Received: 22 August 2020 | Revised: 1 June 2021 | Accepted: 11 July 2021

DOI: 10.1002/prot.26184

RESEARCH ARTICLE

PROTEINS WILEY

Predicting mutant outcome by combining deep mutational scanning and machine learning

Hagit Sarfati[1] | Si Naftaly[2] | Niv Papo[2] | Chen Keasar[1]

[1]Department of Computer Science, Ben-Gurion University of the Negev, Be'er Sheva, Israel

[2]Avram and Stella Goldstein-Goren Department of Biotechnology Engineering and the National Institute of Biotechnology in the Negev, Ben-Gurion University of the Negev, Be'er Sheva, Israel

Correspondence
Chen Keasar, Department of Computer Science, Ben-Gurion University of the Negev, Be'er Sheva, Israel.
Email: keasar@bgu.ac.il

Funding information
H2020 European Research Council, Grant/Award Number: 336041; European Research Council, Grant/Award Number: 336041; Israel Science Foundation, Grant/Award Numbers: 1122/14, 615/14

Abstract

Deep mutational scanning provides unprecedented wealth of quantitative data regarding the functional outcome of mutations in proteins. A single experiment may measure properties (eg, structural stability) of numerous protein variants. Leveraging the experimental data to gain insights about unexplored regions of the mutational landscape is a major computational challenge. Such insights may facilitate further experimental work and accelerate the development of novel protein variants with beneficial therapeutic or industrially relevant properties. Here we present a novel, machine learning approach for the prediction of functional mutation outcome in the context of deep mutational screens. Using sequence (one-hot) features of variants with known properties, as well as structural features derived from models thereof, we train predictive statistical models to estimate the unknown properties of other variants. The utility of the new computational scheme is demonstrated using five sets of mutational scanning data, denoted "targets": (a) protease specificity of APPI (amyloid precursor protein inhibitor) variants; (b-d) three stability related properties of IGBPG (immunoglobulin G-binding β1 domain of streptococcal protein G) variants; and (e) fluorescence of GFP (green fluorescent protein) variants. Performance is measured by the overall correlation of the predicted and observed properties, and enrichment—the ability to predict the most potent variants and presumably guide further experiments. Despite the diversity of the targets the statistical models can generalize variant examples thereof and predict the properties of test variants with both single and multiple mutations.

KEYWORDS
deep mutational scanning, machine learning, mutant outcome, prediction, protein library, protein-protein interactions, random forest, specificity, structural features, structural stability

1 | INTRODUCTION

Mutations in a protein sequence may affect its structural stability and/or affinity to other molecules.[1,2] This effect makes mutations a major driving force of protein evolution, modulators of protein-protein interactions (PPIs), and the key to protein engineering. Thus, deciphering the outcome of mutations is essential, for fields as diverse as the discrimination between deleterious and benign mutations[3] and the design of novel therapeutic proteins.[1,4] A reliable evaluation of mutation outcome requires the expression and purification of protein variants, followed by activity measurements. These experimental tasks are time and resource consuming, and thus efficient and accurate computational methods for the prediction of mutation outcome are desirable.

Abbreviations: APPI, amyloid precursor protein inhibitor; AUC, area under the curve; EMA, estimation of model accuracy; KLK6, kallikrein-6; ML, machine learning; NGS, next generation sequencing; OH, one hot (coding of amino acids); PPI, protein-protein interaction; SI, specificity index.

Until recently the number of characterized variants per protein was relatively small, and prediction methods relied mainly on energy calculations, and chemical knowledge. The available examples were used to fine-tune the energy terms, and for validation. The most rigorous methods for the prediction of mutation outcome are based on the free energy perturbation approach,[5] but they are also the most computationally demanding ones. Other methods such as EGED,[6] FoldX,[7] CC/PBSA,[8] PoPMuSiC & BeAtMuSiC,[9,10] and Rosetta-ddG,[11] offer a more computationally efficient alternative, which is also based on physical force fields and conformational sampling (for recent reviews see Geng et al.[12] and Goldenzweig and Fleishman[13]).

The emergence of deep mutational scanning as a major approach to protein study and engineering,[14,15] opens an alternative route to the prediction of mutation outcome such as changes in protein stability, binding affinity, or selectivity. In a single experiment, deep mutational scans provide quantitative activity measures for tens of thousands of variants.[16] These new riches of data call for machine learning (*ML*) methods that can exploit them. One direction of using large-scale experimental mutagenesis datasets was explored by the developers of the ENVISION predictor of single mutant effects.[17] Aiming at elucidating the outcomes of naturally occurring mutants, the ENVISION study used data from nine mutational scan studies to build a statistical model of single mutation outcome, conditioned on structural and evolutionary background. An alternative route considers each mutational scan separately, aiming to deduce properties of protein variants from the known properties of numerous other variants of the same protein. Such methods may be applied to virtual libraries of the target protein and point at promising regions of sequence space, guiding further experimental work. The leading ML approach to this problem is purely sequence-based. That is, the features are derived from the amino acid sequences of the protein variants.[3,18] Structural features, namely energy terms and components thereof, may augment sequence-based ones. Unlike the latter, such features are directly related to the underlying physical basis of stability and specificity and inherently considers interactions of residues, which are distant in sequence. Masso and co-workers[19-21] used their four-body statistical potential to assign each residue with an environment score and these scores serve as the features of their ML scheme. Here, we suggest a new ML scheme for the prediction of mutant outcome, combining structural and sequence features for high throughput prediction of mutation outcome.

The method that we propose includes three stages: (a) structural modeling of the protein or complex variants; (b) extraction of their features (structural and sequence-based); and (c) prediction of variants' activity, using a ML model. We use our in-house molecular modeling package MESHI[22] for structural modeling and feature extraction. MESHI's unique set of structural features (DATA S1 section F) covers a wide range of phenomena that relate to structural stability of proteins and their complexes (eg, compactness, hydrogen bonding). Combined with machine learning, this set of features proved useful in the context of estimation of model accuracy (EMA).[23,24] Here, we show that these features also carry information regarding the stability of proteins and protein complexes.

The method was developed using the APPI benchmark discussed below. Then we applied it to the two other benchmarks to validate its general applicability: Immunoglobulin G-binding β1 domain of streptococcal protein G (IGBG),[2] and green fluorescent protein (GFP).[25] Each benchmark associates the variants with different values which are either directly measured by the experiment or calculated from experimental values. We will refer to these values as "targets." The paragraph below briefly summarizes the three experimental studies and highlight their targets.

APPI, a small protein of 54 residues, is an inhibitor of KLK6 and mesotrypsin, which are serine proteases involved in the progression and invasion stages of prostate, pancreatic and ovarian cancers.[26,27] The crystal structure of $APPI_{M17G/I18F/F34V}$, a potent mesotrypsin-inclined variant, with mesotrypsin is available[28] (PDB ID 5C67, Figure S1). The contact interface of this complex includes a short APPI loop, residues 11-18. Aiming at a more potent and selective protease inhibitor, Naftaly and her co-workes[16] expended $APPI_{M17G/I18F/F34V}$ to an APPI yeast-display library. Seven interface residues (residues 11-13, 15-18) were mutated to all possible alternatives using saturation mutagenesis (degenerate codons). The remaining (non-interface) residues in the entire APPI gene were sparsely mutated using error-prone PCR. The library was FAX sorted to three gates: KLK6-inclined, neutral, and mesotrypsin inclined. Individual variants were sampled at the gates and were identified using next generation sequencing (NGS) analysis. Selected protease-specific variants were validated by standard biochemical procedures. The target associated with this benchmark is a specificity index (*SI*, defined below) that quantifies the differential preference of the variants to either KLK6 (negative values) or mesotrypsin (positive values).

The heavy bias of this library towards binding loop mutations allowed Naftaly and her co-workers to identify three of them as selectivity switches.[16] That is, their selectivity is almost independent of other mutations. From the viewpoint of the prediction task however, this library bias raises much concern. It may lead to a seemingly good overall prediction performance, due to success in predicting only the outcome of binding loop residues. Further, having a binding loop mutation in the training set might ease the prediction of other variants with the same mutation. We addressed these concerns by computational prediction with independent training and test sets.

IGBG is also a small (56 residues), structurally stable protein, which is often used as a model for protein folding and stability studies. Nisthal and his co-workers[2] chose this protein as their benchmark for protein thermodynamic stability "for its small size, high amount of secondary structure, and well-behaved WT sequence." To this end, they generated almost all the possible single residue mutants, and followed their guanidinium chloride (GdmCl) denaturation by tryptophan fluorescence. For each mutant they measured the denaturant concentration at the denaturation midpoint (C_m), and the fluorescence slope at that point (m), and from both C_m and m they calculated the change in folding free energy ($\Delta\Delta G$) upon mutation. Each of C_m, m, and $\Delta\Delta G$ is a target in this study.

Green fluorescent protein (GFP) is a major workhorse in current molecular and cellular biology. Sarkisyan and his co-workers studied

its fitness-landscape using a library of GFP variants.[25] The fluorescence emission of these variants is the target of this benchmark. In terms of mutation outcome prediction this is the most challenging task. GFP is larger than the other two proteins (225 residues), and its structure and function depend on a hard-to-model chromophore.

The current study applies ML to these targets by associating structural and sequence features of training set variants with their observed target values and predicting these values for test set variants.

2 | DATA AND METHODS

2.1 | Software

Structural manipulations and analysis were performed with the program MUTATE of the MESHI molecular modeling package.[22] The analysis of experimental data and ML were performed in MATLAB (The MathWorks, Inc.) version 2019a.

2.2 | Experimental targets

This study applies a uniform methodology to the prediction of mutational outcome of five targets in three experimental systems. Each system applied high throughput mutagenesis to a different protein and uses a different assay.

2.2.1 | APPI

The experimental data in this study were was generated by Naftaly and her co-workers.[16] They FACS-sorted a yeast-display mutagenesis library of APPI variants for selective protease binding (mesotrypsin vs KLK6). They collected fractions of the sorted population using three affinity gates: two specific gates for mesotrypsin and KLK6 bound variants, and one gate for neutral variants that bind both proteases. The collected APPI variants were analyzed by NGS: three samples from each of the specific gates, and one sample from the neutral one. The lists of their reads (FASTQ format) constitute the raw experimental data. Variants with fewer than three copies in the dataset (all three gates) were discarded as suspected sequencing errors. The remaining 11 983 variants are summarized in Table 1A.

KLK6 vs mesotrypsin selectivity of a variant manifests itself by abundance of NGS reads in the sample taken from one gate, compared with the samples from the other gates. The counts of APPI variants range from thousands to a few, representing both the size differences of the underlying, presorted, populations and the differential binding of the variants to specific proteases. The difference in population sizes can be factored out by considering count ratios rather than the raw numbers, applying the log operator to the ratio allows symmetric treatment of mesotrypsin and KLK6 preferences. The counts at the neutral gate are evenly split between the other two gates leading to a close-to-zero score for variants that have concentrated mainly at that gate. Thus, we define the specificity index of each APPI variant as:

$$SI = ln\left(\frac{\overline{C_{\text{meso}}} + \left(\frac{1}{2}\right)C_{\text{neutral}} + 1}{\overline{C_{\text{KLK6}}} + \left(\frac{1}{2}\right)C_{\text{neutral}} + 1}\right) \quad (1)$$

where $\overline{C}_{meso}$ and $\overline{C}_{\text{KLK6}}$ are the average read-counts in the specific gates (three experimental repeats) and C_{neutral} is the read-count in the neutral gate. The addition of one to both numerator and denominator, known as the Laplace rule of succession, reduces the noise in estimating the probability of rare events and improves numerical stability. Thus, positive indices indicate mesotrypsin inclination, and the indices of KLK6-inclined variants are negative. Non-specific variants, observed either equally on mesotrypsin and KLK6 gates or only in the neutral gate, score zero. Overall, the dataset was by design equally distributed between mesotrypsin and KLK6 inclined variants (Figure S2).

TABLE 1 An overview of the experimental date of (A) APPI, (B) IGBPG, and (C) GFP

	Single mutation	Double mutants	Triple mutations	>3 mutations	Total
(A) APPI[a]			Variants (instances)		
Interface mutants	133 (1987022)	9045 (336999)	2053 (19153)	33 (168)	11 264 (2343342)
Non-Interface	292 (50247)	411 (3437)	13 (62)	3 (17)	719 (770028)
All	425 (2037269)	9456 (340436)	2066 (19215)	36 (185)	11 983 (3113370)
(B) IGBPG[b]			Variants		
	907	0	0	0	907
(C) GFP[c]			Variants		
	1040	11 826	11 019	21 060	44 945

[a]"Variants" refers to the numbers of different variants, "Instances" depicts the number of sequence reads (at least three per-variant), "Interface mutants" have at least one mutation at an interface position. The data set also contains 716 265 instances of $APPI_{M17G/I18F/F34V}$ the mutagenesis starting point. Each of these variants has counts from the KLK6, mesotrypsin, and neutral gates, from which we calculated the variant's specificity index (SI).
[b]The data consists of denaturation midpoint concentration (C_m) and slope (m), as well as $\Delta\Delta G$ values of 907 single mutations.
[c]The data consists of log fluorescence values of 44 945 variants.

2.2.2 | IGBPG

Denaturation midpoint slope (m) and denaturant concentration (C_m), as well as the calculated $\Delta\Delta G$ values were downloaded from the ProtaBank[29] site at https://www.protabank.org/. All 907 variants are single mutations (Table 1B). The template for variant modeling was 1PGA.[30]

2.2.3 | GFP

Experimental data includes the mutations of each variant as well as its fluorescence. These were downloaded from https://doi.org/10.6084/m9.figshare.3102154. Notably this dataset is larger than the other two and includes variants with as much as 15 mutations (Table 1C). The template structure for variant modeling was 2wur.[31]

2.3 | Computational data—variant models and features

For each target we generated three sets of features: 40 structural features (MESHI), $20 * proteinLength$ one-hot sequence features (OH), and $40 + 20 * proteinLength$ combined set.

To derive the structural features we use the MUTATE program of the MESHI package,[22] and crystal structures of the target proteins. For APPI, we used the crystal structure of the of mesotrypsin and $APPI_{M17G/I18F/F34V}$ complex (PDB ID 5C67). For IGBPG and GPF we used the PDB entries 1GPA and 2WUR, respectively. For each protein variant MUTATE (a) substitutes sidechains in the crystal structure according to the variant sequence; (b) uses the MESHI energy function and the LBFGS[32] minimization algorithm, to adjust the structure of the complex to the modified sidechains; and (c) extracts 40 structural features, which are components of the minimized energy function. The whole procedure takes on average 2 min per variant on a standard desktop computer. Once all the variants are analyzed, the features are z-score normalized (zero mean and SD of one) to reduce the influence of range differences between the features and to adjust them to a common scale.

The MESHI structural features (DATA S1, section F) include the energy of a complex, its energy terms, and components thereof. The energy terms include: (a) standard bonded energy terms (eg, a quadratic bond term) that indicate deviations from standard chemical bond geometry; (b) torsion angle terms that quantify the compatibility of the complex models with the Ramachandran plot and rotamer preferences, at both the residue and protein levels[33]; (c) pairwise potentials adopted from the literature[34-37]; (d) knowledge based pair potentials and meta-features that consider their distribution within the models; (e) hydrogen bond energy terms[38]; (f) an atom environment term that assigns a set of neighbors (proximate atoms) to each atom and compares the composition of this set with the composition of similar sets in native structures; (g) contact terms that quantify the compatibility of the models with the observed, length and atom-type dependent average number of contacts per atom.

Fore sequence features we used one-hot (OH) representation of the sequence. That is, each residue is represented by 20 binary features, 19 of which are "zero" and one (specific for each amino-acid type) is "one."

2.4 | Machine learning algorithms

Given a vector of experimental data such as specificity indices (one datum per variant), and a matrix of features (one row per variant) the ML task is to build a predictive model that associates the matrix rows with the corresponding experimental data. We tested six standard ML algorithms: linear regression (*LR*), support vector machine (*SVM*), least absolute shrinkage and selection operator (*LASSO*), K-nearest neighbors (*K-NN*), neural network (*NN*), and random forest (*RF*). We used an in-house MATLAB implementation of K-NN, and standard MATLAB implementations for the other five (Table S5). The performances of these methods were compared in the context of predicting the protease specificity of APPI variants (Figures S3-1 and S3-2). RF outperformed all the other methods, and we mainly discuss its results. LASSO, however, was used to estimate feature importance.

2.5 | Performance measures

This study uses two major performance measures: correlation of the predicted and observed experimental data, and enrichment of the top performing variants within the highest scoring ones (see formal definition below). The APPI system may also be viewed as a classification problem, with each variant classified as either mesotrypsin-inclined, KLK6-inclned or neutral. For this target we also calculated two classification performance measures: accuracy and area under curve (AUC). To this end, we consider as neutral, variants that are found in the neutral gate at least 10 times more than in any of the other gates, resulting in SIs of ±0.2. Variants with SIs above and below these thresholds were classified as mesotrypsin and KLK6 inclined, respectively.

Enrichment compares the ML prediction with a hypothetical random prediction. For each of these two predictions we calculate I_5, the size of the intersection between the 5% variants with the highest predicted SI and the 5% variants with the highest observed SI. With a random "prediction" the expectation is $I_5^{random} = N * 0.05^2$, where N is the size of the test set. Enrichment is the ratio between this random expectation and the actual size.

$$\text{Enrichment} = I_5^{prediction} / I_5^{random} = \frac{400 * I_5^{prediction}}{N} \tag{4}$$

A better-than-random prediction has an enrichment score above 1. The better the predictor is, a larger fraction of the best variants is indeed selected, and the enrichment increases.

The accuracy and AUC measures are adapted to support a 3-class setting instead of the typically binary ones (success vs failure). Accuracy is the percentage of correct classifications. Formally,

$$\text{accuracy} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{2}$$

where N is the number of classified samples, and y_i is 1 for a correct classification and 0 otherwise. Following Hand and Till,[39] we calculate the three-state AUC as the average of the six pairwise AUC values,

$$\text{AUC} = \frac{1}{c(c-1)}\sum_{i \neq j} \text{AUC}_{ij}, \quad 1 \le i,j \le c \tag{3}$$

where c is the number of classes, and AUC_{ij} is the area under the true-positive-rate vs false-positive-rate (aka ROC) curve of variants from classes i and class j.

2.6 | Feature importance

As mentioned earlier (Section 2.3), this study associates each variant with 40 structural features. However, it seems that not all the features contribute equally to the learning process and the contribution of some of them may be negligible. To identify the most informative features, we use LASSO to rank their usefulness (Figure 1), and a series of predictions with increasing number of features (Figure S3-1).

2.7 | The generation of independent training and test sets for APPI

The set of APPI variants constitutes a connected graph of dependencies. The vertices of this graph are the variants, and shared mutations (the same non-wild-type residue at the same position) constitute edges between variants. A random choice of training and test sets results in numerous edges (common mutations) between their variants. To test the usability of our algorithm in unexplored regions of the mutations landscape, we studied its sensitivity to variant dependencies. To this end we developed a greedy, breadth-first algorithm to break the dependency graph into two disjoint subgraphs. That is, no variant in one set shares a mutation with a variant in the other. Most importantly, when a double or triple mutant is in the test set, none of the corresponding single mutants is in the training set, and vice versa. To this end, a set is seeded by one mutation (position + non-wild-type amino acid), all variants that share this mutation form the first layer, variants that share mutation(s) with any of the variants at the first layer form the second layer, etc. Finally, a layer is chosen for removal, leaving the inner and outer subsets disjoint. The removed layer is chosen to maximize the size of the remaining subsets. The larger of the remaining sets serves for training, and the smaller for testing. Homogenous sets in which more than 80% of the training set variants were either KLK6 inclined or mesotrypsin inclined were ignored. This procedure is applied to all mutants and repeated for all possible seeds. Thus, a variant may appear in several test sets, with different predicted SIs. In the analysis below we consider the average predicted SI.

3 | RESULTS

3.1 | Informative structural features

The 40 structural features used in this study were developed for a different task, protein structure prediction.[23,24] Here we verify their usefulness in the context of mutant outcome prediction. Figure 1 depicts the most informative features in predicting our five targets, based on their LASSO weights. Interestingly, the sets of most informative features are rather diverged, apparently reflecting the differences between the targets, even those that are related to the same protein. While only a few structural features are marked as important by LASSO, a systematic experiment that gradually increase the number of used features (Figure S3-1) suggests that at least 15 structural features are required to reach the best performance on APPI, and that more features do not deteriorate the performance of random forest regression and classification. Thus, considering the variability of feature importance between the targets, we decided to consistently use all the structural features.

3.2 | Regression

To estimate the predictive power of the three feature sets (MESHI, OH, and Combined) on the APPI and IGBPG datasets we trained RF models on random subsets that included two thirds of the variants (training set) and tested their performance on the remaining third. The reverse proportions were used for GFP to accommodate its higher computational cost. The three feature sets were applied to the same training/test splits. Each training-and-prediction experiment was repeated 100 times with different random splits. Figure 2 depicts representative predictions for the three proteins using the 40 structural features. Figure 3 summarizes these experiments presenting the method performance on each target of experimental data (two leftmost columns), as well as the time consumed by the prediction method.

Both sequence and structural feature sets provide considerable correlation and enrichment (left and right columns, respectively) with all targets. In most cases a combined feature set provides the best performance. The enrichment in GFP is a major exception, with the best performance achieved by the sequence features.

The right column in Figure 3 depicts the average computation time of each training-and-test round. Notably the computational cost of using the sequence features is far higher than using the structural ones and even the combined set of features.

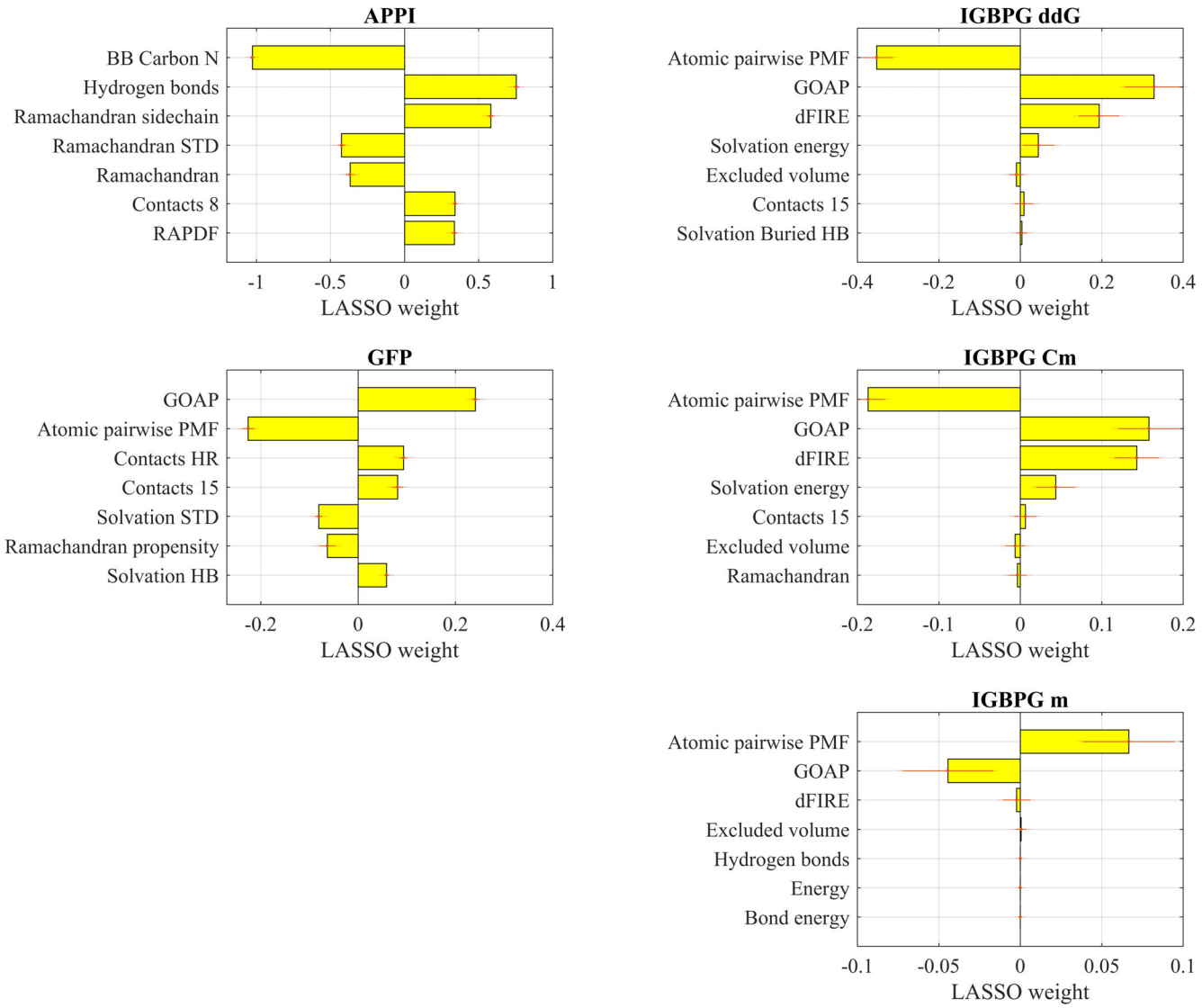

FIGURE 1 Important structural features. Most important structural features for the three target proteins and properties—the structural features are sorted by their LASSO weights. For clarity we show only the top seven features in each case, as the lower rank features have negligible LASSO weights. See DATA S1 section F for the full list of the 40 structural features. Four of these features: Atomic pairwise PMF, GOAP, dFIRE, and RAPDF are MESHI implementations of pairwise energy terms from the literature.[34-37] The Hydrogen bonds,[38] Ramachandran,[33] bond energy, and excluded volume features quantify deviations from ideal geometry of these molecular properties. Ramachandran STD, the SD of the per-residue Ramachandran energy, is an indicator of structural frustration. Ramachandran propensity[33] quantify the propensity of the sequence to the observed backbone torsion angles. Contacts 8 and Contacts 15 are indicators of structural compactness, the average number of contacts (8 and 15 Å, respectively) between carbon atoms. Solvation is a MESHI energy terms that quantifies burial of hydrophobic atoms and hydrogen bonds, and the exposure of hydrophilic and charged atoms. BB carbon N (the number of backbone carbons, including Cβ, in the molecule), Solvation HB, and Solvation buried HB are components of the Solvation energy term. Solvation STD, the SD of per-atom solvation energy, is an indicator of structural frustration

3.3 | Considering the biases of the APPI library

The APPI library is biased, by design, towards binding loop mutations. Figure 4 however demonstrates that we can predict highly specific APPI variants even when they do not include binding loop mutations. Similarly, considering only variants without binding loop mutations (719 variants, all mesotrypsin inclined), resulted in an average correlation of 71% between the observed SIs and the predicted SIs (Figure 5A). To address this issue in a more systematic way we experimented with independent training and test sets, that is no mutation is shared between the sets. Overall, we could find 238 such pairs of sets whose training set includes at least 20% of either KLK6-inclined and mesotrypsin-inclined variants. As the various data set splits are overlapping (a single variant may appear in several test sets) we average each variant's SI predictions. These averaged SI predictions have 72% correlation to the observed SIs, with a 4- and

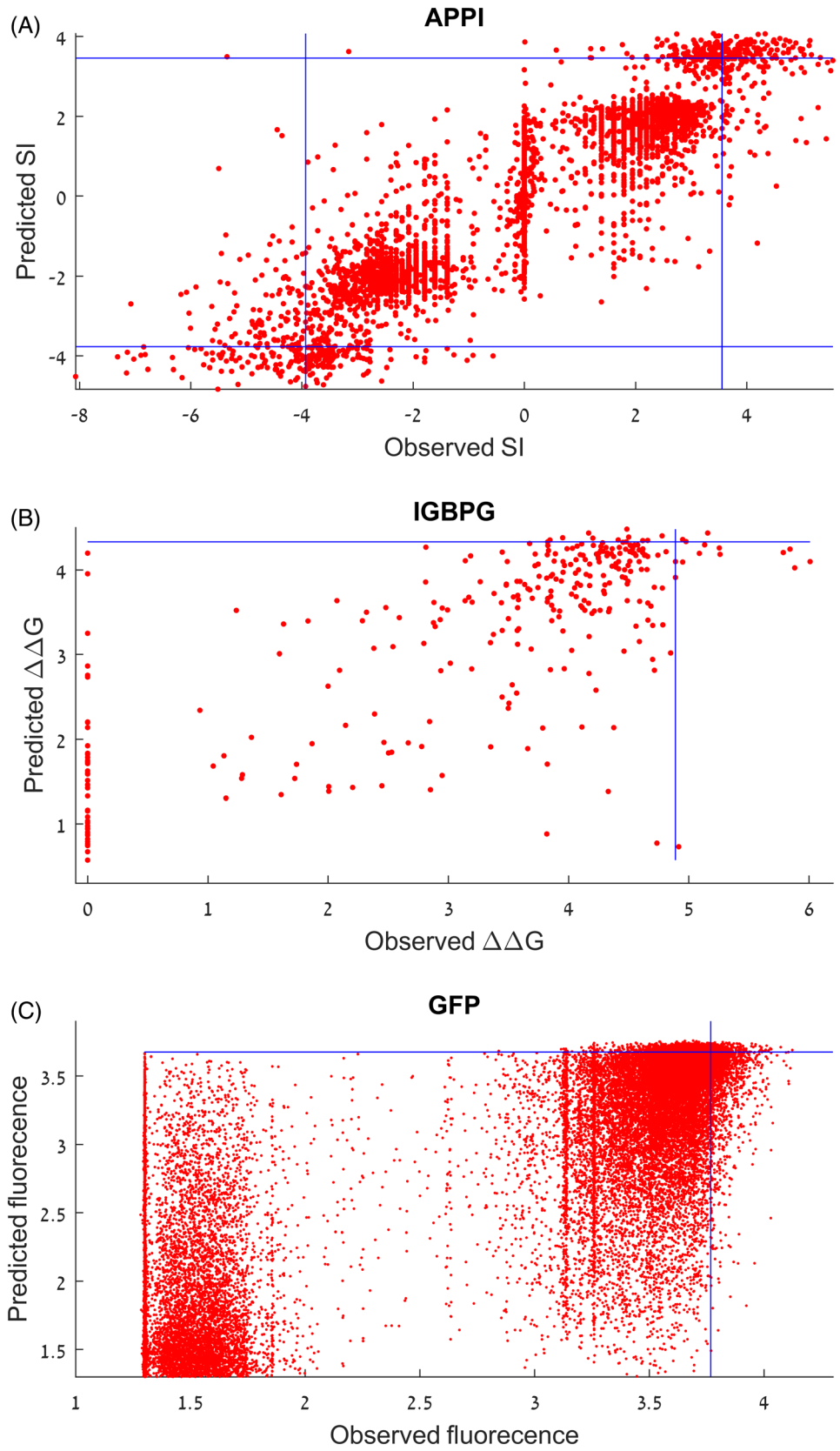

FIGURE 2 Representative predictions of mutation outcome by a random forest regressor using 40 structural features. (A) Predicted vs observed SI values of 3995 randomly selected $APPI_{M17G/I18F/F34V}$ variants (red dots) Positive values imply mesotrypsin inclination, while negative values imply KLK6 preference. The vertical and horizontal blue lines indicate the top (mesotrypsin inclination) and the bottom (KLK6 inclination) 5% observed and predicted variants, respectively. Correlation 0.91, enrichment 10.1 and 8.8 for mesotrypsin and KLK6, respectively. (B) Predicted vs observed $\Delta\Delta G$ values of 303 randomly selected IGBPG variants (red dots). A random forest regressor was trained on the other 604 variants, using 40 structural features. The vertical and horizontal blue lines indicate the top 5% observed and predicted variants, respectively. Correlation 0.75, enrichment 4. (C) Predicted vs observed (log) fluorescence of 29 964 randomly selected GFP variants (red dots). Random forest was trained on 14 982 variants, using 40 structural features. The vertical and horizontal blue lines indicate the top 5% observed and predicted variants, respectively. Correlation 0.82, enrichment 3.5

8-fold enrichment for KLK6 and mesotrypsin, respectively (Figure 5B). As expected, there is a decrease in the prediction quality, in comparison to the full data set, due to the adversarial setting and the considerable reduction in training set size (ranging between 309 to 8813 variants with a median of 2720). Yet evidently, even under these constrains the prediction scheme is very useful (see Section 4).

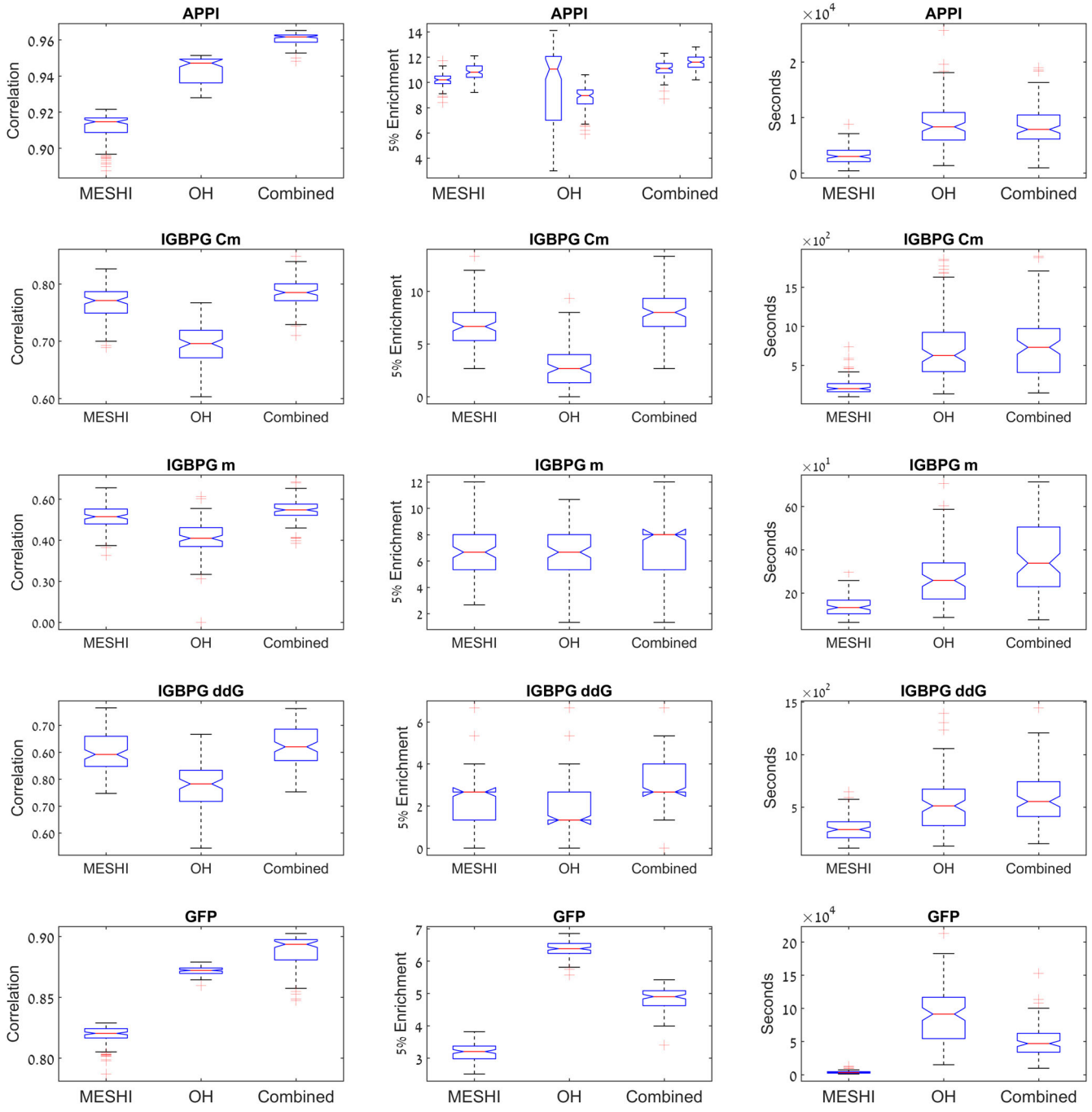

FIGURE 3 Summary of mutation outcome prediction by a random forest regressor. Each box represents 100 rounds of training and test (representative rounds are presented in Figure 1). Each plot depicts an aspect (correlation, enrichment, and computational runtime) of predictions based on 40 structural features (MESHI), one-hot sequence features (OH), and a combination of both (Combined). Each row depicts one of the prediction targets. (A) prediction of APPI selectivity index (SI). (B) Prediction of IGBPG $\Delta\Delta G$, and denaturation midpoint concentration and slope. (C) Prediction of GFP fluorescence. The APPI enrichment is split between KLK6 (left) and mesotrypsin (right)

3.4 | The APPI dataset as a classification problem

The APPI variants may be classified as either neutral (ie, no binding specificity), KLK6-inclined, and mesotrypsin-inclined. In terms of the SI these classes are marked by close-to-zero ($-0.2 \le$ SI ≤ 0.2, 1871 variants), negative (SI < -0.2, 5133 variants), and positive (SI > 0.2, 4979 variants), respectively.

Applying RF to the structural and sequence features provides very similar classification performance, an average classification accuracy of 0.84 and an average area under the ROC curve (AUC) of 0.91.

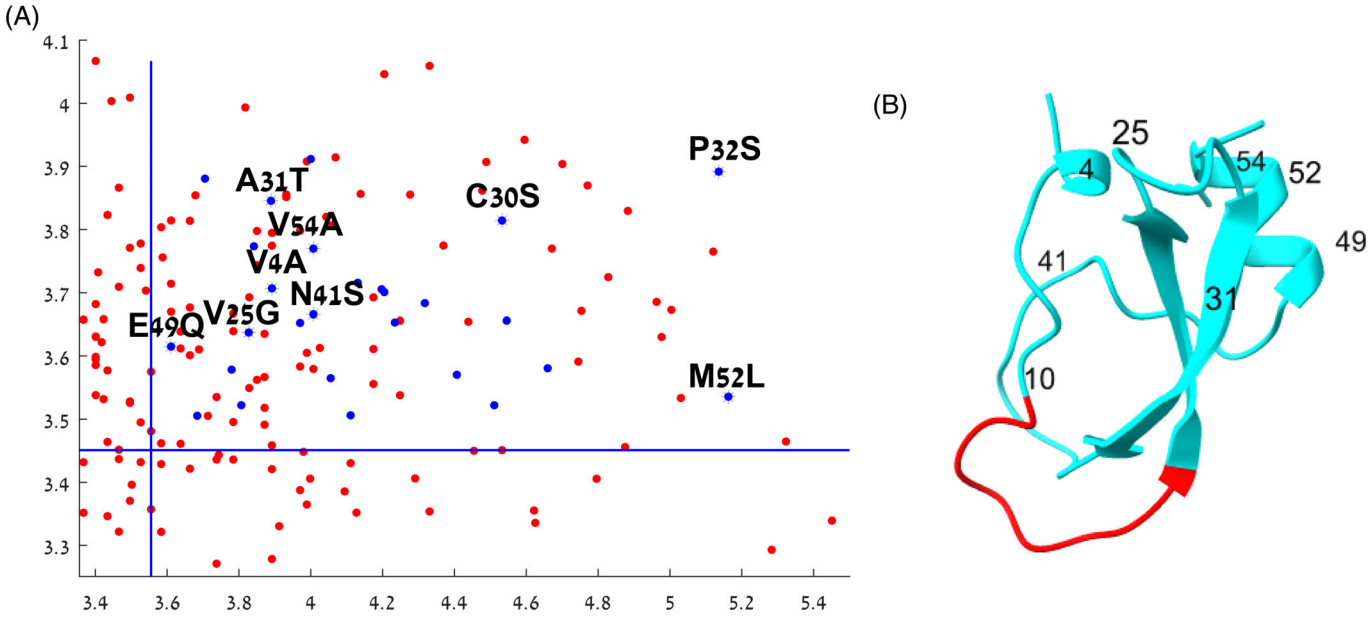

FIGURE 4 A closer look of the top variants. (A) The top-right corner of Figure 1A, the intersection of the 5% top observed SIs and 5% top predicted SIs (blue lines). Overall, at this representative prediction there are 102 such points compared with 10 expected by random, a 10-fold enrichment. Red points indicate variants that have at least one mutant at the eight interface residues. Blue points are the complementary set of variants with mutants only outside the interface. Some of these off-interface variants are explicitly labeled. A complete list of these variants can be found in the supplementary material (Table S4). (B) The positions of the labeled variants in the APPI structure. Note that our machine learning scheme can identify potent mutations away from the binding interface

An even better classification performance is provided by the combined feature set with accuracy of 0.9 and AUC of 0.96 (Table 2, Figure S3-3). The high AUC values indicates that the most pronounced predictions (highest or lowest predicted SI values) are very likely to be assigned to the correct class, in accordance with the regression results above. Low absolute values of predicted SIs, on the other hand, are less decisive.

4 | DISCUSSION

Deep mutational scans provide a wealth of quantitative functional data regarding numerous protein variants. Yet, notwithstanding their power, deep mutational scans can sample only a small fraction of sequence space, even for small proteins. Thus, the discovery of a medically or industrially applicable variants, with the desired properties, may require several rounds of deep scans. Designing such repeated scans requires tough decisions regarding exploration of novel mutations vs exploitation of already identified productive regions of the mutational landscape.

ML may offer shortcuts in this path by extrapolating from the functionally characterized variants to ones that have not been explored yet. Thus, a reliable method for the prediction of mutation outcome may help the design of effective, more focused subsequent library scans.[40] Specifically, an ML model can be used for in silico scanning of virtual libraries exploring stability, binding affinity and specificity landscapes and thus accelerate the search for novel useful proteins. Our enrichment criterion for performance is directly related to this vision. It can be interpreted as the expected benefit of designing a consecutive scan based on a prediction method, compared with a naïve approach of random mutagenesis.

This study demonstrates the ability of ML models to accomplish this promise, by training and testing the models on five target properties of three proteins. Both sequence and structural features proved useful in all targets, and most importantly, they appear to be complementary as the best performance is achieved with both in all, but one tests.

Notably, the best performance measures are for APPI, with median correlation above 90%, and enrichment above 8, for the three feature sets. This may, to some degree, be a consequence of the dataset structure (Figure S2). By design, it includes almost identical number of variants with and without the specificity switch glycine in position 17. Thus, a single OH feature is enough to reach a correlation of 86% and enrichment factor of 2. Interestingly, the structural features seem to perform better in predicting mesotrypsin inclination compared with KLK6 (Figures 2A, 3, top panel, and 5B). We speculate that this is due to the use of a mesotrypsin complex as the template for all the mutant models.

$\Delta\Delta G$ values of IGBPG mutants were also predicted by Nisthal and co-workers[2] using Rosetta,[41,42] FoldX,[7] and PoPMuSiC.[43] The correlations between the observed and predicted values werw 0.65, 0.51, and 0.56, respectively. The substantially better performance of

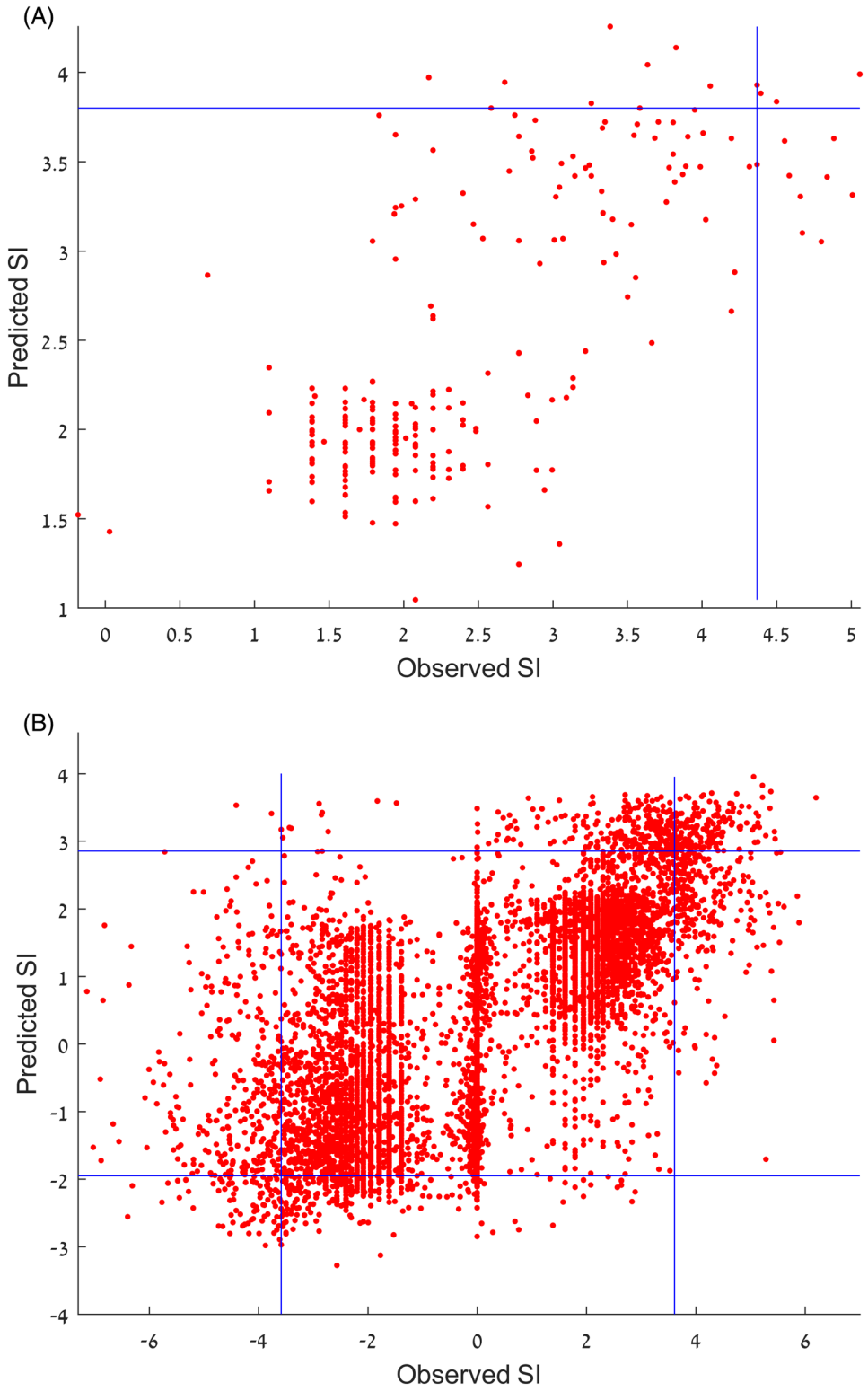

FIGURE 5 Predictions of APPI SI with unbiased training and test sets, and the 40 structural features. (A) Representative predicted vs observed SI values of 237 randomly selected APPI variants (red dots) that do not have mutations in the binding loop. The random forest regressor was trained on the remaining 473 such variants. Correlation is 0.72 with 5-fold enrichment. (B) Prediction with independent training and test sets. Each dot in the figure represents a variant (8147 variants overall). The X-axis represents the variant's observed SI, and the Y-axis depicts its average predicted SI. Correlation is 0.72 and the 5% enrichment is 8-fold for mesotrypsin and 4-fold for KLK6. In this experiment, we discarded training/test pairs of sets, in which more than 80% of the training set variants inclined to the same protease. Notably, most of the variants in the figure are double or triple mutant, predicted without any of their single mutant counterparts in the training set

our method, correlation of 0.76 (Figure 2B and Table 3b), may probably be attributed to the use of data from the same library to parameterize the statistical model.

The fluorescence of GFP variants were predicted by Masso[19] using structural features and data from other variants of the same proteins. The similar approach yielded similar performance (0.83 correlation) to our structural features despite their very different nature (a single four-body potential vs 40 different energy terms and their components). Adding sequence feature allowed us to reach even better results.

The APPI and GFP datasets include numerus variants with two or more mutations and any random split thereof results in variants with shared mutations in the training and test sets. That is, a variant with mutations A and B in the training set and a variant with mutations B and C in the test set. Our success in these experiments suggest that the proposed method can support follow up mutational scans that sample around the most promising variants of a previous round. In this case, the new variants are likely to share at least one mutation with variants of the previous round. The results with the IGBPG dataset that includes only single mutations, and with independent training and tests for APPI (Figure 5B) suggest that our features and ML scheme captures more than additive effects and may also be useful in the design of more exploratory experiments.

An obvious advantage of sequence features, over structural ones, is that they do not require a known structure as the template for variant modeling. Yet, deep mutational scans, the motivation of the current study, are themselves major experimental endeavors applied to biologically, medically, or industrially important proteins. If the structure of such protein is not within the over 150 000 proteins whose structures are already known, it is likely to become the focus of intense structural studies. Further, the results of the last rounds of the Critical Assessment of Structure prediction (CASP)[44] suggest that in the near future structures of proteins might be more readily available. On the other hand, training an RF model with sequence features was computationally expensive for a medium size protein (GFP).

The relative predictive power of the structural and sequence features and their combination differ between the targets and the performance measures (Figure 3). We believe that the three feature types need more exploration before one could generalize their pros and cons. The set of structural features may be missing important aspects of the structure (eg, long range electrostatic interactions) and more sequence features (eg, profile based) would probably prove useful.

Finally, no aspect of the computational scheme was tailored towards the peculiarities of the current targets. Thus, its success suggests that it would be useful when applied to other protein libraries and may add a major lever to the protein design toolbox.

TABLE 2 Classification performance of random forest for APPI protein with three sets of features. Accuracy is the fraction of test-set variants correctly classified to neutral, KLK6-inclined, and mesotrypsin-inclined (SI = ±0.2), (SI < −0.2), and (SI > 0.2), respectively. AUC is the average of the three pairwise AUC values. Each combination was tested 115 times with random splits of training and test sets (7998 and 3995 variants, respectively)

	Structural features (40)	One hot sequence features (54 × 20)	Combined features
Accuracy (±STD)	0.84 (±0.007)	0.85 (±0.015)	0.903 (±0.07)
AUC (±STD)	0.91 (±0.008)	0.91 (±0.014)	0.96 (±0.004)

TABLE 3 Mean performance in the prediction of (A) specificity index (SI) of APPI variants by random forest; (B) $\Delta\Delta G$ and denaturation midpoint concntration (C_m) and slope (m) of IGBPG variants, by random forest; and (C) (log) fluorocence of GFP variants by random forest

	Structural features	One Hot features	Combined features
(A)[a]			
Correlation (±STD)	0.91 (±0.007)	0.86 (±0.005)	0.95 (±0.003)
Enrichment (±STD)	10.8 (±0.633)	7.58 (±0.719)	11.6 (±0.549)
(B)[b]			
$\Delta\Delta G$			
Correlation (±STD)	0.75 (±0.038)	0.69 (±0.040)	0.76 (±0.039)
Enrichment (±STD)	2.63 (±1.568)	1.75 (±1.281)	2.97 (±1.378)
m			
Correlation (±STD)	0.48 (±0.069)	0.40 (±0.080)	0.52 (±0.062)
Enrichment (±STD)	3.57 (±1.677)	1.76 (±1.277)	3.49 (±1.621)
C_m			
Correlation (±STD)	0.77 (±0.028)	0.69 (±0.035)	0.79 (±0.245)
Enrichment (±STD)	7.19 (±2.066)	3.24 (±1.990)	7.72 (±2.113)
(C)[c]			
Correlation (±STD)	0.82 (±0.008)	0.87 (±0.003)	0.89 (±0.014)
Enrichment (±STD)	3.22 (±0.291)	6.38 (±0.241)	4.14 (±0.366)

[a]Each set of features was tested 115 times with random splits of training and test sets (7998 and 3995 variants, respectively).
[b]Each set of features was tested 100 times with random splits of training and test sets (604 and 303 variants, respectively).
[c]Each set of features was tested 100 times with a random split to training and test sets (14 981 and 29 964 variants, respectively).

ACKNOWLEDGMENTS

The authors are grateful for support by Grant no. 1122/14 from the Israel Science Foundation (ISF) to Chen Keasar, and by the European Research Council (ERC) Grant no. 336041 and the Israel Science Foundation (ISF) Grant no. 615/14 to Niv Papo. The authors are also grateful to Tamar Keasar and Eitan Bachmat for thoughtful discussions and support. Finally, the authors thank the anonymous reviewers of this manuscript for their thoughtful and helpful suggestions.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Niv Papo https://orcid.org/0000-0002-7056-2418
Chen Keasar https://orcid.org/0000-0003-0994-1520

REFERENCES

1. Rouet R, Lowe D, Christ D. Stability engineering of the human antibody repertoire. *FEBS Lett*. 2014;588:269-277.
2. Nisthal A, Wang CY, Ary ML, Mayo SL. Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis. *Proc Natl Acad Sci U S A*. 2019;116:16367-16377.
3. Yang KK, Wu Z, Arnold FH. Machine-learning-guided directed evolution for protein engineering. *Nat Methods*. 2019;16:687-694.
4. Vigneri R, Squatrito S, Sciacca L. Insulin and its analogs: actions via insulin and IGF receptors. *Acta Diabetol*. 2010;47:271-278.
5. Seeliger D, de Groot BL. Protein Thermostability calculations using alchemical free energy simulations. *Biophys J*. 2010;98:2309-2316.
6. Pokala N, Handel TM. Energy functions for protein design: adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. *J Mol Biol*. 2005;347:203-227.
7. Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J Mol Biol*. 2002;320:369-387.
8. Benedix A, Becker CM, de Groot BL, Caflisch A, Böckmann RA. Predicting free energy changes using structural ensembles. *Nat Methods*. 2009;6:3-4.
9. Dehouck Y, Grosfils A, Folch B, Gilis D, Bogaerts P, Rooman M. Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics*. 2009;25:2537-2543.
10. Dehouck Y, Kwasigroch JM, Rooman M, Gilis D. BeAtMuSiC: prediction of changes in protein–protein binding affinity on mutations. *Nucleic Acids Res*. 2013;41:W333-W339.
11. Barlow KA, Ó Conchúir S, Thompson S, et al. Flex ddG: Rosetta Ensemble-based estimation of changes in protein–protein binding affinity upon mutation. *J Phys Chem B*. 2018;122:5389-5399.
12. Geng C, Xue LC, Roel-Touris J, Bonvin AMJJ. Finding the ΔΔG spot: are predictors of binding affinity changes upon mutations in protein–protein interactions ready for it? *WIREs Comput Mol Sci*. 2019;9:e1410.
13. Goldenzweig A, Fleishman SJ. Principles of protein stability and their application in computational design. *Annu Rev Biochem*. 2018;87:105-129.
14. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. *Nat Methods*. 2014;11:801-807.
15. Gasperini M, Starita L, Shendure J. The power of multiplexed functional analysis of genetic variants. *Nat Protoc*. 2016;11:1782-1787.
16. Naftaly S, Cohen I, Shahar A, Hockla A, Radisky ES, Papo N. Mapping protein selectivity landscapes using multi-target selective screening and next-generation sequencing of combinatorial libraries. *Nat Commun*. 2018;9:3935.
17. Gray, V. E., Hause, R. J., Luebeck, J., Shendure, J. & Fowler, D. M. Quantitative missense variant effect prediction using large-scale mutagenesis data. *Cell Syst*. 2018;6:116-124.e3.
18. Kumar R, Raghava GPS. Hybrid approach for predicting coreceptor used by HIV-1 from its V3 loop amino acid sequence. *PLoS One*. 2013;8:e61437.
19. Masso M. Accurate and efficient structure-based computational mutagenesis for modeling fluorescence levels of *Aequorea victoria* green fluorescent protein mutants. *Protein Eng Des Sel*. 2020;33:gzaa022. https://doi.org/10.1093/protein/gzaa022.
20. Masso M, Vaisman II. Accurate and efficient gp120 V3 loop structure based models for the determination of HIV-1 co-receptor usage. *BMC Bioinform*. 2010;11:494.
21. Masso M, Vaisman II. Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. *Bioinformatics*. 2008;24:2002-2009.
22. Kalisman N, Levi A, Maximova T, et al. MESHI: a new library of Java classes for molecular modeling. *Bioinformatics*. 2005;21:3931-3932.
23. Elofsson A, Joo K, Keasar C, et al. Methods for estimation of model accuracy in CASP12. *Prot*. 2018;86:361-373.
24. Mirzaei S, Sidi T, Keasar C, Crivelli S. Purely structural protein scoring functions using support vector machine and ensemble learning. *IEEE/ACM Trans Comput Biol Bioinform*. 2019;16(5):1515-1523.
25. Sarkisyan KS, Bolotin DA, Meer MV, et al. Local fitness landscape of the green fluorescent protein. *Nature*. 2016;533:397-401.
26. Hockla A, Miller E, Salameh MA, Copland JA, Radisky DC, Radisky ES. PRSS3/Mesotrypsin is a therapeutic target for metastatic prostate cancer. *Mol Cancer Res*. 2012;10:1555-1566.
27. Sananes A, Cohen I, Shahar A, et al. A potent, proteolysis-resistant inhibitor of kallikrein-related peptidase 6 (KLK6) for cancer therapy, developed by combinatorial engineering. *J Biol Chem*. 2018;293:12663-12680.
28. Cohen I, et al. Combinatorial protein engineering of proteolytically resistant mesotrypsin inhibitors as candidates for cancer therapy. *Biochem J*. 2016;473(10):1329-1341. https://doi.org/10.1042/BJ20151410
29. Wang CY, Chang PM, Ary ML, et al. ProtaBank: a repository for protein design and engineering data. *Protein Sci*. 2019;28:672.
30. Gallagher T, Alexander P, Bryan P, Gilliland GL. Two crystal structures of the B1 immunoglobulin-binding domain of streptococcal protein G and comparison with NMR. *Biochemistry*. 1994;33:4721-4729.
31. Shinobu A, Palm GJ, Schierbeek AJ, Agmon N. Visualizing proton antenna in a high-resolution green fluorescent protein structure. *J Am Chem Soc*. 2010;132:11093-11102.
32. Liu DC, Nocedal J. On the limited memory BFGS method for large scale optimization. *Math Program*. 1989;45:503-528.
33. Amir E-AD, Kalisman N, Keasar C. Differentiable, multi-dimensional, knowledge-based energy terms for torsion angle probabilities and propensities. *Proteins*. 2008;72:62-73.
34. Summa CM, Levitt M. Near-native structure refinement using in vacuo energy minimization. *Proc Natl Acad Sci U S A*. 2007;104:3177-3182.
35. Zhou H, Skolnick J. GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophys J*. 2011;101:2043-2052.
36. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci*. 2002;11:2714-2726.
37. Samudrala R, Moult J. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J Mol Biol*. 1998;275:895-916.

38. Levy-Moonshine A, Amir ED, Keasar C. Enhancement of beta-sheet assembly by cooperative hydrogen bonds potential. *Bioinformatics*. 2009;25:2639-2645.
39. Hand DJ, Till RJ. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Mach Learn*. 2001;45:171-186.
40. Yang K, Wu Z, Arnold F. Machine-learning-guided directed evolution for protein engineering. *Nature methods*. 2019;16:687-694.
41. Das R, Baker D. Macromolecular modeling with Rosetta. *Annu Rev Biochem*. 2008;77:363-382.
42. Kellogg EH, Leaver-Fay A, Baker D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins*. 2011;79:830-838.
43. Dehouck Y, Kwasigroch JM, Gilis D, Rooman M. PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinform*. 2011;12:151.
44. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP): round XIII. *Proteins*. 2019;87:1011-1020.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Sarfati H, Naftaly S, Papo N, Keasar C. Predicting mutant outcome by combining deep mutational scanning and machine learning. *Proteins*. 2021;1-13. https://doi.org/10.1002/prot.26184