Computational and Structural Biotechnology Journal 27 (2025) 1686–1692

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Short communication

Transfer learning towards predicting viral missense mutations: A case study on SARS-CoV-2

Shaylyn Govender [a,1], Emily Morgan [a,1], Rabelani Ramahala [a], Kevin Lobb [b], Nigel T. Bishop [c,d,*], Özlem Tastan Bishop [a,d,**]

[a] Research Unit in Bioinformatics (RUBi), Department of Biochemistry, Microbiology and Bioinformatics, Rhodes University, Makhanda 6139, South Africa
[b] Department of Chemistry, Rhodes University, Makhanda 6139, South Africa
[c] Department of Pure and Applied Mathematics, Rhodes University, Makhanda 6139, South Africa
[d] National Institute for Theoretical and Computational Studies (NITheCS), South Africa

ARTICLE INFO

Keywords:
NSP10-NSP16 complex
Main protease
Papain-like protease
Spike protein
Dynamic residue network
Machine learning

ABSTRACT

Understanding viral evolution and predicting future mutations are crucial for overcoming drug resistance and developing long-lasting treatments. Previously, we established machine learning (ML) models using dynamic residue network (DRN) metric data and leveraging a vast amount of existing mutation data from the SARS-CoV-2 main protease (M^{pro}). Here, we sought to assess the generalizability and robustness of the current models across other SARS-CoV-2 proteins. To achieve this, for the first time, we employed a transfer learning (TL) approach, allowing us to determine the extent to which M^{pro} trained models could be applied to other SARS-CoV-2 proteins. The TL results were highly promising, with artificial neural network (ANN) and random forest (RF) correlation coefficients for M^{pro} closely matching those of NSP10, NSP16, and PL^{pro}. The ANN |R| value for M^{pro} was 0.564, while NSP10, NSP16, and PL^{pro} had values of 0.533, 0.527, and 0.464, respectively. Similarly, the RF |R| value for M^{pro} was 0.673, compared to 0.457, 0.460, and 0.437 for NSP10, NSP16, and PL^{pro}, respectively. Interestingly, we did not observe a strong correlation for the spike (S) protein monomer and its domains. The low p-values that are associated with the correlation |R| values show that the linear correlations between predicted and actual mutation frequencies are statistically significant. This indicates that TL may generalize well across structurally related viral proteins using DRN-derived ML model from M^{pro}. Overall, we aim to develop a universal ML model for predicting missense mutation frequencies in viral proteins, and this study lays the foundation for that goal.

1. Introduction

The COVID-19 pandemic, caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has negatively impacted global health and socio-economic stability [1,2]. The ongoing evolution of the virus emphasizes the urgent need for approaches to understand its progression and predict its future trends [3]. Such efforts are essential to develop effective and long-lasting treatment strategies [4]. More importantly, these approaches should be adaptable to other viruses to enhance our capacity for rapid response and improving overall pandemic preparedness.

Recently, there has been growing momentum in developing robust pipelines to predict genetic mutations of the SARS-CoV-2. Due to the public availability of massive viral genomic data, a modest number of computational models have been developed. To mention a few, Zhou et al. developed a model called transformer-based mutation prediction framework (TEMPO) that samples the SARS-CoV-2 sequence data using phylogenetic trees and predicts mutation sites [5]. Obermeyer et al. designed the PyRo model to predict the fitness, growth and prevalence of new lineages from the large SARS-CoV-2 genomic datasets using hierarchical Bayesian multinomial logistic regression approach [6]. Maher et al. designed a model that predicts spreading SARS-CoV-2 mutations

* Corresponding author at: Department of Pure and Applied Mathematics, Rhodes University, Makhanda 6139, South Africa.
** Corresponding author at: Research Unit in Bioinformatics (RUBi), Department of Biochemistry, Microbiology and Bioinformatics, Rhodes University, Makhanda 6139, South Africa.
E-mail addresses: n.bishop@ru.ac.za (N.T. Bishop), o.tastanbishop@ru.ac.za (Ö. Tastan Bishop).
[1] Equally contributed first authorship

https://doi.org/10.1016/j.csbj.2025.04.029
Received 22 February 2025; Received in revised form 16 April 2025; Accepted 22 April 2025
Available online 22 April 2025

using a logistic regression algorithm with epidemiological features and positive selection features as inputs [7]. Incorporating Machine Learning (ML) algorithms in creating these predictive pipelines improves the performance of these models. ML has been previously used to predict nucleotide mutation rate, explain genetic variability and the effects of these mutations [8]. Saldivar-Espinoza et al. used Artificial Neural Networks (ANN) to forecast the positions to hold a recurrent mutation in the SARS-CoV-2 genome and the actual recurrent mutations [9]. In addition, Choi et al. employed artificial intelligence (AI) models to identify new mutations on the RBD region of the Spike protein using clade information [10].

In our previous study, we established ANN and random forest (RF) ML models to predict both the mutation frequency of main protease (M^{pro}) residues (a regression problem) and whether a given M^{pro} residue would mutate or not (classification) [3]. As we aim to increase the accuracy of these predictions, we also want to assess the generalizability and robustness of these models across different datasets via a transfer learning (TL) approach. While ML enhances the prediction of mutation positions and frequencies within a given dataset, TL extends this capability by effectively applying learned knowledge to other relevant proteins [11]. By using pre-trained models, TL may enable more accurate mutation predictions for less characterized proteins, reducing the dependence on experimental data.

M^{pro} plays a key role in antiviral drug development, but several other proteins within the SARS-CoV-2 genome are equally important and serve as promising drug targets. The proteins of interest in this study include the NSP10-NSP16 complex, papain-like protease (PL^{pro}), and the spike (S) protein. NSP10-NSP16 is a 2′-O-methyltransferase (MTase) involved in viral RNA capping, enabling the virus to evade the immune system in humans [12]. PL^{pro} is responsible in cleavage and maturation of viral polyproteins, assembly of the replicase-transcriptase complex, and disruption of host responses [13]. The S glycoprotein promotes entry of the virus into the host cell [14]. Understanding mutation patterns within these proteins is highly important, as mutations can alter protein function [15–18].

In this study, we extend the application of our trained M^{pro} ML models to these additional SARS-CoV-2 proteins and assess their predictive performance. We demonstrate that, without re-training, M^{pro} model can be applied to other SARS-CoV-2 proteins, and significantly enhances the use of ML in viral mutation frequency prediction. Our analysis highlights the strengths and limitations of TL across different structures. Overall, this work not only investigates the adaptability of pre-trained ML models across diverse viral proteins but also explores strategies to improve model transferability and accuracy. Ultimately, our approach contributes towards development of universal ML models for predicting mutation frequencies across multiple viral targets, an important step towards pandemic preparedness and rational drug design.

2. Methodology

2.1. Data retrieval

The crystal structures of SARS-CoV-2 NSP10-NSP16 complex (PDB ID: 6W4H) and PL^{pro} (PDB ID: 6WZU) were obtained from the RCSB Protein Data Bank [19]. The closed state of trimeric S glycoprotein (PDB ID: 6VXX) was retrieved from BioExcel-CV19 [20,21] (https://covid.molssi.org//structures/#6vxx-spike).

SARS-CoV-2 sequences for the full length of NSP10, NSP16, PL^{pro} and S protein from Alpha, Beta, Gamma, Delta and Omicron lineages were retrieved from the Global Initiative on Sharing Avian Influenza Data (GISAID; https://gisaid.org/) [22] for the period between 1 December 2019 and 24 February 2024. The sequence selection was based on high coverage and the availability of patient status information.

The retrieved sequences were submitted to the GISAID CoVsurver tool [23] (https://gisaid.org/database-features/covsurver-mutations-app) to identify mutations. SNVs specific to each protein were filtered from the CoVsurver output (*.tsv* files) using an in-house Python script, which also computed mutation frequencies within each variants of concern (VOC) dataset to identify key mutations for further analysis.

2.2. Treatment of zinc metals in proteins

Protonation steps were performed using H++ (version 4.0) [24] at pH values of 7.5 for NSP10-NSP16 [25] and 7.0 for PL^{pro} (default parameter) with a salt concentration of 0.15 M. In both proteins, structural zinc ions were retained and coordinating cysteine residues were manually adjusted to deprotonated cysteines (CYM) to maintain correct coordination geometry to the metal. For metal parameterization, metal-containing sites were converted to mol2 format using metalpdb2mol2.py, and parameters were derived using *MCPB.py* [26] in AmberTools [27,28] with quantum mechanical calculations in Gaussian09 [29].

Each system was assembled in AmberTools *tleap* and ff14SB force fields were applied [27]. The systems were solvated using the TIP3P water models, and Na^+ and Cl^- ions were added to maintain a physiological salt concentration of 0.15 M. Final topology and coordinate files were converted to GROMACS format using *ACPYPE* [30]. Periodic boundary conditions were optimized by adjusting the simulation box to cubic shape for NSP10-NSP16 and PL^{pro} using *editconf*.

2.3. Molecular dynamics simulations

MD simulations were conducted using GROMACS on the Centre for High Performance Computing (CHPC) in Cape Town. Energy minimization for NSP10-NSP16 and PL^{pro} was performed using the steepest descent algorithm, terminating when the maximum force dropped below 1000 kJ/mol/nm or after 50,000 steps. Temperature equilibration was conducted in the NVT ensemble at 310 K for 100 ps, using Velocity-rescale (V-rescale) coupling. This was followed by pressure equilibration in the NPT ensemble at 1 atm and 310 K for 100 ps utilizing the Parrinello-Rahman barostat for NSP10-NSP16 and the stochastic cell rescaling (C-rescale) barostat for PL^{pro}.

Four independent 1 μs production runs were conducted for NSP10-NSP16, and four 100 ns runs for PL^{pro}, each with a 2 fs time step. The difference in simulation times reflects the point at which each system reached equilibrium (Figure S1). In all systems, a 1.0 nm cutoff was applied for short-range electrostatics and van der Waals interactions, while long-range electrostatics were handled using the Particle Mesh Ewald (PME) method with a Fourier spacing of 0.16 nm. Periodic boundary conditions and bond constraints were applied using the LINCS algorithm [31].

Trajectory data for the S protein was sourced from BioExcel-CV19 [20,21] (https://covid.molssi.org//simulations/#gromacs-60-ns-md-of-sars-cov-2-spike-trimer-all-atom-model), simulated in a $20 \times 20 \times 20$ nm TIP3P water box (0.1 M NaCl) using the Charmm27 force field. MD simulations were performed using GROMACS 2018.8 on the Puhti cluster at CSC-IT on the isothermal-isobaric (NPT) ensemble, maintaining a pressure at 1 bar and a temperature of 300 K. A 60 ns all-atom trajectory was generated with frames saved every 80 ps (Figure S1).

Periodic conditions were removed after simulation for all proteins, and measures such as root mean squared deviation (RMSD), root mean square fluctuation (RMSF) and radius of gyration (Rg) were calculated using GROMACS tools (*gmx rms*, *gmx rmsf*, and *gmx gyrate*, respectively), visualized using VMD and Python libraries (Seaborn, Pandas, Matplotlib, and NumPy).

2.4. Dynamic residue network metric calculations

DRN calculations were performed for all the proteins following the methodology by Barozi et al. [3], and eight *averaged* centrality metrics

were calculated per MD trajectory: *betweenness centrality* (*BC*), *closeness centrality* (*CC*), *degree of centrality* (*DC*), *eigenvector centrality* (*EC*), *eccentricity* (*ECC*), *Katz centrality* (*KC*), *shortest path* (*L*), and *pagerank* (*PR*). The formulae on how each metric is calculated can be found in Table S1 (adapted from [32]).

Network calculations were performed using the *calc_network.py* tool from the MDM-TASK suite (https://github.com/RUBi-ZA/MD-TASK/tree/mdm-task-web/src) [32,33], utilizing complete MD trajectories (*.xtc*) and corresponding topology files (*.pdb*). For NSP10 and NSP16, PL^{pro}, S protein and the domains of S protein the *calc_network.py* script was used to analyze MD trajectories. For NSP10 (residues 23–133) and NSP16 (residues 1–293), calculations were performed on 5,000-time frames sampled from the 1 million frames generated over each 1 μs simulation, while for PL^{pro}, analysis was conducted on 100 ns trajectories consisting of 5,000-time frames. In both cases, a step size of 5 was applied, and Euclidean distances were set at 6.7 Å. Metrics were averaged across all frames to yield residue-specific values. For the S protein, calculations were performed on a 60 ns trajectory that comprised of 750-time frames using a step size of 5. Only chain A was considered, thereafter divided into S1 subunit (residues 15–685) and S2 subunit (residues 686–1145), the N terminal domain (NTD) (residues 15–305) and the receptor binding domain (RBD) (residues 319–541). Each domain of the S protein, as well as the entire protein (chain A), was analyzed individually.

2.5. Machine learning models

In our previous work [3], we developed ML models to predict mutation positions and frequencies in M^{pro} of SARS-CoV-2. The software packages used include Python version 3.10.13 [34], Keras version 3.0.5 [35] (https://github.com/fchollet/keras), TensorFlow version 2.15.0 [36], and Scikit-learn version 0.24.2 [37]. For each residue, the predictor dataset comprised eight DRN metrics, BLOSUM62 matrix scores, solvent-accessible surface area (SASA), B-factor, and RMSF. These features were calculated using GROMACS v2021.1 [38], forming a matrix where rows represented protein residues and columns represented 31 features. The eight DRN metrics and three specific protein features (RMSF, B-factor and SASA) were calculated for each residue per MD simulation trajectory, and averaged over the MD trajectory replicates per protein. The target values were derived from mutation frequencies of each residue, transformed using $\log_{10}(1 + \text{mutation frequency})$ to address extreme values.

Previously [1], feature scaling was applied using the Scikit-learn StandardScaler. However, for TL the same scaling needs to be used on the datasets of the various proteins. This was achieved using an in-house script that normalizes features using the normalization function (also see Table S2):

$$X_{NORMALIZED} = (X - X_{MIN}) / (X_{MAX} - X_{MIN})$$

Here, X represented the data for each feature, X_{MIN} was the minimum value found within a given X, and X_{MAX} was the maximum value for X. For each feature, X_{MAX} and X_{MIN} were evaluated for M^{pro} and then these values were used for feature normalization of the other proteins. Thus, for M^{pro}, $X_{NORMALIZED}$ was in the range [0,1], but this was not the case for the other proteins.

A consequence of the change of scaling was that the ML models needed to be retrained, which was performed with 400 different seeds (ranging from 0–399) to identify the optimal fit, minimizing overfitting and underfitting. All other aspects of the models are as described in Barozi et al. [3]. The best-performing M^{pro} ANN and RF models and normalization formula were saved as.*pkl* files using the *joblib* module [39]. These files were then applied to the NSP10, NSP16, PL^{pro} and the S protein datasets. Model performance was evaluated by comparing predicted and target values, visualized through a scatterplot. To quantify the results obtained, the Pearson correlation coefficient (R value) (Table S2) and associated p-values were calculated using the *pearsonr* function from SciPy version 1.15.2 package (https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html).

The SHapley Additive exPlanations (SHAP) [40] method was used to evaluate the contribution of each feature to the output of the trained ANN and RF models. SHAP values are numerical scores assigned to each feature for a given prediction [40], representing how much that feature contributes to moving the prediction away from the model's baseline output (typically the mean prediction). These values are calculated by considering all possible combinations of input features and averaging the marginal contribution of a feature across these combinations. For the RF model, a TreeExplainer object was used to compute SHAP values due to the model's tree structure (https://shap.readthedocs.io/en/latest/api.html). In contrast, for the ANN, a model-agnostic Explainer object was chosen due to the network's non-tree architecture. To visualize the results, a summary bar plot was generated for each model, displaying the average absolute SHAP value for each feature indicating global feature importance.

3. Results and discussion

3.1. Correlation between post MD metrics and target data is observed

We investigated the eight *averaged* DRN metric values (*BC, CC, DC, EC, ECC, KC, L* and *PR*), three physicochemical properties (SASA, B-factor and RMSF) and the BLOSUM62 matrix for each residue of all selected proteins to determine if any relationship existed between these features and residue mutation frequencies. This was performed by calculating and analyzing the correlation (R) between these features and the residue mutation frequencies, expressed in logarithmic form as $\log_{10}(1 + \text{mutation frequency})$ for each residue per protein (or per domain in the case of the S protein: Chain A, S1 subunit, S2 subunit, NTD, and RBD) (Fig. 1).

Similar to the M^{pro} |R| values, the selected protein features showed low to moderate correlation with the target data. Among the DRN metrics, the *DC* and *KC* features appeared to be the most consistent, with |R| values above 0.2 across all proteins. Of the selected proteins, M^{pro} had the most consistently high |R| values, and was the only protein with no |R| values below 0.2 for a single metric. The highest |R| values were found for two DRN features of the NSP10 protein (0.52 for *DC* and 0.55 for *PR*), but overall, this protein had a greater variability in |R| values compared to M^{pro}. The S2 subunit appeared to have consistently low |R| values, lower than both the whole S protein Chain A and the S1 domain. The NTD and RBD domains were more variable, with some features showing higher |R| values than the S protein Chain A and S1 domain, while others were lower.

We also observed varying |R| values across other features. For example, residue T in the BLOSUM62 matrix showed the highest values in NSP10 and PL^{pro} (0.46 and 0.30, respectively). Notably, in M^{pro} certain BLOSUM62 residues exhibited higher correlation values than some DRN metrics, such as residues P (0.23), A (0.27), and V (0.23).

From the |R| values alone, it is difficult to determine which proteins will result in the best TL performance, or whether similar performance is to be expected across datasets. A key consideration is that not all features have equal importance in all trained models. Feature importance can be affected by the chosen subset of training samples and ML model type. Consequently, the feature importance for our best ANN and RF models are most likely different.

The feature importance rankings derived from SHAP analysis differ between the chosen RF and ANN models (Figure S2). While the RF model identified averaged BC as the most influential feature, the ANN model assigned the highest importance to the V BLOSUM62 feature. Overall, the magnitude of the average absolute SHAP values was notably lower in the RF model, suggesting that the individual features have a smaller impact on final predictions compared to the ANN. Despite these differences, the *averaged BC* , SASA and V BLOSUM62 features ranked

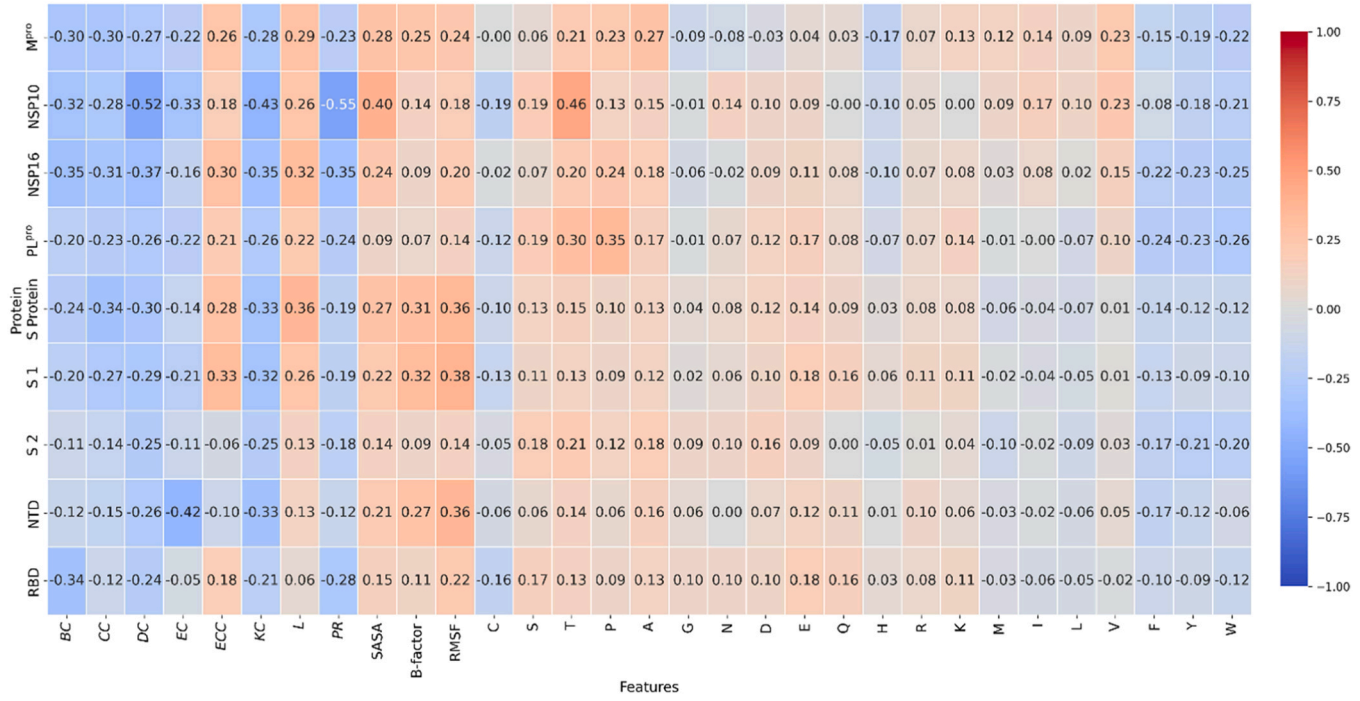

Fig. 1. Heatmap of the R values for the correlation between DRN metrics, post-MD metrics (SASA, B-factor, and RMSF), BLOSUM62, and the target data (observed mutation frequency per residue per protein).

within the top five features with the highest contributions in both models, suggesting their importance in predicting mutation frequency. Conversely, the G BLOSUM62 feature was among the bottom five in both models, implying a consistently small contribution to the predicted mutation frequencies.

The variability in |R| values across proteins for each feature suggests that the feature and target data relationships may likely affect the overall effectiveness of the TL technique. TL application to the selected protein datasets may depend on how well the feature-target relationships generalize. Substantially different correlations between feature and target data may result in limited transferability of these trained models. However, the M^{pro} protein has the most consistent R values across all analyzed features, meaning this data may be a good choice for training the models. Further research is needed to confirm the suitability of this dataset.

3.2. *Transferability of M^{pro} trained ML models to other SARS-CoV-2 proteins*

Next, the ML models developed for M^{pro} were applied to proteins of interest to determine the transferability of the model. The |R| values for each protein for two ML models are given in Table 1, and the scatterplots are presented in Figs. S3-S11. The results demonstrate that transfer ML works better for certain proteins, compared to others.

NSP16 resulted in the best predictive performance using the M^{pro} trained RF model, with a correlation coefficient of 0.460 (RF). NSP10 had the highest |R| value (0.533) for the ANN model among the selected proteins, excluding M^{pro}. Both NSP10 and NSP16 outperformed PL^{pro} in predictive accuracy. This is particularly interesting since M^{pro} and PL^{pro} are both cysteine proteases, making them more functionally similar than NSP10 and NSP16. However, structurally, M^{pro} and the NSP10-NSP16 complex are both dimers, whereas PL^{pro} is a monomer (Fig. 2). This underscores the complexity of mutation prediction, suggesting that it cannot be determined solely by structural or functional similarity alone. We also noted that TL for the S protein monomer and its subunits S1 and S2 gave similar results, indicating that the prediction is not size dependent.

The p-values associated with the |R| values indicate the probability that a correlation value would be as strong or stronger than the observed correlation coefficient if there is no actual correlation. The null hypothesis is that there is no linear correlation between the predicted mutation frequencies and the actual mutation frequencies. It is common practice to regard $p < 0.05$ as indicating a statistically significant correlation. Here all p-values are smaller than 0.004, and in most cases they

Table 1
Correlations and corresponding p-values between actual mutations and predictions of the M^{pro} ML models for each protein/data file. The number of residues in each protein is also provided.

	RF		ANN		Number of residues
	|R| value	**p-value**	**|R| value**	**p-value**	
M^{pro}	0.673	2.773×10^{-41}	0.564	6.468×10^{-27}	304
NSP10	0.457	4.495×10^{-7}	0.533	1.670×10^{-9}	111
NSP16	0.460	8.890×10^{-17}	0.527	2.634×10^{-22}	293
PL^{pro}	0.437	3.886×10^{-16}	0.464	3.438×10^{-18}	315
S protein monomer	0.273	8.908×10^{-21}	0.312	5.284×10^{-27}	1131
S1	0.307	3.957×10^{-16}	0.317	3.703×10^{-17}	671
S2	0.314	5.850×10^{-12}	0.300	5.104×10^{-11}	460
NTD	0.183	1.686×10^{-3}	0.198	6.902×10^{-4}	291
RBD	0.195	3.442×10^{-3}	0.340	2.027×10^{-7}	223

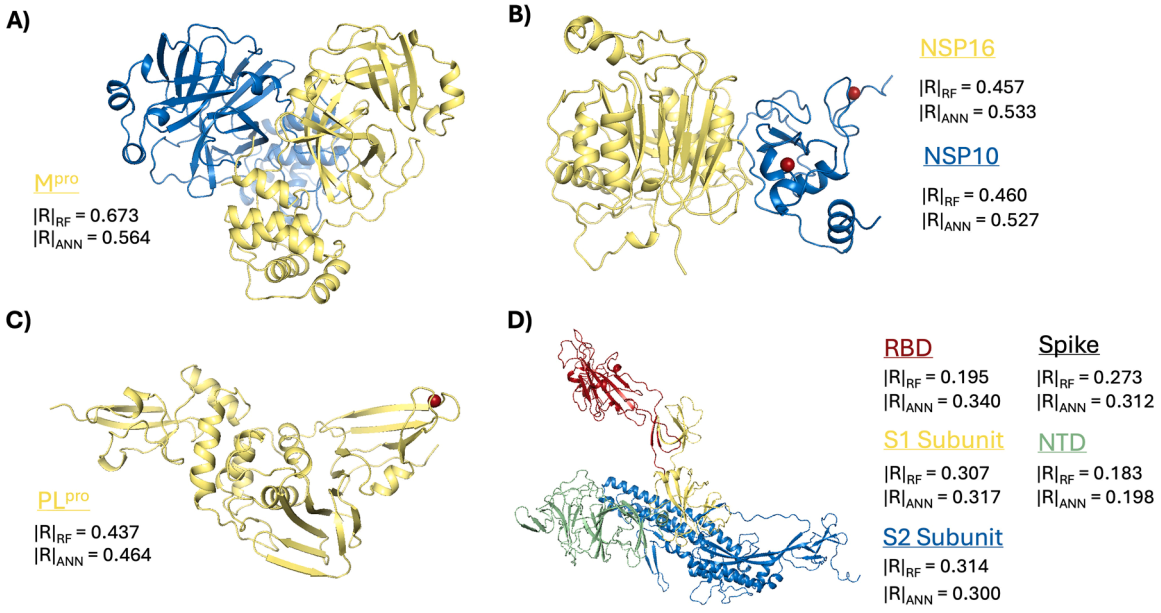

Fig. 2. 3D structural conformation of SARS-CoV-2 proteins investigated with their corresponding |R| values - each ML model denoted as the subscript. **A)** M^{pro} homodimer, **B)** NSP10&16 complex, **C)** PL^{pro} monomeric and **D)** S protein (chain A) monomeric structure. Differentiation of chains are seen in yellow and blue. The S protein domains are colored and labeled accordingly (red is the RBD, green is the NTD, yellow is the rest of S1 subunit which includes the RBD and NTD, blue is the S2 subunit).

are extremely small. Thus, the null hypothesis can be rejected, and there is a statistically significant linear correlation between actual and predicted mutation frequencies for all tested datasets.

Although the performance of models based on the input dataset varies, the overall correlation coefficients show moderate correlation for the globular proteins, i.e. M^{pro}, NSP10, NSP16 and PL^{pro} (Fig. 2). This indicates that TL may generalize well across structurally related viral proteins when using DRN-derived M^{pro} trained ML models. In contrast, the S protein datasets showed lower predictive power, suggesting that M^{pro} derived features may not generalize as well to the highly flexible S protein. Notably, the NTD and RBD displayed the weakest correlation coefficients for RF, possibly indicating that prediction performance is influenced by protein shape and completeness, which impacts residue network construction. However, one must take into consideration that other factors, such as feature importance and differences in target data distribution may also contribute to the lower |R| values observed for these datasets when the RF model is applied. In both cases, the correlation between DRN metric values and mutation frequencies were also variable. The difference in |R| value between the ANN and RF models for the RBD dataset suggests that additional underlying relationships, mainly non-linear dependencies, may be influencing the variation in model performance.

4. Conclusion

While we established the ANN and RF models trained on M^{pro}, their ability to generalize to other SARS-CoV-2 proteins remained untested. To evaluate their robustness and transferability, here, we assessed these models, without further tuning or retraining, on test data from NSP10, NSP16, PL^{pro}, the S monomer protein and its domains, using datasets that shared the same feature set as M^{pro}. This approach allowed us to directly determine the extent to which the M^{pro} trained models could be applied to other SARS-CoV-2 proteins. The outcome of TL was highly promising, as the ANN |R| values for M^{pro} (0.564) and NSP10, NSP16, and PL^{pro} (0.533, 0.527, and 0.464, respectively) were comparable. This was also observed in the RF model.

The features used to develop the ML models are primarily based on network analysis, where residues are treated as nodes connected by edges within a defined distance. Given this approach, it is expected that TL would be more effective for proteins with similar structures. In this study, our ML model was trained on a globular protein (M^{pro}), which may explain why TL performed relatively better on proteins with similar structures, compared to the S protein, which has a highly extended shape.

The results presented in this study pave the way for several promising research directions, including applications to other viral families to predict future mutation frequencies and incorporation of potential mutation information into early drug discovery pipelines. Introduction of additional biochemical features, such as hydrogen bond information, hydrophobic interactions, could further enhance the accuracy of ML models. For the S protein, developing more robust ML models, specifically for this protein, may offer improved predictive power. If such a model proves transferable to other flexible proteins, it could open the door for developing a generalized ML-based mutation predictor that leverages data from proteins with diverse structural characteristics. This study marks an initial step toward achieving that broader objective. Additionally, it presents a reproducible workflow, including MD simulation protocols, DRN metric calculations, feature normalization, and ML training, which can be applied to other similar biochemical research questions.

Funding sources

This project is supported by the Novo Nordisk Foundation and the Pandemic Antiviral Discovery (PAD) Initiative; Grant number: NNF23SA0084504.

Manuscript data

Manuscript data for this article, which includes the SARS-CoV-2 protein datasets, trained ML models and transfer learning script are available in the data.zip file at https://drive.google.com/file/d/1CK5ucc2nulwXC_xsnkXfGvDYjSNZGpFH/view?usp=drive_link. After unzipping this file, a detailed description of all supplementary files and further information can be found in the readme.md file.

The ***datasets*** folder contains the raw dataset files (.csv format) for

SARS-CoV-2 proteins, including M^{pro}, NSP10, NSP16, PL^{pro} and Spike Chain A and its domains. These files are NSP10_apo.csv, NSP16_apo.csv, PLpro_datafile.csv, Target_datafile_Mpro.csv, Target_datafile_NTD.csv, Target_datafile_RBD.csv, Target_datafile_S1.csv, Target_datafile_S2.csv, and Target_datafile_Spike_whole.csv.

The ***models*** folder contains pickled files for the best ANN (ann_model_.pkl) and RF models (rf_model_.pkl), as well as the for the minimum and maximum X values per feature used for normalization during training (data_max_.pkl, data_min_.pkl, X_max_.pkl, and X_min_.pkl).

Additionally, the ***Transfer_Learning.ipynb*** file is present, which is the script for transfer learning using the datasets provided from the ***datasets*** folder and models from the ***models*** folder.

CRediT authorship contribution statement

Govender Shaylyn: Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation. **Morgan Emily:** Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation. **Lobb Kevin:** Methodology. **Ramahala Rabelani:** Methodology, Formal analysis. **Tastan Bishop Özlem:** Writing – original draft, Visualization, Resources, Project administration, Methodology, Funding acquisition, Conceptualization. **Bishop Nigel T.:** Writing – original draft, Validation, Supervision, Methodology, Conceptualization.

Declaration of Competing Interest

Authors declare no conflict of interest.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2025.04.029.

Data availability statements

All data generated or analyzed during this study are included in this published article and its supplementary data (please see https://drive.google.com/file/d/1CK5ucc2nulwXC_xsnkXfGvDYjSNZGpFH/view?usp=drive_link).

References

[1] Cucinotta D, Vanelli M. WHO Declares COVID-19 a pandemic. Acta Bio Med Atenei Parm 2020;91:157–60. https://doi.org/10.23750/abm.v91i1.9397.

[2] COVID-19 Deaths WHO COVID-19 Dashboard n.d. ⟨https://data.who.int/dashboards/covid19/cases⟩.

[3] Barozi V, Chakraborty S, Govender S, Morgan E, Ramahala R, Graham SC, et al. Revealing SARS-CoV-2 Mpro mutation cold and hot spots: dynamic residue network analysis meets machine learning. Comput Struct Biotechnol J 2024;23:3800–16. https://doi.org/10.1016/j.csbj.2024.10.031.

[4] Angius F, Puxeddu S, Zaimi S, Canton S, Nematollahzadeh S, Pibiri A, et al. SARS-CoV-2 evolution: implications for diagnosis, treatment, vaccine effectiveness and development. Vaccines 2024;13:17. https://doi.org/10.3390/vaccines13010017.

[5] Zhou B, Zhou H, Zhang X, Xu X, Chai Y, Zheng Z, et al. TEMPO: a transformer-based mutation prediction framework for SARS-CoV-2 evolution. Comput Biol Med 2023;152:106264. https://doi.org/10.1016/j.compbiomed.2022.106264.

[6] Obermeyer F, Jankowiak M, Barkas N, Schaffner SF, Pyle JD, Yurkovetskiy L, et al. Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. Science 2022;376:1327–32. https://doi.org/10.1126/science.abm1208.

[7] Maher MC, Bartha I, Weaver S, Di Iulio J, Ferri E, Soriaga L, et al. Predicting the mutational drivers of future SARS-CoV-2 variants of concern. Sci Transl Med 2022;14:eabk3445. https://doi.org/10.1126/scitranslmed.abk3445.

[8] Hossain, Pathan MS, AQMSU, Islam MN, Tonmoy MIQ, Rakib MI, Munim MA, 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. Inf Med Unlocked 2021;27:100798. https://doi.org/10.1016/j.imu.2021.100798.

[9] Saldivar-Espinoza B, Macip G, Garcia-Segura P, Mestres-Truyol J, Puigbò P, Cereto-Massagué A, et al. 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.

[10] Choi WJ, Park J, Seong DY, Chung DS, Hong D. A prediction of mutations in infectious viruses using artificial intelligence. Genom Inf 2024;22:15. https://doi.org/10.1186/s44342-024-00019-y.

[11] Shamsi Z, Chan M, Shukla D. TLmutation: predicting the effects of mutations using transfer learning. J Phys Chem B 2020;124:3845–54. https://doi.org/10.1021/acs.jpcb.0c00197.

[12] Klima M, Khalili Yazdi A, Li F, Chau I, Hajian T, Bolotokova A, et al. Crystal structure of SARS-CoV -2 nsp10–nsp16 in complex with small molecule inhibitors, SS148 and WZ16. Protein Sci 2022;31:e4395. https://doi.org/10.1002/pro.4395.

[13] Osipiuk J, Azizi S-A, Dvorkin S, Endres M, Jedrzejczak R, Jones KA, et al. Structure of papain-like protease from SARS-CoV-2 and its complexes with non-covalent inhibitors. Nat Commun 2021;12:743. https://doi.org/10.1038/s41467-021-21060-3.

[14] Banerjee S, Wang X, Du S, Zhu C, Jia Y, Wang Y, et al. Comprehensive role of SARS-CoV-2 spike glycoprotein in regulating host signaling pathway. J Med Virol 2022;94:4071–87. https://doi.org/10.1002/jmv.27820.

[15] Barozi V, Tastan Bishop Ö. Impact of African-Specific ACE2 Polymorphisms on Omicron BA.4/5 RBD binding and allosteric communication within the ACE2–RBD Protein Complex. Int J Mol Sci 2025;26:1367. https://doi.org/10.3390/ijms26031367.

[16] Tastan Bishop Ö, Musyoka TM, Barozi V. Allostery and missense mutations as intermittently linked promising aspects of modern computational drug discovery. J Mol Biol 2022;434:167610. https://doi.org/10.1016/j.jmb.2022.167610.

[17] Sheik Amamuddy O, Afriyie Boateng R, Barozi V, Wavinya Nyamai D, Tastan Bishop Ö. Novel dynamic residue network analysis approaches to study allosteric modulation: SARS-CoV-2 Mpro and its evolutionary mutations as a case study. Comput Struct Biotechnol J 2021;19:6431–55. https://doi.org/10.1016/j.csbj.2021.11.016.

[18] Sheik Amamuddy O, Verkhivker GM, Tastan Bishop Ö. Impact of early pandemic stage mutations on molecular dynamics of SARS-CoV-2 M^{pro}. J Chem Inf Model 2020;60:5080–102. https://doi.org/10.1021/acs.jcim.0c00634.

[19] Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, Di Costanzo L, et al. RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic Acids Res 2019;47:D464–74. https://doi.org/10.1093/nar/gky1004.

[20] Walls AC, Park Y-J, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, function, and antigenicity of the SARS-CoV-2 Spike glycoprotein. Cell 2020;181:281–292.e6. https://doi.org/10.1016/j.cell.2020.02.058.

[21] Desautels T., Zemla A., Lau E., Franco M., Faissol D. Rapid *in silico* design of antibodies targeting SARS-CoV-2 using machine learning and supercomputing 2020. https://doi.org/10.1101/2020.04.03.024885.

[22] Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data – from vision to reality. Eurosurveillance 2017;22. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494.

[23] Khare S, Gurry C, Freitas L, B Schultz M, Bach G, Diallo A, et al. GISAID's role in pandemic response. China CDC Wkly 2021;3:1049–51. https://doi.org/10.46234/ccdcw2021.255.

[24] Anandakrishnan R, Aguilar B, Onufriev AV. H+ + 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. Nucleic Acids Res 2012;40:W537–41. https://doi.org/10.1093/nar/gks375.

[25] Khalili Yazdi A, Li F, Devkota K, Perveen S, Ghiabi P, Hajian T, et al. A high-throughput radioactivity-based assay for screening SARS-CoV-2 nsp10-nsp16 complex. SLAS Discov 2021;26:757–65. https://doi.org/10.1177/24725552211008863.

[26] Li P, Merz KM. MCPB.py: a python based metal center parameter builder. J Chem Inf Model 2016;56:599–604. https://doi.org/10.1021/acs.jcim.5b00674.

[27] Case DA, Aktulga HM, Belfon K, Cerutti DS, Cisneros GA, Cruzeiro VWD, et al. AmberTools. J Chem Inf Model 2023;63:6183–91. https://doi.org/10.1021/acs.jcim.3c01153.

[28] Case DA, Aktulga HM, Belfon K, Ben-Shalom IY, Berryman JT, Brozell SR, et al. Amber 2024. San Francisco: University of California; 2024.

[29] GaussView, Version 9, Roy Dennington, Todd A. Keith, and John M. Millam, Semichem Inc., Shawnee Mission, KS, 2016 2016.

[30] Sousa Da Silva AW, Vranken WF. ACPYPE - AnteChamber PYthon Parser interfacE. BMC Res Notes 2012;5:367. https://doi.org/10.1186/1756-0500-5-367.

[31] Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS: A linear constraint solver for molecular simulations. J Comput Chem 1997;18:1463–72. https://doi.org/10.1002/(SICI)1096-987X(199709)18:12< 1463::AID-JCC4> 3.0.CO;2-H.

[32] Sheik Amamuddy O, Glenister M, Tshabalala T, Tastan Bishop Ö. MDM-TASK-web: MD-TASK and MODE-TASK web server for analyzing protein dynamics. Comput Struct Biotechnol J 2021;19:5059–71. https://doi.org/10.1016/j.csbj.2021.08.043.

[33] Brown DK, Penkler DL, Sheik Amamuddy O, Ross C, Atilgan AR, Atilgan C, et al. MD-TASK: a software suite for analyzing molecular dynamics trajectories. Bioinformatics 2017;33:2768–71. https://doi.org/10.1093/bioinformatics/btx349.

[34] Python Release Python 3.10.0. In: Python.org. n.d.

[35] Kapoor A., Gulli A., Pal S., Chollet F. (2022) Deep Learning with TensorFlow and Keras: Build and deploy supervised, unsupervised, deep, and reinforcement learning models. Packt Publishing Ltd n.d.

[36] Abadi M., Agarwal A., Barham P., et al (2016) TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems n.d.

[37] Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Scikit-learn: Machine Learning in Python 2012. https://doi.org/10.48550/ARXIV.1201.0490.2012.

[38] Páll S, Zhmurov A, Bauer P, Abraham M, Lundborg M, Gray A, et al. Heterogeneous parallelization and acceleration of molecular dynamics simulations in GROMACS. J Chem Phys 2020;153:134110. https://doi.org/10.1063/5.0018516.
[39] Joblib Development Team. Joblib: running Python functions as pipeline jobs 2020.
[40] Lundberg S., Lee S.-I. 2017. A Unified Approach to Interpreting Model Predictions 2017. https://doi.org/10.48550/ARXIV.1705.07874..