BMC Bioinformatics

RESEARCH Open Access

Machine learning framework for cost effective deep mutational scanning through targeted substitution profiling

Emily Morgan[1], Shaylyn Govender[1], Prashant Singh[2], Ian Goodfellow[3], Stephen C. Graham[3*], Nigel T. Bishop[4,5*] and Özlem Tastan Bishop[1,5*]

*Correspondence:
Stephen C. Graham
scg34@cam.ac.uk
Nigel T. Bishop
n.bishop@ru.ac.za
Özlem Tastan Bishop
o.tastanbishop@ru.ac.za

Full list of author information is available at the end of the article

Abstract

Background Deep mutational scanning (DMS) provides comprehensive maps of protein variant effects but remains experimentally intensive. Machine learning (ML) approaches have the potential to reduce experimental burden of DMS by predicting the functional impact of substitutions from limited data.

Results We introduced a ML classifier trained on normalised DMS scores from SARS-CoV-2 main protease (Mpro) to categorise amino acid substitutions as functional (wild-type-like) or non-functional. Using brute-force feature selection, we identified minimal subsets of six substitution scores per residue that enable accurate classification of the remaining substitutions, achieving minimum (worst accuracy) scores exceeding 90%. Models including support vector machines, random forests, and logistic regression were evaluated without retraining (zero-shot prediction) against additional SARS-CoV-2 Mpro datasets and against unrelated datasets. The zero-shot performance of the models was strongest for other enzymes and more modest when applied to DMS systems that assess protein folding and/or protein-protein interactions.

Conclusion The results show that targeted DMS combined with ML can reduce sequencing and reagent costs while preserving classification accuracy, offering a practical route to accelerate variant effect prediction.

Keywords Brute-force approach, Deep mutational scanning, Random Forest, Support vector machine, Zero-shot prediction

Background

Deep mutational scanning (DMS) is a high-throughput experimental approach that systematically measures the effects of a large number of protein variants, enabling functional mapping of protein sequence space at single-residue resolution [1]. In a typical DMS experiment, nearly all single amino acid substitutions (and sometimes multiple concurrent substitutions) are introduced into a protein-coding sequence, producing a mutant library. The mutant sequences are inserted into a suitable expression system in place of the wild-type gene, and the mutant library amplified in a suitable host to

produce a large, diverse pool of constructs [2]. The mutant library is then expressed in a model system, where the functional consequences of each variant are assessed using assays tailored to the protein's function such as enzymatic activity or binding affinity, or via proxy measures like protein expression or organism growth rate [1]. High-throughput sequencing is used to determine the frequency of each variant before and after selection, enabling the calculation of functional or fitness scores for thousands of mutations in parallel.

DMS has been applied in many biological contexts when investigating functional effects of protein variants [1]. For example, it has classified missense mutations in human proteins as either pathogenic or benign, allowing identification of pathogenic variants [3]. Similarly, DMS has been applied to protein fitness landscapes, showing quantitatively how sequence changes affect protein function [4], as well as revealing residues critical for folding, stability, or functional interactions [5, 6].

Machine learning (ML) approaches increasingly use DMS data to predict the functional consequences of protein variants. Sruthi and Prakash investigated strategies for selecting minimal subsets of DMS data such as by percentage and mutation type, in an attempt to identify rational methods for subset identification and to reduce experimental cost [7]. The sequence, structural, and coevolutionary features of the minimal subsets were used to train regression neural network models, predicting untested mutation fitness scores to reconstruct the full DMS landscape of the protein. In a related study, Sruthi, Balaram and Prakash established simple conservation-based thresholds for six physicochemical and evolutionary features to classify mutations as neutral or deleterious [8]. This was performed to identify informative features for classification, and to produce simple, interpretable rules for predicting mutation effects.

The DeMaSk framework was created by Munro et al., which constructed a new amino acid substitution matrix, encoding substitution effects across a range of proteins based on DMS data [9]. The corresponding substitution matrix values, along with conservation values and variant amino acid frequencies across homologs were provided as inputs to a linear model, predicting delta fitness scores per substitution for unseen proteins. Fu et al. incorporated low-throughput mutagenesis data, such as alanine scanning (AS), alongside other protein features, and found this improved variant prediction in certain datasets [10]. Linear regression models were trained with leave-one-protein-out cross-validation, where the Spearman Rank correlation performance metric was calculated as the correlation between the substitution score predictions and the actual DMS scores.

Multiple neural network model types were compared by Gelman et al. on their ability to learn sequence-function relationships from combinations of encoded physicochemical and structural representations, predicting functional DMS scores for unseen variants [11]. Sarfati et al. evaluated six different ML model types on their ability to predict DMS experimental functional outcome values, such as fluorescence or stability properties [12]. A subset of 40 sequence and structural features were provided based on their importance determined from the LASSO (Least Absolute Shrinkage and Selection Operator) model. The correlation between predicted and actual values was assessed, and better-than-random predictions were identified.

DMS datasets have been used to benchmark variant effect predictors and assess their clinical relevance [13–15], evaluating multiple predictors across DMS datasets in diverse

organisms. Also, DMS data has revealed that stability predictors can partially explain functional impacts of mutations [16].

DMS datasets have also been widely utilised in structural and viral contexts. Drake et al. used single-mutant DMS data as inputs to improve protein structure prediction via deep learning [17]. More recently, convolutional neural networks incorporating local environment, sequence and physicochemical features have been trained and tested on SARS-CoV-2 Spike receptor-binding domain (RBD) DMS data to predict ACE2 binding affinity changes [18]. Transfer learning was applied to forecast the effects of multiple mutations upon ACE2 binding for major SARS-CoV-2 variants of concern.

Most frameworks applied to DMS data typically predict variant effects using various supervised ML models. DMS scores measured by function- and protein-specific assays are provided as targets and predicted by combinations of a large variety of features, often with structural, evolutionary and physicochemical properties. Only Munro et al. incorporated DMS scores from independent datasets as an indirect feature through their DMS substitution matrix [9], while Fu et al. included encoded alanine scanning measurements from high-compatibility assays of the tested protein [10]. However, these scores are determined from experimental data from different proteins or assays. Our approach differs from existing DMS-based prediction methods by choosing minimal subsets of experimentally measured substitutions per residue within the same DMS experiment that can accurately predict the remaining substitutions for that residue within a protein.

This choice results in each data sample corresponding to a residue, rather than an individual substitution. This reduces the dataset size substantially, making simpler supervised ML models, namely Random Forest and Support Vector Machines, more suitable than more complex models such as neural networks employed in other research [7, 11–12, 18].

This technique requires the identification of minimal subsets of DMS data that will need to be determined experimentally. To our knowledge, no existing method has been published that has addressed this subset-selection problem in this context. Therefore, a novel, brute-force approach was developed to identify minimal subsets.

Using this approach, we determined that 6-DMS substitution scores per residue are sufficient to predict the remaining substitution scores with approximately 90% accuracy (based on the F1 score), while classifying mutations as either wild-type or loss-of-function. This implies that only seven substitutions per residue need to be measured using the experimental assay (six mutations plus the stop codon), instead of all nineteen substitutions plus stop. This reduces the total number of variants to be generated and experimentally assessed, thus reducing the overall resource burden by simplifying library generation and reducing required sequencing depth. Furthermore, our approach allows for zero-shot prediction, without retraining, across independent DMS datasets that use comparable experimental techniques, where varying performance was achieved.

Methods

Classification ML models were trained to predict either wild-type function or loss-of-function for a given amino acid substitution, using a smaller set of DMS substitution scores as features. Brute-force analysis was applied to identify the optimal combinations of DMS substitution scores to predict the remaining scores for each residue. Subsequently, we evaluated the DMS substitution score feature combinations using zero-shot

prediction when applied to a variety of DMS datasets. A summary of the full workflow is shown in Fig. 1.

Part 1: Brute-Force combination selection

Dataset preparation

The chosen dataset for initial combination selection was produced by Flynn et al. [19], containing SARS-CoV-2 main protease (Mpro) normalised DMS scores per substitution from the first replicate of the FRET assay. This dataset will be referred to as the Mpro FRET replicate 1 dataset. A total of 16 residues with missing substitution values were removed, as all twenty substitution values are required. In addition, the last two Mpro residues (305 and 306) were excluded as they were absent from molecular dynamics simulation data used previously [20, 21], from which some initial features were derived. This resulted in a dataset of 288 records, each with twenty substitution values.

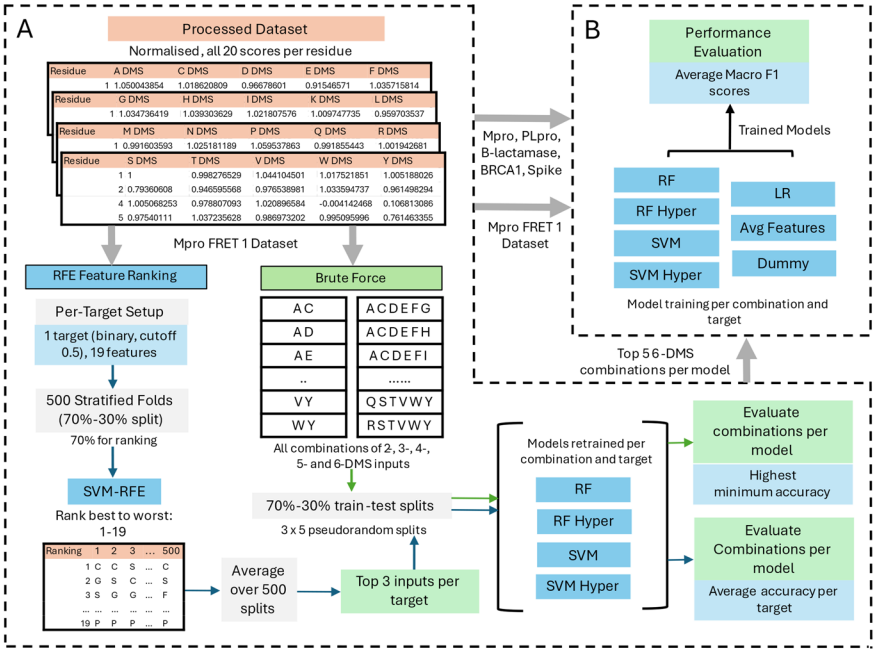

Fig. 1 Methodological workflow for DMS selection and zero-shot mutation effect prediction. **A** Feature selection and model training using the Mpro FRET replicate 1 dataset. Normalised DMS substitution scores were used as input, with all twenty substitution DMS values for each residue. Two separate feature-selection strategies were applied. In the Support Vector Machine Recursive Feature Elimination (SVM-RFE) approach, for each substitution, the remaining substitutions were used as candidate features to predict wild-type or loss-of-function behaviour. The dataset was split into 500 stratified folds (70–30%), and RFE using a linear SVM was applied separately for each target substitution to rank the nineteen candidate features. Feature rankings were averaged across folds, and the top three substitutions per target were selected. A brute-force approach evaluated all possible 2-, 3-, 4-, 5-, and 6-DMS substitution combinations to use as input features, predicting wild-type or loss-of-function behaviour for each of the remaining substitutions. For both approaches, Random Forest and SVM classifiers, with and without hyperparameter tuning, were trained independently for each feature set, target substitution, and dataset split. The resultant average accuracies for the SVM-RFE method and highest minimum scores for the brute-force approach were identified. **B** Zero-shot evaluation on independent DMS datasets. Seven models trained on the Mpro FRET replicate 1 dataset using the top brute-force 6-DMS combinations were applied to multiple independent experimental DMS datasets without retraining. Model performance was shown using average macro F1 scores across predicted substitutions

Initial feature selection

From twenty potential substitutions, the aim was to select a small subset as features that could accurately predict the residue behaviour for the remaining substitutions as wild-type or loss-of-function, defined as mutations retaining ≥ 50% or < 50% of wild-type activity, respectively. Within the dataset, each substitution was treated as the target, with the remaining nineteen DMS substitutions used as features. The dataset was split into 500 stratified folds (using fixed seeds 0–499), where 70% of the data is used for ranking, and the remaining 30% is not used in that fold. Recursive Feature Elimination (RFE) using a Support Vector Machine (SVM) with a linear kernel and default parameters, implemented using the scikit-learn Python package [22], was applied to each train-test split to rank the nineteen features from most to least important. For each target substitution, the top DMS features (optimally three) with the highest ranking when averaged over all runs were selected. To assess whether the 500 stratified folds captured sufficient variability in feature ranking, RFE was rerun using completely random stratified folds, and the majority of top-ranked feature combinations remained identical, shown in Supplementary Table 1.

These top three substitutions were provided as features to Random Forest (RF) and SVM binary classifiers to predict the corresponding target substitution. RF was chosen for its built-in feature selection and robustness as an ensemble model, while SVM was selected for its ability to handle high-dimensional, non-linear spaces. A fixed model seed of 42 was chosen to ensure reproducibility during training of these RF and SVM models.

Model preparation

Each model was trained initially with default parameters and then with hyperparameter tuning. Manual tuning was applied to reduce overfitting, as high variability was observed between dataset splits when using GridSearchCV or *k*-fold cross-validation for the RF model. The most stable results were obtained for the SVM model when the parameters *C*, *degree*, and *coef0* were manually tuned. Then, GridSearchCV was used to select the *kernel* and *gamma* hyperparameters for each substitution prediction. The final hyperparameter values are shown in Supplementary Tables 2 and 3.

Feature selection using brute-force approach

No clear subset of DMS substitution scores was identified through RFE that could accurately predict all other scores. Therefore, a (computationally intensive) brute-force approach was applied, where each combination of 2-, 3-, 4-, 5- and 6-DMS substitution scores was used as features to predict the remaining substitutions. The same two model types with and without hyperparameter tuning were used, where all previously selected hyperparameters were kept constant during this analysis.

The number of substitutions used as features was denoted as *n*. Each model was used to predict one target substitution for all residues. Therefore, the same model was run (20-*n*) times to predict each of the targets for a combination. The total number of combinations is given by:

$$C_n^{20} = \frac{20!}{n!\,(20-n)!}$$

where the total number of amino acids is 20 and the number used as features is n. The model was then run for each unique combination of features, over the two model types with and without hyperparameter tuning, with five pseudorandom dataset splits generated by a fixed seed list. Dataset partitioning was controlled by using predefined seeds, and three independent seed lists were used to assess robustness to dataset partitioning (Seed list 1: 493, 121, 334, 158, 397; Seed list 2: 189, 480, 45, 173, 362; Seed list 3: 49, 402, 377, 350, 113).

For 2-DMS combinations, this equates to (190 combinations × 18 targets × 5 splits × 3 seed lists × 4 models) models trained and tested, and (38760 combinations × 14 targets × 5 splits × 3 seed lists × 4 models) models for the 6-DMS combinations. Training and testing of the 2-, 3- and 4-DMS combinations were performed on a single-core CPU, while the 5- and 6-DMS combination runs were parallelised across 8 processes using Python's *ProcessPoolExecutor*. The actual per-combination wall-clock times were recorded during execution.

Model training and evaluation

Due to the small dataset size, a 70–30% train-test split was employed, forgoing the validation set. With only 288 residues in the dataset and the combinatorial nature of the feature search space, nested cross-validation was not employed for brute-force feature selection. The exhaustive evaluation of feature combinations already required repeated training and testing across multiple predefined train-test splits and seed lists, resulting in many model evaluations per combination. Introducing nested cross-validation would have substantially increased computational cost and further reduced the data points across inner and outer folds, increasing variability in feature selection.

The *MinMaxScaler* from scikit-learn was applied to normalise the input features between zero and one. For each target substitution, a cutoff of 0.5 was chosen, where any values equal or higher were classified as wild-type class (1), with the remainder as loss-of-function effect (0). The DMS scores were strongly bimodal, featuring a valley between modes (see Supplementary Fig. 1A), with the cutoff 0.5 lying approximately at the center of the valley. Stratified sampling was applied during the dataset splitting due to class imbalance, where some target DMS substitutions consisted of a higher proportion of residues with wild-type behaviour.

To mitigate the inherent variation in performance between the train-test splits for this small dataset, each model was trained on five different pseudorandom splits, each producing an accuracy value for a given target substitution. Accuracy was calculated as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where *TP*, *TN*, *FP*, and *FN* represent true positives, true negatives, false positives, and false negatives, respectively. The mean and sample standard deviation accuracy value for each model was obtained from these accuracy values. This was repeated two more times to determine if overall model performance remained relatively stable between dataset splits.

Choice of brute-force top DMS substitution combinations

The minimum score per tested feature combination, corresponding to the lowest mean test accuracy value for a target using that combination, was identified for each model. The five DMS substitution feature combinations with the highest minimum scores across the first five pseudorandom dataset splits were selected for each of the two model types, with and without hyperparameter tuning. This was performed to ensure that the most difficult remaining substitution targets were still reliably predicted. In addition, the five substitution combinations per model with the lowest minimum scores were identified, to demonstrate the least possible accuracy achieved by the model with the provided features.

Part 2: zero-shot prediction for other DMS score datasets

Dataset preparation

Additional independent experimental DMS score datasets from multiple proteins were chosen for testing DMS substitution score combinations. Each dataset was required to meet three criteria to ensure suitability for zero-shot prediction. First, the dataset had to contain DMS scores for each of the twenty possible amino acid substitutions per residue. Second, the DMS substitution scores needed to be normalised, where the average wild-type score is approximately 1, and preferably the average nonsense or stop-codon score is equal to 0. Third, the DMS substitution scores had to be amenable to classification as either wild-type or loss-of-function.

The datasets selected for downstream analysis included the SARS-CoV-2 Mpro functional scores (FRET, TF and Growth assays) [19], SARS-CoV-2 papain-like protease (PLpro) fitness activity scores [23], SARS-CoV-2 Spike RBD expression scores in yeast [24], *Escherichia coli* TEM-1 β-lactamase ampicillin resistance scores [25] and human BRCA1 RING domain E3 ligase activity scores [26]. All datasets consisted of normalised scores with the average wild-type score approximately 1, except for SARS-CoV-2 Spike RDB which was subsequently processed by normalising the average stop-codon scores and wild-type scores. All residues within these datasets with at least one missing substitution score were removed, as performed for the Mpro FRET replicate 1 dataset.

Model application

Each of the models used in Part 1, along with an additional three models (described below), was trained on each of the identified top 6-DMS combinations for all residues in the Mpro FRET replicate 1 dataset. The first additional model, a Logistic Regression model from scikit-learn with default parameters, was included to assess whether comparable performance could be achieved by a simpler, linearly separable model. The second additional model was an averaging-based classifier that predicted the target class based on the mean values of the feature values classified using the 0.5 cutoff. This model was included to evaluate whether the relationships between the selected features and target can be determined accurately without a complex ML model. Lastly, a *DummyClassifier* from scikit-learn with the most frequent strategy was chosen as a control to ensure that the performance metrics for the models were not biased by datasets with imbalanced classes. As the DMS substitution scores were already normalised, scaling using *MinMaxScaler* was removed. These models were evaluated on the remaining experimental DMS score datasets using zero-shot prediction.

An experimental baseline was calculated using classifications from FRET replicate 1 as predictions for the remaining Mpro datasets. For each of the top 6-DMS combinations, the remaining 14 substitution outcomes for a given residue in the other Mpro datasets was assigned directly as the value of the corresponding residue from Mpro FRET replicate 1.

Model performance evaluation

Model performance for each feature combination was characterised by the averaged macro F1 scores for the 14 predicted targets. Macro F1 score was calculated as:

$$F1_{\text{macro}} = \frac{1}{N}\sum_{i=1}^{N} 2 \cdot \frac{\text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}$$

where N is the number of classes (2 in this case, wild-type and loss-of-function), $\text{Precision}_i = \frac{TP_i}{TP_i+FP_i}$ is the fraction of predicted positives that are true positives and $\text{Recall}_i = \frac{TP_i}{TP_i+FN_i}$ is the fraction of true positives that are correctly identified. The average macro F1 score accounts equally for the two target classes, which reduces the likelihood of models appearing to have high performance when only predicting the majority class. In addition, it is a harmonic mean of precision and recall, meaning it equally penalises false negatives and positives, both of which are important in this context. The sample standard deviation over all predicted substitutions per combination was also calculated to quantify the variability in macro F1 scores across targets.

Part 3: software and implementation details

All analyses were performed in Python 3.10.13, with any notebooks run using the JupyterLab 4.3.2 environment. This workflow relied on several scientific Python libraries, including NumPy 1.26.4, Pandas 2.2.3, SciPy 1.14.1 and scikit-learn 1.6.1 for ML model implementation and evaluation. Matplotlib 3.9.3 and Seaborn 0.13.2 were used for data visualisation, while Openpyxl 3.1.5 and xlrd 2.0.2 supported the reading of spreadsheets holding input DMS dataset values. All software dependencies are additionally provided in a supplementary YAML environment file for reproducibility.

The brute-force computations were performed on a high-performance computing cluster using a node equipped with an AMD EPYC 7543P 32-core processor. Individual jobs were executed on a single CPU core, with parallelisation implemented across 8 processes. Memory resources were not explicitly requested per job, so a minimum of 32 GB RAM of the 156 GB available was allocated on the node. All remaining computations were performed on a personal Linux-based laptop equipped with an Intel Core Ultra 7 155 H processor and 32 GB RAM. These analyses were executed on a single CPU core without parallelisation.

Results

Effect of number of substitutions on predictive performance

Predictive performance increased with the number of substitutions provided as input (2, 3, 4, 5 and 6) to the brute-force search, with the highest mean of the average minimum accuracies across the top combinations when using 6 substitutions (Fig. 2). As expected, the runtime of the models increased with the number of features as shown in Supplementary Table 4. When parallelising over eight cores, the 6-DMS combination runtime

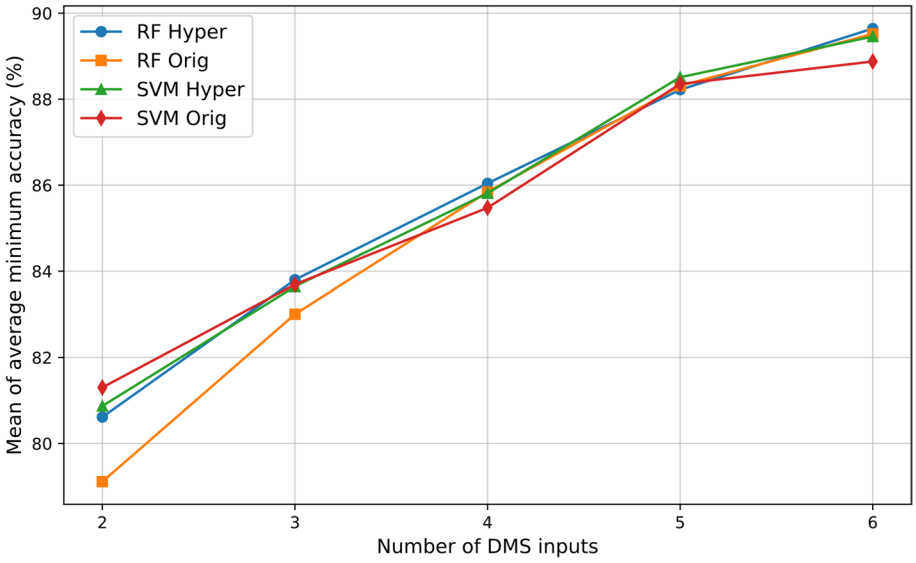

Fig. 2 Predictive performance of four machine learning models. Performance is represented by the mean of the average minimum accuracies for the top five DMS score input combinations per model. These models are Random Forest with hyperparameter tuning (RF Hyper), Random Forest with default hyperparameters (RF Orig), Support Vector Machine with hyperparameter tuning (SVM Hyper), and Support Vector Machine with default hyperparameters (SVM Orig)

was 41.6 days. Additional reasons for limiting the search to 6 substitutions are presented in the Discussion section.

Top substitution combination identification

The top five 6-DMS substitution combinations from the brute-force approach are summarised in Table 1, while the top five 2-, 3-, 4- and 5-DMS combinations for each model type are shown in Supplementary Tables 5–8. The scores corresponding to substitutions to glycine and proline were consistently selected as features in the brute-force 5- and 6-DMS combinations for predicting the remaining substitutions. However, only the proline DMS substitution scores were frequently in the top five 2-, 3-, and 4-DMS combinations per model (Supplementary Tables 5–8). From three features onwards, proline is always present in the selected feature sets.

The pattern of proline and glycine substitutions always being chosen in the 6-DMS combinations warranted further investigation. Therefore, for each target substitution, the prediction accuracy obtained using the top three RFE-selected features was compared to the frequency with which that target appeared in the brute-force combinations for the default SVM model (Supplementary Table 9). As expected, glycine and proline were the worst-predicted targets and were always included in the 6-DMS combinations. However, for all other substitutions, there was no clear correlation between target prediction accuracy and the frequency of that target appearing in the brute-force combinations. Finally, the total frequencies of the default SVM brute-force combinations for the three features identified by RFE were compared against the target accuracies shown in Supplementary Table 9. However, there did not appear to be any patterns in this data.

The remaining substitutions included in the top 6-DMS combinations represent residues with different physicochemical properties, such as polar residues with side chain amide, indole or hydroxyl groups (asparagine, histidine, glutamine, threonine, tyrosine),

Table 1 Highest minimum scores for combinations of 6-DMS substitution scores across ML models

Model	Substitution combination	Highest minimum score	Average minimum score
RF_hyper	E, F, G, H, P, T	90.57 ± 3.49	89.12 ± 4.00
	G, N, P, S, V, Y	90.34 ± 3.00	89.43 ± 3.01
	E, G, N, P, V, Y	90.11 ± 2.65	**90.19 ± 2.60**
	E, F, G, H, P, V	90.11 ± 3.87	89.81 ± 4.17
	E, G, M, P, T, Y	90.11 ± 3.11	89.66 ± 3.28
RF_orig	C, D, F, G, P, Q	90.57 ± 2.97	89.58 ± 2.96
	C, D, G, P, Q, Y	90.34 ± 2.38	89.43 ± 2.26
	C, E, F, G, H, P	90.11 ± 3.69	**89.73 ± 4.22**
	C, D, F, G, P, S	90.11 ± 3.00	89.58 ± 2.32
	C, E, G, N, P, Y	90.11 ± 3.87	89.27 ± 2.52
SVM_hyper	G, N, P, Q, V, Y	90.80 ± 2.69	**89.73 ± 2.99**
	E, F, G, P, Q, V	90.80 ± 3.81	89.43 ± 3.82
	C, D, G, N, P, Y	90.80 ± 2.57	89.35 ± 3.02
	E, F, G, N, P, V	90.34 ± 3.31	89.50 ± 2.52
	C, G, H, P, T, Y	90.34 ± 1.92	89.27 ± 2.77
SVM_orig	C, D, F, G, P, Q	90.34 ± 1.54	**89.81 ± 2.25**
	G, H, P, T, V, Y	90.34 ± 3.95	88.97 ± 4.30
	C, F, G, H, I, P	90.34 ± 2.09	88.97 ± 2.60
	C, F, G, L, P, Q	90.34 ± 1.74	88.43 ± 3.24
	C, F, G, P, Q, V	90.34 ± 2.52	88.20 ± 3.37

The minimum score refers to the lowest mean test set accuracy for a target using that combination

The table shows the highest minimum scores obtained from combinations of six DMS substitution scores for four machine learning models. These models are hyperparameter-tuned RF (RF_hyper), RF with default parameters (RF_orig), hyperparameter-tuned SVM (SVM_hyper) and SVM with default parameters (SVM_orig), respectively. The highest minimum scores are calculated from the first set of 5 pseudorandom seeds. The highest average minimum score over three lists of 5 pseudorandom train-test split seeds per model is highlighted in bold

charged residues (aspartic acid, glutamic acid), and hydrophobic residues (isoleucine, leucine, methionine, phenylalanine, valine).

Bottom substitution combination identification

The bottom five 6-DMS substitution combinations per model with the lowest minimum scores (least accurate) were identified and are summarised in Table 2. Supplementary Tables 10–13 contain the bottom five combinations per model with the lowest minimum scores for the 2-, 3-, 4- and 5-DMS feature sets. The average lowest minimum scores (worst accuracies) ranged between 69% and 71% for all combinations across the SVM models and RF model with hyper parameter tuning, and decreased to between 65% and 67% for the default RF model (Table 2). Therefore, even when using the worst combinations of features, the behaviour of the target is still correctly predicted for around 70% of residues.

The 6-DMS combinations with the lowest minimum scores (least accurate) primarily consisted of the substitutions to alanine, cysteine, phenylalanine, isoleucine, leucine, methionine, threonine and valine, as seen in Table 2. Only one 6-DMS combination identified by the hyperparameter-tuned RF model contained a tyrosine substitution. Additionally, many of the same combinations were identified by multiple model types, and the same eight substitutions always appear in the combinations for the 2-, 3-, 4- and 5-DMS feature sets.

With the exception of threonine with a side chain hydroxyl group, and cysteine with a weakly-polar thiol side chain, the substitutions present in the combinations with the

Table 2 Lowest minimum scores for combinations of 6-DMS substitution scores across ML models

Model	Substitution combination	Lowest minimum score	Average minimum score
RF_hyper	A, C, F, I, L, V	69.43 ± 3.11	**71.11 ± 3.58**
	A, C, I, L, V, Y	70.57 ± 2.88	71.87 ± 3.78
	A, C, F, I, M, V	70.80 ± 2.38	72.26 ± 3.22
	A, I, L, M, T, V	70.80 ± 2.52	72.18 ± 3.23
	C, I, L, M, T, V	70.80 ± 1.74	71.64 ± 3.27
RF_orig	A, C, I, L, T, V	65.52 ± 2.93	**65.44 ± 3.15**
	C, I, L, M, T, V	65.75 ± 2.97	65.98 ± 3.95
	A, C, F, L, M, T	66.90 ± 2.49	67.13 ± 3.83
	A, C, I, L, M, T	67.59 ± 1.50	68.20 ± 3.70
	C, F, I, M, T, V	67.82 ± 4.74	68.43 ± 4.68
SVM_hyper	A, C, F, L, T, V	69.43 ± 2.38	**69.27 ± 3.15**
	A, C, F, I, T, V	70.11 ± 2.93	70.80 ± 3.00
	A, F, L, M, T, V	70.34 ± 0.96	70.57 ± 3.00
	A, C, F, I, L, V	71.03 ± 2.62	71.57 ± 3.08
	A, C, F, I, L, T	71.03 ± 2.21	71.26 ± 2.88
SVM_orig	A, C, F, I, T, V	71.26 ± 1.41	71.34 ± 2.96
	A, C, F, L, M, T	71.26 ± 0.82	**70.80 ± 3.24**
	A, C, I, L, T, V	71.49 ± 2.21	71.72 ± 3.22
	C, F, I, L, T, V	71.49 ± 0.96	71.49 ± 3.01
	F, I, L, M, T, V	71.72 ± 1.03	71.80 ± 2.98

The minimum score refers to the lowest mean test set accuracy for a target using that combination

The table shows the lowest minimum scores obtained from combinations of six DMS substitution scores for four machine learning models. These models are hyperparameter-tuned RF (RF_hyper), RF with default parameters (RF_orig), hyperparameter-tuned SVM (SVM_hyper) and SVM with default parameters (SVM_orig), respectively. The lowest minimum scores are calculated from the first set of 5 pseudorandom train-test split seeds. The lowest average minimum score over three lists of 5 pseudorandom train-test split seeds per model is highlighted in bold

lowest minimum scores are all hydrophobic residues. Cysteine is unusual in that it is found in both the best and worst feature sets.

The highest minimum accuracy scores predicted by each 6-DMS combination were all above 90% for one set of five dataset splits, and above 88% when averaged across the three repetitions using different splits.

Generalisability of selected DMS substitution combinations across additional datasets

To assess the generalisability of the combinations identified during brute-force analysis, each selected combination was tested by training a selection of models on the full Mpro FRET replicate 1 dataset and evaluating them on the remaining experimental DMS score datasets using zero-shot prediction. Tables 3 and 4 show example average macro F1 scores for each dataset, obtained by testing SVM combinations with the default SVM model and RF combinations with the hyperparameter-tuned RF models, respectively. All average macro F1 scores for the twenty combinations over all models for all datasets are provided in Supplementary Tables 14–25. Figure 3 summarises these results, showing average F1 scores across the twenty combinations with error bars indicating the minimum and maximum values, with bars and subplots representing models and datasets (Fig. 3A) or datasets and models (Fig. 3B).

The average macro F1 scores for all datasets of the control model, the dummy classifier, were below 50% (Fig. 3B), indicating that the chosen metric of average macro F1 score appropriately reflects model performance. The overall macro F1 scores (Table 3) are approximately 1–2% higher for the FRET replicate 1 training set when using the default SVM model compared to those produced by the combinations in Table 4 with

Table 3 Mean macro F1 scores ± sample standard deviation for zero-shot predictions using the default SVM model

Combination	FRET rep 1	FRET rep 2	TF rep 1	TF rep 2	Growth rep 1	Growth rep 2	PLpro	β-lactamase	BRCA1	Spike rep 1	Spike rep 2
GNPQVY	91.83 ± 2.00	91.55 ± 2.20	**89.17 ± 2.35**	89.52 ± 2.09	90.80 ± 2.19	90.07 ± 1.85	86.98 ± 3.70	83.56 ± 4.69	69.00 ± 8.41	77.23 ± 9.55	76.70 ± 10.65
EFGPQV	91.71 ± 1.79	91.35 ± 2.00	89.14 ± 2.44	**89.95 ± 1.97**	90.70 ± 2.36	90.22 ± 2.47	86.89 ± 4.33	83.76 ± 4.80	**70.92 ± 8.09**	75.12 ± 14.14	73.58 ± 13.81
CDGNPY	90.89 ± 2.78	90.99 ± 2.39	87.51 ± 4.14	87.73 ± 3.96	90.00 ± 2.69	88.50 ± 2.81	85.40 ± 5.51	82.54 ± 5.44	65.51 ± 5.79	77.44 ± 9.19	76.14 ± 9.71
EFGNPV	91.30 ± 2.40	91.44 ± 2.16	88.71 ± 2.48	89.61 ± 2.11	90.65 ± 2.24	90.49 ± 1.82	86.01 ± 4.70	83.82 ± 5.52	69.32 ± 8.77	72.43 ± 14.15	70.85 ± 13.82
CGHPTY	90.69 ± 2.84	90.71 ± 3.14	87.44 ± 3.80	87.21 ± 3.59	90.07 ± 1.90	88.82 ± 2.68	85.43 ± 4.12	82.25 ± 4.88	63.62 ± 7.27	77.70 ± 9.73	76.89 ± 9.47
CDFGPQ	91.59 ± 2.00	91.47 ± 1.80	88.56 ± 3.96	88.70 ± 4.15	91.00 ± 2.28	89.51 ± 2.48	86.29 ± 5.35	84.05 ± 4.64	66.50 ± 5.67	80.22 ± 9.12	78.25 ± 9.52
GHPTVY	91.74 ± 2.03	91.42 ± 2.17	88.55 ± 3.18	88.68 ± 2.82	91.01 ± 1.53	90.34 ± 1.52	87.38 ± 3.31	**85.05 ± 3.75**	68.35 ± 9.66	77.11 ± 8.61	76.56 ± 8.83
CFGHIP	92.19 ± 1.85	92.20 ± 2.06	88.69 ± 3.57	88.47 ± 3.44	91.65 ± 1.61	90.44 ± 1.48	87.28 ± 3.31	84.26 ± 3.47	64.48 ± 7.91	**81.35 ± 5.82**	**81.70 ± 5.26**
CFGLPQ	92.35 ± 1.73	**92.28 ± 1.98**	88.12 ± 2.61	88.42 ± 3.28	91.41 ± 1.69	90.65 ± 1.51	87.29 ± 3.46	84.28 ± 3.63	64.51 ± 5.80	81.04 ± 6.32	80.13 ± 5.89
CFGPQV	91.70 ± 1.72	91.64 ± 2.06	89.15 ± 2.14	89.06 ± 2.30	**91.85 ± 1.72**	**90.97 ± 1.34**	**87.63 ± 3.07**	84.68 ± 3.58	65.81 ± 7.46	80.08 ± 9.39	79.68 ± 9.39

Mean macro F1 scores across experimental DMS datasets for zero-shot predictions using the default SVM model trained on the Mpro FRET replicate 1 dataset. Each value represents the average macro F1 score over 14 predicted substitutions across experimental DMS datasets. The ten feature combinations shown are the top five identified from brute-force analysis using the hyperparameter-tuned SVM and default SVM models, respectively. Bold values indicate the highest mean macro F1 score for each dataset

Table 4 Mean macro F1 scores ± sample standard deviation for zero-shot predictions using the hyperparameter-tuned RF model

Combination	FRET rep 1	FRET rep 2	TF rep 1	TF rep 2	Growth rep 1	Growth rep 2	PLpro	β-lactamase	BRCA1	Spike rep 1	Spike rep 2
EFGHPT	90.93 ± 1.88	89.63 ± 2.40	88.49 ± 3.25	88.29 ± 3.28	87.65 ± 4.21	87.39 ± 4.33	79.63 ± 10.41	77.30 ± 10.79	61.30 ± 10.89	62.53 ± 17.27	61.60 ± 16.85
GNPSVY	91.02 ± 1.62	89.54 ± 2.88	88.53 ± 2.12	88.53 ± 2.22	88.08 ± 3.21	87.80 ± 2.66	82.02 ± 7.92	80.79 ± 7.47	60.21 ± 9.68	61.63 ± 15.98	61.30 ± 14.78
EGNPVY	90.96 ± 2.30	90.02 ± 2.02	87.73 ± 2.86	87.41 ± 3.38	88.14 ± 2.89	88.66 ± 2.62	81.89 ± 8.33	79.57 ± 9.19	59.51 ± 9.00	56.64 ± 16.32	56.49 ± 15.91
EFGHPV	91.05 ± 1.82	**90.20 ± 2.14**	88.66 ± 2.85	**89.05 ± 3.04**	88.76 ± 3.02	88.33 ± 3.44	80.83 ± 10.29	79.55 ± 8.85	61.87 ± 9.25	59.30 ± 15.07	57.88 ± 14.86
EGMPTY	91.16 ± 1.86	90.17 ± 2.16	**88.74 ± 3.41**	88.19 ± 3.42	**89.16 ± 2.32**	**89.24 ± 2.23**	**84.47 ± 5.15**	79.06 ± 7.17	**63.09 ± 7.04**	59.15 ± 16.81	58.92 ± 16.26
CDFGPQ	90.95 ± 2.95	89.76 ± 3.35	86.18 ± 5.25	86.22 ± 5.47	88.97 ± 3.65	88.78 ± 3.76	81.91 ± 8.40	**82.92 ± 5.37**	60.99 ± 9.13	65.99 ± 17.77	63.91 ± 16.47
CDGPQY	90.70 ± 3.33	89.77 ± 3.33	86.30 ± 4.80	86.01 ± 5.37	88.53 ± 4.30	88.49 ± 4.13	82.50 ± 8.42	80.65 ± 8.31	61.45 ± 8.34	65.46 ± 19.12	64.40 ± 18.29
CEFGHP	90.47 ± 2.43	89.27 ± 3.48	87.12 ± 4.60	87.01 ± 4.66	88.21 ± 3.95	88.73 ± 4.11	81.57 ± 8.44	81.65 ± 5.48	60.84 ± 8.73	63.88 ± 17.14	63.87 ± 17.65
CDFGPS	90.36 ± 3.18	89.26 ± 3.72	85.75 ± 4.89	85.36 ± 4.59	88.55 ± 3.91	88.60 ± 3.79	80.92 ± 7.67	82.78 ± 4.97	61.62 ± 9.95	**67.13 ± 17.92**	**66.64 ± 16.17**
CEGNPY	89.88 ± 2.93	88.98 ± 3.17	85.51 ± 4.87	85.66 ± 5.03	87.78 ± 4.16	87.48 ± 3.98	80.88 ± 9.05	79.38 ± 8.52	57.24 ± 9.45	63.12 ± 19.08	62.48 ± 17.79

Mean macro F1 scores across experimental DMS datasets for zero-shot predictions using the hyperparameter-tuned RF model trained on the Mpro FRET replicate 1 dataset. Each value represents the average macro F1 score over the fourteen predicted substitutions. The ten feature combinations shown are the top five combinations identified from brute-force analysis using the RF hyperparameter-trained model and the default parameter RF model, respectively. The bold value in each column corresponds to the highest mean macro F1 score for a given dataset

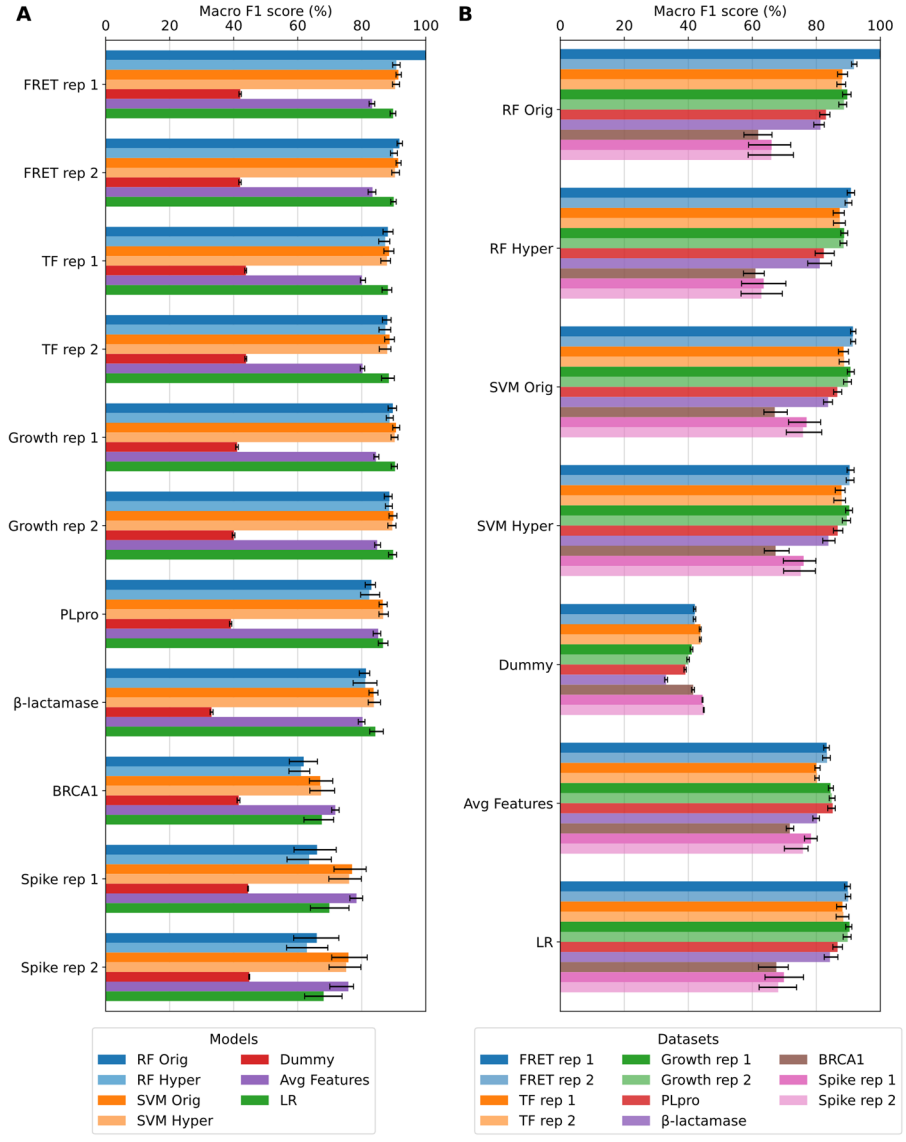

Fig. 3 Average macro F1 scores for top DMS input combinations across (**A**) datasets and (**B**) models. Bar graphs showing results across the twenty top DMS input combinations, with grouped bars representing (**A**) models and (**B**) datasets. Bars show the mean of the average macro F1 scores for each (**A**) model and (**B**) dataset, with error bars representing the range of scores (minimum to maximum values across the twenty combinations)

the hyperparameter-tuned RF model. The dataset with the average macro F1 scores closest to those of the training set for all SVM and RF models was the Mpro FRET replicate 2 dataset (Fig. 3A), with the highest score using the 6-DMS combinations approximately 92%. This is much lower than the macro F1 scores of around 98% achieved by the experimental baseline when using the corresponding substitution values from the Mpro FRET replicate 1 dataset, as seen in Supplementary Tables 26 and 27.

The average macro F1 scores for the other assays used to characterise the functionality of the SARS-CoV-2 Mpro protein are the next highest of all test sets, ranging from approximately 90–91% for the growth assay and around 88–89% for the TF assay (Fig. 3A). Interestingly, the experimental baseline macro F1 scores were similar to the ML model predictions for the TF datasets, with scores of approximately 88–89%, but

slightly higher for the growth datasets, from 92 to 95% (Supplementary Tables 26 and 27). This difference in macro F1 scores between the experimental baseline and ML models is smaller than between the FRET replicates.

The average macro F1 scores for the PLpro activity fitness score dataset were only slightly lower than those for the Mpro datasets, ranging from around 85–87% for all combinations when tested with the SVM models and 79–84% for RF models (Fig. 3A). The TEM-1 β-lactamase dataset was the only bacterial protein dataset used for testing. It had comparably high average macro F1 scores, around 82–85% for the SVM models and 77–82% for the RF models (Fig. 3A).

The average macro F1 scores for the tested Spike RBD replicate datasets were lower than those of any other SARS-CoV-2 proteins, achieving scores around 70–81% for the SVM models and 56–67% for the RF models (Fig. 3A). The dataset with the worst average macro F1 scores was the human BRCA1 RING domain dataset, with ranges of 63–70% for the SVM models and 57–63% for the RF models (Fig. 3A).

In contrast to the Mpro, PLpro, and β-lactamase datasets, where model performance was comparable across algorithms, the SVM models consistently outperformed both the RF and logistic regression models for all feature combinations in the human BRCA1 RING domain and SARS-CoV-2 Spike RBD datasets (Fig. 3B).

The average macro F1 scores for the Mpro and PLpro datasets when using the Logistic Regression model are very similar to those for the SVM and RF models, at around 87–91% for all combinations (Fig. 3A). Additionally, this model achieved the highest average macro F1 score for the TEM-1 β-lactamase dataset, at 88.28%, which is slightly higher than the scores obtained by the SVM model (Fig. 3A). However, the performance of this model on the Spike RBD datasets fell between the RF and SVM models.

In Supplementary Figs. 1 and 2, the distributions of the top 20 DMS score combinations were plotted, comprising 19 independent combinations, as the combination of C, D, F, G, P and Q DMS scores was identified in the top 5 for both RF and SVM models without hyperparameter tuning. For all Mpro, PLpro and β-lactamase datasets, the 20 DMS score distributions were bimodal, with separation between modes occurring at the 0.5 threshold (Supplementary Figs. 1 and 2A, B). However, the distribution of DMS scores for the BRCA1 dataset was most consistent with a unimodal distribution (Supplementary Fig. 2C). For the Spike datasets, a bimodal distribution was observed, but the cutoff separating these two modes appeared to be closer to 0.65 (Supplementary Fig. 2D, E). The distributions of all 20 DMS scores for each dataset were compared to distributions of the 6-DMS score combinations per dataset, and were found to have almost identical distributions (Supplementary Figs. 3 and 4). All distributions of the 20 chosen 6-DMS combinations are almost indistinguishable from each other within each dataset.

Discussion

Our study demonstrates that predictive performance for DMS substitution effects improves with the number of input substitutions, with six features providing mean minimum accuracies of just under 90%. As the search space scales combinatorially with the number of input substitutions, extending beyond six features would lead to rapidly increasing computational cost, while also increasing the number of required experimental substitutions measured in the laboratory. This is contrary to the main objective, which is maintaining a minimal set of substitutions to be measured experimentally while

retaining sufficient performance for predicting the remaining substitutions. Therefore, in this study, the feature-selection framework was restricted to 6-DMS combinations. There is a trade-off between mean minimum accuracy, number of DMS substitutions and cost (both computational and laboratory). Figure 2 indicates that if 80% mean minimum accuracy is acceptable, then only 2 substitutions would be needed, whereas achieving over 90% mean minimum accuracy would require 7 or more substitutions.

Although the 6-DMS feature combination search required substantial computational time, this procedure was performed only once on the Mpro FRET replicate 1 dataset as a feature combination selection step, confining the computational cost to this stage. All identified combinations were then reused for subsequent zero-shot prediction analyses on independent datasets. The search would only need to be repeated if a different DMS training dataset was used, or if a different prediction task such as regression was considered, requiring new feature combinations. The SVM-RFE heuristic method was explored, but only resulted in an ordered feature ranking, and could not identify feasible feature combinations. Other heuristic or combinatorial search strategies may be able to identify suitable subsets more efficiently than the brute-force approach, which may represent a possible direction for future work.

Analysis of the top substitution combinations revealed that proline and glycine were consistently selected in 6-DMS feature sets. This suggests these substitutions are difficult to predict, likely reflecting their distinctive structural properties. The cyclic pyrrolidine side chain of proline limits both its backbone conformational freedom and that of the preceding residue. Additionally, its secondary amide backbone nitrogen restricts the ability of proline to form hydrogen bonds and proline residues are associated with disruption regular α-helical or β-sheet secondary structure. In contrast, glycine is uniquely flexible due to the absence of a β-carbon, allowing it to adopt backbone conformations inaccessible to other amino acid residues.

The remaining substitutions in the top-performing combinations typically included representatives from different physicochemical classes. Incorporating at least one amino acid from each class likely ensures that the selected features capture the diverse aspects of amino acid behaviour, which may contribute to robust prediction performance.

Proline was always present in the top combination feature subsets from 3-DMS inputs onwards, whereas the glycine substitution was absent from some of the top combinations for 2-, 3- and 4-DMS feature subsets. This is likely a reflection of a tradeoff occurring during feature selection. When fewer substitutions are selected as features, the model cannot only include the worst-predicted substitutions but must select features that are good predictors for the majority of the remaining target substitutions.

No patterns were found in the comparison between the RFE and brute-force top substitution combinations, suggesting that the RFE approach could not identify feature combinations found by the brute-force approach. This suggests that brute-force combination testing is necessary to identify optimal feature sets in this context.

In contrast to the top five 6-DMS substitution combinations, all 6-DMS combinations with the lowest minimum scores (least accurate) show high consistency and are restricted to residues from a subset of eight. One reason may be that the target corresponding to these worst accuracies is most likely to be the proline DMS substitution, which is the worst predicted by these eight substitutions.

With the exception of cysteine and threonine, only hydrophobic residue substitutions are present in combinations with the lowest minimum scores. This suggests that hydrophobic residues are poor predictors of polar or charged residues, and these substitutions carry redundant information. Interestingly, cysteine is often found in both the best and worst feature sets, suggesting that the predictive value of this residue is dependent on the context of the other residues in the combination.

This consistent performance of the top substitution combinations (88%) over different splits and multiple repetitions indicate that these results are reliable, as multiple substitution combinations can provide comparable predictive performance. This highlights brute-force combination testing as a suitable strategy for reducing the number of experimental substitutions required for DMS profiling. However, this approach is only likely to be reliable if the DMS profiling uses experimental techniques comparable to those used to produce DMS scores on which the models were trained. This includes using similar normalisation and consistent categorisation of the measured experimental effect of the amino acid substitutions.

After training the models with the top combinations and Mpro FRET replicate 1 dataset, and generalising to the remaining datasets, the best average macro F1 score was achieved when predicting substitutions from the Mpro FRET replicate 2 dataset for all models. This is expected, as this dataset was the closest to the training set, simply being another replicate of the experiment used to create the training set. However, this predictive performance was not as good as when using the experimental scores for the corresponding substitution from the Mpro FRET replicate 1 dataset. This suggests that the limited information provided by 6 input substitution scores constrains the maximum average macro F1 score that can be achieved.

The next best predicted datasets were found to be the Mpro TF and Growth replicate 1 and 2 datasets. This result is not unexpected, as these datasets describe the same SARS-CoV-2 Mpro protein, with DMS scores representing the changes in protease activity resulting from a single mutation. The experimental baseline average macro F1 scores were similar to the scores for these models. It is not surprising that using substitution scores from a different experiment on the same protein can achieve comparable or better predictive performance than using ML models. However, the ML model has the advantage that it reduces the number of experimental substitutions required and is generalisable to other proteins.

The PLpro activity fitness score dataset was the next best predicted after the Mpro datasets, potentially because the proteins have similar function (both proteases) and belong to the same virus. Similar average macro F1 scores were predicted for the TEM-1 β-lactamase dataset, which may be because the Mpro, PLpro and β-lactamase proteins are all enzymes, specifically hydrolases, suggesting that the selected combinations are robust across proteins with similar functional classes.

The dataset that was the worst predicted by all models was the human BRCA1 RING domain dataset. This is most likely caused by the different functional properties being probed by the BRCA1 DMS experiment. Whereas the Mpro and PLpro DMS experiments compared wild-type and mutant enzymes, where it might be anticipated that the wild-type enzyme has evolved optimal activity for a given substrate, the BRCA1 DMS experiment measured the propensity of a chimeric protein comprising BARD1 residues 26–126 plus the BRCA1 RING domain to become autoubiquitinated in the presence of

ubiquitin E1/E2 ligase enzymes and epitope (FLAG)-tagged ubiquitin [26]. However, the BRCA1 RING domain also mediates ubiquitination of many cellular targets via several different E2 ligases [27] and thus it differs from the enzymes, as the DMS assay measures only one of its multiple functions. Additionally, the resultant DMS scores represent both the stability of the BARD1:BRCA1 interaction, which is essential for its E3 ligase activity [28], plus the affinity of BRCA1 for its cognate E2 ligase UbcH5c/UBE2D3. Therefore, these activity scores are derived from changes in protein-protein interactions, which the models have not been trained to recognise.

Both the BRCA1 RING domain and Spike datasets with the lowest average macro F1 scores had DMS score distributions that were least similar to the Mpro FRET 1 dataset. These distribution differences may have contributed to reduced model performance on these datasets. It is not clear whether the decrease in performance for these datasets was driven primarily by the distributional differences or functional differences when compared to the Mpro FRET 1 dataset. Further investigation is necessary to disentangle these effects on predictive performance.

From the observation that distributions of all 20 DMS scores are almost identical to those of the 20 6-DMS combination distributions per dataset, it appears that the distribution for any DMS score dataset can be ascertained using only a minimal 6-DMS subset. In addition, the similarity of the distributions between combinations within a dataset suggests that any of the combinations can be chosen as a suitable experimental subset.

The SVM models showed similar performance to the RF and logistic regression ML models for the Mpro, PLpro and β-lactamase datasets, outperforming all other models on the human BRCA1 RING domain and SARS-CoV-2 Spike RBD datasets. As SVMs with an RBF kernel can capture smooth nonlinear boundaries, this behaviour may have been captured better by the SVM models. Linear models, such as the logistic regression model, were found to be comparable to non-linear ones, only on the hydrolase datasets, outperforming the RF models on the Spike RBD dataset. Overall, the zero-shot prediction SVM models appeared to provide the most consistent and robust performance among the tested zero-shot prediction ML models across the diverse DMS datasets, as shown in Fig. 3B.

Although the feature combinations were identified using a single, small dataset without robust validation techniques, a consistent performance was observed in the zero-shot evaluations across multiple independent DMS datasets. This suggests that the selected feature combinations were not overfitted and not strongly biased towards the Mpro FRET replicate 1 dataset.

In summary, all ML models and methods trained on the Mpro FRET replicate 1 dataset had varying predictive power, reflected by the average macro F1 scores across the evaluated datasets. Higher scores were observed for the Mpro, PLpro and β-lactamase datasets, while performance decreased for the Spike and BRCA1 datasets. While some test sets were better predicted than others, the overall macro F1 scores indicate that the chosen combinations identified by the brute-force approach show promise for classifying substitution behaviour as wild-type or loss-of-function in related proteases and hydrolases.

However, the limitations of these models must be acknowledged. The predictive power of these models is limited to a binary classification of protein behaviour with a single

amino acid substitution as either wild-type or loss-of-function in nature. This binary classification does not capture intermediate functional effects, or the continuous nature of some protein fitness landscapes. For example, the BRCA1 dataset values were classified by Starita et al. into three classes, namely wild-type like, impaired and completely nonfunctional. Therefore, these models are not suitable for applications in which functionality cannot be described by only two classes, the relative ranking of substitutions to one another is required, or the experimental data contains multiple simultaneous substitutions.

The next limitation of the framework is that one fixed target threshold was chosen to distinguish the wild-type and loss-of-function behaviour over all datasets. While this appears appropriate for datasets with DMS scores naturally separating at the fixed threshold, this assumption does not hold for all DMS datasets. As a result, model performance is most likely sensitive to the chosen threshold, and deviations in separation may introduce bias into the classification results.

The last limitation of this framework is that it relies on normalised DMS score datasets. This assumes the same scaling is applied to all datasets, and all information required for this scaling is present, such as stop-codon and nonsense variant scores. However, in their absence, score distributions may be shifted or rescaled inconsistently, making the application of a fixed threshold less reliable and reducing transferability across datasets.

These limitations suggest that the current classification models within this framework may not be optimal for datasets that do not meet all requirements. The use of regression models, which predict continuous fitness scores rather than binary classes, may be more appropriate for these datasets. Further investigation is necessary to determine whether regression models and their corresponding feature combinations could improve generalisibility across diverse DMS datasets.

To enable application of this framework to new proteins, the following methodology is proposed. Any one of the 20 6-DMS combinations can be chosen, where the corresponding DMS scores for all residues are then experimentally obtained, along with stop codon scores. These scores are then normalised, and their distribution calculated. In cases where a clear bimodal distribution separated with a 0.5 threshold is present, the trained classification models from this framework can be applied with minimal modification. However, if the resulting distributions are unimodal, continuous or exhibit shifted or poorly separated modes, model performance may be compromised, and alternative approaches should be considered. Regression models are recommended for datasets that do not exhibit bimodality, while classification models may instead be trained on a dataset that exhibits a similar bimodal distribution and threshold to the target dataset to be predicted.

Conclusion

In this study, we established a brute-force approach that integrates ML with DMS to identify subsets of experimental mutations needed to predict all others, thereby effectively reducing experimental effort and cost. We showed that multiple sets of 6-DMS substitutions can predict the remaining fourteen with approximately 90% accuracy (based on the F1 score), while classifying mutations as either wild-type or loss-of-function. Interestingly, while glycine and proline were consistently included in these substitution sets, the rest of the amino acids spanned diverse physicochemical properties,

capturing various aspects of amino acid behaviour, which may contribute to robust prediction performance.

We noted that all ML models and methods trained on the Mpro FRET replicate 1 dataset had good predictive power when applied to DMS of proteins of similar functional classes, including other proteases (PLpro) and hydrolases (β-lactamase). The high overall macro F1 scores indicated that the chosen combinations are reliable, provided that datasets are carefully prepared, requiring both normalisation and bimodal, binary functional classification as either wild-type or loss-of-function with a defined separation between classes at DMS scores around 0.5.

Our results indicate that using subsets of DMS scores to predict the remaining substitutions with ML models shows promise, yet generalisability across protein families requires improvement. Additional accuracy gains may be achievable by integrating additional information sources and approaches such as protein language models [29].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-026-06473-5.

Supplementary Material 1

Acknowledgements
Prashant Singh and Özlem Tastan Bishop are members of ARUA/ The GUILD African-European Cluster of Research Excellence for AI (CoRE-AI; https://africaeuropecoreai.org/). Emily Morgan and Shaylyn Govender acknowledge Uppsala University for providing the travel grant for the research visit as part of the CoRE-AI activity.

Author contributions
Conceptualization, S.C.G., N.T.B. and Ö.T.B.; Funding acquisition, I.G, S.C.G. and Ö.T.B.; Project administration, S.C.G. and Ö.T.B.; Resources, P.S. and Ö.T.B.; Methodology: E.M., S.G., P.S., N.T.B.; Formal analysis, E.M., S.G., P.S., I.G., S.C.G, N.T.B. and Ö.T.B.; Writing - original draft: E.M.; Writing – review and editing: E.M., S.G., P.S., I.G., S.C.G, N.T.B. and Ö.T.B.

Funding
This project is supported by the Novo Nordisk Foundation (NNF) and the Pandemic Antiviral Discovery (PAD) Initiative; Grant Number: NNF23SA0084504. E.M. and S.G are supported by National Research Foundation (NRF) of South Africa PhD scholarship via National Institute for Theoretical and Computational Sciences (NITheCS).

Data availability
The raw DMS substitution data used in this study were obtained from the Supplementary and Source Data sections of previously published work [19, 23–26]. Links and instructions for accessing these files, together with all processed datasets, scripts, results and figures necessary for full reproducibility are publicly available in a FigShare repository (https://doi.org/10.21504/RUR.32101051) and a GitHub repository (https://github.com/RUBi-ZA/ml-dms-targeted-substitution-profiling.git).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Research Unit in Bioinformatics (RUBi), Department of Biochemistry, Microbiology and Bioinformatics, Rhodes University, Makhanda 6139, South Africa
[2]Department of Information Technology, Science for Life Laboratory, Scientific Computing, Uppsala University, Uppsala, Sweden
[3]Division of Virology, Department of Pathology, University of Cambridge, Cambridge CB2 1QP, UK
[4]Department of Pure and Applied Mathematics, Rhodes University, Makhanda 6139, South Africa
[5]National Institute for Theoretical and Computational Sciences (NITheCS), Stellenbosch, South Africa

Received: 12 February 2026 / Accepted: 5 May 2026
Published online: 19 May 2026

References

1. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nat Methods. 2014;11:801–7. https://doi.org/10.1038/nmeth.3027.
2. Wei H, Li X. Deep mutational scanning: a versatile tool in systematically mapping genotypes to phenotypes. Front Genet. 2023;14:1087267. https://doi.org/10.3389/fgene.2023.1087267.
3. Shepherdson JL, Granas DM, Li J, Shariff Z, Plassmeyer SP, Holehouse AS, et al. Mutational scanning of CRX classifies clinical variants and reveals biochemical properties of the transcriptional effector domain. Genome Res. 2024;34:1540–52. https://doi.org/10.1101/gr.279415.124.
4. Li C, Qian W, Maclean CJ, Zhang J. The fitness landscape of a tRNA gene. Science. 2016;352:837–40. https://doi.org/10.1126/science.aae0568.
5. Küng C, Protsenko O, Vanella R, Nash MA. Deep mutational scanning reveals a de novo disulfide bond and combinatorial mutations for engineering thermostable myoglobin. Protein Sci Publ Protein Soc. 2025;34:e70112. https://doi.org/10.1002/pro.70112.
6. Newberry RW, Leong JT, Chow ED, Kampmann M, DeGrado WF. Deep mutational scanning reveals the structural basis for α-synuclein activity. Nat Chem Biol. 2020;16:653–9. https://doi.org/10.1038/s41589-020-0480-6.
7. Sruthi CK, Prakash M. Deep2Full: evaluating strategies for selecting the minimal mutational experiments for optimal computational predictions of deep mutational scan outcomes. PLoS ONE. 2020;15:e0227621. https://doi.org/10.1371/journal.pone.0227621.
8. Sruthi CK, Balaram H, Prakash MK. Toward developing intuitive rules for protein variant effect prediction using deep mutational scanning data. ACS Omega. 2020;5:29667–77. https://doi.org/10.1021/acsomega.0c02402
9. Munro D, Singh M. DeMaSk: a deep mutational scanning substitution matrix and its use for variant impact prediction. Bioinforma Oxf Engl. 2021;36:5322–9. https://doi.org/10.1093/bioinformatics/btaa1030.
10. Fu Y, Bedő J, Papenfuss AT, Rubin AF. Integrating deep mutational scanning and low-throughput mutagenesis data to predict the impact of amino acid variants. GigaScience. 2022;12:giad073. https://doi.org/10.1093/gigascience/giad073.
11. Gelman S, Fahlberg SA, Heinzelman P, Romero PA, Gitter A. Neural networks to learn protein sequence-function relationships from deep mutational scanning data. Proc Natl Acad Sci U S A. 2021;118:e2104878118. https://doi.org/10.1073/pnas.2104878118.
12. Sarfati H, Naftaly S, Papo N, Keasar C. Predicting mutant outcome by combining deep mutational scanning and machine learning. Proteins. 2022;90:45–57. https://doi.org/10.1002/prot.26184.
13. Livesey BJ, Marsh JA. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. Mol Syst Biol. 2020;16:e9380. https://doi.org/10.15252/msb.20199380.
14. Livesey BJ, Marsh JA. Interpreting protein variant effects with computational predictors and deep mutational scanning. Dis Model Mech. 2022;15:dmm049510. https://doi.org/10.1242/dmm.049510.
15. Livesey BJ, Marsh JA. Updated benchmarking of variant effect predictors using deep mutational scanning. Mol Syst Biol. 2023;19:e11474. https://doi.org/10.15252/msb.202211474.
16. Gerasimavicius L, Livesey BJ, Marsh JA. Correspondence between functional scores from deep mutational scans and predicted effects on protein stability. Protein Sci Publ Protein Soc. 2023;32:e4688. https://doi.org/10.1002/pro.4688.
17. Drake ZC, Day EH, Toth PD, Lindert S. Deep-learning structure elucidation from single-mutant deep mutational scanning. Nat Commun. 2025;16:6874. https://doi.org/10.1038/s41467-025-62261-4.
18. Xia H, Wei D, Guo Z, Chung LW. Machine learning on the impacts of mutations in the SARS-CoV-2 spike RBD on binding affinity to human ACE2 based on deep mutational scanning data. Biochemistry. 2025;64:3790–800. https://doi.org/10.1021/acs.biochem.4c00587.
19. Flynn JM, Samant N, Schneider-Nachum G, Barkan DT, Yilmaz NK, Schiffer CA, 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.
20. Barozi V, Chakraborty S, Govender S, Morgan E, Ramahala R, Graham SC, et al. Revealing SARS-CoV-2 Mpro mutation cold and hot spots: dynamic residue network analysis meets machine learning. Comput Struct Biotechnol J. 2024;23:3800–16. https://doi.org/10.1016/j.csbj.2024.10.031.
21. Govender S, Morgan E, Ramahala R, Lobb K, Bishop NT, Tastan Bishop Ö. Transfer learning towards predicting viral missense mutations: a case study on SARS-CoV-2. Comput Struct Biotechnol J. 2025;27:1686–92. https://doi.org/10.1016/j.csbj.2025.04.029.
22. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
23. Wu X, Go M, Nguyen JV, Kuchel NW, Lu BGC, Zeglinski K, et al. Mutational profiling of SARS-CoV-2 papain-like protease reveals requirements for function, structure, and drug escape. Nat Commun. 2024;15:6219. https://doi.org/10.1038/s41467-024-50566-9.
24. Starr TN, Greaney AJ, Hilton SK, Ellis D, Crawford KHD, Dingens AS, et al. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. Cell. 2020;182:1295–e131020. https://doi.org/10.1016/j.cell.2020.08.012.
25. Firnberg E, Labonte JW, Gray JJ, Ostermeier M. A comprehensive, high-resolution map of a gene's fitness landscape. Mol Biol Evol. 2014;31:1581–92. https://doi.org/10.1093/molbev/msu081.
26. Starita LM, Young DL, Islam M, Kitzman JO, Gullingsrud J, Hause RJ, et al. Massively parallel functional analysis of BRCA1 RING domain variants. Genetics. 2015;200:413–22. https://doi.org/10.1534/genetics.115.175802.
27. Witus SR, Stewart MD, Klevit RE. The BRCA1/BARD1 ubiquitin ligase and its substrates. Biochem J. 2021;478:3467–83. https://doi.org/10.1042/BCJ20200864.
28. Hashizume R, Fukuda M, Maeda I, Nishikawa H, Oyake D, Yabuki Y, et al. The RING heterodimer BRCA1-BARD1 Is a ubiquitin ligase inactivated by a breast cancer-derived mutation. J Biol Chem. 2001;276:14537–40. https://doi.org/10.1074/jbc.C000881200.

29. Vieira LC, Handojo ML, Wilke CO. Medium-sized protein language models perform well at transfer learning on realistic datasets. Sci Rep. 2025;15:21400. https://doi.org/10.1038/s41598-025-05674-x.

Publisher's note