Machine Learning in the Identification of Key Residues of Variants and Polymorphisms in the Interaction of ACE2 Proteins with Spike of SARS-CoV2

Ana Carolina Damasceno Sanches, Ana Luisa Rodrigues de Avila, Levy Bueno Alves and Silvana Giuliatti
Departament of Genetics. Ribeirao Preto Medical School
University of Sao Paulo, USP
Ribeirao Preto, Brazil
e-mail: silvana@fmrp.usp.br

Abstract—**The binding affinity between the Spike protein and the angiotensin-converting enzyme 2 receptor (ACE2) is one of the main determining factors in the replication rate of Severe Acute Respiratory Syndrome of Coronavirus-2 that directly affects the clinical condition of the patient. The presence of multiple variants indicates a high mutation rate of the virus. Furthermore, genetic variations within the coding regions of ACE2 can impact the susceptibility, severity, and progression of the disease. However, the effect of these mutations on the stability and affinity of the Spike-ACE2 interaction is not well understood. To gain insight into this interaction, molecular dynamics simulations are used. Although these simulations produce a large amount of data, they do not make easy to identify residues that play a significant role in the interaction between the proteins. To overcome this issue, we combined molecular dynamics simulations and supervised machine learning techniques to identify the residues that have the most impact on the interaction and dynamics of the complexes. The molecular dynamics simulations showed slight variations in complex trajectories, but highlighted key residues and loop region residues. Despite stable behavior among variants with only minor differences, the machine learning methods identified critical residues in ACE2 and Spike proteins that can affect virus-host interaction.**

Keywords—COVID-19; Bioinformartics; Molecular Docking; Polymorphism; Variants.

I. Introduction

On March 11, 2020, the World Health Organization characterized COVID-19 [1] as a pandemic, an infectious disease caused by the Severe Acute Respiratory Syndrome of Coronavirus-2 (SARS-CoV-2). To date, November 2022, more than 630 million cases have been confirmed, including more than 6.6 million deaths globally. In Brazil alone, there are more than 35 million cases with almost 690,000 deaths [2]. COVID-19 is a respiratory disease, transmitted by the epithelial cells of the lung through aerosols, which can lead from mild viral pneumonia to Acute Respiratory Distress Syndrome, and in even more serious cases leading to multiple organ failure [3]. It mainly affects individuals with comorbidities and/or some type of immunosuppression. Some people develop the severe form of COVID-19, while others are asymptomatic [3][4]. The entry of the virus into the cell is one of the most important processes in viral infection, being the target in the development of vaccines and drugs. The invasion of SARS-CoV-2 into host cells depends on the interaction of the Spike structural protein with the human protein, present in the cell membrane, angiotensin-II converting enzyme [5] and variants of this protein have been associated with susceptibility to SARS-CoV-2 [3][4].

SARS-CoV-2 has a high probability of mutating and adapting better to the environment [4]. The current Variant of Concern (VOC) is Omicron (B.1.1.529 - several countries), prior to this, also classified as VOC: Alpha (B.1.1.7 - United States), Beta (B.1.351 - Africa do Sul), Gamma (P.1 - Brazil) and Delta (B.1.617.2 - India) [1]. P2 variant (or Zeta variant) (B.1.1.28.2) was detected in the city of Rio de Janeiro in October 2020. The mutations suffered by SARS-CoV-2 observed in its variants, as well as the polymorphisms observed in the ACE2 protein, raise questions such as whether genetic variability of the virus and the host could explain the different degrees of severity in cases of infection. How these mutations contribute to improving the stability and affinity between Spike-ACE2 complexes is not a process fully understood.

Molecular Dynamics simulations have been used to assess the stability and affinity between complexed structures. The trajectories resulting from these simulations generate large amounts of data from thousands of atoms at each time interval. The stability of the complex is analyzed by calculating the root-mean-square deviation and also by the number and type of contact between the structures. However, the high-dimensional nature and noisy output of the simulations make it extremely difficult to extract meaningful features from the trajectories, thereby hindering a deeper understanding of molecular processes

Machine learning techniques are employed to analyze vast data sets. These methods assist in identifying key differences between the trajectories obtained from molecular dynamics simulations. Fleetwood et al. [6] demonstrated the usefulness and potential of machine learning techniques in comprehending biomolecular processes by applying both supervised and unsupervised techniques to three different biological systems. Inspired by this work, we utilized molecular dynamics simulations to evaluate the stability of complexes and applied supervised machine learning techniques using the resulting trajectories as input data to investigate the effect of genetic variability in SARS-CoV-2 and ACE2 polymorphisms on the interaction region between these proteins.

II. Material and Methods

In this section, we will outline the methods employed to perform molecular dynamics simulations and implement machine learning architectures.

A. Molecular Dynamics

The tertiary structure of the complex Spike and ACE2 (PDB ID: 6LZG) was obtained from the Protein Data Bank [7]. Modeller software v9.23 [8] was used to fill the missing atoms.

GROMACS package version 2019.3 [9] was used in the MD simulations of complexes. The force field used was CHARMM36 [10]. The molecules were solvated with TIP3P water molecules and neutralized by adding the appropriate number of Na+Cl ions considering the ionic concentration of 0.15 M. The energy minimisation was performed using the steepest descent method with a maximum force of 1000 Kj/mol.nm. After minimization, the systems were equilibrated in two stages: a canonical NVT set followed by an isothermal-isobaric NPT set. The NVT equilibrium was performed with a constant temperature of 300 K for 500 ps. The NPT equilibrium was performed with a constant pressure of 1 bar and a constant temperature of 300 K for 500 ps. The production step was conducted at 300 K for 100 ns and the trajectories were saved every 10 ps. Four complexes ACE2-Spike complexes were analised: ACE2-Spike (wild) and 3 ACE-Spike (Omicron, Delta and P2 – Zeta variant).

B. Machine Learning

Based on Fleetwood et. al [6], we employed molecular dynamics trajectories as input for supervised ML techniques. To reduce the influence of a single model and enhance the stability of our results, we utilized two differing supervised machine learning classification strategies: Multilayer Perceptron (MLP) and Random Forest (RF). These methods were used to identify residues that most significantly contribute to the difference in the dynamic behavior between the complexes (Fig. 1). A multiplayer perceptron is a type of artificial neural network has multiple layers between input and output layers. Meanwhile, Random Forest is an ensemble learning technique that is used for classification by building many decision trees and finding the mode of the classes of each tree. We chose to use both RF and MLP because they are powerful and commonly-used supervised machine learning algorithms. RF excels at performing both regression and classification tasks and is well-known for its robust performance and handling of noisy and missing data. MLP, a feedforward neural network, can handle regression and classification problems, and is frequentetly used for complex, non-linear relationships

The input features for these algorithms include the contact distances between ACE2 residues and Spike. These distances were calculated as the minimum distance between the heavy atoms of residues in the interaction region. Only distances less than 15 Å were considered in forming our dataset. The values were then inverted, normalized and used as inputs.

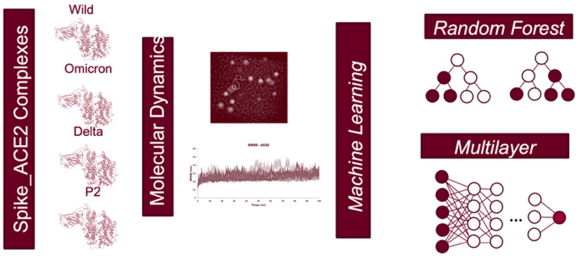

Figure 1. Flowchart of the Machine Learning methods used on this study.

The MLP was implemented using the open-source machine learning library Scikit-learn in Python [11]. We also used the data analysis and manipulation library Pandas [12], and the numerical computing library NumPy [13]. Scikit-learn is a widely-used, well-documented, and efficient machine learning library that provides quick prototyping and testing,

We employed 8 hidden layers with 100, 75, 50, 40, 30, 20, 10, and 5 neurons respectively, with ReLU activation. ReLU is a popular activation function in deep learning that is known for its effectiveness. The labels were one-hot encoded to represent categorical data numerically. The training process used the Adam optimizer [14] to adjust the node weights. This optimizer is frequentely used due to its demonstrated efficacy.

We created the first profile by building a correlation matrix for training and testing. Four additional profiles were generated through bootstrapping and features with strong correlation were discarded using a 0.9 threshold. As a result, 5 profiles were obtained with 1828, 1907, 1925, 1909 and 1934 features respectively, each with 40 thousand frames. The MLP was trained with each of these profiles, resulting in 10 total MLPs. We used a train-test split to evaluate the performance of the ML algorithms, with 80% of the data in the training set and 20% in the test set.

Important features for classification were determined using Layer-Wise Relevance Propagation (LRP) [15] with the LRP-0 rule. LRP assigns relevance scores to input features, making it possible to visualize which inputs have the most impact on a specific prediction made by the model. This enhances transparency and confidence in the decision-making of neural networks

Our Random Forest model utilized the Gini impurity coefficient, which ranges from 0 to 1, with 0 indicating a pure split and 1 representing maximum impurity. The aim was to choose splits that would lower Gini impurity, resulting in more homogeneous class distribution in the tree's leaves. RF The RF classifier uses an internal bootstrapping process to produce consistent profiles. The model consisted of 100 decision trees, with 3201 features and 40 thousand frames. The one-versus-the-rest method was employed to claculate feature importance, a strategy in multi-class classification that plits the problem into several binary classification problems. RF was implemented using the Scikit-learn library.

III. Results and Discussion

The outcomes achieved at each step of our work will be detailed in the subsequent sub-sections.

A. Molecular Dynamics

We sought differences in the interactions between SARS-CoV-2 variants and the ACE2 protein through 100 ns molecular dynamics simulations for each complex. The simulation data was used to compute Root Mean Square Deviation (RMSD) and Root Mean Square Fluctuation (RMSF). Fig. 2A and fig. 2B show RMSD values for ACE2 and Spike proteins, respectively.

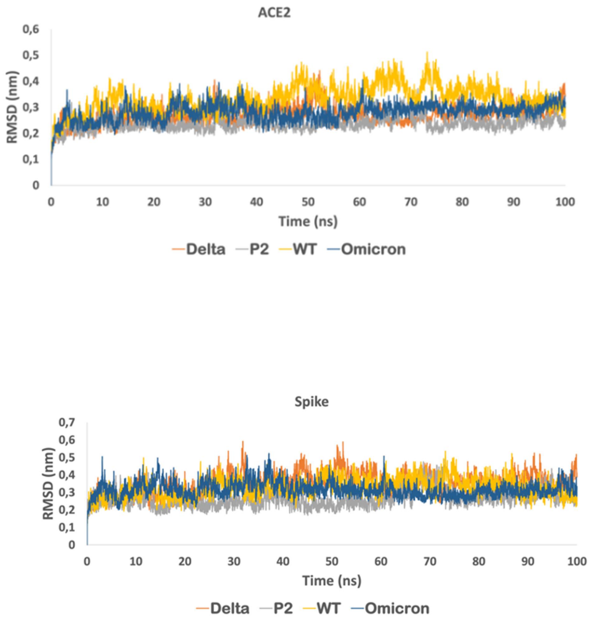

Figure 2. Analysis of the trajectories obtained in the MD simulation. (A) RMSD of ACE2 (B) RMSD of Spike.

Results demonstrate stability in the ACE2 protein for all complexes at around 10ns, with similar RMSD values ranging from 0.2 to 0.4 nm.

The RMSF analysis of the ACE2 trajectory (Fig. 3A) showed no significant fluctuations, limited, wich were limited to loop regions.

The Spike protein (Fig. 3B) showed that the Lys444 residue of the delta variant had the highest fluctuation peak of 0.20 nm, followed by the omicron variant (0.16 nm), P2 variant (0.16 nm), and wild-type (0.14 nm). Lys444 is located close to Gly446, Tyr449, and Gln498, which have polar interactions with ACE2, according to a study by Sironi et al. [16].

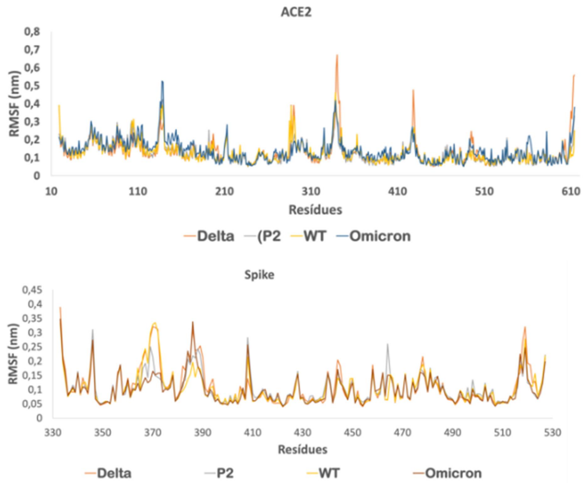

Figure 3. Analysis of the trajectories obtained in the MD simulation. (A) RMSF of ACE2 (B) RMSF of Spike.

The Tyr449 residue is situated near to Leu452, which was mutated to arginine in the delta variant. Other residues with high fluctuation peaks are located in loop regions.

B. MLP

Table 1 shows the five most significants pairs for each complex. The residue importance values for each pair were determined by finding the average LRP-0 value assigned to these pairs in the generated MLPs. Key residues responsible for differences in binding between Spike variants and ACE2 have been identified. Some of these were previously noted in previous studies.

TABLE I. Residues Importance obtained from MLP

Variant	MLP	
	Residue Pair (ACE, Spike)	Importance Value
Wild	(SER106, GLY485) (VAL107, PHE486) (GLN89, SER477) (SER19, PRO479) (ALA71, GLU484)	1.0 0.99 0.98 0.89 0.82
Delta	(ASP30, GLU484) (GLN24, LYS417) (GLY352, ARG408) (ALA65, SER443) (ASN33, GLN498)	1.0 0.71 0.68 0.62 0.60
Omicron	(GLU329, SER438) (GLN42, SER349) (TYR381, GLY502) (GLY352, ASN448) (GLY354, GLY504)	1.0 0.96 0.92 0.91 0.9
P2	(PRO321, ARG403) (SER19, ASN477) (SER19, PRO479) (GLN325, SER371) (GLU37, THR415)	1.0 0.92 0.87 0.84 0.81

The analysis of the results highlights the role of key ACE2 residues GLN24, GLN42, GLN325, GLU329, and

GLY354 in interaction with protein S. Moreover, ACE2 residue SER19, which was commonly seen among the pairs, is also important. A mutationin this residue S19 to P increased the interaction between ACE2 and Spike protein. However, mutations in residues ASN33 (N33I) and GLY352 (G352V) were found to reduce this interaction [17].

Our results highlight the key residues in the Spike protein that can contribute to variations in binding between Spike variants and ACE2. The mutation of the LYS417 to K417N, increases virus transmissibility. The SER477 residue (S477N mutation) enhances binding affinity. The GLU484 residue, when mutated to E484K, has been linked to antibody resistance [4]. The only exception is the simultaneous presence of the (SER19, PRO479) pair in both the Wild and P2 variants, as no other pair of residues showed were significance across other variants.

C. *RF*

Table 2 displays the top five residues that were determined by the Random Forest model, based on their Gini importance values.

TABLE II. RESIDUES IMPORTANCE OBTAINED FROM RF.

Variant	RF	
	Residue Pair (ACE, Spike)	*Importance Value*
Wild	(SER19, VAL483) (SER19, CYS488) (SER19, CYS480) (SER44, TYR505) (SER19, GLN474)	1.0 0.95 0.60 0.52 0.52
Delta	(ALA36, ASN501) (GLY66, ASN501) (ALA342, THR500) (ASN103, TYR505) (LYS68, ASN501)	1.0 0.98 0.93 0.71 0.64
Omicron	(ALA25, ASN417) (GLN24, ASN417) (ILE21, ASN417) (LYS353, ARG498) (THR27, ASN417)	1.0 0.98 0.97 0.94 0.88
P2	(SER106, LYS484) (SER19, CYS480) (SER105, ASN487) (GLY104, ASN487) (SER105, LYS484)	1.0 0.90 0.87 0.80 0.79

The residues pairs identified by the Random Forest model differed from those identified by the MLP model. However, some residues were identified by both methods. Several of these residues have been previously reported, including TYR505 in the Spike protein, whose mutation can increase transmission [4], and ARG498 in the Omicron variant, which leads to increased binding affinity with ACE2 [18]. SER19, LYS353, and THR27 are crucial residues in ACE2 [17]. SER19 was found repeatedly among pairs and variants. The exception was the (SER19, CYS480) pair, which was present in both the Wild and P2 variants, but no other residue pair was present in multiple variants.

IV. CONCLUSIONS AND FUTURE WORK

The interaction between the Spike and ACE2 proteins is crucial in determining the replication rate of SARS-CoV-2 and affects the progression of the disease in infected patients. SARS-CoV-2 exhibits a high mutation rate, as evidenced by the emergence of various variants over the past two years. Polymorphisms in the coding regions of ACE2 may impact a patient's susceptibility to the disease, as well as its severity, and clinical outcome. However, the impact of mutations and polymorphisms on the stability and interaction between the SARS-CoV2-ACE2 complex is not yet fully understood.

In our work, we combined molecular dynamics simulations and machine learning techniques to examine the interaction between SARS-CoV-2 variants and human ACE2. The simulations provided insight into the protein complex interaction, while ML methods identified important residues in the binding region.

Our molecular dynamics simulations showed stability similarities among the variants. The ACE2 protein complex with Spike-Wild showed slightly lower stability, as indicated by RMSD values, compared to the SARS-CoV-2 variant complexes. This aligns with the expectation that mutations in the Spike interaction region increase stability. The ACE2 protein in the wild-type complex is therefore more flexible and less stable. The Spike protein in the Delta variant had slightly higher RMSF values, with a peak at Tyr444 near key residues that interact with ACE2, including Tyr449 near the L452R mutation. Replacing the hydrophobic Leucine with the polar Arginine may enhance intra- and intermolecular interactions.

We achieved an accuracy score of 1 and loss values less than 0.005 for both machine learning methods using the test set. High accuracy and low loss on test data suggest that the model is performing well, not guarantee that the model is not overfitting. Further evaluation using other data sources, such as cross-validation, is needed to determine if overfitting is present.

The ML approaches successfully identified key residues from both proteins responsible for differences in binding region, some of which have been previously reported in the literature. This demonstrates that our method was able to identify residues that significantly contribute to the distinction between virus and host interaction due to mutations in Spike (variants) and ACE2 polymorphisms.

Our study shows that machine learning can simplify the complexity of virus-host interactions by reducing dimensionality and identifying crucial residues. Our findings indicate that there may be additional important residues beyond those previously considered, which can impact the interaction between Spike and ACE2 proteins. These residues could account for differences in stability and affinity, leading to varying levels of susceptibility to SARS-CoV-2 and resulting in varying degrees of disease severity. In our work, we aim to gain a deeper understanding of the relationship between mutations and the affinity between

Spike-ACE2 by not only exploring other variants, but also incorporating various machine learning methods.

ACKNOWLEDGMENT

Fundacao de Amparo a Pesquisa do Estado de São Paulo (FAPESP) for Sponsorship (2022/02782-5; 2022/03038-8) and Superintendence of Technology and Information at USP for the High-Performance Computing (HPC) Computing Resources.

REFERENCES

[1] WHO Director General's Speeches. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. WHO Director General's speeches, n. March, p. 4, 2020. Available from: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 2023.01.31

[2] WHO. Tracking SARS-CoV-2 variants. Who. Disponível em: Avalilable from: https://www.who.int/activities/tracking-SARS-CoV-2-variants. 2023.01.31

[3] S. Choudhary, K. Sreenivasulu, P. Mitra, S. Misra and P. Sharma. "Role of genetic variants and gene expression in the susceptibility and severity of COVID-19". Annals of Laboratory Medicine, vol. 41, no. 2, pp. 129–138, 2020.

[4] J. Zepeda-Cervantes et al. "Implications of the Immune Polymorphisms of the Host and the Genetic Variability of SARS-CoV-2 in the Development of COVID-19."Viruses, vol. 14, no 1, pp. 1-34, 2022.

[5] R. Peng, L-A. Wu, Q. Wang, J. Qi and G. F. Gao " Cell entry by SARS-CoV-2." Trends in Biochemical Sciences, vol. 46, no.10, pp 848–860, 2021

[6] O. Fleetwood, M. A. Kasimova, A. M. Westerlund and L. Delemotte. "Molecular Insights from Conformational Ensembles via Machine Learning." Biophysical Journal, vol. 118, no. 3, pp. 765–780, 2020.

[7] H. M. Berman et al. "The Protein Data Bank", Nucleic Acids Research, vol. 28, no 1, pp. 235-242, 2000.

[8] A. Sali and T. L. Blundell "Comparative protein modelling by satisfaction of spatial restraints", J. Mol. Biol., vol. 234, no. 1, pp. 779-815, 1993.

[9] M. Mohammadi, M. Shayestehpour and H. MirzaeiI "The impact of spike mutated variants of SARS-CoV2 [Alpha, Beta, Gamma, Delta, and Lambda] on the efficacy of subunit recombinant vaccines." Brazilian Journal of Infectious Diseases, vol. 25, no. 4, pp. 101606, 2021.

[10] J. Huang and A. D. MacKerell "CHARMM36 all-atom additive protein force field: Valida-tion based on comparison to NMR data." J. Comput. Chem., vol. 34, no 1, pp. 2135-2145 2013.

[11] Scikit-learn: Machine learning in Python. Available from https://scikit-learn.org/stable/index.html. 2023.01.31

[12] Pandas-Python Data Analysis Library. Available form https://pandas.pydata.org/. 2023.01.31

[13] Numpy. Available from https://numpy.org/ .2023.01.31

[14] D. P. Kingma and J. Ba. 2014. Adam: a method for stochastic optimization. arXiv, arXiv:1412.6980. [Published as a conference paper at ICLR 2015].

[15] G. S. Montavon, S. Bach, A. Binder, W. Samek and K.-R. M"uller. "Explaining nonlinear classification decisions with deep Taylor decomposition." Pattern Recognit. vol. 65, pp. 211–222, 2017.

[16] M. Sironi et al. "SARS-CoV-2 and COVID-19: A genetic, epidemiological, and evolutionary perspective." Infection, Genetics and Evolution, vol. 84, 104384, pp. 1-15, 2020.

[17] K. Suryamohan et al. "Human ACE2 receptor polymorphisms and altered susceptibility to SARS-CoV-2." Communications Biology, vol. 4, no. 1, pp. 1–11, 2021.