Supplementary Material - 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]

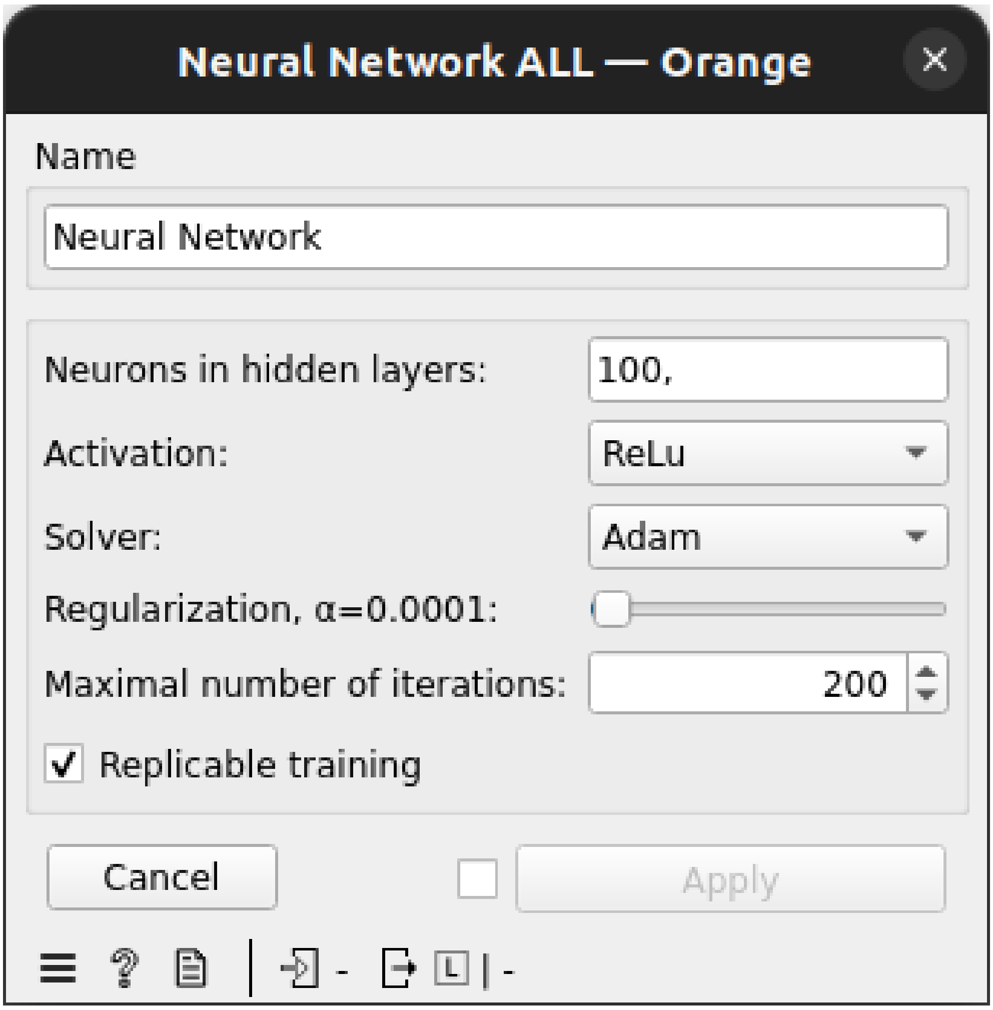
1 Laboratory of Bioinformatics and Systems (LBS)
Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil
2 Laboratory of Molecular Modeling and Bioinformatics (LAMMB)
Universidade Federal São João Del Rey (UFSJ), Sete Lagoas, Minas Gerais, Brazil
∗Same contribution level
raquelcm@dcc.ufmg.br

Table S1. SARS-CoV-2 VOCs and their defining mutations.

Variants	Scientific name	Spike Glycoprotein Defining Mutations
Alpha	B.1.1.7	Del 69-70; Del 144; N501Y; A570D; D614G; P681H; T716I; S982A; D1118H.
Beta	B.1.351	D80A; D215G; Del 241-243; K417N; E484K; N501Y; D614G; A701V.
Gamma	P.1	L18F; T20N; P26S; D138Y; R190S; K417T; E484K; N501Y; D614G; H655Y; T1027I; V1176F.
Delta	B.1.617.2	T19R; G142D; Del 156-157; R158G; L452R; T478K; D614G; P681R; D950N.
Omicron	BA.1	A67V; Del 69-70; T95I; Del 142-144; Y145D; Del 211; L212I; G339D; S371L; S373P; S375F; K417N; N440K; G446S; S477N; T478K; E484A; Q493R; G496S; Q498R; N501Y; Y505H; T547K; D614G; H655Y; N679K; P681H; N764K; D796Y; N856K; Q954H; N969K; L981F.
Omicron	BA.2	T19I; Del 24-26; A27S; G142D; V213G; G339D; S371F; S373P; S375F; T376A; D405N; R408S;

		K417N; N440K; S477N; T478K; E484A; Q493R; Q498R; N501Y; Y505H; D614G; H655Y; N679K; P681H; N764K; D796Y; Q954H; N969K.
Omicron	BA.2.12.1	T19I; Del 24-26; A27S; G142D; V213G; G339D; S371F; S373P; S375F; T376A; D405N; R408S; K417N; N440K; L452Q; S477N; T478K; E484A; Q493R; Q498R; N501Y; Y505H; D614G; H655Y; N679K; P681H; S704L; N764K; D796Y; Q954H; N969K.
Omicron	BA.4	T19I; Del 24-26; A27S; Del 69-70; G142D; V213G; G339D; S371F; S373P; S375F; T376A; D405N; R408S; K417N; N440K; L452R; S477N; T478K; E484A; F486V; Q498R; N501Y; Y505H; D614G; H655Y; N679K; P681H; N764K; D796Y; Q954H; N969K.

Source: (HODCROFT, 2021).



Neural Network ALL — Orange

Name

Neural Network

Neurons in hidden layers: 100,

Activation: ReLu

Solver: Adam

Regularization, α=0.0001:

Maximal number of iterations: 200

☑ Replicable training

Figure S1. Parameters used in the neural network analysis of Orange Data Mining.

Table S1. Formula for analysis metrics.

Methods	Formulas
Accuracy	(True Positives + True Negatives) / Total elements
F1-Score	2 x (Precision and Recall) / (Precision + Recall)
Precision	True Positives / (True Positives + False Positives)
Recall (or Sensitivity)	True Positives / (True Positives + False Negatives)

Source: Adapted from MARIANO *et al.* (2021).

Table S2. RMSD values between selected templates.

	7CWL	7KRQ	7N1Q	7N1U	7SBK	7SBP	7SBS	7TNW	8D55
7CWL		1.672	2.308	1.133	1.249	1.333	1.513	2.176	1.887
7KRQ	1.672		0.810	1.167	0.660	0.540	1.081	0.620	0.485
7N1Q	2.308	0.810		1.732	1.174	0.888	0.589	1.124	0.876
7N1U	1.133	1.167	1.732		0.784	0.792	0.869	1.560	1.258
7SBK	1.249	0.660	1.174	0.784		0.388	0.644	0.904	0.868
7SBP	1.333	0.540	0.888	0.792	0.388		0.858	0.732	0.649
7SBS	1.513	1.081	0.589	0.869	0.644	0.858		0.925	1.043
7TNW	2.176	0.620	1.124	1.560	0.904	0.732	0.925		0.484
8D55	1.887	0.485	0.876	1.258	0.868	0.649	1.043	0.484	

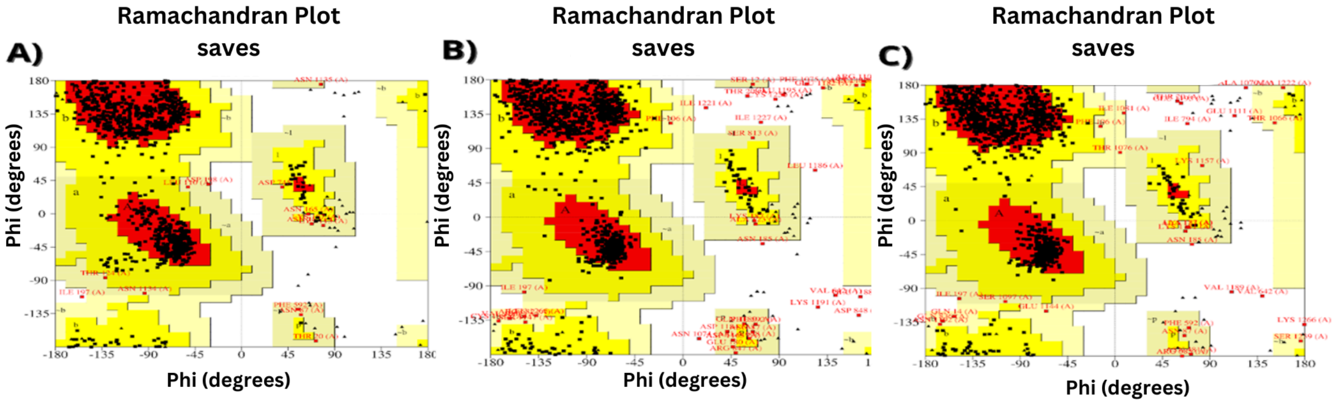


Figure S2. Ramachandran plots. Figure A shows the Ramachandran plot of the protein PDB ID 7CWL; Figure B shows the plot of the Spike protein of the wild-type strain modeled by the MODELLER tool; and Figure C shows the plot of the protein modeled with the single mutation N501Y. In the Ramachandran plot analysis for the template protein 7CWL (PDB ID) (Supplementary material), 85.9% of the 1,073 residues in chain A were in allowed regions, with only one residue (0.1%) in a disallowed region. Additionally, 12.7% of residues were in additionally allowed regions, and 1.4% in generously allowed regions. VERIFY 3D analysis showed that 69.33% of residues had a 3D-1D score of 0.1 or higher, serving as a reference for model selection. In the wild-type model (Figure X, Supplementary Material), 83.5% of the 1,273 residues were in allowed regions, 13.4% in additionally allowed, 2.0\% in generously allowed, and 1.1\% in disallowed regions. Due to the large number of generated models and their similarity to the template, no further refinement was performed.

References

HODCROFT, E. B. **CoVariants: SARS-CoV-2 Mutations and Variants of Interest**. Disponível em: <https://covariants.org/>. Acesso em: 30 nov. 2022.

PIRES, D. E. V. et al. Cutoff Scanning Matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns. **BMC genomics**, v. 12 Suppl 4, n. Suppl 4, p. S12, 22 dez. 2011.

MARIANO, D. Métricas de avaliação em machine learning: acurácia, sensibilidade, precisão, especificidade e F-score. Em: MARIANO, D. et al. (Eds.). **BIOINFO - Revista Brasileira de Bioinformática e Biologia Computacional**. 1. ed. [s.l.] Alfahelix, 2021