FEW-SHOT PREDICTION OF THE EXPERIMENTAL FUNCTIONAL MEASUREMENTS FOR PROTEINS WITH SINGLE POINT MUTATIONS

Michael Bikman, Rachel Kolodny & Margarita Osadchy *

ABSTRACT

We use few-shot learning to train transformer-based models to predict the experimental measurements of the effects of single point mutations. Using sequence-based inputs, we supervise the pre-training of our models to predict 150,000 normalized measurements derived from various experiments (MAVE data). The pre-training is done on proteins distinct from the one under evaluation. The inputs include embeddings from a protein foundation model, sequence conservation features, and protein stability features. After pre-training, for each experiment, we rely on 2%-5% randomly sampled values in the tested protein to estimate the normalization transform and to fine-tune the model for that experiment. We transform the normalized predictions of the fine-tuned model of all the single point mutations using the inverse of the estimated normalization transform, thus predicting the experimental values, and compare their Mean Absolute Error (MAE) to the true experimental values. We compare the accuracy of predictions in different settings, including ablation to study the contributions of the different inputs.

1 INTRODUCTION

Quantitatively predicting the impact of single point mutations on experimentally measured function is a specific and well-defined instance of the fundamental, multi-faceted, challenge of computationally predicting protein function from sequence. The value of accurate prediction lies in its ability to approximate the functional landscape of sequence space near wild-type proteins. This can inform us of mutational paths that are beneficial to function, which is useful for protein engineering and directed evolution (Luo et al., 2021; Yang et al., 2019; Diaz et al., 2023), or detrimental to function, which has wide spread implications for human health (Stein et al., 2019; Richards et al., 2015). The structures of the protein, both wild type and those of point mutation variants, dictate the functional outcome, including the many cases where loss of function is due to loss of structural stability (Høie et al., 2022). Nonetheless, the success of state-of-the-art machine learning methods in efficiently predicting the wild-type structure from sequence (Jumper et al., 2021; Lin et al., 2023) implies that functional predictions can still be made even when only the sequence is available.

Two relevant data sources for training machine learning (ML) models are all known proteins and MAVE experimental data. Here, the latter is the dataset Høie et al. (2022) curated of 39 multiplexed assays of variant effects, also referred to as deep mutational scans (DMS) for 29 proteins. Each assay includes measurements in a specific experimental setting (e.g., competitive growth assay, antibiotics resistance, two-hybrid assay) of (practically) all possible single point mutations of the assayed protein. The challenge is formidable, as these measurements vary within each assay. Furthermore, the measured values also vary among assays, with markedly different distributions, including even the ranges of values, across different dataset. Protein databases and MAVE differ from each other both in size and in their described functionality. Protein databases are huge, with hundreds of millions of proteins, in stark contrast to only a few dozen MAVE experiments. Protein databases include only well-functioning proteins while MAVE experiments include single point mutations for the full range of function, from well-functioning like those in the protein databases, to non-functioning ones. Namely, the MAVE datasets includes data about mutated proteins that

*Department of Computer Science, University of Haifa, Haifa, 31905, Israel
bikman.m@gmail.com, {trachel,rita}@cs.haifa.ac.il

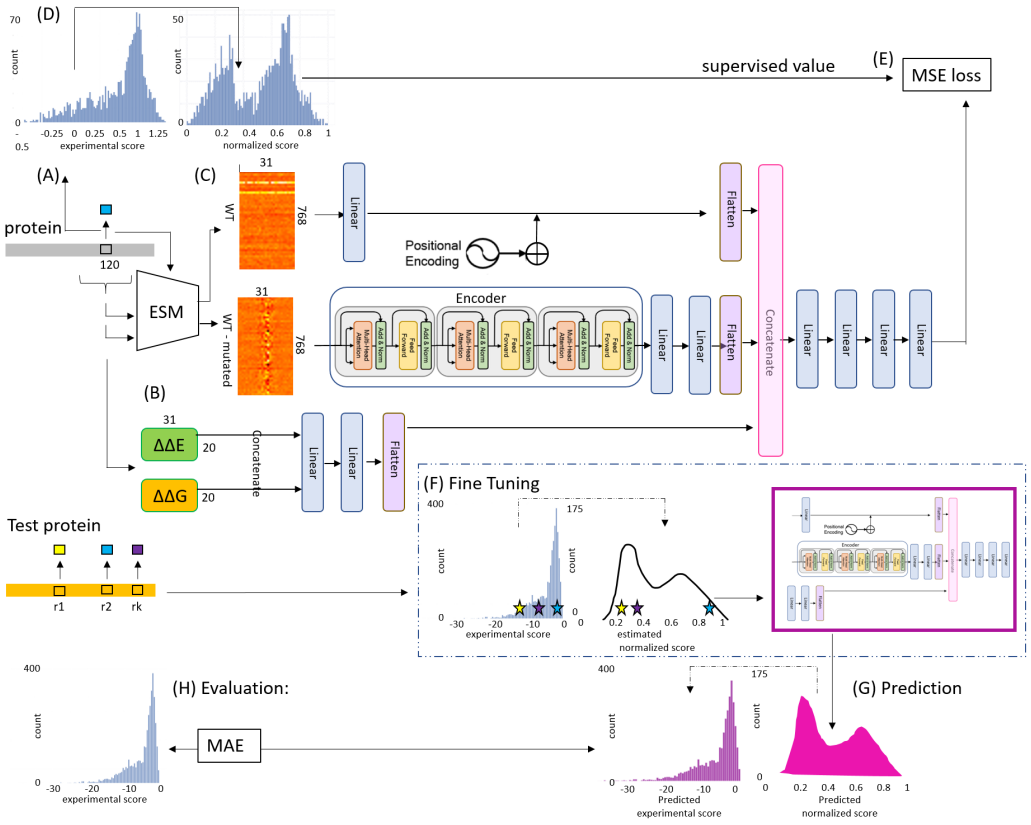

Figure 1: Few-shot prediction model of experimental values of single point mutation variants. The goal in pre-training is to predict the normalized experimental score for a single point mutation (A). We input ddE, ddG values for the single point mutations in a window of 31 residues encompassing the mutation (B), and the ESM-1b embeddings (C). During FT, a few mutations are sampled at random (shown as stars). We estimate a transform that can normalize these values from the sampled experimental scores, and use these normalized scores to fine-tune the model. Finally, we use the FT model (shown boxed in purple) to predict normalized scores to all single point mutations, and the inverse of the estimated transform to convert these to experimental scores (G). These are compared to the true values using and MAE (H)

would have not survived evolutionary selection Tokuriki & Tawfik (2009), and hence cannot appear in the protein databases. Previous studies trained models to address this challenge using data from the protein databases (Riesselman et al., 2018; Marquet et al., 2022; Meier et al., 2021) and adding MAVE experimental data in a supervised manner (Luo et al., 2021; Bepler & Berger, 2021; Høie et al., 2022). As previously highlighted, a relevant signal is the evolutionary record as reflected in the multiple sequence alignment (MSA) of the mutated protein (Laine et al., 2019; Marquet et al., 2022; Meier et al., 2021). Notice that another relevant dataset, not used in this study, is human single point mutation VUS data, that includes human proteins that function poorly (Landrum et al., 2016).

An important distinction from the experimental perspective is whether prediction is zero-shot or few-shot. Zero-shot means no single point mutation scores of the test-set protein are used, and thus there is no need to experimentally collect them. For example, the setup suggested by Høie et al. (2022) relied on MAVE experimental data for supervised training and predicts all functional scores for a protein not in the training set is zero-shot. Models that do not use any MAVE data (Riesselman et al., 2018; Marquet et al., 2022; Meier et al., 2021), the test-set protein in particular, are also zero-shot predictors. Notice that this setup is different from an x-fold cross validation where both training and test sets have a partial set of mutations from all proteins (Luo et al., 2021; Bepler & Berger, 2021).

Zero-shot functional score predictors were evaluated using Spearman's rank correlation, a measure of the strength of association between two ranked variables. Namely, the predicted scores themselves are not evaluated, and effectively have no significance, rather, it is only their relative order that is evaluated. Our goal is to predict experimental measurements of proteins with single point mutations. Because the distributions of experimental measurements vary among the experiments, it is hard to imagine that one could predict the actual scores in zero-shot Diaz et al. (2023). Thus, we shift to using few-shot prediction, predict the experimental values themselves and evaluate our predictions w.r.t to the true values rather than predicting only ordinality.

The most related to our approach is the work by Høie et al. (2022) who used MAVE data to train a random forest (RF) in a supervised setting. Hoie et al. hand-crafted features based on two calculated measures for mutated variants: ddG – Rosetta (Leman et al., 2020) predicted changes in thermodynamic stability, and ddE – GEMME (Laine et al., 2019) -based evolutionary score, which quantifies how likely is a mutation given the evolutionary record as reflected in the MSA. For each protein in their set, they trained a RF based on the data of all other proteins, using normalized scores and then predicted normalized, as opposed to true, scores. Protein Gym(Notin et al., 2023) classify Hoie et al.'s method as a hybrid between zero-shot and supervised learning and thus do not benchmark it; they do not consider nor benchmark our few-shot approach. While Hoie et al. operate in a zero-shot setting and evaluate only the relative order of their predictions, we show here that their normalized scores can be converted into true scores if we one has an estimation of the range of values in the experiment. We did not add these few additional samples and retrain the RF because as RF cannot be fine-tuned, one expects the impact of this very small set of additional samples to be negligible.

Here, we trained a supervised transformer model (Chandra et al., 2023) to predict experimental measurements of all single point mutations, that benefits from few-shot prediction. We train our model on all available training data [1], and further fine-tune it, with a small sample of mutations and their experimental measurements from the tested protein. The input to our model are the ESM-1b embedding (Rives et al., 2021) of a window encompassing the mutated residue, both of the wild-type sequence and the mutated sequence, GEMME ddE scores, and Rosetta ddG scores. As expected, using a few shots of data from the evaluated protein both allows to predict the actual experimental values, and more accurately than in a zero-shot setting. We use a transformer model, as it readily lends itself to fine tuning the model, thus improving predictions when more data becomes available. Working in a few-shot regime which allows predicting the actual experimental values also allows evaluating the predictions with the more straight-forward MAE measure. Finally, we carry out an ablation study to reason about the contribution of the ESM embeddings, ddE, and ddG values to the predictions. We envision using few-shot predictors like the one we present to reduce the work load of carrying out a MAVE experiments to measuring the functional scores of only a few mutations, and having nonetheless good approximations of the true functional scores of single point mutations.

2 Methods and Results

Figure 1(A-C) shows an overview of our attention-based neural network for predicting the experimental values of all single point mutations. Our network is trained in a supervised setting to predict normalized scores from four input types, all confined to a 31-residue window centered around the mutation site (Figure 1(A)). The ddE and ddG input features capture the evolutionary and biophysical properties of single point mutations: ddE is MSA conservation score (Laine et al., 2019) and ddG is the Rosetta score of the thermodynamic stability of the mutated protein (Leman et al., 2020). We input the values of all 31*20 possible mutations in the window (Figure 1(B)), as this signal may be informative to more robustly identify positions in the protein structure that have a significant impact on the experimental prediction. We rely on the pre-calculated values by Høie et al. (2022), and replace missing values by 0s. We also input the protein representation by the protein foundation model ESM-1b, and the difference between this representation and the representation for the mutated protein sequence; we restrict the input to the embedding difference in the 31-residue window encompassing the mutation site (Figure 1(C)). To inform the model about the position of the window in the sequence we encoded the positions of the residues in the window with standard positional encoding (Vaswani et al., 2017).

[1] in leave-one-out setting

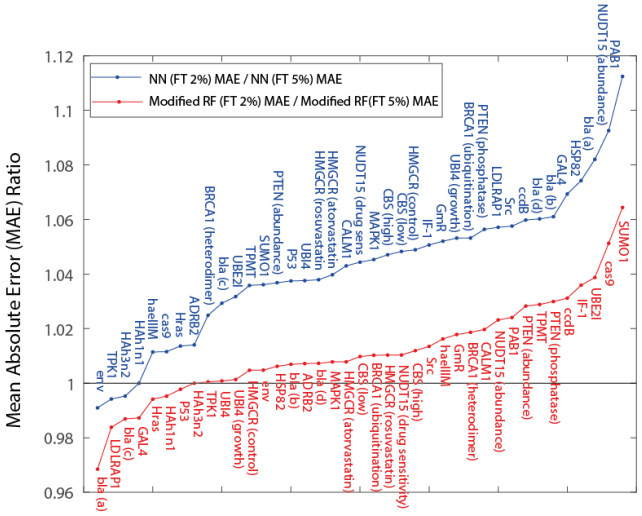

Figure 2: The improvement in MAE when using more random samples. We calculate the MAE for predictions of the experimental values for models FT with 2%,5% of the data, and show the ratio between the MAE of the two NN models (blue). Better models have lower MAE, and thus cases where more data improved the predictions are above 1. As a baseline, we also show the ratio of MAE for the RF predictions transformed into experimental values, where more data points contribute to a more accurate calculation of the transform from normalized values (red).

All inputs are embedded into a common feature space after processing separately each type of information, concatenating the embeddings, and processing them jointly with a 4-layer MLP with ReLu activations. ddE and ddG inputs are concatenated and embedded using 2-layer MLP with ReLu. The ESM-1b embedding difference is processed by a 3-layer standard transformer block encoder, followed by 2 fully connected layers with ReLu activations. ESM-1b embedding of the wild-type is first passed through a fully connected layer with ReLu and then added to the sin-cos positional encoding. Our model has approximately 37.3 M parameters.

Training procedure: We train the model in two stages: supervised pre-training and fine-tuning. We use the MAVE dataset curated by Høie et al. (2022) with 155,000 experimental measurements for 39 experiments for 29 different proteins. We trained 39 models, each geared to test one protein/experiment. During the pre-training of a model, we use all available data, excluding only information associated with the test protein. Supplementary figure 3 shows the histograms of the experimental values from the 39 experiments (blue), highlighting that these scores differ significantly in distribution and even range. To assist the model during training, we normalize the data in the fixed [0,1] range, with the quantile transform normalization as in Høie et al. (2022) [2](1(D)). The normalized score is used with an MSE loss to guide the training of the network (1 (E)). During FT (1 (F)), we sample a small set of experimental data from the tested protein (indicated by colored stars), and use it to FT the pre-trained model. Namely, we employ experimental values of random mutations to estimate the parameters of the quantile transform normalization. The data is then normalized using this estimated transform. Subsequently, these normalized values are utilized to fine-tune the model for 30 epochs (updated network shown encased in a purple box 1(F)). Finally (1(G)), the fine-tuned model predicts normalized values for all mutations, and the inverse of the estimated transform is applied to convert these predictions into actual values. The evaluation is conducted by comparing the predicted experimental values with the true values.

Experiments: We measured the MAE of experimental values vs. those derived from normalized predictions of the fine-tuned transformer models, and the RFs Høie et al. (2022), using 2%, 3%, and 5%, and for models with restricted inputs (for the ablation study). The lengths of the proteins in the dataset range from 65-1023 residues (mean 351), implying that the number possible mutations range from 1,234-19,437 (mean 6663), and hence the number of random samples used in fine-tuning range from 12-194 for every percent. For the transformer models, the experimental values of the randomly sampled points are used to fine-tune its learned weights. In addition, it also serves the purpose of estimating the normalizing transform and its inverse, so that we can convert the normalized predicted values to true experimental values. Supplementary figure 3 also shows the distribution of the predictions of the NN model FT with 5% of the data (green), and the RF predictions (red) that were transformed into experimental values with the same calculated inverse transform. Because the

[2] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.QuantileTransformer.html

results vary so much between experiments, we focused on comparing different prediction settings to better understand viable strategies for accurate prediction. The ranges of the experimental scores, and consequently the MAE, vary in scale. Thus, to compare two settings, we study the (sorted along the x-axis) ratios between MAEs, so as to identify the more accurate predictions with ratio > 1.

Figure 2 compares the settings depending on the amount of sampled data from the tested protein. The ratio comparison (blue) of 2% vs. 5% FT shows that more data improves predictions. Improvement is due both to a more accurately estimated transform, and to better normalized value prediction. The contribution of the improved estimated transform is shown as the improvement in the MAE of the transformed RF (normalized) predictions (red). For the RF model, fine-tuning is not possible, and we did not re-train the RF with the additional sampled data, as a fraction of a few percent is not expected to have a significant impact, thus, we transform the same normalized predictions with the inverse of the estimated transforms. Table A lists the MAE values. We emphasize that using Høie et al. (2022) predictions in this few-shot setting with the inverse normalizing transform, is an extension of their model and we show it here to better understand the properties of the predictions. We note that Høie et al. (2022) did not optimize these predictions (rather, they focused on evaluating normalized predictions via Spearman correlation with true ranking). The contribution of the few-shot samples is two-fold: estimation of the normalizing transform and FT, while the FT part is important, as it improves predictions beyond the improvement in the estimation of the normalizing transform.

Figure 5 shows our ablation study, comparing the FT model (with 5% samples) to models in which the embedding difference, the ddE, and the ddG inputs are omitted. We show in blue the comparison of 2% and 5% FT. In the left panel, we show each of the ratio values sorted for the clarity of presentation; on the right, the order of the tested proteins is maintained. The improvement of the FT model is most dramatic with respect to the model in which the difference embedding is omitted (red), then, in which both ddE and ddG are omitted (cyan). Among the ddG and ddE, it is the omission of ddG has the smaller impact (green).

3 DISCUSSION

Predicting the effects of single point mutations on the experimental values measuring a phenotype is an important yet difficult challenge. As Diaz et al. (2023) highlighted, each experimental system has its own specific biochemistry and is influenced by many factors including, but not limited to, stability. The limited amount of data in each setting is the fundamental challenge to using a straightforward supervised learning solution. Our contribution is to tackle this challenge with few-shot learning and to fine-tune the model to the experiment at-hand.

The inputs pass information learned from the protein databases into the model: the ESM-1b embeddings, ddE, and ddG. Given their public availability (Rives et al., 2021; Laine et al., 2019; Leman et al., 2020), these are easy to compute. Notice that we use the ESM-1b embedding (i.e., the 768 dimensional vector per residue from the one-before-last layer), as opposed to the probabilities the ESM-1b model assigns to different amino acids used by Meier et al. (2021) and Brandes et al. (2023). We use the scarcer and harder to obtain MAVE data to supervise the pre-training of our attention-based model. Finally, we use the most expensive to obtain random samples from the test-protein, both to estimate an inverse transform that allows us to convert normalized predictions into experimental values and to fine-tune the model to the tested protein. Because for the problem at hand there are different types of relevant and available data, some of it is abundant and some very scarce, we believe that any model that is designed to address this challenge must consider ways of incorporating these data (Biswas et al., 2021). Our analysis demonstrates that the fine-tuning is important for improving the accuracy of predictions of the experimental values and that the input that contributes most to the accuracy is the change in the foundation model embedding upon mutation.

While predictions of experimental values are far from satisfying, our study offers a promising path to address this challenge with a few shot learning. We anticipate several future directions will improve predictions: As more MAVE data becomes available, pre-training will be more effective. Meta-learning can help the network adapt better to the FT step, and in this case, even as scarce, randomly sampled, data for different proteins becomes available, the trained models can improve. Experimental tools developed for directed evolution Cobb et al. (2013) can offer such samples. The advantage of having such models is immense, as it will reduce the experimental load on these experiments, and offer a way to better characterize the functional landscape of proteins.

REFERENCES

Tristan Bepler and Bonnie Berger. Learning the protein language: Evolution, structure, and function. *Cell systems*, 12(6):654–669, 2021.

Surojit Biswas, Grigory Khimulya, Ethan C Alley, Kevin M Esvelt, and George M Church. Low-n protein engineering with data-efficient deep learning. *Nature methods*, 18(4):389–396, 2021.

Nadav Brandes, Grant Goldman, Charlotte H Wang, Chun Jimmie Ye, and Vasilis Ntranos. Genome-wide prediction of disease variant effects with a deep protein language model. *Nature Genetics*, 55(9):1512–1522, 2023.

Abel Chandra, Laura Tünnermann, Tommy Löfstedt, and Regina Gratz. Transformer-based deep learning for predicting protein properties in the life sciences. *Elife*, 12:e82819, 2023.

Ryan E Cobb, Ran Chao, and Huimin Zhao. Directed evolution: past, present, and future. *AIChE Journal*, 59(5):1432–1440, 2013.

Daniel J Diaz, Anastasiya V Kulikova, Andrew D Ellington, and Claus O Wilke. Using machine learning to predict the effects and consequences of mutations in proteins. *Current Opinion in Structural Biology*, 78:102518, 2023.

Magnus Haraldson Høie, Matteo Cagiada, Anders Haagen Beck Frederiksen, Amelie Stein, and Kresten Lindorff-Larsen. Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and conservation. *Cell reports*, 38(2), 2022.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Elodie Laine, Yasaman Karami, and Alessandra Carbone. Gemme: a simple and fast global epistatic model predicting mutational effects. *Molecular biology and evolution*, 36(11):2604–2619, 2019.

Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Jeffrey Hoover, et al. Clinvar: public archive of interpretations of clinically relevant variants. *Nucleic acids research*, 44(D1):D862–D868, 2016.

Julia Koehler Leman, Brian D Weitzner, Steven M Lewis, Jared Adolf-Bryfogle, Nawsad Alam, Rebecca F Alford, Melanie Aprahamian, David Baker, Kyle A Barlow, Patrick Barth, et al. Macromolecular modeling and design in rosetta: recent methods and frameworks. *Nature methods*, 17 (7):665–680, 2020.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Yunan Luo, Guangde Jiang, Tianhao Yu, Yang Liu, Lam Vo, Hantian Ding, Yufeng Su, Wesley Wei Qian, Huimin Zhao, and Jian Peng. Ecnet is an evolutionary context-integrated deep learning framework for protein engineering. *Nature communications*, 12(1):5743, 2021.

Céline Marquet, Michael Heinzinger, Tobias Olenyi, Christian Dallago, Kyra Erckert, Michael Bernhofer, Dmitrii Nechaev, and Burkhard Rost. Embeddings from protein language models predict conservation and variant effects. *Human genetics*, 141(10):1629–1647, 2022.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34:29287–29303, 2021.

Pascal Notin, Aaron Kollasch, Daniel Ritter, Lood van Niekerk, Steffanie Paul, Han Spinner, Nathan Rollins, Ada Shaw, Rose Orenbuch, Ruben Weitzman, Jonathan Frazer, Mafalda Dias, Dinko Franceschi, Yarin Gal, and Debora Marks. Proteingym: Large-scale benchmarks for protein fitness prediction and design. In A. Oh, T. Neumann,

A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 64331–64379. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/cac723e5ff29f65e3fcbb0739ae91bee-Paper-Datasets_and_Benchmarks.pdf.

Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genetics in medicine*, 17(5):405–423, 2015.

Adam J Riesselman, John B Ingraham, and Debora S Marks. Deep generative models of genetic variation capture the effects of mutations. *Nature methods*, 15(10):816–822, 2018.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

Amelie Stein, Douglas M Fowler, Rasmus Hartmann-Petersen, and Kresten Lindorff-Larsen. Biophysical and mechanistic models for disease-causing protein variants. *Trends in biochemical sciences*, 44(7):575–588, 2019.

Nobuhiko Tokuriki and Dan S Tawfik. Stability effects of mutations and protein evolvability. *Current opinion in structural biology*, 19(5):596–604, 2009.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Kevin K Yang, Zachary Wu, and Frances H Arnold. Machine-learning-guided directed evolution for protein engineering. *Nature methods*, 16(8):687–694, 2019.

A Appendix

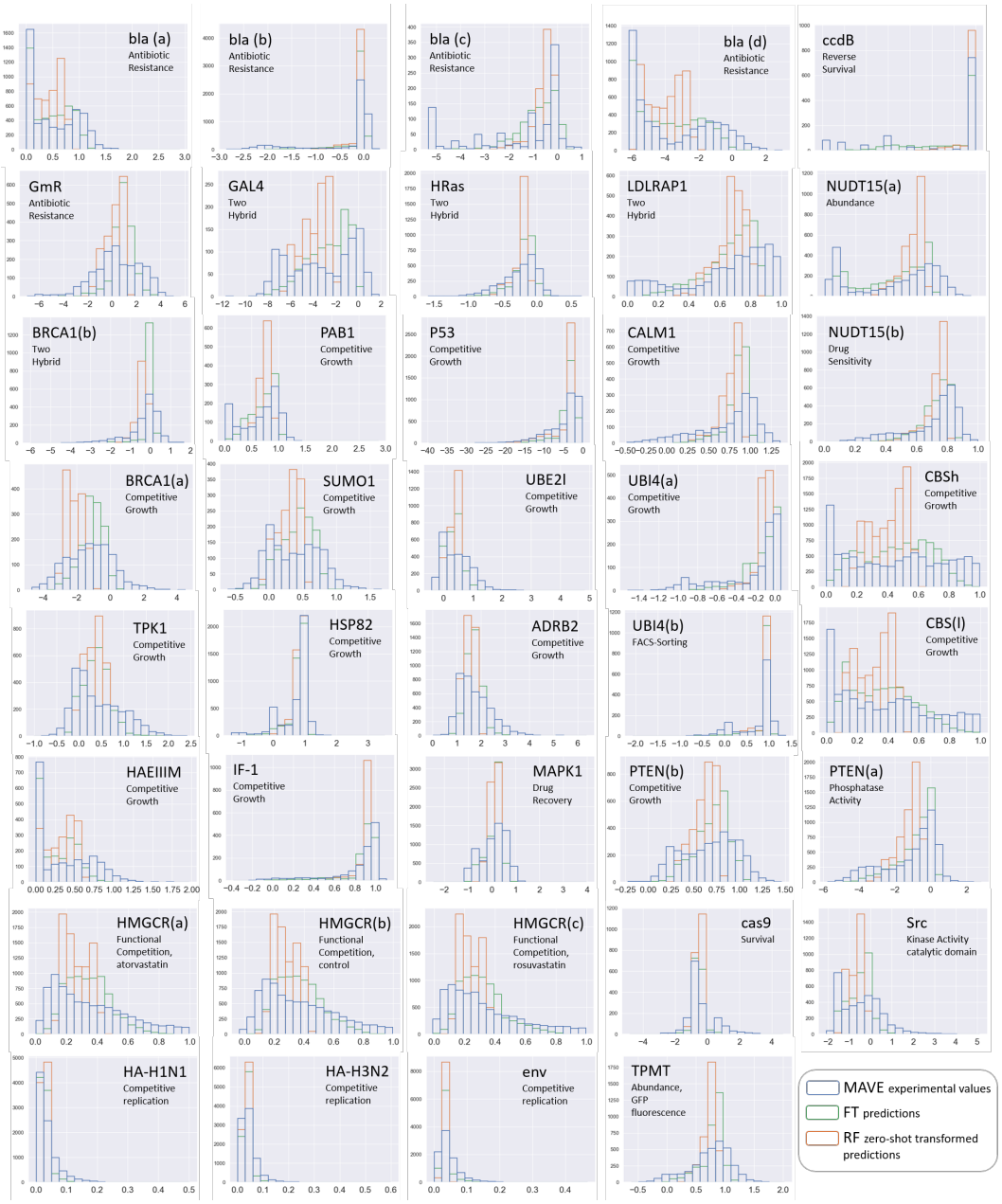

Figure 3: Histograms of Predicted and Experimental Values: MAVE experimental values are shown in blue. FT predictions (in green) are obtained using NN fine-tuned on 5% of the experimental data from the target protein. RF zero-shot transformed values (in red) correspond to zero shot predictions produces by Random Forest (Høie et al., 2022) transformed to the experimental distribution using the parameters estimated using 5% of the experimental data from the target protein.

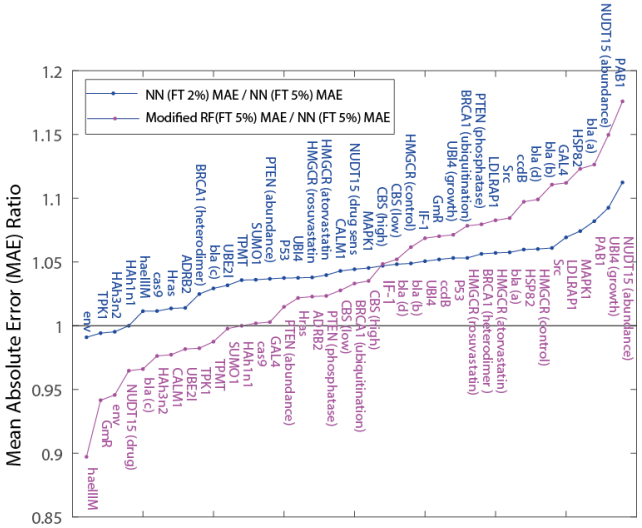

Figure 4: The improvement in MAE comparing the NN vs. the modified RF using 5% of the data sampled at random to calculate the experimental values. We calculate the MAE for predictions of the experimental values for the NN FT model, and the RF predictions transformed into experimental values (magenta). Better models have lower MAE, and thus cases where more data improved the predictions are above 1. For reference, we show in blue again (the same data as above) the improvement in MAE with the addition of data.

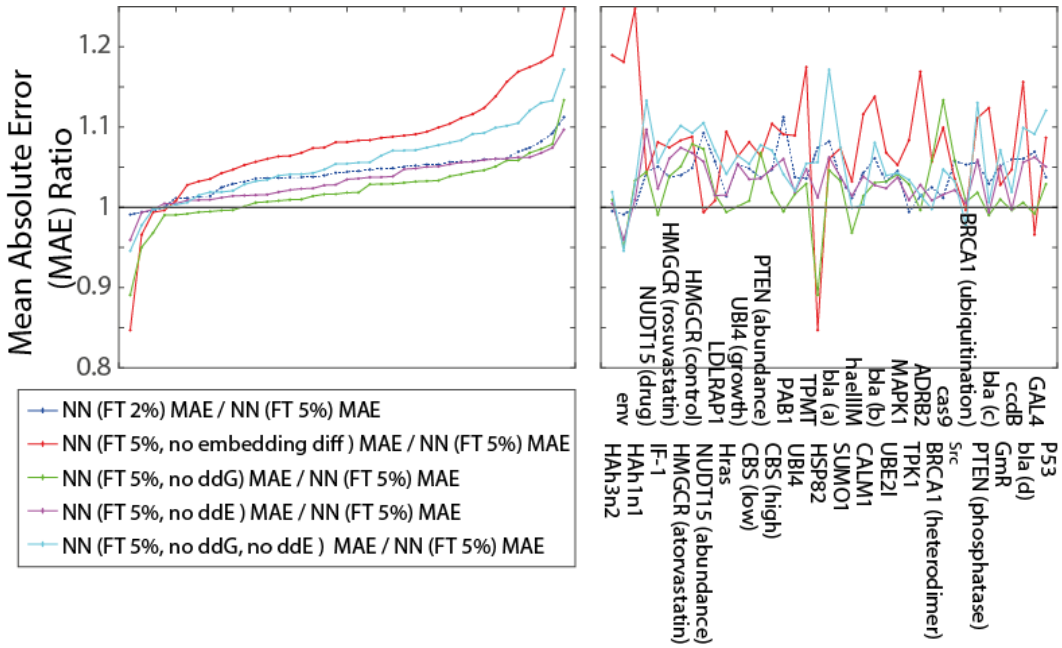

Figure 5: Ablation Study comparing the improvement in MAE of the FT model using 5% data to models that are missing part of the input: the embedding diff (red), ddG (green), ddE (magenta), both ddG and ddE (cyan). For reference, we show the improvement to the MAE when using 5% of the data rather than 2% (in dashed blue, this is the same curve as in 2 above). In the left panel, we sort in each case the MAE ratios, so that the presentation will be clearer. However, the order of the experiments is different depending on the case of the reduced inputs. Thus, on the right, we show the MAE ratios in the same order, with the names of the proteins listed along the x-axis

		MAE					
		Our Network			Modified RF		
Name	Organism (Assay)	2%	3%	5%	2%	3%	5%
bla (a) (P62593)	E.coli (Antibiotics resist.)	0.2691	0.2646	0.2487	0.2612	0.2657	0.2697
bla (c) (P62593)	E.coli (Antibiotics resist.)	1.3524	1.3119	1.3139	1.2527	1.2181	1.2693
bla (d) (P62593)	E.coli (Antibiotics resist.)	1.6604	1.6204	1.5661	1.6597	1.645	1.6477
GAL4 (P04386)	Yeast (Two-hybrid)	2.1349	2.0919	1.9965	1.9768	2.0194	2.0023
PAB1 (P04147)	Yeast (Comp. growth)	0.2543	0.243	0.2286	0.2637	0.2623	0.2575
NUDT15 (b) (Q9NV35)	Human (Drug senst., FACS)	0.1154	0.1148	0.1105	0.1077	0.1103	0.1066
TPMT (P51580)	Human (Abundance, GFP)	0.24	0.2341	0.2317	0.2354	0.2309	0.2288
PTEN (b) (P60484)	Human (Comp. growth)	0.9836	0.9614	0.9311	0.9814	0.9629	0.9529
P53 (P04637)	Human (Comp. growth, P53)	2.6436	2.5878	2.5481	2.7238	2.6633	2.7299
GmR (Q53396 template)	E.coli (Antibiotics resist.)	1.478	1.4286	1.4049	1.3464	1.3409	1.3228
HMGCR (a) (P04035)	Human (atorvastatin)	0.1596	0.1573	0.1535	0.1675	0.1661	0.1662
infA (P69222)	E.coli (Comp. growth)	0.1452	0.1466	0.1382	0.1501	0.1443	0.1449
CBS (a) (P35520)	Human (Comp. growth)	0.2292	0.2265	0.2189	0.2293	0.2289	0.2266
SUMO1 (P63165)	Human (Compt. growth, yeast)	0.2726	0.2682	0.2631	0.2794	0.2654	0.2625
PTEN (a) (P60484)	Human (Abundance, GFP)	0.2168	0.2169	0.2091	0.2182	0.2143	0.2122
HMGCR (c) (P04035)	Human (rosuvastatin)	0.1588	0.1563	0.153	0.1667	0.1651	0.165
UBE2I (P63279)	Human (Comp. growth, yeast)	0.3446	0.3388	0.334	0.3406	0.3357	0.3279
HMGCR (b) (P04035)	Human (Func. complimentation)	0.1609	0.1573	0.1534	0.1694	0.1682	0.1686
bla (b) (P62593)	E.coli (Antibiotics resist.)	0.3322	0.322	0.3131	0.3347	0.3342	0.3324
CBS (b) (P35520)	Human (Comp. growth)	0.2193	0.2142	0.2092	0.2171	0.2175	0.215
ADRB2 (P07550)	Human (Comp. growth)	0.4696	0.4664	0.4631	0.4771	0.4783	0.4737
haelllM (P20589)	H. Aegyptius (Comp. growth)	0.2921	0.2904	0.2888	0.2633	0.2599	0.2591
Src (b) (P12931)	S. Cerevisae (Kinase activity)	0.6429	0.6326	0.6079	0.6843	0.6812	0.6752
UBI4 (a) (P0CG63)	Yeast (Comp. growth)	0.204	0.2038	0.1937	0.223	0.2215	0.2227
MAPK1 (P28482)	Human (Comp. growth, doxycycline)	0.3575	0.3555	0.342	0.3871	0.3886	0.3841
HSP82 (P02829)	Yeast (Comp. growth)	0.285	0.2791	0.2653	0.2929	0.2915	0.2911
ccdB (P62554)	E.coli (Reverse survival)	1.5578	1.5356	1.4699	1.6221	1.707	1.5731
UBI4 (b) (P0CG63)	Yeast (FACS-sorting)	0.2402	0.2393	0.2315	0.2476	0.2617	0.2474
HA-H1N1 (A0A2Z5-U3Z0)	Influenza virus (Comp. growth)	0.021	0.021	0.021	0.0209	0.0209	0.021
NUDT15 (a) (Q9NV35)	Human (Abundance, GFP)	0.1924	0.1868	0.1761	0.2119	0.2072	0.2071
LDLRAP1 (Q5SW96)	Human (2-hybrid, OBFC1)	0.1943	0.1885	0.1838	0.2011	0.2032	0.2044
Hras (P01112)	Human (Two-hypbrid)	0.1861	0.1866	0.1836	0.1865	0.1837	0.1876
BRCA1 (b) (P38398)	Human (yeast 2-hybrid, phage)	0.5602	0.5649	0.5466	0.6011	0.5884	0.5901
env (P03377)	HIV virus (Comp. replication)	0.0219	0.022	0.0221	0.021	0.0209	0.0209
cas9 (Q99ZW2)	S. Pyrogenes (survival)	0.5868	0.5844	0.5801	0.611	0.5793	0.5812
HA-H3N2 (A0A097-PF60)	Influenza virus (Comp. growth)	0.021	0.021	0.0211	0.0206	0.0206	0.0206
TPK1 (Q9H3S3)	Human (Comp. growth, yeast)	0.4289	0.4323	0.4314	0.424	0.4219	0.4238
BRCA1 (a) (P38398)	Human (Comp. growth, HAP1)	1.0452	1.027	0.9924	1.0358	1.0526	1.0253
CALM1 (P0DP23)	Human (Comp. growth, yeast)	0.2984	0.2905	0.2861	0.2851	0.2837	0.2796

Table 1: MAE values with respect to true values for all proteins in our dataset when fine-tuning with 2%, 3%, and 5% our networks, and when considering the modified RF