Computational and Structural Biotechnology Journal 23 (2024) 3800–3816

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

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

Research Article

Revealing SARS-CoV-2 M^{pro} mutation cold and hot spots: Dynamic residue network analysis meets machine learning

Victor Barozi [a,1], Shrestha Chakraborty [b,1], Shaylyn Govender [a,1], Emily Morgan [a,1], Rabelani Ramahala [a,1], Stephen C. Graham [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] Division of Virology, Department of Pathology, University of Cambridge, Cambridge CB2 1QP, UK
[c] Department of Pure and Applied Mathematics, Rhodes University, Makhanda 6139, South Africa
[d] National Institute for Theoretical and Computational Sciences (NITheCS), South Africa

ARTICLE INFO

Keywords:
Drug design
Drug resistance
Pathogen evolution
Network analysis
Artificial neural networks
Decision tree
Chymotrypsin-like protease
3CLpro
Nsp5
M^{pro}
SARS-CoV
COVID-19

ABSTRACT

Deciphering the effect of evolutionary mutations of viruses and predicting future mutations is crucial for designing long-lasting and effective drugs. While understanding the impact of current mutations on protein drug targets is feasible, predicting future mutations due to natural evolution of viruses and environmental pressures remains challenging. Here, we leveraged existing mutation data during the evolution of the SARS-CoV-2 protein drug target main protease (M^{pro}) to test the predictive power of dynamic residue network (DRN) analysis in identifying mutation cold and hot spots. We conducted molecular dynamics simulations on the M^{pro} of SARS-CoV-2 (Wuhan strain) and calculated eight DRN metrics (*averaged BC*, *CC*, *DC*, *EC*, *ECC*, *KC*, *L*, *PR*), each of which identifies a unique network feature within the protein. The sets of residues with the highest and lowest values for each metric, comprising potential cold and hot spots, were compared to published biochemical analyses and per residue mutation frequencies observed across five SARS-CoV-2 lineages, encompassing a total of 191,878 sequences. Individual DRN metrics displayed only modest power to predict the mutation frequency of individual residues. However, integrating the eight DRN metrics with additional structural and sequence-derived metrics allowed us to develop machine learning models which significantly improved the prediction of residue mutation frequency. While further refinements should enhance accuracy, we demonstrated a robust method to understand pathogen evolution. This approach can also guide the development of long-lasting drugs by targeting functional residues located in and near active site, and allosteric sites, that are less prone to mutations.

1. Introduction

Understanding the effects of known mutations and predicting the evolution of future mutations, along with their potential impacts on the protein drug targets of viruses, is crucial in designing effective and long-lasting antivirals [1]. This also applies to the design of new immunotherapies or vaccines that target conserved regions of viral proteins that are less likely to mutate [2–4]. However, this is a highly challenging task due to the complexity of proteins and the nature of virus evolution. Selective pressure for effective replication in a given host can trigger systematic changes in viral protein structure and dynamics, favouring specific patterns of amino acid substitution and leading to the emergence of new variants, as we observed with SARS-CoV-2 [5–7]. Distinct mutations can yield proteins with similar physical properties, which might not be evident from the sequence alone, but for variants to retain fitness, the critical properties of each protein must be conserved [8,9]. Mutations may impact the overall dynamics and structure of a protein [10–12], and modifications at or near the active site of enzymes are especially important since they can modify both catalysis and protein-substrate interactions [13,14]. For example, mutations that strengthen side chain interactions within the active site may limit the dynamics of catalytic residues and inhibit function [12]. Structural dynamics is essential for enzyme catalysis [15], and several studies have delineated long-range effects of residue substitution, implying that

* Corresponding author.
** Corresponding authors at: National Institute for Theoretical and Computational Sciences (NITheCS), South Africa.
E-mail addresses: scg34@cam.ac.uk (S.C. Graham), n.bishop@ru.ac.za (N.T. Bishop), o.tastanbishop@ru.ac.za (Ö. Tastan Bishop).
[1] Equally contributed first authorship (in alphabetic order).

https://doi.org/10.1016/j.csbj.2024.10.031
Received 15 August 2024; Received in revised form 19 October 2024; Accepted 19 October 2024
Available online 22 October 2024

catalytically efficient alterations might transpire across the entire protein [16–19].

Recent studies have explored the impact of known missense mutations and predicted future changes on proteins from various computational perspectives, ranging from graph theory to machine learning (ML). For example, Miotto et al. [20] applied graph theory through the representation of proteins as network graphs, to identify thermostable proteins by using energetically weighted links. Prabantu et al. [21] used network representation of proteins to study the changes in residue contacts (networks) as a consequence of mutations. Additionally, mutation effect prediction tools such as mcSM-PPI2 [22] use graph-based structural signatures to model variations in inter-residue interaction networks. DynaMut, a protein stability prediction tool, integrates graph-based signatures with normal mode dynamics to model mutational effects on protein stability [23]. Similar tools, like PolyPhen-2 [24] and SIFT [25], also focus on assessing the impact of existing mutations on protein stability. In contrast, FoldX [26] predicts the structural and energetic impact of mutations on proteins, estimating the effects based on protein stability rather than mutation likelihood. Alternatively, a number of studies have applied ML based algorithms to predict mutation effects using methods such as support vector machines (SVM) [27–29], neural networks [30,31], and multiple regression and classification techniques [32]. Additionally, some computational approaches for predicting the likelihood and functional effects of mutations mainly rely on human genetic data. Combined Annotation-Dependent Depletion (CADD) [33] and MutPred [34], for instance, use machine learning models trained on human variants and evolutionary conservation to infer mutation likelihood. Similarly, MutSigCV [35] and SigProfilerAssignment [36] analyse background mutation rates and mutational signatures from human cancer genomes to identify regions more prone to mutations, particularly in cancer. OncoKB [37] emphasizes protein function and conservation, utilizing human cancer data to identify likely driver mutations.

The integration of graph-based methods, which provide a robust framework for modelling complex network interactions, with the predictive power of efficient ML algorithms could dramatically enhance our understanding of the intricate relationship between protein sequence, structure, and biological function [38,39]. This study builds upon our previous research using dynamic residue network (DRN) analysis to decipher the effects of missense mutations [40–43]. We previously proposed that the different DRN metrics provide distinct, complementary information on protein dynamics, and that holistic analysis of these DRN metrics could be used to identify the regions prone to changes (hot spots) or resilient to mutations (cold spots) within proteins [42,43]. Our goal in this study is to evaluate the mutation prediction capability of DRN, both alone and in combination with ML approaches. For this purpose, we used the SARS-CoV-2 main protease (M^{pro}) as a case study, primarily due to the availability of a vast amount of evolutionary mutation data since the initial SARS-CoV-2 strain was identified in Wuhan [44,45]. Additionally, M^{pro} is one of the key SARS-CoV-2 antiviral drug targets [46–48], and its constant evolution due to mutations under selective pressure could impair the future efficacy of drugs that target this protease. A systematic and comprehensive computational analysis to monitor existing and emergent mutations in M^{pro} could predict both fast-evolving regions and residues that are unlikely to mutate, thereby highlighting regions of the protein that could be targeted by therapeutics where there would be a high intrinsic barrier to the evolution of drug resistance.

Our start point was M^{pro} protein from the SARS-CoV-2 virus (Wuhan strain). We performed six molecular dynamics simulations (MD) and calculated eight DRN metrics for each simulation (*averaged BC*, *CC*, *DC*, *EC*, *ECC*, *L*, *PR*, *KC*), the values of which were then averaged across the six simulations. We evaluated the predictive power of these averaged network analysis metrics to identify residues likely to have low or high tolerance of mutations (cold spots and hot spots, respectively). These predictions were compared to per-residue mutation rates obtained from population-scale SARS-CoV-2 sequencing and in vitro investigations of M^{pro} activity. The integration of these eight DRN metrics with additional protein features within an ML framework significantly improved our ability to predict whether a residue would mutate. DRN analysis combined with ML, thus, offers a novel and effective technique for predicting future protein mutations. The method proposed here holds significant potential for drug discovery against evolving pathogens, facilitating the design of inhibitors that target residues less prone to mutation and thus with a higher barrier to the evolution of drug resistance.

2. Methods

2.1. Data retrieval

The complete human SARS-CoV-2 M^{pro} variant of concern (VOC) sequences (Alpha, Beta, Delta, Gamma and Omicron) of high coverage and with patient status were retrieved from the Global Initiative on Sharing Avian Influenza Data (GISAID; https://gisaid.org/) [49] for dates between 1st December 2019 and 24th February 2024. Retrieved sequences were submitted to the GISAID CoVsurver tool [50] (https://gisaid.org/database-features/covsurver-mutations-app) for mutation identification: deletion, insertion, and single nucleotide variations (SNV). SNVs specific to the M^{pro} protein were then filtered from the CoVsurver output (*.tsv* files) using an in-house Python script which also computed the mutation frequency of unique SNVs in each VOC dataset, highlighting the most prevalent mutations.

Function scores for M^{pro} activity were obtained from published deep scanning mutagenesis data [51] by taking an average of the functional scores obtained from FRET, transcription factor inactivation and growth assays for each residue (excluding stop mutations). Scores were sorted to obtain the set of residues that are least mutation tolerant (functional scores < 0.3) or most mutation tolerant (functional scores ≈ 1).

2.2. Molecular dynamic simulations

The dimeric form of the M^{pro} reference structure (from Wuhan strain) was prepared in our previous study [43] using the crystal structure (PDB ID: 5RFV [52]). The structure was then protonated at a pH of 7.0 using the PROPKA tool from PDB2QR [53] before the MD simulations.

GROMACS software v2021.1 [54] was employed for the system simulations using Amber03 force fields [55]. The systems were simulated in a triclinic box of 1.5 nm clearance and filled with single point charge 216 (SPC216) water model [56]. The charge in the box was neutralized using NaCl ions at 0.15 M concentration. Prior to equilibration, systems were minimized using the steepest descent algorithm until a minimum energy of 1000.0 kJ/mol/nm was achieved within a maximum of 50000 steps. The NVT (constant number of particles, volume, and temperature) temperature equilibration ensemble was achieved using V-rescale at 310 K for 0.1 ns, whereas the NPT (constant number of particles, pressure and temperature) pressure ensemble was achieved using the C-rescale at 1 atm and 310 K, both for 0.1 ns. Production simulations of 100 ns with a time step of 2 fs followed, where hydrogen bonds were constrained using the LINCS algorithm [57]. For the long-range electrostatics calculations, Particle Mesh Ewald (PME) electrostatics were used with a Fourier spacing of 0.16 nm. For the short-range Coulomb and van der Waals interactions, a cut-off distance of 1.1 nm was used. Six independent 100 ns MD simulations were performed.

Following production simulations, the periodic boundary conditions (PBC) were removed for each simulation, and the trajectories analysed using the GROMACs *gmx rms*, *gmx rmsf*, and *gmx gyrate* functions for root mean square deviation (RMSD), root mean square fluctuation (RMSF) and radius of gyration (Rg). RMSD, RMSF and Rg analyses showed the MD simulation to be reproducible. Additionally, each simulation achieved equilibration within 100 ns. The RMSD line and violin plots (Fig. S1A and B) displayed closely clustered mean squared deviations

and unimodal distributions, respectively. Furthermore, local residue fluctuation remained consistent across all simulations, and the systems maintained a consistent degree of compaction throughout the simulations (Fig. S1C and D).

2.3. Dynamic residue network analysis

DRN capitalizes on the principals of network analysis in graph theory to compute the average network properties per residue within a protein structure across an MD trajectory [40,41]. In a DRN, protein residues (C_α and C_β atoms) are the nodes/vertices and connections between nodes within a Euclidian distance of $\leq$ 6.7 Å are treated as an edge [40,58]. This distance is optimal as previously shown [58]. However, we also demonstrated that adjusting the distance between the nodes may slightly alter the metric values without impacting the overall data trend [59].

With the graph representation of proteins, various network metric values can be computed to address different aspects of the protein residue networks [42,43]. Here, we calculated eight metric values per MD run, namely *averaged 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 metrics were computed for each frame of the MD trajectory and then averaged across the entire trajectory to obtain a single *averaged* metric value per residue, using the formulas shown in Table 1. This process was repeated for all six MD simulations, and the average of these six simulations was used for further analysis.

Each DRN metric provides specific insights into the protein network characteristics. *Averaged BC*, for instance, measures the extent to which a node/residue lies on the shortest paths between other nodes within the network. Residues with high *BC* values are considered important in controlling the information flow (gatekeepers) in a network [40,41]. *Averaged CC* measures the closeness of a given residue to all other residues in the network. Interface and protein core residues are associated with high *CC* values [42,43]. Residues with high *CC* are considered good disseminators of information. *Averaged DC* describes the local connectivity of a node/residue in network through measuring the number of incident edges to the node in question, whereas *averaged EC* describes the influence of a residue in the network based on the centrality of both the high and low scoring neighbors [60]. Residues with high *DC* have values have a high local influence. *Averaged ECC* is the measure of the longest path from a node to any other node in a network [40,41]. Much like *averaged EC*, *averaged KC* assesses the relative impact of a node within a network by iteratively gauging node centrality through the centrality of its neighboring nodes. *KC* assigns centrality via adjacency damping coefficient α and a basal adjacency β of a node's immediate connectivity. *L* measures the shortest paths of a node from all the other nodes in a network. *L* is measured by the sum of the geodesic distance (d_i) to every other node (u) divided by the number of nodes less by one [40,59] (Table 1). In essence, signals from residues with high *L* values travel longer distance to reach other residues. *Averaged PR* is an adjusted version of *KC* that iteratively assigns importance/influence on a node based on of its connected neighboring nodes. The centrality of each neighbor is normalized by its own degree, *D*, for each iteration. As in *KC*, *PR* also includes a damping factor α and a constant β [40,41] (Table 1).

The DRN calculations were performed using the *calc_network.py* tool (https://github.com/RUBi-ZA/MD-TASK/tree/mdm-task-web/src) from the MDM-TASK suite [40,41]. This tool computed the eight DRN metric values for each of the six wild type (WT) systems. Complete 100 ns MD trajectories (*xtc*, periodicity completed) and the WT topology file (*pdb*) were used as input for network calculations, with a step size of 5 and a Euclidian threshold of 6.7 Å. Thus, we calculated DRN for each MD trajectory, in which residue interaction networks are constructed for every 5th frame of the trajectory using 10 ps time intervals. We also repeated the DRN calculations with a step size 10 to check if this introduces significant changes. We assessed the DRN values per metric per residue using a Mann-Whitney U test. The p-value for each *averaged* metric was *BC*: 0.974, *CC*: 0.919; *DC*: 0.995; *EC*: 0.830; *ECC*: 0.934; *KC*: 0.974; *L*: 0.909; *PR*: 0.976. As the results are highly similar, we proceeded using the step size 5 results.

The *averaged* DRN metric results for each residue were then averaged over the six WT simulations to mitigate variability introduced by randomness of MD simulation seeds and initial velocities.

Density plots were calculated for each DRN metric per chain, and chain A and B were compared for their similarity (Fig. S2 A–H). Additionally, the Mann-Whitney U test [61], a non-parametric test used to determine whether there is a significant difference between two independent distributions, was utilized to compared the distributions of each metric between Chain A and Chain B (Fig. S2 A–H). The p-values for each are all well above the typical significance threshold ($p < 0.05$). This suggests that there is no statistically significant difference between the distributions of Chain A and Chain B across all centrality metrics. These p-values strongly support the hypothesis that the distributions of these metrics for both chains are similar. Thus, only the DRN values for chain A were used for subsequent analysis.

2.4. Machine learning (ML)

Several ML predictors were built independently via Artificial Neural Network (ANN) and Random Forest (RF) algorithms using both regression and classification models, so enabling cross-checking between the models. In each case, a number of models were built and the models with the best performance (as described later) are presented in Section 3.4. This is common practice because the error regarded as a function of the internal model parameters has many local minima, and a training run may converge to a value well above the global minimum (e.g., see [62]).

Table 1
Formulas for the DRN centrality metrics (adopted from [40,41]).

Centrality metric	Formula	Note
Averaged BC	$\overline{BC}(v) = \frac{1}{m}\sum_{i=1}^{m}\sum_{s,t\in V}\frac{\sigma(s_i,t_i\|v_i)}{\sigma(s_i,t_i)}$	V is the complete set of nodes; m is the number of frames; $\sigma(s,t)$ is the number of shortest paths connecting nodes s and t; $\sigma(s,t\|v)$ is the number of these paths passing another node v; and i is the frame number.
Averaged CC	$\overline{CC}(v) = \frac{n-1}{m}\sum_{i=1}^{m}\sum_{u=1}^{n-1}\frac{1}{d_i(v,u)}$	$d(v,u)$ is the shortest-path distance between v and u, and n is the number of nodes in the graph.
Averaged DC	$\overline{DC}(i) = \frac{1}{m(n-1)}\sum_{k=1}^{m}\sum_{j=1,j\neq i}^{n}A_{ijk}$	n is the number of nodes; A_{ijk} is the jk^{th} adjacency for the i^{th} frame.
Averaged EC	$A\cdot\overrightarrow{EC} = \lambda\cdot\overrightarrow{EC}$ (a) $\overline{EC}(i) = \frac{1}{m}\sum_{k=1}^{m}EC_{ik}$ (b)	(a) *EC* is the eigenvector, and lambda is the eigenvalue for the eigen decomposition of adjacency matrix A. In NetworkX, this is obtained by power iteration. (b) Averaged *EC* is computed for i^{th} residue by computing the vector for each MD frame and averaging.
Average ECC	$\overline{ECC}(j) = \frac{1}{m}\sum_{i=1}^{m}\max_{i,k\ \epsilon V}d_i(j,k)$	*ECC* is the measure of the longest path from a node to any other node in a network
Averaged KC	$KC(i) = \alpha\sum_{j=1}^{n}A_{ij}KC_j+\beta$ (a) $\overline{KC}(i) = \frac{1}{m}\sum_{k=1}^{m}KC_{ik}$ (b)	*KC* is a modification of *EC* that employs a dampening coefficient and a constant in order to influence adjacency values.
Averaged L	$L(v) = \frac{1}{m(n-1)}\sum_{i=1}^{m}\sum_{u=1}^{n-1}d_i$	v is a node at a time i; d_i is the geodesic distance to every node u.
Averaged PR	$PR(i) = \alpha\sum_{j=1}^{n}\frac{A_{ij}}{D_j}PR_j+\beta$ (a) $\overline{PR}(i) = \frac{1}{m}\sum_{k=1}^{m}PR_{ik}$ (b)	*PR* is an adjusted version of *KC* where centrality is assigned based on that of the neighbors while still applying the dampening factor and a constant.

The software packages utilized were MATLAB [63] (version 2023b) Deep Learning Toolbox, and Python version 3.10.13 [64] together with Keras version 3.0.5 [65] (https://github.com/fchollet/keras), TensorFlow version 2.15.0 [66] and the Scikit-learn library version 0.24.2 [67]. In all cases we used data for each of the 304 residues in chain A of the M^{pro} structure. For each residue, the predictor data comprised the eight DRN metric values for *averaged BC*, *CC*, *DC*, *EC*, *ECC*, *KC*, *L* and *PR*, the BLOSUM62 matrix values of the M^{pro} structure (20 values), and RMSF. SASA and atomic displacements parameters ("B-factors") were calculated using GROMACS v2021.1 and included in the prediction data. Specifically, the B-factor for each residue was determined using the *rmsf* command with the *-oq* option, selecting the C_α atoms. The SASA was computed using the *gmx sasa* command, also selecting the C_α atoms to ensure calculations were performed on a residue basis. Each predictor data was averaged over six MD simulations. Consequently, the predictor data was a matrix of size 304 × 32. The target data was, for each residue, calculated from the mutation frequencies, i.e., the number of sequences that contained a mutation at that residue, and was obtained as described in Section 2.1. The set of data points was randomly divided into 3 subsets, training (70 %), validation (15 %) and testing (15 %). The validation set was used during model training to fine tune the hyperparameters and prevent overfitting, while the testing set was used after training was completed to provide an unbiased evaluation of the model's performance on unseen data.

For the **regression models,** since some of the mutation frequencies are very large (up to 67,139), we found that the accuracy of the predictors was improved when using a target of $\log_{10}$ (1 + mutation frequency) rather than the raw mutation frequency. Model performance was evaluated by comparing the predicted and target values and then producing a scatterplot and calculating the correlation or regression "R" value and its associated "p-value'. The R value indicates the extent to which the two datasets are linearly related: it has a value of 1 for a perfect predictor and if the predictions are made using a random number generator, then R will be close to 0. (Note that R can also be 0 for certain nonlinear relationships between the two datasets). The p-value is the probability of the null hypothesis that the predicted and target values are not linearly related, and normally the R value is regarded as statistically significant provided $p < 0.05$. Further details are given in many texts on statistics (e.g. Sec. 12.4 of [68]).

For the **classification models**, the target was adjusted to be 1 for mutation frequencies larger than the cut-off value (set at 20), and 0 for frequencies less than or equal to the cut-off value. The performance of the model was evaluated by calculating the confusion matrix, as well as the receiver operating characteristic (ROC) curve and the area under the curve (AUC), which is 1 when the predictor is perfect and 0.5 when it is equivalent to using random numbers. Note that the ROC curve (and thus the AUC) is a plot in which the cut-off value is varied, and thus it measures the performance of the ML algorithm; the confusion matrix shows the accuracy of the classifier for the specific value used for the cut-off.

2.4.1. Artificial neural networks

The first step in the construction of an ANN is to determine the network structure, and specifically the number of nodes in the hidden layer. As the number of nodes is increased the accuracy of the model increases, but at some stage the problem of overfitting will occur with the model losing generalizability and the accuracy of the model on the independent testing set will deteriorate when compared to the accuracy on the training set. After some trial and error, it was found that optimal performance occurred with 10 nodes, which is often used as the default number.

Regression models were constructed using a default network architecture comprising of a hidden layer of 10 nodes using the sigmoid activation function, while the linear activation function was chosen for the single node in the output layer. Training the ANN involved adjusting the weights and biases in the network to minimize the mean squared error (MSE); built-in algorithms; namely Levenberg-Marquardt in MATLAB and Adams optimization in Keras were used for this purpose.

The regression ANN models were re-purposed into **classification models** by mapping the predicted mutation frequency to 1 if it was greater than the cut-off value, and to 0 if it was smaller or equal to the cutoff.

2.4.2. Random Forest

Random Forest is a supervised ensemble ML method that utilizes a combination of tree predictors, or decision trees, to solve both classification and regression problems. Under RF, the model constructs a forest of decision trees and aggregates their predictions to yield a more robust and accurate ensemble model [69]. RF models take labelled data (predictor and targets) to develop, train and test for model efficiency and accuracy.

RF in **MATLAB** is performed using TreeBagger [70], a class specifically designed for building ensembles of decision trees. The number of trees in the forest was set to 100, and otherwise default parameters were used. The **regression** and **classification** models were constructed independently.

Using the **Python Scikit-learn** tool [67], model parameters were optimized using the *RandomizedSearchCV* module in Scikit-learn and the model run on both the class balanced and imbalanced dataset. This was done to address the effect of imbalanced data classes in the data. The **classification model** was initiated using the Scikit-learn *RandomForestClassifier* with a random seed 42, 100 estimators, a minimum sample split of 2, a sample leaf of 2 and $\log_2$ for maximum features with no maximum depth. The **regression model** was run with a minimum sample split of 13, minimum sample leaf of 1, 100 estimators, no maximum depth, and a random state of 1 as identified by the randomized parameter search.

3. Results and discussion

3.1. Uncovering cold spots: dynamic residue network analysis supported by literature insights

Previously, we defined the protein cold spots as regions where mutations are not tolerated and are thus rare or not observed in the population [42,43]. Identifying residues in and near the active site that are less prone to mutations is crucial for designing drugs with high resistance barriers. We proposed that these residues can be identified by DRN analysis, where the centrality hubs (residues) with the highest metric values, representing key roles within the network, would be the cold spots. Furthermore, we showed how combining information from each metrics brings deeper insights, as each identifies a unique network feature within the protein [18,42,43]. For each metric, we identified 15 residues (i.e., 5 %) with the highest metric values. The selection of 5 % reflects a common practice [18,42,43] ensuring equal set sizes across the different metrics. However, alternative selection strategies, such as using two standard deviations from the mean, can be applied if the distribution is Gaussian. In our case, most metrics did not follow a Gaussian distribution (see Fig. S2 A–H). Further, please note that cold spots for *averaged L* (as defined in Table 1) are residues with the lowest, rather than highest, metric values.

The residues identified by one or more of the metrics are presented in Table 2, as are the prevalence of mutations at these residues across the analysed SARS-CoV-2 sequences. Among these, only one residue, **L115**, was consistently identified by all eight metrics, and we call it a *persistent hub* [42,43]. Additionally, **S10** was highlighted by six metrics, while **A7**, **S113** and **V125** were recognized by five metrics (Table 2). Notably, 25 residues, including **S10, S113** and **L115**, had no mutations across five lineages. The remaining residues showed mutation frequencies ranging from 5.2×10^{-6} (**F3**, **M6**, **C117** and **A206**) to 6.5×10^{-4} (**M17**) across all SARS-CoV-2 sequences. Furthermore, in our previous study [43], we identified communication paths between an allosteric pocket and the

Table 2
Cold spots as predicted by DRN metrics – residues with the highest metric values (top 5 %). The second and third columns list the residue mutation(s) and their locations within the M^{pro} structure. The tick indicates that the residue falls within the top 5 % of the DRN metric (for *ECC* and *L* it is the residues with the lowest metric values). The table also provides mutation rates for each residue across different lineages (Alpha, Beta, Gamma, Delta and Omicron) and the total mutation rate across all lineages. The numbers given in each lineage column-heading specify the number of unique M^{pro} sequences per that lineage, and the "Total" is for all the sequences analysed in this study. Light grey marks the residues with no mutations.

Residue	Mutations	Location	*BC*	*CC*	*DC*	*EC*	*ECC*	*KC*	*L*	*PR*	Alpha (35302)	Beta (4317)	Gamma (8268)	Delta (121477)	Omicron (22514)	Total (191 878)
F3	F3V	N-Finger/Dimer interface	√											**1** ($8.2x10^{-6}$)		**1** ($5.2x10^{-6}$)
R4	R4K	Dimer interface		√					√					**13** ($1.1x10^{-4}$)		**13** ($6.8x10^{-5}$)
K5		N-Finger/ Dimer interface		√			√		√							
M6	M6I	N-Finger/ Dimer interface		√			√		√					**1** ($8.2x10^{-6}$)		**1** ($5.2x10^{-6}$)
A7	A7T, A7V	N-Finger/ Dimer interface		√		√	√	√	√		**3** ($8.5x10^{-5}$)			**43** ($3.5x10^{-4}$)	**1** ($4.4x10^{-5}$)	**47** ($2.4x10^{-4}$)
F8	F8L	N-Finger/ Dimer interface					√							**14** ($1.2x10^{-4}$)		**14** ($7.3x10^{-5}$)
P9		N-Finger/ Dimer interface		√		√	√		√							
S10		Dimer interface		√	√	√	√	√	√							
G11		Dimer interface	√			√										
V13	V13I	Dimer interface				√					**1** ($2.8x10^{-5}$)		**1** ($1.2x10^{-4}$)			**2** ($1.0x10^{-5}$)
E14		Dimer interface	√			√										
M17	M17V, M17I, M17T	Domain 1	√								**16** ($4.5x10^{-4}$)	**2** ($4.6x10^{-4}$)	**39** ($4.7x\ 10^{-3}$)	**64** ($5.3x10^{-4}$)	**4** ($1.8x10^{-4}$)	**125** ($6.5x10^{-4}$)
V18	V18L, V18A, V18I	Domain 1	√		√						**2** ($5.7x10^{-5}$)			**72** ($5.9x10^{-4}$)		**74** ($3.9x10^{-4}$)
G29	G29C	Domain 1						√						**9** ($7.4x10^{-5}$)		**9** ($4.7x10^{-5}$)
V36	V36I	Domain 1			√			√		√				**8** ($6.6.x10^{-5}$)		**8** ($4.2x10^{-5}$)
C38	C38G	Domain 1						√			**3** ($8.5x10^{-5}$)				**2** ($8.9x10^{-5}$)	**5** ($2.6x10^{-5}$)
P39		Domain 1	√		√			√								
L57		Domain 1								√						
V86	V86I, V86L, V86A	Domain 1			√						**3** ($8.5x10^{-5}$)		**3** ($3.6\ x\ 10^{-4}$)	**19** ($1.6\ x\ 10^{-4}$)	**5** ($2.2x10^{-4}$)	**30** ($1.6x10^{-4}$)
L89	L89I, L89F	Domain 1								√	**15** ($4.2x10^{-4}$)	**1** ($2.3x10^{-4}$)	**17** ($2.1x10^{-3}$)	**24** ($2.0x10^{-4}$)	**12** ($5.3x10^{-4}$)	**69** ($3.6x10^{-4}$)
V91	V91I	Domain 1			√					√				**2** ($1.6x10^{-5}$)		**2** ($1.0x10^{-5}$)
T111		Domain 1	√													
F112		Domain 1	√													
S113		Domain 2		√		√	√	√	√							
V114	V114L	Domain 2		√					√					**5** ($4.1x10^{-5}$)	**1** ($4.2x10^{-5}$)	**6** ($3.1x10^{-5}$)
L115		Domain 2	√	√	√	√	√	√	√	√						
A116	A116V, A116T, A116S	Domain 2		√					√		**10** ($2.8x10^{-4}$)	**1** ($2.3x10^{-4}$)	**2** ($2.4\ x\ 10^{-4}$)	**53** ($4.4x10^{-4}$)	**5** ($2.2x10^{-4}$)	**71** ($3.7x10^{-4}$)
C117	C117S	Dimer interface			√	√		√							**1** ($4.2x10^{-5}$)	**1** ($5.2x10^{-6}$)
P122	P122S	Dimer interface				√	√				**2** ($5.7x10^{-5}$)			**9** ($7.4x10^{-5}$)		**11** ($5.7x10^{-5}$)
G124		Domain 2		√		√	√		√							
V125	V125I	Domain 2		√		√	√	√	√					**7** ($5.8x10^{-5}$)		**7** ($3.6x10^{-5}$)
Y126		Dimer interface		√			√		√							
Q127		Dimer interface		√			√		√							
C128	C128S	Domain 2	√	√			√		√					**2** ($1.6x10^{-5}$)	**1** ($4.2x10^{-5}$)	**3** ($1.6x10^{-5}$)
S139		Dimer interface	√													
F140		Active site	√													
G146		Active site	√		√			√								
S147		Dimer interface	√			√		√								
V148	V148I, V148A	Domain 2	√					√			**1** ($2.8x10^{-5}$)		**2** ($2.4\ x\ 10^{-4}$)	**2** ($1.7x10^{-5}$)		**5** ($2.6x10^{-5}$)
G149		Domain 2				√										
F150		Domain 2			√	√		√		√						
V157	V157I, V157L	Domain 2								√	**4** ($1.1x10^{-4}$)			**73** ($6.0x10^{-4}$)	**4** ($1.8x10^{-4}$)	**81** ($4.2x10^{-4}$)
L167		Domain 2								√						
N203		Domain 3			√					√						
A206	A206S	Domain 3			√			√		√				**1** ($8.2x10^{-6}$)		**1** ($5.2x10^{-6}$)
F230		Domain 3								√						
L250		Domain 3			√					√						
L253	L253F	Domain 3			√					√				**3** ($2.5x10^{-6}$)		**3** ($1.6x10^{-5}$)
A266	A266S, A266V	Domain 3								√	4 ($1.1x10^{-4}$)	1 ($2.3x10^{-4}$)		**21** ($1.7x10^{-4}$)		**26** ($1.4x10^{-4}$)
S267	S267L	Domain 3								√				**2** ($1.6x10^{-5}$)		**2** ($1.0x10^{-5}$)
D295		Domain 3					√									
V296	V296I	Domain 3			√									**108** ($8.9\ x10^{-4}$)		**108** ($5.6x10^{-4}$)

active site in the presence of potential allosteric modulators. These pathways traversed the interface residues of domain I and II and were identified by combining averaged *EC* hubs. The involved residues include **7**, **9-11**, **13**, **14**, **17**, **28**, **29**, **38**, **113**, **115-117**, **122**, **124**, **125**, **146-150**, many of which are also listed in Table 2. Please note that the residues identified as cold or hot spots in this study by DRN metrics are shown in **bold**, and residues not identified as hot or cold spots are **bold and underlined**.

Identifying residues in and near the active site that are less prone to mutations is crucial for designing drugs with high resistance barriers. Visual inspection of the top 5 % metric hubs for each of the four the most commonly used metrics (*BC*, *CC*, *DC*, *EC*) mapped onto the M^{pro} structure revealed that these hubs are located in key functional and core regions of the protein, including in the vicinity of the active site, the dimerization region and a previously identified allosteric pocket [10] (Fig. 1**A–D**). In the rest of this section, we report the evidence from literature on the function of these residues and consequences of their mutation for protein stability and catalysis.

N-finger residue **R4** promotes dimerization SARS-CoV-1 M^{pro}, with mutation of this residue destabilizing the active dimer conformation and reducing enzymatic activity to a variable extent. Chen et al. demonstrated that **R4A** substitution of a SARS-CoV-1 M^{pro} resulted in a 20 % decrease in dimerization and the enzyme retained only 10 % of its activity [71,72]. Conversely, another study found that mutating **R4** in SARS-CoV-1 M^{pro} did not affect the activity of the enzyme but reduced the dimerization efficacy by 80 % relative to the wild type [73]. Mutational studies combined with computational analysis on SARS-CoV-2 identified **R4** as a minor player in dimerization, as evidenced by mass photometry measurements, crystallographic structure determination and MD data [74]. Another study demonstrated that substituting this residue likely disrupts the salt-bridge between **R4** and **E290** of domain III, resulting in decreased catalytic efficiency while maintaining an inactive dimeric state [75].

Another crucial residue, **P9**, interacts with residues **P122** and **S123** of the adjacent protomer. Residues **S10**, **K12**, and **E14** are in a one-turn α-helix at the end of the N-finger, which forms a contact point at the dimer interface [75]. **S10** is involved in a hydrogen bond, where the backbone NH of **S10** may interact with the hydroxyl group of **S10** on the opposite protomer. Disruption of this bond may trigger reduced catalytic efficiency [75]. **G11** is highly conserved in both SARS-CoV-2 and SARS-CoV-1, and it interacts with the side chain of **E14** on the opposite monomer, forming a hydrogen bond. A Gly-to-Ala mutation at this position in SARS-CoV-1 completely disrupted dimerization and eliminated catalytic activity [76]. Similarly, the **E14A** substitution in SARS-CoV-1 resulted in a 50 % reduction in dimerization and only residual (4 %) enzymatic activity [71]. A recent mutational study of SARS-CoV-2 has also demonstrated importance of this residue, as substituting of **E14** with A/D/S/Q impaired catalytic activity, suggesting that the distance between the N-fingers of the monomers and the polarity of the side chain of **E14** are important for catalytic activity [75].

An extensive mutational study involving deep sequence scanning and a high-throughput fluorescent reporter assay in yeast demonstrated that N-finger residues **P9**, **S10**, **G11**, **E14** and subsequent domain I, II and III residues **T111**, **S113**, **G146**, **S147**, **G149**, **F150**, and **N203** exhibited low mutation tolerance, with substitution resulting in null-like function [51]. Another report indicated that the side chains of residues **L115** and **F150** are involved in strong hydrophobic interactions with **P9** of opposite protomer, and substitution with polar residues led to reduced protease activity [77].

In the SARS-CoV-2 M^{pro}, **S139** in domain II forms critical hydrogen bonds with **Q299** in domain III of the paired monomer, at distances of 2.9 and 3.1 Å, respectively. These bonds are essential for maintaining the integrity of the oxyanion hole within the S1 subsite, which is crucial for catalysis [75,78]. The **S139A** mutation results in substantial catalytic impairment, reducing activity to 46 % of the wild type (Wuhan), and causes a two to three-fold decrease in the turnover number, while still preserving dimer formation, as observed through analytical size exclusion chromatography (SEC) experiments [75]. Beyond assessing direct contributions of SARS-CoV-2 residues to catalysis, a recent study characterised the activity of mutant enzymes in the presence of antiviral drugs (nirmatrelvir and ensitrelvir) to identify mutations that could confer antiviral drug resistance [79]. Substitution of **S139** to P/L/Q conferred strong drug resistance, and the authors suggest that substitution of **S139** with proline triggers a conformational change in the

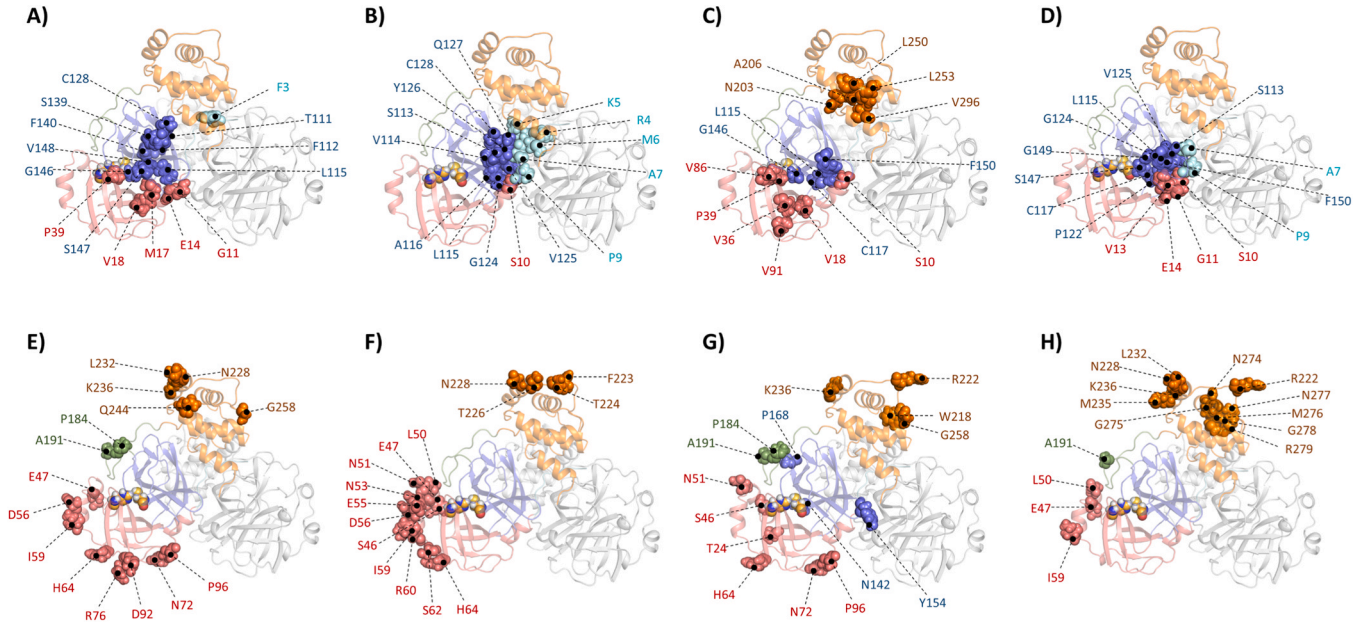

Fig. 1. Cold spot residues per DRN metric (**A-D**: *BC*, *CC*, *DC*, *EC*, respectively) and hot spot residues per metric (**E–H**: *BC*, *CC*, *DC*, *EC*, respectively). The structural domains I (residues 10–99), II (residues 100–183) and III (residues 183–197) are coloured in pink, blue and orange, respectively. The N-finger is shown in cyan, and the linker loop is shown in green. The catalytic residues (H41 and C145) are coloured based on their atom types, where the carbon atoms are coloured light grey, oxygen atoms in red, nitrogen atoms in blue, sulfur atoms in yellow. The cold spots are depicted as spheres in **A-D**, while hot spots are depicted as spheres in **E–H**. Chain B is coloured grey.

active site that specifically disrupts the interactions with nirmatrelvir [79].

Dimerization is critical for M^{pro} activity, with both residues at the dimer interface plus residues near the substrate binding surface contributing to dimer formation. Mutation of SARS-CoV-1 residues **S139** and **F140**, which belong to the oxyanion hole of the S1 substrate binding pocket, disrupts or destabilizes the dimer formation [80]. Similarly, **S147** [81] and **E166** [82] are crucial for SARS-CoV-1 dimerization as substitution of these residues with alanine disrupted dimer formation. Both **S147** and **E166** are located beyond the dimerization interface and are considered active site residues, indicating that the functions of these two sites are inter-dependent. A recent study on SARS-CoV-2 suggested that there is a physical interaction of residues **H163**, **S147**, **L115**, and **S10**, which have low mutation tolerance and form a bridge from the active site to the dimerization site. Each of these mutation-sensitive residues is reported to be crucial for catalysis and is strongly conserved among homologs [51]. Study of a **S284A/T285A/I286A** triple mutant revealed tighter dimer packing, which caused perturbations

Table 3
Hot spots – residues with the lowest metric values (bottom 5 %). The second and third columns list the residue mutation(s) and their locations within the M^{pro} structure. The tick indicates that the residue falls within the bottom 5 % of the DRN metric (for *ECC* and *L* it is the residues with the highest metric values). Additionally, the mutation rates for each residue in each lineage (Alpha, Beta, Delta, Gamma and Omicron) and the total mutation rate across all lineages. The numbers given in each lineage column heading specify the number of unique M^{pro} sequences per that lineage, and the "Total" is for all the sequences analysed in this study. Light grey marks the residues with no mutations.

Residue	Mutations	Location	BC	CC	DC	EC	ECC	KC	L	PR	Alpha (35302)	Beta (4317)	Gamma (8268)	Delta (121477)	Omicron (22514)	Total (191 878)
T24	T24A, T24I	Domain 1			√					√	**7** ($2x10^{-4}$)	**3** ($6.9x10^{-4}$)	**4** ($4.8x10^{-4}$)	**23** ($1.9x10^{-4}$)	**15** ($6.7x10^{-4}$)	**52** ($2.7x10^{-4}$)
S46	S46F, S46A, S46P	Domain 1		√	√		√	√	√		**6** ($1.7x10^{-4}$)			**61** ($5.0x10^{-4}$)	**11** ($4.9x10^{-4}$)	**78** ($4.1x10^{-4}$)
E47	E47N, E47K, E47A	Domain 1	√	√		√	√	√	√		**36** ($1.0x10^{-3}$)	**4** ($9.3x10^{-4}$)	**1** ($1.2x10^{-4}$)	**4** ($3.3x10^{-5}$)	**1** ($4.4x10^{-5}$)	**46** ($2.4x10^{-4}$)
D48	D48N	Domain 1					√				**1** ($2.8x10^{-5}$)	**1** ($2.3x10^{-4}$)	**1** ($1.2x10^{-4}$)	**1** ($8.2x10^{-6}$)		**4** ($2.1x10^{-5}$)
L50	L50F	Domain 1		√		√	√	√	√		**11** ($3.1x10^{-4}$)	**2** ($4.6x10^{-4}$)	**1** ($1.2x10^{-4}$)	**67** ($5.5x10^{-4}$)	**5** ($2.2x10^{-4}$)	**86** ($4.5x10^{-4}$)
N51	N51S	Domain 1		√	√		√	√	√						**1** ($4.4x10^{-5}$)	**1** ($5.2x10^{-6}$)
P52		Domain 1					√									
N53		Domain 1		√			√		√							
E55	E55D	Domain 1		√			√		√		**2** ($5.7x10^{-5}$)		**1** ($1.2x10^{-4}$)	**29** ($2.3x10^{-4}$)		**32** ($1.7x10^{-4}$)
D56		Domain 1	√	√			√		√							
I59	I59L	Domain 1	√	√		√	√	√	√		**2** ($5.7x10^{-5}$)			**8** ($6.6.x10^{-5}$)		**10** ($5.2x10^{-5}$)
R60	R60L, R60C	Domain 1		√			√		√		**1** ($2.8x10^{-5}$)	**1** ($2.3x10^{-4}$)		**22** ($1.8x10^{-4}$)		**24** ($1.3x10^{-6}$)
S62		Domain 1		√			√		√							
H64	H64Y	Domain 1	√	√	√			√	√	√	**1** ($2.8x10^{-5}$)					**1** ($5.2x10^{-6}$)
N72	N72S	Domain 1	√		√			√		√				**9** ($7.4x10^{-5}$)	**1** ($4.4x10^{-5}$)	**10** ($5.2x10^{-5}$)
R76	R76K	Domain 1	√											**1** ($8.2x10^{-6}$)	**1** ($4.4x10^{-5}$)	**2** ($1.0x10^{-5}$)
S81	S81A, S81F, S81T, S81Y	Domain 1								√	**7** ($2.0x10^{-4}$)	**1** ($2.3x10^{-4}$)	**3** ($3.6x10^{-4}$)	**36** ($3.0x10^{-4}$)	**2** ($8.9x10^{-5}$)	**49** ($2.6x10^{-4}$)
D92	D92N, D92G	Domain 1	√								**5** ($1.4x10^{-4}$)			**16** ($1.3x10^{-4}$)	**1** ($4.4x10^{-5}$)	**22** ($1.1x10^{-4}$)
P96	P96S, P96L, P96H, P96T	Domain 1	√		√			√		√	**103** ($2.9x10^{-3}$)		**7** ($8.5x10^{-4}$)	**82** ($6.8x10^{-4}$)	**9** ($4.0x10^{-4}$)	**201** ($1.0x10^{-3}$)
R105	R105H, R105L	Domain 1								√				**16** ($1.3x10^{-4}$)		**16** ($1.3x10^{-4}$)
Q107	Q107R	Domain 1								√			**2** ($2.4x10^{-4}$)	**15** ($1.2x10^{-4}$)		**17** ($8.8x10^{-5}$)
N142	N142S	Domain 2			√			√		√				**4** ($3.3x10^{-5}$)		**4** ($2.1x10^{-5}$)
Y154		Domain 2			√			√		√						
D155	D155E, D155N	Domain 2						√			**2** ($5.7x10^{-5}$)		**1** ($1.2x10^{-4}$)	**4** ($3.3x10^{-5}$)		**7** ($3.6x10^{-5}$)
P168		Domain 2			√					√						
P184	P184S, P184H, P184L, P184T	Linker loop	√		√			√		√	**11** ($3.1x10^{-4}$)	**1** ($2.3x10^{-4}$)		**183** ($1.5x10^{-3}$)	**8** ($3.6x10^{-4}$)	**203** ($1.1x10^{-3}$)
Q189	Q189K	Linker loop					√								**1** ($4.4x10^{-5}$)	**1** ($5.2x10^{-6}$)
T190	T190I, T190N	Linker loop					√				**8** ($2.3x10^{-4}$)		**2** ($2.4x10^{-4}$)	**28** ($2.3x10^{-4}$)	**4** ($1.8x10^{-4}$)	**42** ($2.2x10^{-4}$)
A191	A191S, A191V, A191T	Linker loop	√		√	√	√	√		√	**37** ($1.0x10^{-3}$)		**2** ($2.4 x 10^{-4}$)	**150** ($1.2x10^{-3}$)	**24** ($3.6x10^{-3}$)	**213** ($1.1x10^{-3}$)
W218		Domain 3			√					√						
R222	R222Q, R222L	Domain 3			√	√				√	**5** ($1.4x10^{-4}$)			**11** ($9.1x10^{-5}$)	**3** ($1.3x10^{-4}$)	**19** (9.9x10-5)
F223	F223L, F223V, F223S	Domain 3		√					√		**2** ($5.7x10^{-5}$)		**8** ($9.7x10^{-4}$)	**21** ($1.7x10^{-4}$)	**25** ($1.1x10^{-3}$)	**56** ($2.9x10^{-4}$)
T224	T224N, T224A, T224I	Domain 3		√					√		**1** ($2.8x10^{-5}$)			**39** ($3.2x10^{-4}$)		**40** ($2.1x10^{-4}$)
T226	T226S, T226N, T226A	Domain 3		√					√		**4** ($1.1x10^{-4}$)			**7** ($5.8x10^{-5}$)		**11** ($5.7x10^{-5}$)
N228	N228S	Domain 3	√	√		√			√					**6** ($4.9x10^{-5}$)	**2** ($8.9x10^{-5}$)	**8** ($4.2x10^{-5}$)
L232	L232F, L232I	Domain 3	√			√					**3** ($8.5x10^{-5}$)	**11** ($2.5x10^{-4}$)	**2** ($2.4 x 10^{-4}$)	**30** ($2.5x10^{-4}$)	**1** ($4.4x10^{-5}$)	**47** ($2.4x10^{-4}$)
M235	M235I	Domain 3				√					**2** ($5.7x10^{-5}$)			**2** ($1.7x10^{-5}$)		**4** ($2.1x10^{-5}$)
K236	K236R	Domain 3	√		√	√		√			**12** ($3.4x10^{-4}$)	**1** ($2.3x10^{-4}$)	**17** ($2.1x10^{-3}$)	**23** ($1.9x10^{-4}$)	**4** ($1.8x10^{-4}$)	**57** ($3.0x10^{-4}$)
Q244		Domain 3	√													
G258		Domain 3	√		√					√						
N274	N274D, N274S, N274T	Domain 3				√					**62** ($1.8x10^{-3}$)	**2** ($4.6x10^{-4}$)	**1** ($1.2x10^{-4}$)	**48** ($3.9x10^{-4}$)	**16** ($7.1x10^{-4}$)	**129** ($6.7x10^{-4}$)
G275	G275S	Domain 3				√					**1** ($2.8x10^{-5}$)			**22** ($1.8x10^{-4}$)		**23** ($1.2x10^{-4}$)
M276	M276L, M276I, M276V, M276T	Domain 3				√					**4** ($1.1x10^{-4}$)			**8** ($6.6.x10^{-5}$)		**12** ($6.3x10^{-5}$)
N277	N227T, N227D, N227S	Domain 3				√					**3** ($8.5x10^{-5}$)		**1** ($1.2x10^{-4}$)	**25** ($2.1x10^{-4}$)	**6** ($2.7x10^{-4}$)	**35** ($1.8x10^{-4}$)
G278	G278E, G278R	Domain 3				√							**1** ($1.2x10^{-4}$)	**1** ($8.2x10^{-6}$)	**6** ($2.7x10^{-4}$)	**8** ($4.2x10^{-5}$)
R279	R279H, R279C, R279L	Domain 3				√		√			**12** ($3.4x10^{-4}$)	**1** ($2.3x10^{-4}$)		**40** ($1.8x10^{-4}$)	**1** ($4.4x10^{-5}$)	**54** ($2.8x10^{-4}$)

of the N-finger and helix A that were transmitted to distal residues including **T111**, **S113**, **L115-Y118** and **S123**, leading to enhanced catalytic activity [83]. These findings emphasise the importance of residues networks in maintaining both the structural integrity and functional efficacy of the viral protease.

Collectively, as shown here and in our previous work [18,42,43], DRN metrics can effectively identify key residues in functional regions of proteins, including M^{pro} of SARS-CoV-2. We extracted top 5 % residues with the highest metric values, and thus we considered them as potential cold spots. Many of these residues have been demonstrated in other studies to support M^{pro} dimerization and/or catalytic activity.

3.2. Revealing hot spots: dynamic residue network analysis supported by prior literature studies

With the rapid emergence of SARS-CoV-2 variants, it is important to identify not only cold spot residues that may have core functions in maintaining enzyme activity, but also hot spot residues that could alter protein dynamics, modify enzyme activity or confer resistance to M^{pro} inhibitors. We applied a similar approach as described above to extract the potential hot spots, using the eight DRN metrics to identify a total of 46 hub residues (Table 3). These residues have the lowest metric values, representing the bottom 5 % per metric, except for *L* where we extracted the top 5 %. We did not identify any *persistent hubs*, but there were four residues identified by six metrics (**E47**, **I59**, **H64**, **A191**) and three residues recognized by five metrics (**S46**, **L50**, **N51**) (Table 3). Eight of the hot spot residues exhibited no mutations across five SARS-CoV-2 lineages. 80.4 % of the hot spot residues across these metrics underwent mutation in at least one SARS-CoV-2 lineage, with mutation frequencies ranging from 5.2×10^{-6} (**N51** and **H64**) to 1.1×10^{-3} (**A191**) (Table 3).

We also selected the 5 % of residues that are most frequently mutated in M^{pro} sequences, focusing on their metric specific positions (Table 4), to evaluate the accuracy of the DRN in identifying potential mutational hot spots. Among the 15 most frequently mutated residues analysed (Table 4), only three (**P96**, **P184** and **A191**) were identified by the DRN metrics as being in the bottom 5 %. **P96** and **P184** were detected by *BC*, *DC*, *KC and PR* metrics, while **A191** was identified by *six* metrics. Interestingly, the DRN metrics also identified eight residues neighbouring the most frequently mutated residues, including residues **D92**, **A191**, **Q244**, **Q107**, **R76**, **N72**, **T190**, **Q244** and **T24** that are adjacent to **K90**, **A193**, **P241**, **P108**, **L75**, **V73**, **A191, H246** and **T21** (listed in descending order of mutation frequency). The most frequently mutated residue in M^{pro} during the evolution of SARS-CoV-2 was **P132**, followed by **K90** and **A193** (Table 4), none of which ranked within the bottom 5 % of the eight DRN metrics analysed. **P132** was best detected by the *averaged EC* metric at position 218, ranking in the 28.29th percentile from the bottom. **K90** was most effectively identified by the *averaged BC* and *averaged L* metrics. **A193** was best selected by the *averaged DC* and *KC* metrics at position 274, corresponding to the bottom 9.87 % percentile in both metrics (Table 4).

Visual inspection of the bottom 5 % metric hubs mapped onto the M^{pro} structure for each of the four commonly used metrics revealed that these hubs are predominantly located in the loops and peripheral regions of the protein, particularly within domains I and III, and are distant from the N finger region and dimerization interface (Fig. 1**E-H**). This observation suggests that the DRN metrics have the capability to identify potential hot spot regions that are relatively distant from structurally and functionally critical areas of the protein. Sequence analysis revealed that there are 12 key residue substitutions between SARS-CoV-1 and SARS-CoV-2. Most of the substitutions are found in the β-strand rich domains I and II, while 4 substitutions were found in domain III [84]. Among the hot spot residues identified in our study the residue **S46** was substituted from alanine in SARS CoV-1. Among the different variants identified in previous studies and strains, the most concerning ones are those residing at or in the vicinity of the active site, among which **L50F**, **E47K**, **E47N**, and **S46F**, are identified as occurring at hot spot residues (Table 3). The variant **E47N** is a prominent mutation that appeared with high prevalence in the SARS-CoV-2 Alpha VOC [85]. Available experimental reports state that the two variant of the residue **E47**, **E47N** and **E47K**, have significantly different consequences on the protease activity. Substitution of this glutamic acid with asparagine causes modest changes in the conformations of multiple substrate-binding residues, leading to a two-fold decrease in substrate binding affinity, whereas substitution with the positively charged residue lysine significantly increases substrate binding efficacy, possibly due to dramatic change in charge around the S3′ binding site [85]. A recent study identified that **L50F**, which is a part of an 'active site gateway', had a lower K_M value and slightly elevated k_{cat}, which resulted in 1.6-fold higher proteolytic efficiency as determined by FRET-based cleavage assays [85]. An independent study reported 1.7-fold increased enzymatic activity for the **L50F** mutant compared to wild type (Wuhan) M^{pro} [86], but a third independent study reported that the **L50F** substitution leads to almost complete loss of enzymatic activity [87]. The naturally occurring **L50F** variant demonstrated nirmatrelvir resistance in cell culture plus high fitness in cell-based infection systems [88].

The active site gateway comprises of two loops, L50–Y54 and D187–A191, which stabilizes the substrate binding pocket. Along with **L50**, the residue **A191** is among the identified hot spot residues and **A191V** is among the most frequently observed naturally occurring variants. FRET based assay reveal increased catalytic efficiency in **A191T** mutant, while **A191V** displayed comparable activity to wild type (Wuhan) SARS-CoV-2 M^{pro} [89]. A rare clinical mutant isolate with substitution **F185S** had a significant (30-fold) decrease in activity. **F185** is in proximity to the active site and forms several interactions that stabilizes the loop that connects domain II to domain III. **P184**, which is one of the identified hot spots, is involved in pi-stacking interaction with **F185**, contributing to stabilizing the loop. Differential scanning fluorimetry revealed that mutation of **P184** to serine has a destabilizing effect on the molecule [85].

Multiple studies have identified that mutation of hot spot residue **N142** does not dramatically decrease M^{pro} activity, but instead leads to unchanged or increased catalytic activity [77,86,90,91]. *In silico* screening coupled with biochemical analysis suggested that the **N142L** substitution increased catalytic activity and reduced susceptibility to the covalent M^{pro} inhibitor nirmatrelvir [91], and substitution of **N142** to serine or aspartic acid has been reported to modestly reduce susceptibility to inhibition by nirmatrelvir [89]. However, another study observed no significant change in either catalytic activity or nirmatrelvir resistance when **N142** was substituted with nine different amino acids, including serine, leucine and aspartic acid [90]. Although not concordant in their assessment of nirmatrelvir resistance, these studies do agree that mutation of the active-site proximal hotspot residue **N142** is well tolerated by the enzyme. Interestingly, **N142** lies close to the binding site of pelitinib and changes conformation dramatically upon pelitinib binding [92], suggesting that **N142** mutations may also alter pelitinib efficacy. A deep mutagenic scanning study concluded that **P168** is highly tolerant of mutation [77]. Kinetic characterisation of **P168M** substituted M^{pro} demonstrated a slight reduction in enzymatic activity and no resistance to the effects of the inhibitor nirmatrelvir [91].

Overall, these analyses show that DRN analysis can identify hot spot residues that have a high propensity for mutation. However, there is not a perfect correspondence between the set of hot spot residues and those residues that most frequently vary in SARS-CoV-2 sequences. While hot spot residue mutations often have only mild consequences for M^{pro} activity, there have been reports that hot spot mutations can alter M^{pro} susceptibility to inhibitor compounds such as nirmatrelvir [88,89,91], although we note that there is not universal support for these effects [90]. This highlights the need for additional large-scale enzymatic studies to probe the effects M^{pro} mutations upon enzyme activity and

Table 4
The top 5 % most frequently mutated M^{pro} residues during the course of SARS-CoV-2 evolution with their metric specific positions listed in the descending and their corresponding percentiles in ascending order.

Residue	Mutations	Location	*BC*	*CC*	*DC*	*EC*	*ECC*	*KC*	*L*	*PR*	Alpha (35,302)	Beta (4317)	Gamma (8268)	Delta (121,477)	Omicron (22,514)	Total (191,878)
P132	P132H, P132L, P132S	Domain 2	111 (63.49 %)	143 (52.96 %)	188 (38.16 %)	218 (28.29 %)	175 (42.43 %)	196 (35.53 %)	143 (52.96 %)	182 (40.13 %)	**41** (1.2×10^{-3})	**2** (4.6×10^{-4})	**10** (1.2×10^{-3})	**325** (2.7×10^{-3})	**22,219** (9.9×10^{-1})	**22,597** (1.2×10^{-1})
K90	K90N, K90R	Domain 1	244 (19.74 %)	248 (18.42 %)	118 (61.18 %)	112 (63.16 %)	241 (20.72 %)	116 (61.84 %)	247 (18.75 %)	124 (59.21 %)	**695** (2.0×10^{-2})	**4297** 0.99	**160** (1.9×10^{-2})	**3191** (2.6×10^{-2})	**123** (5.5×10^{-3})	**8466** (4.4×10^{-2})
A193	A193T, A193V, A193S	Domain 2	219 (27.96 %	198 (34.87 %)	274 (9.87 %)	260 (14.47 %)	258 (15.13 %)	274 (9.87 %)	198 (34.87 %)	260 (14.47 %)	**5** (1.4×10^{-4})	**773** (1.8×10^{-1})	**2** (2.4×10^{-4})	**78** (6.4×10^{-4})	**26** (1.2×10^{-3})	**884** (4.6×10^{-3})
P241	P241H, P241L, P241S, P241T	Domain 3	212 (30.26 %)	220 (27.63 %)	235 (22.70 %)	257 (15.46 %)	218 (28.29 %)	227 (25.33 %)	220 (27.63 %)	248 (18.42 %)	**232** (6.6×10^{-3})	**32** (7.4×10^{-3})	**90** (1.1×0^{-2})	**315** (2.6×10^{-3})	**32** (1.4×10^{-3})	**701** (3.7×10^{-3})
A260	A260S, A260T, A260V	Domain 3	260 (14.47 %)	239 (21.38 %)	207 (31.91 %)	224 (26.32 %)	190 (37.50 %)	174 (42.76 %)	239 (21.38 %)	228 (25.00 %)	**58** (1.6×10^{-3})	**8** (1.9×10^{-3})	**9** (1.1×10^{-3})	**385** (3.2×10^{-3})	**10** (4.4×10^{-4})	**470** (2.4×10^{-3})
P108	P108L, P108Q, P108S, P108T	Domain 2	201 (33.88 %)	121 (60.20 %)	205 (32.57 %)	197 (35.20 %)	151 (50.33 %)	212 (30.26 %)	120 (60.53 %)	214 (29.61 %)	**76** (2.2×10^{-3})	**32** (7.4×10^{-3})	**72** (8.7×10^{-3})	**241** (2.0×10^{-3})	**18** (8.0×10^{-4})	**439** (2.3×10^{-3})
L75	L75F, L75I, L75S	Domain 1	224 (26.32 %)	234 (23.03 %)	117 (61.51 %)	121 (60.20 %)	213 (29.93 %)	128 (57.89 %)	235 (22.70 %)	84 (72.37 %)	**106** (3.0×10^{-3})	**14** (3.2×10^{-3})	**25** (3.0×10^{-3})	**83** (6.8×10^{-4})	**109** (4.8×10^{-3})	**337** (1.8×10^{-3})
V73	V73A, V73F, V73I, V73L, V73K, V73T	Domain 1	262 (13.82 %)	246 (19.08 %)	258 (15.13 %)	212 (30.26 %)	222 (26.97 %)	275 (9.54 %)	246 (19.08 %)	213 (29.93 %)	**18** (5.1×10^{-4})	**0**	**4** (4.8×10^{-4})	**288** (2.4×10^{-3})	**1** (4.4×10^{-5})	**311** (1.6×10^{-3})
I213	I213T, I213V, I213M	Domain 3	118 (61.18 %)	99 (67.43 %)	152 (50.00 %)	152 (50.00 %)	102 (66.45 %)	119 (60.86 %)	99 (67.43 %)	181 (40.46 %)	**21** (5.9×10^{-4})	**0**	**1** (1.2×10^{-4})	**245** (2.0×10^{-3})	**4** (1.8×10^{-4})	**271** (1.4×10^{-3})
A191	A191S, A191V, A191T	Linker loop	292 (3.95 %)	269 (11.51 %)	303 (0.33 %)	303 (0.33 %)	297 (2.30 %)	304 (0.00 %)	270 (11.18 %)	301 (0.99 %)	**37** (1.0×10^{-3})	**0**	**2** (2.4×10^{-4})	**150** (1.2×10^{-3})	**24** (1.1×10^{-3})	**213** (1.1×10^{-3})
P184	P184S, P184H, P184L, P184T	Linker loop	297 (2.30 %)	229 (24.67 %)	301 (0.99 %)	273 (10.20 %)	283 (6.91 %)	300 (1.32 %)	229 (24.67 %)	302 (0.66 %)	**11** (3.1×10^{-4})	**1** (2.3×10^{-4})	**0**	**183** (1.5×10^{-3})	**8** (3.6×10^{-4})	**203** (1.1×10^{-3})
P96	P96S, P96L, P96H	Domain 1	304 (0.00 %)	216 (28.95 %)	304 (0.00 %	226 (25.66 %)	186 (38.82 %)	301 (0.99 %)	216 (28.95 %)	304 (0.00 %)	**103** (2.9×10^{-3})	**0**	**7** (8.5×10^{-4})	**82** (6.8×10^{-4})	**9** (4.0×10^{-4})	**201** (1.0×10^{-3})
H246	H246Y, H246N, H246R	Domain 3	174 (42.76 %)	189 (37.83 %)	140 (53.95 %)	200 (34.21 %)	124 (59.21 %)	153 (49.67 %)	188 (38.16 %)	126 (58.55 %)	**32** (9.1×10^{-4})	**7** (1.6×10^{-3})	**119** (1.4×10^{-2})	**34** (2.8×10^{-4})	**6** (2.7×10^{-4})	**198** (1.0×10^{-3})
K100	K100N, K100R	Domain 1	265 (12.83 %)	166 (45.39 %)	257 (15.46 %)	170 (44.08 %)	160 (47.37 %)	260 (14.47 %)	166 (45.39 %)	243 (20.07 %)	**1** (2.8×10^{-5})	**159** (3.7×10^{-2})	**0**	**11** (9.1×10^{-5})	**0**	**171** (8.9×10^{-4})
T21	T21I, T21A	Domain 1	193 (36.51 %)	208 (31.58 %)	189 (37.83 %)	123 (59.54 %)	204 (32.89 %)	178 (41.45 %)	207 (31.91 %)	185 (39.14 %)	**34** (9.6×10^{-4})	**7** (1.6×10^{-3})	**33** (4.0×10^{-3})	**80** (6.6×10^{-4})	**17** (7.6×10^{-4})	**171** (8.9×10^{-4})

inhibitor susceptibility.

3.3. Comparing computational and mutational screening to estimate mutation tolerance

Deep mutational scanning (DMS) is a powerful functional method to define the mutation tolerance of specific residues within a protein [51, 77]. We compared the set of residues that were identified as cold spots or hot spots by DRN using any of the eight metrics tested (Tables 2 and 3) with the set of residues identified by DMS of M^{pro} [51] as being most mutation sensitive (greatest loss of function when mutated) or least mutation sensitive, respectively (Fig. 2). The number of residues per set was defined by the number identified in the DRN analyses, representing 17 % (cold spots) or 15 % (hot spots) of resides in M^{pro}. For both comparisons we observed more overlap between the DRN and DMS sets than would be expected by random chance: 22 residues that are both cold spots and highly mutation sensitive, versus 9 residues (2.9 %) to be expected by random selection, and 20 residues that are both hot spots and highly mutation tolerant, versus 7 residues (2.3 %) to be expected from random selection. The comparison of hot spots and mutation-tolerant residues is confounded by the fact that DMS analysis shows approximately half of the residues in M^{pro} to have WT-like functional scores (between 0.9 and 1.0), indicating high mutation tolerance [51]. Of the 46 residues that are identified as hot spots in M^{pro} (Table 3), 42 have WT-like functional scores (> 0.9). Three others have high functional scores (0.89 for **T190**, 0.87 for **P52** and 0.78 for **Q189**), indicating significant residual activity. No DMS data was obtained for **G258** by Flynn and colleagues [51], but another DMS study showed this residue to be highly mutation tolerant [77]. Similarly, 48 of the 52 cold spot residues have functional scores < 0.9, indicating either intermediate or null-like catalytic function [51]. These results suggest that DRN analysis can readily identify hot spot residues dispensable for protein activity. DRN analysis can also identify cold spot residues that are likely to have important functional roles, but the metric scores from DRN analysis and catalytic sensitivity of residues to mutation (as determined using DMS) are not perfectly correlated.

We also compared the hot spots and most mutation tolerant residues with the set of residues having the highest mutation frequency in SARS-CoV-2 sequences (Fig. 2**B**), revealing only modest overlap with both sets. This may again arise from the fact that approximately half the residues within M^{pro} are highly mutation tolerant in vitro, making the relative ranking of these residues somewhat arbitrary. However, we wondered whether the rich information on protein dynamics that is accessible via DRN might facilitate greater discrimination of relative propensity to mutate for these mutation-tolerant residues. Given the importance of predicting mutation propensity for drug resistance, we thus proceeded to investigate whether the power of DRN analysis could be improved by combining it with ML approaches.

3.4. Boosting DRN predictive power with advanced ML models

To enhance the predictive power of the DRN, we incorporated individual residue values per metric, along with other predictors, into our ML models as outlined in the Methodology section. Here, we present eight ML models that comprise all combinations of ANN or RF, Regression or Classification, MATLAB or Scikit-Learn. In all cases, the models split the data randomly into training (70 %), validation (15 %) and testing (15 %) sets, and we set the initial state for training in a random way. For reproducibility, we set values for the random seeds; varying these values can lead to varying performance in the predictive power of the models, particularly because our datasets are small, e.g., the validation and testing datasets typically have 45 elements.

The performance of the models is summarized in Table 5A, which gives the "R" values for each regression model for each data subset, as well as the AUC values for each classification model for each data subset. The "Combined" subset is all data for the ANN models but comprises only the testing and validation data for the RF models. This is because RF overfits to the training data (as seen by the AUC values of 1 in the Table 5A), so its performance needs to be evaluated on other data. We have also produced ROC plots for each classification model, see Fig. 3. The regression plots and confusion matrices for the Scikit-Learn ANN models are given in Fig. 4; and for the other cases in the Supplementary Data that represent the best models – Fig. S3 for MATLAB ANN, Fig. S4 for MATLAB RF, and Fig. S5 for Scikit-Learn RF. In addition to this, the intermediate model examples (Fig. S6 for Scikit-Learn ANN, Fig. S7 for MATLAB ANN, and Fig. S8 for MATLAB RF) and worst model examples (Fig. S9 for Scikit-Learn ANN, Fig. S10 for MATLAB ANN, and Fig. S11 for MATLAB RF) are given in the Supplementary Data.

Recalling that AUC = 0.5, R = 0 for a random predictor and AUC = 1, R = 1 for a perfect predictor, it is seen that in all cases the models are significantly better than random, but far from excellent. Importantly, for the ANN models, the R and AUC values for the testing and other datasets are quite similar (Scikit case) or only a little different (MATLAB case), so there has not been significant overfitting. With training data excluded, the range in R values is between 0.421 and 0.673, and the range in AUC values between 0.650 and 0.820. All the models could be described as being reasonable predictors though not ones of high quality.

3.5. Comparison of DRN and ML predictions

Next, we compared the predictive power of DRN both independently and when integrated with the ML models.

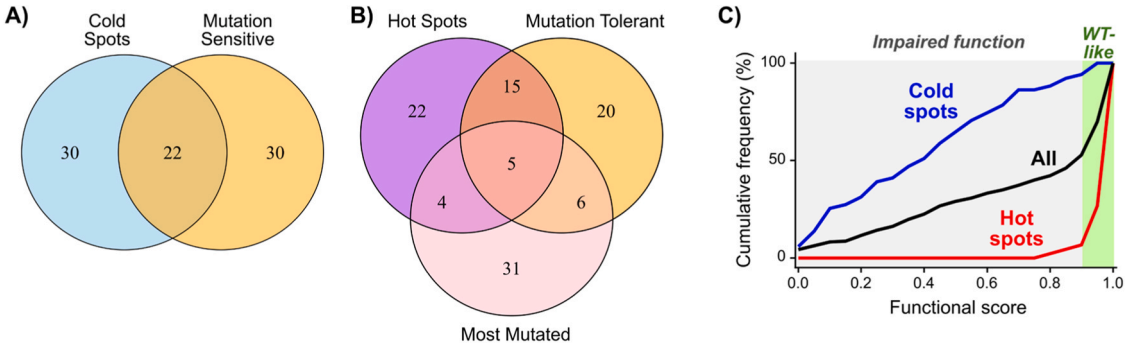

Fig. 2. Graphical comparison of predicted cold and hot spots from DRN analysis with functional scores from deep mutational scanning (DMS) [51]. **A)** Overlap of cold spots identified via DRN metrices (Table 2) versus least mutation-tolerant residues in M^{pro} (52 residues in each set). B) Hot spots identified by DRN analysis (Table 3) are compared with the most mutation-tolerant residues from DMS and the residues most frequently mutated in deposited sequences (46 residues in each set). **C)** Cumulative frequency distribution of functional scores from DMS for all residues versus residues identified as cold or hot spots by DRN analysis. Residues with functional scores above 0.9 have "WT-like" function, while residues with lower scored have impaired function (either "intermediate" or "null-like" functions) [51].

Table 5A
Evaluating the predictive power of the ML model using the correlation coefficient or "R" value for the regression models, and the AUC value for ML classification models.

Method & Program	Regression: R value				Classification: AUC value			
	Training	Validation	Testing	Combined	Training	Validation	Testing	Combined
ANN MATLAB	0.752	0.531	0.613	0.692	0.845	0.650	0.720	0.798
ANN Scikit	0.653	0.673	0.664	0.656	0.833	0.860	0.820	0.836
RF MATLAB	0.862	0.503	0.421	0.460	1.000	0.732	0.754	0.738
RF Scikit	0.651	0.489	0.581	0.533	1.000	0.736	0.759	0.740

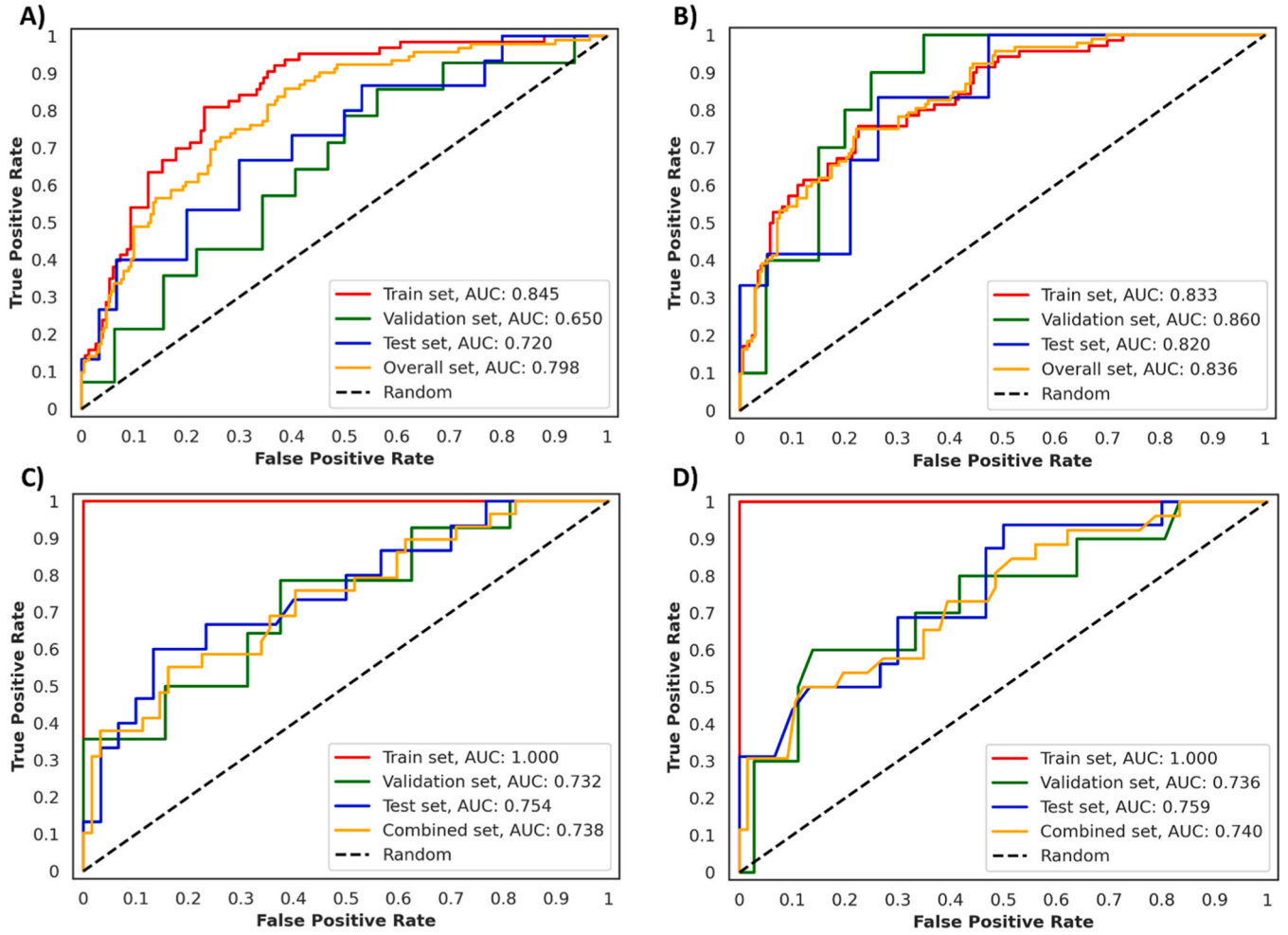

Fig. 3. ROC plots showing the true (y-axis) and false positive (x-axis) rates represented for each classification model; **A)** MATLAB ANN, **B)** Python Sci-kit learn ANN, **C)** MATLAB RF (using TreeBagger) and **D)** Python Sci-kit learn RF. Each subplot consists of curves and computed area under the curve (AUC) for the training (red), validation (green), test (blue) and the overall/combined datasets (orange). In each case, the dashed diagonal line corresponds to a random predictor.

3.5.1. Correlation between predictions (by DRN or ML) and target values

We calculated the correlation, or "R" values, between various metrics and $\log_{10}(1 + \text{mutation frequency})$, i.e. adjusting the observed or target data as was done for the ML regression models. The results are presented in Table 5B and vary between 0.22 and 0.3. The "R" values between predictions and targets for the ML models are given in Table 5A: the values for independent data sets vary between 0.421 and 0.637, with an average of 0.55. These results clearly demonstrate that ML is able to combine information from the different metrics and make predictions that are substantially more accurate than those from the individual metrics.

3.5.2. Comparison of DRN and ML predictions of cold and hot spot residues

There are 109 residues in M^{pro} that remained unmutated throughout the evolution of SARS-CoV-2. The protein analysed has 304 residues, and we aimed to determine the location of these non-mutated residues within each metric. To do this, we divided the metric specific datasets into three bins: the top bin containing 102 residues with the highest metric values (and *ECC* and *L* with the lowest metric values), the middle bin with 101 residues, and the bottom bin with 101 residues having the lowest metric values (with *ECC* and *L* being the highest in this case). To assess the significance of non-mutating residue enrichment in the top 102 residues (top third) as well as the significance of their depletion in the bottom 101 residues (bottom third) bins for each DRN metric, hypergeometric p-values were calculated. The p-values were computed for two mutation frequency cut-offs: a cut-off of 0, where non-mutated residues have a mutation frequency of 0 in the used GISAID dataset, and a cut-off of 20, where non-mutated residues have a mutation

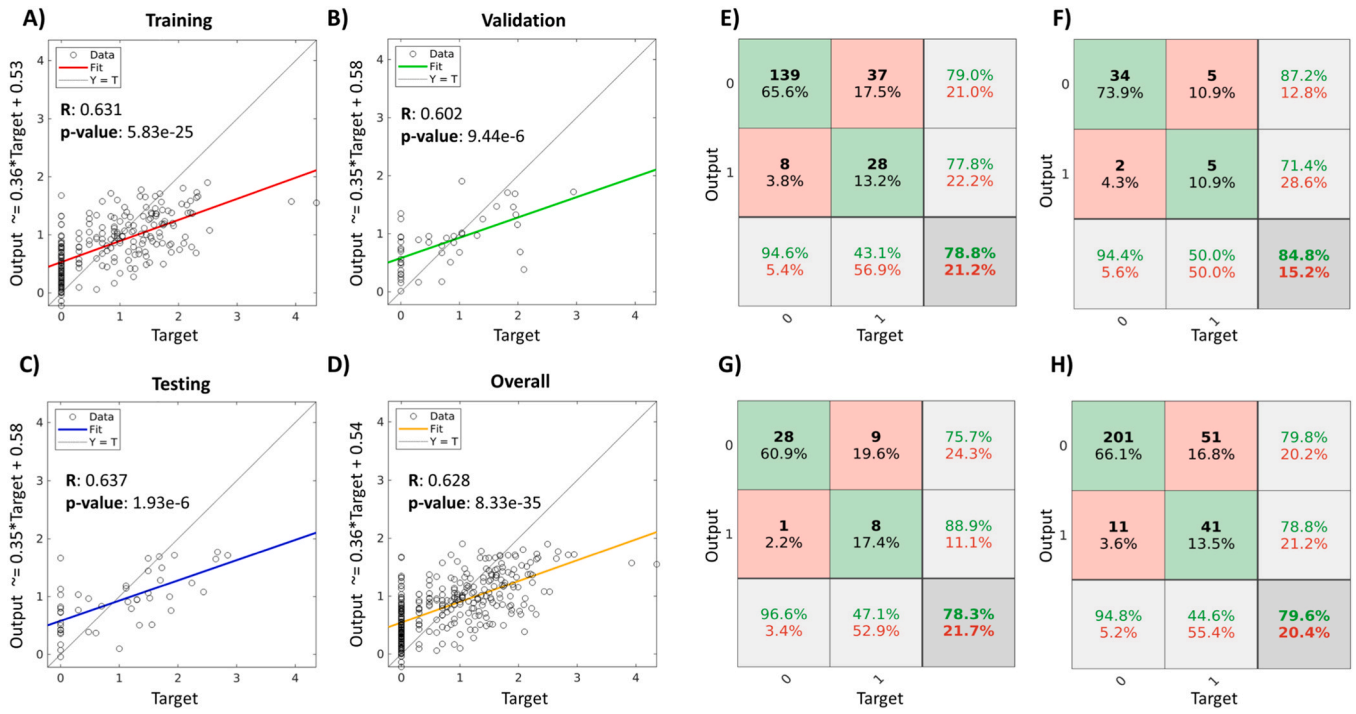

Fig. 4. Assessment of Python Scikit-learn ANN model. The correlation (scatter plots) between model predictions and target values for the regression model are shown across the datasets: **A)** training, **B)** validation, **C)** testing, and **D)** overall. The x-axis represents the target values, while the y-axis shows the output or predicted values. As discussed in Section 2.4, the target data is $\log_{10}$ (1 + mutation frequency) and varies between 0 and about 4. Each plot features a line of best fit, and the R-value. For the classification model in which each residue is regarded as either mutated or not-mutated, **E-H** illustrate the confusion matrices for the training, validation, testing, and overall datasets, respectively. The x-axis denotes the target values, and the y-axis the predicted outputs. In each classification subplot, the first two green diagonal cells display the number and percentage of correct classifications, and the red off-diagonal cells show the number and percentage of incorrect classifications. For the overall dataset, 79.6 % of classifications were correct and 20.4 % were wrong.

Table 5B
The "R" value for the correlation between various DRN metrics and the target data.

Metric	*Averaged BC*	*Averaged CC*	*Averaged CC*	*Averaged EC*	*Averaged ECC*	*Averaged KC*	*Averaged L*	*Averaged PR*	SASA	B factor	RMSF
|R| value	0.30	0.30	0.27	0.22	0.26	0.28	0.29	0.23	0.28	0.25	0.24

frequency ranging from 0 to 20. For the top 102 and bottom 101 residue bins, hypergeometric analysis showed highly significant enrichment of non-mutating residues in the top bin and depletion in the bottom bin, with p-values less than 0.01 for both the 0 and 20 mutation frequency cut-offs (Table 6).

In the DRN analysis, residues were ranked in descending order based

Table 6
Enrichment fold and the enrichment/depletion p-values from hypergeometric test of non-mutated residues in the top and bottom third of the DRN metrics (non-mutated cut-off of 20). **Notes**: p-value: * p < .05, ** p < .01, *** p < .001.

Averaged DRN metric	No. of non-mutated residues in the top bin	Enrichment fold	P value		No. of non-mutated residues in the bottom bin	Depletion fold	P value	
Cut-off of 0								
BC	56	1.53	9.78×10^{-7}	***	16	2.26	1.05×10^{-7}	***
CC	54	1.48	1.05×10^{-5}	***	21	1.72	6.67×10^{-5}	***
DC	54	1.48	1.05×10^{-5}	***	21	1.72	6.67×10^{-5}	***
EC	48	1.31	2.97×10^{-3}	**	23	1.57	5.09×10^{-4}	***
ECC	53	1.45	3.11×10^{-5}	***	26	1.39	6.31×10^{-3}	**
KC	51	1.39	2.31×10^{-4}	***	18	2.01	1.78×10^{-6}	***
L	54	1.48	1.05×10^{-5}	***	21	1.72	6.67×10^{-5}	***
PR	51	1.39	2.31×10^{-4}	***	20	1.81	2.16×10^{-5}	***
Cut-off of 20								
BC	89	1.25	8.08×10^{-7}	***	53	1.33	4.77×10^{-6}	***
CC	88	1.24	3.37×10^{-6}	***	50	1.41	9.51×10^{-8}	***
DC	85	1.19	1.38×10^{-4}	***	59	1.19	2.06×10^{-3}	**
EC	85	1.19	1.38×10^{-4}	***	54	1.30	1.54×10^{-5}	***
ECC	87	1.22	1.28×10^{-5}	***	54	1.30	1.54×10^{-5}	***
KC	86	1.21	4.40×10^{-5}	***	54	1.30	1.54×10^{-5}	***
L	88	1.24	3.37×10^{-6}	***	50	1.41	9.51×10^{-8}	***
PR	81	1.14	5.92×10^{-3}	**	60	1.17	4.55×10^{-3}	**

on their metric values (except for metrics *ECC* and *L*), with higher values indicating greater functional importance, and potentially being less prone to mutations. The analysis is presented as an alluvial plot in Fig. 5**A**, which illustrates the transitions of the unmutated residues between the bins for eight DRN metrics. The Supporting information details the unmutated residues found within each bin across the eight DRN metrics (Table S1).

For the ANN ML models, residues were ranked in ascending order, as the ML models predict the number of mutations. Fig. 5**B** represents the flow of the unmutated residues between each bin of the two ML models. The unmutated residues predicted within each bin for the two models are listed in Table S2. We focused on the two ANN ML models only because, as previously discussed, RF tends to overfit the training data. To compare with the DRN analysis, we need to consider the entire dataset of 304 residues, which is represented by the combined dataset. Here, we looked at the combined dataset of the best model.

Notably, while DRN metrics accurately predicted approximately half of the residues with no mutations in the top bin (56, 54, 54, 48, 53, 51, 53, 51 for averaged *BC*, *CC*, *DC*, *EC, ECC, KC, L, PR* respectively; see Table S1), the prediction rate significantly increased with the incorporation of ML ANN models (72 for both Python and MATLAB ANN; see Table S2). This demonstrates the effectiveness of ML approaches. As expected, there was a progressive reduction in unmutated residues in the middle bin followed by the bottom bin, with ML models showing improved predictive power in these two bins as well (Tables S1 and S2). The alluvial plots in Fig. S12 illustrate distinct residue flow patterns across DRN metrics for Python and MATLAB ANN models. Both ML models show significant residue prediction similarity seen across the three bins. Furthermore, notable differences in distribution are evident, particularly in the middle and bottom bins between ML models and DRN metrics.

Secondly, we revisited the top 5 % most frequently mutated residues of M^{pro} (Table 4). Although the combination of DRN metrics identified three hot spot residues (**P96**, **P184** and **A191**) out of 15 (Fig. 1 and Table 3), integrating DRN with ML models significantly increased the prediction accuracy (Fig. 6). Specifically, the Python ANN model predicted five highly mutated residues (**A193**, **P241**, **A260**, **V73**, **P184**), while the MATLAB ANN model included **P132** (the most mutated residue in the evolution of protein), **K90**, **A193**, **A260**, **P108** and **V73**. Altogether, the ML models predicted eight residues, demonstrating roughly three-fold increase in prediction power.

Lastly, we compared the overlap between predicted cold and hot spots identified by DRN (top and bottom 5 % residues) (Table 2, Table 3 and Fig. 1) and those identified by the ML ANN models (bottom and top 5 % residues respectively) (Fig. 6). ML models commonly predicted residues **N28**, **C128**, **H163**, **N203**, **N214** and **W218** as cold spots while eight DRN metrics commonly identified only **L115**. While there were no common hot spot residues identified among the eight DRN metrics, ML models agreed on residues **V73**, **A193** and **A260** as hot spots.

4. Conclusion

Evolutionary mutations of viruses and other pathogens pose significant challenges to drug discovery by altering the structure, function, and interactions of drug targets. To overcome these challenges, drug discovery efforts must continuously adapt by monitoring pathogen mutations and updating therapeutic strategies accordingly. Identifying conserved mutation patterns that lead to drug resistance mechanisms, for example, can help in designing more effective and long-lasting drugs [93,94]. Identifying mutation-resilient (cold spots) and mutation-prone regions (hot spots), and thereafter developing drugs that target cold spots, can guide the development of broad spectrum therapies that are effective against multiple viral strains.

While understanding the effects of current mutations at the protein structure level is achievable, predicting future mutations remains a significant challenge. Here, we took advantage of the extensive mutation data available for SARS-CoV-2 M^{pro}, along with a vast amount of literature to test our hypothesis regarding the mutation position predictive power of DRN metrics alone and together with ML approaches. We used the M^{pro} protein from SARS-CoV-2 (Wuhan strain) to perform DRN analysis across eight metrics (*averaged BC*, *CC*, *DC*,*EC, ECC, KC, L, PR*) and evaluated how well the combined metric values correlate with per residue mutation frequencies observed during the evolution SARS-CoV-2. We then used ML to combined these DRN with other sequence- and structure-based metrics (BLOSUM62 matrix, RMSF, SASA and B-factor), with the aim of enhancing our ability to predict residue mutation likelihood.

Our key observations in this study can be summarized as follows: **(1)** We identified the top 5 % of residues with the highest metric values across eight metrics as potential cold spots, resulting in 52 unique residues. Mutation rates were calculated for each residue across five lineages, totalling 191,878 unique M^{pro} sequences. The DRN metrics identified 25 residues out of 52 with zero mutations throughout the evolution of virus up to February 24, 2024. Thus, 23 % (25 out of 109) of the non-mutated M^{pro} residues were within the top 5 % of high centrality residues. **(2)** Mapping potential cold spots onto 3D structure of

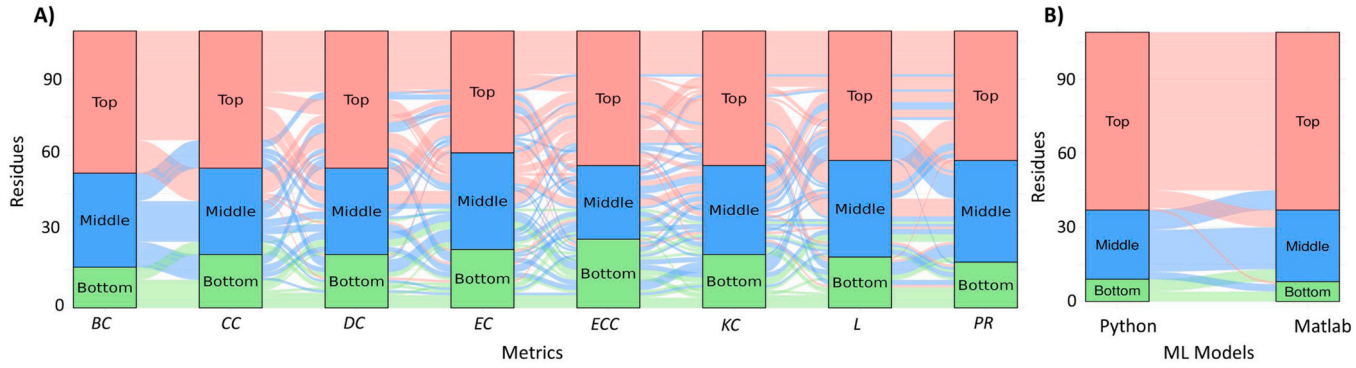

Fig. 5. Alluvial plot depicting the classification of the 109 residues with no mutations into top, middle, and bottom groups across **A)** eight DRN centrality metrics and **B)** two machine learning models**.** Each column represents one metric in **A)**, with the leftmost column representing *BC* and the rightmost column representing *PR*. The columns in **B)** represent the Python ANN and MATLAB ANN machine learning models. The flows between the columns illustrate how the unmutated residues transition between the top, middle, and bottom groups for each metric (**A)**) or model (**B)**). Each flow is coloured according to the groups of the first column. The thickness of the flows corresponds to the number of residues moving between the groups. The top group (pink) contains the first 102 residues of the M^{pro} protein with the highest values of the DRN metric (**A)**) or lowest predicted mutation frequencies (B)). The middle group (blue) contains 101 residues with the next highest DRN metric values (**A)**), or next lowest mutation frequency predictions (**B)**). The bottom group (green) contains the last 101 residues with the lowest values of the DRN metric (**A)**), or highest mutation frequencies (**B)**).

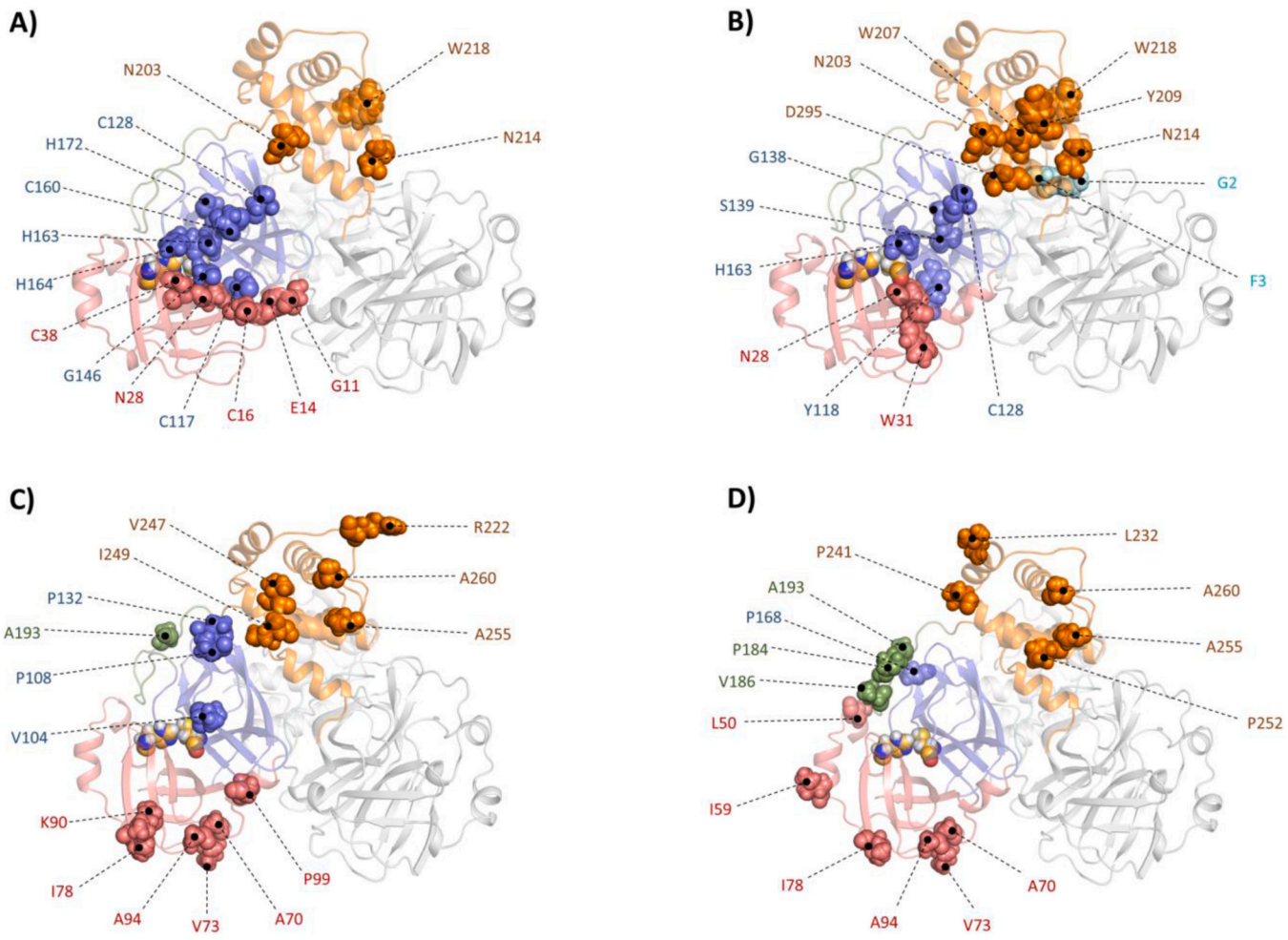

Fig. 6. Cold spot residues per ML-ANN models (bottom 5 %) (**A)** MATLAB ANN; **B)** Python ANN) and hot spot residues per ML-ANN models (top 5 %) (**C)** MATLAB ANN; **D)** Python ANN). The structural domains I (residues 10–99), II (residues 100–183) and III (residues 183–197) are coloured in pink, blue and orange, respectively. The N-finger is shown in cyan, and the linker loop is shown in green. The catalytic residues (H41 and C145) are coloured based on their atom types, where the carbon atoms are coloured light grey, oxygen atoms in red, nitrogen atoms in blue, sulfur atoms in yellow. The cold spots are depicted as spheres in **A–B**, while hot spots are depicted as spheres in **C-D**. Chain B is coloured grey.

the protein revealed that these hubs are located in key functional and core regions of the protein. Our literature review supports the finding that many of these residues are crucial for dimerization and catalytic activity. **(3)** Similar to cold spot identification, potential hot spot residues (bottom 5 % of residues with the lowest metric values) were also extracted. Of these, 80.4 % underwent mutation in at least one SARS-CoV-2 lineage. However, only three of these low-centrality residues (P96, A191, P184) overlapped with the top 5 % most mutated M^{pro} residues **(4)** We identified potential hot spots which were mainly located in the peripheral parts of the protein suggesting that DRN metrics can identify residues that are distant from structurally and functionally critical areas of the protein. This is supported by literature which shows that the SARS-CoV-2 M^{pro} has multiple allosteric sites distal to the active site linked to enzyme activity [95]. **(5)** Although the 5 % subset of data analysis was not statistically significant for all DRN metrics, the goal was to provide a snapshot of the data for comparison with existing literature. Importantly, each DRN metric showed a meaningful correlation with mutation frequencies in the overall dataset. While the R-values were relatively low, ranging from 0.22 to 0.30, the p-value tests indicated that all correlations were statistically significant. **(6)** We further compared the cold and hot spots with the residues that were least/most tolerant of mutation in the DMS. The cumulative frequency plot demonstrates that almost all hot spot residues (42/46) have WT-like activity, while almost all cold-spot residues (48/52) have impaired activity (either null-like or intermediate). We conclude that DRN is competent to identify residues that are mutation sensitive or mutation insensitive. However, it struggles to determine the severity of catalytic defect caused by mutation of a mutation-sensitive residue. **(7)** Machine learning was then used to construct predictors that combine the DRN data with other data, and it was found that the correlation to the mutation frequencies improves with R-values on the testing dataset of order 0.6. Furthermore, all models had Pearsons correlation p-values < 0.05 between the predicted and target data for the testing dataset implying great model prediction (Scikit learn ANN p-value: 1.93e-6, MATLAB ANN: 2.08e-2, MATLAB RF: 3.98e-3 and Python RF: 2.30e-5). **(8)** Integrating DRN with ML models significantly increased the prediction accuracy of most frequently mutated residues. We also observed close agreement between the ANN models in Python and MATLAB in terms of cold and hot spot residue distribution signifying consistency. Python ANN model predicted five while MATLAB ANN model predicted six including residue **P132**, the most mutated residue in the evolution of the protein. Altogether, the ML models predicted eight residues, demonstrating roughly 3-fold increase in prediction power. **(9)** Integration of DRN with ML models also increased the prediction power of cold spots. While DRN metrics identified ~ 50 non-mutated residues within the top bin that contained 102 residues with the highest metric values, both Python and MATLAB ANN models could identify 72 nonmutated residues.

While the combination of DRN metrics and ML models has improved

our ability to predict whether specific residues in M^{pro} will have a high mutation frequency in patient-derived samples, it is not perfect. There are some fundamental biological considerations that may limit the achievable accuracy of such predictions. Firstly, the enzymatic activity of M^{pro} may not be the only determinant of its contribution to virus fitness. In the crowded context of an infected cell, M^{pro} may need to interact with additional binding partners [96] or it may need to avoid spurious interactions with other proteins that would be detrimental to virus fitness. Secondly, if hot spot mutations are assumed to have neutral or only mild effects on M^{pro} function and to arise via genetic drift, then some mutations might be over- or underrepresented in populations due to founder effects [97] and genetic bottlenecks [98]. These potential pitfalls notwithstanding, DRN represents an orthogonal approach to identify residues that are structurally amenable to mutations. This could inform selection of promising leads from a panel of drug-like lead compounds, to avoid compounds that bind to mutation-prone residues, or might help identify residues with mutational rates far below that expected from DRN analysis, potentially highlighting residues with additional functional roles. Furthermore, in this study we showed that the DRN metrics identified residues neighbouring the most frequently mutated ones. Thus, with fine-tuning of the DRN calculations, we may be able to improve the predictive power.

Overall, this work lays a foundation for understanding protein evolution and holds promising potential for practical applications in drug discovery and pathogen evolution. By identifying and targeting residues that are less prone to mutations, this approach can guide the development of drugs with higher resistance barriers. While aiming to improve the predictive power of DRN analysis, our future work will also extend this approach to other proteins. We would expect that the DRN metric values are correlated to mutation frequencies, but a key question to investigate is whether the ML predictors developed here can be used unchanged to make reasonable predictions about the mutation frequencies of other proteins.

CRediT authorship contribution statement

Stephen C. Graham: Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition, Conceptualization. **Nigel T. Bishop:** Writing – original draft, Validation, Supervision, Methodology, Conceptualization. **Emily Morgan:** Writing – original draft, Visualization, Methodology, Formal analysis, Data curation. **Rabelani Ramahala:** Visualization, Formal analysis, Data curation. **Shrestha Chakraborty:** Writing – original draft, Validation, Formal analysis, Data curation. **Shaylyn Govender:** Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation. **Victor Barozi:** Writing – original draft, Visualization, Methodology, Formal analysis, Data curation. **Özlem Tastan Bishop:** Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization.

Funding sources

This project is supported by the Novo Nordisk Foundation and the Pandemic Antiviral Discovery (PAD) Initiative; Grant no.: NNF23SA0084504.

Abbreviations

ANN, artificial neural networks; AUC, area under the curve; B-factor, atomic displacements parameters; *BC*, *betweenness centrality*; *CC*, *closeness centrality*; DRN, dynamic residue network; DSM, deep scanning mutagenesis; *EC*, *eigenvector centrality*; *ECC*, *eccentricity*; FRET, Fluorescence resonance energy transfer; GISAID, Global Initiative on Sharing Avian Influenza Data; *KC*, *katz centrality*; *L*, *shortest path*; MD, molecular dynamics; ML, machine learning; M^{pro}, main protease; MSE, mean squared error; PME, Particle Mesh Ewald; *PR*, *pagerank*; RF, random forest; Rg, radius of gyration; RMSD, root mean square deviation; RMSF, root mean squared fluctuation; ROC, receiver operating characteristic; SARS-CoV-1, severe acute respiratory syndrome coronavirus 1; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; SASA, solvent accessible surface area; SEC, size exclusion chromatography; SVM, support vector machines; VOC, variants of concern; WT, wild type.

Declaration of Competing Interest

The authors declare no competing financial interests.

Acknowledgment

Authors acknowledge the use of Centre for High Performance Computing (CHPC), South Africa. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

Conflicts of interest

The authors declare that no conflicts of interest exist regarding the publication of this paper.

Requirements for Existing Software

The SARS-CoV-2 variants of concern sequences were retrieved from the GISAID database (https://gisaid.org/). The VOC mutations were identified from the sequences using the CoVsurver tool (https://gisaid.org/database-features/covsurver-mutations-app). The SARS-CoV-2 M^{pro} protein structure (PDB ID 5RFV [52]) was downloaded from the RCSB PDB website (https://www.rcsb.org/). PyMol (version 2.4) (free to download at https://www.pymol.org) was used to remove any non-protein molecules and to reconstitute the biological unit as chains A and B and prepare protein structure figures. The PROPKA tool version 3.5.1 (free to download: https://pypi.org/project/propka) was then to protonate the M^{pro} structure. The GROMACS software (version 2021.1) was used for the production simulations. It is free to download (https://manual.gromacs.org/documentation/2021.1/download.html). VMD version 1.9.3 (https://www.ks.uiuc.edu/Research/vmd/vmd-1.9.3/) was used for visualization of MD simulations. The GROMACS version 2020.1 was also used to calculate RMSD, RSMF, Rg, B-factor and SASA. MD-TASK (freely available at https://md-task.readthedocs.io/en/latest/home.html) was used for DRN calculations. Python version 3.10.13 (free to download: https://www.python.org/downloads/) was used for generating Python RF and ANN machine learning models. In Python, the free to download packages Keras version 3.0.5 (https://github.com/fchollet/keras), TensorFlow version 2.15.0 (https://pypi.org/project/tensorflow/) and Scikit-learn library version 0.24.2 (https://pypi.org/project/scikit-learn/) were used for model generation. MATLAB version 2023b was used with TreeBagger version for classification models and Deep Learning Toolbox for ANN models. Alluvial plots were generated in R version (4.3.2) free to download (https://cran.r-project.org/bin/windows/base/). MD simulations will be made available upon request.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2024.10.031.

References

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

[2] Olukitibi TA, Ao Z, Warner B, et al. Significance of conserved regions in coronavirus spike protein for developing a novel vaccine against SARS-CoV-2 infection. Vaccines 2023;11:545. https://doi.org/10.3390/vaccines11030545.

[3] Wu W-L, Chiang C-Y, Lai S-C, et al Monoclonal antibody targeting the conserved region of the SARS-CoV-2 spike protein to overcome viral variants. JCI Insight. Vol. 7, e157597. ⟨https://doi.org/10.1172/jci.insight.157597⟩.

[4] Ao D, He X, Liu J, Xu L. Strategies for the development and approval of COVID-19 vaccines and therapeutics in the post-pandemic period. Signal Transduct Target Ther 2023;8:1–17. https://doi.org/10.1038/s41392-023-01724-w.

[5] Markov PV, Ghafari M, Beer M, et al. The evolution of SARS-CoV-2. Nat Rev Microbiol 2023;21:361–79. https://doi.org/10.1038/s41579-023-00878-2.

[6] López-Cortés GI, Palacios-Pérez M, Veledíaz HF, et al. The spike protein of SARS-CoV-2 is adapting because of selective pressures. Vaccines 2022;10:864. https://doi.org/10.3390/vaccines10060864.

[7] Jaroszewski L, Iyer M, Alisoltani A, et al. The interplay of SARS-CoV-2 evolution and constraints imposed by the structure and functionality of its proteins. PLoS Comput Biol 2021;17:e1009147. https://doi.org/10.1371/journal.pcbi.1009147.

[8] Kandwal S, Fayne D. Genetic conservation across SARS-CoV-2 non-structural proteins – insights into possible targets for treatment of future viral outbreaks. Virology 2023;581:97–115. https://doi.org/10.1016/j.virol.2023.02.011.

[9] Shitrit A, Zaidman D, Kalid O, et al. Conserved interactions required for inhibition of the main protease of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Sci Rep 2020;10:20808. https://doi.org/10.1038/s41598-020-77794-5.

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

[11] Barozi V, Musyoka TM, Sheik Amamuddy O, Tastan Bishop Ö. Deciphering isoniazid drug resistance mechanisms on dimeric Mycobacterium tuberculosis KatG via post-molecular dynamics analyses including combined dynamic residue network metrics. ACS Omega 2022. https://doi.org/10.1021/acsomega.2c01036.

[12] Diessner EM, Takahashi GR, Cross TJ, et al. Mutation effects on structure and dynamics: adaptive evolution of the SARS-CoV-2 main protease. Biochemistry 2023;62:747–58. https://doi.org/10.1021/acs.biochem.2c00479.

[13] Chebon-Bore L, Sanyanga TA, Manyumwa CV, et al. Decoding the molecular effects of atovaquone linked resistant mutations on Plasmodium falciparum Cytb-ISP complex in the phospholipid bilayer membrane. Int J Mol Sci 2021;22:2138. https://doi.org/10.3390/ijms22042138.

[14] Punnatin P, Chanchao C, Chunsrivirot S. Molecular dynamics reveals insight into how N226P and H227Y mutations affect maltose binding in the active site of α-glucosidase II from European honeybee, Apis mellifera. PLoS One 2020;15: e0229734. https://doi.org/10.1371/journal.pone.0229734.

[15] Eisenmesser EZ, Millet O, Labeikovsky W, et al. Intrinsic dynamics of an enzyme underlies catalysis. Nature 2005;438:117–21. https://doi.org/10.1038/nature04105.

[16] Guarnera E, Berezovsky IN. Allosteric drugs and mutations: chances, challenges, and necessity. Curr Opin Struct Biol 2020;62:149–57. https://doi.org/10.1016/j.sbi.2020.01.010.

[17] Tee W-V, Guarnera E, Berezovsky IN. On the allosteric effect of nsSNPs and the emerging importance of allosteric polymorphism. J Mol Biol 2019;431:3933–42. https://doi.org/10.1016/j.jmb.2019.07.012.

[18] Barozi V, Edkins AL, Tastan Bishop Ö. Evolutionary progression of collective mutations in Omicron sub-lineages towards efficient RBD-hACE2: allosteric communications between and within viral and human proteins. Comput Struct Biotechnol J 2022. https://doi.org/10.1016/j.csbj.2022.08.015.

[19] Sheik Amamuddy O, Musyoka TM, Boateng RA, et al. Determining the unbinding events and conserved motions associated with the pyrazinamide release due to resistance mutations of Mycobacterium tuberculosis pyrazinamidase. Comput Struct Biotechnol J 2020;18:1103–20. https://doi.org/10.1016/j.csbj.2020.05.009.

[20] Miotto M, Olimpieri PP, Di Rienzo L, et al. Insights on protein thermal stability: a graph representation of molecular interactions. Bioinformatics 2018;35:2569–77. https://doi.org/10.1093/bioinformatics/bty1011.

[21] Prabantu VM, Naveenkumar N, Srinivasan N. Influence of disease-causing mutations on protein structural networks. Front Mol Biosci 2021;7. https://doi.org/10.3389/fmolb.2020.620554.

[22] Rodrigues CHM, Myung Y, Pires DEV, Ascher DB. mCSM-PPI2: predicting the effects of mutations on protein–protein interactions. Nucleic Acids Res 2019;47: W338–44. https://doi.org/10.1093/nar/gkz383.

[23] Rodrigues CH, Pires DE, Ascher DB. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic Acids Res 2018;46: W350–5. https://doi.org/10.1093/nar/gky300.

[24] Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet 2013;Chapter 7 (Unit7.20). https://doi.org/10.1002/0471142905.hg0720s76.

[25] Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 2003;31:3812–4.

[26] Buß O, Rudat J, Ochsenreither K. FoldX as protein engineering tool: better than random based approaches? Comput Struct Biotechnol J 2018;16:25–33. https://doi.org/10.1016/j.csbj.2018.01.002.

[27] Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 2005;33: W306–10. https://doi.org/10.1093/nar/gki375.

[28] Cheng J, Randall A, Baldi P. Prediction of protein stability changes for single-site mutations using support vector machines. Proteins 2006;62:1125–32. https://doi.org/10.1002/prot.20810.

[29] Masso M, Vaisman II. Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. Bioinform Oxf Engl 2008;24:2002–9. https://doi.org/10.1093/bioinformatics/btn353.

[30] Gelman S, Fahlberg SA, Heinzelman P, et al. Neural networks to learn protein sequence–function relationships from deep mutational scanning data. Proc Natl Acad Sci 2021;118:e2104878118. https://doi.org/10.1073/pnas.2104878118.

[31] Tsuchiya Y, Tomii K. Neural networks for protein structure and function prediction and dynamic analysis. Biophys Rev 2020;12:569–73. https://doi.org/10.1007/s12551-020-00685-6.

[32] Agajanian S, Oluyemi O, Verkhivker GM. Integration of random forest classifiers and deep convolutional neural networks for classification and biomolecular modeling of cancer driver mutations. Front Mol Biosci 2019;6:44. https://doi.org/10.3389/fmolb.2019.00044.

[33] Rentzsch P, Witten D, Cooper GM, et al. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res 2019;47:D886–94. https://doi.org/10.1093/nar/gky1016.

[34] Pejaver V, Urresti J, Lugo-Martinez J, et al. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. Nat Commun 2020;11:5918. https://doi.org/10.1038/s41467-020-19669-x.

[35] Lawrence MS, Stojanov P, Polak P, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013;499:214–8. https://doi.org/10.1038/nature12213.

[36] Díaz-Gay M, Vangara R, Barnes M, et al. Assigning mutational signatures to individual samples and individual somatic mutations with SigProfilerAssignment. bioRxiv 2023:2023.07.10.548264. https://doi.org/10.1101/2023.07.10.548264.

[37] Chakravarty D, Gao J, Phillips SM, et al. OncoKB: a precision oncology knowledge base. JCO Precis Oncol 2017 2017:PO.17.00011. https://doi.org/10.1200/PO.17.00011.

[38] Hatano N, Kamada M, Kojima R, Okuno Y. Network-based prediction approach for cancer-specific driver missense mutations using a graph neural network. BMC Bioinform 2023;24:383. https://doi.org/10.1186/s12859-023-05507-6.

[39] Gaudelet T, Day B, Jamasb AR, et al. Utilizing graph machine learning within drug discovery and development. Brief Bioinform 2021;22:bbab159. https://doi.org/10.1093/bib/bbab159.

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

[41] 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.

[42] Okeke CJ, Musyoka TM, Sheik Amamuddy O, et al. Allosteric pockets and dynamic residue network hubs of falcipain 2 in mutations including those linked to artemisinin resistance. Comput Struct Biotechnol J 2021;19:5647–66. https://doi.org/10.1016/j.csbj.2021.10.011.

[43] Sheik Amamuddy O, Afriyie Boateng R, Barozi V, et al. 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.

[44] Guo Y-R, Cao Q-D, Hong Z-S, et al. The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak – an update on the status. Mil Med Res 2020;7:11. https://doi.org/10.1186/s40779-020-00240-0.

[45] Singh D, Yi SV. On the origin and evolution of SARS-CoV-2. Exp Mol Med 2021;53: 537–47. https://doi.org/10.1038/s12276-021-00604-z.

[46] Li G, Hilgenfeld R, Whitley R, De Clercq E. Therapeutic strategies for COVID-19: progress and lessons learned. Nat Rev Drug Discov 2023;22:449–75. https://doi.org/10.1038/s41573-023-00672-y.

[47] Narayanan A, Narwal M, Majowicz SA, et al. Identification of SARS-CoV-2 inhibitors targeting Mpro and PLpro using in-cell-protease assay. Commun Biol 2022;5:1–17. https://doi.org/10.1038/s42003-022-03090-9.

[48] She Z, Yao Y, Wang C, et al. Mpro-targeted anti-SARS-CoV-2 inhibitor-based drugs. J Chem Res 2023;47:17475198231184799. https://doi.org/10.1177/17475198231184799.

[49] Shu Y, McCauley J. GISAID: global initiative on sharing all influenza data – from vision to reality. Eurosurveillance 2017;22:30494. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494.

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

[51] Flynn JM, Samant N, Schneider-Nachum G, et al. Comprehensive fitness landscape of SARS-CoV-2 Mpro reveals insights into viral resistance mechanisms. eLife 2022; 11:e77433. https://doi.org/10.7554/eLife.77433.

[52] Douangamath A, Fearon D, Gehrtz P, et al. Crystallographic and electrophilic fragment screening of the SARS-CoV-2 main protease. Nat Commun 2020;11:5047. https://doi.org/10.1038/s41467-020-18709-w.

[53] Dolinsky TJ, Nielsen JE, McCammon JA, Baker NA. PDB2PQR: an automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations. Nucleic Acids Res 2004;32:W665–7. https://doi.org/10.1093/nar/gkh381.

[54] Páll S, Zhmurov A, Bauer P, 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.

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

[56] Mark P, Nilsson L. Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K. J Phys Chem A 2001;105:9954–60. https://doi.org/10.1021/jp003020w.

[57] 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.

[58] Atilgan AR, Akan P, Baysal C. Small-world communication of residues and significance for protein dynamics. Biophys J 2004;86:85–91. https://doi.org/10.1016/S0006-3495(04)74086-2.

[59] Penkler DL, Atilgan C, Tastan Bishop Ö. Allosteric modulation of human Hsp90α conformational dynamics. J Chem Inf Model 2018;58:383–404. https://doi.org/10.1021/acs.jcim.7b00630.

[60] Liang Z, Verkhivker GM, Hu G. Integration of network models and evolutionary analysis into high-throughput modeling of protein dynamics and allosteric regulation: theory, tools and applications. Brief Bioinform 2020;21:815–35. https://doi.org/10.1093/bib/bbz029.

[61] Nachar N. The Mann-Whitney U: a test for assessing whether two independent samples come from the same distribution. Tutor Quant Methods Psychol 2008;4. https://doi.org/10.20982/tqmp.04.1.p013.

[62] Silva IN da, Spatti DH, Flauzino RA, et al. Artificial neural networks: a practical course. Springer; 2016.

[63] MathWorks – Makers of MATLAB and Simulink. ⟨https://uk.mathworks.com/⟩. [Accessed 18 March 2024]..

[64] Python Release Python 3.10.0. In: Python.org. ⟨https://www.python.org/downloads/release/python-3100/⟩. [Accessed 25 May 2024].

[65] Kapoor A, Gulli A, Pal S, Chollet F. Deep Learning with TensorFlow and Keras: build and deploy supervised, unsupervised, deep, and reinforcement learning models. Packt Publishing Ltd; 2022.

[66] Abadi M, Agarwal A, Barham P, et al. TensorFlow: large-scale machine learning on heterogeneous distributed systems. 2016.

[67] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825–30.

[68] Collaborative Statistics, Connexions – Google Search. ⟨https://www.google.com/search?q=Collaborative+Statistics%2C+Connexions&client=firefox-b-d&sca_esv=c768f71cdf49a14c&sca_upv=1&sxsrf=ADLYWIIECmDIc7yM84xqMZe3rUeBn9frJA%3A1727438320647&ei=8J32ZtCXJ_m6hbIPieTFoQ8&ved=0ahUKEwjQz-SbieOIAxV5XUEAHQlyMfQQ4dUDCA8&uact=5&oq=Collaborative+Statistics%2C+Connexions&gs_lp=Egxnd3Mtd2l6LXNlcnAiJENvbGxhYm9yYXRpdmUgU3RhdGlzdGljcywgQ29ubmV4aW9uczIEECMYJzIIEAAYgAQYogQyCBAAGIAEGKIEMggQABiABBiiBEjYVlCUU1iUU3ACeACQAQCYAWmgAWmqAQMwLjG4AQPIAQD4AQL4AQGYAgGgAm6YAwCIBgGSBwMwLjGgB6UE&sclient=gws-wiz-serp⟩. [Accessed 27 September 2024]..

[69] Breiman L. Random forests. Mach Lang 2001;45:5–32. https://doi.org/10.1023/A:1010933404324.

[70] Works M. TreeBagger. 2016.

[71] Chen S, Zhang J, Hu T, et al. Residues on the dimer interface of SARS coronavirus 3C-like protease: dimer stability characterization and enzyme catalytic activity analysis. J Biochem 2008;143:525–36. https://doi.org/10.1093/jb/mvm246.

[72] Hsu W-C, Chang H-C, Chou C-Y, et al. Critical assessment of important regions in the subunit association and catalytic action of the severe acute respiratory syndrome coronavirus main protease. J Biol Chem 2005;280:22741–8. https://doi.org/10.1074/jbc.M502556200.

[73] Chou C-Y, Chang H-C, Hsu W-C, et al. Quaternary structure of the severe acute respiratory syndrome (SARS) coronavirus main protease. Biochemistry 2004;43:14958–70. https://doi.org/10.1021/bi0490237.

[74] Lis K, Plewka J, Menezes F, et al. SARS-CoV-2 Mpro oligomerization as a potential target for therapy. Int J Biol Macromol 2024;267:131392. https://doi.org/10.1016/j.ijbiomac.2024.131392.

[75] Ferreira JC, Fadl S, Rabeh WM. Key dimer interface residues impact the catalytic activity of 3CLpro, the main protease of SARS-CoV-2. J Biol Chem 2022;298:102023. https://doi.org/10.1016/j.jbc.2022.102023.

[76] Chen S, Hu T, Zhang J, et al. Mutation of Gly-11 on the dimer interface results in the complete crystallographic dimer dissociation of severe acute respiratory syndrome coronavirus 3C-like protease: crystal structure with molecular dynamics simulations. J Biol Chem 2008;283:554–64. https://doi.org/10.1074/jbc.M705240200.

[77] Iketani S, Hong SJ, Sheng J, et al. Functional map of SARS-CoV-2 3CL protease reveals tolerant and immutable sites. Cell Host Microbe 2022;30:1354–1362.e6. https://doi.org/10.1016/j.chom.2022.08.003.

[78] Lee J, Worrall LJ, Vuckovic M, et al. Crystallographic structure of wild-type SARS-CoV-2 main protease acyl-enzyme intermediate with physiological C-terminal autoprocessing site. Nat Commun 2020;11:5877. https://doi.org/10.1038/s41467-020-19662-4.

[79] Flynn JM, Huang QYJ, Zvornicanin SN, et al. Systematic analyses of the resistance potential of drugs targeting SARS-CoV-2 main protease. ACS Infect Dis 2023;9:1372–86. https://doi.org/10.1021/acsinfecdis.3c00125.

[80] Hu T, Zhang Y, Li L, et al. Two adjacent mutations on the dimer interface of SARS coronavirus 3C-like protease cause different conformational changes in crystal structure. Virology 2009;388:324–34. https://doi.org/10.1016/j.virol.2009.03.034.

[81] Barrila J, Bacha U, Freire E. Long range cooperative interactions modulate dimerization in SARS 3CLpro. Biochemistry 2006;45:14908–16. https://doi.org/10.1021/bi0616302.

[82] Cheng S-C, Chang G-G, Chou C-Y. Mutation of Glu-166 blocks the substrate-induced dimerization of SARS coronavirus main protease. Biophys J 2010;98:1327–36. https://doi.org/10.1016/j.bpj.2009.12.4272.

[83] Lim L, Shi J, Mu Y, Song J. Dynamically-driven enhancement of the catalytic machinery of the SARS 3C-like protease by the S284-T285-I286/A mutations on the extra domain. PLoS One 2014;9:e101941. https://doi.org/10.1371/journal.pone.0101941.

[84] Parmar M, Thumar R, Patel B, et al. Structural differences in 3C-like protease (Mpro) from SARS-CoV and SARS-CoV-2: molecular insights revealed by Molecular Dynamics Simulations. Struct Chem 2022:1–18. https://doi.org/10.1007/s11224-022-02089-6.

[85] Chen SA, Arutyunova E, Lu J, et al. SARS-CoV-2 Mpro protease variants of concern display altered viral substrate and cell host target galectin-8 processing but retain sensitivity toward antivirals. ACS Cent Sci 2023;9:696–708. https://doi.org/10.1021/acscentsci.3c00054.

[86] Jm F, Sn Z, T T, et al. Contributions of hyperactive mutations in Mpro from SARS-CoV-2 to drug resistance. ACS Infect Dis 2024;10. https://doi.org/10.1021/acsinfecdis.3c00560.

[87] Jochmans D, Liu C, Donckers K, et al. The substitutions L50F, E166A, and L167F in SARS-CoV-2 3CLpro are selected by a protease inhibitor in vitro and confer resistance to nirmatrelvir. mBio 2023;14:e0281522. https://doi.org/10.1128/mbio.02815-22.

[88] Zhou Y, Gammeltoft KA, Ryberg LA, et al. Nirmatrelvir-resistant SARS-CoV-2 variants with high fitness in an infectious cell culture system. Sci Adv 2022;8:eadd7197. https://doi.org/10.1126/sciadv.add7197.

[89] Noske GD, de Souza Silva E, de Godoy MO, et al. Structural basis of nirmatrelvir and ensitrelvir activity against naturally occurring polymorphisms of the SARS-CoV-2 main protease. J Biol Chem 2023;299(3):103004. https://doi.org/10.1016/j.jbc.2023.103004.

[90] Hu Y, Lewandowski EM, Tan H, et al. Naturally occurring mutations of SARS-CoV-2 main protease confer drug resistance to nirmatrelvir. ACS Cent Sci 2023;9:1658–69. https://doi.org/10.1021/acscentsci.3c00538.

[91] Sasi VM, Ullrich S, Ton J, et al. Predicting antiviral resistance mutations in SARS-CoV-2 main protease with computational and experimental screening. Biochemistry 2022;61:2495–505. https://doi.org/10.1021/acs.biochem.2c00489.

[92] Günther S, Reinke PYA, Fernández-García Y, et al. X-ray screening identifies active site and allosteric inhibitors of SARS-CoV-2 main protease. Science 2021;372:642–6. https://doi.org/10.1126/science.abf7945.

[93] Sheik Amamuddy O, Musyoka TM, Boateng RA, et al. Determining the unbinding events and conserved motions associated with the pyrazinamide release due to resistance mutations of Mycobacterium tuberculosis pyrazinamidase. Comput Struct Biotechnol J 2020;18:1103–20. https://doi.org/10.1016/j.csbj.2020.05.009.

[94] Sheik Amamuddy O, Bishop NT, Tastan Bishop Ö. Characterizing early drug resistance-related events using geometric ensembles from HIV protease dynamics. Sci Rep 2018;8:17938. https://doi.org/10.1038/s41598-018-36041-8.

[95] El-Baba TJ, Lutomski CA, Kantsadi AL, et al. Allosteric inhibition of the SARS-CoV-2 main protease: insights from mass spectrometry based assays*. Angew Chem Int Ed Engl 2020;59:23544–8. https://doi.org/10.1002/anie.202010316.

[96] Gordon DE, Jang GM, Bouhaddou M, et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature 2020;583:459–68. https://doi.org/10.1038/s41586-020-2286-9.

[97] Rambaut A, Posada D, Crandall KA, Holmes EC. The causes and consequences of HIV evolution. Nat Rev Genet 2004;5:52–61. https://doi.org/10.1038/nrg1246.

[98] Bendall EE, Callear AP, Getz A, et al. Rapid transmission and tight bottlenecks constrain the evolution of highly transmissible SARS-CoV-2 variants. Nat Commun 2023;14:272. https://doi.org/10.1038/s41467-023-36001-5.