Predicting Mutation-Driven Changes in the SARS-CoV-2 Spike Protein Using Structural Signatures and Neural Networks

Eduardo U. M. Moreira [1]*, Leandro Morais [1]*, Sheila C. Araujo[1,2]*, Rafael P. Lemos[1], Ana Luísa A. Bastos[1], Alessandra Lima[1], Diego Mariano[1], Raquel C. de Melo-Minardi [1]

[1] Laboratory of Bioinformatics and Systems (LBS)
Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil
[2] Laboratory of Molecular Modeling and Bioinformatics (LAMMB)
Universidade Federal São João Del Rey (UFSJ), Sete Lagoas, Minas Gerais, Brazil
*These authors have contributed equally

raquelcm@dcc.ufmg.br

Abstract. *COVID-19, caused by the SARS-CoV-2 virus, has led to a global pandemic since 2020, resulting in nearly 7 million deaths. The virus's rapid spread is due to more transmissible variants, many with spike glycoprotein mutations, which are key for cell invasion and a vaccine target. Understanding these mutations is crucial for preventing more dangerous variants. This study developed a computational method to predict the impact of mutations on the spike protein. Using data from 23,472 mutations, molecular modeling, graph-based structural signatures, and a machine-learning approach based on neural networks, the model analyzed 318 proteins, showing the methodology's effectiveness in assessing the potential of new variants.*

1. Introduction

In December 2019, SARS-CoV-2 emerged in Wuhan, China, and rapidly spread globally [Alsharif and Qurashi 2021], with the World Health Organization (WHO) declaring a pandemic in 2020. SARS-CoV-2 is a coronavirus that causes the respiratory disease now called Coronavirus disease 2019 (COVID-19). In May 2023, the Public Health Emergency of International Concern (PHEIC) ended. However, COVID-19 remains a threat, with over 770 million cases and 7 million deaths as of August 2023 [WHO 2023].

Coronaviruses belong to the *Coronaviridae* family, in the *Betacoronavirus* genera. These viruses possess the largest RNA genome, ranging from 27 to 32 kilobases (kb), with SARS-CoV-2's genome being about 29.9 kb long. The viral structure includes a capsid, envelope, and spike protein (S) [Wang et al. 2020], with mutations in these proteins being essential for understanding viral behavior and developing treatments [Weiss and Navas-Martin 2005, Yang and Rao 2021, Nieto-Torres et al. 2015]. The Spike (S) protein, which binds to the ACE2 receptor, facilitates the rapid spread of SARS-CoV-2, mainly due to the furin cleavage site. For example, mutations such as D614G and alterations in glycosylation increase transmissibility and immune evasion, making the S protein central to pathogenicity and transmission studies [Rabaaan et al. 2020, Cueno and Imai 2021].

Since 2019, SARS-CoV-2 has evolved into more than ten variants with altered transmissibility and severity. WHO has classified some of them as Variants of Concern (VOC) (Supplementary Table S1). As of March 2023, only Variants of Interest (VOIs) and Variants Under Monitoring (VUMs) remain. Mutations in the Receptor-Binding Domain (RBD) of the S protein have been crucial for immune escape, with key residues identified as responsible for antibody resistance. [Harvey et al. 2021, Weisblum et al. 2020]. Therefore, scientists have made a great effort to understand the structural relationships between these molecules.

Bioinformatics can be a helpful tool for analyzing viral genomes and the effects of mutations, facilitating studies of SARS-CoV-2 [Moreira et al. 2024]. With advances in sequencing and data management, these techniques have accelerated discoveries about viral evolution and protein function, helping to combat threats such as COVID-19 [Bayat 2002, Ibrahim et al. 2018, Paiva et al. 2022]. Structural alignment and molecular modeling are essential bioinformatic approaches to studying SARS-CoV-2 variants. They can reveal how mutations impact amino acids and protein function [Shukla et al. 2023]. These techniques identify critical regions and predict changes in the S protein, helping to combat emerging variants [Shukla et al. 2023, Ribeiro et al. 2023].

Another strategy that can be adopted to understand the structural role of these molecules is structural signatures. Studying structural signatures is essential for predicting the effects of mutations on function and stability [Hilario et al. 2004, Pires et al. 2011]. Thus, we hypothesized that this methodology could be applied to protein S, identifying patterns linked to transmissibility and pathogenicity [Zatorski et al. 2022]. These signatures provide a more detailed analysis than sequence-based methods, aiding in understanding viral evolution [Pires et al. 2011]. We wonder whether structural patterns in sets of mutations (mainly in spike protein) increase the epidemiological prevalence of SARS-COV-2. If such patterns exist, could they be detected using structural signatures? To answer these questions, we modeled a series of spike protein mutants and calculated their structural signatures.

This study examines the impact of mutations in the SARS-CoV-2 spike protein, aiming to predict how structural changes influence infection and immune evasion. Understanding these mutations is essential to developing effective vaccines and treatments in the face of more transmissible and resistant variants. Using structural bioinformatics algorithms such as the atomic Cutoff Scanning Matrix (aCSM, [Pires et al. 2013]), we created structural signatures of the S protein to model and predict the impacts of mutations. We also developed a neural network model that analyzes how these mutations provide biological advantages to SARS-CoV-2, contributing to improved control measures and applicable to studying other emerging pathogens. Figure 1 presents an overview of the methodology adopted in this study.

2. Methodology

2.1. Data Collection

Data for the Spike protein was gathered from UniProt using the P0DTC2 code, which provided details on the wild-type sequence, mutagenesis, and protein modifications. Mutation frequency data, especially for the RBD region, were sourced from GISAID. Addi-

tionally, unique mutations across the entire S protein were selected from the Bioinformatics Institute of Singapore database (ASTAR Singapore)[1].

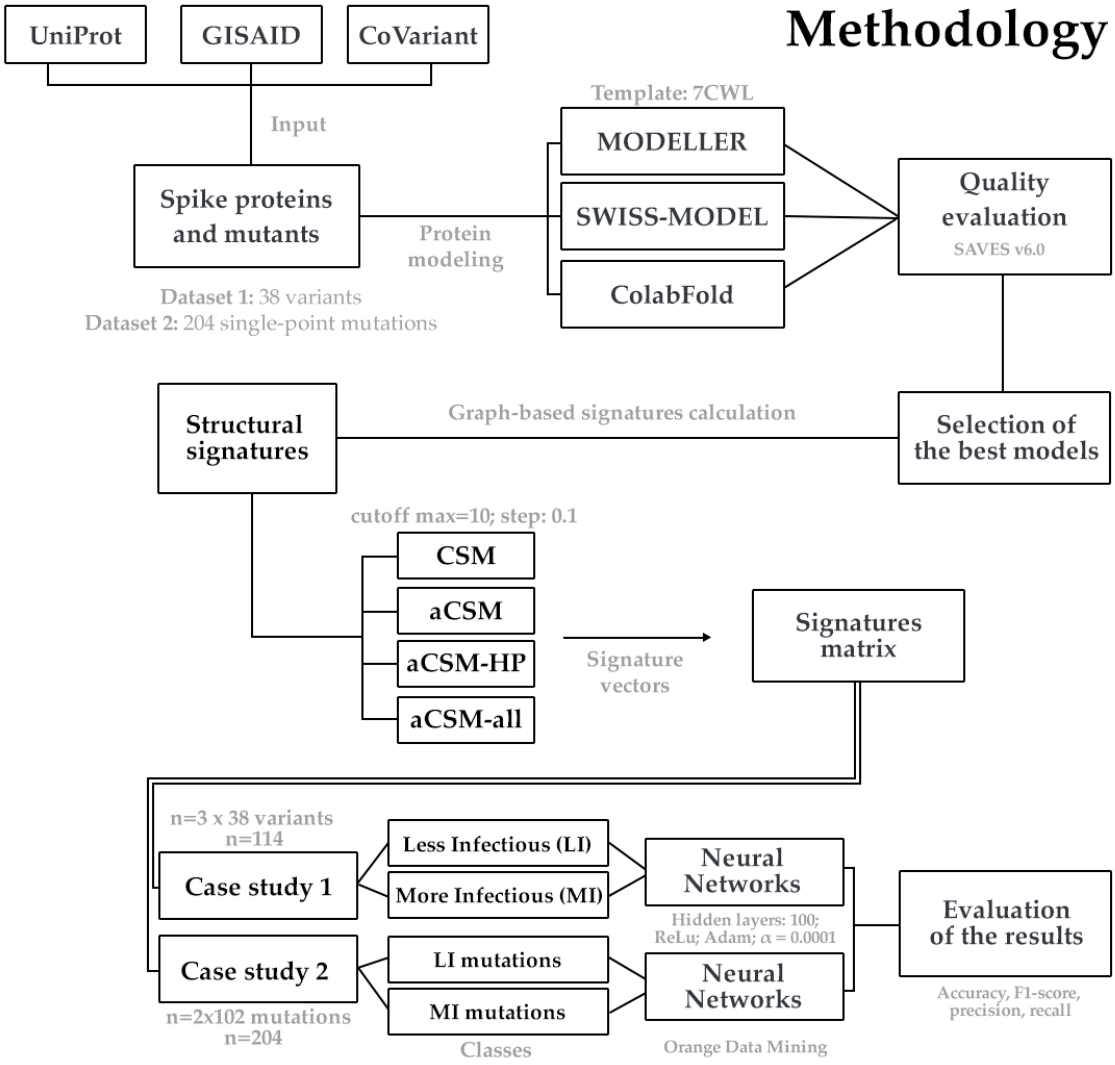

Figure 1. Overview of the methodology discussed in this work. The workflow illustrates the process where Spike protein sequences and mutation data were collected, followed by modeling using MODELLER, SWISS-MODEL, and ColabFold. Quality evaluation was performed to select the best models. Structural signatures were then calculated using four algorithms (CSM, aCSM, aCSM-HP, and aCSM-ALL). These signature vectors were processed through neural networks, classifying mutations of interest and evaluating their impact.

An in-house Python script was created to process the Wuhan-Hu-1 strain's sequence and the 37 mutations of the prevalent variants from the CoVariants website [2], generating a ".txt" file with the mutated sequences. The sequences were manually checked in the Clustal Omega to confirm the mutations.

[1]Available at https://mendel.bii.a-star.edu.sg/METHODS/corona/current/MUTATIONS/hCoV-19_Human_2019_WuhanWIV04/hcov19_Spike_mutations_table.html.

[2]Available at https://covariants.org

2.2. Molecular 3D Modeling

For comparison purposes, the models were generated using three tools: MODELLER and SWISS-MODEL, based on comparative modeling; and ColabFold, based on deep learning.

The MODELLER v.10.4 procedure started with a BLAST search [Altschul et al. 1990] to identify proteins resolved in the PDB, selecting models with $\geq 25\%$ identity [Webb and Sali 2016]. The structures with PDB ID 7CWL, 7KRQ, 7N1Q, 7N1U, 7SBK, 7SBP, 7SBS, 7TNW, and 8D55 were downloaded and modified in PyMOL v.2.5.4, isolating monomers from homotrimers. Five models were created for each protein and the best model was selected based on the lowest DOPE value, which is calculated based on a sample of native protein structures, generating a statistical potential that varies depending on the atomic distance [Shen and Sali 2006].

Using SWISS-MODEL, the mutated sequences were directly inputted, and the tool suggested models for construction. The eight best models were selected, considering the best sequence coverage. The models generated were evaluated based on the GMQE value (quality assessment that integrates characteristics of the alignment between the model, the objective and the structure of the model itself), with higher scores being considered better[3].

ColabFold v.1.5.2-patch [Mirdita et al. 2022][4] was used with default parameters to generate five models per sequence. The best model was selected on the basis of the lDDT (local Distance Difference Test), which assesses how well local atomic interactions in the structure of the reference protein are reproduced in the prediction.

2.3. Model Quality Check

The modeled structures were evaluated in PyMOL by visual analysis and alignment with standard models to prevent inconsistencies. The Ramachandran plot [Ramachandran et al. 1963], generated by PROCHECK [Laskowski et al. 1993] on the SAVES server[5], was essential to validate the 3D models (Supplementary Figure S2) [Bowie et al. 1991, Lüthy et al. 1992]. The results were compared to PDB models and used as a reference for new models.

Molecular modeling was chosen due to the limited availability of database variants, thus avoiding the high computational cost of molecular dynamics. The PDB ID 7CWL was utilized as the template for modeling the variants, including single mutations.

We then divided this work into two parts: (i) evaluating the 38 variants and (ii) evaluating each single-point mutation individually.

The structural models were also evaluated using the VERIFY 3D tool, which compares the atomic 3D structure with the linear amino acid sequence. The tool classifies the structure based on its conformation and environment (such as α-helix, β-sheet, etc.) and compares these data with reference structures, allowing the identification of the most appropriate model and the evaluation of its quality [Bowie et al. 1991, Ramachandran et al. 1963]. Further details are provided in the Supplementary Material.

[3] Available at https://swissmodel.expasy.org/docs/help

[4] Available at https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb

[5] Available at https://saves.mbi.ucla.edu/

2.4. Structural Signatures

We used the SIGNA Python library (https://github.com/LBS-UFMG/signa) to generate the structural signatures. Four algorithms were used: CSM, aCSM, aCSM-HP, and aCSM-ALL [Pires et al. 2011, Pires et al. 2013]. A total of 114 structural signatures were obtained, one for each protein structure modeled using the previously mentioned tools. The output was a ".csv" file containing the variant's name, followed by the numeric vector representing the structural signature and the cutoff step legend for each column. The parameters used for signature extraction included a cutoff limit of 10 Å, and a cutoff step of 0.1 Å [Mariano et al. 2019]. The resulting vector sizes varied depending on the signature type.

2.5. Prediction Analysis

We then built a model to classify variants using a neural network algorithm. The following parameters were used: 100 neurons in hidden layers; ReLU activation; Adam solver; regularization $\alpha = 0.0001$; the maximal number of iterations of 200; and the option "replicate training". The input file for classification consisted of structural signatures, which represent the predictive attributes. Specifically, these signatures capture the patterns in the numerical vectors of the mutations' structural features, and the neural network algorithm was used to identify and rank the importance of each segment. The target attribute for the model was the prior classification of the variants into two categories: Strong variants ("Yes" in the dataset) and Weak variants ("No" in the dataset), corresponding to more prevalent Variants of Concern (VOCs) as defined by the World Health Organization - MI mutations - and variants with less clinical impact - LI mutations (Table 1). We used the Orange Data Mining software v.3.34.1 [Demšar et al. 2013] for the analysis. The models' accuracy, F1-Score, precision, and recall metrics (Supplementary Figure S1 and Table S1) were compared to identify the best-performing one.

Table 1. Classification of Variants. The table lists the More Infectious (MI) and Less Infectious (LI) variants selected through the CoVariants website. The full mutation list is available at https://github.com/LBS-UFMG/s-variant-signatures.

Class	Variant
More Infectious (MI)	20H_(Beta_V2), 20I_(Alpha_V1), 20J_(Gamma_V3), 21A_(Delta), 21I_(Delta), 21J_(Delta), 21K_(Omicron), 21L_(Omicron), 22A_(Omicron), 22B_(Omicron), 22C_(Omicron), 22D_(Omicron), 22E_(Omicron), 22F_(Omicron), 23A_(Omicron), 23B_(Omicron), 23C_(Omicron), 23D_(Omicron), 23E_(Omicron), 23F_(Omicron)
Less Infectious (LI)	20A.EU2, 20A_98F, 20A_126A, 20A_439K, 20B_626S, 20B_732A, 20B_1122L, 20C_80Y, 20E_(EU1), 21B_(Kappa), 21C_(Epsilon), 21D_(Eta), 21F_(Iota), 21G_(Lambda), 21H_(Mu), 677H.Robin1, 677P.Pelican, Wild_Variant

.

3. Results and Discussions

3.1. Spike Protein Data

The S protein obtained from UniProt consists of 1,273 amino acids and has a molecular weight of 141,178 Da. The RBD region spans amino acids 319 to 541 [Consortium 2023].

Through GISAID, a global data-sharing initiative, mutation frequency data for this protein was obtained. Since January 2020, over 5 million SARS-CoV-2 genetic sequences from 194 countries have been publicly available via GISAID's EpiCoV database. This data is crucial for developing diagnostic and preventive measures and monitoring emerging variants and mutations [Chen et al. 2021].

The most prevalent SARS-CoV-2 variants were identified via CoVariants, an open-source project using Nextstrain clade nomenclature. It provides an overview of variants, their defining mutations, impacts, and geographic spread [Goujon et al. 2010, Sievers et al. 2011]. The selected dataset includes Binding Free Energy (BFE) variations for the S protein-ACE2 complex, where negative values mean weak binding and positive values mean strong binding, making the variant more infectious [Chen et al. 2021]. Thus, MI and LI were selected based on this.

Figure 2 shows the structure of the SARS-CoV-2 S protein (PDB ID: 7CWL), highlighting examples of MI and LI mutations. MI mutations include position N501, shown in green at the top, while the mutation N501Y is shown in magenta. This mutation significantly affects the protein's interaction with the ACE2 receptor, thereby influencing viral transmissibility. LI mutations include position F43, shown in green at the bottom, and the mutation F43M in magenta. This mutation has a minor effect on the overall stability and function of the Spike protein compared to MI mutations. The entire protein structure illustrates how these changes alter specific residues in the Spike protein.

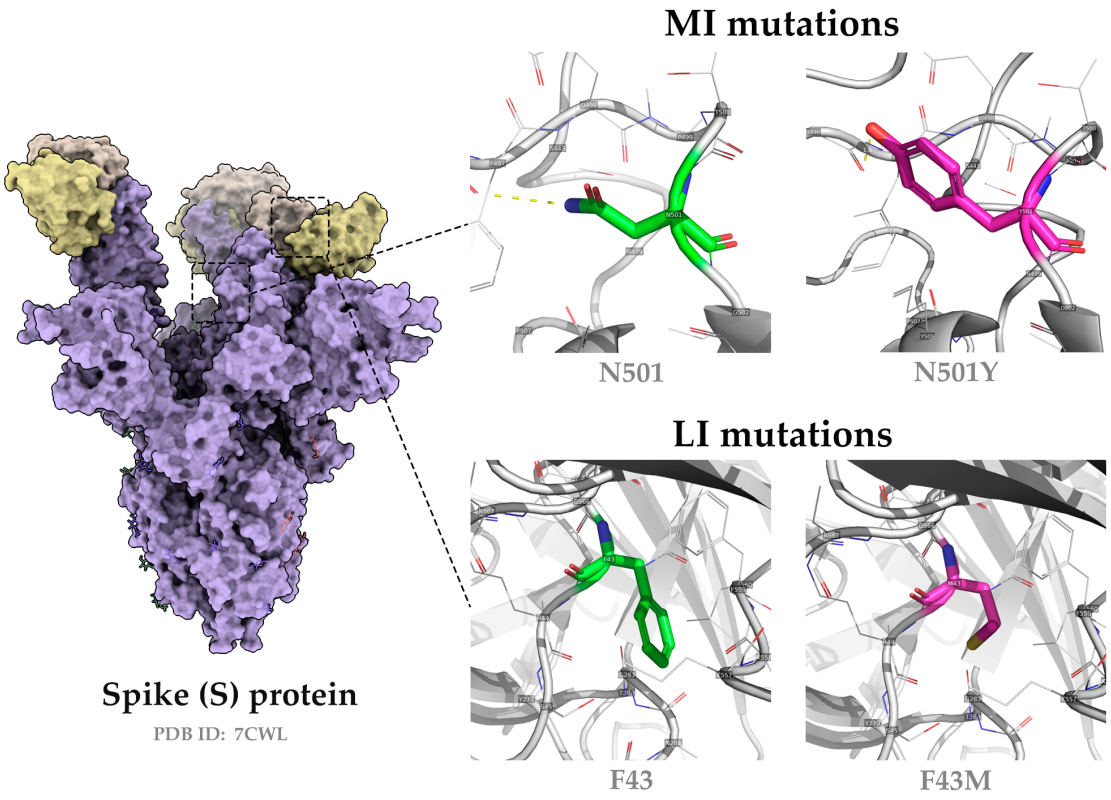

Figure 2. Spike structure and examples of more infectious (MI) and less infectious (LI) mutations. The MI N501Y mutation is shown at the top, and the LI mutation F43M is shown at the bottom.

3.2. Part 1: Evaluation of the 38 Variants

The 38 variant models were generated through molecular modeling (Table 1). Although the tools used are not designed to predict the impact of mutations directly, they were employed as the most computationally viable option. Alternatives like molecular dynamics simulations would be significantly more computationally expensive for large-scale analyses.

The generated structural models were evaluated using the VERIFY 3D tool. It verified the congruence between the three-dimensional (3D) atomic model and its corresponding one-dimensional (1D) amino acid sequence. The structure was categorized based on its arrangement and environment, such as $\alpha - helix$, $\beta - sheet$, loop, polar, and nonpolar, and the results were compared with reference structures to select the best model and evaluate its quality [Bowie et al. 1991][Lüthy et al. 1992]. We generated Ramachandran plots for the PDB wild structure and our models. Preliminary results indicate that our models are as good as native structures. For example, the Ramachandran plot analysis of 7CLW indicated that 84.7% of residues in this model were in allowed regions, while VERIFY 3D confirmed 62.48% of residues had a 3D-1D score above 0.1 (Supplementary Figure S2).

After modeling the mutants and calculating the signatures, we build a neural network model using Orange Data Mining software (Table 2). This was necessary for comparing the performance of the models and evaluating metrics such as Accuracy, F1-Score, Precision, and Recall. As mentioned, variants were classified into MI (VOCs, as defined by the WHO) and LI. The values of the structural signatures were used to determine the class of a modified SARS-CoV-2 variant based on this binary classification. Table 2 presents the best values for each metric across the different modeling tools.

SWISS-MODEL coupled with aCSM-HP performed consistently well in almost all comparisons, with an accuracy of 92%. This is an indication that there are structural patterns that explain more infectious and less infectious variants, which corroborates our initial hypothesis. The aCSM-HP is a signature that considers atomic positions and classifies atoms into two types: polar and hydrophobic. This signature is considered simpler than the more complete version, aCSM-ALL, which considers eight atomic types: acceptor, donor, hydrophobic, positive, negative, sulfide, aromatic, and neutral. This interesting result indicates that only the polarity of atoms is sufficient to classify the structures.

3.3. Part 2: Obtaining Models with Single Mutations

After building a model to classify variants based on their signature, we want to know whether single-point mutations are sufficient to classify our structures. Thus, we analyzed 23,472 unique mutations across the S protein from the literature. From this, 102 mutations found in the most prevalent SARS-CoV-2 variants were selected (More Infectious - MI mutations). To balance the sampling, another 102 mutations were randomly chosen from the remaining 23,370 mutations (Less Infectious - LI mutations) (Table 3).

The 204 selected single-point mutations were modeled only using the MODELLER tool due to its efficient automation capability for handling multiple structures. Five models were generated for each of the 204 mutations, and the model with the lowest DOPE score was selected. The structural signatures of the best models were obtained and subsequently analyzed alongside the variant signatures using a neural network model

Table 2. Model metrics were obtained through analysis of each tool's signatures. Grey lines indicate the signature method used (CSM, aCSM, aCSM-HP and aCSM-ALL, respectively), and Class columns indicate the three modeling tools used (SWISS-MODEL, MODELLER, and ColabFold, respectively).

CSM				
Class	**Accuracy**	**F1-Score**	**Precision**	**Revocation**
SWISS-MODEL	0.868	0.869	0.870	0.868
MODELLER	0.895	0.895	0.895	0.895
ColabFold	0.763	0.763	0.772	0.763
aCSM				
Class	**Accuracy**	**F1-Score**	**Precision**	**Revocation**
SWISS-MODEL	0.868	0.869	0.870	0.868
MODELLER	0.842	0.839	0.857	0.842
ColabFold	0.474	0.471	0.470	0.474
aCSM-HP				
Class	**Accuracy**	**F1-Score**	**Precision**	**Revocation**
SWISS-MODEL	0.921	0.921	0.922	0.921
MODELLER	0.868	0.869	0.870	0.868
ColabFold	0.658	0.654	0.659	0.658
aCSM-ALL				
Class	**Accuracy**	**F1-Score**	**Precision**	**Revocation**
SWISS-MODEL	0.895	0.895	0.895	0.895
MODELLER	0.868	0.869	0.870	0.868
ColabFold	0.895	0.895	0.900	0.895

built with the Orange Data Mining software. The objective here is to verify the impact of each mutation on the created model.

As shown in Table 4, due to the more detailed data patterns — involving individual mutations in a structure with 1,273 amino acid residues — the aCSM-ALL structural signature showed superior results compared to the other signatures. This can be attributed to the ability of this signature to provide more detailed structural information, differentiating between various classes of atoms.

Table 3. More Infectious (MI) and Less Infectious mutations (LI). Single mutations selected and characterized as Most Infectious (MI) in the first column and Least Infectious (LI) in the second column according to prevalence.

MI mutations	LI mutations
L18F, T19R, T19I, T20N, L24del, P25del, P26del, P26S, A27S, Q52H, A67V, H69del, V70del, D80A, V83A, T95I, D138Y, G142D, G142del, V143del, Y144del, Y145D, H146Q, K147E, W152R, E156del, F157L, F157del, R158G, E180V, Q183E, R190S, I210V, N211del, L212I, V213G, V213E, D215G, A222V, L241del, L242del, A243del, G252V, D253G, G257S, G339H, G339D, R346T, L368I, S371F, S371L, S373P, S375F, T376A, D405N, R408S, K417N, K417T, N440K, K444T, V445P, G446S, L452R, L452Q, F456L, N460K, S477N, T478K, T478R, E484A, E484K, F486V, F486S, F486P, F490S, Q493R, G496S, Q498R, N501Y, Y505H, P521S, T547K, A570D, D614G, H655Y, N679K, P681H, P681R, A701V, S704L, T761N, N764K, D796Y, N856K, D950N, Q954H, N969K, L981F, S982A, T1027I, D1118H, V1176F	F43M, T51S, G75N, V83C, P85Q, W104C, T109D, D111A, L118K, R158S, Y170R, M177C, V193Q, F194H, H207R, Q218H, R246Y, T259Y, Q271T, T274N, T274A, K278S, K278Q, E281K, D287Y, T315del, T315N, S316V, S316del, C391G, D398H, P412P, N349L, R457I, R466del, D467S, P479D, C480S, N481T, V483D, T500Y, G504P, L533F, V534G, N540T, F543F, N544I, N544Q, V551S, V551A, N556L, Q607V, L650T, V705G, A706Q, D737C, M740Y, G769D, D775V, D775N, D775K, Q784G, Q836V, A846V, D848M, P863del, A871S, L878V, I882Y, A893Y, M900N, L922L, A930G, A944I, V959L, L959M, F970L, S975I, I980Y, L984Q, Y1007L, I1018I, M1029H, E1031Q, V1033K, S1037L, A1056P, Y1067E, Y1067I, V1104Q, Q1113Y, I1179E, I1183S, K1191T, K1205I, Q1208M, G1219L, L1224I, I1227Y, M1229W, C1235P, C1243M

Table 4. Metrics obtained for prediction with different types of structural signatures. The white cells indicate the signature method (CSM, aCSM, aCSM-HP and aCSM-ALL, respectively) used in the models resulting from the MODELLER tool, where the classes addressed in the prediction method were More Infectious (MI) and Less Infectious (LI) and the gray colored line indicates the metrics analyzed.

Class		Accuracy	F1-Score	Precision	Recall
CSM	MI	0.579	0.680	0.567	0.850
	LI	0.579	0.385	0.625	0.278
MODELLER		**0.579**	**0.540**	**0.594**	**0.579**
aCSM	MI	0.368	0.520	0.433	0.650
	LI	0.368	0.077	0.125	0.056
MODELLER		**0.368**	**0.310**	**0.287**	**0.368**
aCSM-HP	MI	0.632	0.741	0.588	1.000
	LI	0.632	0.364	1.000	0.222
MODELLER		**0.632**	**0.562**	**0.783**	**0.632**
aCSM-ALL	MI	0.816	0.829	0.810	0.850
	LI	0.816	0.800	0.824	0.778
MODELLER		**0.816**	**0.815**	**0.816**	**0.816**

Based on these results, it is possible to randomly combine mutations in a potential SARS-CoV-2 variant and predict whether, with certain combined mutations, this potential variant tends to be more or less prevalent, according to the data from variants that, in the

real-world context, proved to be of greater importance to global public health.

4. Conclusion

This study explored the ability of the SARS-CoV-2 S protein to undergo mutations that may enhance its pathogenicity. Understanding these factors is crucial, given the severe impact of COVID-19. The main goal was to develop a met0.562hod to predict the impact of mutations in the S protein, identifying those that increase infectivity, transmissibility, and pathogenic, similar to the most prevalent variants. Molecular models were generated, filtered for quality, and analyzed using aCSM structural signatures processed by neural networks. The results sMIhowed high sensitivity and specificity in classifying variants, with good accuracy in distinguishing more pathogenic from less prevalent ones. Our results indicate that aCSM-HP and models built with SWISS-MODEL were better for this case study. When trained with single-point mutation signatures, the model accurately predicted the impact of new mutations, offering a method to forecast the formation of more pathogenic variants. In this case, the best signature was aCSM-ALL. This indicates that this signature with the parameters max cutoff 10 and cutoff step of 0.1 are good strategies to represent SARS-COV-2 spike protein. This can be used for example to assess the impact of new variants.

5. Acknowledgements

The authors would like to thank the research funding agencies CAPES, FAPEMIG, and CNPq. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

6. Supplementary Material

All the Supplementary material is available at https://github.com/LBS-UFMG/s-variant-signatures/

References

Alsharif, W. and Qurashi, A. (2021). Effectiveness of covid-19 diagnosis and management tools: A review. *Radiography*, 27(2):682–687.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410.

Bayat, A. (2002). Bioinformatics. *BMJ: British Medical Journal*, 324(7344):1018–1022.

Bowie, J. U., Lüthy, and Eisenberg, D. (1991). a method to identify protein sequences that fold into a known three-dimensional structure. *science*, 253(5016):164–170.

Chen, J., Wang, R., Wang, M., and Wei, G. (2021). prediction and mitigation of mutation threats to covid-19 vaccines and antibody therapies. *chemical science*, 12(20):6929–6948.

Consortium, T. U. (2023). uniprot: the universal protein knowledgebase in 2023. *nucleic acids research*, 51(d1):d523–d531.

Cueno, M. E. and Imai, K. (2021). Structural comparison of the sars cov 2 spike protein relative to other human-infecting coronaviruses. *Frontiers in Medicine*, 7:1–8.

Demšar, J., Zupan, B., Leban, G., Curk, T., Starič, A., Petkovšek, E., Kavšek, B., and Polajnar, M. (2013). orange: data mining toolbox in python. *journal of machine learning research*, 14(71):2349–2353.

Goujon, M., McWilliam, H., Li, W., Valentin, F., Squizzato, S., Paern, J., and Lopez, R. (2010). A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res.*, 38(Web Server issue):W695–9.

Harvey, W. T. et al. (2021). Sars-cov-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology*, 19(7):409–424.

Hilario, M. et al. (2004). Classifying protein fingerprints. In boulicaut, j.-f. et al., editors, *Knowledge Discovery in Databases: PKDD 2004*, volume 3202 of *Lecture Notes in Computer Science*, pages 209–220. Springer.

Ibrahim, B. et al. (2018). A new era of virus bioinformatics. *Virus Research*, 251:86–90.

Laskowski, R. A., Macarthur, M. W., Moss, D. S., and Thornton, J. M. (1993). procheck: a program to check the stereochemical quality of protein structures. *j. appl. crystallogr.*, 26(2):283–291.

Lüthy, R., Bowie, J. U., and Eisenberg, D. (1992). assessment of protein models with three-dimensional profiles. *nature*, 356(6364):83–85.

Mariano, D. C. B., Santos, L. H., Machado, K. D. S., Werhli, A. V., de Lima, L. H. F., and de Melo-Minardi, R. C. (2019). A computational method to propose mutations in enzymes based on structural signature variation (SSV). *Int. J. Mol. Sci.*, 20(2):333.

Mirdita, M. et al. (2022). colabfold: making protein folding accessible to all. *nature methods*, 19(6):679–682.

Moreira, E. U. M., Mariano, D. C. B., and de Melo-Minardi, R. C. (2024). computational analysis of mutations in sars-cov-2 variants spike protein and protein interactions. In *features, transmission, detection, and case studies in covid-19*, pages 123–139. elsevier.

Nieto-Torres, J. L. et al. (2015). Severe acute respiratory syndrome coronavirus e protein transports calcium ions and activates the nlrp3 inflammasome. *Virology*, 485:330–339.

Paiva, V. d. A. et al. (2022). Protein structural bioinformatics: An overview. *Computers in Biology and Medicine*, 147:105695.

Pires, D. E. V. et al. (2011). Cutoff scanning matrix (csm): structural classification and function prediction by protein inter-residue distance patterns. *BMC Genomics*, 12(Suppl 4):S12.

Pires, D. E. V. et al. (2013). acsm: noise-free graph-based signatures to large-scale receptor-based ligand prediction. *bioinformatics*, 29(7):855–861.

Rabaaan, A. A. et al. (2020). Sars-cov-2, sars-cov, and mers-cov: A comparative overview. *Le Infezioni in Medicina*, 28(2):174–184.

Ramachandran, G. N., Ramakrishnan, C., and Sasisekharan, V. (1963). stereochemistry of polypeptide chain configurations. *j. mol. biol.*, 7(1):95–99.

Ribeiro, R. et al. (2023). Molecular modeling study of natural products as potential bioactive compounds against sars-cov-2. *Journal of Molecular Modeling*, 29(6):183.

Shen, M.-y. and Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein Science*, 15(11):2507–2524.

Shukla, N. et al. (2023). Covid variants, villain and victory: A bioinformatics perspective. *Microorganisms*, 11(8):2039.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., and Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Mol. Syst. Biol.*, 7(1):539.

Wang, M.-Y. et al. (2020). Sars-cov-2: Structure, biology, and structure-based therapeutics development. *Frontiers in Cellular and Infection Microbiology*, 10:587269.

Webb, B. and Sali, A. (2016). Comparative protein structure modeling using modeller. *Current Protocols in Bioinformatics*, 54:5.6.1–5.6.37.

Weisblum, Y. et al. (2020). Escape from neutralizing antibodies by sars-cov-2 spike protein variants. *eLife*, 9:e61312.

Weiss, S. R. and Navas-Martin, S. (2005). Coronavirus pathogenesis and the emerging pathogen severe acute respiratory syndrome coronavirus. *Microbiology and Molecular Biology Reviews*, 69(4):635–664.

WHO (2023). Who coronavirus (covid-19) dashboard. `https://covid19.who.int`. Accessed: 2023-09-05.

Yang, H. and Rao, Z. (2021). Structural biology of sars-cov-2 and implications for therapeutic development. *Nature Reviews Microbiology*, 19(11):685–700.

Zatorski, N. et al. (2022). Structural signatures: a web server for exploring a database of and generating protein structural features from human cell lines and tissues. *Database: The Journal of Biological Databases and Curation*, 2022:baac053.