WHAT MAKES THE EFFECT OF PROTEIN MUTATIONS DIFFICULT TO PREDICT?

A PREPRINT

Floris van der Flier[1]
floris.vanderflier@wur.nl

Dave Estell[2]
david.estell@iff.com

Sina Pricelius[3]
sina.pricelius@iff.com

Lydia Dankmeyer[3]
lydia.dankmeyer@iff.com

Sander van Stigt Thans[3]
sander.van.stigt.thans@iff.com

Harm Mulder[3]
harm.mulder@iff.com

Rei Otsuka[3]
rei.otsuka@iff.com

Frits Goedegebuur[3]
frits.goedegebuur@iff.com

Laurens Lammerts[3]
laurens.lammerts@iff.com

Diego Staphorst[3]
diego.staphorst@iff.com

Aalt D.J. van Dijk[1]
aaltjan.vandijk@wur.nl

Dick de Ridder[1]
dick.deridder@wur.nl

Henning Redestig[3]
henning.redestig@iff.com

September 25, 2023

[1]Department of Plant Sciences, Wageningen University & Research Wageningen, 6708 PB, The Netherlands
[2]Health & Biosciences, International Flavors and Fragrances, Palo Alto, 94304 CA, USA
[3]Health & Biosciences, International Flavors and Fragrances, Oegstgeest, 2342 BG, The Netherlands

ABSTRACT

Protein engineering increasingly relies on machine learning models to computationally pre-screen variants to identify those that meet the target requirements. Although machine learning approaches have proven effective, their performance on prospective screening data has room for improvement. Prediction accuracy can vary greatly from one variant to the next. So far, it is unclear what characterizes variants that are associated with large model error. We designed and generated a dataset that can be stratified based on four structural characteristics (buriedness, number of contact residues, proximity to the active site and presence of secondary structure), to answer this question. We found that variants with multiple mutations that are buried, closely connected with other residues or close to the active site, which we call challenging mutations, are harder to model than their counterparts (i.e. exposed, loosely connected, far from the active site). This effect emerges only for variants with multiple challenging mutations, since single mutations at these sites were not harder to model. Our findings indicate that variants with challenging mutations are appropriate benchmarking targets for assessing model quality and that stratified dataset design can be leveraged to highlight areas of improvement for machine learning guided protein engineering.

Keywords Protein engineering · Machine learning · Mutation predictability

1 Introduction

Enzymes are nature's catalysts, allowing reaction rates that are orders magnitude higher than those of their non-biological counterparts (1). They are also versatile, catalyzing many different reactions that are relevant in addressing societal challenges, such as carbon capture (2) and plastic degradation (3). Enzymes are used in many commercially relevant applications, such as stain removal on fabric or dish surfaces, starch degradation for beer brewing or glucose production from plant waste. Protein engineering seeks to optimize the performance of enzymes in the context of an application through the introduction of mutations. These mutations are often be selected by variant effect prediction (VEP), a field that is of interest not only to protein engineering, but also in medicine (4) and agriculture (5), where mutations induce changes in phenotype.

Machine learning models that predict the effect of mutations can guide protein engineering by computationally pre-screening protein variants in numbers that far exceed the capacity of experimental screens. Machine learning models used in protein engineering set themselves apart from other domains of VEP in that the properties/functions of interest are often non-native, i.e. not aligned with what would confer increased fitness of the organism. The exact desired biochemical mode of improvement is often undefined, but rather characterized with carefully designed assays. The implication of this is that VEP models for protein engineering are generally supervised in nature (6; 7; 8).

Advances of machine learning in the field of structural biology have led to unprecedented accomplishments, like the prediction of structure from sequence (9), sequence from structure (10) and even the joint generation of both sequence and structure (11). However, accurate prediction of the effects of mutations on structure and function remains a challenge (12; 13). Pak et al. tried to use Alphafold to infer the effect of mutations on protein stability and found that it holds little predictive power for this task (14). This result is unsurprising, as Alphafold is trained to predict the structure of a sequence under the assumption that the sequence has a corresponding structure. This assumption is violated when inferring the effect of arbitrary mutations. Although language models are capable of predicting the effect of mutations on native properties (15), their training objective is similarly misaligned with the objective of protein engineering by considering only natural sequences. Finally, property measurements in protein variant screens often suffer from far greater levels of noise than other protein sequence modelling tasks, which can arise from both systematic and stochastic errors.

There is still limited knowledge about the mechanisms that models fail to capture when they make errors. Gaining a better understanding of this may enable us to provide our models with the necessary inductive biases and incorporate additional training modalities, such as structural, dynamic or physicochemical features, that better capture these mechanisms, ultimately leading to more efficient protein engineering practices.

As a first step towards improving property prediction models, we seek to identify when they fail to generalize to test data: what makes the effects of some mutations harder to predict than others? Our analysis focuses on relating the predictability of variant properties to the structural characteristics of their mutated residues. Knowing what determines predictability is useful, because this information can be used to propose variant sequences that models are less likely to make errors on, thereby increasing the detection rate of enzymes with improved properties. Furthermore, it can help improve property prediction models and support benchmarking machine learning models on sequences known to have low predictability, much like how proteins with remote structural homology to training data can be used to assess structure prediction models (16).

To this end, we designed a framework to select positions for mutation with different structural characteristics, and used these to contrast variants with mutations that exhibit those characteristics with variants harboring mutations that do not. In particular, we compared variants with mutations that are 1) buried or exposed, 2) closely or loosely connected with other residues, 3) close to or distant from the active site and 4) part of helices/sheets or part of loops. We hypothesized that mutations that have large effects on the structure and function are more difficult to model and that those mutations are largely defined by their structural characteristics. Buried mutations are surrounded by many residues and may cause more displacement, closely connected mutations are subject to many interactions, mutations close to the active site may perturb the activity to a larger extent and mutations in helices or sheets may disrupt conserved structural patterns.

For each of these four structural characteristics of protein mutations, we ask: "Is there a difference in predictability of variants with mutations that have this structural characteristic when compared to variants devoid of those mutations?". We answered this question by selecting a natural alpha-amylase enzyme with potential use for stain removal of starch residues, creating a large set of variants with multiple mutations and measuring their activity. We repeated this analysis on a dataset with single mutation variants of the same protein, to determine whether it is the interaction of mutations at certain sites that affects predictability.

2 Results

Here, we investigate the influence of different structural properties of mutations on the predictability of their effect, in the context of engineering enzymes for stain removal. We do this by creating a dataset specifically for this purpose, composed of variants of an alpha-amylase that all belong exclusively to one structural class; we will refer to the variants in this dataset as combinatorial variants. We then test the stain removal activity of these variants and analyse the relation between mutation characteristics and predictability of stain removal activity by machine learning. To our knowledge, no dataset is currently available that allows multiple classes of high-order combinatorial variants — i.e. more than two mutations — to be partitioned and contains a sufficient number of samples for the training of machine learning models.

We then repeat our analysis on a site saturation mutagenesis (SSM) dataset containing only single mutation variants, to assess whether the effects we find occur in isolation or whether it is the introduction of multiple mutations in certain classes that causes differences in performance. These variants will be referred to as single variants.

2.1 A protein variant dataset partitioned by structural characteristics of mutations

We consider four types of structural characteristics: buriedness; number of contact residues; distance to the active site; and presence of secondary structure. The aim is to create a dataset that can be split in half based on every characteristic. From here on out, we shall refer to the splits for every structural characteristic as "structural classes", e.g. the structural class of variants with only buried mutations. The dataset design is visualized in Figure 1(b). In short, we assign every position binary labels for all structural classes that each indicate whether it has the corresponding structural characteristic (1) or not (0). We proceeded by creating sets of positions for every unique combination of structural characteristics, $2^4 = 16$ in total. Variants sequences were then obtained by sampling up to eight mutations per variant from the positions in each set individually, resulting in 16 bins of variants with unique combinations of structural characteristics. These sequences were later cloned and screened, resulting in a dataset of 3,706 variants with associated stain removal activity measurements. Further details on the creation of the dataset are presented in the Methods.

2.2 Beneficial and detrimental mutations are found everywhere

We initially set out to determine whether all structural classes contain variants with improved properties. To this end we plotted stain removal activity for all structural classes. The results, shown in Figure 2, show that property distributions are class specific. While activity differences between variants with mutations at helices or sheets and loop regions are not strong, variants with only buried mutations often have minimal activity when compared to variants with exposed mutations, possibly occurring due to conformational rearrangements and loss of stability (17). Despite the differences in activity distributions, all classes contain variants with higher activity than the reference, which is indicated by the dashed line. This demonstrates that the structural characteristics of mutations are weak predictors of their effects, and that using such characteristics to limit the scope of positions to mutate potentially restricts the identification of variants with improved properties. It is therefore of interest to understand how these characteristics influence predictability, enabling targeted model improvement on variants with low predictability.

2.3 Complex models leverage different features than baselines models

The aim of this work is to determine how structural characteristics of mutations influence model predictability. Since the effect may be specific to the model, we evaluate a total of four models. All models consist of a fixed encoder that maps sequences to a numerical representation, followed by a linear regression model that predicts stain removal activity from the encoded sequences. We chose to evaluate encoders spanning a wide range of complexity, from one-hot encodings to language model embeddings. We included one model, the RITA regressor, that generates variant features based on the sequence as a whole, in contrast to the other three models that generate features for all amino acids in the sequences. Sequence level encoding ensures that features of amino acids are not modelled additively and could help account for interactions between residues. The models are summarized in Table 1 and more extensively described in the Methods.

To ensure the relevance of comparing the baseline models with the more complex models, we questioned whether these models do in fact leverage different information, or if they simply learn to reconstruct the sequence from their embeddings and make predictions in the same way as the baseline models. We therefore compared the performance of these models on sequences with mutations at positions that were not mutated in the training set with their performance on random folds of the training set. Since the baseline models can only generalize to mutations at positions encountered in the training set, the extent to which these models rely on the additional features available to them should be reflected by the performance on test sequences with mutations at positions not encountered in the training set.

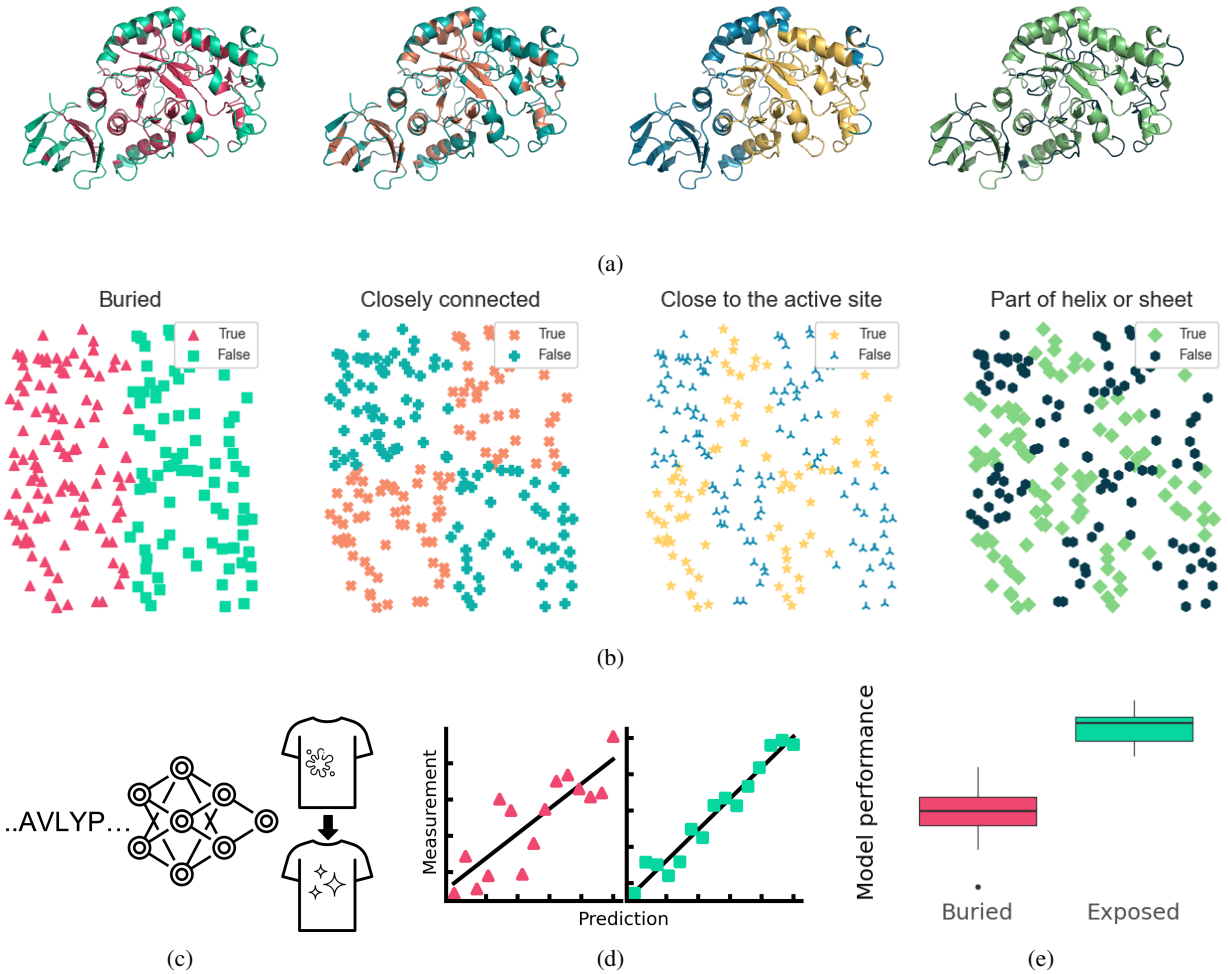


Figure 1: Overview of the approach. (a) We select mutations based on four structural characteristics: buriedness; number of contact residues; distance to the active site; and presence of secondary structure. Every mutation position was then assigned "positive" or "negative" for each characteristic. (b) Variant sequences were then designed by sampling mutations such that the resulting set of sequences can be divided in two parts of roughly equal size that are positive and negative for each structural characteristic. These variants are cloned, expressed and screened for stain removal activity. (c, d) Machine learning models are trained to predict the stain removal activity of the variants based on their sequences. Every structural class of variants is individually trained and evaluated. (e) The performance on held out training data of positive and negative samples is then compared to determine the influence of each characteristic on predictability.

We evaluated the Potts regressor and the RITA regressor on each of the 16 bins, after training on the remaining 15 bins, i.e. stratified cross-validation. The factorial design of our dataset ensures that every bin contains variants with mutated positions that are not mutated in any other bin. We also evaluated these same models by doing 16-fold random cross validation on the entire dataset. Figure 3 compares the performance of these models and shows that the more complex models indeed identify additional features that are independent of position information.

2.4 Effects of combined mutations with specific characteristics are harder to predict

Next, we examined how these characteristics affect the predictability of activity by machine learning models. Differences in prediction performance for a given split provide insight into the difficulty of modelling the corresponding structural characteristic of mutations. The factorial design of the dataset enabled us to train and evaluate models separately on all structural characteristics, and compare the performance between models trained on both splits individually for a given structural characteristic, e.g. buried and exposed. To minimize the potential for confounding effects, our cloning strategy ensured that all structural classes have a comparable number of samples and a similar distribution of the number

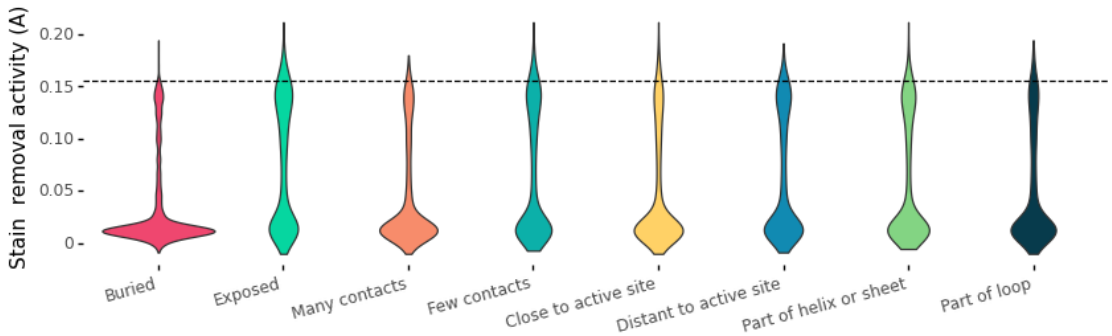


Figure 2: Violin plots of stain removal activity measured by absorbance (A) for every structural characteristic. Stain removal activity of the reference enzyme is indicated by the dashed black line. Every structural split of the data contains improved variants.

Table 1: Four models are selected with different encoding strategies. Pretraining refers to the data used to train the embedding models. L: maximum length of protein sequence.

	Encoded by	**Pretraining**	**Dimensionality**
Residue Agnostic Regressor	Residue	None	L
Partial Least Squares	Residue	None	$L \times 20$
Potts regressor	Residue	MSA	L
RITA regressor	Sequence	UniRef-100 + MGnify + Metaclust	4096

of mutations (Table 2) and degree of diversity within each structural class, as measured by edit distance (Supplementary Figure S6).

Figure 4(a) shows large differences in median performance between class positive and negative variants for the characteristics buriedness, number of contact residues and distance to the active site. Apparently, combinatorial variants are more difficult to model when their mutated sites are buried, close to the active site or in contact with many residues. Going forward, we will refer to mutations at these sites as "challenging mutations". We observe slightly better predictions for mutations occurring at helices or sheets versus loop regions.

The results shown in Figure 4(a) are almost identical across all models (Supplementary Figure S1), which indicates that for difficult classes of variants, there is only marginal improvement in prediction performance compared to the residue agnostic baseline. Although evolutionary models and language models produce rich sequence features and model mutations in the context of all other residues, these results show that the mechanism by which mutations affect stain removal activity in variants with challenging mutations is not well accounted for by these models.

2.5 Mutations with opposite characteristics are harder to predict in single mutation variants

The results above raised the question whether these sites are inherently difficult to model, or whether the combinatorial nature of the variants is the driving factor of the modelling difficulty. To investigate this, we formulated a subsequent experiment where we repeated the analysis on a previously published site saturation mutagenesis dataset containing only single mutation variants for the same alpha-amylase and a comparable assay.

Figure 4(b) displays the results of the predictability analysis carried out on this single variant dataset. We again find significant differences in median performance, but this time in the opposite direction: sites that we found to be less predictable for combinatorial variants are more predictable in isolation. Another finding is that the residue agnostic baseline performs equally well as models trained on rich sequence encodings (Supplementary Figure S2). Like for the combinatorial variants, richer encodings provide only marginal improvements in performance and do not influence the difference in predictability. This suggests that the effects of mutations at sites that are buried, closely connected to other residues or close to the active site are typically less dependent on the amino acid than effects of mutations at opposing sites (exposed, loosely connected and distant to the active site).

The contrasting results between single and combinatorial variants suggest that mutations at sites that are buried, closely connected or close to the active site are not inherently more difficult to predict. Rather, it is their occurrence alongside

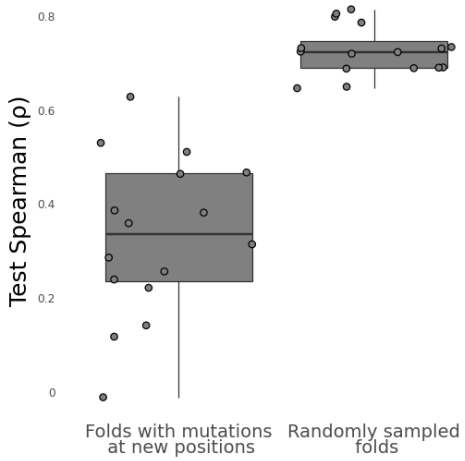


Figure 3: Test Spearman correlations of the RITA regressor over 16 different folds of the combinatorial variant data. Left ("Folds with mutations at novel positions"): performance on variants with mutations at positions not observed in the training data. Right ("Randomly sampled folds"): performances of models on randomly split folds. Similar results are obtained with the Potts regressor (Supplementary Figure S4).

other mutations with the same structural characteristic that makes effect prediction harder, possibly resulting from epistasis (18).

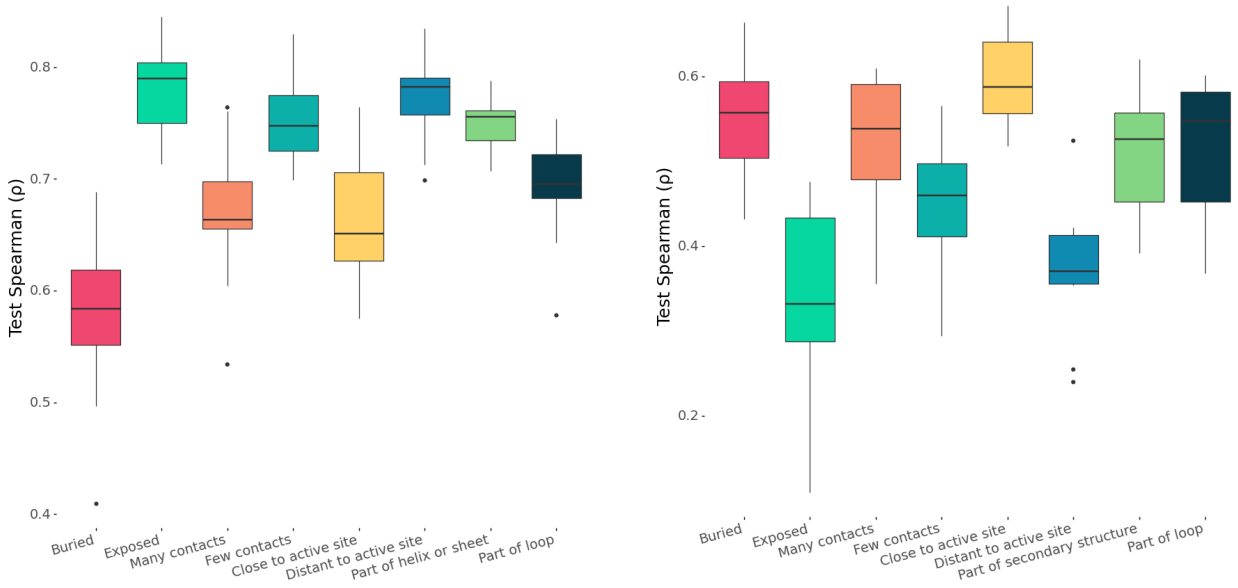
To check whether the difference observed between combinatorial and single variants did not result from the different properties that were predicted, we repeated this experiment on a site saturation mutagenesis dataset of an entirely different enzyme and assay (19), and observed similar effects (Supplementary Figure S3).

3 Discussion

In this study, we have shown that mutation effects of combinatorial protein variants are harder to predict at certain sites than at others. In particular, we observe that variants with mutations at sites that are buried, in close contact with other residues and close to the active site are harder to predict than their counterparts. This is not true for variants with single mutations; in fact, there we observe that mutation effects at the aforementioned sites are easier to predict than their counterparts. These findings remain when using different sequence embedding techniques and even when using a regression model trained solely on information about which residue is mutated. From this we conclude that effects of single mutations at positions that are buried, in close contact with many residues or close to the active site are often similar, regardless of which amino acid is substituted in. On the other hand, the introduction of multiple mutations at such sites negatively impacts performance, which may be the result of epistasis.

We expected our baseline regression models to be affected the most by mutations that interact with other mutations, since they model the input features in an additive manner. While Potts model embeddings do not model residues in isolation, they are limited to second order interactions and may not capture higher order interactions that could be present in our data. Autoregressive language models theoretically do have the ability to capture such interactions, but the RITA regressor predicts the effects of challenging mutations only marginally better. A possible explanation could be that the effect of interactions are specific to the protein property we try to model, and that fixed language model embeddings are not salient enough for predicting these effects for a given property. It may thus be necessary to tailor language model embeddings to properties of interest, for example by finetuning them to directly predict properties of interest.

Further insights could be obtained by evaluating new predictions strategies on all structural classes of combinatorial variants. We assume that all classes of variants are affected by the same noise distributions caused by stochastic and systematic error. A perfect model should therefore perform equally well on positive and negative classes of variants. A goal of using these different classes of variants as benchmarks would be to minimize the performance difference, while preserving the performance on the class negatives, i.e. variants with mutations that are exposed, loosely connected and far away from the active site.



(a) (b)

Figure 4: Test Spearman correlation score distributions of the RITA regressor on 10 folds of the combinatorial training data (a) and the single variant training data (b). Similar results are observed for all other models (Supplementary Figure S1 and Supplementary Figure S2). The models exhibit significantly worse prediction performance on combinatorial variants with mutations at sites that are buried, close to the active site, and in contact with many residues. On the single variant training data in contrast, the models exhibit significantly *better* prediction performance on variants with mutations at sites that are buried, close to the active site, and in contact with many residues. Only variants with mutations occurring within helices or sheets are slightly more difficult to predict than their counterparts.

3.1 Alternative approaches to protein engineering

A common strategy in protein engineering is to generate site saturation mutagenesis (SSM) libraries (20). Here, every possible mutation of a reference sequence for a set of properties of interest is screened. New sequences are designed by combining mutations that yield the highest improvement in all properties, assuming additivity. This study reveals that a large portion of residues interact in non-trivial ways with potential for beneficial effects, which could be left undetected when relying on SSM approaches.

De novo protein design has advanced successfully by designing functional proteins with new folds, often with low sequence identity to existing proteins (21; 22; 10; 23). This approach holds promise for the discovery of new enzymes with desired functions. However, given that even single mutations can improve properties significantly, there is no indication yet that the identification of proteins with improved properties requires the design of sequences with low similarity to natural sequences.

3.2 Limitations of existing protein variant datasets

To date, there are many protein variant datasets for various proteins and accompanying properties that resulted from different methods of sampling the mutational landscape (24; 25). The majority are comprised of single mutation variants, although some datasets are combinatorial, i.e. contain variant sequences with more than one mutation. Some protein variant datasets are limited to a small subset of residue positions (26), while others have no constraints on the positions from which mutations are sampled (27).

An inherent limitation of combinatorial protein variant datasets to address our research questions is that stratification by mutated positions is hindered by variants with overlapping mutated positions. For instance, when comparing variants with mutations exclusively in the core of the protein with those with mutations only on the surface, all variants that contain mutations both in the core and on the surface would have to be disregarded. Tsuboyama et al. recently published a large folding stability dataset comprising both single and double mutants of 331 natural and 148 de novo designed proteins (28). While the dataset presented here is far smaller, it is of an enzyme of significant size (425 aa) instead of

single domains, with higher-order combinatorial variants and mutations covering nearly all residues as opposed to small subsets – all features often encountered in protein engineering.

Our dataset is unique in its potential for stratification by mutation characteristic, while also being an exemplar of a typical protein engineering dataset encountered in industrial scenarios. Both the dataset itself, as well as the design strategy serve as instruments to advance machine learning guided protein engineering.

3.3 Challenges and prospects for future research

One of the central challenges in protein engineering is predicting the fitness of variants that lie beyond the training data domain. To design proteins with improved fitness, it is essential to predict the effect of new mutations as well as properties that exceed the domain of the training data (29). Fannjiang et al. make the case that both choices in training set design and providing models with domain-specific inductive biases hold the most potential in achieving good extrapolation to novel mutations and improved properties.

Furthermore, the nature of the training data poses additional issues. Sequences are not sampled independently, as most protein variant datasets are generated with the purpose of property optimization. Variants generated using new design strategies will have a different distribution than the original dataset. This introduces a selection bias that breaks the independent sampling assumption (i.i.d.) underlying most machine learning methods. Models trained on the original dataset may thus encounter domain shift when used to make predictions on the new dataset.

Finally, there is still room for improvement in performance evaluation. New methods in machine learning guided protein engineering are often evaluated by their rank correlation on randomly held out data. This does not satisfactorily address the capacity to extrapolate, nor does it deal well with the non-i.i.d. nature of protein variant data. Yang et al. have proposed strategies of evaluating extrapolation by creating data splitting strategies based on the dependent variable and number of mutations (30). Still, these strategies are not yet widely adopted and can be improved upon, e.g. by assuring mutually exclusive mutated positions in the splits. Our work demonstrates how dataset design plays an important role in illuminating data characteristics that cause modelling challenges, and provides data for targeted model improvement.

4 Methods

4.1 Data

4.1.1 Sampling of combinatorial variant sequences

We created a dataset of variants with multiple mutations that can be stratified into two sets of roughly the same size for four structural characteristics. The number of samples per structural characteristic of mutations is described in Table 2.

Calculation of structural characteristics for each position We started by computing all structural characteristics on PDB structure 1UA7 (31), the latest out of the two available structures for the amylase we measure, for every residue on a continuous scale. Buriedness was quantified as the shortest distance of the distances of all atoms in the residue to the convex hull of the protein, which was computed using the ConvexHull class of the Spatial module of scipy version 1.9.2. The degree to which residues are connected was computed by counting the number of C_α atoms within a given radius from the C_α atom of the considered residue. This radius was set to the average distance of neighboring C_α atoms plus two corresponding standard deviations, which in the case of this enzyme resulted in 7.3 Ångström. The distance to the active site was found by computing the distances of a residue C_α atom to the C_α atoms of the two residues at the active site (D176, E208), and then taking the shortest of these distances. Lastly, secondary structure was assessed with the ProDy python package version 2.1.2. (32).

Assignment of binary structural characteristics to positions To turn these continuous structural characteristics into binary values, we computed the median of the respective characteristics and then assigned them a binary label using this median as a threshold. In the case of buriedness, all positions with a value higher than the median were assigned a positive label and negative otherwise. Positions were assigned a positive label for "many contacts" if they had more contacts than the median number. Positions closer to the active site than the median distance were assigned a positive label for being "close to the active site" and negative otherwise. Finally, secondary structure labels were assigned based on the secondary structure assessment as described in the previous paragraph.

Grouping of positions To ensure that we can split our data for every structural characteristic, we opted for a factorial design strategy: for every unique combination of structural characteristic labels, we grouped the positions into sets

Table 2: Number of variants in the combinatorial dataset per structural class in total and per number of mutations.

Structural characteristic	Number of Variants								
	Total	1 Mut.	2 Mut.	3 Mut.	4 Mut.	5 Mut.	6 Mut.	7 Mut.	8 Mut.
Exposed	1,823	254	519	504	320	168	46	10	2
Buried	1,883	235	528	561	354	142	47	16	0
Few contacts	1,905	232	490	532	374	202	61	12	2
Many contacts	1,801	257	557	533	300	108	32	14	0
Distant to active site	1,924	255	537	574	343	157	45	12	1
Close to active site	1,782	234	510	491	331	153	48	14	1
Part of loop	1,747	248	506	501	313	125	42	11	1
Part of helix or sheet	1,959	241	541	564	361	185	51	15	1

associated with a bin. Every bin contains variants with mutations that occur only at positions of the associated set. This results in a total of $4^2 = 16$ bins. Distributions of the structural characteristic values per bin are presented in Supplementary Figure S5.

Sampling of variant sequences The final step in creating a dataset of variant sequences that can be split for every structural characteristic was to sample variants from each bin. To obtain a near equal amount of variants for all negative and positive splits, we sampled equal numbers of variants from all 16 bins. For a given bin, variants were generated by uniformly sampling from each set of positions and mutating residues randomly. We sampled residues of 75% of the variants from protein language model (PLM) likelihood scores of ESM1v (esm1v_t33_650m_ur90s_1) in an effort to optimize the yield of the cloning pipeline, as such scores have been shown to correlate with expression (15). The remaining 25% of variants were sampled uniformly over all possible alternative residues, ensuring that some fraction of the variants are not constrained to diversity that is biased towards higher PLM likelihoods. This yielded 21,436 candidate alpha-amylase variants, of which 3,706 were successfully synthesized, cloned and tested for stain removal activity.

4.1.2 Cloning

Expression cassettes encoding the enzyme variants were constructed through fragment assembly of synthetic DNA sourced from Integrated DNA Technologies. The gene encoding the amylase variant was fused to a signal sequence to drive secretion of the amylase, and flanking sequences to allow for integration into the genome of *Bacilus subtilis* upon transformation. The DNA of obtained transformants was sequenced using Oxford Nanopore Technologies for validation. Culture supernatants containing the secreted amylase variants were used for downstream assays without further purification.

4.1.3 Stain removal screening

Enzyme solutions were screened for stain removal in triplicate on 96-well plates that were prepared with swatches of fabric with starch-based stains (Rice Starch, CS28 `https://www.cftbv.nl/`) applied (33). Non-normalized stain removal activity of all alpha-amylase variants was measured as released soil through absorbance measurements of the supernatant of the incubated swatch-enzyme wells. Final non-normalized activity data were obtained by subtracting a blank absorbance measurement from the raw measurements and taking the average of the triplicate measurements, resulting in values between -1.03×10^{-2} and 2.12×10^{-1}. Concentration normalized stain activity data were also obtained, but were not included in this analysis due to erroneous concentration measurements.

4.2 Single mutation variant datasets

The single variant dataset presented here is an SSM dataset. Out of all 8,500 possible variants, 7,467 were successfully cloned and screened for their activity on glucose polymers with a degree of polymerization of 3. These activity values were subsequently normalized by the measured activity of the reference (unmutated) enzyme and used for fitting the regression models. The dataset contains variants with single mutations at each position, with the exception of position 4, 233, 399 and 417 (34; 35; 36).

We reproduced the differences in predictability of single mutation variants using the ERK2 SSM dataset previously released by Brenan et al. (19).

4.3 Model definitions

We selected four linear regression models, each trained on embeddings ranging in complexity. To allow for higher model complexity, non-linear relations between in- and outputs and utilization of unlabeled sequence data, we used two types of pretrained models to embed our training data for downstream regression of our linear models - the RITA regressor and the Potts regressor. These models leverage rich representations of our sequences and model interactions of mutated residues in our sequences:

- The Potts regressor is a ridge regression model trained on embeddings of a direct coupling analysis (DCA) model (37). DCA uses maximum likelihood to obtain an emission parameter for every possible residue and every possible pair of residues in a multiple sequence alignment (38). Residue energies are then obtained by summing all emission parameters for the observed amino acid in a given sequence, where lower energies reflect amino acids more likely to be observed given the multiple sequence alignment. We generate embeddings by taking the residue energy of all residues in the variant as features.
- The RITA regressor is ridge regression model trained on embeddings from an autoregressive language model. Autoregressive language models are commonly used to generate embeddings of input sequences by taking the final hidden state of the sequence. This method is effective because each hidden state is conditioned on the previous hidden state, allowing for the modeling of interactions between all tokens in the sequence, and thus introducing sensitivity to interactions between mutations. Note that the hidden state is aggregated over the entire sequence. Features thus no longer correspond to individual residues, but to properties of the entire sequence. The regression model is trained to capture the effect of learned sequence features on the mutation effect, rather than simply additive contributions from individual residues.

These two more complex embedding strategies are compared to two baseline approaches that are only trained on one-hot encodings of the variant sequences. The difference between the two baseline models is that the Partial Least Squares models is trained on one hot encodings of the entire sequences, whereas the Residue Agnostic Regressor is trained on one hot encodings of only the positions in the sequences which are mutated. Comparison to these baselines provides information on the effect of additional signal obtained from residue type on modelling difficulty.

The Residue Agnostic Regressor, Potts regressor and RITA regressor were all implemented using the Ridge class from the linear_model module of scikit-learn version 1.1.2 (39). All parameters were left to their defaults. The Partial Least Squares model was also implemented in scikit-learn. We used the PLSRegression class of the cross_decomposition module and left all parameters to their defaults, except for the n_components parameter, which was set to 20. The Potts regressor embeds sequences using DCA algorithm GREMLIN_CPP v1.0 with default parameters, except for the gap_cutoff parameter. We set this parameter to 1 to ensure that emission parameters were computed for every residue in the sequence. The RITA regressor embeds sequences using RITA_xl from the huggingface repository lightonai/RITA_xl (40). RITA_xl produces embeddings with a hidden state dimensionality of 2048 for both sequence directions. We obtained the final embeddings by concatenating the embeddings for both directions, resulting in an embedding dimensionality of 4096.

4.4 Model evaluation

For combinatorial variants, we obtained model performance estimates by training on both the negative and positive variant classes separately for all four mutation characteristics, using 10-fold cross validation and taking the mean Spearman correlation coefficients between predicted and measured stain removal activity over all the test-folds as a performance measure.

For the single mutation variants, we made one adjustment to the 10-fold cross validation procedure: instead of randomly sampling test-folds, we sampled folds by taking random fractions of variants with mutations at each site individually, ensuring that no sites are mutated in the test folds that are not also mutated in the corresponding training folds.

References

[1] J. Hagen, *Industrial catalysis: a practical approach*. John Wiley & Sons, 2015.

[2] A. C. Pierre, "Enzymatic carbon dioxide capture," *International Scholarly Research Notices*, vol. 2012, 2012.

[3] V. Tournier, S. Duquesne, F. Guillamot, H. Cramail, D. Taton, A. Marty, and I. André, "Enzymes' power for plastics degradation," *Chemical Reviews*, vol. 123, no. 9, pp. 5612–5701, 2023.

[4] G. Alliance, D. of Columbia Department of Health, *et al.*, "Understanding genetics: a district of columbia guide for patients and health professionals," 2010.

[5] J. Gressel and A. A. Levy, "Agriculture: the selector of improbable mutations," *Proceedings of the National Academy of Sciences*, vol. 103, no. 33, pp. 12215–12216, 2006.

[6] K. K. Yang, Z. Wu, and F. H. Arnold, "Machine-learning-guided directed evolution for protein engineering," *Nature methods*, vol. 16, no. 8, pp. 687–694, 2019.

[7] Y. Luo, G. Jiang, T. Yu, Y. Liu, L. Vo, H. Ding, Y. Su, W. W. Qian, H. Zhao, and J. Peng, "Ecnet is an evolutionary context-integrated deep learning framework for protein engineering," *Nature communications*, vol. 12, no. 1, p. 5743, 2021.

[8] B. Li, Y. T. Yang, J. A. Capra, and M. B. Gerstein, "Predicting changes in protein thermodynamic stability upon point mutation with deep 3d convolutional neural networks," *PLoS computational biology*, vol. 16, no. 11, p. e1008291, 2020.

[9] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[10] J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. Wicky, A. Courbet, R. J. de Haas, N. Bethel, *et al.*, "Robust deep learning–based protein sequence design using proteinmpnn," *Science*, vol. 378, no. 6615, pp. 49–56, 2022.

[11] S. L. Lisanza, J. M. Gershon, S. W. K. Tipps, L. Arnoldt, S. Hendel, J. N. Sims, X. Li, and D. Baker, "Joint generation of protein sequence and structure with rosettafold sequence space diffusion," *bioRxiv*, pp. 2023–05, 2023.

[12] J. Horne and D. Shukla, "Recent advances in machine learning variant effect prediction tools for protein engineering," *Industrial & engineering chemistry research*, vol. 61, no. 19, pp. 6235–6245, 2022.

[13] F. Pucci, M. Schwersensky, and M. Rooman, "Ai challenges for predicting the impact of mutations on protein stability," *arXiv preprint arXiv:2111.04208*, 2021.

[14] M. A. Pak, K. A. Markhieva, M. S. Novikova, D. S. Petrov, I. S. Vorobyev, E. S. Maksimova, F. A. Kondrashov, and D. N. Ivankov, "Using alphafold to predict the impact of single mutations on protein stability and function," *Plos one*, vol. 18, no. 3, p. e0282689, 2023.

[15] J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, and A. Rives, "Language models enable zero-shot prediction of the effects of mutations on protein function," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29287–29303, 2021.

[16] R. Chowdhury, N. Bouatta, S. Biswas, C. Floristean, A. Kharkar, K. Roy, C. Rochereau, G. Ahdritz, J. Zhang, G. M. Church, *et al.*, "Single-sequence protein structure prediction using a language model and deep learning," *Nature Biotechnology*, vol. 40, no. 11, pp. 1617–1623, 2022.

[17] W. A. Lim, D. C. Farruggio, and R. T. Sauer, "Structural and energetic consequences of disruptive mutations in a protein core," *Biochemistry*, vol. 31, no. 17, pp. 4324–4333, 1992.

[18] T. N. Starr and J. W. Thornton, "Epistasis in protein evolution," *Protein science*, vol. 25, no. 7, pp. 1204–1218, 2016.

[19] L. Brenan, A. Andreev, O. Cohen, S. Pantel, A. Kamburov, D. Cacchiarelli, N. S. Persky, C. Zhu, M. Bagul, E. M. Goetz, *et al.*, "Phenotypic characterization of a comprehensive set of mapk1/erk2 missense mutants," *Cell reports*, vol. 17, no. 4, pp. 1171–1183, 2016.

[20] M. T. Reetz and J. D. Carballeira, "Iterative saturation mutagenesis (ism) for rapid directed evolution of functional enzymes," *Nature protocols*, vol. 2, no. 4, pp. 891–903, 2007.

[21] D. Repecka, V. Jauniskis, L. Karpus, E. Rembeza, I. Rokaitis, J. Zrimec, S. Poviloniene, A. Laurynenas, S. Viknander, W. Abuajwa, *et al.*, "Expanding functional protein sequence spaces using generative adversarial networks," *Nature Machine Intelligence*, vol. 3, no. 4, pp. 324–333, 2021.

[22] I. Anishchenko, S. J. Pellock, T. M. Chidyausiku, T. A. Ramelot, S. Ovchinnikov, J. Hao, K. Bafna, C. Norn, A. Kang, A. K. Bera, *et al.*, "De novo protein design by deep network hallucination," *Nature*, vol. 600, no. 7889, pp. 547–552, 2021.

[23] J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, *et al.*, "Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models," *bioRxiv*, pp. 2022–12, 2022.

[24] P. Notin, M. Dias, J. Frazer, J. M. Hurtado, A. N. Gomez, D. Marks, and Y. Gal, "Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval," in *International Conference on Machine Learning*, pp. 16990–17017, PMLR, 2022.

[25] P. Notin, M. Dias, J. Frazer, J. Marchena-Hurtado, A. N. Gomez, D. S. Marks, and Y. Gal, "Proteingym."

[26] N. C. Wu, L. Dai, C. A. Olson, J. O. Lloyd-Smith, and R. Sun, "Adaptation in protein fitness landscapes is facilitated by indirect paths," *Elife*, vol. 5, p. e16965, 2016.

[27] P. Kennouche, A. Charles-Orszag, D. Nishiguchi, S. Goussard, A.-f. Imhaus, M. Dupré, J. Chamot-Rooke, and G. Duménil, "Deep mutational scanning of the neisseria meningitidis major pilin reveals the importance of pilus tip-mediated adhesion," *The EMBO Journal*, vol. 38, no. 22, p. e102145, 2019.

[28] K. Tsuboyama, J. Dauparas, J. Chen, E. Laine, Y. Mohseni Behbahani, J. J. Weinstein, N. M. Mangan, S. Ovchinnikov, and G. J. Rocklin, "Mega-scale experimental analysis of protein folding stability in biology and design," *Nature*, pp. 1–11, 2023.

[29] C. Fannjiang and J. Listgarten, "Is novelty predictable?," *arXiv preprint arXiv:2306.00872*, 2023.

[30] C. Dallago, J. Mou, K. E. Johnston, B. J. Wittmann, N. Bhattacharya, S. Goldman, A. Madani, and K. K. Yang, "Flip: Benchmark tasks in fitness landscape inference for proteins," *bioRxiv*, pp. 2021–11, 2021.

[31] M. Kagawa, Z. Fujimoto, M. Momma, K. Takase, and H. Mizuno, "Crystal structure of bacillus subtilis α-amylase in complex with acarbose," *Journal of bacteriology*, vol. 185, no. 23, pp. 6981–6984, 2003.

[32] A. Bakan, L. M. Meireles, and I. Bahar, "Prody: protein dynamics inferred from theory and experiments," *Bioinformatics*, vol. 27, no. 11, pp. 1575–1577, 2011.

[33] L. G. Cascao-Pereira, R. Chin, W. A. Cuevas, D. A. Estell, S.-K. Lee, M. J. Pepsin, S. D. Power, S. W. Ramer, C. A. Requadt, A. Shaw, A. R. Toppozada, and L. Wallace, "Uses of an alpha-amylase from bacillus subtilis."

[34] L. G. Cascao-Pereira, W. A. Cuevas, D. A. Estell, S.-K. Lee, S. D. Power, S. W. Ramer, A. Toppozada, and L. Wallace, "Variant alpha-amylases from bacillus subtilis and methods of uses thereof." US Patent, Dec 2012.

[35] L. G. Cascao-Pereira, W. A. Cuevas, D. A. Estell, S.-K. Lee, S. D. Power, S. W. Ramer, A. Toppozada, and L. Wallace, "Variant alpha-amylases from bacillus subtilis and methods of uses thereof." US Patent, Mar 2015.

[36] W. A. Cuevas, S.-K. Lee, S. W. Ramer, A. Shaw, A. R. Toppozada, D. E. Estell, L. Wallace, R. Chin, C. A. Requadt, S. D. Power, and M. J. Pepsin, "Variant alpha-amylases from bacillus subtilis and methods of use thereof." US Patent, Jul 2015.

[37] S. Ovchinnikov, "Gremlin_cpp," 2022.

[38] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, "Direct-coupling analysis of residue coevolution captures native contacts across many protein families," *Proceedings of the National Academy of Sciences*, vol. 108, no. 49, pp. E1293–E1301, 2011.

[39] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[40] D. Hesslow, N. Zanichelli, P. Notin, I. Poli, and D. Marks, "Rita: a study on scaling up generative protein sequence models," *arXiv preprint arXiv:2205.05789*, 2022.

Acknowledgements We thank Viktor Alekseyev for meticulously reviewing the manuscript and providing extensive feedback. We would also like to express our gratitude to the members of the Bioinformatics group at Wageningen University for discussion and input.

Data and code availability Both the data and code used generate our findings and accompanying figures are made available at `https://github.com/florisvdf/mutation-predictability`.

A Supplementary Materials

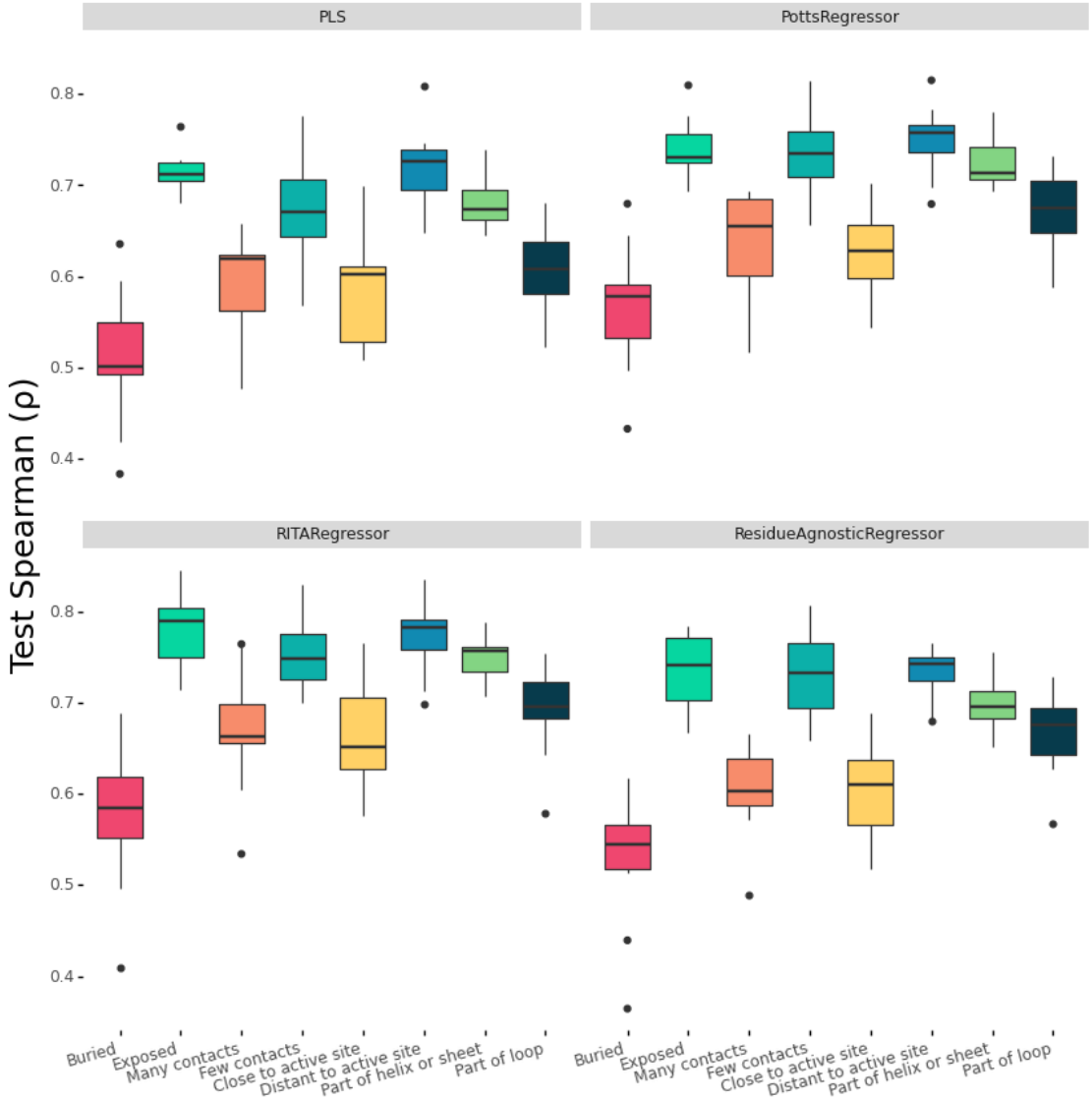


Figure S1: Test Spearman correlation score distributions of all models on 10 folds of the combinatorial training data. The models exhibit significantly worse prediction performance on combinatorial variants with mutations at sites that are buried, close to the active site, and in contact with many residues.

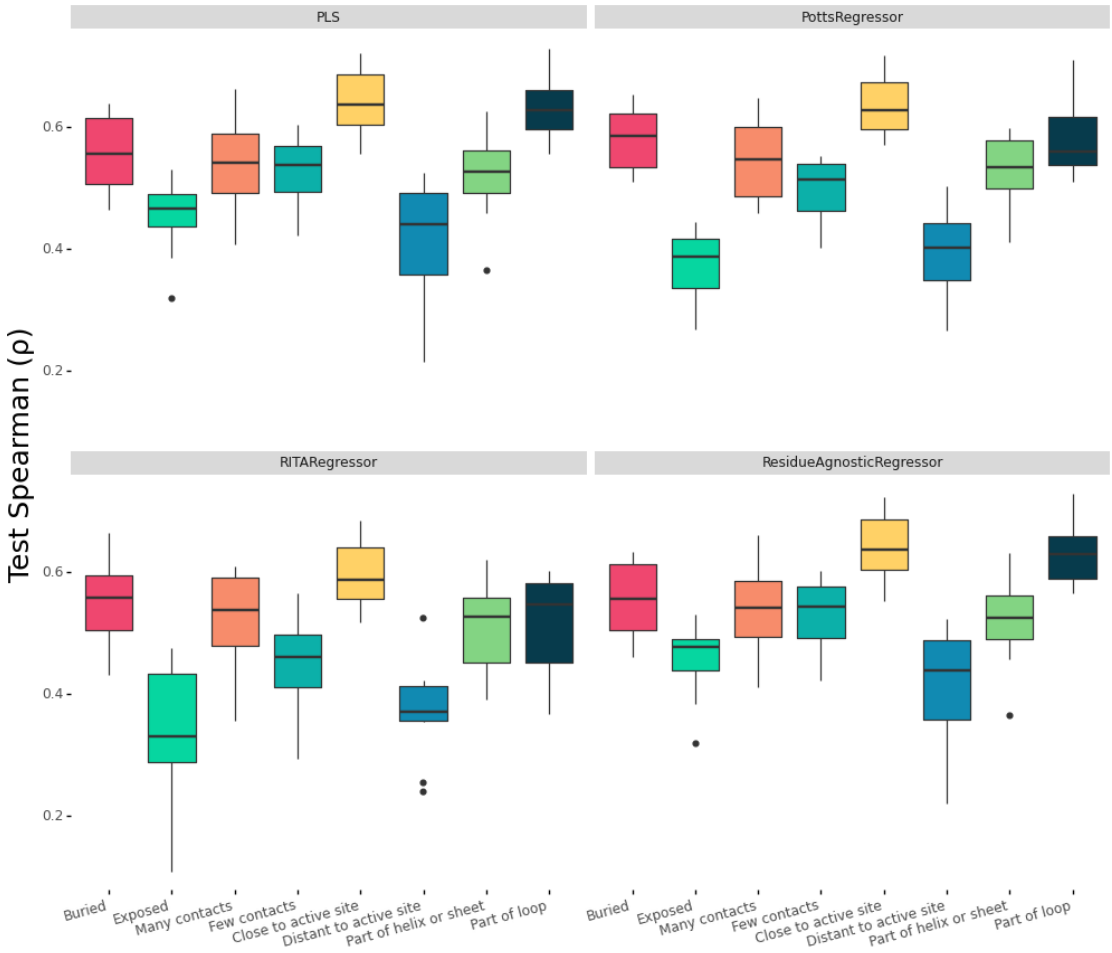


Figure S2: Test Spearman correlation score distributions of all models on 10 folds of the single variant training data. The models exhibit significantly *better* prediction performance on variants with mutations at sites that are buried, close to the active site, and in contact with many residues. Only variants with mutations occurring within helices or sheets are slightly more difficult to predict than their counterparts.

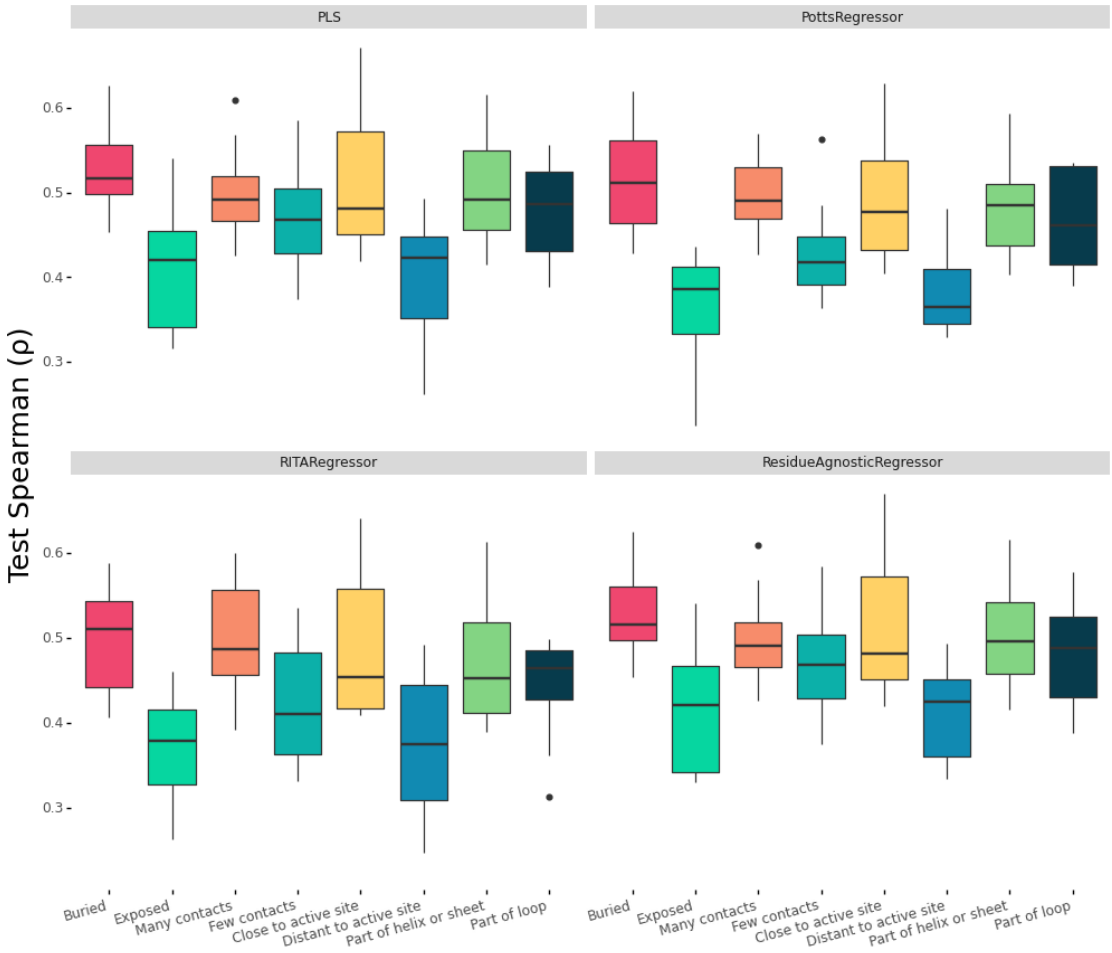


Figure S3: Test Spearman correlation score distributions of all models on 10 folds of the ERK2 SSM dataset (19). Models were trained to predict cell response to doxycycline. These results are to a large extent in accordance with the observations in our single variant dataset, with the exception of predictability differences observed for secondary structure.

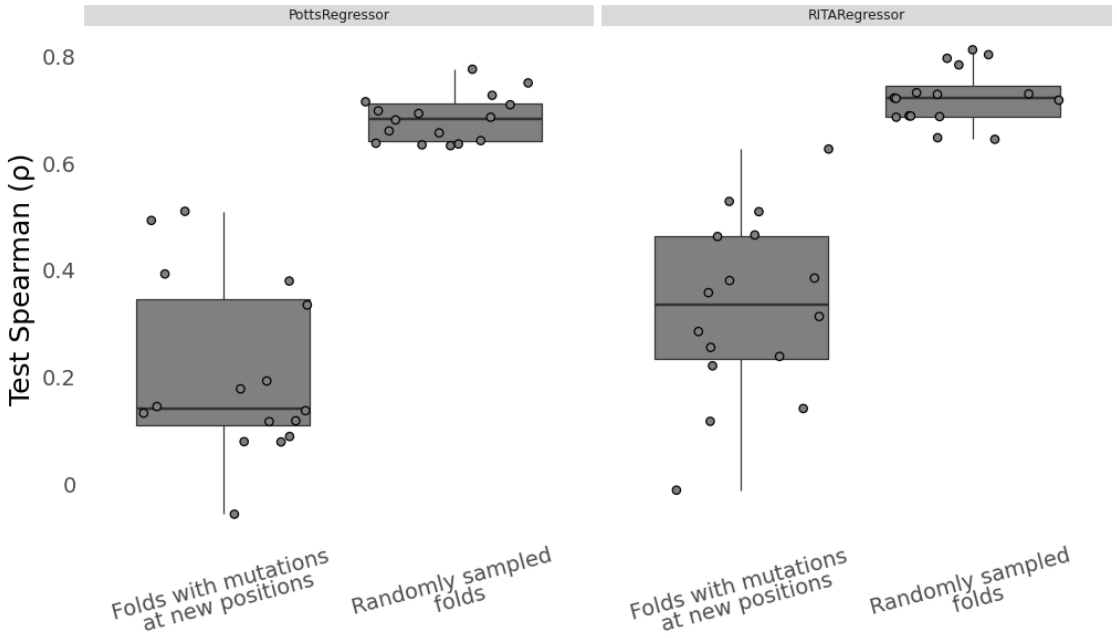


Figure S4: Test Spearman correlations of the Potts regressor and RITA regressor over 16 different folds of the combinatorial variant data. Left ("Folds with mutations at novel positions"): performance on variants with mutations at positions not observed in the training data. Right ("Randomly sampled folds"): performances of models on randomly split folds.

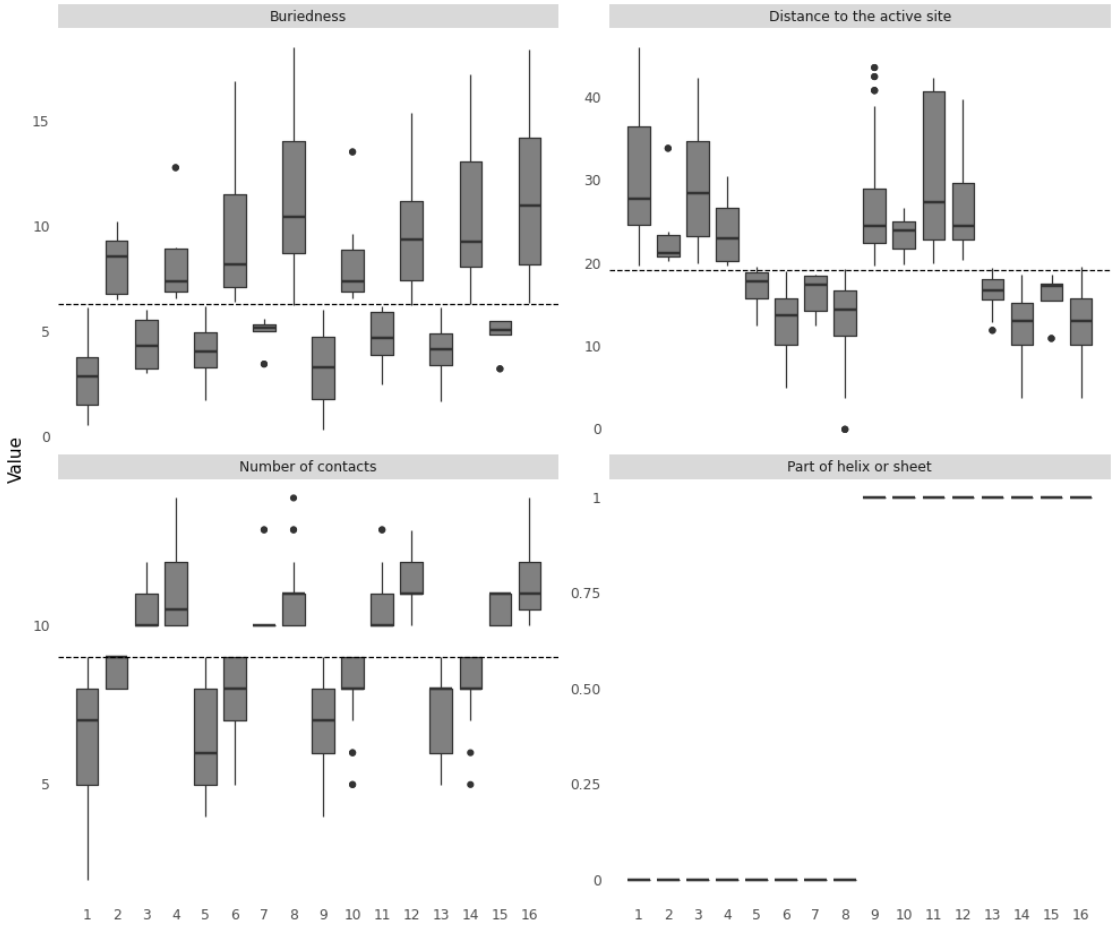


Figure S5: The calculated values of all structural characteristics for every bin. The median value used as threshold in assigning binary labels is visualized with the dashed line. For structural characteristics "Buriedness", "Distance to the active site" and "Number of contacts", bins with values below the median contain positions where residues are buried, are close to the active site and have few contact residues, respectively.

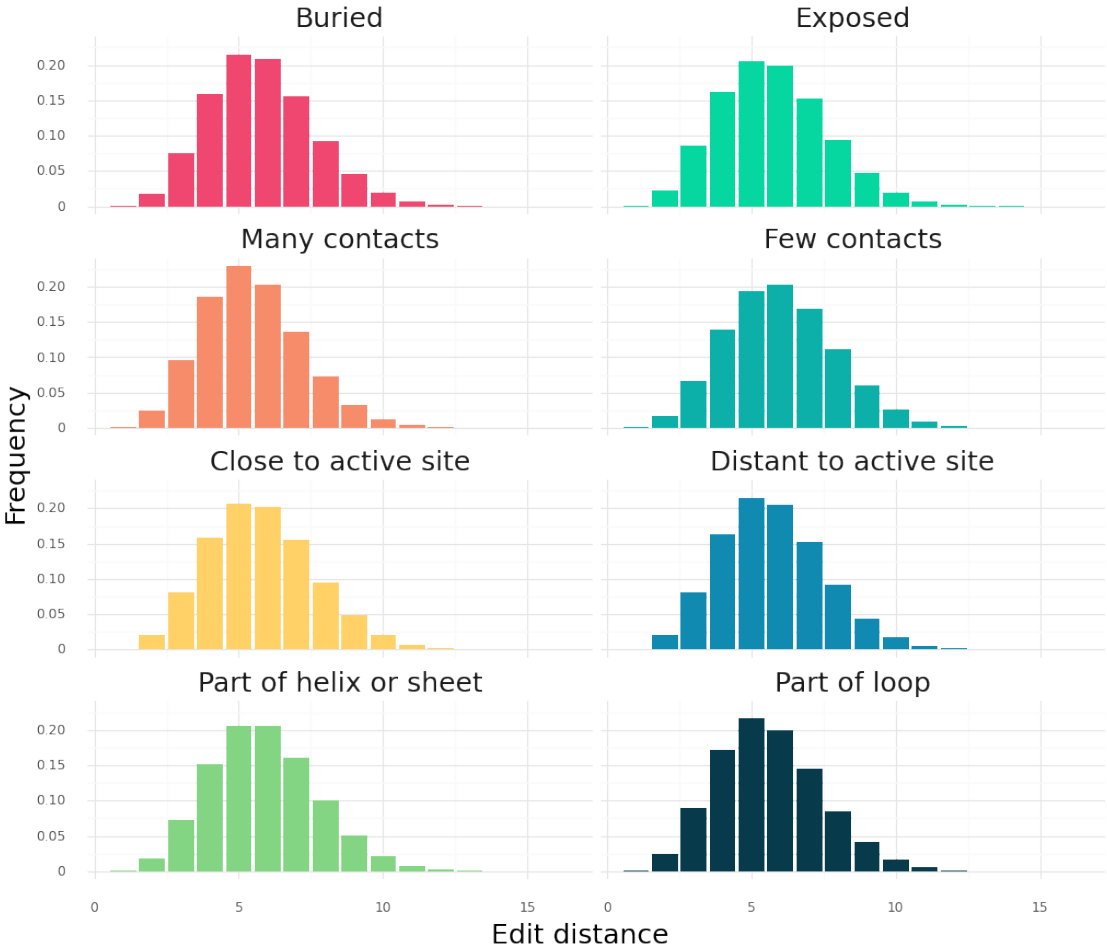


Figure S6: Diversity of the variants in each structural class presented as the distributions of all versus all edit distances within each structural class. Nearly identical distributions of edit distances demonstrate that the degree of variant diversity does not vary significantly among structural classes.