPlanation (SHAP) [51] values from the resulting XGBoost [47] model (product of TPOT). We also applied Sensitivity Analysis [52,53] on XGBoost to see how the predicted values change when a variable changes its value. This is present in **Table S3.**

REFERENCES

1. Dirlikov, E.; Zhou, S.; Han, L.; Li, Z.; Hao, L.; Millman, A.; Marston, B. Use of Public Data to Describe COVID-19 Contact Tracing in Hubei Province and Non-Hubei Provinces in China between 20 January and 29 February 2020. West. Pac. Surveill. Response 2021, 12, 6–6, doi:10.5365/wpsar.2021.12.3.808.
2. COVID-19 Map Available online: https://coronavirus.jhu.edu/map.html (accessed on 27 September 2023).
3. WHO Coronavirus (COVID-19) Dashboard Available online: https://covid19.who.int (accessed on 27 September 2023).
4. Wu, F.; Zhao, S.; Yu, B.; Chen, Y.-M.; Wang, W.; Song, Z.-G.; Hu, Y.; Tao, Z.-W.; Tian, J.-H.; Pei, Y.-Y.; et al. A New Coronavirus Associated with Human Respiratory Disease in China. Nature 2020, 579, 265–269, doi:10.1038/s41586-020-2008-3.
5. Ali, I.; Alharbi, O.M.L. COVID-19: Disease, Management, Treatment, and Social Impact. Sci. Total Environ. 2020, 728, 138861, doi:10.1016/j.scitotenv.2020.138861.
6. Naseer, S.; Khalid, S.; Parveen, S.; Abbass, K.; Song, H.; Achim, M.V. COVID-19 Outbreak: Impact on Global Economy. Front. Public Health 2023, 10.
7. Zhu, C.; Zhang, T.; Li, Q.; Chen, X.; Wang, K. Depression and Anxiety During the COVID-19 Pandemic: Epidemiology, Mechanism, and Treatment. Neurosci. Bull. 2023, 39, 675–684, doi:10.1007/s12264-022-00970-2.
8. Li, F. Impact of COVID-19 on the Lives and Mental Health of Children and Adolescents. Front. Public Health 2022, 10.
9. Bughrara, M.S.; Swanberg, S.M.; Lucia, V.C.; Schmitz, K.; Jung, D.; Wunderlich-Barillas, T. Beyond COVID-19: The Impact of Recent Pandemics on Medical Students and Their Education: A Scoping Review. Med. Educ. Online 2023, 28, 2139657, doi:10.1080/10872981.2022.2139657.
10. Cohen-Mansfield, J. The Impact of COVID-19 on Long-Term Care Facilities and Their Staff in Israel: Results from a Mixed Methods Study. J. Nurs. Manag. 2022, 30, 2470–2478, doi:10.1111/jonm.13667.
11. Geetha, D.; Kronbichler, A.; Rutter, M.; Bajpai, D.; Menez, S.; Weissenbacher, A.; Anand, S.; Lin, E.; Carlson, N.; Sozio, S.; et al. Impact of the COVID-19 Pandemic on the Kidney Community: Lessons Learned and Future Directions. Nat. Rev. Nephrol. 2022, 18, 724–737, doi:10.1038/s41581-022-00618-4.
12. Rizvi, A.A.; Kathuria, A.; Al Mahmeed, W.; Al-Rasadi, K.; Al-Alawi, K.; Banach, M.; Banerjee, Y.; Ceriello, A.; Cesur, M.; Cosentino, F.; et al. Post-COVID Syndrome, Inflammation, and Diabetes. J. Diabetes Complications 2022, 36, 108336, doi:10.1016/j.jdiacomp.2022.108336.
13. Ata, B.; Vermeulen, N.; Mocanu, E.; Gianaroli, L.; Lundin, K.; Rautakallio-Hokkanen, S.; Tapanainen, J.S.; Veiga, A. SARS-CoV-2, Fertility and Assisted Reproduction. Hum. Reprod. Update 2023, 29, 177–196, doi:10.1093/humupd/dmac037.
14. Gao, X.; Lv, F.; He, X.; Zhao, Y.; Liu, Y.; Zu, J.; Henry, L.; Wang, J.; Yeo, Y.H.; Ji, F.; et al. Impact of the COVID-19 Pandemic on Liver Disease-Related Mortality Rates in the United States. J. Hepatol. 2023, 78, 16–27, doi:10.1016/j.jhep.2022.07.028.
15. Khan, M.M.; Odoi, A.; Odoi, E.W. Geographic Disparities in COVID-19 Testing and Outcomes in Florida. BMC Public Health 2023, 23, 79, doi:10.1186/s12889-022-14450-9.
16. Das, P.; Igoe, M.; Lenhart, S.; Luong, L.; Lanzas, C.; Lloyd, A.L.; Odoi, A. Geographic Disparities and Determinants of COVID-19 Incidence Risk in the Greater St. Louis Area, Missouri (United States). PLOS ONE 2022, 17, e0274899, doi:10.1371/journal.pone.0274899.
17. Abbasi, B.A.; Chanana, N.; Palmo, T.; Pasha, Q. Disparities in COVID-19 Incidence and Fatality Rates at High-Altitude. PeerJ 2023, 11, e14473, doi:10.7717/peerj.14473.

18. Vicetti Miguel, C.P.; Dasgupta-Tsinikas, S.; Lamb, G.S.; Olarte, L.; Santos, R.P. Race, Ethnicity, and Health Disparities in US Children With COVID-19: A Review of the Evidence and Recommendations for the Future. J. Pediatr. Infect. Dis. Soc. 2022, 11, S132–S140, doi:10.1093/jpids/piac099.

19. Siegel, M.; Critchfield-Jain, I.; Boykin, M.; Owens, A.; Muratore, R.; Nunn, T.; Oh, J. Racial/Ethnic Disparities in State-Level COVID-19 Vaccination Rates and Their Association with Structural Racism. J. Racial Ethn. Health Disparities 2022, 9, 2361–2374, doi:10.1007/s40615-021-01173-7.

20. Guay, M.; Maquiling, A.; Chen, R.; Lavergne, V.; Baysac, D.-J.; Kokaua, J.; Dufour, C.; Dubé, E.; MacDonald, S.E.; Gilbert, N.L. Sociodemographic Disparities in COVID-19 Vaccine Uptake and Vaccination Intent in Canada. Health Rep. 2022, 33, 37–54, doi:10.25318/82-003-x202201200004-eng.

21. Zhao, M.; Hamadi, H.Y.; Haley, D.R.; Xu, J.; Tafili, A.; Spaulding, A.C. COVID-19 Deaths and the Impact of Health Disparities, Hospital Characteristics, Community, Social Distancing, and Health System Competition. Popul. Health Manag. 2022, 25, 807–813, doi:10.1089/pop.2022.0144.

22. Er, S.; Yang, S.; Zhao, T. COUnty aggRegation Mixup AuGmEntation (COURAGE) COVID-19 Prediction. Sci. Rep. 2021, 11, 14262, doi:10.1038/s41598-021-93545-6.

23. Olshen, A.B.; Garcia, A.; Kapphahn, K.I.; Weng, Y.; Vargo, J.; Pugliese, J.A.; Crow, D.; Wesson, P.D.; Rutherford, G.W.; Gonen, M.; et al. COVIDNearTerm: A Simple Method to Forecast COVID-19 Hospitalizations. J. Clin. Transl. Sci. 2022, 6, e59, doi:10.1017/cts.2022.389.

24. Ruck, D.J.; Bentley, R.A.; Borycz, J. Early Warning of Vulnerable Counties in a Pandemic Using Socio-Economic Variables. Econ. Hum. Biol. 2021, 41, 100988, doi:10.1016/j.ehb.2021.100988.

25. Boserup, B.; McKenney, M.; Elkbuli, A. Disproportionate Impact of COVID-19 Pandemic on Racial and Ethnic Minorities. Am. Surg. 2020, 86, 1615–1622, doi:10.1177/0003134820973356.

26. Menda, K.; Laird, L.; Kochenderfer, M.J.; Caceres, R.S. Explaining COVID-19 Outbreaks with Reactive SEIRD Models. Sci. Rep. 2021, 11, 17905, doi:10.1038/s41598-021-97260-0.

27. McCoy, D.; Mgbara, W.; Horvitz, N.; Getz, W.M.; Hubbard, A. Ensemble Machine Learning of Factors Influencing COVID-19 across US Counties. Sci. Rep. 2021, 11, 11777, doi:10.1038/s41598-021-90827-x.

28. Clouston, S.A.P.; Natale, G.; Link, B.G. Socioeconomic Inequalities in the Spread of Coronavirus-19 in the United States: A Examination of the Emergence of Social Inequalities. Soc. Sci. Med. 2021, 268, 113554, doi:10.1016/j.socscimed.2020.113554.

29. Formanack, A.; Doshi, A.; Valdez, R.; Williams, I.; Moorman, J.R.; Chernyavskiy, P. Race, Class, and Place Modify Mortality Rates for the Leading Causes of Death in the United States, 1999–2021. J. Gen. Intern. Med. 2023, doi:10.1007/s11606-023-08062-1.

30. Mollalo, A.; Rivera, K.M.; Vahedi, B. Artificial Neural Network Modeling of Novel Coronavirus (COVID-19) Incidence Rates across the Continental United States. Int. J. Environ. Res. Public. Health 2020, 17, 4204, doi:10.3390/ijerph17124204.

31. Hawkins, R.B.; Charles, E.J.; Mehaffey, J.H. Socio-Economic Status and COVID-19–Related Cases and Fatalities. Public Health 2020, 189, 129–134, doi:10.1016/j.puhe.2020.09.016.

32. Tan, S.B.; deSouza, P.; Raifman, M. Structural Racism and COVID-19 in the USA: A County-Level Empirical Analysis. J. Racial Ethn. Health Disparities 2022, 9, 236–246, doi:10.1007/s40615-020-00948-8.

33. Zheng, Z.; Peng, F.; Xu, B.; Zhao, J.; Liu, H.; Peng, J.; Li, Q.; Jiang, C.; Zhou, Y.; Liu, S.; et al. Risk Factors of Critical & Mortal COVID-19 Cases: A Systematic Literature Review and Meta-Analysis. J. Infect. 2020, 81, e16–e25, doi:10.1016/j.jinf.2020.04.021.

34. Marvel, S.W.; House, J.S.; Wheeler, M.; Song, K.; Zhou, Y.; Wright, F.A.; Chiu, W.A.; Rusyn, I.; Motsinger-Reif, A.; Reif, D.M. The COVID-19 Pandemic Vulnerability Index (PVI) Dashboard: Monitoring County-Level Vulnerability Using Visualization, Statistical Modeling, and Machine Learning 2020, 2020.08.10.20169649.

35. Conrad, T.; Matzel, M.; Schwarzenberger, M.; Wiemer, H.; Ihlenfeldt, S. Benchmarking AutoML for Regression Tasks on Small Tabular Data in Materials Design. Sci. Rep. 2022, 12, 19350, doi:10.1038/s41598-022-23327-1.

36. Scriven, A.; Kedziora, D.J.; Musial, K.; Gabrys, B. The Technological Emergence of AutoML: A Survey of Performant Software and Applications in the Context of Industry Available online: https://arxiv.org/abs/2211.04148v1 (accessed on 8 October 2023).

37. University of Wisconsin Population Health Institute, C.H.R.& R. | National Data & Documentation: 2010-2020 Available online: https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation/national-data-documentation-2010-2019 (accessed on 13 April 2023).

38. Centers for Disease Control And Prevention, (CDC) Rates and Trends in Hypertension-Related Cardiovascular Disease Mortality Among US Adults (35+) by County, Age Group, Race/Ethnicity, and Sex – 2000-2019 | Chronic Disease and Health Promotion Data & Indicators Available online: https://chronicdata.cdc.gov/Heart-Disease-Stroke-Prevention/Rates-and-Trends-in-Hypertension-related-Cardiovas/uc9k-vc2j (accessed on 13 April 2023).

39. Centers for Disease Control And Prevention, (CDC) Provisional COVID-19 Deaths by County, and Race and Hispanic Origin | Data | Centers for Disease Control and Prevention Available online: https://data.cdc.gov/NCHS/Provisional-COVID-19-Deaths-by-County-and-Race-and/k8wy-p9cg/ (accessed on 13 April 2023).

40. Health Factors | County Health Rankings & Roadmaps Available online: https://www.countyhealthrankings.org/explore-health-rankings/county-health-rankings-model/health-factors (accessed on 4 February 2024).

41. Carlson, S.J.; Andrews, M.S.; Bickel, G.W. Measuring Food Insecurity and Hunger in the United States: Development of a National Benchmark Measure and Prevalence Estimates. J. Nutr. 1999, 129, 510S-516S, doi:10.1093/jn/129.2.510S.

42. Other Primary Care Providers* | County Health Rankings & Roadmaps Available online: https://www.countyhealthrankings.org/explore-health-rankings/county-health-rankings-model/health-factors/clinical-care/access-to-care/other-primary-care-providers (accessed on 4 February 2024).

43. Social Associations | County Health Rankings & Roadmaps Available online: https://www.countyhealthrankings.org/explore-health-rankings/county-health-rankings-model/health-factors/social-economic-factors/family-and-social-support/social-associations (accessed on 4 February 2024).

44. Income Inequality | County Health Rankings & Roadmaps Available online: https://www.countyhealthrankings.org/explore-health-rankings/county-health-rankings-model/health-factors/social-economic-factors/income/income-inequality (accessed on 4 February 2024).

45. Plevris, V.P.; Solorzano, G.S.; Bakas, N.B.; Seghier, M.E.A.B.S. Investigation of Performance Metrics in Regression Analysis and Machine Learning-Based Prediction Models. ECCOMAS Congr. 2022 - 8th Eur. Congr. Comput. Methods Appl. Sci. Eng. 2022, Computational Solid Mechanics, doi:10.23967/eccomas.2022.155.

46. Chicco, D.; Warrens, M.J.; Jurman, G. The Coefficient of Determination R-Squared Is More Informative than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis Evaluation. PeerJ Comput. Sci. 2021, 7, e623, doi:10.7717/peerj-cs.623.

47. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Association for Computing Machinery: New York, NY, USA, August 13 2016; pp. 785–794.

48. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. ArXiv170603762 Cs 2017.

49. Lin, T.; Wang, Y.; Liu, X.; Qiu, X. A Survey of Transformers 2021.

50. Amatriain, X.; Sankar, A.; Bing, J.; Bodigutla, P.K.; Hazen, T.J.; Kazi, M. Transformer Models: An Introduction and Catalog 2023.

51. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the Advances in Neural Information Processing Systems; Curran Associates, Inc., 2017; Vol. 30.

52. Lek, S.; Delacoste, M.; Baran, P.; Dimopoulos, I.; Lauga, J.; Aulagnier, S. Application of Neural Networks to Modelling Nonlinear Relationships in Ecology. Ecol. Model. 1996, 90, 39–52, doi:10.1016/0304-3800(95)00142-5.

53. Pizarroso, J.; Portela, J.; Muñoz, A. NeuralSens: Sensitivity Analysis of Neural Networks. J. Stat. Softw. 2022, 102, 1–36, doi:10.18637/jss.v102.i07.

54. Basu, S.; Berkowitz, S.A.; Phillips, R.L.; Bitton, A.; Landon, B.E.; Phillips, R.S. Association of Primary Care Physician Supply With Population Mortality in the United States, 2005-2015. JAMA Intern. Med. 2019, 179, 506–514, doi:10.1001/jamainternmed.2018.7624.

55. Food Environment Index | County Health Rankings & Roadmaps Available online: https://www.countyhealthrankings.org/explore-health-rankings/county-health-rankings-model/health-factors/health-behaviors/diet-and-exercise/food-environment-index (accessed on 5 February 2024).

56. Food Insecurity | County Health Rankings & Roadmaps Available online: https://www.countyhealthrankings.org/explore-health-rankings/county-health-rankings-model/health-factors/health-behaviors/diet-and-exercise/food-insecurity (accessed on 5 February 2024).*

57. Limited Access to Healthy Foods | County Health Rankings & Roadmaps Available online: https://www.countyhealthrankings.org/explore-health-rankings/county-health-rankings-model/health-factors/health-behaviors/diet-and-exercise/limited-access-to-healthy-foods (accessed on 5 February 2024).*

58. Low Birthweight | County Health Rankings & Roadmaps Available online: https://www.countyhealthrankings.org/explore-health-rankings/county-health-rankings-model/health-outcomes/quality-of-life/low-birthweight (accessed on 5 February 2024).

59. Kalyani, R.R.; Everett, B.M.; Perreault, L.; Michos, E.D. Heart Disease and Diabetes. In Diabetes in America; Lawrence, J.M., Casagrande, S.S., Herman, W.H., Wexler, D.J., Cefalu, W.T., Eds.; National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK): Bethesda (MD), 2023.

60. Bureau, U.C. 2017 National Population Projections Tables: Main Series Available online: https://www.census.gov/data/tables/2017/demo/popproj/2017-summary-tables.html (accessed on 24 October 2023).

61. USDA ERS - Food Environment Atlas Available online: https://www.ers.usda.gov/data-products/food-environment-atlas/ (accessed on 24 October 2023).

62. CDC - NCHS - National Center for Health Statistics Available online: https://www.cdc.gov/nchs/index.htm (accessed on 24 October 2023).

63. CDC - BRFSS Available online: https://www.cdc.gov/brfss/index.html (accessed on 24 October 2023).

64. University, © Stanford; Stanford; California 94305 Stanford Education Data Archive (SEDA) Available online: https://exhibits.stanford.edu/data/catalog/db586ns4974 (accessed on 24 October 2023).

65. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. J. Mach. Learn. Res. 2011, 12, 2825–2830.

66. Buuren, S. van; Groothuis-Oudshoorn, K. Mice: Multivariate Imputation by Chained Equations in R. J. Stat. Softw. 2011, 45, 1–67, doi:10.18637/jss.v045.i03.

67. Buck, S.F. A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with an Electronic Computer. J. R. Stat. Soc. Ser. B Methodol. 1960, 22, 302–306.

68. Sklearn.Impute.IterativeImputer Available online: https://scikit-learn/stable/modules/generated/sklearn.impute.IterativeImputer.html (accessed on 20 March 2022).

69. Dash, Ch.S.K.; Behera, A.K.; Dehuri, S.; Ghosh, A. An Outliers Detection and Elimination Framework in Classification Task of Data Mining. Decis. Anal. J. 2023, 6, 100164, doi:10.1016/j.dajour.2023.100164.

70. Płońska, A.; Płoński, P. MLJAR: State-of-the-Art Automated Machine Learning Framework for Tabular Data. Version 0.10.3 2021.

71. Le, T.T.; Fu, W.; Moore, J.H. Scaling Tree-Based Automated Machine Learning to Biomedical Big Data with a Feature Set Selector. Bioinformatics 2020, 36, 250–256, doi:10.1093/bioinformatics/btz470.

72. Hollmann, N.; Müller, S.; Eggensperger, K.; Hutter, F. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second Available online: https://arxiv.org/abs/2207.01848v4 (accessed on 4 May 2023).

SUPPLEMENTARY MATERIAL

Table S1. Used variables sorted by feature importance

Variable	Feature importance (SHAP)		Correlation against	
	Normalized	Cumulative	COVID-19 deaths/population	COVID-19 deaths
other primary care providers rate	0.088	0.088	0.410	0.086
primary care physicians ratio	0.064	0.152	-0.122	-0.161
other primary care providers ratio	0.062	0.214	-0.332	-0.094
Median household income|HIousehold Income	0.061	0.274	-0.444	0.164
Physical inactivity|% Physically Inactive	0.050	0.324	0.407	-0.180
children in poverty %	0.048	0.372	0.404	-0.024
Long commute - driving alone|% Long Commute - Drives Alone	0.043	0.415	-0.367	0.183
Driving alone to work|% Drive Alone	0.041	0.456	0.167	-0.310
Diabetes prevalence| % Diabetic	0.033	0.489	0.380	-0.178
Demographics|% Female	0.031	0.520	0.187	0.149
Premature age-adjusted mortality| Age-Adjusted Mortality	0.028	0.548	0.428	-0.188
Uninsured children|% Uninsured	0.027	0.575	0.153	-0.027
Social associations| Association Rate	0.026	0.601	0.319	-0.292
Demographics|% 65 and over	0.026	0.627	0.212	-0.174
Demographics|% African American	0.023	0.649	0.157	0.135
Unemployment|Labor Force	0.019	0.669	-0.104	0.951
Sexually transmitted infections|Chlamydia Rate	0.018	0.687	0.187	0.211

ht_Ages 65+ years Overall Male	0.017	0.704	0.179	0.009
teen births rate	0.015	0.718	0.377	-0.118
Demographics|% Asian	0.014	0.732	-0.244	0.369
ht_Ages 35-64 years Overall Female	0.014	0.746	0.329	-0.016
violent crime rate	0.014	0.759	0.234	0.266
some college %	0.013	0.772	-0.263	0.170
Uninsured adults|% Uninsured	0.013	0.785	0.222	0.027
Demographics|% Native Hawaiian/Other Pacific Islander	0.013	0.797	-0.125	0.032
excessive drinking %	0.013	0.810	-0.319	0.063
Preventable hospital stays|Preventable Hosp. Rate	0.011	0.821	0.226	-0.042
Food environment index|Food Environment Index	0.011	0.833	-0.322	0.061
primary care physicians rate	0.011	0.844	0.113	0.199
Access to exercise opportunities|% With Access	0.011	0.855	-0.131	0.341
Adult obesity|% Obese	0.011	0.866	0.281	-0.284
Children in poverty|% Children in Poverty (Hispanic)	0.010	0.876	0.238	-0.041
Unemployment|% Unemployed	0.010	0.886	0.147	-0.031
Children in single-parent households|% Single-Parent Households	0.010	0.896	0.339	0.073
Alcohol-impaired driving deaths|% Alcohol-Impaired	0.009	0.905	-0.116	-0.039
Income inequality| Income Ratio	0.009	0.914	0.211	0.183
Food insecurity|% Food Insecure	0.008	0.922	0.333	-0.016
Flu vaccinations|% Vaccinated	0.008	0.930	-0.100	-0.011
Life expectancy|Life	0.008	0.938	-0.415	0.232

Expectancy				
Residential segregation - non-white/white| Segregation Index	0.008	0.946	0.130	0.185
Adult smoking|% Smokers	0.008	0.954	0.344	-0.246
Homeownership|% Homeowners	0.007	0.961	-0.107	-0.352
Limited access to healthy foods|% Limited Access	0.006	0.967	0.180	-0.109
insufficient sleep %	0.006	0.974	0.214	0.088
Motor vehicle crash deaths|MV Mortality Rate	0.006	0.980	0.235	-0.271
Children in poverty|% Children in Poverty (White)	0.005	0.985	0.299	-0.215
injury deaths rate	0.005	0.991	0.301	-0.197
frequent mental distress %	0.005	0.996	0.349	-0.124
Demographics|% Not Proficient in English	0.003	0.998	-0.112	0.399
Low birthweight|% LBW	0.002	1.000	0.347	0.048

Table S2. Sensitivity Analysis. Variation of COVID-19 deaths/population by changing each variable's values. Impact of changing the values of the most important variables on the number of COVID-19 caused deaths by county population. On the vertical axis are the most important variables sorted top to bottom by feature importance. On the horizontal axis is how much each feature was changed. The values are the effect of increasing/decreasing , in percentages.

Variables\Change of the variable	-25%	-20%	-15%	-10%	-5%	+5%	+10%	+15%	+20%	+25%
other primary care providers rate	-8.13	-7.28	-6.44	-5.78	-4.78	-3	-2.27	-1.69	-0.87	-0.27
primary care physicians ratio	-0.19	-1.22	-1.75	-2.38	-3.11	-4.39	-4.79	-5.31	-5.82	-6.28
other primary care providers ratio	-1.52	-1.92	-2.39	-2.97	-3.47	-4.18	-4.55	-4.92	-5.32	-5.77
Median household income| Household Income	-0.35	-1.02	-1.68	-2.3	-3.07	-4.58	-5.16	-5.9	-6.63	-7.4
Physical inactivity|% Physically Inactive	-7.88	-7.58	-6.84	-6.03	-5.44	-3	-2.27	-1.53	-0.96	-0.49
children in poverty %	-5.43	-5.06	-4.75	-4.43	-4.06	-3.87	-3.58	-3.32	-3.23	-2.91
Long commute - driving alone|% Long Commute - Drives Alone	-0.85	-1.4	-2.07	-2.46	-2.97	-4.19	-4.66	-4.99	-5.38	-5.74
Driving alone to work|% Drive Alone	-10.01	-10.69	-10.86	-10.52	-8.11	-2.72	-1.95	-1.7	-1.64	-1.66
Diabetes prevalence|% Diabetic	-5.38	-5.3	-5.11	-4.81	-4.47	-3.8	-2.76	-2.73	-1.53	-1.07
Demographics|% Female	-6.59	-6.59	-6.59	-6.59	-6.55	0.27	0.95	1.04	1.05	1.05
Premature age-adjusted mortality| Age-Adjusted Mortality	-6.13	-5.73	-5.26	-4.71	-4.33	-3.56	-3.65	-3.6	-3.71	-3.96
% Uninsured children	-4.88	-4.71	-4.65	-4.38	-4.38	-3.8	-3.73	-3.7	-3.29	-3.13
Social associations rate	-4.75	-4.72	-4.61	-4.52	-4.16	-3.43	-3.19	-2.96	-2.54	-2.08
Demographics|% 65 and over	-6.06	-5.78	-5.35	-4.71	-4.27	-3.31	-2.94	-2.52	-2.24	-2.04
Demographics|% African American	-3.77	-3.73	-3.72	-3.76	-3.81	-3.81	-3.82	-3.88	-3.9	-3.95
Unemployment|Labor Force	-3.26	-3.58	-3.7	-3.71	-3.77	-3.9	-3.91	-3.87	-3.87	-3.8
Sexually transmitted infections| Chlamydia Rate	-2.78	-2.93	-3.25	-3.46	-3.51	-4.01	-4.04	-4.15	-4.25	-4.34
ht_Ages 65+ years Overall Male	-5.21	-4.82	-4.49	-4.18	-3.95	-3.58	-3.31	-3.22	-3.12	-2.93
teen births rate	-4.25	-4.15	-4.01	-3.97	-3.95	-3.77	-3.7	-3.63	-3.58	-3.5
Demographics|% Asian	-3.7	-3.73	-3.8	-3.7	-3.82	-3.81	-3.87	-3.88	-3.69	-3.73
ht_Ages 35-64 years Overall Female	-4.29	-4.2	-4.05	-4.07	-3.9	-3.75	-3.7	-3.62	-3.52	-3.41
violent crime rate	-4.08	-4.07	-4.03	-3.98	-3.93	-3.7	-3.65	-3.62	-3.6	-3.61
some college %	-1.98	-2.34	-2.68	-2.9	-3.4	-4.04	-4.34	-4.56	-4.84	-5.04
Uninsured adults|% Uninsured	-4.48	-4.4	-4.26	-4.1	-4	-3.8	-3.62	-3.5	-3.15	-3.17
Demographics|% Native Hawaiian/Other Pacific Islander	-3.6	-3.65	-3.65	-3.65	-3.65	-3.8	-3.8	-3.8	-3.8	-3.81

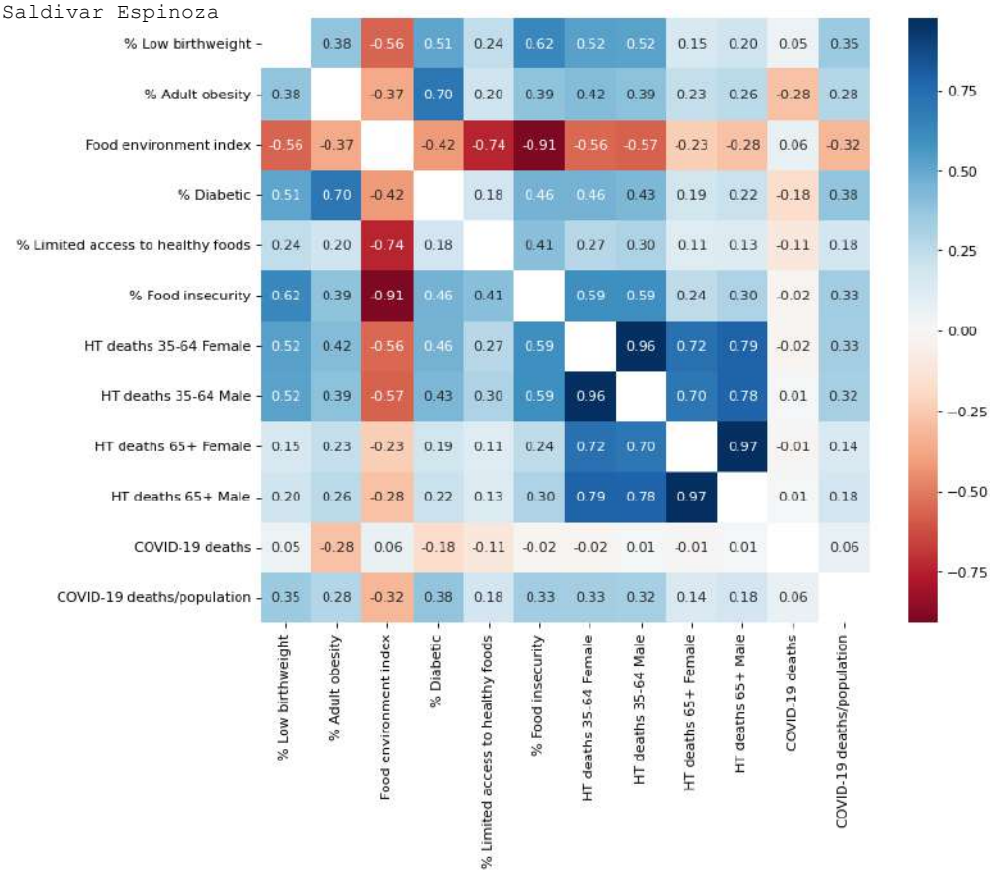


Fig S1. Correlations between nutritional variables and COVID-19 deaths.

Manuscript 2

The Mutational Landscape of SARS-CoV-2

Bryan Saldivar-Espinoza (1), Pol Garcia-Segura (1), Nil Novau-Ferré (1), Guillem Macip (1), Ruben Martínez (2), Pere Puigbò (3,4,5), Adrià Cereto-Massagué (6), Gerard Pujadas (1,*) and Santiago Garcia-Vallve (1,*)

1 Departament de Bioquímica i Biotecnologia, Research Group in Cheminformatics & Nutrition, Campus de Sescelades, Universitat Rovira i Virgili, 43007 Tarragona, Spain

2 Institut La Guineueta, 08042 Barcelona, Spain

3 Department of Biology, University of Turku, 20500 Turku, Finland

4 Department of Biochemistry and Biotechnology, Rovira i Virgili University, 43007 Tarragona, Spain

5 Eurecat, Technology Centre of Catalonia, Unit of Nutrition and Health, 43204 Reus, Spain

6 EURECAT Centre Tecnològic de Catalunya, Centre for Omic Sciences (COS), Joint Unit Universitat Rovira i Virgili-EURECAT, Unique Scientific and Technical Infrastructures (ICTS), 43204 Reus, Spain

* Authors to whom correspondence should be addressed. (gerard.pujadas@gmail.com, anti.garcia-vallve@urv.cat)

International Journal of *Molecular Sciences*

MDPI

Article

The Mutational Landscape of SARS-CoV-2

Bryan Saldivar-Espinoza [1,†], Pol Garcia-Segura [1], Nil Novau-Ferré [1,†], Guillem Macip [1], Ruben Martínez [2], Pere Puigbò [3,4,5], Adrià Cereto-Massagué [6], Gerard Pujadas [1,*] and Santiago Garcia-Vallve [1,*]

1 Departament de Bioquímica i Biotecnologia, Research Group in Cheminformatics & Nutrition, Campus de Sescelades, Universitat Rovira i Virgili, 43007 Tarragona, Spain; bsaldivar.emc2@gmail.com (B.S.-E.); polgarse2@gmail.com (P.G.-S.); nnovauf@gmail.com (N.N.-F.); guillem.macip@gmail.com (G.M.)
2 Institut La Guineueta, 08042 Barcelona, Spain; rmartbernabe@gmail.com
3 Department of Biology, University of Turku, 20500 Turku, Finland; pepuav@utu.fi
4 Department of Biochemistry and Biotechnology, Rovira i Virgili University, 43007 Tarragona, Spain
5 Eurecat, Technology Centre of Catalonia, Unit of Nutrition and Health, 43204 Reus, Spain
6 EURECAT Centre Tecnològic de Catalunya, Centre for Omic Sciences (COS), Joint Unit Universitat Rovira i Virgili-EURECAT, Unique Scientific and Technical Infrastructures (ICTS), 43204 Reus, Spain; ssorgatem@gmail.com
* Correspondence: gerard.pujadas@gmail.com (G.P.); santi.garcia-vallve@urv.cat (S.G.-V.)
† These authors contributed equally to this work.

Abstract: Mutation research is crucial for detecting and treating SARS-CoV-2 and developing vaccines. Using over 5,300,000 sequences from SARS-CoV-2 genomes and custom Python programs, we analyzed the mutational landscape of SARS-CoV-2. Although almost every nucleotide in the SARS-CoV-2 genome has mutated at some time, the substantial differences in the frequency and regularity of mutations warrant further examination. C>U mutations are the most common. They are found in the largest number of variants, pangolin lineages, and countries, which indicates that they are a driving force behind the evolution of SARS-CoV-2. Not all SARS-CoV-2 genes have mutated in the same way. Fewer non-synonymous single nucleotide variations are found in genes that encode proteins with a critical role in virus replication than in genes with ancillary roles. Some genes, such as spike (*S*) and nucleocapsid (*N*), show more non-synonymous mutations than others. Although the prevalence of mutations in the target regions of COVID-19 diagnostic RT-qPCR tests is generally low, in some cases, such as for some primers that bind to the *N* gene, it is significant. Therefore, ongoing monitoring of SARS-CoV-2 mutations is crucial. The SARS-CoV-2 Mutation Portal provides access to a database of SARS-CoV-2 mutations.

Keywords: SARS-CoV-2 mutations; COVID-19; molecular evolution

Citation: Saldivar-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. *Int. J. Mol. Sci.* **2023**, *24*, 9072. https://doi.org/10.3390/ijms24109072

Academic Editors: Kamalendra Singh and Christian Lorson

Received: 28 March 2023
Revised: 12 May 2023
Accepted: 16 May 2023
Published: 22 May 2023

1. Introduction

Mutation (including insertions and deletions) and recombination are two important mechanisms that generate genomic variability in SARS-CoV-2 variants [1]. Most SARS-CoV-2 mutations are expected to be either neutral or mildly deleterious [2]. Highly deleterious mutations, such as those that prevent the virus from invading the host, are unlikely to occur. However, SARS-CoV-2 is under selective pressure because of vaccines and antiviral drugs [3]. Mutations that improve virulence, infectivity, transmissibility, increase viral replication, or aid in immune evasion are expected to be fixed and spread. However, the high frequency of certain mutations is not always due to a mutation's beneficial effect. It can also be caused by a founder effect, which occurs when a mutation appears early in the evolution of a pandemic and is transmitted to all of its descendants [4] or when a mutation is found in a variant that also carries an additional advantageous mutation. Genetic diversification of the SARS-CoV-2 virus has led to the emergence of new clades and variants [5,6]. Variants of concern (VOC) are SARS-CoV-2 variants for which there is

evidence of an increase in transmissibility or virulence, a detrimental charge in COVID-19 epidemiology, or a decrease in the effectiveness of available diagnostic tools, vaccines, or therapeutics. Omicron is the only currently circulating VOC (in March 2023). It was first identified in November 2021 and has since been responsible for the vast majority of COVID-19 cases worldwide [7]. This variant has undergone significant mutations in comparison to previous variants [8]. Alpha, beta, delta, and gamma VOCs have all previously been in circulation. In December 2020, a rapidly growing lineage (the alpha variant) was identified in the UK [1] and increased in prevalence worldwide in the following months. Soon after, other rapidly growing variants, beta and gamma, appeared [1] but were soon overtaken by the delta variant that appeared in India, spread widely in numerous countries, and become the predominant variant in the second part of 2021 until the emergence of Omicron. All of these variants contained the spike mutation D614G that resulted in increased SARS-CoV-2 infectivity [9–11].

Mutations in SARS-CoV-2 can be caused by RNA-dependent RNA polymerase (RdRp) replication errors or by host deaminases that deaminate unpaired nitrogenous bases [1]. At the start of the pandemic, the prevalence of C>U mutations and other evidence suggested that RNA editing is the major source of SARS-CoV-2 mutation [12–17]. Since then, the role of RNA editing in SARS-CoV-2 evolution has been experimentally demonstrated [18–20]. Because of the prevalence of RNA editing in SARS-CoV-2 evolution, some recurrent mutations in SARS-CoV-2 can be predicted [21]. Mammalian apolipoprotein B mRNA editing catalytic polypeptide-like (APOBEC) enzymes deaminate cytosines into uracils in single-stranded DNA (ssDNA) and ssRNA [22]. When APOBEC enzymes act on the SARS-CoV-2 genome's negative strand, G>A mutations occur on the positive strand [23]. Adenosine deaminases acting on RNA (ADAR) deaminate adenines into inosines (A>I) in double-stranded RNA (dsRNA) [24]. As inosine preferentially pairs with cytidine, A>I mutations cause A>G and U>C transitions on the positive strand of the SARS-CoV-2 genome [23,25]. The presence of SARS-CoV-2 mutations may affect the test performance of COVID-19 diagnostic tests [26–28]. For this reason, they must be monitored. To provide guidelines and recommendations for assessing the potential effects of current and future viral mutations of SARS-CoV-2 on COVID-19 tests, in February 2021 the FDA published the "Policy for Evaluating Impact of Viral Mutations on COVID-19 Tests", which was updated in January 2023. Molecular tests are intended to detect viruses by focusing on a specific region of the viral genome. False-negative results can occur if there are mutations that reduce the ability of these tests to detect the virus's RNA genome. The gold-standard test to detect COVID-19 is the quantitative RT-PCR (RT-qPCR), which uses forward and reverse primers to amplify a specific region of the SARS-CoV-2 genome. Probes bind downstream of one of the primers and give a fluorescent signal proportional to the number of amplicons synthesized. Various primers and probe sets have been reported for the detection of SARS-CoV-2 by RT-qPCR [29]. SNVs, mutations, and insertions can affect primer and probe hybridization and cause amplification failure [26], especially if they cause mismatches with the template DNA near the 3′-end of a primer [30]. The effects of mutations are less pronounced in tests designed to detect multiple SARS-CoV-2 genes than in tests designed to detect a single target. For example, the use of a multiplex RT-qPCR made it possible to identify the alpha variant for the first time in the UK. The 69-70 deletion in this variant prevents the spike (*S*) gene from being amplified in the Thermo Fisher TaqPath COVID-19 PCR assay, resulting in *S*-gene target failure in the test results [31].

The study of SARS-CoV-2 mutations is critical for detecting and treating SARS-CoV-2 and developing vaccines, and it must be carried out on a regular basis. Throughout the pandemic, the SARS-CoV-2 mutational landscape has been analyzed recurrently [32–36], though occasionally with a small number of genomes or focusing on amino acid changes or a specific gene, country, or variant. In this article, we analyze the mutational landscape of SARS-CoV-2 using data from more than 5 million SARS-CoV-2 genomes, collected after more than two years of the pandemic. We focus on nucleotide-level changes, and

in particular, we analyze their distribution among SARS-CoV-2 genes, the most common mutations and types of mutations, and their potential impact on COVID diagnostic tests.

2. Results and Discussion

2.1. SARS-CoV-2 Genomes Analyzed

We analyzed 5,340,569 SARS-CoV-2 genomes available from the GISAID database [37]. They are complete, high-coverage SARS-CoV-2 genomes isolated from humans and were available on 27 June 2022. Since the rates of genome sequencing in different nations fluctuate significantly, it is important to keep in mind that there is a bias in the genomes examined. The USA and the United Kingdom sequenced 51.9% of all genomes (Figure S1). In terms of continents, Europe (55.1%) and North America (34.1%) accounted for the majority of genomes (Table S1). However, this bias does not invalidate the results reported herein. The genomes analyzed were collected between December 2019 and June 2022 (Figure 1). The number of genomes increased from 2020 as sequencing efforts in different countries and the number of cases increased. At the end of 2020, the alpha variant emerged, and throughout the first few months of 2021, it predominated, although it did not completely replace earlier varieties. The delta variant caused an exponential rise in the number of cases, and by the end of 2021, it was the most common variety. Then, at the start of 2022, the omicron variant took its place (Figure 1).

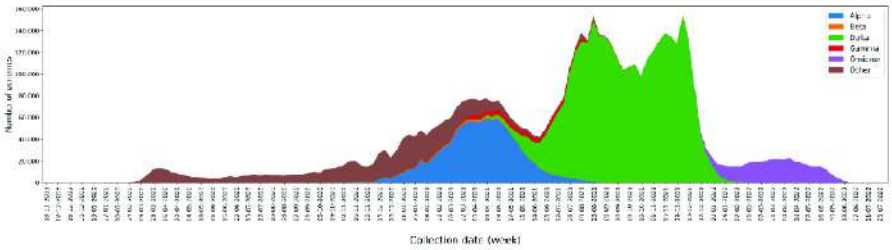


Figure 1. The number of genomes collected per week and classified by a variant of concern (VOC): alpha in blue, beta in orange, delta in green, gamma in red, omicron in purple, and others in brown.

2.2. Mutations, Deletions, and Insertions per Genome per Week

Among the mutations, the most frequent were single nucleotide variants (SNVs): i.e., those that exchange one nucleotide for another (Table S2). As expected, the number of SNVs per genome per week increased during the pandemic (Figure S2). Until mid-May 2020, the average number of SNVs per genome was less than 10 (Figure S2). In June 2020, the average was around 7 [33] but by the beginning of January 2022, it had increased to 50. It then increased again when the omicron variant expanded, and by early June 2022, the average number of SNVs per genome was around 72 (Figure S2). In terms of variants, alpha, beta, delta, and gamma VOCs contain a median of 29 to 41 SNVs per genome (Figure 2). The omicron variant is the most highly mutated VOC, with over 60 SNVs per genome (Figure 2) that potentially improve transmissibility, immunological evasion, and virulence [38,39].

The number of deletions per genome per week was quite low until early 2021 when there was an increase (Figure S3). Since then, they have remained at an average of three deletions per genome. Some deletions are conserved in SARS-CoV-2 variants and have a significant regional preference, possibly to prevent neutralizing antibodies from binding to their target and thus cause immune escape [40–42]. Thus, although SNVs outnumber deletions, deletions have a significant influence on the evolution of viruses and may contribute to the evasion of immune responses and the evolution of highly transmissible variants [43,44]. Over the course of the pandemic, there have been few insertions, an average of 0.2 per genome (Figure S4). Questions have been raised about whether some of the insertions observed in the SARS-CoV-2 genomes were insertions or sequencing arti-

facts [45]. Figures S5 and S6 show that the most common lengths of deletions and insertions in the coding regions of the SARS-CoV-2 genome are multiples of three nucleotides (3, 6, 9, ...). This suggests that some of the deletions and insertions are caused by real viral variation and not by sequencing errors. Single nucleotide deletions are relatively frequent (Figure S5), but 26% of them occur in *ORF7a* or *ORF8* genes. Deletions that truncate the *ORF7a* or *ORF8* genes have been observed and associated with a milder infection [43,46]. Because insertions and deletions can affect the antigenic properties of SARS-CoV-2 proteins, they had to be monitored [40,45].

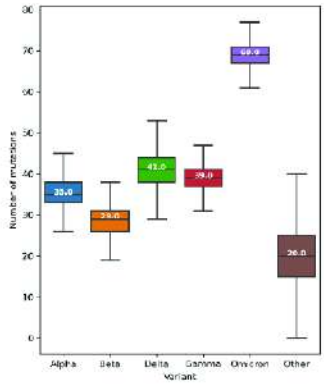


Figure 2. Boxplots of the number of single nucleotide variants (SNVs) per genome and VOC. The boxes show the first, second (median), and third quartiles, and the whiskers show the minimum and maximum values, excluding outliers. The median of each VOC is shown in white.

2.3. Most Frequent SNVs

A total of 73,464 different SNVs were found in the 5,340,569 SARS-CoV-2 genomes analyzed. Of these, 1842 were mutations from untranslated regions (UTRs), 51,467 were non-synonymous, 18,413 were synonymous, and 1742 were only observed in conjunction with another mutation affecting the same codon (Table 1). Although there are more non-synonymous than synonymous mutations, synonymous mutations are generally more frequent (Figure S7 and median values in Table 1). This is to be expected because synonymous mutations have fewer restrictions and do not alter the coded protein. However, codon usage and the maintenance of the RNA secondary structure are two forces that can cause some selection pressure on synonymous mutations [47]. The distribution of synonymous mutations and mutations from UTRs are comparable (Figure S7).

Table 1. Unique SNV counts and the median number of genomes for each mutation type. Frequency in % is shown in parentheses.

Mutation Type	Count	Median Number of Genomes (%)
in UTRs	1842	120 (2.2×10^{-3} %)
non-synonymous	51,467	20 (3.8×10^{-4} %)
synonymous	18,413	135 (2.6×10^{-3} %)
not alone	1742	1 (1.9×10^{-5} %)

Not all SNVs are equally frequent and many are low frequency [48]. In fact, 23.69%, 8.19%, and 4.61% of SNVs have been found in only one, two, or three genomes (Figure S8). These percentages decrease as the number of genomes increases, but 27.25% of SNVs have

been found in more than 100 genomes (Figure S8). The most frequent SNVs are the A23403G spike mutation (present in 99.47% of SARS-CoV-2 genomes analyzed), the C14408U RdRp mutation (present in 99.35% of genomes), the C3037U synonymous mutation (present in 99.27% of genomes), and the C241U UTR mutation (present in 97.96% of genomes) (Figure 3 and Table 2). These mutations appeared early in the pandemic (January 2020) and have since become dominant [33,49]. The A23403G mutation causes the D614G mutation at the S protein, which has been associated with enhanced infectivity [9–11]. The C14408U mutation in the *RdRp* gene causes the non-synonymous mutation P323L (Table 2). According to recent studies, this mutation confers transmission advantages and was crucial to the P323L/D614G genotype becoming established early in the pandemic [50]. C3037U and C241U mutations are most likely neutral [51]. C3037U is a synonymous mutation that affects the *nsp3* gene and C241U is found in an unpaired six-base loop in the conserved 5′-UTR SL5B secondary structure [51]. The most frequent mutations in each SARS-CoV-2 gene are shown in Figure S9. Of the 24 genes, 12 (i.e., the S, *RdRp*, *nsp3*, *nsp4*, nucleocapsid (*N*), *M*, *ORF3a*, *helicase*, *ORF7a*, *ORF8*, *nsp6*, and *exonuclease*) have some mutations with a prevalence higher than 50% (Figure S9).

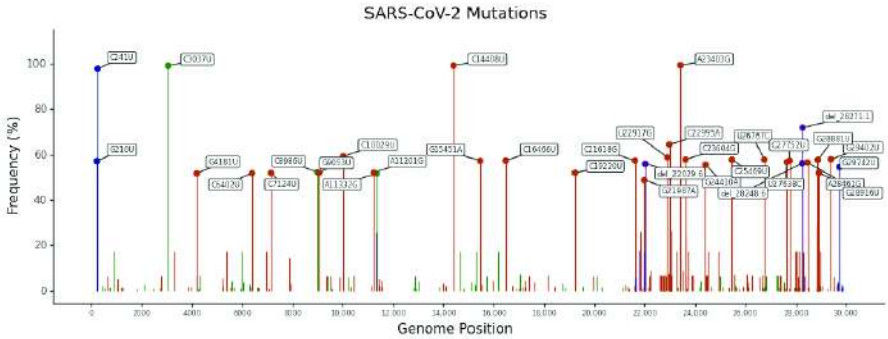


Figure 3. Lolliplot of the most frequent SARS-CoV-2 mutations. Synonymous, non-synonymous, and UTR mutations are shown in green, dark orange, and blue, respectively. Deletions are in purple.

Table 2. The 10 most common SARS-CoV-2 mutations.

Mutation	Gene	Mutation Type	AA Change	No. Genomes (%)
A23403G	*S*	non-synonymous	D614G	5,312,457 (99.48%)
C14408U	*RdRp*	non-synonymous	P323L	5,306,009 (99.35%)
C3037U	*nsp3*	synonymous	F106F	5,301,673 (99.27%)
C241U	*5′-UTR*	in UTRs	-	5,231,432 (97.96%)
del_28271:1	*N*	deletion	-	3,835,041 (71.81%)
C22995A	*S*	non-synonymous	T478K	3,456,098 (64.71%)
C10029U	*nsp4*	non-synonymous	T492I	3,176,761 (59.48%)
U22917G	*S*	non-synonymous	L452R	3,149,402 (58.97%)
G29402U	*N*	non-synonymous	D377Y	3,105,411 (58.15%)
C23604G	*S*	non-synonymous	P681R	3,097,770 (58.00%)

Some of the SARS-CoV-2 mutations are specific to some SARS-CoV-2 variants and have been used for early identification of SARS-CoV-2 variants through amplification [52,53]. Table 3 shows some of the variant-specific mutations from the spike protein and their frequency among some variants. L452R, W152C, K417T, and K417N mutations are particularly

specific to Delta, Epsilon, Gamma, and Omicron variants, respectively (Table 3). These and other mutations (and combinations of them) have been proposed to identify variants, but erroneous identifications can occur when using only single specific mutations [52]. Therefore, sequencing is currently the gold standard method for variant identification [52].

Table 3. Total frequency and variant frequency of some of the *S* mutations used in the classification of SARS-CoV-2 variants.

Mutation	No. Genomes (% [a])	No. Genomes (% [b]) in Variants					
		Alpha	Beta	Delta	Epsilon	Gamma	Omicron
Δ69/70	945,256 (17.70%)	895,448 (94.73%)	14 (0.001%)	3850 (0.41%)	287 (0.03%)	189 (0.02%)	19,260 (2.04%)
W152C	45,304 (0.85%)	9 (0.02%)	0 (0%)	29 (0.06%)	45,192 (99.75%)	2 (0.004%)	1 (0.002%)
K417N	383,722 (7.18%)	137 (0.04%)	24,366 (6.35%)	5353 (1.40%)	10 (0.003%)	2 (0.0005%)	352,543 (91.87%)
K417T	88,480 (16.32%)	32 (0.04%)	0 (0%)	59 (0.07%)	1 (0.001%)	86,902 (98.22%)	1426 (1.61%)
L452R	3,149,260 (58.97%)	450 (0.01%)	28 (0.0009%)	3,075,838 (97.67%)	45,457 (1.44%)	42 (0.001%)	1847 (0.06%)
E484A	357,227 (6.69%)	20 (0.006%)	0 (0%)	2595 (0.73%)	1 (0.0003%)	0 (0%)	354,118 (99.13%)
N501Y	1,396,003 (26.14%)	914,121 (65.48%)	25,377 (1.82%)	1520 (0.11%)	25 (0.0002%)	89,927 (6.44%)	348,897 (24.99%)

[a] The percentage is calculated in relation to the total number of genomes. [b] The percentage of each variant is calculated in relation to the total number of genomes containing that mutation.

Not all SARS-CoV-2 genes have accumulated the same number of mutations. As mutation rates per nucleotide are small, our calculations were based on 100 nucleotides (Figure 4). The number of synonymous mutations per 100 nucleotides is quite similar across all SARS-CoV-2 genes (Figure 4). On average, there are 63.6 synonymous SNVs per 100 nucleotides. Other types of mutations are more variable. The number of non-synonymous SNVs per 100 nucleotides ranges between 147.5 and 298.5 (Figure 4). There are fewer non-synonymous SNVs in genes that encode proteins that play critical roles in virus replication, e.g., *helicase*, *RdRp*, and main protease (*M-pro*), than in genes with accessory functions (e.g., *ORF7a*, *ORF8*, and *ORF6*). This is consistent with previous observations from mid-2020, which indicates that there is a tendency to conserve important structural and functional features in SARS-CoV-2 proteins [35]. Genes encoding S and N proteins have more non-synonymous SNVs than other genes (Figure 4). We expected the *S* gene to contain more non-synonymous mutations. Mutations in the S protein may enhance its interaction with ACE2, help it to escape from the immune system, or improve furin cleavage [2,3,54,55]. It has also been suggested that the *S* gene is more likely to be single-stranded than other SARS-CoV-2 genes, thus making it a favourable target for C>U deamination and leading to an excessively high mutation rate [56]. The high mutation frequency of the *N* gene may be due to its higher G+C percentage [57]. This gene is frequently used as a target for RT-qPCR diagnostic tests and it has been suggested that it be part of future vaccines against COVID-19 [58]. Nonetheless, its high mutation frequency must be considered since any changes in this gene may render vaccines or diagnostic tests ineffective [59]. However, mutations in the *N* gene are not uniformly distributed, and a leucine-rich sequence (LRS) from amino acids 218 to 231 is a conserved region that may provide a new path for the development of pan-coronavirus therapeutics and vaccines [60,61].

The number of insertions and deletions among SARS-CoV-2 genes is also highly variable (Figure 4). Genes that encode proteins essential for viral replication contain fewer insertions and deletions (Figure 4). It is worth noting a large number of deletions in accessory genes, such as *ORF7a*, *ORF8*, and *ORF6* (Figure 4). It has been suggested that deletions in these genes may eventually lead to more effective variants that produce a milder infection [43,44,46]. In all genes, insertions are less common than deletions (Figure 4).

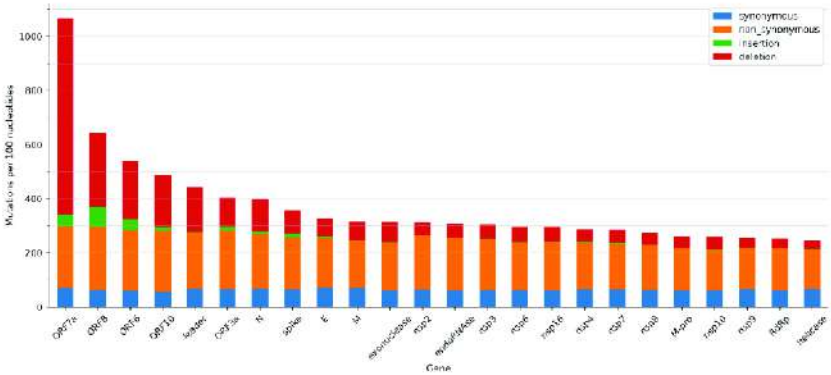


Figure 4. Mutations per 100 nucleotides in the SARS-CoV-2 genes. Synonymous, non-synonymous mutations, insertions, and deletions are shown in blue, orange, green, and red, respectively.

2.4. *SNV Signature Analysis*

Of the 73,464 SNVs analyzed, transversions—i.e., an SNV in which a purine is exchanged for a pyrimidine or vice versa—are more frequent than transitions (61.72% vs. 38.28%). The most prevalent mutations are U>C and A>G (Table 4). However, because the SARS-CoV-2 genome is richer in As and Us than in Gs and Cs (its G+C content is 37.97%), the C>U mutation stands out when the fraction of each type of nucleotide that has mutated is calculated (Table 4). A total of 97.4% of all Cs in the SARS-CoV-2 genome have mutated at some time to a U, but only 65.2% of them have mutated to a G (Table 4). This is consistent with the C>U mutation being the most common SNV at the beginning of the pandemic (Figure 5) [15,33,34,62]. By mid-April 2020, 70% of all C>U mutations had already been observed (Figure 5). In addition, C>U mutations are the most frequent mutations on average [17], and they have been observed in the largest number of variants, pangolin lineages, and countries (Figure S10). All of this evidence supports the role of C>U mutations as a driving mechanism in the evolution of SARS-CoV-2 [63]. The second most remarkable SNV type is the A>G mutation (Table 4). A total of 94.0% of all As in the SARS-CoV-2 genome have mutated at some time to a G (Table 4), and 70% of total A>G mutations were first observed by the end of September 2020 (Figure 5). The prevalence of C>U and A>G mutations is consistent with the predominant role of host deaminases in causing a significant portion of SARS-CoV-2 mutations [14,17,18,64].

Table 4. SNV counts showing the initial nucleotide (from) and the new nucleotide (to). The percentage of the total number of initial bases in the SARS-CoV-2 genome is displayed in parentheses.

		To Nucleotide				
		A	**G**	**C**	**U**	**Total SNVs**
From nucleotide	**A**	0	8416 (94.0%)	7008 (78.3%)	6982 (78.0%)	22,406
	G	5420 (92.4%)	0	4059 (69.2%)	5475 (93.4%)	14,954
	C	4780 (87.0%)	3580 (65.2%)	0	5351 (97.4%)	13,711
	U	6791 (70.8%)	6666 (69.5%)	8936 (93.1%)	0	22,393
Total SNVs		16,991	18,662	20,003	17,808	73,464

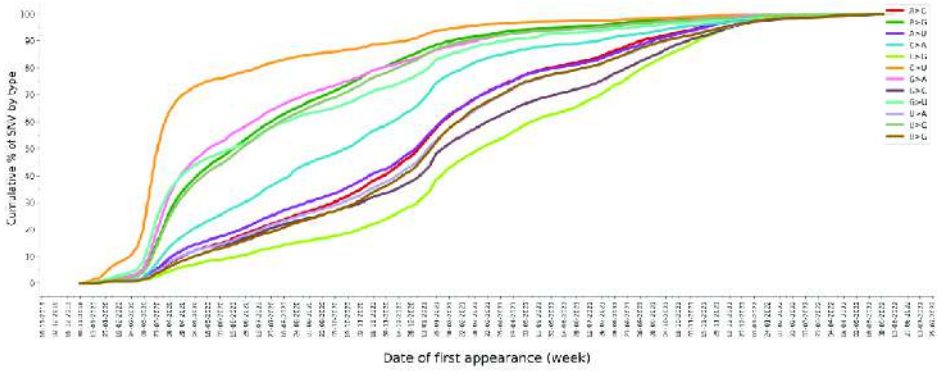


Figure 5. Cumulative percentage of SNV types by date of first appearance.

2.5. *Mutations in the Target Regions of the COVID-19 Diagnostic RT-qPCR Tests*

Table 5 and Table S3 show the number of different mutations found in the primer and probe regions used in the RT-qPCR for COVID-19 diagnosis. Although the frequency of mutations is usually low (Figure S11), in some cases they are important. For example, the total frequency of the Charite-RdRp primer/probe set is 60.84% (Table 5), or 57.57% when the SNVs were in the last 5 nucleotides of the 3′-end of the forward primer (Table S3). For the China-CDC-N set, the total frequency is 141.29% (Table 5), mainly due to three missense mutations: (i) the G28881U mutation that is found in 57.8% of the genomes analyzed; (ii) the two simultaneous mutations G28881A and G28882A that affect the same codon, with a frequency of 29.3% and (iii) the G28883C mutation, with a frequency of 28.1%. The *N* gene is highly conserved in coronavirus. For this reason, it has been extensively used by RT-qPCR as a target region to detect COVID-19. However, the *N* gene is one of the SARS-CoV-2 genes with the most reported mutations (Figure 4). Some *N* gene mutations, such as the SNVs G29140U, G29179U, and C29200U, and deletions have been reported to affect RT-qPCR results [65–72]. Therefore, using primers and probes that hybridize to a region of the *N* gene is not an optimal choice [73]. A negative result in one of the target genes in a multiplex RT-qPCR assay used to detect COVID-19 is not interpreted as a negative test result, but it may render the assay susceptible to diagnostic failure. Consequently, continued surveillance of SARS-CoV-2 mutations is critical [74]. However, the lack of information about the primers and probes used by some commercial RT-qPCR kits is a drawback for this type of analysis. To reduce the impact of SARS-CoV-2 mutations on COVID-19 surveillance, new primers, and probes targeting the most conserved regions of the SARS-CoV-2 genome or specific regions of a SARS-CoV-2 variant have been suggested [74].

Table 5. The number of different mutations found in SARS-CoV-2 regions that hybridize with probes and forward and reverse primers from some COVID-19 diagnostic RT-qPCR tests.

Name	Gene	Region Amplified	No. Different Mutations Found in Forward and Reverse Primers and Probe	Total No. Mutations and Total Frequency (%)
nCoV_IP2	*RdRp*	12,690–12,797	46 | 68 | 64	178 (1.75%)
nCoV_IP4	*RdRp*	14,080–14,186	50 | 66 | 65	181 (3.95%)
Charite-E	*E*	26,269–26,381	89 | 81 | 143	313 (7.15%)
N-Sarbeco	*N*	28,706–28,833	68 | 93 | 116	277 (2.14%)
Charite-RdRp	*RdRp*	15,431–15,528	67 | 52 | 64	183 (60.84%)

Table 5. *Cont.*

Name	Gene	Region Amplified	No. Different Mutations Found in Forward and Reverse Primers and Probe	Total No. Mutations and Total Frequency (%)
HKU-ORF1ab	*ORF1ab*	18,778–18,909	60 | 73 | 60	193 (1.18%)
HKU-N	*N*	29,145–29,254	145 | 222 | 167	534 (3.25%)
China-CDC-ORF1ab	*ORF1ab*	13,342–13,460	58 | 59 | 103	220 (0.79%)
China-CDC-N	*N*	28,881–28,979	156 | 118 | 86	360 (141.28%)
US-CDC-N1	*N*	28,287–28,358	102 | 111 | 131	344 (14.59%)
US-CDC-N2	*N*	29,164–29,230	154 | 184 | 189	527 (2.73%)
US-CDC-N3	*N*	28,681–28,752	88 | 91 | 90	269 (3.36%)
Japan-N	*N*	29,125–29,282	116 | 234 | 211	561 (2.02%)
Thailand-N	*N*	28,320–28,376	104 | 112 | 78	294 (2.46%)
Sigma-Aldrich	*N*	28,750–28,860	96 | 96 | -[1]	192 (2.66%)

[1] It does not use a probe.

2.6. SARS-CoV-2 Mutation Portal

We have created a database of all the mutations discovered in the more than five million SARS-CoV-2 genomes analyzed. The SARS-CoV-2 Mutation Portal (http://sarscov2-mutation-portal.urv.cat/, accessed on May 2023) provides access to this database, which contains information on over 100,000 mutations (including point mutations, insertions, and deletions). For each mutation, it gives a variety of information, such as the type of mutation, its location, effect, frequency, the number of countries, lineages, and variants in which it has been found. The mutations are shown in the form of a table and a scatter diagram (Figures S12–S14).

3. Materials and Methods

Origin and Characterization of the SARS-CoV-2 Genomes Analyzed

A FASTA file containing the multiple sequence alignment of 10,417,619 complete SARS-CoV-2 genomes were downloaded from the GISAID database [37] on 27 June 2022. In this multi-alignment file, the SARS-CoV-2 sequence NC_045512.2, isolated from Wuhan and submitted to the GenBank database on 17 January 2020, was used as a reference. Only sequences labelled as "high coverage" (i.e., sequences containing: (a) less than 1% of unidentified bases (Ns), (b) less than 0.05% of unique amino acid mutations, to withdraw possible sequencing artefacts, and (c) no insertions and/or deletions, unless verified by the submitter) and obtained from human samples were considered. Thus, the initial number of SARS-CoV-2 genomes was reduced to 5,340,569 sequences. For each sequence, information about the collection date, location, pango lineage [75], and VOC was extracted from a metadata file available in GISAID. For each sequence, single mutations, insertions, and deletions were extracted and numbered relative to the reference genome. Mutations were classified as mutations from UTRs, synonymous mutations (i.e., mutations that do not affect the encoded amino acid), and non-synonymous mutations (which include missense and nonsense mutations). Mutation frequencies were calculated as the number of specific mutations in the total number of genomes. All analyses and figures were created with custom programs in Python 3.9.

4. Conclusions

Although almost every nucleotide in the SARS-CoV-2 genome has mutated at some time, the frequency and regularity of the mutations vary significantly. C>U mutations are the most prevalent mutations. They are found in the largest number of variants, pangolin lineages, and countries. The predominance of C>U mutations during the early stages of the pandemic suggested that host deaminases were responsible for a considerable percentage

of SARS-CoV-2 mutations. Since then, the predominant role of host deaminases on SARS-CoV-2 evolution has been demonstrated experimentally. Not all SARS-CoV-2 genes have accumulated the same number of mutations. Non-synonymous SNVs are less common in genes encoding proteins that have key roles in virus replication than in genes with accessory functions. Genes encoding S and N proteins are among the genes with the most non-synonymous SNVs. Although the prevalence of mutations in the target regions of COVID-19 diagnostic RT-qPCR tests is generally low, it is significant in some cases, such as for some primers that bind to the *N* gene. For this reason, SARS-CoV-2 mutations must be tracked. However, the lack of information about the primers and probes used by some commercial RT-qPCR kits is a drawback for this type of analysis. The SARS-CoV-2 Mutation Portal (at http://sarscov2-mutation-portal.urv.cat/, accessed on 10 May 2023) gives access to a database of all the mutations (including point mutations, insertions, and deletions) that have been analyzed here.

Supplementary Materials: The supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms24109072/s1.

Author Contributions: Conceptualization, B.S.-E., P.G.-S., N.N.-F. and S.G.-V.; methodology, B.S.-E., P.G.-S., N.N.-F. and S.G.-V.; formal analysis, B.S.-E., P.G.-S., N.N.-F. and S.G.-V.; investigation, B.S.-E., P.G.-S., N.N.-F. and S.G.-V.; data curation, B.S.-E., S.G.-V. and R.M.; software, A.C.-M. and R.M. writing—original draft preparation, B.S.-E. and S.G.-V.; writing—review and editing, B.S.-E., P.G.-S., N.N.-F., P.P., G.P. and S.G.-V.; visualization, B.S.-E., G.M., P.G.-S., N.N.-F. and S.G.-V.; supervision, P.P., A.C.-M., G.P. and S.G.-V.; project administration, S.G.-V. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie, grant agreement No. 713679, and by the Universitat Rovira i Virgili, grant 2021PFR-URV-96.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: We have created the database SARS-CoV-2 Mutation Portal (http://sarscov2-mutation-portal.urv.cat/SARS-CoV-2_mutation-portal, accessed on 10 May 2023) with all mutations discovered in the more than five million genomes analyzed.

Acknowledgments: We would like to acknowledge the authors, both from the submitting and originating laboratories, for the sequences from the GISAID database used in this study. We acknowledge our University's English language service for proofreading and correcting this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tao, K.; Tzou, P.L.; Nouhin, J.; Gupta, R.K.; de Oliveira, T.; Kosakovsky Pond, S.L.; Fera, D.; Shafer, R.W. The Biological and Clinical Significance of Emerging SARS-CoV-2 Variants. *Nat. Rev. Genet.* **2021**, *22*, 757–773. [CrossRef] [PubMed]
2. Harvey, W.T.; Carabelli, A.M.; Jackson, B.; Gupta, R.K.; Thomson, E.C.; Harrison, E.M.; Ludden, C.; Reeve, R.; Rambaut, A.; COVID-19 Genomics UK (COG-UK) Consortium; et al. SARS-CoV-2 Variants, Spike Mutations and Immune Escape. *Nat. Rev. Microbiol.* **2021**, *19*, 409–424. [CrossRef] [PubMed]
3. Souza, P.F.N.; Mesquita, F.P.; Amaral, J.L.; Landim, P.G.C.; Lima, K.R.P.; Costa, M.B.; Farias, I.R.; Belém, M.O.; Pinto, Y.O.; Moreira, H.H.T.; et al. The Spike Glycoprotein of SARS-CoV-2: A Review of How Mutations of Spike Glycoproteins Have Driven the Emergence of Variants with High Transmissibility and Immune Escape. *Int. J. Biol. Macromol.* **2022**, *208*, 105–125. [CrossRef] [PubMed]
4. Lauring, A.S.; Hodcroft, E.B. Genetic Variants of SARS-CoV-2-What Do They Mean? *JAMA* **2021**, *325*, 529–531. [CrossRef]
5. Alkhatib, M.; Svicher, V.; Salpini, R.; Ambrosio, F.A.; Bellocchi, M.C.; Carioti, L.; Piermatteo, L.; Scutari, R.; Costa, G.; Artese, A.; et al. SARS-CoV-2 Variants and Their Relevant Mutational Profiles: Update Summer 2021. *Microbiol. Spectr.* **2021**, *9*, e0109621. [CrossRef]
6. Chakraborty, C.; Sharma, A.R.; Bhattacharya, M.; Agoramoorthy, G.; Lee, S.-S. Evolution, Mode of Transmission, and Mutational Landscape of Newly Emerging SARS-CoV-2 Variants. *mBio* **2021**, *12*, e0114021. [CrossRef]
7. Callaway, E. Beyond Omicron: What's next for COVID's Viral Evolution. *Nature* **2021**, *600*, 204–207. [CrossRef]

8. Kannan, S.; Shaik Syed Ali, P.; Sheeza, A. Omicron (B.1.1.529)—Variant of Concern—Molecular Profile and Epidemiology: A Mini Review. *Eur. Rev. Med. Pharmacol. Sci.* **2021**, *25*, 8019–8022. [CrossRef]
9. Zhang, L.; Jackson, C.B.; Mou, H.; Ojha, A.; Peng, H.; Quinlan, B.D.; Rangarajan, E.S.; Pan, A.; Vanderheiden, A.; Suthar, M.S.; et al. SARS-CoV-2 Spike-Protein D614G Mutation Increases Virion Spike Density and Infectivity. *Nat. Commun.* **2020**, *11*, 6013. [CrossRef]
10. Korber, B.; Fischer, W.M.; Gnanakaran, S.; Yoon, H.; Theiler, J.; Abfalterer, W.; Hengartner, N.; Giorgi, E.E.; Bhattacharya, T.; Foley, B.; et al. Tracking Changes in SARS-CoV-2 Spike: Evidence That D614G Increases Infectivity of the COVID-19 Virus. *Cell* **2020**, *182*, 812–827.e19. [CrossRef]
11. Plante, J.A.; Liu, Y.; Liu, J.; Xia, H.; Johnson, B.A.; Lokugamage, K.G.; Zhang, X.; Muruato, A.E.; Zou, J.; Fontes-Garfias, C.R.; et al. Spike Mutation D614G Alters SARS-CoV-2 Fitness. *Nature* **2021**, *592*, 116–121. [CrossRef]
12. Simmonds, P. 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. *mSphere* **2020**, *5*, e00408-20. [CrossRef]
13. Di Giorgio, S.; Martignano, F.; Torcia, M.G.; Mattiuz, G.; Conticello, S.G. Evidence for Host-Dependent RNA Editing in the Transcriptome of SARS-CoV-2. *Sci. Adv.* **2020**, *6*, eabb5813. [CrossRef]
14. Ratcliff, J.; Simmonds, P. Potential APOBEC-Mediated RNA Editing of the Genomes of SARS-CoV-2 and Other Coronaviruses and Its Impact on Their Longer Term Evolution. *Virology* **2021**, *556*, 62–72. [CrossRef]
15. Simmonds, P.; Ansari, M.A. Extensive C->U Transition Biases in the Genomes of a Wide Range of Mammalian RNA Viruses; Potential Associations with Transcriptional Mutations, Damage- or Host-Mediated Editing of Viral RNA. *PLoS Pathog.* **2021**, *17*, e1009596. [CrossRef]
16. Li, J.; Lai, S.; Gao, G.F.; Shi, W. The Emergence, Genomic Diversity and Global Spread of SARS-CoV-2. *Nature* **2021**, *600*, 408–418. [CrossRef]
17. Wang, J.; Wu, L.; Pu, X.; Liu, B.; Cao, M. Evidence Supporting That C-to-U RNA Editing Is the Major Force That Drives SARS-CoV-2 Evolution. *J. Mol. Evol.* **2023**, *91*, 214–224. [CrossRef]
18. Kim, K.; Calabrese, P.; Wang, S.; Qin, C.; Rao, Y.; Feng, P.; Chen, X.S. The Roles of APOBEC-Mediated RNA Editing in SARS-CoV-2 Mutations, Replication and Fitness. *Sci. Rep.* **2022**, *12*, 14972. [CrossRef]
19. Song, Y.; He, X.; Yang, W.; Wu, Y.; Cui, J.; Tang, T.; Zhang, R. Virus-Specific Editing Identification Approach Reveals the Landscape of A-to-I Editing and Its Impacts on SARS-CoV-2 Characteristics and Evolution. *Nucleic Acids Res.* **2022**, *50*, 2509–2521. [CrossRef]
20. Nakata, Y.; Ode, H.; Kubota, M.; Kasahara, T.; Matsuoka, K.; Sugimoto, A.; Imahashi, M.; Yokomaku, Y.; Iwatani, Y. Cellular APOBEC3A Deaminase Drives Mutations in the SARS-CoV-2 Genome. *Nucleic Acids Res.* **2023**, *51*, 783–795. [CrossRef]
21. Saldivar-Espinoza, B.; Macip, G.; Garcia-Segura, P.; Mestres-Truyol, J.; Puigbò, P.; Cereto-Massagué, A.; Pujadas, G.; Garcia-Vallve, S. Prediction of Recurrent Mutations in SARS-CoV-2 Using Artificial Neural Networks. *Int. J. Mol. Sci.* **2022**, *23*, 14683. [CrossRef] [PubMed]
22. Harris, R.S.; Dudley, J.P. APOBECs and Virus Restriction. *Virology* **2015**, *479–480*, 131–145. [CrossRef] [PubMed]
23. Graudenzi, A.; Maspero, D.; Angaroni, F.; Piazza, R.; Ramazzotti, D. Mutational Signatures and Heterogeneous Host Response Revealed via Large-Scale Characterization of SARS-CoV-2 Genomic Diversity. *iScience* **2021**, *24*, 102116. [CrossRef] [PubMed]
24. Eisenberg, E.; Levanon, E.Y. A-to-I RNA Editing—Immune Protector and Transcriptome Diversifier. *Nat. Rev. Genet.* **2018**, *19*, 473–490. [CrossRef]
25. Vlachogiannis, N.I.; Verrou, K.-M.; Stellos, K.; Sfikakis, P.P.; Paraskevis, D. The Role of A-to-I RNA Editing in Infections by RNA Viruses: Possible Implications for SARS-CoV-2 Infection. *Clin. Immunol.* **2021**, *226*, 108699. [CrossRef]
26. Zimmermann, F.; Urban, M.; Krüger, C.; Walter, M.; Wölfel, R.; Zwirglmaier, K. In Vitro Evaluation of the Effect of Mutations in Primer Binding Sites on Detection of SARS-CoV-2 by RT-QPCR. *J. Virol. Methods* **2022**, *299*, 114352. [CrossRef]
27. Jian, M.-J.; Chung, H.-Y.; Chang, C.-K.; Lin, J.-C.; Yeh, K.-M.; Chen, C.-W.; Lin, D.-Y.; Chang, F.-Y.; Hung, K.-S.; Perng, C.-L.; et al. SARS-CoV-2 Variants with T135I Nucleocapsid Mutations May Affect Antigen Test Performance. *Int. J. Infect. Dis.* **2022**, *114*, 112–114. [CrossRef]
28. Mentes, A.; Papp, K.; Visontai, D.; Stéger, J.; VEO Technical Working Group; Csabai, I.; Medgyes-Horváth, A.; Pipek, O.A. Identification of Mutations in SARS-CoV-2 PCR Primer Regions. *Sci. Rep.* **2022**, *12*, 18651. [CrossRef]
29. Corman, V.M.; Landt, O.; Kaiser, M.; Molenkamp, R.; Meijer, A.; Chu, D.K.; Bleicker, T.; Brünink, S.; Schneider, J.; Schmidt, M.L.; et al. Detection of 2019 Novel Coronavirus (2019-NCoV) by Real-Time RT-PCR. *Eurosurveillance* **2020**, *25*, 2000045. [CrossRef]
30. Dong, H.; Wang, S.; Zhang, J.; Zhang, K.; Zhang, F.; Wang, H.; Xie, S.; Hu, W.; Gu, L. Structure-Based Primer Design Minimizes the Risk of PCR Failure Caused by SARS-CoV-2 Mutations. *Front. Cell. Infect. Microbiol.* **2021**, *11*, 741147. [CrossRef]
31. Vogels, C.B.F.; Breban, M.I.; Ott, I.M.; Alpert, T.; Petrone, M.E.; Watkins, A.E.; Kalinich, C.C.; Earnest, R.; Rothman, J.E.; Goes de Jesus, J.; et al. Multiplex QPCR Discriminates Variants of Concern to Enhance Global Surveillance of SARS-CoV-2. *PLoS Biol.* **2021**, *19*, e3001236. [CrossRef]
32. Wang, R.; Hozumi, Y.; Yin, C.; Wei, G.-W. Decoding SARS-CoV-2 Transmission and Evolution and Ramifications for COVID-19 Diagnosis, Vaccine, and Medicine. *J. Chem. Inf. Model.* **2020**, *60*, 5853–5865. [CrossRef]
33. Mercatelli, D.; Giorgi, F.M. Geographic and Genomic Distribution of SARS-CoV-2 Mutations. *Front. Microbiol.* **2020**, *11*, 1800. [CrossRef]

34. Wang, R.; Hozumi, Y.; Zheng, Y.-H.; Yin, C.; Wei, G.-W. Host Immune Response Driving SARS-CoV-2 Evolution. *Viruses* **2020**, *12*, 1095. [CrossRef]
35. Jaroszewski, L.; Iyer, M.; Alisoltani, A.; Sedova, M.; Godzik, A. The Interplay of SARS-CoV-2 Evolution and Constraints Imposed by the Structure and Functionality of Its Proteins. *PLoS Comput. Biol.* **2021**, *17*, e1009147. [CrossRef]
36. Abbasian, M.H.; Mahmanzar, M.; Rahimian, K.; Mahdavi, B.; Tokhanbigli, S.; Moradi, B.; Sisakht, M.M.; Deng, Y. Global Landscape of SARS-CoV-2 Mutations and Conserved Regions. *J. Transl. Med.* **2023**, *21*, 152. [CrossRef]
37. Khare, S.; Gurry, C.; Freitas, L.; Schultz, M.B.; Bach, G.; Diallo, A.; Akite, N.; Ho, J.; Lee, R.T.; Yeo, W.; et al. GISAID's Role in Pandemic Response. *China CDC Wkly.* **2021**, *3*, 1049–1051. [CrossRef]
38. Tian, D.; Sun, Y.; Xu, H.; Ye, Q. The Emergence and Epidemic Characteristics of the Highly Mutated SARS-CoV-2 Omicron Variant. *J. Med. Virol.* **2022**, *94*, 2376–2383. [CrossRef]
39. Focosi, D.; Quiroga, R.; McConnell, S.; Johnson, M.C.; Casadevall, A. Convergent Evolution in SARS-CoV-2 Spike Creates a Variant Soup from Which New COVID-19 Waves Emerge. *Int. J. Mol. Sci.* **2023**, *24*, 2264. [CrossRef]
40. McCarthy, K.R.; Rennick, L.J.; Nambulli, S.; Robinson-McCarthy, L.R.; Bain, W.G.; Haidar, G.; Duprex, W.P. Recurrent Deletions in the SARS-CoV-2 Spike Glycoprotein Drive Antibody Escape. *Science* **2021**, *371*, 1139–1142. [CrossRef]
41. Weng, S.; Zhou, H.; Ji, C.; Li, L.; Han, N.; Yang, R.; Shang, J.; Wu, A. Conserved Pattern and Potential Role of Recurrent Deletions in SARS-CoV-2 Evolution. *Microbiol. Spectr.* **2022**, *10*, e0219121. [CrossRef] [PubMed]
42. Akaishi, T.; Fujiwara, K. Insertion and Deletion Mutations Preserved in SARS-CoV-2 Variants. *Arch. Microbiol.* **2023**, *205*, 154. [CrossRef] [PubMed]
43. Rogozin, I.B.; Saura, A.; Bykova, A.; Brover, V.; Yurchenko, V. Deletions across the SARS-CoV-2 Genome: Molecular Mechanisms and Putative Functional Consequences of Deletions in Accessory Genes. *Microorganisms* **2023**, *11*, 229. [CrossRef] [PubMed]
44. Venkatakrishnan, A.J.; Anand, P.; Lenehan, P.J.; Ghosh, P.; Suratekar, R.; Silvert, E.; Pawlowski, C.; Siroha, A.; Chowdhury, D.R.; O'Horo, J.C.; et al. Expanding Repertoire of SARS-CoV-2 Deletion Mutations Contributes to Evolution of Highly Transmissible Variants. *Sci. Rep.* **2023**, *13*, 257. [CrossRef] [PubMed]
45. Garushyants, S.K.; Rogozin, I.B.; Koonin, E.V. Template Switching and Duplications in SARS-CoV-2 Genomes Give Rise to Insertion Variants That Merit Monitoring. *Commun. Biol.* **2021**, *4*, 1343. [CrossRef]
46. Young, B.E.; Fong, S.-W.; Chan, Y.-H.; Mak, T.-M.; Ang, L.W.; Anderson, D.E.; Lee, C.Y.-P.; Amrun, S.N.; Lee, B.; Goh, Y.S.; et al. Effects of a Major Deletion in the SARS-CoV-2 Genome on the Severity of Infection and the Inflammatory Response: An Observational Cohort Study. *Lancet* **2020**, *396*, 603–611. [CrossRef]
47. Bai, H.; Ata, G.; Sun, Q.; Rahman, S.U.; Tao, S. Natural Selection Pressure Exerted on "Silent" Mutations during the Evolution of SARS-CoV-2: Evidence from Codon Usage and RNA Structure. *Virus Res.* **2022**, *323*, 198966. [CrossRef]
48. Martínez-González, B.; Soria, M.E.; Vázquez-Sirvent, L.; Ferrer-Orta, C.; Lobo-Vega, R.; Mínguez, P.; de la Fuente, L.; Llorens, C.; Soriano, B.; Ramos-Ruíz, R.; et al. SARS-CoV-2 Mutant Spectra at Different Depth Levels Reveal an Overwhelming Abundance of Low Frequency Mutations. *Pathogens* **2022**, *11*, 662. [CrossRef]
49. Yang, H.-C.; Chen, C.-H.; Wang, J.-H.; Liao, H.-C.; Yang, C.-T.; Chen, C.-W.; Lin, Y.-C.; Kao, C.-H.; Lu, M.-Y.J.; Liao, J.C. Analysis of Genomic Distributions of SARS-CoV-2 Reveals a Dominant Strain Type with Strong Allelic Associations. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 30679–30686. [CrossRef]
50. Goldswain, H.; Dong, X.; Penrice-Randal, R.; Alruwaili, M.; Shawli, G.T.; Prince, T.; Williamson, M.K.; Raghwani, J.; Randle, N.; Jones, B.; et al. The P323L Substitution in the SARS-CoV-2 Polymerase (NSP12) Confers a Selective Advantage during Infection. *Genome Biol.* **2023**, *24*, 47. [CrossRef]
51. Jungreis, I.; Sealfon, R.; Kellis, M. SARS-CoV-2 Gene Content and COVID-19 Mutation Impact by Comparing 44 Sarbecovirus Genomes. *Nat. Commun.* **2021**, *12*, 2642. [CrossRef]
52. Berno, G.; Fabeni, L.; Matusali, G.; Gruber, C.E.M.; Rueca, M.; Giombini, E.; Garbuglia, A.R. SARS-CoV-2 Variants Identification: Overview of Molecular Existing Methods. *Pathogens* **2022**, *11*, 1058. [CrossRef]
53. Specchiarello, E.; Matusali, G.; Carletti, F.; Gruber, C.E.M.; Fabeni, L.; Minosse, C.; Giombini, E.; Rueca, M.; Maggi, F.; Amendola, A.; et al. Detection of SARS-CoV-2 Variants via Different Diagnostics Assays Based on Single-Nucleotide Polymorphism Analysis. *Diagnostics* **2023**, *13*, 1573. [CrossRef]
54. Cassari, L.; Pavan, A.; Zoia, G.; Chinellato, M.; Zeni, E.; Grinzato, A.; Rothenberger, S.; Cendron, L.; Dettin, M.; Pasquato, A. SARS-CoV-2 S Mutations: A Lesson from the Viral World to Understand How Human Furin Works. *Int. J. Mol. Sci.* **2023**, *24*, 4791. [CrossRef]
55. He, X.; He, C.; Hong, W.; Yang, J.; Wei, X. Research Progress in Spike Mutations of SARS-CoV-2 Variants and Vaccine Development. *Med. Res. Rev.* **2023**, *in press*. [CrossRef]
56. Liu, X.; Liu, X.; Zhou, J.; Dong, Y.; Jiang, W.; Jiang, W. Rampant C-to-U Deamination Accounts for the Intrinsically High Mutation Rate in SARS-CoV-2 Spike Gene. *RNA* **2022**, *28*, 917–926. [CrossRef]
57. Ravi, V.; Swaminathan, A.; Yadav, S.; Arya, H.; Pandey, R. SARS-CoV-2 Variants of Concern and Variations within Their Genome Architecture: Does Nucleotide Distribution and Mutation Rate Alter the Functionality and Evolution of the Virus? *Viruses* **2022**, *14*, 2499. [CrossRef]
58. Oronsky, B.; Larson, C.; Caroen, S.; Hedjran, F.; Sanchez, A.; Prokopenko, E.; Reid, T. Nucleocapsid as a Next-Generation COVID-19 Vaccine Candidate. *Int. J. Infect. Dis.* **2022**, *122*, 529–530. [CrossRef]

59. Saldivar-Espinoza, B.; Macip, G.; Pujadas, G.; Garcia-Vallve, S. Could Nucleocapsid Be a Next-Generation COVID-19 Vaccine Candidate? *Int. J. Infect. Dis.* **2022**, *125*, 231–232. [CrossRef]
60. Zhao, H.; Nguyen, A.; Wu, D.; Li, Y.; Hassan, S.A.; Chen, J.; Shroff, H.; Piszczek, G.; Schuck, P. Plasticity in Structure and Assembly of SARS-CoV-2 Nucleocapsid Protein. *PNAS Nexus* **2022**, *1*, pgac049. [CrossRef]
61. Zhao, H.; Wu, D.; Hassan, S.A.; Nguyen, A.; Chen, J.; Piszczek, G.; Schuck, P. A Conserved Oligomerization Domain in the Disordered Linker of Coronavirus Nucleocapsid Proteins. *Sci. Adv.* **2023**, *9*, eadg6473. [CrossRef] [PubMed]
62. De Maio, N.; Walker, C.R.; Turakhia, Y.; Lanfear, R.; Corbett-Detig, R.; Goldman, N. Mutation Rates and Selection on Synonymous Mutations in SARS-CoV-2. *Genome Biol. Evol.* **2021**, *13*, evab087. [CrossRef] [PubMed]
63. Li, Y.; Hou, F.; Zhou, M.; Yang, X.; Yin, B.; Jiang, W.; Xu, H. C-to-U RNA Deamination Is the Driving Force Accelerating SARS-CoV-2 Evolution. *Life Sci. Alliance* **2023**, *6*, e202201688. [CrossRef] [PubMed]
64. Ringlander, J.; Fingal, J.; Kann, H.; Prakash, K.; Rydell, G.; Andersson, M.; Martner, A.; Lindh, M.; Horal, P.; Hellstrand, K.; et al. Impact of ADAR-Induced Editing of Minor Viral RNA Populations on Replication and Transmission of SARS-CoV-2. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2112663119. [CrossRef]
65. Ziegler, K.; Steininger, P.; Ziegler, R.; Steinmann, J.; Korn, K.; Ensser, A. SARS-CoV-2 Samples May Escape Detection Because of a Single Point Mutation in the N Gene. *Eurosurveillance* **2020**, *25*, 2001650. [CrossRef]
66. Vanaerschot, M.; Mann, S.A.; Webber, J.T.; Kamm, J.; Bell, S.M.; Bell, J.; Hong, S.N.; Nguyen, M.P.; Chan, L.Y.; Bhatt, K.D.; et al. Identification of a Polymorphism in the N Gene of SARS-CoV-2 That Adversely Impacts Detection by Reverse Transcription-PCR. *J. Clin. Microbiol.* **2020**, *59*, e02369-20. [CrossRef]
67. Hasan, R.; Hossain, M.E.; Miah, M.; Hasan, M.M.; Rahman, M.; Rahman, M.Z. Identification of Novel Mutations in the N Gene of SARS-CoV-2 That Adversely Affect the Detection of the Virus by Reverse Transcription-Quantitative PCR. *Microbiol. Spectr.* **2021**, *9*, e0054521. [CrossRef]
68. Zannoli, S.; Dirani, G.; Taddei, F.; Gatti, G.; Poggianti, I.; Denicolò, A.; Arfilli, V.; Manera, M.; Mancini, A.; Battisti, A.; et al. A Deletion in the N Gene May Cause Diagnostic Escape in SARS-CoV-2 Samples. *Diagn. Microbiol. Infect. Dis.* **2022**, *102*, 115540. [CrossRef]
69. Laine, P.; Nihtilä, H.; Mustanoja, E.; Lyyski, A.; Ylinen, A.; Hurme, J.; Paulin, L.; Jokiranta, S.; Auvinen, P.; Meri, T. SARS-CoV-2 Variant with Mutations in N Gene Affecting Detection by Widely Used PCR Primers. *J. Med. Virol.* **2022**, *94*, 1227–1231. [CrossRef]
70. Miller, S.; Lee, T.; Merritt, A.; Pryce, T.; Levy, A.; Speers, D. Single-Point Mutations in the N Gene of SARS-CoV-2 Adversely Impact Detection by a Commercial Dual Target Diagnostic Assay. *Microbiol. Spectr.* **2021**, *9*, e0149421. [CrossRef]
71. Isabel, S.; Abdulnoor, M.; Boissinot, K.; Isabel, M.R.; de Borja, R.; Zuzarte, P.C.; Sjaarda, C.P.; Barker, R.K.; Sheth, P.M.; Matukas, L.M.; et al. Emergence of a Mutation in the Nucleocapsid Gene of SARS-CoV-2 Interferes with PCR Detection in Canada. *Sci. Rep.* **2022**, *12*, 10867. [CrossRef]
72. Kami, W.; Kinjo, T.; Hashioka, H.; Arakaki, W.; Uechi, K.; Takahashi, A.; Oki, H.; Tanaka, K.; Motooka, D.; Nakamura, S.; et al. Impact of G29179T Mutation on Two Commercial PCR Assays for SARS-CoV-2 Detection. *J. Virol. Methods* **2023**, *314*, 114692. [CrossRef]
73. Wang, R.; Hozumi, Y.; Yin, C.; Wei, G.-W. Mutations on COVID-19 Diagnostic Targets. *Genomics* **2020**, *112*, 5204–5213. [CrossRef]
74. Marchini, A.; Petrillo, M.; Parrish, A.; Buttinger, G.; Tavazzi, S.; Querci, M.; Betsou, F.; Elsinga, G.; Medema, G.; Abdelrahman, T.; et al. New RT-PCR Assay for the Detection of Current and Future SARS-CoV-2 Variants. *Viruses* **2023**, *15*, 206. [CrossRef]
75. Rambaut, A.; Holmes, E.C.; O'Toole, Á.; Hill, V.; McCrone, J.T.; Ruis, C.; du Plessis, L.; Pybus, O.G. A Dynamic Nomenclature Proposal for SARS-CoV-2 Lineages to Assist Genomic Epidemiology. *Nat. Microbiol.* **2020**, *5*, 1403–1407. [CrossRef]

Supplementary Information for
The mutational landscape of SARS-CoV-2

Bryan Saldivar-Espinoza[1,†], Pol Garcia-Segura[1], Nil Novau-Ferré[1,†], Guillem Macip[1], Ruben Martinez[2], Pere Puigbò [3,4,5], Adrià Cereto-Massagué[6], Gerard Pujadas[1,*] and Santiago Garcia-Vallve[1,*]

[1] Departament de Bioquímica i Biotecnologia, Research group in Cheminformatics & Nutrition, Campus de Sescelades, Universitat Rovira i Virgili, 43007 Tarragona, Spain; bsaldivar.emc2@gmail.com (B.S.-E.); polgarse2@gmail.com (P.G.-S.); nnovauf@gmail.com (N.N.-F.); guillem.macip@gmail.com (G.M.)

[2] Institut La Guineueta. 08042 Barcelona, Spain; rmartbernabe@gmail.com

[3] Department of Biology, University of Turku, 20500 Turku, Finland; pepuav@utu.fi

[4] Department of Biochemistry and Biotechnology, Rovira i Virgili University, 43007 Tarragona, Spain

[5] Eurecat Technology Centre of Catalonia, Unit of Nutrition and Health, 43204 Reus, Spain

[6] EURECAT Centre Tecnològic de Catalunya, Centre for Omic Sciences (COS), Joint Unit Universitat Rovira i Virgili-EURECAT, Unique Scientific and Technical Infrastructures (ICTS), 43204 Reus, Spain. ssorgatem@gmail.com

This PDF file includes:

Figures S1 to S14
Tables S1 to S3

Figure S1. Bar chart of the percentage of analyzed genomes sequenced for each country.

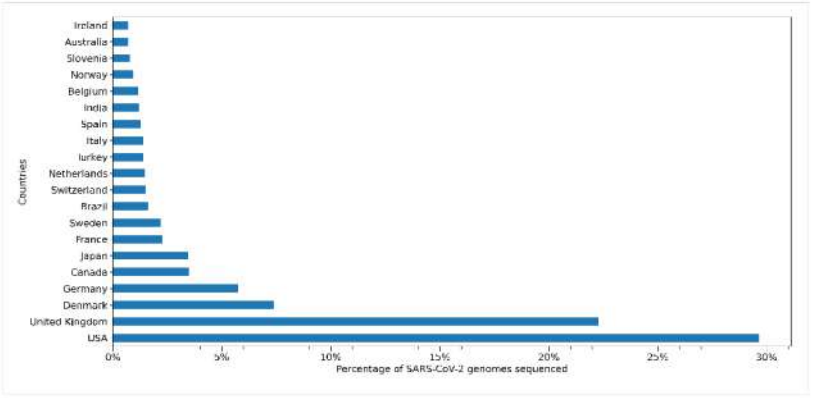


Figure S2. Mean number of SNVs per genome and week.

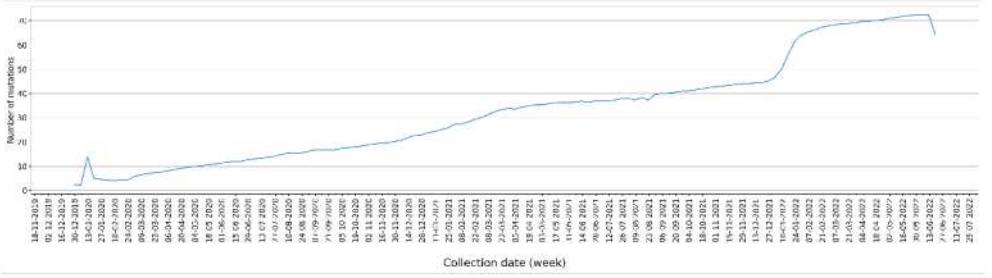


Figure S3. Mean number of deletions per genome and week.

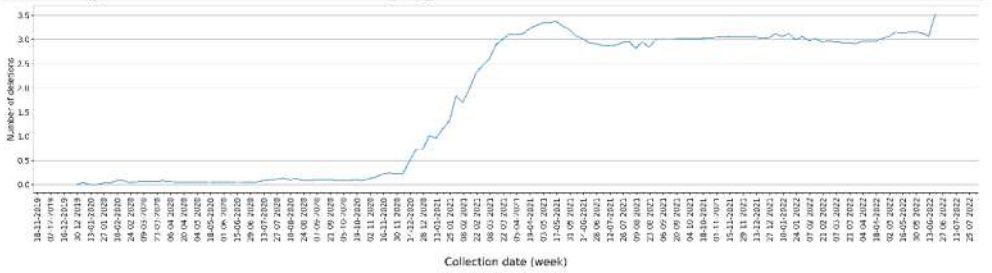


Figure S4. Mean number of insertions per genome and week.

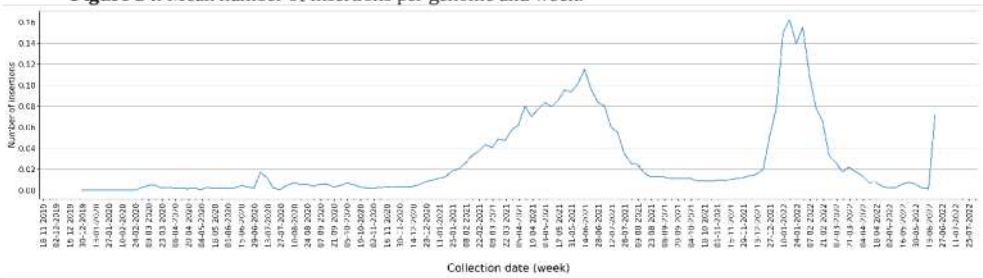


Figure S5. Histogram of the length of deletions in SARS-CoV-2 genomes. Only deletions with a relative frequency greater than 10% were analyzed.

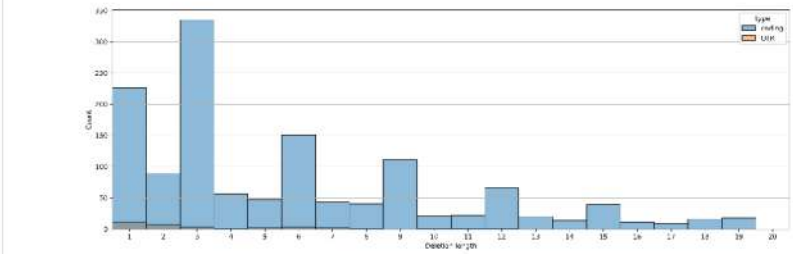


Figure S6. Histogram of the length of insertions in SARS-CoV-2 genomes. Only deletions with a relative frequency greater than 10% were analyzed.

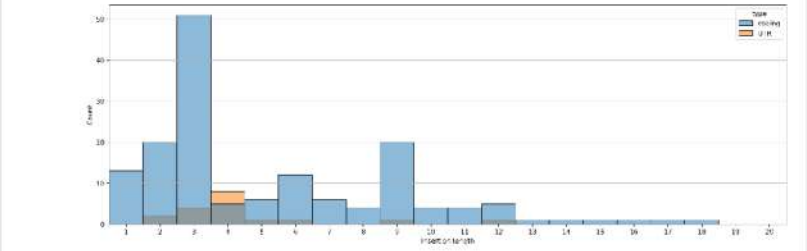


Figure S7. Box plots of the relative frequency of SNV types.

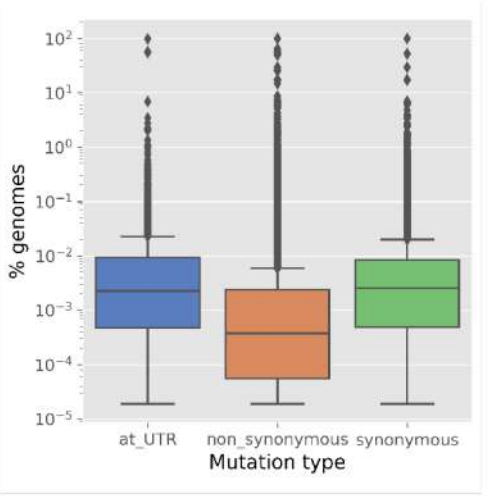
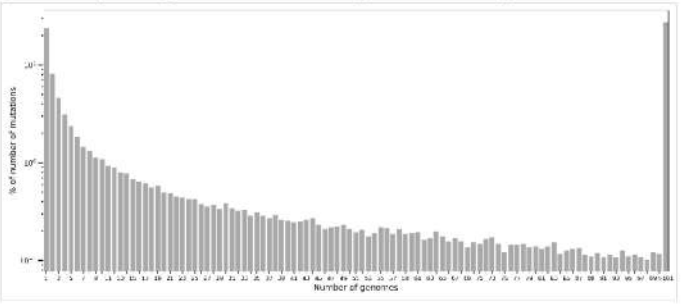
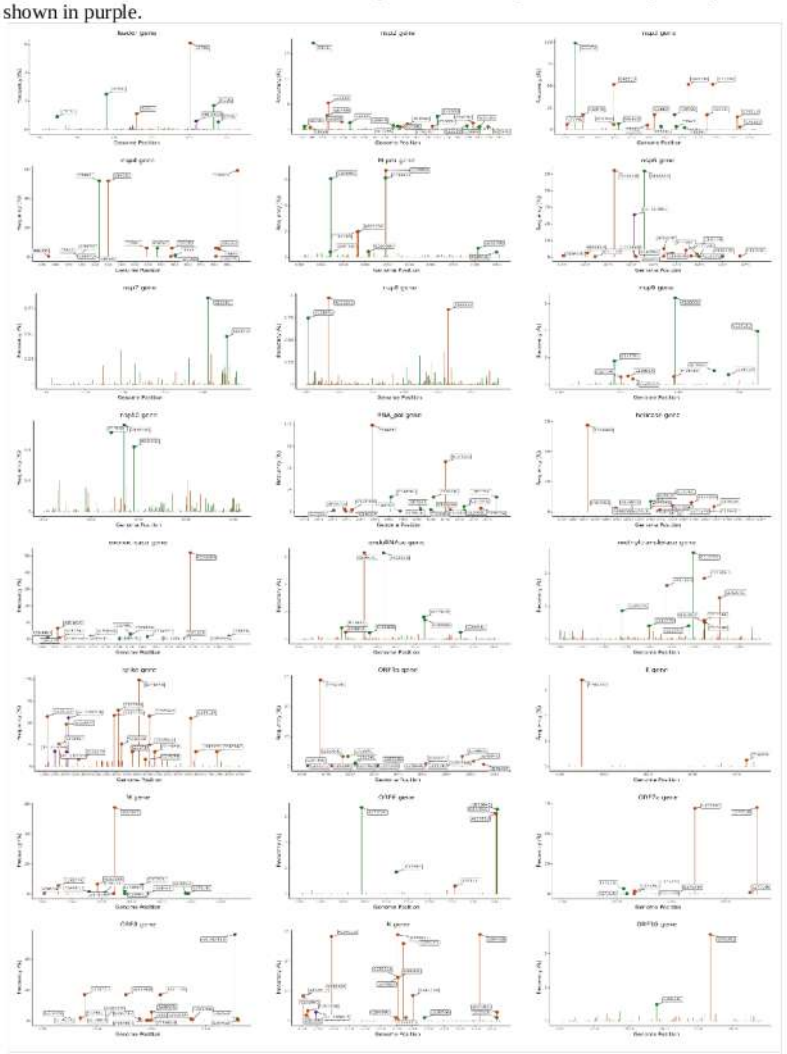


Figure S8. Histogram of the percentage of SNVs found in a given number of genomes.



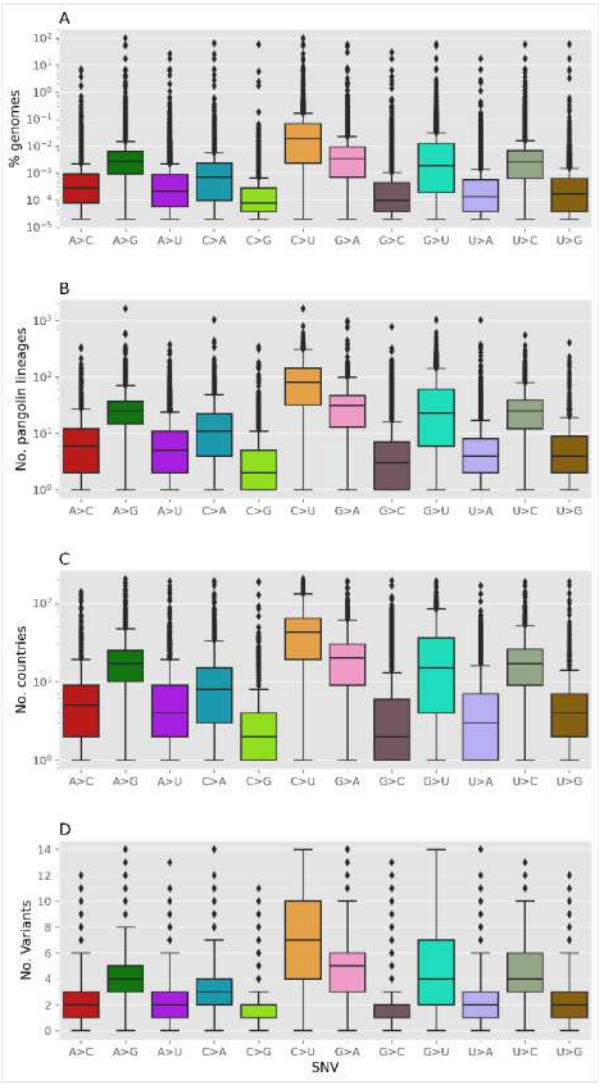
4

Figure S9. Lolliplots of the most frequent mutations in each SARS-CoV-2 gene. Synonymous, non-synonymous and UTR mutations are shown in green, dark orange and blue, respectively. Deletions are shown in purple.



5

Figure S10. Box plots of the frequency (A), number of pangolin lineages (B), number of countries (C) and number of variants (D) of SNV types.



6

Figure S11. Box plots of the frequency (%) of the mutations found in SARS-CoV-2 regions that hybridize with probes and forward and reverse primers from some COVID-19 diagnostic PCR tests.

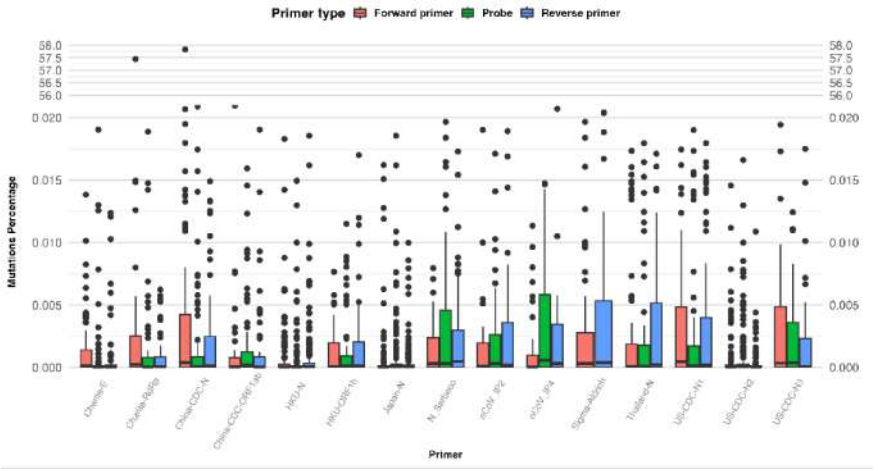
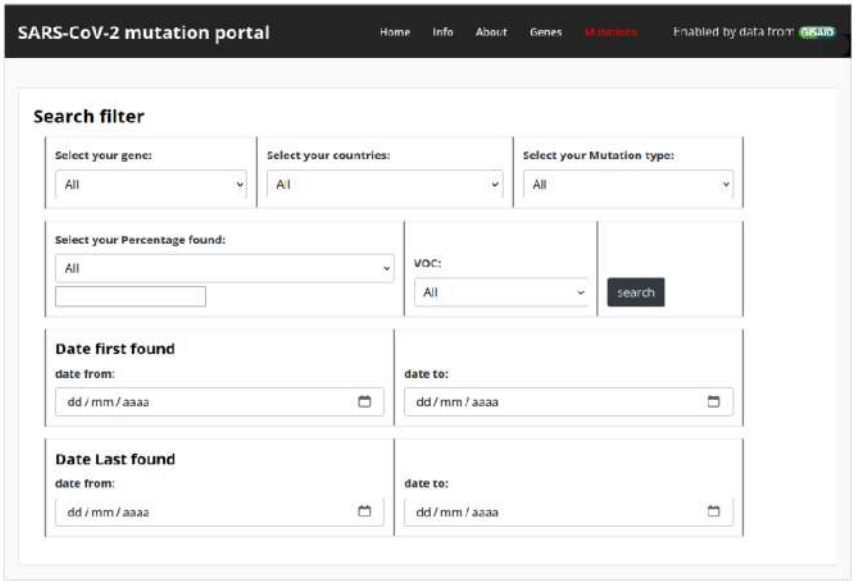


Figure S12. Search engine of the SARS-CoV-2 mutation portal at http://sarscov2-mutation-portal.urv.cat/



8

Figure S13. Results of a search of the SARS-CoV-2 mutation portal at http:// sarscov2-mutation-portal.urv.cat/

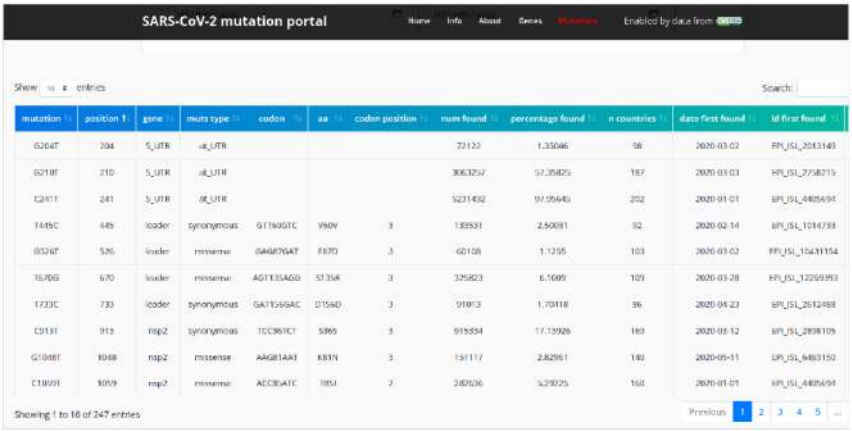


Figure S14. Scatter plot of the above search

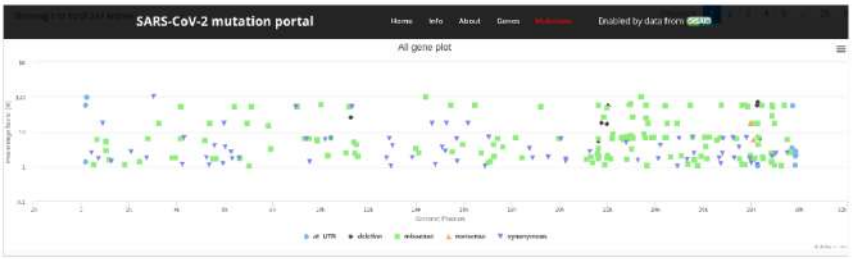


Table S1. Distribution by continent of the 5,340,569 SARS-CoV-2 genomes analyzed.

Continent	Genomes sequenced	Percentage (%)
Europe	2,942,990	55.1
North America	1,819,981	34.1
Asia	371,704	7.0
South America	126,872	2.4
Oceania	44,942	0.8
Africa	34,080	0.6

Table S2. Number of different SNVs, insertions and deletions found in the 5,340,569 SARS-CoV-2 genomes analyzed.

Mutation	Total number[1]	Coding regions	UTRs
SNVs	73,464	71,622	1842
deletions	21,712	21,464	248
insertions	1820	1700	120

[1]Number of different SNV, deletions or insertions

Table S3. Mutations found in SARS-CoV-2 regions that hybridize with probes and forward and reverse primers from some COVID-19 diagnostic PCR tests.

Name	Gene	Forward primer position	Forward primer count[2]	Forward primer %[1]	Forward primer 5 last count[1,3]	Forward primer 5 last %[1,3]	Reverse primer position	Reverse primer count[2]	Reverse primer %[1]	Reverse primer 5 last count[1,3]	Reverse primer 5 last %[1,3]	Probe position	Probe count[2]	Probe %[1]	Total count	Total %[1]
nCoV_IP2	RdRp	12,690-12,707	46 (5|0)	0.07	14 (1|0)	0.039	12,780-12,797	68 (13|2)	1.25	18 (5|1)	0.14	12,717-12,737	64 (12|0)	0.44	178	1.75
nCoV_IP4	RdRp	14,080-14,098	50 (10|0)	0.12	13 (3|0)	0.068	14,167-14,186	66 (6|0)	0.85	15 (2|0)	0.038	14,105-14,123	65 (4|1)	2.97	181	3.94
Charite-E	E	26,269-26,294	89 (16|0)	6.91	15 (5|0)	0.018	26,360-26,381	81 (22|1)	0.10	30 (15|1)	0.003	26,332-26,357	143 (32|0)	0.14	313	7.15
N_Sarbeco	N	28,706-28,724	68 (8|0)	0.87	22 (5|0)	0.23	28,814-28,833	93 (18|0)	0.75	34 (14|0)	0.17	28,753-28,777	116 (23|3)	0.52	277	2.14
Charite-RdRp	RdRp	15,431-15,452	67 (10|0)	57.85	24 (5|0)	57.57	15,505-15,528	52 (9|0)	2.76	15 (6|0)	0.01	15,470-15,494	64 (15|0)	0.23	183	60.84
HKU-ORF1b	ORF1ab	18,778-18,797	60 (3|0)	0.31	11 (0|0)	0.07	18,889-18,909	73 (13|0)	0.70	19 (3|0)	0.14	18,849-18,872	60 (5|0)	0.17	193	1.18
HKU-N	N	29,145-29,166	145 (79|0)	0.59	93 (77|0)	0.08	29,236-29,254	222 (151|0)	1.95	154 (136|0)	0.16	29,177-29,196	167 (104|1)	0.71	534	3.25
China-CDC-ORF1ab	ORF1ab	13,342-13,362	58 (11|1)	0.27	12 (4|0)	0.009	13,442-13,460	59 (13|0)	0.24	16 (6|0)	0.04	13,377-13,404	103 (21|1)	0.29	220	0.79
China-CDC-N	N	28,881-28,902	156 (35|3)	120.51	48 (23|2)	0.26	28,958-28,979	118 (23|4)	20.42	30 (13|0)	0.56	28,934-28,953	86 (20|2)	0.35	360	141.28
US-CDC-N1	N	28,287-28,306	102 (15|4)	5.29	26 (7|0)	0.06	28,335-28,358	111 (23|2)	0.56	29 (10|1)	0.11	28,309-28,332	131 (25|2)	8.75	344	14.59
US-CDC-N2	N	29,164-29,183	154 (92|1)	1.11	104 (90|1)	0.50	29,213-29,230	184 (131|0)	0.66	129 (117|0)	0.13	29,188-29,210	189 (115|0)	0.95	527	2.73
US-CDC-N3	N	28,681-28,702	88 (14|0)	0.96	22 (7|0)	0.11	28,732-28,752	91 (17|2)	0.75	25 (8|0)	0.15	28,704-28,727	90 (10|0)	1.64	269	3.36
Japan-N	N	29,125-29,144	116 (63|0)	0.52	77 (60|0)	0.35	29,263-29,282	234 (173|0)	0.92	174 (159|0)	0.24	29,222-29,241	211 (140|0)	0.59	561	2.02
Thailand-N	N	28,320-28,339	104 (21|2)	1.25	27 (8|1)	0.09	28,358-28,376	112 (29|0)	0.82	35 (17|0)	0.17	28,341-28,356	78 (18|1)	0.39	294	2.46
Sigma-Aldrich	N	28,750-28,771	96 (18|1)	0.39	26 (11|0)	0.10	28,842-28,860	96 (17|2)	2.27	30 (8|0)	0.14	-	0	0	192	2.66

[1] Total accumulated frequencies in %. [2] The numbers in brackets indicate the number of deletions and insertions. [3]Only the last 5 bases of the 3'-end have been taken into account

Manuscript 3

Prediction of Recurrent Mutations in SARS-CoV-2 Using Artificial Neural Networks

Bryan Saldivar-Espinoza (1), Guillem Macip (1), Pol Garcia-Segura (1), Júlia Mestres-Truyol (1), Pere Puigbò (2,3,4), Adrià Cereto-Massagué (5), Gerard Pujadas (1) and Santiago Garcia-Vallve 1,*

1 Research Group in Cheminformatics & Nutrition, Departament de Bioquímica i Biotecnologia, Campus de Sescelades, Universitat Rovira i Virgili, 43007 Tarragona, Spain

2 Department of Biology, University of Turku, 20500 Turku, Finland

3 Department of Biochemistry and Biotechnology, Rovira i Virgili University, 43007 Tarragona, Spain

4 Nutrition and Health Unit, Eurecat Technology Centre of Catalonia, 43204 Reus, Spain

5 EURECAT Centre Tecnològic de Catalunya, Centre for Omic Sciences (COS), Joint Unit Universitat Rovira i Virgili-EURECAT, Unique Scientific and Technical Infrastructures (ICTS), 43204 Reus, Spain

* Author to whom correspondence should be addressed. (santi.garcia-vallve@urv.cat)

International Journal of *Molecular Sciences*

MDPI

Article

Prediction of Recurrent Mutations in SARS-CoV-2 Using Artificial Neural Networks

Bryan Saldivar-Espinoza [1], Guillem Macip [1], Pol Garcia-Segura [1], Júlia Mestres-Truyol [1], Pere Puigbò [2,3,4], Adrià Cereto-Massagué [5], Gerard Pujadas [1] and Santiago Garcia-Vallve [1,*]

[1] Research Group in Cheminformatics & Nutrition, Departament de Bioquímica i Biotecnologia, Campus de Sescelades, Universitat Rovira i Virgili, 43007 Tarragona, Spain
[2] Department of Biology, University of Turku, 20500 Turku, Finland
[3] Department of Biochemistry and Biotechnology, Rovira i Virgili University, 43007 Tarragona, Spain
[4] Nutrition and Health Unit, Eurecat Technology Centre of Catalonia, 43204 Reus, Spain
[5] EURECAT Centre Tecnològic de Catalunya, Centre for Omic Sciences (COS), Joint Unit Universitat Rovira i Virgili-EURECAT, Unique Scientific and Technical Infrastructures (ICTS), 43204 Reus, Spain
* Correspondence: santi.garcia-vallve@urv.cat

Abstract: Predicting SARS-CoV-2 mutations is difficult, but predicting recurrent mutations driven by the host, such as those caused by host deaminases, is feasible. We used machine learning to predict which positions from the SARS-CoV-2 genome will hold a recurrent mutation and which mutations will be the most recurrent. We used data from April 2021 that we separated into three sets: a training set, a validation set, and an independent test set. For the test set, we obtained a specificity value of 0.69, a sensitivity value of 0.79, and an Area Under the Curve (AUC) of 0.8, showing that the prediction of recurrent SARS-CoV-2 mutations is feasible. Subsequently, we compared our predictions with updated data from January 2022, showing that some of the false positives in our prediction model become true positives later on. The most important variables detected by the model's Shapley Additive exPlanation (SHAP) are the nucleotide that mutates and RNA reactivity. This is consistent with the SARS-CoV-2 mutational bias pattern and the preference of some host deaminases for specific sequences and RNA secondary structures. We extend our investigation by analyzing the mutations from the variants of concern Alpha, Beta, Delta, Gamma, and Omicron. Finally, we analyzed amino acid changes by looking at the predicted recurrent mutations in the M-pro and spike proteins.

Keywords: SARS-CoV-2; COVID-19; machine learning; mutations

Citation: Saldivar-Espinoza, B.; Macip, G.; Garcia-Segura, P.; Mestres-Truyol, J.; Puigbò, P.; Cereto-Massagué, A.; Pujadas, G.; Garcia-Vallve, S. Prediction of Recurrent Mutations in SARS-CoV-2 Using Artificial Neural Networks. *Int. J. Mol. Sci.* **2022**, *23*, 14683. https://doi.org/10.3390/ijms232314683

Academic Editor: João R. Mesquita

Received: 27 October 2022
Accepted: 22 November 2022
Published: 24 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

SARS-CoV-2 is the coronavirus that causes COVID-19. It has a positive sense single-stranded RNA (ssRNA) genome of around 29,900 nucleotides that codifies 11 genes: ORF1ab, spike (S), ORF3a, envelope (E), membrane (M), ORF6, ORF7a, ORF7b, ORF8, nucleocapsid (N), and ORF10 [1,2]. The ORF1ab gene encodes the polyproteins pp1a and pp1ab, which are further cleaved by the main protease (M-pro) and papain-like protease (PLpro) [3]. Pp1ab includes pp1a, and its synthesis requires a ribosomal frameshift [3]. When pp1ab cleaves, it gives rise to 15 proteins: a lead protein, nsp2, nsp3 (PLpro), nsp4, nsp5 (M-pro), nsp6, nsp7, nsp8, nsp9, nsp10, nsp12 (an RNA-dependent RNA polymerase, RdRp), nsp13 (a helicase), nsp14 (a 3′-5′ exonuclease), nsp15 (an endoRNAse) and nsp16 (a 2′-O-ribose methyltransferase) [3].

Like other viruses, the SARS-CoV-2 genome mutates. Mutations can lead to enhanced viral fitness and the emergence of virus variants [4]. However, recombination and re-assortment are also important mechanisms to generate genomic variability [5]. Virus mutation rates vary widely [6], but coronaviruses have a proofreading activity (due to the nsp14 gene) [7] that may explain their abnormally large genome compared to other ssRNA viruses [5]. Mutations can be caused by RNA polymerase errors during virus replication or

by the deamination of unpaired nitrogenous bases caused by host deaminases [6,8–10]. In mammalian species, apolipoprotein B mRNA editing catalytic polypeptide-like (APOBEC) enzymes deaminate cytosines into uracils (C > U) in single-stranded DNA (ssDNA) and ssRNA [11]. Recent experiments have demonstrated that APOBEC3A, APOBEC1, and APOBEC3G can effectively cause C > U mutations in the SARS-CoV-2 genome at specific sites [12]. Cytosines in UC and AC motifs showed the highest mutation rate, modulated by features of the RNA structure around these motifs [12]. This is consistent with previous results [13,14]. For example, the 5′-[U | A]C>U mutation occurs more frequently than 5′-[C | G]C > U ($p = 0.0501$) in the SARS-CoV-2 genome [14]. If APOBEC enzymes were to act on the negative strand of the SARS-CoV-2 genome, it would be reflected on the positive strand as G>A mutations [15]. Adenosine deaminases acting on RNA (ADAR) deaminate adenines into inosines (A > I) in double-stranded RNA (dsRNA) [16]. As inosine preferentially base pairs with cytidine, A > I mutations cause A > G and U > C transitions on the positive strand of the SARS-CoV-2 genome [15,17]. Most SARS-CoV-2 mutations are expected to be neutral, but some may be advantageous or deleterious to the virus [18]. Viruses experience selection pressure from their host's immune system, defense mechanisms, antiviral drugs, and vaccines [5]. Highly deleterious mutations, such as those that prevent the virus from invading the host, are unlikely to be observed [18]. The high frequency of some mutations is not always due to an advantageous mutation. It can also be caused by a founder effect, which is when a mutation emerges early in the evolution of a pandemic and is transmitted to all of its descendants [19] or when they are found in a variant that carries an additional advantageous mutation.

During the COVID-19 pandemic, the number of new SARS-CoV-2 variants, including the variants of concern (VoCs), has steadily increased [20,21]. VoCs are variants that exhibit increased transmissibility; more severe disease; significantly decreased neutralization by antibodies developed from previous infection or immunization; reduced efficacy of therapies or vaccines, or failures of detection at diagnosis [22]. Therefore, it is very important to understand the mutational patterns in the evolution of SARS-CoV-2 and to predict its mutations in order to devise better antiviral treatments [23]. Due to the random nature of these mutations, predicting SARS-CoV-2 mutations caused by replication errors can be difficult. However, it is feasible to predict mutations driven by the host, such as those caused by host deaminases [12]. These mutations are expected to be recurrent, i.e., to appear multiple times independently and be present in several SARS-CoV-2 lineages. In this paper, we use machine learning (ML) to predict recurrent mutations that will emerge repeatedly and independently as the virus adapts to humans [18,24]. Before the pandemic, ML was used extensively in biology [25–28], for example, to predict mutations of influenza A viruses by predicting which AA position will mutate [29] and to predict recurrent mutations in cancer [30]. ML has been used throughout the SARS-CoV-2 pandemic as a tool to assist vaccine development and predict epitope hotspots [31]; the binding affinity of antibodies to mutations in the spike RBD [32]; the binding affinity of chemical compounds as inhibitors against the M-pro protein [33,34]; the clinical disease severity based on the virus genome mutations [35]; the mutation rate of nucleotide substitution (e.g., A > T) [36]; the subsequent nucleotide given a sequence of the SARS-CoV-2 genome, and also given a pair of sequences to indicate the location of the changes [37]; the antibody escape mutations of the spike protein [38]; the spread of spike protein mutation, based on fold-change per country [39]; future domain-specific spike mutations [40]; anti-SARS-CoV-2 activities from molecular structure [41]; and many more [42,43]. In this article, we start by showing some descriptive statistics of SARS-CoV-2 mutations. We continue by defining recurrent mutations. We then use ML models to predict which positions of the genome will have a recurrent mutation, showing the performance metrics of the models and variables that are more important for the ML models. Subsequently, we extend our investigation to predict which mutations will become recurrent and how our work can be used with the variants of concern Alpha, Beta, Delta, Gamma, and Omicron. Finally, we analyze amino acid changes by looking at the

predicted recurrent mutations in the M-pro and spike proteins, evaluated with recent data from 2022.

2. Results and Discussion

2.1. SARS-CoV-2 Mutation Description

The GISAID database [44] had 877,086 SARS-CoV-2 genomes as of 19 April 2021. From these genomes, we found 25,353,899 mutations (including insertions and deletions), of which 52,160 were unique single nucleotide variants (SNVs) found in one or more genomes. Among the unique SNVs, adenine and uracil were the nucleotides with the most SNVs, 15,898 and 15,313, respectively (Figure 1A). Because the SARS-CoV-2 genome is richer in adenines and uracils (its G + C content is 37.97%), in SNVs, it is expected to find more adenines and uracils than guanines and cytosines. Transitions, i.e., U > C, A > G, C > U, and G > A, are more frequent than transversions (Figure 1B). When uracil mutates, 51% of the time, it mutates into a cytosine, with similar percentages found regarding other transitions (Figure 1B), with the exception of G > A, which has a slightly lower frequency (Figure 1B). C > G and G > C transversions are observed with lower frequencies (Figure 1B). Only 14% and 17% of cytosine and guanine SNVs are, respectively, C > G and G > C transversions. This may reflect the CpG avoidance that has been described for SARS-CoV-2 and other coronaviruses [45,46]. This CpG dinucleotide suppression is thought to be due to the fact that it is evading the zinc-finger antiviral protein (ZAP) that specifically binds CpG dinucleotides in ssRNA and causes its degradation [47,48].

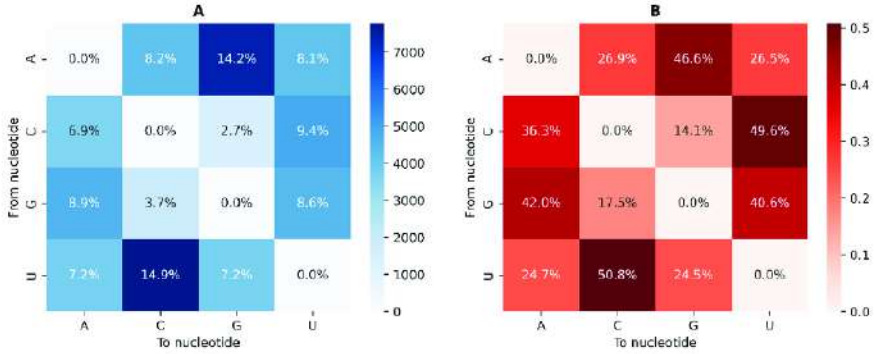


Figure 1. Nucleotide change count of unique single nucleotide mutations. (**A**) Total count of changes among unique mutations. (**B**) Nucleotide change normalized through the horizontal axis. Each row adds up to 1.

2.2. Recurrent Mutations in SARS-CoV-2

Recurrent mutations (RM) are mutations that occur independently and many times throughout a virus' evolution. They could be the result of host RNA-editing mechanisms or ongoing selection [18,24]. After analyzing 46,723 and 7710 SARS-CoV-2 genomes from July and April 2020, van Dorp et al. [18,24] identified 5710 and 198 RM, respectively. Among the RM, they found no evidence for increased transmissibility, suggesting that RMs were caused by RNA editing [18]. To identify RMs, van Drop et al. [18,24] used a multiple alignment and a maximum likelihood tree. However, due to the large number of analyzed sequences, we used another strategy. We used Pango nomenclature that classifies SARS-CoV-2 genomes into lineages [49,50]. The Pango nomenclature is a hierarchical and dynamic classification system based on phylogenetic evidence that uses ML to assign each SARS-CoV-2 genome to a lineage [49,50]. Although this system is not intended to represent every evolutionary change in SARS-CoV-2 [49], we have used it to count the

number of different lineages in which each mutation is found. Taking advantage of the hierarchical nature of the Pango system, we then reduced this number by grouping related lineages and counting the number of distantly-related lineages (NDRL) (see Materials and Methods). This provided us with an estimate of the number of times a mutation emerged independently. Then, we defined RM for a set of NDRL thresholds of 5, 10, and 15. We used different NDRL thresholds to overcome potential sequencing errors, artefactual biases, and other causes, such as recombination, which may lead to homoplasies [14,24]. We found 22,738, 11,275, and 6767 RM for the 5, 10, and 15 NDRL thresholds. Dataset S1 contains all the mutations found and the number of NDRLs for each mutation. Mutations present in almost all Pango lineages that appeared early in the pandemic, such as the A23403G mutation that results in the D614G substitution of the spike protein, are not considered to be RM because they have an NDRL value of 1. As expected and based on previous work [14,18,46,51,52], RMs are rich in C > U mutations (Figure S1). For instance, for the NDRL threshold of 15, 47% of the 6767 mutations are C > U, while 19% are G > U. U > C, G > A, and A > G each constitute 10% (Figure S1).

2.3. Prediction of Whether a Given Position in the SARS-CoV-2 Genome Will Be Affected by a Recurrent Mutation

To predict whether a given position in the SARS-CoV-2 genome holds an RM, as defined by the NDRL thresholds of 1, 5, 10, and 15, a deep learning/machine learning model was trained using the artificial neural network/multi-layer perceptron architecture. The variables used to train the models were the SARS-CoV-2 genome sequence, the prediction of the secondary structure of the SARS-CoV-2 genome, the RNA normalized 2′-hydroxyl acylation analyzed by primer extension (SHAPE) reactivity [53], and the translated AA sequences of the coding parts of the genome. The genome variables were split into 13 position windows, with the central window position indicating the location of the possible mutation. The data split for the machine learning setup included a group of 16 genes for training and four different genes for validation (Figure S2). To evaluate the model predictions, a separate test set was used. The test set was not used at any moment during training or model tuning. Given their relevance, the M-pro, spike, PLpro, and RNA polymerase genes were included in the test set [54–56] (Figure S2). The genes in the validation and test sets were chosen in order to have a similar number of mutations per nucleotide between them (Figures S2 and S3).

We decided to prioritize sensitivity (true positive rate) over specificity (true negative rate) in choosing the best prediction model. We chose the model that achieved the highest specificity with a minimum sensitivity of 0.85 in the validation set. As the NDLR threshold increased, the performance of the trained model on the testing, validation, and training set improved (Figure 2). This is shown by the increase in the area under the curve (AUC) of the receiver operating characteristic (ROC) curve. The AUC values for the training set were between 0.82 and 0.87 and, as expected, were higher than the values for the validation and testing sets. Interestingly, the values for the validation and testing sets were similar. The best AUC for the testing set was achieved for the NDRL threshold of 15, with an AUC of 0.81. This shows that it is possible to predict the position of recurrent SARS-CoV-2 mutations. When analyzing the model's performance on the test set genes separately for the four genes included in this set (M-pro, spike, PLpro, and RNA polymerase), for the NDRL threshold of 15, the prediction is worse for the spike gene (with an AUC value of 0.77) (Figure S4). This is not uncommon, as mutations in the spike gene can have a high impact on the infectious power of the virus and these mutations are the most difficult to predict.

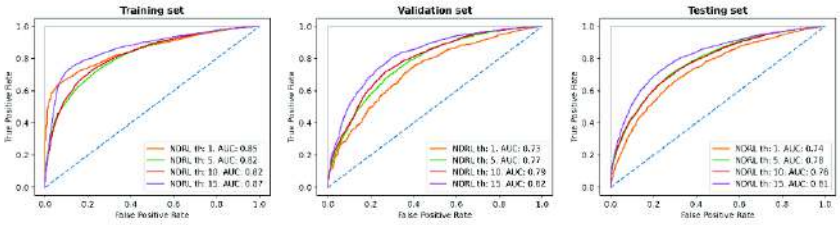


Figure 2. Receiver operating characteristic (ROC) curve for the testing, validation, and training set using 1, 5, 10, and 15 as thresholds for the NDRL. The blue dashed diagonal line represents how a random model would behave.

Figure 3 shows the sensitivity, specificity, and confusion matrix of the test set across the four NDRL thresholds. The four predictive models showed similar sensitivity values, but as the NDRL thresholds increased, specificity also increased from 0.46 to 0.69. Confusion matrices show that when the degree of RM is low, 1 or 5 NDRL, more positions in the SARS-CoV-2 genome have an RM. In this case, predictive models perform well for predicting true positive cases but perform worse for predicting true negative cases. When the NDRL threshold increases, the number of RM decreases, but predictive models are able to predict reasonably well the positions in the genome that do or do not have an RM.

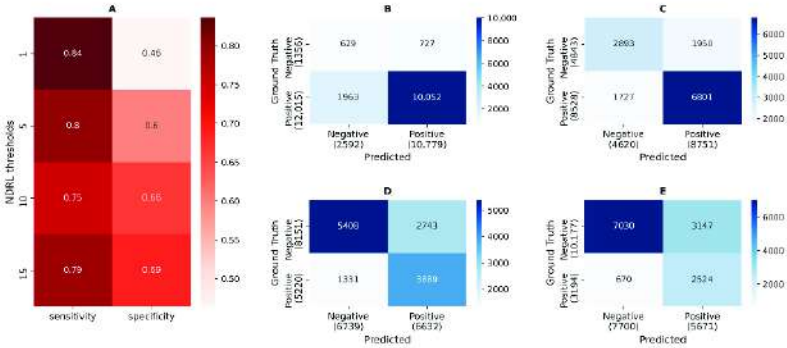


Figure 3. Sensitivity, specificity, and confusion matrix of the test set using the thresholds 1, 5, 10, and 15 for the NDRL. **(A)** shows sensitivity and specificity. **(B–E)** show, respectively, the confusion matrix using the NDRL thresholds of 1, 5, 10, and 15. The values are not normalized. Therefore the color cannot be compared between subplots. The ground truth, true categories, is placed on the left, and predicted values on the bottom. True positives are on the bottom right, and true negatives are on the top left.

We hypothesized that some positions predicted by the model as false positives might become true positives later on. To test this hypothesis, we used model predictions trained with data from 19 April 2021 but with updated ground truth from 6 January 2022. Table 1 shows the percentage of the predicted false positives that turned into true positives and other variables for various NDRL thresholds for considering a mutation as RM in the January 2022 ground truth. We used different NDRL thresholds because the number of lineages for each mutation in the January 2022 data is three to four times higher than in the April 2021 data. The AUC and sensitivity of the RM position prediction increase as the NDRL threshold increases (Table 1). When using an NDRL threshold of 45, 17.7% of the false positives predicted for the NDRL threshold of 15 turns into true positives in January

2022. At this NDRL threshold, the AUC, sensitivity, and specificity of the RM position prediction are 0.8, 0.747, and 0.716, respectively (Table 1). All these metrics correspond to the testing set. These data confirm our hypothesis that some of our predicted false positives become true positives later on.

Table 1. Performance metrics of models trained with data from 19 April 2021 and evaluated with data from 6 January 2022. This table shows the metrics (ROC-AUC, sensitivity, specificity, accuracy), false positives from 2021 (fps in 2021) that turn into true positives (fps in 2021 to tps in 2022), and the proportion of this conversion (fps to tps ratio) using different NDRL thresholds for the data from 2022 (th true January 2022).

NDRL Threshold Pred 04/2021	NDRL Threshold True 01/2022	ROC-AUC	Sensitivity	Specificity	Accuracy	Fps in 2021	Fps in 2021 to Tps in 2022	Fps to Tps Ratio
15	15	0.644	0.481	0.724	0.549	3147	2119	0.673
15	30	0.728	0.597	0.743	0.671	3147	1402	0.446
15	45	0.800	0.747	0.716	0.726	3147	557	0.177
15	60	0.848	0.853	0.681	0.715	3147	99	0.031
15	75	0.873	0.910	0.655	0.691	3147	14	0.004
15	90	0.879	0.936	0.636	0.668	3147	5	0.002
15	105	0.877	0.939	0.622	0.647	3147	2	0.001
15	120	0.880	0.949	0.612	0.634	3147	2	0.001
15	135	0.883	0.953	0.606	0.625	3147	0	0

2.4. Global Feature Importance of the Prediction of Whether a Given Position in the SARS-CoV-2 Genome Will Be Affected by a Recurrent Mutation

Neural networks are often described as black-box models when the influence of each input variable on the success of the model is unknown. We used the Shapely Additive exPlanations (SHAP) [57] to determine the influence of each variable on whether a position in the trained model would mutate or not. The most important features are those with the highest normalized SHAP values (see Materials and methods). We analyzed four models with NDRL thresholds of 1, 5, 10, and 15 from April 2021. The nucleotide in the central position (P0) of each evaluated window of 13 positions (P-6 to P6) is the most important variable in predicting the position of the SARS-CoV-2 genome where an RM will take place (Figure S5). Other important variables are the nucleotides in other positions (e.g., P1, P-1, P2) and the in vivo and in vitro RNA SHAPE-Seq reactivity data [53]. When the NDRL threshold is higher, the most relevant variables become more important. Mainly cytosines, and to a lesser extent, guanines, are more prone to being RM (Figures 4 and S6). False positives have either a guanine (35%) or a cytosine (25%), and true negatives have mainly adenine (46%) and uracil (45%) (Figure 4). Regarding the nucleotides surrounding the nucleotide that mutates, at an NDRL threshold of 15, the upstream and downstream positions (P-1 and P1, respectively) are the most relevant. In general, the other positions are of little importance (Figure 4 and Figure S6). In 44% and 27% of the true positives, there is an adenine or an uracil at P-1, and in 37% of the cases, there is an uracil at P1. This is consistent with evidence that the cytosines of the UC and AC motifs of the SARS-CoV-2 genome are preferentially deaminated by the APOBEC3A and APOBEC1 enzymes [12]. The importance given to the SHAPE-Seq reactivity comes after that of the nucleotides (Figure S5). However, the magnitude of their importance is at least five times lower. Low SHAPE-Seq reactivity values, in the range of 0 to 0.69, do not promote mutagenesis at most positions (Figure S7). However, higher SHAPE-Seq reactivity values lead to mutations (Figure S7). This analysis of the most important variables is compatible with a model that mainly predicts cytosines of the ACU pattern as RM in a region with an RNA structure that makes this cytosine more reactive. This is consistent with the SARS-CoV-2 mutational bias pattern and the preference of some host deaminases for specific sequences and RNA secondary structures [11,12,14].

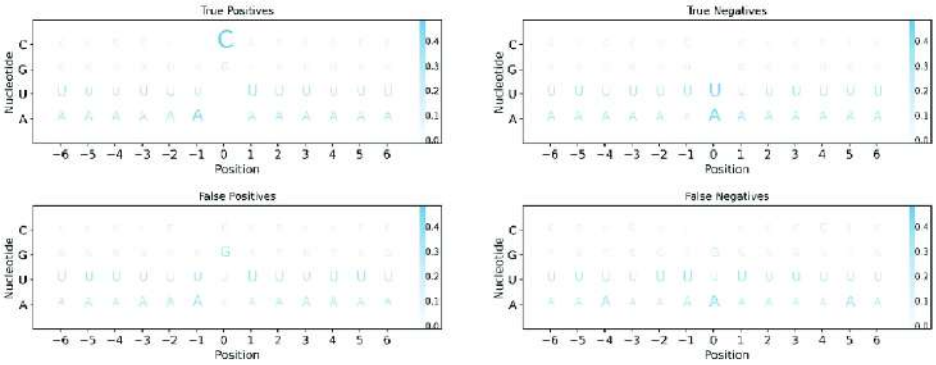


Figure 4. Logos of the true positives, true negatives, false positives, and false negatives when the NDRL threshold is 15. Each subplot shows the position on the horizontal axis and a row per nucleotide on the vertical axis. Each column/position for every subplot is normalized vertically, so they add up to 1. The intensity of the blue correlates with the size of each letter, so a more frequent nucleotide appears in a darker blue and a larger letter size.

2.5. *Prediction of Whether a Given Mutation Will Be a Recurrent Mutation*

We developed another machine learning method, this time to predict the NDRL in which we can find a specific mutation, i.e., whether a specific mutation will become an RM. The data were split into training, validation, and testing sets, in the same manner as described before. Similarly, the model selection was also chosen by maintaining a minimum value of 0.8 for the sensitivity in the validation set and selecting the model that achieved the highest specificity. The performance of this prediction method was similar to the previous one. The ROC-AUC of the prediction of whether a mutation will be found in more than 15 NDRLs was 0.88, 0.83, and 0.84 for, respectively, the training, validation, and testing sets (Figures S8 and S9). In the testing set, once again, the worst prediction was found in the spike gene (AUC 0.82, Figure S9). The most important variables for predicting the NDRL of a mutation were the starting nucleotide, towards which it mutates, and the in vitro SHAPE-Seq reactivity (Figure S10). For the NDRL threshold of 15, (a) the most important variable is when a nucleotide mutates into an uracil (>U at Figure S10), and (b) adenine and cytosine were the most relevant starting nucleotides (A> and C> at Figure S10). Again, this is compatible with a model that predicts the mutations C > U and A > G to be recurrent.

2.6. *Evaluation of the Models with the Variants of Concern*

A good way to test the usefulness of our predictions is to check whether our models could have predicted the mutations we found in the variants of concern (VoCs). The identification of the positions of the testing set that mutate in the Alpha, Beta, Delta, Gamma, and Omicron VoCs has an accuracy of 0.636, 0.600, 0.778, 0.80, and 0.697, respectively, when using the ground truth of January 2022 and an NDRL threshold of 45 (Table 2). The accuracy for predicting the mutations of these VoCs is 0.545, 0.40, 0.33, 0.733, 0.636, and 0.60 (Table 2).

Table 2. Performance over variants of concern.

Variant of Concern	No. of Mutations	No. of Mutations NDRL45	Accuracy	Sensitivity	Specificity
Position Prediction. NDRL 15/45 * (2022)					
Alpha	11	8	0.636	0.750	0.333
Beta	10	8	0.600	0.625	0.500
Delta	9	7	0.778	0.857	0.500
Gamma	15	11	0.800	0.818	0.750
Omicron	33	24	0.697	0.708	0.667
Combined	61	49	0.697	0.776	0.667
Mutation Prediction. NDRL 15/45 * (2022)					
Alpha	11	8	0.545	0.500	0.667
Beta	10	8	0.400	0.375	0.500
Delta	9	7	0.333	0.286	0.500
Gamma	15	11	0.733	0.727	0.750
Omicron	33	17	0.636	0.471	0.812
Combined	61	42	0.607	0.500	0.842

* 15/45 means that the NDRL threshold of 15 was used for the prediction, but it was evaluated with the ground truth from January 2022, using an NDRL threshold of 45.

Several mutations of the testing set from the VoC are correctly predicted by our two prediction methods (Table 3 and Table S1). This is the case for the C3267U, C3828U and G5230U mutations of the PLpro gene, the G15451A mutation of the RNA polymerase, the C21614U, C21638U, C21762U, C21846U, G21974U, G22132U, C22686U, G22813U, G22898A, C23525U, C23604A, C23664U, C23709U, G23948U, C24642U, G24914C and G25088U mutations of the spike gene. Our method predicts that the C14408U mutation, present in all VoCs and that codes for the RNA polymerase P323L shift, is an RM. As this mutation was found early in the pandemic, it is found in more than 99% of SARS-CoV-2 genomes available until January 2022. This mutation is present in all Pango lineages and therefore it is not considered to be an RM. As a result, this mutation is a false positive of our predictions. Mutations A5648C and A22812C from the VoC Gamma and U6515A, G8393A, A23055G, U23075C, A23403G, and A24424C from the VoC Omicron are true negative predictions of our position and mutation prediction models. These mutations are not recurrent because they are found in less than 45 Pango lineages. Mutations C10449A, U23599G, C23854A, and C24130A (Omicron) are true positives of the position prediction and true negatives of the mutation prediction. This means that these positions contain RMs, but the particular mutations observed in these VoCs are not recurrent. It has been described that the VoC Omicron contains many mutations not observed with a high frequency in other SARS-CoV-2 genomes [58]. Other VoC mutations were false negatives of our predictions. This is the case with the A2832G and U6954C mutations from the PLpro and the A21801C, U22679C, U22917G, G23012A, A23013C, and A23063U mutations of the spike gene. The G23012A mutation from the receptor binding domain (RBD) of the spike protein causes the AA change E484K, which reduces serum neutralization efficiency [59]. The A23063U mutation is a missense mutation present in the VoCs Alpha, Beta, Gamma, and Omicron that results in the AA substitution N501Y of the spike protein's RBD. This substitution enhances SARS-CoV-2 infection and transmission and occurs convergently in several lineages [60]. The U22917G mutation causes the AA substitution L452R that increases spike stability, viral infectivity, and viral fusogenicity and thereby promotes viral replication [61]. Although the A23063U and U22917G mutations were present, respectively, in more than 1 million and 2 million of SARS-CoV-2 genomes available up until January 2022 and in more than 280 pangolin lineages, neither of our two prediction methods predicted these positions or mutations as recurrent. These kinds of mutations, which enhance SARS-CoV-2 infection and transmission, are the most interesting ones but the most difficult to predict because they could not be caused by host deaminases. Our current prediction models are not specifically trained to detect them. Other interesting cases are those that are false positives

of our predictions. The C14408C (RNA polymerase) and C24503U (spike) mutations are found in a few SARS-CoV-2 genomes but are now in the VoC Omicron. They are false positives of our predictions because they are found in very few cases until January 2022. They could be mutations that were not observed because they have a negative impact on SARS-CoV-2, or they could be mutations that may be recurrent in the future, and it would therefore be interesting to monitor them.

Table 3. Summary of some VoC predictions on position (pos) and mutation (mut). See Table S1 for a complete table.

Position	VoC *	Gene	Mutation	AA	N [i]	Countries [i]	NL [i,†]	NDRL [i]		Prediction 15/45 [‡]	
								pos.	mut.	pos.	mut.
3267	A	Plpro	C3267U	T183I	903,866	164	246	241	238	tp	tp
21614	G	S	C21614U	L18F	167,687	145	428	399	397	tp	tp
21762	O	S	C21762U	A67V	13,723	103	244	248	244	tp	tp
23709	A	S	C23709U	T716I	904,197	167	247	234	234	tp	tp
14408	A,B,D,G,O	RNA pol	C14408U	P323L	4,577,014	193	1450	1	1	fp	fp
6515	O	Plpro	U6515A	L1266I	61	4	3	15	3	tn	tn
23403	A,B,D,G,O	S	A23403G	D614G	4,589,366	193	1460	1	1	tn	tn
24424	O	S	A24424C	Q954H	5	4	4	30	4	tn	tn
8393	O	Plpro	G8393A	A1892T	722	30	33	43	32	tn	tn
10449	O	M-pro	C10449A	P132H	1064	32	33	173	31	tp	tn
23599	O	S	U23599G	N679K	2425	38	36	138	34	tp	tn
23854	O	S	C23854A	N764K	849	27	26	200	24	tp	tn
24130	O	S	C24130A	N856K	658	32	32	314	31	tp	tn
21801	B	S	A21801C	D80A	25,012	108	88	133	84	fn	fn
22917	D	S	U22917G	L452R	2,844,958	171	321	154	137	fn	fn
23063	A,B,G,O	S	A23063U	N501Y	1,020,863	175	280	243	242	fn	fn
21801	B	S	A21801C	D80A	25,012	108	88	133	84	fn	fn

* A: Alpha, B: Beta, D: Delta, G: Gamma, and O: Omicron VoC. [i] On 6 January 2022. [†] Number of Pango lineages. [‡] 15/45 means that the NDRL threshold of 15 was used for the prediction, but it was evaluated with the ground truth from January 2022, using an NDRL threshold of 45. tp, fp, tn, and fn mean true positive, false positive, true negative, and false negative, respectively.

2.7. Prediction of AA Changes Caused by Recurrent Mutations in the M-Pro and Spike Proteins

We used our model to predict whether a specific mutation is recurrent to evaluate all possible mutations in the M-pro and spike proteins. The predicted mutations obtained with the model trained with data from April 2021 produced a set of possible AAs that were compared with the AA found in the ground truths from April 2021 and January 2022. We obtained a ROC-AUC of 0.849 and 0.687 for the M-pro and spike proteins, respectively (Table 4). For this calculation, we took all AAs that were neither observed nor predicted to mutate as true negatives. The 8 and 102 AA positions for M-pro and spike proteins among the false positives of the RM prediction became true positives with the data from January 2022 (Table 4).

Table 4. Amino acid change predictions in M-pro and spike proteins.

Gene	Year [†]	tp	fp	fn	tn	tnp	acc	spec	Sens	roc-auc
spike	2021	371	1880	471	24,032	113	0.912	0.927	0.441	0.684
	2022	473	1778	596	23,907	103	0.911	0.931	0.442	0.687
M-pro	2021	133	492	26	5775	22	0.919	0.921	0.836	0.879
	2022	141	484	41	5760	22	0.918	0.922	0.775	0.849

[‡] Date of the ground truth used to evaluate the model. 2021 means the ground truth from 19 April 2021, and 2022 means the ground truth from 6 January 2022. The columns acc, spec, sens, tp, fp, fn, tn, and tnp stand for accuracy, specificity, sensitivity, true positives, false positives, false negatives, true negatives, and true negative positions, respectively.

The comparison of the predicted AA changes with the mutations observed up until January 2022 shows that more than 77% of the observed recurrent AA changes and recurrent synonymous mutations observed in the M-pro protein are well predicted by our method (Figure 5). False positives (shown in red in Figure 5) could be recurrent AA changes that will be observed in the future and are interesting to monitor. AAs that have mutated and

that are thought to have other possibilities as predicted by our method, such as Ala94, Arg105, Pro108, Ala116, Ala129, Cys160, Met162, Pro168, Ala191, Ala193, Ala234, Val247, Ala260, Ala261, Arg279, and Ala285, are positions that tolerate diverse AA substitutions because they do not affect protein function [62]. Among M-pro AAs, such as Thr25, Thr26, His41, Met49, Phe140, Gly143, Cys145, His163, His164, Met165, Glu166, Pro168, His172, Asp187 and Gln189, which usually make intermolecular interactions with covalent and non-covalent inhibitors [63], only Gly143 and Pro168 show significant AAs changes caused by RM (Figure 5). In addition, in order to evaluate the performance of our prediction method, it is important to bear in mind that the false positives and false negatives of our predictions may include negatively or positively selected positions. Among the false positives, there are also deleterious mutations that are not expected to occur. Among these, there are nonsense mutations that lead to the appearance of a premature stop codon and mutations of the catalytic Cys145 and His41 [62]. The first and last AAs (a serine and a glutamine, respectively) from the M-pro are also false positives of our prediction. These two AAs are not expected to mutate because these AAs are recognized by the M-pro itself to cut the polyprotein 1a and 1b to generate the mature M-pro. Other false positives are the AAs between positions 143 and 149 (Figure 5). This region corresponds to the conserved GSCGSxG motif, which has been identified as important for initiating catalysis in SARS-CoV and MERS-CoV [64]. Among the false negatives (shown in dark yellow in Figure 5), they could be recurrent mutations. Instead of being recurrent because the host deaminases have caused them, they have been positively selected, and when they do occur, they confer a beneficial effect on virus transmission. Asn274 has several recurrent AA substitutions that our prediction method was unable to predict.

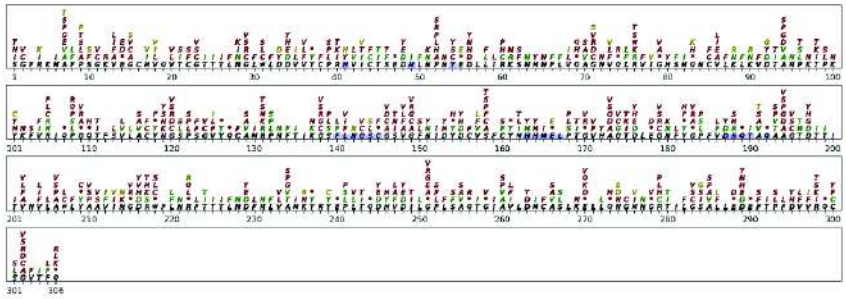


Figure 5. Comparison between M-pro AA changes predicted by our model and changes observed up until January 2022 (NDRL $\geq$ 45). The reference M-pro AA sequence is shown in black, just above the AA positions. The possible AAs produced by the predicted mutations of NDLR $\geq$ 15 are stacked over the reference sequence. True positives, false positives, and false negatives are shown in green, red, and dark yellow/gold, respectively. * represents a stop codon, and the same AA represents a synonymous mutation. The AAs with a blue background correspond to the subsites S1, S1′, S2, and S3.

The sensitivity of our predictions is only 44.2% for the spike protein, showing that the AA changes for the spike proteins are more difficult to predict (Figure S11). One of the main reasons for this low sensitivity is the high number of false negatives (Table 4). The RBD is a key functional part of the spike protein that is responsible for ACE2 binding [65]. Our prediction model showed that 46% of the recurrent AA changes and recurrent synonymous mutations observed for the RBD until January 2022 are true positives (AA in green in Figure 6A). Among the false positives (shown in red in Figure 6), there are nonsense mutations that were not expected to occur. Other false positives may include AA changes that are not observed in enough lineages to be considered RM or mutations not observed because they are deleterious. Among the false negatives (shown in dark yellow in Figure 6),

there are mutations that our method had not predicted as recurrent but that gives an advantage to the virus. These include some of the mutations observed in some of the VOCs discussed earlier, such as L452R [61], E484K [59], and N501Y [60]. Another interesting region of the spike protein to be studied is the furin cleavage site, which plays a key role in the cell tropism and pathogenesis of SARS-CoV-2 [66]. This cleavage site contains the residues PRRARS at positions 681–686 of the spike protein. Figure 6B shows our mutation predictions for this region. Some of the mutations in this region are expected to be rare because they may reduce the cleavage caused by the furin protein [66]. This is the case with R682, R685, and S686. The AAs substitutions R682L and R682W are predicted by our methodology to be caused by the RM G23607T and C23606T, respectively. They are observed in a few SARS-CoV-2 lineages and are false positives of our prediction (Figure 6B). The R685C, R685S, and S686C changes are also false positives of our predictions for the same reason. R683 seems to be not so important. AA changes of R683 to other AAs, i.e., L, Q, and W, are recurrent, as our methodology correctly predicted (Figure 6B). Our methodology also correctly predicted that the P681H substitution observed in the alpha variant was caused by an RM. This substitution may slightly increase the furin cleavage, but it has no effect on viral entry or cell-cell spread [67]. However, the P681R substitution observed in the delta variant caused by the C23604G mutation is a false negative of our prediction.

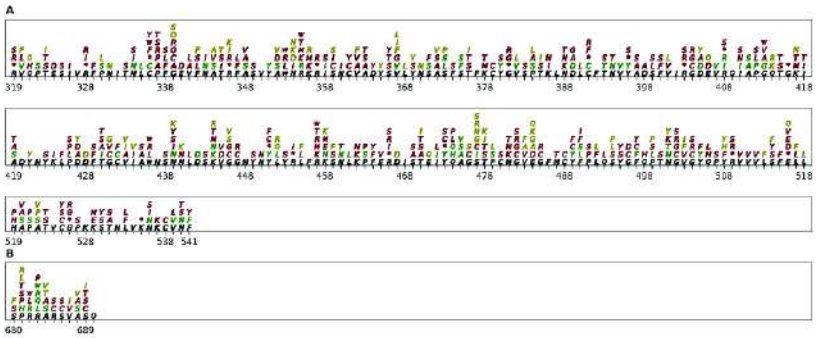


Figure 6. Comparison between AA changes predicted by our model and changes observed up until January 2022 from the RBD and furin cleavage site of the spike protein. The sequence from the RBD (**A**) and furin cleavage site (**B**) from the spike protein is shown in black. The AA changes predicted by our model are stacked over the reference sequence. The true positives, false positives, and false negatives are shown in green, red, and dark yellow/gold. * represents a stop codon, and the same AA represents a synonymous mutation.

3. Materials and Methods

We used 877,086 SARS-CoV-2 genomes from the GISAID database [44,68] available until 19 April 2021, to create the predictive model, and 4,616,059 SARS-CoV-2 genomes from 6 January 2022 to evaluate the model. Only genomes with a high coverage were considered. The NC_045512.2 genome [69] was set as a reference genome in order to align and identify mutations. The mutations, date, pangolin lineage, and genome ID were captured for each genome. Insertions and deletions were not taken into account. Only mutations from A, G, C, and U to A, G, C, and U were considered. For each mutation, we took the position and calculated the number of different pangolin lineages where this mutation was observed. We applied an algorithm to group the lineages that were linked together so that the whole group could be counted as one, thereby reducing the number of lineages for each mutation. We then calculated the NDRL.

3.1. NDRL Algorithm

We established a set of thresholds to define when a mutation belongs to a lineage and a group of linked lineages.

Th1: Threshold that defines when a mutation (grouped by the position that mutates) belongs to a lineage. If a mutation is present in at least th1% of the genomes that belong to that lineage, we say that it belongs to that lineage or that those mutation-genomes are related. In our calculations, we considered that a mutation belongs to a lineage if it is in at least 60% of that lineage's genomes; therefore, Th1 is 0.6.

Th2: Threshold that defines when a mutation belongs to a group of related lineages. If a mutation belongs (marked by Th1) to at least th2% of the lineages of related lineages, we say it belongs to all those lineages for that mutation/position. In our calculations, we considered that a mutation belongs to a group of related lineages if it belongs to at least 60% of them. Thus, Th2 is 0.6 as well.

A group of related lineages is a lineage and all its descendants. For example, A.1.* means all the lineages that begin with A.1. [A.1.1, A.1.2, . . . , A.1.10]. When a mutation belongs to a group of related lineages, the NDRL count is equal to one for that whole group. Therefore, it is easier to count from parent to children, from a more general, bigger group to a more specific one.

For each mutation, we visited each lineage, parents first, and evaluated which complied with the Th1 and Th2 values. If the parent lineage complied, it was grouped with all its children and counted as 1. All these children were then excluded from further evaluations. If there was no group of related lineages, then the NDRL count was equivalent to the number of lineages where the mutation was present.

3.2. Data Set Composition

Our main focus was finding future mutations in the genes M-pro, spike, PLpro, and RNA_pol. Therefore, these genes became the test set. Among the other genes, those that have a similar length are helicase, nsp6, endoRNAse, and M. Thus, we used these for the validation set. This means the training set was composed of the remaining genes: leader, nsp2, nsp4, nsp7, nsp8, nsp9, nsp10, exonuclease, methyltransferase, ORF3a, E, ORF6, ORF7a, ORF7b, ORF8, N and ORF10. For each position in the selected genes, a window of six positions was taken on each side, before and after. Therefore, the input of the models had 13 positions of the genome, the central position being the one under analysis. A higher number of positions did not improve the performance of the trained models. A set of features were considered for each position of every window: mRNA nucleotide, RNA normalized 2′-hydroxyl acylation analyzed by primer extension (SHAPE) reactivity [53], secondary structure information calculated using Vienna RNAfold [70], and the AA to which it is going to be translated. The secondary structure information was composed using a forgi file format, and if it was connected to another nucleotide, we stated to which one it was connected. We converted the variables that did not have a numerical representation to a one-hot encoding representation. There were some missing values in the reactivity data, so we used a multivariate imputation method [71–73]. For the position prediction, we grouped positives and negatives into four groups. These groups contained mutations that were at least in 1, 5, 10, or 15 NDRL. The NDRL was defined using the Th1 and Th2 equal to 0.6. Therefore, when the threshold was set to 5, those positions with an NDRL lower than 5 were set as negatives and those with a higher value as positives. Therefore, each mutation with a threshold (th) of 15 was also a mutation in the group with a threshold of 10, 5, and 1. When higher NDRL values were evaluated, the performance increased, but the number of mutations decreased substantially. For the mutation prediction models, we followed the same steps but introduced a few changes. We only had three groups, NDRL 5, 10, and 15. NDRL 1 was excluded because we only worked with registered mutations. Positions with no registered mutations were not included, so it was not possible to define a negative category for the NDRL lower than 1. The other change was the addition of the nucleotide to which the position in the center of the input window would mutate.

3.3. Machine Learning

We used an artificial neural network (ANN) and multi-layer perceptron (MLP) architecture. To find the best hyper-parameters, such as the number of layers and neurons per layer (Table S2), we used the Scikit-Optimize library [74]. We used a search space range between 1 and 14 layers and between 1 and 2048 neurons per layer. The search space limits were set up so that it could be tested in less than a week and fit into a 12 GB GPU Memory. We used early stopping as the regularization technique. Our model selection criteria consisted in considering only models with at least 0.8 of sensitivity and the highest possible specificity. Details about the metric implementation can be found in the file model_selection_metric.py at https://github.com/bsaldivaremc2/sarscov2_rm_prediction (accessed on 27 October 2022). We also tried convolutional neural networks (CNN) and transformers [75] architectures. The metrics obtained were comparable. However, MLP training was faster than training a transformer. In order to understand the models' feature importance, an MLP was simpler to integrate with the SHapely Additive exPlanation (SHAP) [57] library than CNNs. We also tried a non-ANN approach with TPOT [76], but the performance was worse. In addition, a similar AUC was obtained using Autokeras [77], but it lacked the flexibility to be integrated with our model selection criteria while maintaining good results and explainability. We used the mljar-supervised package [78] to generate a baseline of ensemble machine-learning models so that we could compare the performance of our models to other methods (including traditional machine learning models). A comprehensive list of the performance indicators for our chosen models and this baseline can be found in Tables S3 and S4. Our model outperforms the baseline in terms of meeting our model selection criteria (Tables S3 and S4). By using McNemar's test [79,80], we demonstrate in Table S5 that the differences between our models and the baseline are significant. The uncertainty quality of the models, measured with the Brier score [81], is available in Table S6. To obtain the most important features, we used the SHAP values. One SHAP value was extracted per sample. Therefore, in order to obtain the general importance of a specific feature, we took the absolute value of all SHAP values and added those values to each feature (Equation (1)).

$$F_j = \sum_{i=1}^{i=N} (V_i), F_{nj} = \frac{F_j}{\sum_{i=1}^{i=M} F_i} \quad (1)$$

F_j is Importance of Feature j.
V_i is the SHAP values of sample i.
N is the number of samples.
F_{nj} is the normalized Importance of Feature j.
M is the number of features.

To evaluate the predictions of our models with the test set genes, we used updated data from 6 January 2022. We used this data as a new ground truth, as shown in the Results and discussion section. Nevertheless, the number of lineages and the NDRL had changed for the known mutations from 2021 (with which we trained our models). Therefore, we calculated the growth factor for these known mutations NDRL2022/NRDL2021. The majority grew by a factor of three (15% between 2.75 and 3.25, 26% between 2.5 and 3.5). So, we multiplied the NDRL threshold from 2021 by three, which gave us a correspondence of 1/3, 5/15, 10/30, and 15/45 for 2021/2022. This resulted in an NDRL threshold of 45 instead of 15.

We obtained the list of variants of concern and the mutations that define them from the WHO [82] and covariant [83] websites. For the development of the machine learning models, we used a computer with 32 CPU threads, a 12 GB GPU, and 64 GB RAM.

4. Conclusions

Overall, we have created a novel methodology that uses an artificial neural network capable of predicting RM in the SARS-CoV-2 genome. We have used the SARS-CoV-2

genome sequence, SHAPE-Seq reactivity values, and other variables to predict the position that mutates, the mutation that occurs, and the AA changes involved. We have validated our predictions using a test set of four genes that includes the M-pro and the spike genes, as well as using a real-case scenario such as the prediction of RM in VoCs. The model is robust enough to predict mutations in the long term, as some false positives within a limited time frame become true positives in an extended period of time. The predictive method also may be useful for finding positively and negatively selected positions in the SARS-CoV-2 genome. Among false positives, there are deleterious mutations that were not expected to occur. Among false negatives, there could be positions that have been positively selected, and when they occur, they confer a beneficial effect on virus transmission. These results can be used to find antiviral drugs that will be effective against future SARS-CoV-2 mutations.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms232314683/s1.

Author Contributions: Conceptualization, B.S.-E., P.P., A.C.-M., G.P. and S.G.-V.; methodology, B.S.-E. and S.G.-V.; validation, B.S.-E.; formal analysis, B.S.-E.; investigation, B.S.-E. and S.G.-V.; data curation, B.S.-E., G.M., P.G.-S. and J.M.-T.; writing—original draft preparation, B.S.-E., P.P., G.P. and S.G.-V.; writing—review and editing, B.S.-E., P.P., G.P. and S.G.-V.; visualization, B.S.-E., G.M., P.G.-S. and J.M.-T.; supervision, P.P., A.C.-M., G.P. and S.G.-V.; project administration, S.G.-V. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie, grant agreement No. 713679, and by the Universitat Rovira i Virgili, grant 2021PFR-URV-96.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The machine learning models and information on their use are available in the GitHub repository https://github.com/bsaldivaremc2/sarscov2_rm_prediction (accessed on 27 October 2022).

Acknowledgments: We would like to acknowledge the authors, both from the submitting and originating laboratories, for the sequences from the GISAID database used in this study. We acknowledge our University's English language service for proofreading and correcting this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wu, F.; Zhao, S.; Yu, B.; Chen, Y.-M.; Wang, W.; Song, Z.-G.; Hu, Y.; Tao, Z.-W.; Tian, J.-H.; Pei, Y.-Y.; et al. A New Coronavirus Associated with Human Respiratory Disease in China. *Nature* **2020**, *579*, 265–269. [CrossRef] [PubMed]
2. Kim, D.; Lee, J.-Y.; Yang, J.-S.; Kim, J.W.; Kim, V.N.; Chang, H. The Architecture of SARS-CoV-2 Transcriptome. *Cell* **2020**, *181*, 914–921.e10. [CrossRef] [PubMed]
3. Chen, Y.; Liu, Q.; Guo, D. Emerging Coronaviruses: Genome Structure, Replication, and Pathogenesis. *J. Med. Virol.* **2020**, *92*, 418–423. [CrossRef] [PubMed]
4. Wang, R.; Hozumi, Y.; Zheng, Y.-H.; Yin, C.; Wei, G.-W. Host Immune Response Driving SARS-CoV-2 Evolution. *Viruses* **2020**, *12*, 1095. [CrossRef] [PubMed]
5. Carrasco-Hernandez, R.; Jácome, R.; López Vidal, Y.; Ponce de León, S. Are RNA Viruses Candidate Agents for the Next Global Pandemic? A Review. *ILAR J.* **2017**, *58*, 343–358. [CrossRef]
6. Duffy, S.; Shackelton, L.A.; Holmes, E.C. Rates of Evolutionary Change in Viruses: Patterns and Determinants. *Nat. Rev. Genet.* **2008**, *9*, 267–276. [CrossRef] [PubMed]
7. Eckerle, L.D.; Becker, M.M.; Halpin, R.A.; Li, K.; Venter, E.; Lu, X.; Scherbakova, S.; Graham, R.L.; Baric, R.S.; Stockwell, T.B.; et al. Infidelity of SARS-CoV Nsp14-Exonuclease Mutant Virus Replication Is Revealed by Complete Genome Sequencing. *PLOS Pathog.* **2010**, *6*, e1000896. [CrossRef]
8. Simmonds, P.; Ansari, M.A. Extensive C->U Transition Biases in the Genomes of a Wide Range of Mammalian RNA Viruses; Potential Associations with Transcriptional Mutations, Damage- or Host-Mediated Editing of Viral RNA. *PLOS Pathog.* **2021**, *17*, e1009596. [CrossRef]
9. Ratcliff, J.; Simmonds, P. Potential APOBEC-Mediated RNA Editing of the Genomes of SARS-CoV-2 and Other Coronaviruses and Its Impact on Their Longer Term Evolution. *Virology* **2021**, *556*, 62–72. [CrossRef]

10. Di Giorgio, S.; Martignano, F.; Torcia, M.G.; Mattiuz, G.; Conticello, S.G. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. *Sci. Adv.* **2020**, *6*, eabb5813. [CrossRef]
11. Harris, R.S.; Dudley, J.P. APOBECs and Virus Restriction. *Virology* **2015**, *479–480*, 131–145. [CrossRef]
12. Kim, K.; Calabrese, P.; Wang, S.; Qin, C.; Rao, Y.; Feng, P.; Chen, X.S. The Roles of APOBEC-Mediated RNA Editing in SARS-CoV-2 Mutations, Replication and Fitness. *Sci. Rep.* **2022**, *12*, 14972. [CrossRef]
13. Simmonds, P. 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. *mSphere* **2020**, *5*, e00408-20. [CrossRef] [PubMed]
14. Turakhia, Y.; Maio, N.D.; Thornlow, B.; Gozashti, L.; Lanfear, R.; Walker, C.R.; Hinrichs, A.S.; Fernandes, J.D.; Borges, R.; Slodkowicz, G.; et al. Stability of SARS-CoV-2 Phylogenies. *PLoS Genet.* **2020**, *16*, e1009175. [CrossRef] [PubMed]
15. Graudenzi, A.; Maspero, D.; Angaroni, F.; Piazza, R.; Ramazzotti, D. Mutational Signatures and Heterogeneous Host Response Revealed via Large-Scale Characterization of SARS-CoV-2 Genomic Diversity. *iScience* **2021**, *24*, 102116. [CrossRef] [PubMed]
16. Eisenberg, E.; Levanon, E.Y. A-to-I RNA Editing—Immune Protector and Transcriptome Diversifier. *Nat. Rev. Genet.* **2018**, *19*, 473–490. [CrossRef]
17. Vlachogiannis, N.I.; Verrou, K.-M.; Stellos, K.; Sfikakis, P.P.; Paraskevis, D. The Role of A-to-I RNA Editing in Infections by RNA Viruses: Possible Implications for SARS-CoV-2 Infection. *Clin. Immunol.* **2021**, *226*, 108699. [CrossRef]
18. van Dorp, L.; Richard, D.; Tan, C.C.S.; Shaw, L.P.; Acman, M.; Balloux, F. No Evidence for Increased Transmissibility from Recurrent Mutations in SARS-CoV-2. *Nat. Commun.* **2020**, *11*, 5986. [CrossRef]
19. Lauring, A.S.; Hodcroft, E.B. Genetic Variants of SARS-CoV-2—What Do They Mean? *JAMA* **2021**, *325*, 529–531. [CrossRef]
20. Khateeb, J.; Li, Y.; Zhang, H. Emerging SARS-CoV-2 Variants of Concern and Potential Intervention Approaches. *Crit. Care* **2021**, *25*, 244. [CrossRef]
21. Rochman, N.D.; Wolf, Y.I.; Faure, G.; Mutz, P.; Zhang, F.; Koonin, E.V. Ongoing Global and Regional Adaptive Evolution of SARS-CoV-2. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2104241118. [CrossRef] [PubMed]
22. CDC. Coronavirus Disease 2019 (COVID-19). Available online: https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html (accessed on 8 November 2021).
23. Salama, M.A.; Hassanien, A.E.; Mostafa, A. The Prediction of Virus Mutation Using Neural Networks and Rough Set Techniques. EURASIP J. Bioinforma. *Syst. Biol.* **2016**, *2016*, 10. [CrossRef]
24. van Dorp, L.; Acman, M.; Richard, D.; Shaw, L.P.; Ford, C.E.; Ormond, L.; Owen, C.J.; Pang, J.; Tan, C.C.S.; Boshier, F.A.T.; et al. Emergence of Genomic Diversity and Recurrent Mutations in SARS-CoV-2. *Infect. Genet. Evol.* **2020**, *83*, 104351. [CrossRef]
25. Greener, J.G.; Kandathil, S.M.; Moffat, L.; Jones, D.T. A Guide to Machine Learning for Biologists. *Nat. Rev. Mol. Cell Biol.* **2022**, *23*, 40–55. [CrossRef] [PubMed]
26. Xu, C.; Jackson, S.A. Machine Learning and Complex Biological Data. *Genome Biol.* **2019**, *20*, 76. [CrossRef]
27. Tng, S.S.; Le, N.Q.K.; Yeh, H.-Y.; Chua, M.C.H. Improved Prediction Model of Protein Lysine Crotonylation Sites Using Bidirectional Recurrent Neural Networks. *J. Proteome Res.* **2022**, *21*, 265–273. [CrossRef]
28. Le, N.Q.K.; Ho, Q.-T.; Ou, Y.-Y. Using Two-Dimensional Convolutional Neural Networks for Identifying GTP Binding Sites in Rab Proteins. *J. Bioinform. Comput. Biol.* **2019**, *17*, 1950005. [CrossRef]
29. Yan, S.; Wu, G. Application of Neural Network to Predict Mutations in Proteins from Influenza A Viruses—A Review of Our Approaches with Implication for Predicting Mutations in Coronaviruses. *J. Phys. Conf. Ser.* **2020**, *1682*, 012019. [CrossRef]
30. Yang, W.; Bang, H.; Jang, K.; Sung, M.K.; Choi, J.K. Predicting the Recurrence of Noncoding Regulatory Mutations in Cancer. *BMC Bioinform.* **2016**, *17*, 492. [CrossRef]
31. Malone, B.; Simovski, B.; Moliné, C.; Cheng, J.; Gheorghe, M.; Fontenelle, H.; Vardaxis, I.; Tennøe, S.; Malmberg, J.-A.; Stratford, R.; et al. Artificial Intelligence Predicts the Immunogenic Landscape of SARS-CoV-2 Leading to Universal Blueprints for Vaccine Designs. *Sci. Rep.* **2020**, *10*, 22375. [CrossRef]
32. Liu, X.; Luo, Y.; Li, P.; Song, S.; Peng, J. Deep Geometric Representations for Modeling Effects of Mutations on Protein-Protein Binding Affinity. *PLoS Comput. Biol.* **2021**, *17*, e1009284. [CrossRef] [PubMed]
33. Hu, F.; Wang, L.; Hu, Y.; Wang, D.; Wang, W.; Jiang, J.; Li, N.; Yin, P. A Novel Framework Integrating AI Model and Enzymological Experiments Promotes Identification of SARS-CoV-2 3CL Protease Inhibitors and Activity-Based Probe. *Brief. Bioinform.* **2021**, *22*, bbab301. [CrossRef] [PubMed]
34. Mekni, N.; Coronnello, C.; Langer, T.; Rosa, M.D.; Perricone, U. Support Vector Machine as a Supervised Learning for the Prioritization of Novel Potential SARS-CoV-2 Main Protease Inhibitors. *Int. J. Mol. Sci.* **2021**, *22*, 7714. [CrossRef] [PubMed]
35. Nagy, Á.; Ligeti, B.; Szebeni, J.; Pongor, S.; Győrffy, B. COVIDOUTCOME—Estimating COVID Severity Based on Mutation Signatures in the SARS-CoV-2 Genome. *Database* **2021**, *2021*, baab020. [CrossRef]
36. Hossain, M.S.; Pathan, A.Q.M.S.U.; Islam, M.N.; Tonmoy, M.I.Q.; Rakib, M.I.; Munim, M.A.; Saha, O.; Fariha, A.; Reza, H.A.; Roy, M.; et al. Genome-Wide Identification and Prediction of SARS-CoV-2 Mutations Show an Abundance of Variants: Integrated Study of Bioinformatics and Deep Neural Learning. *Inform Med. Unlocked* **2021**, *27*, 100798. [CrossRef]
37. Nawaz, M.S.; Fournier-Viger, P.; Shojaee, A.; Fujita, H. Using Artificial Intelligence Techniques for COVID-19 Genome Analysis. *Appl. Intell.* **2021**, *51*, 3086–3103. [CrossRef]
38. Hie, B.; Zhong, E.D.; Berger, B.; Bryson, B. Learning the Language of Viral Evolution and Escape. *Science* **2021**, *371*, 284–288. [CrossRef]

39. Maher, M.C.; Bartha, I.; Weaver, S.; Iulio, J.D.; Ferri, E.; Soriaga, L.; Lempp, F.A.; Hie, B.L.; Bryson, B.; Berger, B.; et al. Predicting the Mutational Drivers of Future SARS-CoV-2 Variants of Concern. *Sci. Transl. Med.* **2022**, *14*, eabk3445. [CrossRef]
40. Sangeet, S.; Sarkar, R.; Mohanty, S.K.; Roy, S. Quantifying Mutational Response to Track the Evolution of SARS-CoV-2 Spike Variants: Introducing a Statistical-Mechanics-Guided Machine Learning Method. *J. Phys. Chem. B* **2022**, *126*, 7895–7905. [CrossRef]
41. Kc, G.B.; Bocci, G.; Verma, S.; Hassan, M.M.; Holmes, J.; Yang, J.J.; Sirimulla, S.; Oprea, T.I. A Machine Learning Platform to Estimate Anti-SARS-CoV-2 Activities. *Nat. Mach. Intell.* **2021**, *3*, 527–535. [CrossRef]
42. Arora, G.; Joshi, J.; Mandal, R.S.; Shrivastava, N.; Virmani, R.; Sethi, T. Artificial Intelligence in Surveillance, Diagnosis, Drug Discovery and Vaccine Development against COVID-19. *Pathogens* **2021**, *10*, 1048. [CrossRef] [PubMed]
43. Alyasseri, Z.A.A.; Al-Betar, M.A.; Doush, I.A.; Awadallah, M.A.; Abasi, A.K.; Makhadmeh, S.N.; Alomari, O.A.; Abdulkareem, K.H.; Adam, A.; Damasevicius, R.; et al. Review on COVID-19 Diagnosis Models Based on Machine Learning and Deep Learning Approaches. *Expert Syst.* **2022**, *39*, e12759. [CrossRef] [PubMed]
44. Khare, S.; Gurry, C.; Freitas, L.; Schultz, M.B.; Bach, G.; Diallo, A.; Akite, N.; Ho, J.; Lee, R.T.; Yeo, W.; et al. GISAID's Role in Pandemic Response. *China CDC Wkly.* **2021**, *3*, 1049–1051. [CrossRef] [PubMed]
45. Daron, J.; Bravo, I.G. Variability in Codon Usage in Coronaviruses Is Mainly Driven by Mutational Bias and Selective Constraints on CpG Dinucleotide. *Viruses* **2021**, *13*, 1800. [CrossRef] [PubMed]
46. Forni, D.; Cagliani, R.; Pontremoli, C.; Clerici, M.; Sironi, M. The Substitution Spectra of Coronavirus Genomes. *Brief. Bioinform.* **2022**, *23*, bbab382. [CrossRef] [PubMed]
47. Takata, M.A.; Gonçalves-Carneiro, D.; Zang, T.M.; Soll, S.J.; York, A.; Blanco-Melo, D.; Bieniasz, P.D. CG Dinucleotide Suppression Enables Antiviral Defence Targeting Non-Self RNA. *Nature* **2017**, *550*, 124–127. [CrossRef]
48. Xia, X. Extreme Genomic CpG Deficiency in SARS-CoV-2 and Evasion of Host Antiviral Defense. *Mol. Biol. Evol.* **2020**, *37*, 2699–2705. [CrossRef]
49. Rambaut, A.; Holmes, E.C.; O'Toole, Á.; Hill, V.; McCrone, J.T.; Ruis, C.; du Plessis, L.; Pybus, O.G. A Dynamic Nomenclature Proposal for SARS-CoV-2 Lineages to Assist Genomic Epidemiology. *Nat. Microbiol.* **2020**, *5*, 1403–1407. [CrossRef]
50. O'Toole, Á.; Scher, E.; Underwood, A.; Jackson, B.; Hill, V.; McCrone, J.T.; Colquhoun, R.; Ruis, C.; Abu-Dahab, K.; Taylor, B.; et al. Assignment of Epidemiological Lineages in an Emerging Pandemic Using the Pangolin Tool. *Virus Evol.* **2021**, *7*, veab064. [CrossRef]
51. Yi, K.; Kim, S.Y.; Bleazard, T.; Kim, T.; Youk, J.; Ju, Y.S. Mutational Spectrum of SARS-CoV-2 during the Global Pandemic. *Exp. Mol. Med.* **2021**, *53*, 1229–1237. [CrossRef]
52. 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. *Mol. Biol. Evol.* **2021**, *38*, 67–83. [CrossRef]
53. Manfredonia, I.; Nithin, C.; Ponce-Salvatierra, A.; Ghosh, P.; Wirecki, T.K.; Marinus, T.; Ogando, N.S.; Snijder, E.J.; van Hemert, M.J.; Bujnicki, J.M.; et al. Genome-Wide Mapping of SARS-CoV-2 RNA Structures Identifies Therapeutically-Relevant Elements. *Nucleic Acids Res.* **2020**, *48*, 12436–12452. [CrossRef]
54. Macip, G.; Garcia-Segura, P.; Mestres-Truyol, J.; Saldivar-Espinoza, B.; Pujadas, G.; Garcia-Vallvé, S. A Review of the Current Landscape of SARS-CoV-2 Main Protease Inhibitors: Have We Hit the Bullseye Yet? *Int. J. Mol. Sci.* **2022**, *23*, 259. [CrossRef]
55. Petushkova, A.I.; Zamyatnin, A.A. Papain-Like Proteases as Coronaviral Drug Targets: Current Inhibitors, Opportunities, and Limitations. *Pharmaceuticals* **2020**, *13*, 277. [CrossRef]
56. Chen, J.; Ali, F.; Khan, I.; Zhu, Y.Z. Recent Progress in the Development of Potential Drugs against SARS-CoV-2. Curr. Res. Pharmacol. *Drug Discov.* **2021**, *2*, 100057. [CrossRef]
57. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
58. Mallapaty, S. Where Did Omicron Come from? Three Key Theories. *Nature* **2022**, *602*, 26–28. [CrossRef]
59. Jangra, S.; Ye, C.; Rathnasinghe, R.; Stadlbauer, D.; Alshammary, H.; Amoako, A.A.; Awawda, M.H.; Beach, K.F.; Bermúdez-González, M.C.; Chernet, R.L.; et al. SARS-CoV-2 Spike E484K Mutation Reduces Antibody Neutralisation. *Lancet Microbe* **2021**, *2*, e283–e284. [CrossRef]
60. Liu, Y.; Liu, J.; Plante, K.S.; Plante, J.A.; Xie, X.; Zhang, X.; Ku, Z.; An, Z.; Scharton, D.; Schindewolf, C.; et al. The N501Y Spike Substitution Enhances SARS-CoV-2 Infection and Transmission. *Nature* **2022**, *602*, 294–299. [CrossRef]
61. Motozono, C.; Toyoda, M.; Zahradnik, J.; Saito, A.; Nasser, H.; Tan, T.S.; Ngare, I.; Kimura, I.; Uriu, K.; Kosugi, Y.; et al. SARS-CoV-2 Spike L452R Variant Evades Cellular Immunity and Increases Infectivity. *Cell Host Microbe* **2021**, *29*, 1124–1136.e11. [CrossRef]
62. Flynn, J.M.; Samant, N.; Schneider-Nachum, G.; Barkan, D.T.; Yilmaz, N.K.; Schiffer, C.A.; Moquin, S.A.; Dovala, D.; Bolon, D.N. Comprehensive Fitness Landscape of SARS-CoV-2 Mpro Reveals Insights into Viral Resistance Mechanisms. *eLife* **2022**, *11*, e77433. [CrossRef]
63. Gimeno, A.; Mestres-Truyol, J.; Ojeda-Montes, M.J.; Macip, G.; Saldivar-Espinoza, B.; Cereto-Massagué, A.; Pujadas, G.; Garcia-Vallvé, S. Prediction of Novel Inhibitors of the Main Protease (M-pro) of SARS-CoV-2 through Consensus Docking and Drug Reposition. *Int. J. Mol. Sci.* **2020**, *21*, 3793. [CrossRef] [PubMed]

64. Wang, H.; He, S.; Deng, W.; Zhang, Y.; Li, G.; Sun, J.; Zhao, W.; Guo, Y.; Yin, Z.; Li, D.; et al. Comprehensive Insights into the Catalytic Mechanism of Middle East Respiratory Syndrome 3C-Like Protease and Severe Acute Respiratory Syndrome 3C-Like Protease. *ACS Catal.* **2020**, *10*, 5871–5890. [CrossRef] [PubMed]
65. Lan, J.; Ge, J.; Yu, J.; Shan, S.; Zhou, H.; Fan, S.; Zhang, Q.; Shi, X.; Wang, Q.; Zhang, L.; et al. Structure of the SARS-CoV-2 Spike Receptor-Binding Domain Bound to the ACE2 Receptor. *Nature* **2020**, *581*, 215–220. [CrossRef] [PubMed]
66. Chan, Y.A.; Zhan, S.H. The Emergence of the Spike Furin Cleavage Site in SARS-CoV-2. *Mol. Biol. Evol.* **2022**, *39*, msab327. [CrossRef] [PubMed]
67. Lubinski, B.; Fernandes, M.H.V.; Frazier, L.; Tang, T.; Daniel, S.; Diel, D.G.; Jaimes, J.A.; Whittaker, G.R. Functional Evaluation of the P681H Mutation on the Proteolytic Activation of the SARS-CoV-2 Variant B.1.1.7 (Alpha) Spike. *iScience* **2022**, *25*, 103589. [CrossRef]
68. Elbe, S.; Buckland-Merrett, G. Data, Disease and Diplomacy: GISAID's Innovative Contribution to Global Health. *Glob. Chall.* **2017**, *1*, 33–46. [CrossRef]
69. Severe Acute Respiratory Syndrome Coronavirus 2 Isolate Wuhan-Hu-1, Complete Genome. Available online: https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2 (accessed on 20 March 2022).
70. Lorenz, R.; Bernhart, S.H.; Höner zu Siederdissen, C.; Tafer, H.; Flamm, C.; Stadler, P.F.; Hofacker, I.L. ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **2011**, *6*, 26. [CrossRef]
71. Buck, S.F. A Method of Estimation of Missing Values in Multivariate Data Suitable for Use with an Electronic Computer. *J. R. Stat. Soc. Ser. B Methodol.* **1960**, *22*, 302–306. [CrossRef]
72. van Buuren, S.; Groothuis-Oudshoorn, K. Mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* **2011**, *45*, 1–67. [CrossRef]
73. Sklearn.Impute.IterativeImputer. Available online: https://scikit-learn/stable/modules/generated/sklearn.impute.IterativeImputer.html (accessed on 20 March 2022).
74. Scikit-Optimize. Available online: https://github.com/scikit-optimize/scikit-optimize (accessed on 20 March 2022).
75. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762. [CrossRef]
76. Le, T.T.; Fu, W.; Moore, J.H. Scaling Tree-Based Automated Machine Learning to Biomedical Big Data with a Feature Set Selector. *Bioinformatics* **2020**, *36*, 250–256. [CrossRef] [PubMed]
77. Jin, H.; Song, Q.; Hu, X. Auto-Keras: An Efficient Neural Architecture Search System. *arXiv* **2018**, arXiv:1806.10282v3. [CrossRef]
78. Płońska, A.; Płoński, P. MLJAR: State-of-the-Art Automated Machine Learning Framework for Tabular Data. Version 0.10.3. 2021. Available online: https://github.com/mljar/mljar-supervised (accessed on 12 November 2022).
79. McNemar, Q. Note on the Sampling Error of the Difference between Correlated Proportions or Percentages. *Psychometrika* **1947**, *12*, 153–157. [CrossRef] [PubMed]
80. Dror, R.; Baumer, G.; Shlomov, S.; Reichart, R. The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Melbourne, Australia, 2018; pp. 1383–1392.
81. Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J.V.; Lakshminarayanan, B.; Snoek, J. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. *arXiv* **2019**, arXiv:1906.02530v2. [CrossRef]
82. Tracking SARS-CoV-2 Variants. Available online: https://www.who.int/activities/tracking-SARS-CoV-2-variants (accessed on 20 April 2022).
83. CoVariants. Available online: https://covariants.org/ (accessed on 20 April 2022).

Supplementary Information for
Prediction of recurrent mutations in SARS-CoV-2 using artificial neural networks.

Bryan Saldivar-Espinoza,[1] Guillem Macip,[1] Pol Garcia-Segura,[1] Júlia Mestres-Truyol,[1] Pere Puigbò,[2,3,4] Adrià Cereto-Massagué,[5] Gerard Pujadas,[1*] and Santiago Garcia-Vallve[1*]

[1] Departament de Bioquímica i Biotecnologia, Research group in Cheminformatics & Nutrition, Campus de Sescelades, Universitat Rovira i Virgili, 43007 Tarragona, Catalonia, Spain.
[2] Department of Biology, University of Turku, 20500 Turku, Finland.
[3] Department of Biochemistry and Biotechnology, Rovira i Virgili University, 43007 Tarragona, Catalonia, Spain.
[4] Nutrition and Health Unit, Eurecat Technology Centre of Catalonia, 43204 Reus, Catalonia, Spain
[5] EURECAT Centre Tecnològic de Catalunya, Centre for Omic Sciences (COS), Joint Unit Universitat Rovira i Virgili-EURECAT, Unique Scientific and Technical Infrastructures (ICTS)

*Correspondence: Santiago Garcia-Vallve

Email: santi.garcia-vallve@urv.cat (S.G.-V.)

This PDF file includes:

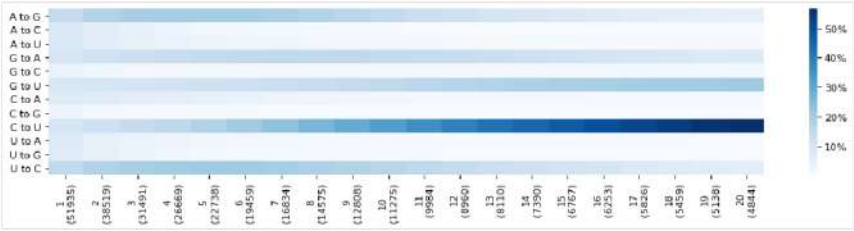
Figures S1 to S11
Tables S1 to S6

Other supplementary materials for this manuscript include the following:

Dataset S1 is in the file SI_Datasets.xlsx

Figure S1. Recurrent mutation changes at nucleotide level per Number of Distantly-Related Lineages thresholds (NDRL).

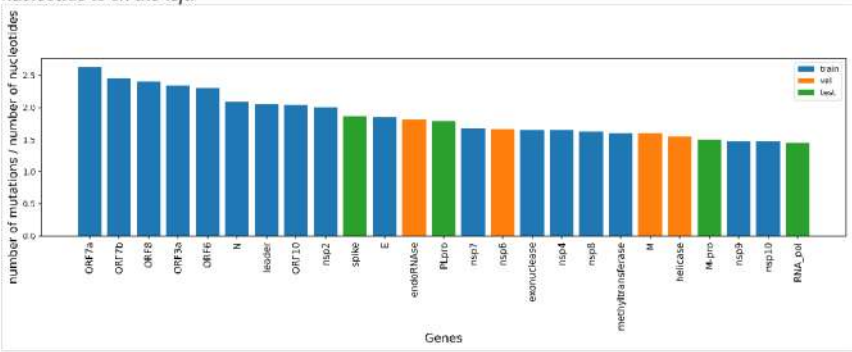
This plot shows each nucleotide change divided by the total number of mutations per NDRL thresholds. Each column adds up to 100%. The horizontal axis include in parentheses the number of total mutations per each considered threshold.



2

Figure S2. Number of mutations per number of nucleotides per gene across training, validation and testing set.

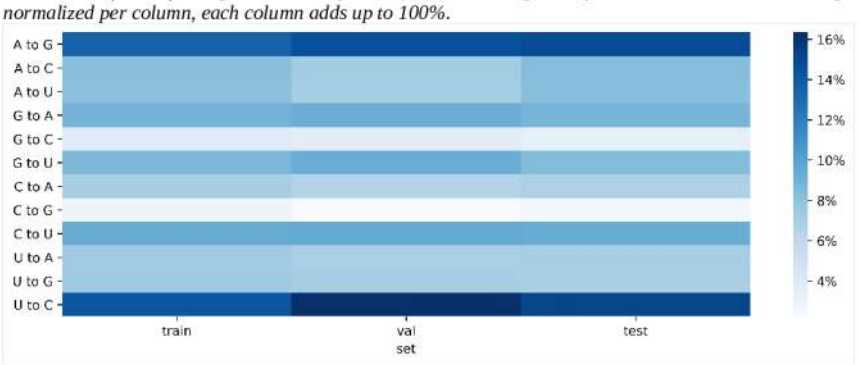
This plot shows the genes used in the training set (in blue), validation set (in orange) and testing set (in green). They are sorted from left to right so the gene with the highest number of mutations per nucleotide is on the left.



3

Figure S3. Mutations per set.

The columns from left to right are training, validation and testing. The plot shows the nucleotide change normalized per column, each column adds up to 100%.



4

Figure S4. Receiver Operating Characteristic (ROC) curve of the testing set genes using 1 (panel A), 5 (panel B), 10 (panel C) and 15 (panel D) as thresholds for the NDRL.
The subplots A, B, C and D correspond, respectively, to the NDRL thresholds of 1, 5, 10 and 15. Orange, green, red and purple lines correspond, respectively, to the Mpro, spike, PLpro and RNA_pol genes. The horizontal axis corresponds to the False Positive rate and the vertical axis to the True Positive rate. The values for the area under the curve (AUC) for a perfect prediction and for a random prediction are 1.0 and 0.5 respectively. The AUC for the genes M-pro, Spike, PLpro, RNA_pol are 0.73, 0.73, 0.76, 0.72 with the threshold 1; 0.79, 0.76, 0.77 and 0.8 with the threshold 5, 0.79, 0.71, 0.78, and 0.8 with the threshold 10 and 0.82, 0.75, 0.83 and 0.82 with the threshold 15.

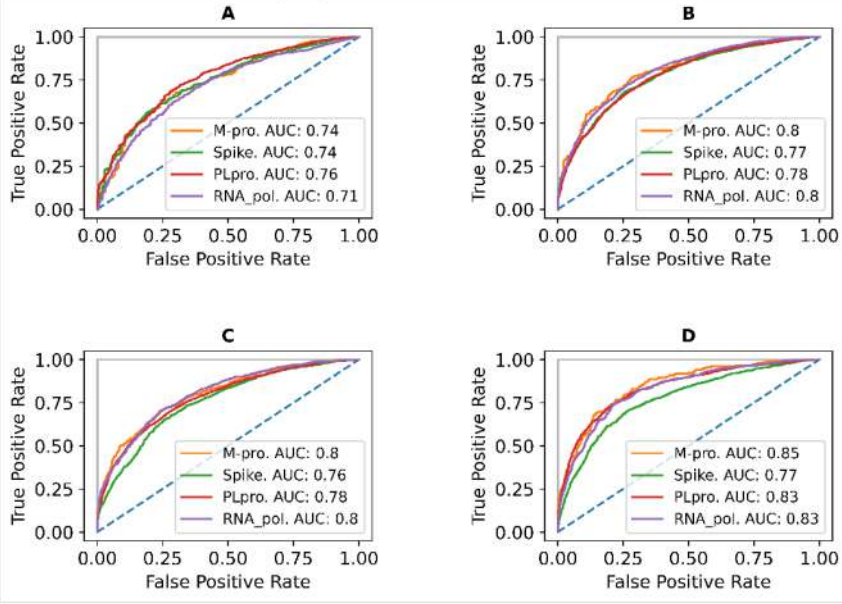
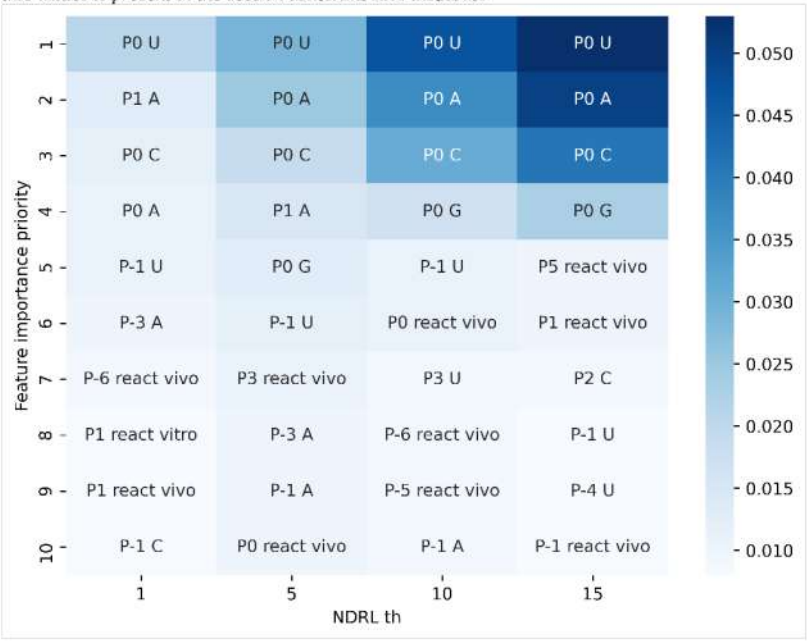


Figure S5. Top 10 most important variables in predicting the position of the SARS-CoV-2 genome where a RM will take place. Calculated across 4 NDRL **thresholds on the test set according to the SHAP values extracted from the model.**

In this Figure the most important variables appear on the top, a priority of 1 means more important and 10 less important. On the horizontal axis, from left to right the degree (threshold) of the recurrent mutation increases. A dark blue color means more important and a lighter blue less important. Each column was normalized among all the variables before cutting the top 10. The procedure to calculate this values is present in the section ***Materials and methods****.*



6

Figure S6. Mutation promoters or inhibitors' nucleotides, by position and NDRL thresholds on the test set according to the model SHAP values.

The subplots A, B, C and D correspond to the feature importance across the thresholds 1, 5, 10 and 15 of the NDRL *respectively. The values are normalized over the maximum absolute value across all the variables' SHAP values. Each box takes a positive value if that variable promotes mutations or otherwise takes a negative value.*

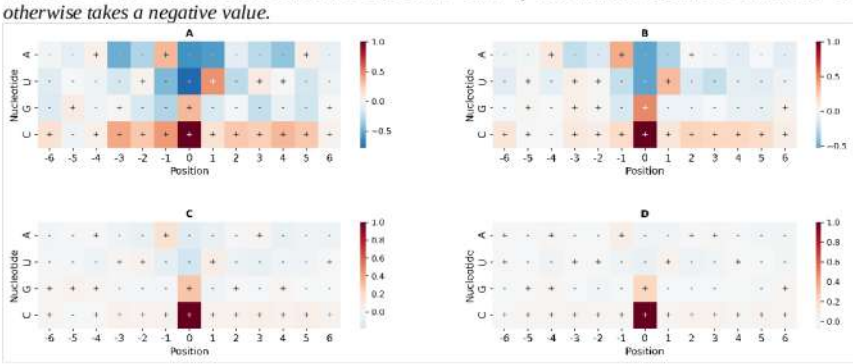
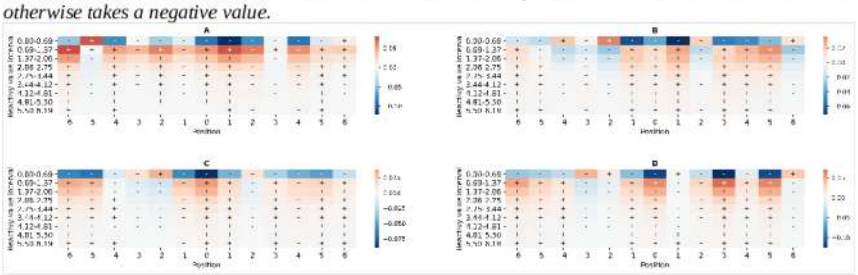


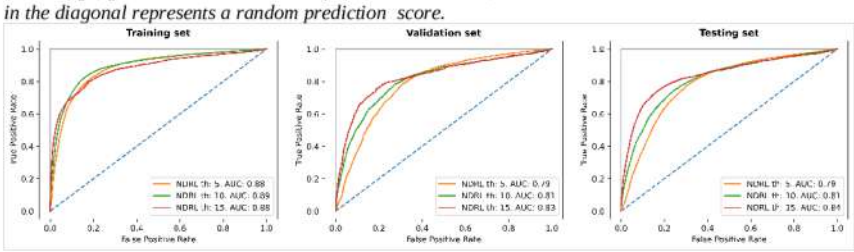
Figure S7. Mutation promoters or inhibitors' RNA normalized in vivo reactivity intervals, by position and NDRL thresholds on the test set according to the model SHAP values.
The subplots A, B, C and D correspond to the feature importance across the thresholds 1, 5, 10 and 15 of the NDRL respectively. The values are normalized over the maximum absolute value across all the variables' SHAP values. Each box takes a positive value if that variable promotes mutations or otherwise takes a negative value.



8

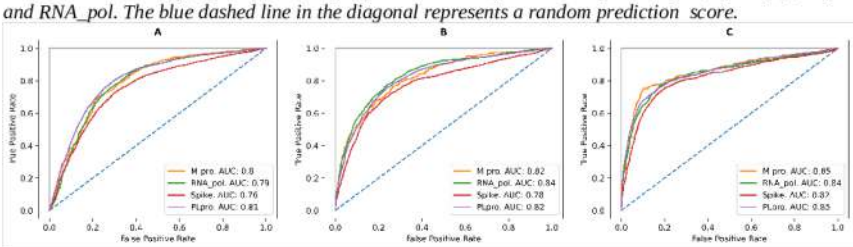
Figure S8. Receiver Operating Characteristic (ROC) curve for the testing, validation and training set using 5, 10 and 15 as thresholds for the Mutations NDRL.

The orange, green and red lines correspond to the NDRL 5, 10 and 15 thresholds. The blue dashed line in the diagonal represents a random prediction score.



9

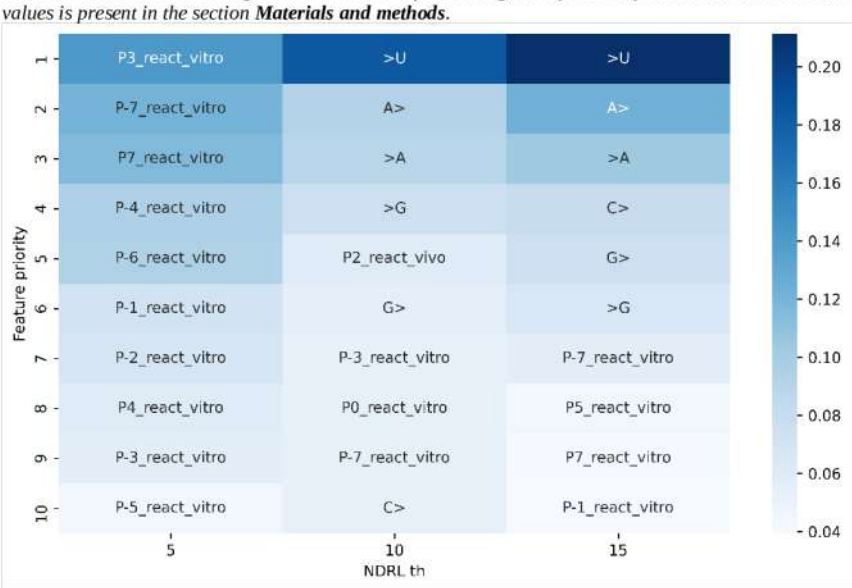
Figure S9. Receiver Operating Characteristic (ROC) curve of the testing set genes using 5, 10 and 15 as thresholds for the Mutations NDRL. *Subplots A, B and C correspond to the NDRL 5, 10 and 15 thresholds. The orange, green, red and purple lines correspond to the genes M-pro, Spike (S), PLpro and RNA_pol. The blue dashed line in the diagonal represents a random prediction score.*



10

Figure S10. Top 10 most important variables in predicting the NDRL of a mutation of the SARS-CoV-2. Calculated across 3 NDRL thresholds on the test set according to the SHAP values extracted from the model.

In this Figure the most important variables appear on the top, a priority of 1 means more important and 10 less important. On the horizontal axis, from left to right the degree (threshold) of the recurrent mutation increases. A dark blue color means more important and a lighter blue less important. Each column was normalized among all the variables before cutting the top 10.The procedure to calculate this values is present in the section ***Materials and methods****.*



11

Figure S11. Spike Predicted amino acid changes, ground truth from 2022, NDRL 15.

*This Figure shows the reference Spike protein amino acid sequence in black, in the first line of each subplot. Each subplot contains the sequence split into a maximum number of 100 residues. The possible amino acids (AA) produced by the predicted mutations, of NDRL≥15, are stacked over the reference sequence. True positives, false positives and false negatives are shown in green, red and dark yellow/gold, respectively. A start * represents a stop codon and a same AA represents a synonym mutation.*

13

Table S1. VoC mutations of the testing set.

Position	VOC	Gene	Mutation	AA	N[i]	Countries[j]	NL[↓†]	NDRL[↓†]		Prediction position[‡]				Prediction mutation[‡]		
								position	mutation	1/3	5/15	10/30	15/45	5/15	10/30	15/45
2832	omicron	nsp3	A2832G	K38R	1331	39	47	48	46	tp	tp	fn	fn	tp	tp	fn
3267	alpha	nsp3	C3267U	T183I	903,866	164	246	241	238	tp	tp	tp	tp	tp	tp	tp
3828	gamma	nsp3	C3828U	S370L	90,140	100	219	198	197	tp	tp	tp	tp	tp	tp	tp
5230	beta	nsp3	G5230U	K837N	30,479	120	237	240	233	tp	tp	tp	tp	tp	tp	tp
5388	alpha	nsp3	C5388A	A890D	899,293	163	116	149	108	tp	tp	tp	tp	tp	fn	fn
5648	gamma	nsp3	A5648C	K977Q	84,776	80	60	43	39	fn	fn	fn	tn	fn	fn	tn
6515	omicron	nsp3	U6515A	L1266I	61	4	3	15	3	tp	fn	tn	tn	tn	tn	tn
6954	alpha	nsp3	U6954C	I1412T	896,419	163	159	152	151	tp	tp	tp	fn	fn	fn	fn
8393	omicron	nsp3	G8393A	A1892T	722	30	33	43	32	tp	tp	fn	tn	tp	tp	tn
10323	beta	M-pro	A10323G	K90R	96,647	157	519	514	514	tp	tp	fn	tp	tp	tp	fn
10449	omicron	M-pro	C10449A	P132H	1064	32	33	173	31	tp	tp	tp	tp	fn	fn	tn
14408	omicron gamma delta beta alpha	RNA_pol	C14408U	P323L	4,577,014	193	1450	1	1	fp	fp	fp	fp	fp	fp	fp
15451	delta	RNA_pol	G15451A	G671S	2,769,305	167	302	130	124	tp	tp	tp	tp	tp	tp	tp
21614	gamma	S	C21614U	L18F	167,687	145	428	399	397	tp	tp	tp	tp	tp	tp	tp
21618	delta	S	C21618G	T19R	2,779,017	167	237	128	58	tp	tp	tp	tp	fn	fn	fn
21621	gamma	S	C21621A	T20N	83,978	84	74	223	52	tp	tp	tp	tp	fn	fn	fn
21638	gamma	S	C21638U	P26S	94,133	109	238	222	216	tp	tp	tp	tp	tp	tp	tp
21762	omicron	S	C21762U	A67V	13,723	103	244	248	244	tp	tp	tp	tp	tp	tp	tp
21801	beta	S	A21801C	D80A	25,012	108	88	133	84	tp	tp	fn	fn	fn	fn	fn
21846	omicron	S	C21846U	T95I	1,253,114	160	452	427	420	tp	tp	tp	tp	tp	tp	tp
21974	gamma	S	G21974U	D138Y	90,868	112	261	290	239	tp	fn	tp	tp	tp	tp	tp
21995	omicron	S	U21995G	Y145D	90	8	18	110	18	tp	fn	fn	fn	fn	tn	tn
22034	delta	S	A22034G	R158G	4072	40	124	127	124	tp	fn	tp	tp	fn	tp	fn
22132	gamma	S	G22132U	R190S	83,999	91	80	110	59	tp	tp	tp	tp	tp	tp	tp
22206	beta	S	A22206G	D215G	25,381	115	138	142	134	tp	tp	tp	tp	tp	tp	fn
22578	omicron	S	G22578A	G339D	1130	44	64	69	62	tp	fn	tp	tp	tp	tp	fn
22679	omicron	S	U22679C	S373P	1003	38	58	59	56	tp	fn	fn	fn	tp	fn	fn
22686	omicron	S	C22686U	S375F	888	32	47	46	45	tp	tp	tp	tp	tp	tp	tp
22812	gamma	S	A22812C	K417T	81,007	80	36	32	15	tp	tp	fn	tn	fn	tn	tn
22813	omicron beta	S	G22813U	K417N	29,436	116	131	144	125	tp	tp	tp	tp	tp	tp	tp
22882	omicron	S	U22882G	N440K	4754	60	57	91	54	tp	tp	tp	tp	fn	fn	fn
22898	omicron	S	G22898A	G446S	1194	53	88	92	88	tp	fn	tp	tp	tp	fn	tp

22917	delta	S	U22917G	L452R	2,844,958	171	321	154	137	fn	fn	fn	fn	fn	fn	fn
22992	omicron	S	G22992A	S477N	49,484	109	235	276	205	tp	tp	tp	tp	tp	fn	fn
22995	omicron delta	S	C22995A	T478K	2,802,366	168	270	123	91	tp	tp	tp	tp	fn	fn	fn
23012	gamma beta	S	G23012A	E484K	160,657	152	306	344	269	fn	fn	fn	fn	tp	tp	fn
23013	omicron	S	A23013C	E484A	2057	58	87	118	85	tp	tp	fn	fn	fn	fn	fn
23040	omicron	S	A23040G	Q493R	899	33	37	70	35	tp	fn	fn	fn	tp	fn	tn
23048	omicron	S	G23048A	G496S	894	37	27	31	27	tp	tp	tp	fp	tp	tn	tn
23055	omicron	S	A23055G	Q498R	793	29	38	37	36	fn	tp	tp	tn	tp	fn	tn
23063	omicron gamma beta alpha	S	A23063U	N501Y	1,020,863	175	280	243	242	tp	fn	fn	fn	fn	fn	fn
23075	omicron	S	U23075C	Y505H	770	33	23	21	21	tp	fn	tn	tn	tp	fp	tn
23202	omicron	S	C23202A	T547K	1082	43	81	251	80	tp	tp	tp	tp	fn	fn	fn
23271	alpha	S	C23271A	A570D	898,365	163	94	149	86	tp	tp	tp	tp	fn	fn	fn
23403	omicron gamma delta beta alpha	S	A23403G	D614G	4,589,366	193	1460	1	1	fp	tn	tn	tn	fp	fp	tn
23525	omicron gamma	S	C23525U	H655Y	92,088	130	326	299	299	tp	tp	tp	tp	tp	tp	tp
23599	omicron	S	U23599G	N679K	2425	38	36	138	34	tp	tp	tp	tp	fn	fn	tn
23604	omicron delta alpha	S	C23604A	P681H	946,888	169	309	278	290	tp	tp	tp	tp	fn	fn	tp
23664	beta	S	C23664U	A701V	53,917	128	188	184	184	tp	tp	tp	tp	tp	tp	tp
23709	alpha	S	C23709U	T716I	904,197	167	247	234	234	tp	tp	tp	tp	tp	tp	tp
23854	omicron	S	C23854A	N764K	849	27	26	200	24	tp	tp	tp	tp	fn	tn	tn
23948	omicron	S	G23948U	D796Y	3967	74	183	216	181	tp	tp	tp	tp	tp	tp	tp
24130	omicron	S	C24130A	N856K	658	32	32	314	31	tp	tp	tp	tp	fn	tp	tn
24410	delta	S	G24410A	D950N	2,689,287	169	252	107	72	tp	tp	tp	tp	fn	fn	fn
24424	omicron	S	A24424C	Q954H	5	4	4	30	4	tp	fn	fn	tn	tn	tn	tn
24469	omicron	S	U24469G	N969K	8	4	8	73	8	tp	tp	tp	fn	tn	tn	fp
24503	omicron	S	C24503U	L981F	626	28	23	31	22	tp	tp	tp	fp	tp	fp	fp
24506	alpha	S	U24506G	S982A	898,085	163	48	40	40	tp	fn	tp	fp	fn	fn	tn
24642	gamma	S	C24642U	T1027I	91,978	106	210	193	187	tp	tp	tp	tp	tp	tp	tp
24914	alpha	S	G24914C	D1118H	896,647	163	124	194	116	tp	tp	tp	tp	fn	tp	tp
25088	gamma	S	G25088U	V1176F	98,400	111	193	169	166	tp	tp	tp	tp	tp	tp	tp

[†] *On January 6, 2022*
[‡] *Number of Pango lineages*
[♯] *15/45 means that the NDRL threshold of 15 was used for the prediction, but it was evaluated with the ground truth from January 2022, using a NDRL threshold of 45. tp, fp, tn and fn mean true positive, false positive, true negative and false negative, respectively.*

Table S2. Number of neurons per layer for each model (Position prediction)

Threshold	L1	L2	L3	L4	L5	L6	L7	L8
1	1272	1963	866	451	1263	427	1764	1136
5	1927	1962	1236	2043	451	292	2048	882
10	530	2048	454	839	1508	1427	2048	219
15	1493	1676	588	413	1226	451	63	887

Table S3. Model's performance metrics - Position prediction

Our selected model													
th	set	sens	spec	auc	acc	tp	tn	fp	fn	Static sens	Raw static sens	Log loss	brier score
1	train	0.84	0.62	0.73	0.82	8929	510	309	1763	0.79	1.42	6.22	0.18
1	val	0.84	0.46	0.65	0.80	3260	223	260	637	0.70	1.26	7.07	0.20
1	test	0.84	0.46	0.65	0.80	10,052	629	727	1963	0.70	1.26	6.95	0.20
5	train	0.82	0.64	0.73	0.77	6786	2071	1186	1468	0.80	1.44	7.96	0.23
5	val	0.81	0.58	0.70	0.72	2188	982	697	513	0.77	1.38	9.54	0.28
5	test	0.80	0.60	0.70	0.73	6801	2893	1950	1727	0.44	0.80	9.50	0.27
10	train	0.80	0.69	0.74	0.74	4365	4138	1902	1106	0.44	0.80	9.03	0.26
10	val	0.79	0.63	0.71	0.69	1277	1752	1019	332	0.44	0.79	10.65	0.31
10	test	0.75	0.66	0.70	0.70	3889	5408	2743	1331	0.41	0.75	10.52	0.30
15	train	0.83	0.73	0.78	0.76	3013	5787	2105	606	0.85	1.53	8.13	0.24
15	val	0.82	0.68	0.75	0.71	829	2299	1068	184	0.82	1.48	9.87	0.29
15	test	0.79	0.69	0.74	0.71	2524	7030	3147	670	0.44	0.79	9.86	0.29
Baseline with mljar-supervised, compete mode													
th	set	sens	spec	auc	acc	tp	tn	fp	fn	Static sens	Raw static sens	Log loss	brier score
1	train	1.00	0.01	0.51	0.93	10,692	11	808	0	0.45	0.81	2.42	0.07
1	val	1.00	0.00	0.50	0.89	3897	0	483	0	0.44	0.80	3.81	0.11
1	test	1.00	0.00	0.50	0.90	12,015	2	1354	0	0.45	0.80	3.50	0.10
5	train	0.83	0.53	0.68	0.74	6833	1733	1524	1421	0.74	1.33	8.84	0.26
5	val	0.85	0.52	0.69	0.73	7291	2507	2336	1237	0.73	1.32	9.23	0.27
5	test	0.85	0.52	0.69	0.73	7291	2507	2336	1237	0.73	1.32	9.23	0.27
10	train	0.62	0.80	0.71	0.71	3377	4851	1189	2094	0.34	0.62	9.85	0.29
10	val	0.66	0.78	0.72	0.74	1057	2175	596	552	0.36	0.66	9.05	0.26
10	test	0.65	0.79	0.72	0.73	3367	6444	1707	1853	0.36	0.65	9.20	0.27

15	train	0.43	0.95	0.69	0.79	1574	7504	388	2045	0.24	0.43	7.30	0.21
15	val	0.43	0.95	0.69	0.79	1574	7504	388	2045	0.24	0.43	7.30	0.21
15	test	0.45	0.94	0.69	0.82	1425	9567	610	1769	0.25	0.45	6.15	0.18

***Th**: recurrent mutation threshold. **Sens**: sensitivity. **Spec**: specificity. **Auc**: area under the curve, **Acc**: accuracy. **Tp**: true positives. **Tn**: true negatives. **Fp**: false positives. **Fn**: false negatives. **Static sens**: Model selection metric, does not count spec while sens is lower than 0.8. Afterwards, spec is added to the maximum allowed sens 0.8. Value between 0 and 1, 1 better. **raw static sens**: Same as static sens, but no normalization. If lower than 0.8 only sens was considered. Maximum value is 1.8. **log loss**: logistic loss sum. **brier score**: uncertainty quality measurement.*

Table S4. Model's performance metrics - Mutation prediction

Our selected model													
th	set	sens	spec	auc	acc	tp	tn	fp	fn	static sens	Raw static sens	Log loss	brier score
5	train	0.81	0.80	0.81	0.80	8135	9343	2397	1852	0.89	1.60	6.75	0.20
5	val	0.80	0.63	0.72	0.71	2450	2623	1518	594	0.80	1.43	10.15	0.29
5	test	0.78	0.66	0.72	0.71	7582	8750	4563	2116	0.43	0.78	10.03	0.29
10	train	0.87	0.77	0.82	0.79	4471	12,713	3852	691	0.87	1.57	7.22	0.21
10	val	0.80	0.66	0.73	0.69	1172	3778	1950	285	0.81	1.46	10.74	0.31
10	test	0.79	0.67	0.73	0.70	3670	12,366	5993	982	0.44	0.79	10.47	0.30
15	train	0.83	0.76	0.79	0.77	2609	14,183	4384	551	0.87	1.56	7.85	0.23
15	val	0.80	0.70	0.75	0.71	712	4419	1881	173	0.83	1.50	9.87	0.29
15	test	0.80	0.72	0.76	0.73	2186	14,571	5720	534	0.84	1.52	9.39	0.27
Baseline with mljar-supervised, compete mode													
th	set	sens	spec	auc	acc	tp	tn	fp	fn	Static sens	Raw static sens	Log loss	brier score
5	train	0.76	0.83	0.79	0.80	7582	9732	2008	2405	0.42	0.76	7.02	0.20
5	val	0.80	0.76	0.78	0.78	2431	3154	987	613	0.44	0.80	7.69	0.22
5	test	0.77	0.78	0.77	0.77	7453	10,361	2952	2245	0.43	0.77	7.80	0.23
10	train	0.48	0.96	0.72	0.85	2497	15890	675	2665	0.27	0.48	5.31	0.15
10	val	0.49	0.94	0.72	0.85	720	5374	354	737	0.27	0.49	5.24	0.15
10	test	0.51	0.94	0.73	0.86	2371	17,348	1011	2281	0.28	0.51	4.94	0.14
15	train	0.51	0.96	0.74	0.90	1626	17,905	662	1534	0.29	0.51	3.49	0.10
15	val	0.55	0.91	0.73	0.87	491	5739	561	394	0.31	0.55	4.59	0.13
15	test	0.55	0.95	0.75	0.91	1491	19,357	934	1229	0.30	0.55	3.25	0.09

***Th**: recurrent mutation threshold. **Sens**: sensitivity. **Spec**: specificity. **Auc**: area under the curve, **Acc**: accuracy. **Tp**: true positives. **Tn**: true negatives. **Fp**: false positives. **Fn**: false negatives. **Static sens**: Model selection metric, does not count spec while sens is lower than 0.8. Afterwards, spec is added to the maximum allowed sens 0.8. Value between 0 and 1, 1 better. **raw static sens**: Same as static sens, but no normalization. If lower than 0.8 only sens was considered. Maximum value is 1.8. **log loss**: logistic loss sum. **brier score**: uncertainty quality measurement.*

Table S5. McNemar's test between our models and baseline models with mljar-supervised.

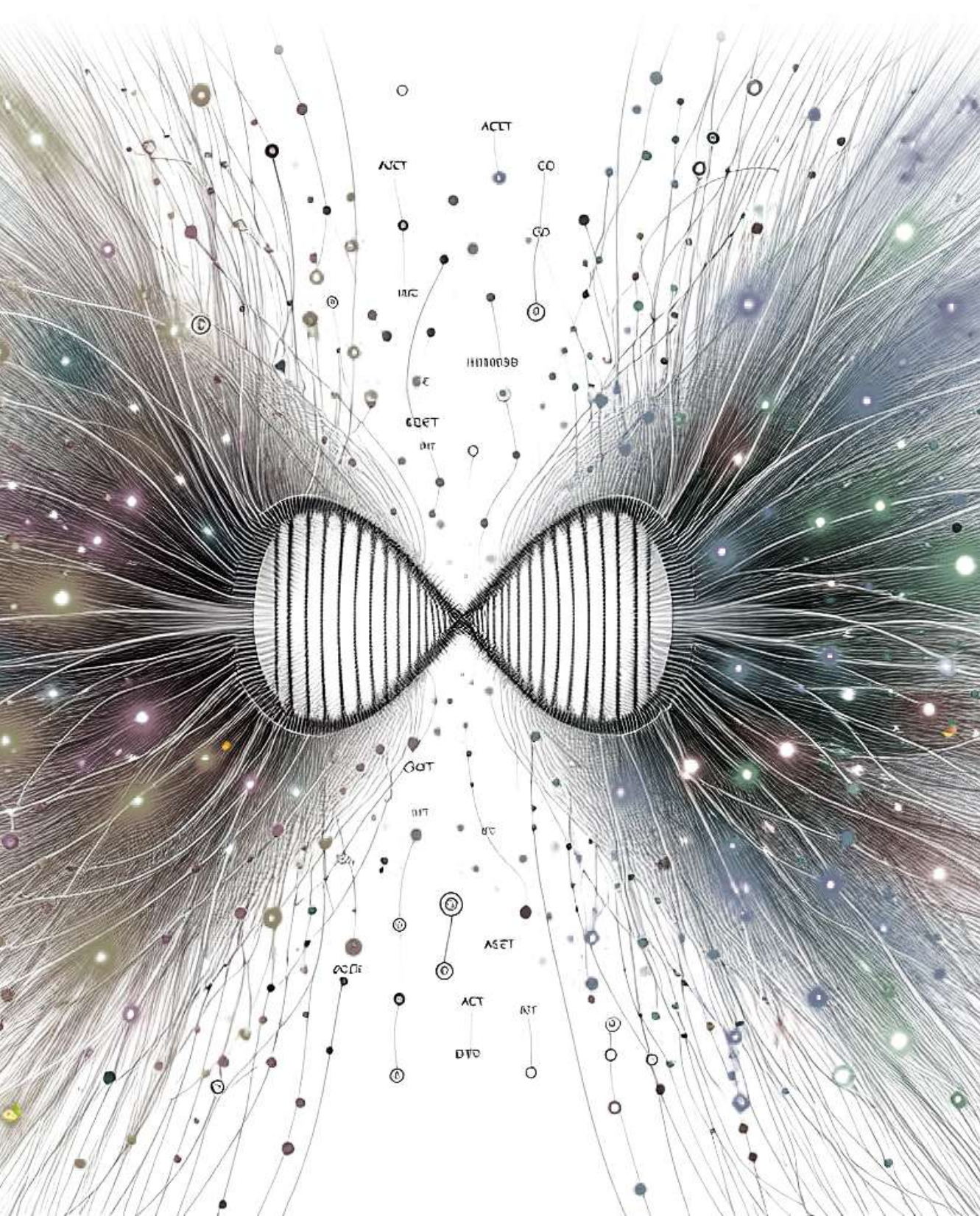
Position prediction					
th	set	chi2	p<0.05	p value round	p value
1	train	705.203	True	0.000000	2.209E-155
1	val	198.336	True	0.000000	4.818E-45
1	test	688.118	True	0.000000	1.147E-151
5	train	38.141	True	0.000000	6.582E-10
5	val	0.059	False	0.807799	8.078E-1
5	test	4.006	True	0.045327	4.534E-2
10	train	29.710	True	0.000000	5.018E-08
10	val	44.497	True	0.000000	2.547E-11
10	test	100.523	True	0.000000	1.170E-23
15	train	23.993	True	0.000001	9.669E-07
15	val	164.567	True	0.000000	1.137E-37
15	test	560.524	True	0.000000	6.467E-124
Mutation prediction					
th	set	chi2	p<0.05	p value round	p value
5	train	6.933	True	0.008460	8.4598E-3
5	val	188.671	True	0.000000	6.199E-43
5	test	512.707	True	0.000000	1.633E-113
10	train	263.506	True	0.000000	2.953E-59
10	val	603.720	True	0.000000	2.598E-133
10	test	2031.030	True	0.000000	0
15	train	1559.527	True	0.000000	0
15	val	763.524	True	0.000000	4.601E-168
15	test	3001.633	True	0.000000	0

Table S6. Model's uncertainty quality. Brier score.

Position prediction			
th	train	val	test
1	0.131931	0.159333	0.155308
5	0.172322	0.207497	0.203409
10	0.204331	0.246146	0.243646
15	0.202455	0.250927	0.249452
Mutation prediction			
th	train	val	test
5	0.248211	0.248999	0.248900
10	0.249022	0.249823	0.249600
15	0.249590	0.250531	0.250194

The brier score goes from 0 to 1. The lower the better.

Discussion



Predictive modeling can be used to understand better the transmission dynamics of SARS-CoV-2, specially when diagnostic testing practices change over time [1] and to evaluate the effects of public health interventions [2]. It can also be useful to predict demand for healthcare resources, among intensive care unit beds [3], renal replacement therapy [4], ventilators [3,4], COVID-19 diagnosis [5–12] and prognosis [3,4,13,14], SARS-CoV-2 variants of interest/variants of concern [15–17] and identification of potential inhibitors against SARS-CoV-2 [18–20]. All these applications contribute in public health planning, by helping policy makers to define better strategies [21], and as a result stop epidemic growth [22]. Among these strategies (public policies), predictive modeling also helps by finding optimal ways to prioritize vaccine distribution [23,24].

These studies use a plethora of machine learning methods. Some of them benchmark the performance across several methods and in others use just one or a few on which they leverage to reach their objectives. We have included the usage of AutoML frameworks to simplify and accelerate this benchmarking process of traditional machine learning algorithms for predicting COVID-19 mortality (Manuscript 1) and SARS-CoV-2 recurrent mutations (Manuscript 3). Nevertheless, for predicting the recurrent mutations, besides the usage of an AutoML framework we opted for a custom Artificial Neural Network (ANN) because it provided us more flexibility in how the model is trained to achieve our goal. Nevertheless, before predicting SARS-CoV-2 mutations we needed to characterize its mutational landscape (Manuscript 2) in order to know the coverage and limitations of such tasks. With this information, we were able to build a ground truth dataset against which we can validate our predictions and provide context and relevance. After characterizing the mutational landscape of SARS-CoV-2 (Manuscript 2) we observed that almost every nucleotide in the SARS-CoV-2 genome has mutated at some time. The impact of this information lies on the low feasibility to predict all mutations. Nonetheless, recurrent mutations, because of their independent origin, could be predicted.

For predicting the number of COVID-19 fatalities (Manuscript 1), we used socioeconomic, health and nutritional data from United States counties as possible predictive variables. In addition, we used the number of COVID-19 fatalities updated until September 2023 as the variable to predict. Nevertheless, the number of COVID-19 fatalities has incorporated the population factor. We noticed this when we performed correlations between the predictive variables and the number of fatalities. Many correlations had a value above 0.9, including the population variable. After applying any machine learning model for predicting this number of fatalities, the predictions from all the models had a correlation above 0.95 against the real number of COVID-19 deaths. This result made us notice that we needed to remove the population component from the target variable, COVID-19 fatalities. Therefore we normalized the number of COVID-19 fatalities by the population of each county. As a result we observed more modest correlations reaching 0.4 and -0.4 between the predictive variables and the target variable. In addition to this processing step, we also removed variables with more than 10% missing values, filled the missing values with an iterative multivariate imputation method and removed outliers. We expected that these steps were all required to perform the analyses, but there was one more pending action. After training the AutoML frameworks, we analyzed the importance of each variable applying the method of sensitivity analysis [25]. With this method, we modified one variable at a time, keeping the others with their original

values in order to know the quantity of variation in the predicted values as we vary one variable. With this information we knew that increasing some variables will also lead to an increase in the number of COVID-19 fatalities. Nevertheless, we compared this values with the correlations we first calculated. We observed that some variables had a contradictory behavior between the importance taken by the model and the correlation sign. This observation lead us into hypothesizing a problem of multicolinearity between variables and that variables with low values of magnitude in the correlation could be giving us a wrong idea of the real relationship against the target variable. In this regard, we went back to remove variables that correlate highly between each other, keeping only those with the highest correlation against the target variable. In addition we also removed those variables with a correlation magnitude lower than 0.15. Finally, we repeated the processing and training and obtained values of feature importance that agreed with the correlation sign (positive, negative). We originally planned on using sensitivity analysis [25] since the result of an AutoML framework is a collection of stacking, bagging and featuring engineering and this can not be integrated with the standard SHAP values [26] procedure for obtaining feature importance. Nonetheless, the best performing AutoML framework obtained as the best performing solution the XGBoost model, with no additional steps, just a custom configuration for the same. As a consequence, we picked XGBoost with the found optimal configuration as the final ML method for prediction, because its simplicity, equivalent predictive power as the best AutoML framework and its compatibility with the standard SHAP values procedure. The optimization process for all ML methods were done targeting the reduction of the Mean Squared Error (MSE) between the real and predicted values of the models. Once the training and validation steps were done, we obtained a correlation of 0.715 between the real and predicted number of COVID-19 fatalities. We consider the correlation instead of the MSE because it provides a more practical understanding of how all predictions are aligned with the real values. With this metric of performance in mind, we dove into the importance of the 50 used variables for training the ML method.

We found that the proportion in a population of primary care physicians (PCP) and providers (others different than physicians) were the most predictive variables. These variables included the ratio (population/PCP) and the rate (amount per 100,000 population) of these two types. Contrary to our intuition, when the amount of physicians or providers increases respect to the population size, the COVID-19 fatalities increases. This apparently contradictory relationship is also visible by just correlating these variables against the target variable (COVID-19 fatalities/Population). As is evident, correlation does not imply causation. We hypothesize that this effect is the product of population density, which will exhibit more primary care providers and physicians per population, but more concentrated and therefore promoting contagion. After these variables, we found as most important variables the median household income, the percentage of people that are physically inactive, the percentage of children in poverty, the percentage of people that drive long commutes alone and the percentage of people with diabetes. Taking into account all these variables, the accumulated importance attributed by the model adds up to 50%, from a total of 50 used variables. Besides the percentage of people with diabetes (10th position sorted by importance), among the variables related to nutrition we can find in the 18th and 22th position of importance the number of hypertension deaths of males older than 65 and the number of hypertension deaths of females between 35 and 64 years old. These two variables have an importance of

1.7% and 1.4%. An increment of any of these three variables also increase the number of COVID-19 fatalities. This behavior also agrees with the positive correlation between those variables and the target variable. The relevance of these 3 nutritional related variables also match the behavior and importance given to them for predicting COVID-19 hospitalization, need for ICU and mortality in previous research [3]. Thus, showing the value they carry for understanding disease impact, in contrast with the other nutrition related variables that we included, which have an importance lower than 1.1% after the 30th position of importance in our analysis. Previous studies that have analyzed a similar dataset for analyzing COVID-19 mortality found that a higher proportion of African-Americans in a population is a predictor of more fatalities [27–33]. Nevertheless, we took into account this variable and it is as not as predictive as previously suggested. This variable, with a correlation of 0.157 against the target, has 2% of importance (the most important variable has 8.8%) and occupies the 16th position sorted by importance. A limitation of this study is that we did not account for population density. This information could have given us a better understanding of the behavior of the proportion of primary care providers and physicians in densely populated counties. In addition, a possible extension of this study would have been how our findings could be taken into account for better epidemiological policies for populations with different characteristics. As a result of that, a better control over the propagation of the disease and its impact in the population. Another possible exercise with the employed data is to use non-supervised machine learning algorithms to create clusters and see if neighboring clusters have similar number of registered fatalities. This could provide a readiness characterization of populations against a pandemic.

Even though we have studied this pandemic through the analysis of society factors, we still needed to include in this investigation the main component of it, the virus. It is known, but required to emphasize, that a virus can change through mutations and consequently the generation of variants of it, such is the case of SARS-CoV-2 [34]. These changes belong to a mutational profile of the virus, a mutational landscape [35]. We care about this landscape of changes because some mutations might lead to higher infectivity [36] (higher affinity with host receptors), resistance against antibodies, failures in diagnostic tests and lower anti-viral response with newly identified anti-viral candidates [37]. During the pandemic we have observed these described cases embodied in Variants of Concern (VoCs), such as Alpha, Delta and Omicron [38,39]. Some VoCs like Alpha, were more transmissible than previous versions and quickly spread globally [40]. In this regard, researchers have worked on predicting amino acid mutations in SARS-CoV-2 that might contribute to future VoCs [41]. Other work was centered in predicting the impact of possible variants in the binding with the host cell receptor ACE2 and escape from antibodies [42]. To join these efforts we have analyzed the mutational landscape of SARS-CoV-2 (Manuscript 2) and incorporated this information to predict recurrent mutations (Manuscript 3). Both actions including also an analysis on VoCs.

To characterize SARS-CoV-2 mutations, we obtained around 10.4 million complete genomes of SARS-CoV-2 from the Global Initiative on Sharing All Influenza Data (GISAID) database [43], collected during more than two years of the pandemic. These genome sequences were aligned using as a reference the NC_045512.2 sequence, which was isolated from Wuhan and submitted to the GenBank database on 17 January 2020.

From these 10.4 million genomes we used only 5.3 million. The remaining part were excluded because they have low coverage. 51.9% of the high coverage genomes had as their origin USA or United Kingdom. Meanwhile, at a broader scope, all Europe accounted for 55.1% of the 5.3 million genomes. Together with North America they reached 89.2% of the total high coverage genomes. Single nucleotide variants (SNV) were the most common type of mutation, in comparison to insertions and deletions. Moreover, synonymous mutations were more frequent than non-synonymous mutations. The total number of unique SNV, unique deletions and unique insertions were approximately 73.4 thousand, 21.7 thousand and 1.8 thousand, respectively. The amount of unique SNVs, deletions and insertions that fell into untranslated regions were 1842, 248 and 120, respectively. From the SNV, 51,467 were non-synonymous, 18,413 were synonymous and 1742 were mutations that were only observed simultaneously with another mutation in the same codon. The most frequent SNV was C>U, which was found in most variants, pangolin lineages and countries. The prevalence of C>U mutations at the beginning of the pandemic suggest that host deaminases were behind an important percentage of these observed mutations. Afterwards, the leading role of host deaminases on the virus evolution has been demonstrated experimentally. Additionally, the genes that showed the highest number of mutations were the accessory genes ORF7a, 8, 6, 10, 3a and leader. From them, the gene ORF7a has the most amount of mutations, being most of them deletions. After these accessory genes, the genes encoding for the S and N proteins, showed the highest amount of non-synonymous SNVs. We expected that the S gene will have more non-synonymous mutations, since they might enhance its interaction with the Angiotensin-converting enzyme 2 (ACE2), help it to escape from the immune system, or improve furin cleavage [44–47]. On the other hand, the elevated mutation rate of the N gene could be attributed to its higher G+C proportion [48]. This high mutation rate for the N gene was far from trivial, since this gene was frequently used as a target for RT-qPCR diagnostic tests. We examined the regions that are being used by 15 of these tests, from which 9 were examining zones in the N gene. In average, there were 373 mutations in the hybridization zones used by the forward and reverse primers and probe in these 9 kits. Even though the frequency of mutations was usually low for these zones, some N gene mutations, such as the SNVs G29140U, G29179U, and C29200U, and deletions have been reported to affect RT-qPCR results [49–56]. Therefore, using primers and probes that hybridize to a region of the N gene is not an optimal choice [57]. Given these reasons, we considered tracking SARS-CoV-2 mutations as a must. In that line, we implemented the SARS-CoV-2 Mutation Portal (http://sarscov2-mutation-portal.urv.cat/, accessed on 10 May 2023) where we have registered all mutations (including point mutations, insertions, and deletions) that have been analyzed in our study.

The mutations that we have reported can be the result of errors by the RNA polymerase during virus replication or as product of host deaminases that deaminate unpaired nitrogenous bases [58–61]. Predicting those mutations produced during virus replication turns difficult because of its random-like nature. Nonetheless, predicting non-random mutations, like those caused by host deaminases [62] is feasible. Some of these non-random mutations might be recurrent mutations (RM), which are mutations that occur independently and many times throughout a virus' evolution. RMs could be the result of host RNA-editing mechanisms, like deamination, or ongoing selection [63,64]. To identify RMs, it is required to use multiple sequence alignment and a maximum likelihood tree

[63,64]. Nonetheless, this procedure is computational expensive and since our number of sequences was numerous, we used the Pango nomenclature. This nomenclature is a hierarchical and dynamic classification system based on phylogenetic evidence, that assigns each SARS-CoV-2 genome to a lineage through the usage of machine learning [65,66]. We used this nomenclature to localize and quantify the lineages where each mutation was present. Once this was done, we needed to count for each mutation in how many distantly-related lineages (DRL) they were present to quantify how recurrent they were. We developed an algorithm that counts the number of distant-related lineages (NDRL) to assign a recurrence level to mutations that emerged independently.

With this ground truth information we decided to train machine learning models that predict the position of the genome that will have a recurrent mutation and also predict how recurrent a mutation will be (Manuscript 3). First, we split the genome in two parts, a testing set formed by very important genes that were the focus in ongoing literature (M-pro, spike, PLpro, and RNA polymerase) and a training set integrated by all the other genes. The training set was futher subdivided into a training and a validation set (formed by the endoRNAse, nsp6, M and helicase genes), ensuring that the validation set comprised a comparable number of genes and had similar lengths to the testing set. The testing set was isolated during the whole training process and only used at the very end to evaluate the generalization capacity of the trained models and for evaluating the importance given to each used variable. To train the ML models, we used windows of 13 positions of the genome. We numbered the positions from -6 to +6, corresponding to the 5' to 3' direction of the genome. Each window had at the center (position 0) the position being evaluated for become recurrent. For each position in these windows, we used the corresponding nucleotide, the prediction of the secondary structure of the SARS-CoV-2 genome, the RNA normalized 2-hydroxyl acylation analyzed by primer extension (SHAPE) reactivity [67], and the translated AA sequences of the coding parts of the genome.

For predicting the position that will have a recurrent mutation, we trained 4 models. One model for predicting if a position will have a mutation with at least a NDRL of 1 (non-recurrent mutations), and three more models for predicting a NDRL of 5, 10 and 15. In addition, three more models for predicting if a mutation will have a NDRL of 5, 10 and 15. We used different NDRL thresholds to overcome potential sequencing errors, artefactual biases, and other causes, such as recombination, which may lead to homoplasies [64,68]. During the evaluation of the models in the validation set, we observed that increasing the NDRL improved performance, particularly in specificity (true negative rate). Consequently, we developed a custom loss function to maintain a sensitivity (true positive rate) close to 0.8 while optimizing for the highest possible specificity under this constrain. We implemented this approach because we anticipated having false positives in our predictions, as some mutations we predicted as recurrent had not yet been observed in the population. Therefore, we believed it was more prudent not to optimize based on sensitivity and instead enhance specificity. Subsequently, the use of an updated test set validated our hypothesis, revealing that a significant portion of the initial false positives transformed into true positives with more current data. We tried the MLJAR [69] AutoML framework in the compete mode (optimized for a better performance disregarding interpretability), Convolutional Neural Networks (CNN) [70], transformers (self-attention) [71] and an Artificial Neural Network (ANN). The

performance between the CNN, transformers and the ANN was similar. Nevertheless, MLJAR had a considerable lower performance for low values of NDRL in comparison to the others, for higher values (NDRL>=10) it was more close, but still lower. It did not provide any advantage with the additional cost of low interpretability. We opted for the ANN method since it provided more information for feature interpretation than a CNN and it was faster to train in comparison with a transformer. For searching the best architecture, we tried a Gaussian Optimization process and a random search. Theoretically the Gaussian Optimization was expected to provide better results but we found the better performance with the random search. Since the predictions made by the resulting ANN and MLJAR for higher NDRLs was apparently similar, we performed a McNemar's test [72] to compare two classifiers, showing that the differences are significant ($p<0.05$).

The prediction whether a position will undergo a mutation in at least 1 NDRL (non-recurrent mutation) yielded an area under the curve (AUC) of the receiver operating characteristic (ROC) of 0.74. Meanwhile, for NDRL values of 5, 10 and 15, the ROC-AUC were 0.78, 0.78 and 0.81, respectively. The corresponding specificity for those NDRL were 0.46, 0.6, 0.66 and 0.69. These results showed that mutations which are more recurrent are easier to predict. On the other hand, mutations with a more random-like nature (NDRL of 1) are more difficult to predict. Moreover, predicting if a specific mutation (e.g. C1764U), not just the position, will have a NDRL of 15, is much easier and has 5% higher performance than predicting if that position will have any RM with that NDRL. This is reflected in the ROC-AUC of 0.79, 0.81 and 0.84 that we obtained for the NDRLs of 5, 10 and 15, respectively. All the reported metrics are for the testing set (M-pro, spike, PLpro, and RNA polymerase genes).

We hypothesized that some of the false positives in our predictions would become true positives later. Thus, we evaluated the model predictions from April 2021 (training date) against data updated after 9 months (january 2022). Nevertheless, first we needed to calibrate the equivalence of NDRLs between both dates and we saw that mutations that had a NDRL of 15 in April 2021, had in average a NDRL of 45 in January 2022. Therefore, we evaluated the predictions of the model trained with a NDRL of 15 against the updated ground truth with a NDRL of 45. The ROC-AUC for predicting positions with a NDRL of 45 was 0.8, and the number of false positives that turned into true positives were 557, confirming our speculation. In addition, we also evaluated our predictions in the VoCs Alpha, Beta, Delta, Gamma and Omicron, using the ground truth of January 2022. This means, evaluating if our predictions could hold in the future for anticipating VoCs. We obtained a sensitivity of 0.776 and a specificity of 0.667, considering all mutations that all mentioned VoC have joined together. These metrics correspond to how well we predicted the positions that define each variant for a NDRL of 45. On the other hand, for the same value of NDRL, we obtained a sensitivity of 0.5 and a specificity of 0.842 for predicting the precise nucleotide changes (e.g C14408U). The mutation C14408U, present in all VoCs and that codes for the RNA polymerase P323L shift, is a RM. This mutation was found in the early stages of the pandemic and was spread and found in more than 99% of SARS-CoV-2 genomes available until January 2022. We predicted this mutation as a RM. Nevertheless, since this mutation is present in all Pango lineages, it is not categorized as a RM.

Some challenging examples are the A23063U and U22917G mutations, that were present in more than 1 million and 2 million of SARS-CoV-2 genomes (by January 2022) and in more than 280 pangolin lineages. None of our models was able to predict them correctly. This kind of mutations, that strengthen SARS-CoV-2 infection and transmission, are the most difficult to predict because they could not be caused by host deaminases and we did not tailor our prediction models to detect them.

After these evaluations, we were interested in finding out what variables were taken into account for making a prediction. We analyzed with more attention mutations with a NDRL of 15, since they deviate further from a random-like nature. We obtained how important is each variable for the models through the usage of SHapley Additive exPlanation (SHAP) values. When predicting the positions that will have a RM, at all levels of NDRL, the models gave more importance to the nucleotide in the center position. It is important to mention that there was no hard-coded information, in the data used for training, about which of the nucleotides in the window was the one that might have the mutation. All models learned this through the training process. The main difference across models trained for different NDRLs lies on the magnitude attributed to this center position. For low values of NDRL, the importance given to other variables is more homogeneous and closer to the importance given to the nucleotide in the middle. On the other hand, as the NDRL increases, the total importance is more concentrated on the nucleotide in the central position and the difference against other variables is higher and far from subtle.
Analyzing the predictions for positions with a NDRL of 15, we saw that true positives have in the central position a cytosine and with a lower degree a guanine. Meanwhile uracils and adenines in this position are uncommon. The nucleotide in the central position in false positives are guanine (35%) or a cytosine (25%), and in true negatives are adenine (46%) or uracil (45%). Nucleotides in other positions are less relevant, but those in the upstream and downstream positions (P-1 and P1, respectively), from the position holding the RM, are the most relevant. It has been reported that cytosines of the UC and AC motifs of the SARS-CoV-2 genome are preferentially deaminated by the APOBEC3A and APOBEC1 enzymes [62]. This matches our findings, since 44% and 27% of the true positives have an adenine or an uracil at P-1, and in 37% of the cases, we find an uracil at P1.

After nucleotides, across all levels of NDRL, the most important type of variable is the SHAPE-Seq reactivity. Even though, it has about a fifth of nucleotide's importance. At most positions, low values of this variable do not promote mutagenesis. Nevertheless, higher values do. The importance given to nucleotides in the central position and reactivity values match with a model that mainly predicts cytosines in a ACU motif as a RM in zones of the RNA structure that makes the cytosine more reactive. These predictions, patterns and attributed importance given by the model agrees with the SARS-CoV-2 mutational bias pattern and the predilection of some host deaminases for particular sequences and RNA secondary structures [62,68,73].

REFERENCES

1. Khan, A.A.; Sbihi, H.; Irvine, M.A.; Jassem, A.N.; Joffres, Y.; Klaver, B.; Janjua, N.; Bharmal, A.; Ng, C.H.; Wilmer, A.; et al. Prediction of SARS-CoV-2 Transmission Dynamics Based on Population-Level Cycle Threshold Values: A Machine Learning and Mechanistic Modeling Study 2023, 2023.03.06.23286837.
2. Wang, W.; Chen, Y.; Wang, Q.; Cai, P.; He, Y.; Hu, S.; Wu, Y.; Wenxiang, W. The Transmission Dynamics of SARS-COV-2 in China: Modeling Study and the Impact of Public Health Interventions 2020.
3. Wollenstein-Betech, S.; Cassandras, C.G.; Paschalidis, I.Ch. Personalized Predictive Models for Symptomatic COVID-19 Patients Using Basic Preconditions: Hospitalizations, Mortality, and the Need for an ICU or Ventilator. Int. J. Med. Inf. 2020, 142, 104258, doi:10.1016/j.ijmedinf.2020.104258.
4. Rodriguez, V.A.; Bhave, S.; Chen, R.; Pang, C.; Hripcsak, G.; Sengupta, S.; Elhadad, N.; Green, R.; Adelman, J.; Metitiri, K.S.; et al. Development and Validation of Prediction Models for Mechanical Ventilation, Renal Replacement Therapy, and Readmission in COVID-19 Patients. J. Am. Med. Inform. Assoc. 2021, 28, 1480–1488, doi:10.1093/jamia/ocab029.
5. Harikrishnan, N.B.; Pranay, S.Y.; Nagaraj, N. Classification of SARS-CoV-2 Viral Genome Sequences Using Neurochaos Learning. Med. Biol. Eng. Comput. 2022, 60, 2245–2255, doi:10.1007/s11517-022-02591-3.
6. Gecgel, O.; Ramanujam, A.; Botte, G.G. Selective Electrochemical Detection of SARS-CoV-2 Using Deep Learning. Viruses 2022, 14, 1930, doi:10.3390/v14091930.
7. Du, R.; Tsougenis, E.D.; Ho, J.W.K.; Chan, J.K.Y.; Chiu, K.W.H.; Fang, B.X.H.; Ng, M.Y.; Leung, S.-T.; Lo, C.S.Y.; Wong, H.-Y.F.; et al. Machine Learning Application for the Prediction of SARS-CoV-2 Infection Using Blood Tests and Chest Radiograph. Sci. Rep. 2021, 11, 14250, doi:10.1038/s41598-021-93719-2.
8. Gómez-Rojas, S.; Segura, G.P.; Ollé, J.; Carreño Gómez-Tarragona, G.; Medina, J.G.; Aguado, J.M.; Guerrero, E.V.; Santaella, M.P.; Martínez-López, J. A Machine Learning Tool for the Diagnosis of SARS-CoV-2 Infection from Hemogram Parameters. J. Cell. Mol. Med. 2023, 27, 3423–3430, doi:10.1111/jcmm.17864.
9. Salman, A.O.; Geman, O. Evaluating Three Machine Learning Classification Methods for Effective COVID-19 Diagnosis. Int. J. Math. Stat. Comput. Sci. 2023, 1, 1–14, doi:10.59543/ijmscs.v1i.7693.
10. Tschoellitsch, T.; Dünser, M.; Böck, C.; Schwarzbauer, K.; Meier, J. Machine Learning Prediction of SARS-CoV-2 Polymerase Chain Reaction Results with Routine Blood Tests. Lab. Med. 2021, 52, 146–149, doi:10.1093/labmed/lmaa111.
11. Dritsas, E.; Trigka, M. Supervised Machine Learning Models to Identify Early-Stage Symptoms of SARS-CoV-2. Sensors 2023, 23, 40, doi:10.3390/s23010040.
12. Monaghan, C.K.; Larkin, J.W.; Chaudhuri, S.; Han, H.; Jiao, Y.; Bermudez, K.M.; Weinhandl, E.D.; Dahne-Steuber, I.A.; Belmonte, K.; Neri, L.; et al. Machine Learning for Prediction of Patients on Hemodialysis with an Undetected SARS-CoV-2 Infection. Kidney360 2021, 2, 456, doi:10.34067/KID.0003802020.
13. Casano, N.; Santini, S.J.; Vittorini, P.; Sinatti, G.; Carducci, P.; Mastroianni, C.M.; Ciardi, M.R.; Pasculli, P.; Petrucci, E.; Marinangeli, F.; et al. Application of Machine Learning Approach in Emergency Department to Support Clinical Decision Making for SARS-CoV-2 Infected Patients. J. Integr. Bioinforma. 2023, 20, doi:10.1515/jib-2022-0047.
14. Zucco, A.G.; Agius, R.; Svanberg, R.; Moestrup, K.S.; Marandi, R.Z.; MacPherson, C.R.; Lundgren, J.; Ostrowski, S.R.; Niemann, C.U. Personalized Survival Probabilities for SARS-CoV-2 Positive Patients by Explainable Machine Learning. Sci. Rep. 2022, 12, 13879, doi:10.1038/s41598-022-17953-y.
15. Li, L.; Li, C.; Li, N.; Zou, D.; Zhao, W.; Xue, Y.; Zhang, Z.; Bao, Y.; Song, S. Machine Learning Detection of SARS-CoV-2 High-Risk Variants 2023, 2023.04.19.537460.

16. Nicora, G.; Mahmi, S.; Salemi, M.; Bellazzi, R. Dynamic Prediction of Non-Neutral SARS-Cov-2 Variants Using Incremental Machine Learning. In Challenges of Trustable AI and Added-Value on Health; IOS Press, 2022; pp. 654–658.
17. Mwanga, M.J.; Obura, H.O.; Evans, M.; Awe, O.I. Enhanced Deep Convolutional Neural Network for SARS-CoV-2 Variants Classification 2023, 2023.08.09.552643.
18. Gawriljuk, V.O.; Zin, P.P.K.; Puhl, A.C.; Zorn, K.M.; Foil, D.H.; Lane, T.R.; Hurst, B.; Tavella, T.A.; Costa, F.T.M.; Lakshmanane, P.; et al. Machine Learning Models Identify Inhibitors of SARS-CoV-2. J. Chem. Inf. Model. 2021, 61, 4224–4235, doi:10.1021/acs.jcim.1c00683.
19. Bucinsky, L.; Bortňák, D.; Gall, M.; Matúška, J.; Milata, V.; Pitoňák, M.; Štekláč, M.; Végh, D.; Zajaček, D. Machine Learning Prediction of 3CLpro SARS-CoV-2 Docking Scores. Comput. Biol. Chem. 2022, 98, 107656, doi:10.1016/j.compbiolchem.2022.107656.
20. Liang, J.; Zheng, Y.; Tong, X.; Yang, N.; Dai, S. In Silico Identification of Anti-SARS-CoV-2 Medicinal Plants Using Cheminformatics and Machine Learning. Molecules 2023, 28, 208, doi:10.3390/molecules28010208.
21. Payedimarri, A.B.; Concina, D.; Portinale, L.; Canonico, M.; Seys, D.; Vanhaecht, K.; Panella, M. Prediction Models for Public Health Containment Measures on COVID-19 Using Artificial Intelligence and Machine Learning: A Systematic Review. Int. J. Environ. Res. Public. Health 2021, 18, 4499, doi:10.3390/ijerph18094499.
22. Pataro, I.M.L.; Oliveira, J.F.; Morato, M.M.; Amad, A.A.S.; Ramos, P.I.P.; Pereira, F.A.C.; Silva, M.S.; Jorge, D.C.P.; Andrade, R.F.S.; Barreto, M.L.; et al. A Control Framework to Optimize Public Health Policies in the Course of the COVID-19 Pandemic. Sci. Rep. 2021, 11, 13403, doi:10.1038/s41598-021-92636-8.
23. Piraveenan, M.; Sawleshwarkar, S.; Walsh, M.; Zablotska, I.; Bhattacharyya, S.; Farooqui, H.H.; Bhatnagar, T.; Karan, A.; Murhekar, M.; Zodpey, S.; et al. Optimal Governance and Implementation of Vaccination Programmes to Contain the COVID-19 Pandemic. R. Soc. Open Sci. 2021, 8, 210429, doi:10.1098/rsos.210429.
24. Bertsimas, D.; Ivanhoe, J.; Jacquillat, A.; Li, M.; Previero, A.; Lami, O.S.; Bouardi, H.T. Optimizing Vaccine Allocation to Combat the COVID-19 Pandemic 2020, 2020.11.17.20233213.
25. Pizarroso, J.; Portela, J.; Muñoz, A. NeuralSens: Sensitivity Analysis of Neural Networks. J. Stat. Softw. 2022, 102, 1–36, doi:10.18637/jss.v102.i07.
26. Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. ArXiv170507874 Cs Stat 2017.
27. Ruck, D.J.; Bentley, R.A.; Borycz, J. Early Warning of Vulnerable Counties in a Pandemic Using Socio-Economic Variables. Econ. Hum. Biol. 2021, 41, 100988, doi:10.1016/j.ehb.2021.100988.
28. Boserup, B.; McKenney, M.; Elkbuli, A. Disproportionate Impact of COVID-19 Pandemic on Racial and Ethnic Minorities. Am. Surg. 2020, 86, 1615–1622, doi:10.1177/0003134820973356.
29. McCoy, D.; Mgbara, W.; Horvitz, N.; Getz, W.M.; Hubbard, A. Ensemble Machine Learning of Factors Influencing COVID-19 across US Counties. Sci. Rep. 2021, 11, 11777, doi:10.1038/s41598-021-90827-x.
30. Clouston, S.A.P.; Natale, G.; Link, B.G. Socioeconomic Inequalities in the Spread of Coronavirus-19 in the United States: A Examination of the Emergence of Social Inequalities. Soc. Sci. Med. 2021, 268, 113554, doi:10.1016/j.socscimed.2020.113554.
31. Hawkins, R.B.; Charles, E.J.; Mehaffey, J.H. Socio-Economic Status and COVID-19–Related Cases and Fatalities. Public Health 2020, 189, 129–134, doi:10.1016/j.puhe.2020.09.016.
32. Tan, S.B.; deSouza, P.; Raifman, M. Structural Racism and COVID-19 in the USA: A County-Level Empirical Analysis. J. Racial Ethn. Health Disparities 2022, 9, 236–246, doi:10.1007/s40615-020-00948-8.

33. Marvel, S.W.; House, J.S.; Wheeler, M.; Song, K.; Zhou, Y.; Wright, F.A.; Chiu, W.A.; Rusyn, I.; Motsinger-Reif, A.; Reif, D.M. The COVID-19 Pandemic Vulnerability Index (PVI) Dashboard: Monitoring County-Level Vulnerability Using Visualization, Statistical Modeling, and Machine Learning 2020, 2020.08.10.20169649.

34. CDC Coronavirus Disease 2019 (COVID-19) Available online: https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-info.html (accessed on 8 November 2021).

35. Sasaki, A.; Nowak, M.A. Mutation Landscapes. J. Theor. Biol. 2003, 224, 241–247, doi:10.1016/S0022-5193(03)00161-9.

36. Plante, J.A.; Liu, Y.; Liu, J.; Xia, H.; Johnson, B.A.; Lokugamage, K.G.; Zhang, X.; Muruato, A.E.; Zou, J.; Fontes-Garfias, C.R.; et al. Spike Mutation D614G Alters SARS-CoV-2 Fitness. Nature 2021, 592, 116–121, doi:10.1038/s41586-020-2895-3.

37. Singh, J.; Samal, J.; Kumar, V.; Sharma, J.; Agrawal, U.; Ehtesham, N.Z.; Sundar, D.; Rahman, S.A.; Hira, S.; Hasnain, S.E. Structure-Function Analyses of New SARS-CoV-2 Variants B.1.1.7, B.1.351 and B.1.1.28.1: Clinical, Diagnostic, Therapeutic and Public Health Implications. Viruses 2021, 13, 439, doi:10.3390/v13030439.

38. Farhud, D.D.; Mojahed, N. SARS-COV-2 Notable Mutations and Variants: A Review Article. Iran. J. Public Health 2022, doi:10.18502/ijph.v51i7.10083.

39. Khatri, R.; Siddqui, G.; Sadhu, S.; Maithil, V.; Vishwakarma, P.; Lohiya, B.; Goswami, A.; Ahmed, S.; Awasthi, A.; Samal, S. Intrinsic D614G and P681R/H Mutations in SARS-CoV-2 VoCs Alpha, Delta, Omicron and Viruses with D614G plus Key Signature Mutations in Spike Protein Alters Fusogenicity and Infectivity. Med. Microbiol. Immunol. (Berl.) 2023, 212, 103–122, doi:10.1007/s00430-022-00760-7.

40. Perez-Gomez, R. The Development of SARS-CoV-2 Variants: The Gene Makes the Disease. J. Dev. Biol. 2021, 9, 58, doi:10.3390/jdb9040058.

41. Maher, M.C.; Bartha, I.; Weaver, S.; Iulio, J. di; Ferri, E.; Soriaga, L.; Lempp, F.A.; Hie, B.L.; Bryson, B.; Berger, B.; et al. Predicting the Mutational Drivers of Future SARS-CoV-2 Variants of Concern; 2021; p. 2021.06.21.21259286;

42. Taft, J.M.; Weber, C.R.; Gao, B.; Ehling, R.A.; Han, J.; Frei, L.; Metcalfe, S.W.; Yermanos, A.; Kelton, W.; Reddy, S.T. Predictive Profiling of SARS-CoV-2 Variants by Deep Mutational Learning 2021, 2021.12.07.471580.

43. Khare, S.; Gurry, C.; Freitas, L.; Schultz, M.B.; Bach, G.; Diallo, A.; Akite, N.; Ho, J.; Lee, R.T.; Yeo, W.; et al. GISAID's Role in Pandemic Response. China CDC Wkly. 2021, 3, 1049–1051, doi:10.46234/ccdcw2021.255.

44. Harvey, W.T.; Carabelli, A.M.; Jackson, B.; Gupta, R.K.; Thomson, E.C.; Harrison, E.M.; Ludden, C.; Reeve, R.; Rambaut, A.; Peacock, S.J.; et al. SARS-CoV-2 Variants, Spike Mutations and Immune Escape. Nat. Rev. Microbiol. 2021, 19, 409–424, doi:10.1038/s41579-021-00573-0.

45. Souza, P.F.N.; Mesquita, F.P.; Amaral, J.L.; Landim, P.G.C.; Lima, K.R.P.; Costa, M.B.; Farias, I.R.; Belém, M.O.; Pinto, Y.O.; Moreira, H.H.T.; et al. The Spike Glycoprotein of SARS-CoV-2: A Review of How Mutations of Spike Glycoproteins Have Driven the Emergence of Variants with High Transmissibility and Immune Escape. Int. J. Biol. Macromol. 2022, 208, 105–125, doi:10.1016/j.ijbiomac.2022.03.058.

46. Cassari, L.; Pavan, A.; Zoia, G.; Chinellato, M.; Zeni, E.; Grinzato, A.; Rothenberger, S.; Cendron, L.; Dettin, M.; Pasquato, A. SARS-CoV-2 S Mutations: A Lesson from the Viral World to Understand How Human Furin Works. Int. J. Mol. Sci. 2023, 24, 4791, doi:10.3390/ijms24054791.

47. He, X.; He, C.; Hong, W.; Yang, J.; Wei, X. Research Progress in Spike Mutations of SARS-CoV-2 Variants and Vaccine Development. Med. Res. Rev. 2023, 43, 932–971, doi:10.1002/med.21941.

48. Ravi, V.; Swaminathan, A.; Yadav, S.; Arya, H.; Pandey, R. SARS-CoV-2 Variants of Concern and Variations within Their Genome Architecture: Does Nucleotide Distribution and Mutation Rate Alter the Functionality and Evolution of the Virus? Viruses 2022, 14, 2499, doi:10.3390/v14112499.

49. Ziegler, K.; Steininger, P.; Ziegler, R.; Steinmann, J.; Korn, K.; Ensser, A. SARS-CoV-2 Samples May Escape Detection Because of a Single Point Mutation in the N Gene. Eurosurveillance 2020, 25, 2001650, doi:10.2807/1560-7917.ES.2020.25.39.2001650.

50. Vanaerschot, M.; Mann, S.A.; Webber, J.T.; Kamm, J.; Bell, S.M.; Bell, J.; Hong, S.N.; Nguyen, M.P.; Chan, L.Y.; Bhatt, K.D.; et al. Identification of a Polymorphism in the N Gene of SARS-CoV-2 That Adversely Impacts Detection by Reverse Transcription-PCR. J. Clin. Microbiol. 2020, 59, 10.1128/jcm.02369-20, doi:10.1128/jcm.02369-20.

51. Hasan, R.; Hossain, M.E.; Miah, M.; Hasan, M.M.; Rahman, M.; Rahman, M.Z. Identification of Novel Mutations in the N Gene of SARS-CoV-2 That Adversely Affect the Detection of the Virus by Reverse Transcription-Quantitative PCR. Microbiol. Spectr. 2021, 9, 10.1128/spectrum.00545-21, doi:10.1128/spectrum.00545-21.

52. Zannoli, S.; Dirani, G.; Taddei, F.; Gatti, G.; Poggianti, I.; Denicolò, A.; Arfilli, V.; Manera, M.; Mancini, A.; Battisti, A.; et al. A Deletion in the N Gene May Cause Diagnostic Escape in SARS-CoV-2 Samples. Diagn. Microbiol. Infect. Dis. 2022, 102, 115540, doi:10.1016/j.diagmicrobio.2021.115540.

53. Laine, P.; Nihtilä, H.; Mustanoja, E.; Lyyski, A.; Ylinen, A.; Hurme, J.; Paulin, L.; Jokiranta, S.; Auvinen, P.; Meri, T. SARS-CoV-2 Variant with Mutations in N Gene Affecting Detection by Widely Used PCR Primers. J. Med. Virol. 2022, 94, 1227–1231, doi:10.1002/jmv.27418.

54. Miller, S.; Lee, T.; Merritt, A.; Pryce, T.; Levy, A.; Speers, D. Single-Point Mutations in the N Gene of SARS-CoV-2 Adversely Impact Detection by a Commercial Dual Target Diagnostic Assay. Microbiol. Spectr. 2021, 9, e01494-21, doi:10.1128/Spectrum.01494-21.

55. Isabel, S.; Abdulnoor, M.; Boissinot, K.; Isabel, M.R.; de Borja, R.; Zuzarte, P.C.; Sjaarda, C.P.; R. Barker, K.; Sheth, P.M.; Matukas, L.M.; et al. Emergence of a Mutation in the Nucleocapsid Gene of SARS-CoV-2 Interferes with PCR Detection in Canada. Sci. Rep. 2022, 12, 10867, doi:10.1038/s41598-022-13995-4.

56. Kami, W.; Kinjo, T.; Hashioka, H.; Arakaki, W.; Uechi, K.; Takahashi, A.; Oki, H.; Tanaka, K.; Motooka, D.; Nakamura, S.; et al. Impact of G29179T Mutation on Two Commercial PCR Assays for SARS-CoV-2 Detection. J. Virol. Methods 2023, 314, 114692, doi:10.1016/j.jviromet.2023.114692.

57. Wang, R.; Hozumi, Y.; Yin, C.; Wei, G.-W. Mutations on COVID-19 Diagnostic Targets. Genomics 2020, 112, 5204–5213, doi:10.1016/j.ygeno.2020.09.028.

58. Duffy, S.; Shackelton, L.A.; Holmes, E.C. Rates of Evolutionary Change in Viruses: Patterns and Determinants. Nat. Rev. Genet. 2008, 9, 267–276, doi:10.1038/nrg2323.

59. Simmonds, P.; Ansari, M.A. Extensive C->U Transition Biases in the Genomes of a Wide Range of Mammalian RNA Viruses; Potential Associations with Transcriptional Mutations, Damage- or Host-Mediated Editing of Viral RNA. PLOS Pathog. 2021, 17, e1009596, doi:10.1371/journal.ppat.1009596.

60. Ratcliff, J.; Simmonds, P. Potential APOBEC-Mediated RNA Editing of the Genomes of SARS-CoV-2 and Other Coronaviruses and Its Impact on Their Longer Term Evolution. Virology 2021, 556, 62–72, doi:10.1016/j.virol.2020.12.018.

61. Giorgio, S.D.; Martignano, F.; Torcia, M.G.; Mattiuz, G.; Conticello, S.G. Evidence for Host-Dependent RNA Editing in the Transcriptome of SARS-CoV-2. Sci. Adv. 2020, doi:10.1126/sciadv.abb5813.

62. Kim, K.; Calabrese, P.; Wang, S.; Qin, C.; Rao, Y.; Feng, P.; Chen, X.S. The Roles of APOBEC-Mediated RNA Editing in SARS-CoV-2 Mutations, Replication and Fitness 2022, 2021.12.18.473309.

63. van Dorp, L.; Richard, D.; Tan, C.C.S.; Shaw, L.P.; Acman, M.; Balloux, F. No Evidence for Increased Transmissibility from Recurrent Mutations in SARS-CoV-2. Nat. Commun. 2020, 11, 5986, doi:10.1038/s41467-020-19818-2.

64. van Dorp, L.; Acman, M.; Richard, D.; Shaw, L.P.; Ford, C.E.; Ormond, L.; Owen, C.J.; Pang, J.; Tan, C.C.S.; Boshier, F.A.T.; et al. Emergence of Genomic Diversity and Recurrent Mutations in SARS-CoV-2. Infect. Genet. Evol. 2020, 83, 104351, doi:10.1016/j.meegid.2020.104351.

65. Rambaut, A.; Holmes, E.C.; O'Toole, Á.; Hill, V.; McCrone, J.T.; Ruis, C.; du Plessis, L.; Pybus, O.G. A Dynamic Nomenclature Proposal for SARS-CoV-2 Lineages to Assist Genomic Epidemiology. Nat. Microbiol. 2020, 5, 1403–1407, doi:10.1038/s41564-020-0770-5.

66. O'Toole, Á.; Scher, E.; Underwood, A.; Jackson, B.; Hill, V.; McCrone, J.T.; Colquhoun, R.; Ruis, C.; Abu-Dahab, K.; Taylor, B.; et al. Assignment of Epidemiological Lineages in an Emerging Pandemic Using the Pangolin Tool. Virus Evol. 2021, 7, veab064, doi:10.1093/ve/veab064.

67. Manfredonia, I.; Nithin, C.; Ponce-Salvatierra, A.; Ghosh, P.; Wirecki, T.K.; Marinus, T.; Ogando, N.S.; Snijder, E.J.; van Hemert, M.J.; Bujnicki, J.M.; et al. Genome-Wide Mapping of SARS-CoV-2 RNA Structures Identifies Therapeutically-Relevant Elements. Nucleic Acids Res. 2020, 48, 12436–12452, doi:10.1093/nar/gkaa1053.

68. Turakhia, Y.; Maio, N.D.; Thornlow, B.; Gozashti, L.; Lanfear, R.; Walker, C.R.; Hinrichs, A.S.; Fernandes, J.D.; Borges, R.; Slodkowicz, G.; et al. Stability of SARS-CoV-2 Phylogenies. PLOS Genet. 2020, 16, e1009175, doi:10.1371/journal.pgen.1009175.

69. Plońska, A.; Ploński, P. MLJAR: State-of-the-Art Automated Machine Learning Framework for Tabular Data. Version 0.10.3 2021.

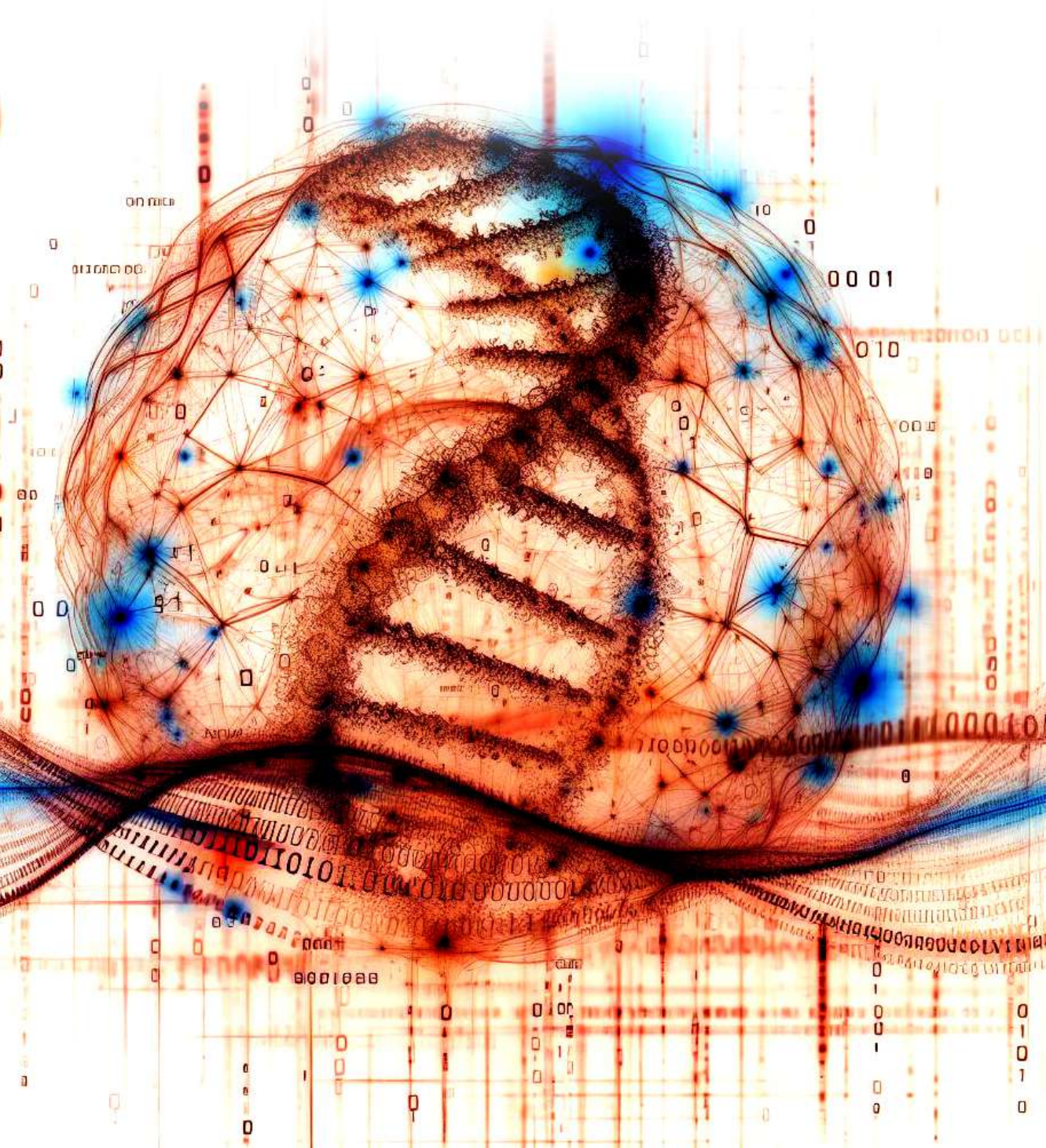
70. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-Based Learning Applied to Document Recognition. Proc. IEEE 1998, 86, 2278–2324, doi:10.1109/5.726791.

71. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. ArXiv170603762 Cs 2017.

72. Sundjaja, J.H.; Shrestha, R.; Krishan, K. McNemar And Mann-Whitney U Tests. In StatPearls; StatPearls Publishing: Treasure Island (FL), 2024.

73. Harris, R.S.; Dudley, J.P. APOBECs and Virus Restriction. Virology 2015, 479–480, 131–145, doi:10.1016/j.virol.2015.03.012.

Conclusions



From this thesis and in alignment with the objective 1, *to develop a predictive model for COVID-19 mortality*, we can conclude:

- It is possible to predict the COVID-19 mortality at the US county level through the integration of health, socioeconomic and nutritional data. Our developed predictive model, utilizing XGBoost, exhibits a notable correlation of 0.715 with the actual count of COVID-19 deaths.
- Alternative AutoML predictive models, such as MLJAR, TabPFN, and TPOT, demonstrated comparable or slightly superior performances compared to XGBoost alone. Nevertheless, XGBoost was chosen for its simplicity and the straightforward ability to obtain the importance of the variables utilized.

In addition, regarding the objective 2, *to to identify the most influential health, socioeconomic, and nutritional factors that have a significant impact on COVID-19 mortality*, we can draw the following conclusions:

- The SHAP values of the selected predictive model enable us to pinpoint the most influencial variables. The trends in variable importance and their impact were consistent with the correlation of these variables with COVID-19 mortality rates.
- The variables with the most predictive power were those related to the proportion of primary care physicians and other health providers (nurse practitioners, physician assistants and clinical nurse specialists) relative to the population. The variables 'primary care providers ratio' and 'other primary care providers rate/ratio' together contribute to 21% of the overall importance, captured from a total of 50 variables. Surprisingly, an elevated rate of other primary care providers (other primary care providers per 100,000 population) was linked to an increase in deaths caused by COVID-19.
- 50% of the overall importance of the predictive model is captured with the addition of the following variables: 'median household income', 'percentage of people that are physically inactive', 'percentage of children in poverty', 'percentage of people that drives alone and have long commute time' and 'percentage of people with diabetes'.
- Variables related to metabolism or nutrition showed little importance in influencing the predictive power. Notably, the 'percentage of people with diabetes' ranked as the 10th most influential variable in predicting US county COVID-19 fatalities, with an importance of 3%, compared to the most critical variable at 8.8%. Additionally, the 'number of hypertension-related deaths in males older than 65' and 'females between 35 and 64 years old' claimed the 18th and 22nd positions in importance, respectively. Importantly, an increase in these three variables correlates with a corresponding rise in the number of COVID-19 fatalities.
- In comparison to previous studies that found the proportion of African-Americans in the population as one of the most predictive variables for COVID-19 fatalities, we did not find it but until the 16th position, sorted by importance.

Furthermore, considering the goal of objective 3, *to characterize SARS-CoV-2 mutations and identify recurrent mutations*, we arrive to these conclusions:

- From the analysis of more than 5.3 million complete and high-coverage SARS-CoV-2 genomes from the GISAID database available on June 2022, we found 73,464 single nucleotide variants (SNV). 51,467 of them were non-synonymous and 18,413 were synonymous mutations. However, synonymous mutations in SARS-CoV-2 were the most frequent.

• The most frequent SARS-CoV-2 mutation was C>U. These mutations are present in the largest number of variants, pangolin lineages, and countries. The prominence of C>U mutations during the initial phases of the COVID-19 pandemic suggests that host deaminases played a substantial role in generating a considerable percentage of these mutations.

• Not all SARS-CoV-2 genes have accumulated the same number of mutations. There are fewer non-synonymous SNV in genes that encode proteins that play critical roles in virus replication, e.g., helicase, RdRp, and main protease (M-pro), than in genes with accessory functions (e.g., ORF7a, ORF8, and ORF6). Genes encoding S and N proteins have more non-synonymous SNVs than other genes.

• Although the prevalence of mutations in the target regions of COVID-19 diagnostic RT-qPCR tests is generally low, it is significant in some cases, such as for some primers that bind to the N gene.

• We developed an algorithm that identify recurrent mutations as mutations that are in distant-related lineages. The mutations with higher degree of recurrence where C>U followed by G>U.

Additionally, turning attention to the objective 4, *to predict SARS-CoV-2 recurrent mutations using a machine learning model*, we conclude that:

• It is possible to predict some of the SARS-CoV-2 recurrent mutations. We built an Artificial Neural Network (ANN) that predicts positions that will have a recurrent mutation in the SARS-CoV-2 genome. In addition, we also developed an ANN for predicting the recurrence level (presence in a Number of Distantly-Related Lineages, NDRL) of a particular mutation. Both models use a window of 13 positions of the genome, where its center is the position being evaluated for a recurrent mutation.

• Both models perform better as the degree of recurrence increases. When predicting positions that will have a mutation in at least 1 NDRL (non-recurrent mutations) the ROC-AUC was 0.74. In contrast, predicting positions that will have a mutation in at least 15 NDRL results in an improved ROC-AUC of 0.81. Moreover, when predicting the recurrence level of a mutation, the model achieves a ROC-AUC of 0.84 for predicting mutations in at least 15 NDRL. These metrics were consistently observed across the testing set.

• We additionally evaluated the model performance with an updated test set, collected nine 9 months after the initial evaluation. Notably, some false positive mutations from the initial model, become true positives, demonstrating the robustness of the model. The ROC-AUC for predicting positions that will hold a recurrent mutation in at least 45 NDRL and 90 NDRL were 0.8 and 0.88, respectively. These results validate the choice to employ a custom loss function during the model construction, aimed at maintaining a sensitivity (true positive rate) close to 0.8 while optimizing for the highest possible specificity (true negative rate) under this constraint.

Moreover, delving into objective 5, *to identify the most important factors for predicting recurrent mutations in SARS-CoV-2*, we draw these conclusions:

- Using SHapley Additive exPlanation (SHAP) values, we observed that the nucleotide in the central position (P0) within each evaluated window of 13 positions is the most important variable in predicting the position of recurrent mutations in the SARS-CoV-2 genome. Other important variables include the nucleotides surrounding the mutating nucleotide (P-1 and P1), as well as the in vivo and in vitro RNA SHAPE-Seq reactivity data. With an NDRL threshold of 15, 44% and 27% of the true positives exhibit either adenine or uracil at position P-1. This finding is consistent with evidence that cytosines in the UC and AC motifs of the SARS-CoV-2 genome are preferentially deaminated by the APOBEC3A and APOBEC1 enzymes. When predicting the recurrence level of a mutation, lower recurrence levels prioritize the importance of in vitro RNA reactivity. Conversely, for higher degrees of recurrence, the importance shifts towards when a cytosine or adenyne mutates to a uracil or adenine. Once again this observation agrees with the prevalence of C>U mutations originated by host deaminases.

Publication List

1. **Saldivar-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. Int J Mol Sci. 2023 May 22;24(10):9072. doi: 10.3390/ijms24109072.

2. **Saldivar-Espinoza B**, Macip G, Garcia-Segura P, Mestres-Truyol J, Puigbò P, Cereto-Massagué A, Pujadas G, Garcia-Vallve S. Prediction of Recurrent Mutations in SARS-CoV-2 Using Artificial Neural Networks. Int J Mol Sci. 2022 Nov 24;23(23):14683. doi: 10.3390/ijms232314683.

3. **Saldivar-Espinoza B**, Macip G, Pujadas G, Garcia-Vallve S. Could nucleocapsid be a next-generation COVID-19 vaccine candidate? Int J Infect Dis. 2022 Dec;125:231-232. doi: 10.1016/j.ijid.2022.11.002.

4. Macip G, Garcia-Segura P, Mestres-Truyol J, **Saldivar-Espinoza B**, Pujadas G, Garcia-Vallvé S. A Review of the Current Landscape of SARS-CoV-2 Main Protease Inhibitors: Have We Hit the Bullseye Yet? Int J Mol Sci. 2021 Dec 27;23(1):259. doi: 10.3390/ijms23010259.

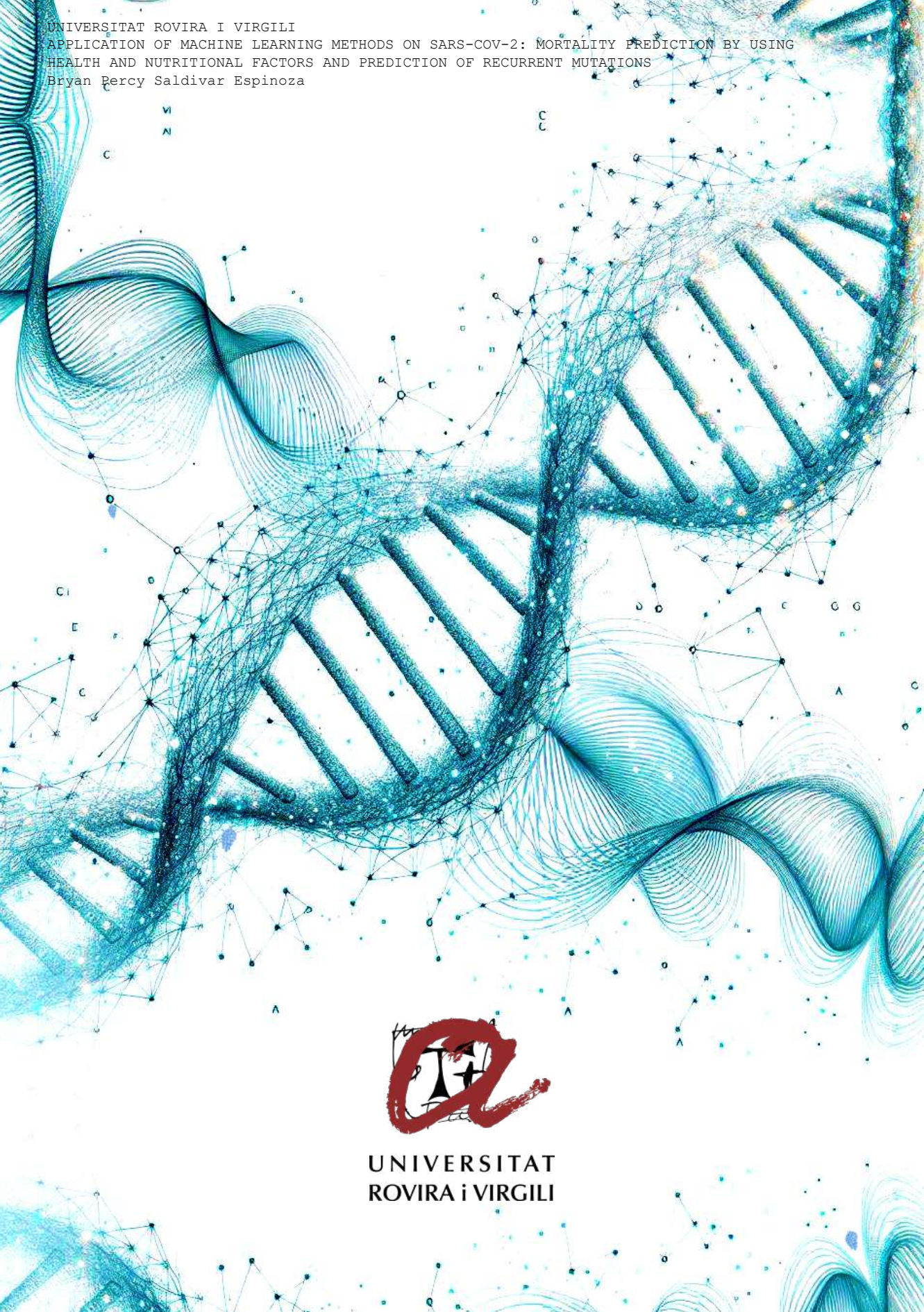
5. Macip G, Garcia-Segura P, Mestres-Truyol J, **Saldivar-Espinoza B**, Ojeda-Montes MJ, Gimeno A, Cereto-Massagué A, Garcia-Vallvé S, Pujadas G. Haste makes waste: A critical review of docking-based virtual screening in drug repurposing for SARS-CoV-2 main protease (M-pro) inhibition. Med Res Rev. 2022 Mar;42(2):744-769. doi: 10.1002/med.21862.

6. Gimeno A, Mestres-Truyol J, Ojeda-Montes MJ, Macip G, **Saldivar-Espinoza B**, Cereto-Massagué A, Pujadas G, Garcia-Vallvé S. Prediction of Novel Inhibitors of the Main Protease (M-pro) of SARS-CoV-2 through Consensus Docking and Drug Reposition. Int J Mol Sci. 2020 May 27;21(11):3793. doi: 10.3390/ijms21113793.

UNIVERSITAT ROVIRA I VIRGILI
APPLICATION OF MACHINE LEARNING METHODS ON SARS-COV-2: MORTALITY PREDICTION BY USING HEALTH AND NUTRITIONAL FACTORS AND PREDICTION OF RECURRENT MUTATIONS
Bryan Percy Saldivar Espinoza

UNIVERSITAT ROVIRA I VIRGILI
APPLICATION OF MACHINE LEARNING METHODS ON SARS-COV-2: MORTALITY PREDICTION BY USING HEALTH AND NUTRITIONAL FACTORS AND PREDICTION OF RECURRENT MUTATIONS
Bryan Percy Saldivar Espinoza

UNIVERSITAT ROVIRA I VIRGILI
APPLICATION OF MACHINE LEARNING METHODS ON SARS-COV-2: MORTALITY PREDICTION BY USING HEALTH AND NUTRITIONAL FACTORS AND PREDICTION OF RECURRENT MUTATIONS
Bryan Percy Saldivar Espinoza



UNIVERSITAT
ROVIRA i VIRGILI