Language models enable zero-shot prediction of the effects of mutations on protein function

Anonymous Author(s)
Affiliation
Address
email

Abstract

Modeling the effect of sequence variation on function is a fundamental problem for understanding and designing proteins. Since evolution encodes information about function into patterns in protein sequences, unsupervised models of variant effects can be learned from sequence data. The approach to date has been to fit a model to a family of related sequences. The conventional setting is limited, since a new model must be trained for each prediction task. We show that using only zero-shot inference, without any supervision from experimental data or additional training, protein language models capture the functional effects of sequence variation, performing at state-of-the-art.

1 Introduction

Proteins have a myriad of diverse functions that underlie the complexity of life. Protein sequences encode function via structure through the spontaneous folding of the sequence into the three dimensional structure of the protein [12]. The effects of sequence mutations on function form a landscape that reveals how function constrains sequence. Alterations at some sites in a protein sequence cannot be tolerated because they are essential to the protein's function. Other sites evolve together because the structure and function is determined by them collectively. Mutations can enhance the activity of a protein, attenuate it or leave it unchanged.

The functional effect of sequence variations can be measured through deep mutational scanning experiments [15]. Consisting of thousands to hundreds of thousands of measurements of protein function, mutational scans give insight into the intrinsic constraints on a protein's structure and function. Due to cost and difficulty of implementing such experiments, compilations of deep mutational scanning data include experiments on a few dozens of proteins at most, relative to the tens of thousands of proteins encoded in the human genome, and the millions more across the tree of life that we would like to understand.

A model that learns the landscape linking sequence to function, and can make accurate predictions of the effect of mutations for new proteins, can provide insight into function without having to do experiments. Unsupervised models of mutational effects can be learned from sequences [13, 20]. Statistical patterns in a family of evolutionarily related protein sequences contain information about structure and function [58, 4, 5]. This is because the properties of a protein act as constraints on the selection of sequences through evolution [16].

Recently in the natural language modeling community, there has been interest in zero-shot transfer of models to new tasks. Massive language models can solve tasks they haven't been directly trained on [10, 37, 38]. Recently protein language models have achieved state-of-the-art in various structure prediction tasks [45, 40, 11]. Work to date has mainly focused on transfer in the classical representation learning setting, using the model features with supervision on a task by task basis.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

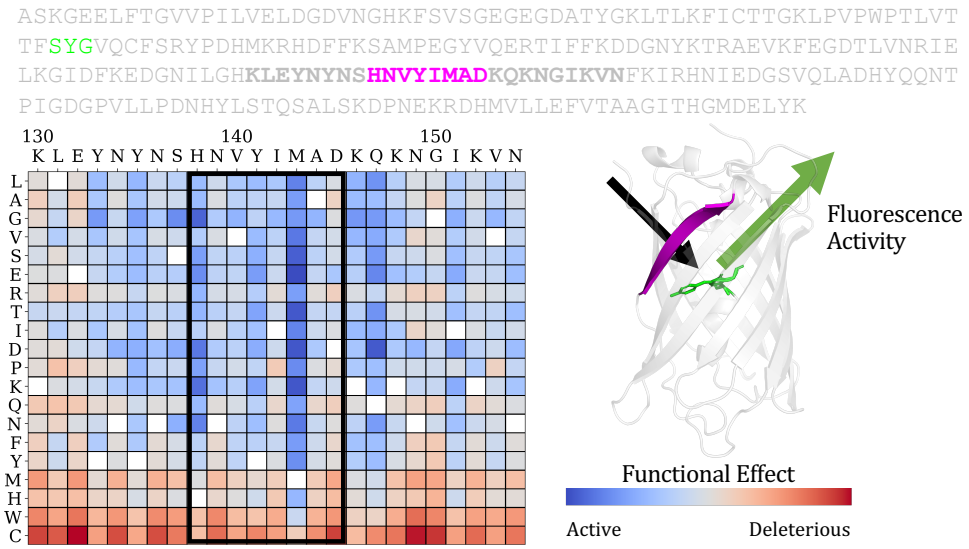

Figure 1: An example mutational effect prediction task. The objective is to score the effect of sequence mutations on the function of a protein. Shown here is green fluorescent protein (GFP, pdbid: 1GFL [55]). Sequence determines structure which is the molecular basis for the protein's function. Deep mutational scanning experiments provide groundtruth experimental measurements of the protein's function for a large set of single mutations or combinations of mutations. For each protein, the prediction task is to score each possible mutation and rank its relative activity. Predictions for single substitutions can be described in a score matrix. The columns are the positions in the sequence. The rows are the possible mutations at each position. In the case of multiple mutations, zero-shot methods use an additive model for scoring.

We are interested in understanding whether language models trained on large and diverse protein sequence databases transfer without further supervision to predicting measurements of protein function. Prior work has focused on transferring the representations using supervision from experimental data [21, 9]. We find that language models can transfer to predict functional measurements without supervision. Language models perform *zero-shot* and *few-shot* prediction of mutational effects across a variety of widely differing functions. The performance is comparable to state-of-the-art methods. Performance can be further improved by fine-tuning the model with sequences from the protein family. Predictions capture the functional landscape of the protein, correlating with amino acid conservation patterns in the core and surface, and identifying residues responsible for binding and activity.

2 Zero-shot transfer

Zero-shot learning has classically described the extension of a classifier to a new set of classes that haven't been seen before in training [25]. In natural language processing this idea has been extended to describe the transfer of models to entirely new tasks without further training. Proposed as zero-data learning by Larochelle et al. [27], this perspective on transfer has been at the center of recent work understanding the generalization capabilities of large language models [37, 10, 1, 38]. The distinction from representation learning is that the models are used *directly* without additional supervision for the task. This means that the tasks must be learned purely from pre-training.

In biology large language models trained on protein sequences have been investigated in the setting of representation learning [45, 11]. Two recent studies found a direct correspondence between the internal representations of the models and the physical structure of the protein, even though there has been no training on structures [54, 41]. Rao et al. [41] used this finding to extract an unsupervised contact predictor from the attention of the language model without making a formal connection to

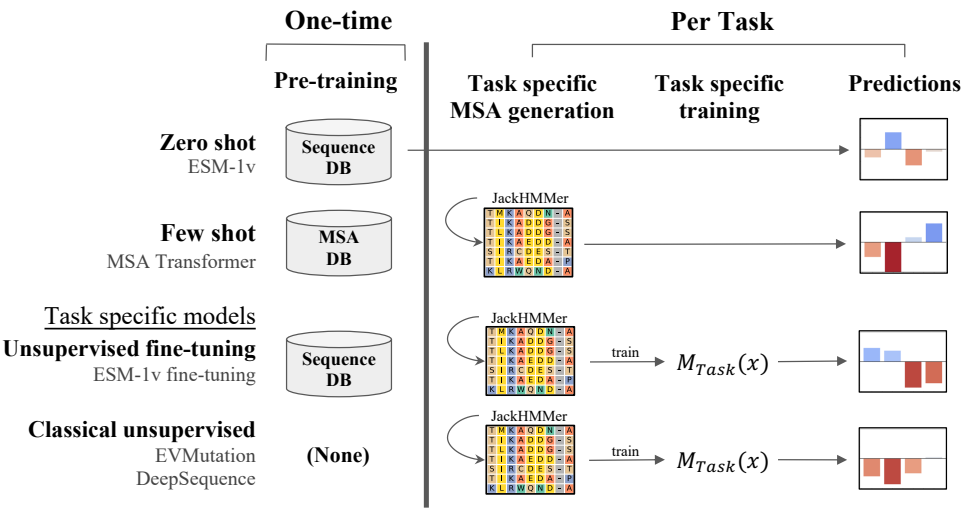

Figure 2: We highlight the pipeline steps involved in various function prediction methods. Compared with EVMutation and DeepSequence, MSA Transformer and ESM-1v require no additional model training. Moreover, ESM-1v requires no MSA generation.

zero-shot transfer. Similar to recent findings in NLP, contact prediction represents a capability that models trained on protein sequences learn without explicit supervision.

Zero-shot transfer represents a fundamentally new unsupervised learning capability that protein language models can bring to the computational biology toolkit. The standard practice for unsupervised inference from protein sequences is to train an independent model for each instance of the problem. In this view every protein is an independent prediction task. If protein language models can learn the biological properties needed to solve the task from pre-training, then they can be applied *directly* to new instances of the task, without specialization. This would mean that in practice a single general purpose model can be trained once and then applied directly to a variety of possible tasks.

Measurements of function, a property of central importance to the understanding and design of proteins, provides an interesting and practical ground for studying the generalization capability of protein language models. Deep mutational scanning experiments provide a useful set of tasks, as they measure the effect of thousands to hundreds of thousands of mutations on a single protein, and have been performed on a variety of proteins having different functions, and using various forms of experimental measurement. The data from these experiments make it possible to assess the zero-shot transfer of protein language models to scoring the effect of mutations on protein function.

Supervised methods, trained with data from experimental measurements [21, 9], and unsupervised methods trained only on sequences [13, 20], have both been developed for prediction of mutational effect. Unsupervised mutational effect predictors are trained as task specific models on sequences from an individual protein family. Hsu et al. [21] observes that even though they are unsupervised, such models have weak supervision on the task through the MSA, which describes the fitness landscape of the protein through positive examples. The zero-shot setting constitutes a further step in the difficulty of generalization: not only is the prediction unsupervised, but it is also performed without any specialization to the task. In the zero-shot setting the model must learn the functional landscape in pre-training without further supervision from sequences or experimental measurements related to the task.

3 Method

The masked language modeling objective trains the model to predict the probability that an amino acid occurs at a position in a protein given the surrounding context. We use this capability to score sequence variations. For a given mutation we can consider the amino acid in the wildtype protein as a

reference state, comparing the probability assigned to the mutated amino acid with the probability assigned to the wildtype.

We score mutations using the log odds ratio at the mutated position, assuming an additive model when multiple mutations exist in the same sequence:

$$\log p(x_t = x_t^{mt} | x_{-t}) - \log p(x_t = x_t^{wt} | x_{-t}) \tag{1}$$

3.1 Inference efficiency

Inference with ESM-1v is more efficient than previous state-of-the-art methods, due to two important differences: (i) the effect of mutations can be inferred directly without training a task-specific model; (ii) fitness landscapes can be predicted with a single forward pass. Fig. 6 summarizes time requirements.

Zero-shot and few-shot transfer As ESM-1v and the MSA Transformer are respectively zero-shot and few-shot, there is no need to train an independent model for each protein family Fig. 2. While the pre-training cost for protein language models is significant, they relax the need to fit a new model for each protein family going forward, making inference for each new family fast.

Fast scoring with ESM-1v The best performing scoring scheme requires one forward pass for each position protein, as compared to existing methods requiring a single forward pass per mutation. Moreover, we also find that ESM-1v can compute entire fitness landscapes in a single forward pass with only a minor 1% drop in absolute performance Table 5.

4 Results

4.1 Experimental setup

Evaluation Datasets We evaluate models on a set of 41 deep mutational scans collected by [44], which comprise a variety of tasks assessing a diverse set of proteins. Across tasks, the experimental data differs widely in the functions tested and in the experimental measurements performed. Of the 41 datasets, 37 are single-mutation only, 1 is double-mutation only, and the rest contain a variable number of mutations per sequence between 1 and 28. The median number of mutations is 2979, and the average is 16822; the smallest dataset has 37 mutations, and the largest has 496137. We treat each deep mutational scanning dataset as an separate prediction task, scoring each of the variants in the dataset with the model. We evaluate performance by comparing the scores from the model with the experimental measurements and computing the Spearman rank correlation. From the deep mutational scans in [44], we exclude the tRNA as it is not a protein and the toxin-antitoxin complex as it comprises multiple proteins.

Of these 41 deep mutational scans, we randomly select 9 single-mutation and 1 double-mutation deep mutation scan as a validation set for hyperparameter tuning and model selection. After tuning the models on this set, we then evaluate the final set of models on the full set, breaking up results between full set and blind test set. Therefore, all ablations reported are on the validation set and only the final set of models are evaluated on the blind test set.

Baseline reproduction We compare to the published DeepSequence and EVMutation results using the individual predictions provided by the authors in the supplement of those works. However, the MSAs used for training are generated from the 2017-10 version of UniRef100, whereas the models we study are trained on sequences from UniRef90 2020-03. In the case of MSA Transformer, the model is pre-trained on the 2018-03 Uniref, but we use 2020-03 MSAs for inference. In order to provide a fair comparison, we regenerate MSAs against the 2020-03 Uniref according to the methodology in Hopf et al. [20] and retrain EVMutation and DeepSequence on these datasets using their open-source codebases. Overall, we did not find a significant change in the results after re-training DeepSequence on the 2020-03 MSAs.

In addition to a comparison with DeepSequence and EVMutation, we also compare to previous protein language models: ESM-1b [45], ProtBERT-BFD [11], UniRep [3], and TAPE [39].

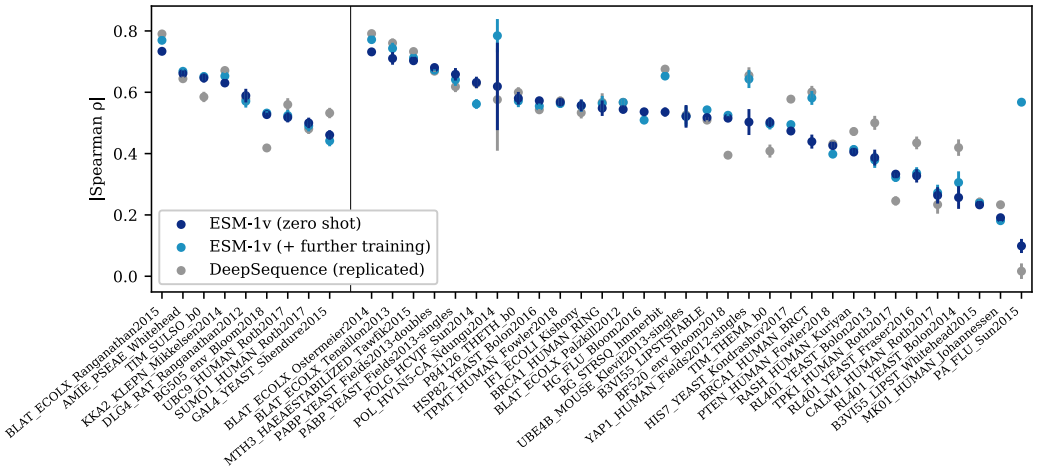

Figure 3: Pointplot of DeepSequence replication, ESM-1v, and finetuned ESM-1v on all 41 deep mutational scans. Points are mean $\pm$ std of 20 bootstrapped samples. Validation proteins are shown to the left of the dividing line and blind test proteins to the right. In half of the tasks, ESM-1v zero-shot has a higher Spearman rho than DeepSequence. The 41 bootstrapped means of the two methods on all data sets were not statistically distinguishable via a paired t-test.

4.2 Language models enable zero-shot prediction of the effects of mutations

Table 1 compares predictions from ESM-1v, the MSA Transformer, and other baselines to the set of 41 mutational scans described above. We find that ESM-1v, even its zero-shot mode, correlates better than DeepSequence with experimental measurements on 20 of the 41 datasets. The 41 bootstrapped means of the two methods are not statistically distinguishable via a paired *t*-test (Fig. 3).

We also find that ESM-1v significantly outperforms existing protein language models TAPE, UniRep, ProtBERT-BFD, and ESM-1b (Table 2 and Figure 7).

4.3 Model and Datasets

Pre-training datasets ESM-1v follows the ESM-1b architecture described in Rives et al. [45], but is pre-trained on a larger sequence dataset. To determine the best pre-training dataset, we pre-train models on Uniref clustered at various sequence identity thresholds (30%, 50%, 70%, 90%, 100%) and examine the change in average Spearman ρ on the validation set over the course of pre-training (Fig. 8, Table 4). The best performing model on the validation set uses Uniref90 and is selected as ESM-1v. We then train five models with different seeds to produce an ensemble.

Models	Full	Blind test
PSSM	0.460	0.461
EVMutation (published)	0.508	0.500
EVMutation (replication)	0.511	0.503
DeepSequence (published)	0.514	0.505
DeepSequence (replication)	0.526	0.510
MSA Transformer	0.542	0.530
ESM-1v (zero shot)	0.510	0.489
ESM-1v (finetuned)	0.537	0.522

Table 1: Average |Spearman ρ| on full and blind test sets of protein language models compared to existing methods for mutation effect prediction. DeepSequence and ESM-1v models are each the ensemble of 5 models. MSA Transformer is a single model, but is ensembled across 5 random samples of the MSA. PSSM and EVMutation are fit deterministically and are not ensembled.

Models	Full	Blind test
Unirep	0.155	0.154
TAPE	0.171	0.172
ProtBert	0.428	0.408
ESM-1b	0.458	0.432
ESM-1v†	0.487	0.465
ESM-1v*	0.510	0.489

Table 2: Average |Spearman ρ| on full and blind test sets of zero-shot protein language models. For ESM-1v we test both *ensemble of five ESM-1v models; † average performance of the five models. The remaining models are not ensembles, as those methods did not release an ensemble of models.

Scoring methods We experiment with a variety of scoring methods on the validation set, including wildtype marginals, masked marginals, mutant marginals, and psuedolikelihood. The masked marginal approach performs best. The motivation and mathematical derivation for these methods are described in Appendix A.

Parameter count We examine the scaling laws of our models in terms of parameter count. We train models using the same width, depth, and learning rate as described in [18], finding a meaningful improvement with scale (Fig. 9). Our findings suggest that continued scaling of language models will further improve our results.

4.4 MSA Transformer

We score mutations for MSA Transformer using the same log odds ratio and additive model as in ESM-1v. However, since MSA Transformer uses a set of sequences for inference, we replace the first sequence in the MSA with the masked wildtype sequence and compute the model's prediction of the probabilities at those positions. As with ESM-1v, this directly parallels the MSA Transformer training process as the MSA Transformer was also trained with masked language modeling.

Due to GPU memory limitations and compute constraints, it is not feasible to use the entire MSA during MSA Transformer inference. We find that the model is highly sensitive to the diversity of the input sequence (Fig. 10, Table 9). Therefore, we explore various inference time strategies to control the diversity of the input sequence, including the strategies described in Rao et al. [40] as well as a weighting scheme we introduce here. In the sequence reweighting scheme, we compute sequence weights [32] for the MSA and sample sequences without replacement according to those weights. We find that the sequence reweighting scheme performs best (Section 7). To further optimize performance, we experiment with the optimal number of sequences for inference, and find that 384 sequences performs best (Fig. 11. We use this strategy for all other experiments. Finally, although an ensemble of MSA Transformer models is not available, we take five subsamples of the MSA and ensemble the predictions.

4.5 Fine-tuning on MSAs

While ESM-1v has zero-shot prediction capability, we also examine how it performs in the conventional setting of training a new model for each deep mutation scanning prediction task. We observe that naively fine-tuning the model on the MSA results in overfitting and poor performance on the prediction tasks (Fig. 12). Therefore, we experiment with regularizing the model by freezing parameters, learning a projection onto a PSSM, or spiking pre-training sequences into the MSA (details in Appendix B). Spiking pre-training sequences into the MSA performs best. We use a batch size 50% of that used in pre-training and sample 0.1% of the sequences in each batch from the MSA; the remaining are drawn from the Uniref90 pre-training dataset.

5 Analysis of models

Protein structure and function The model's understanding of protein sequences corresponds to function. We use the entropy of the model's predictions for a position as a measure of its estimation

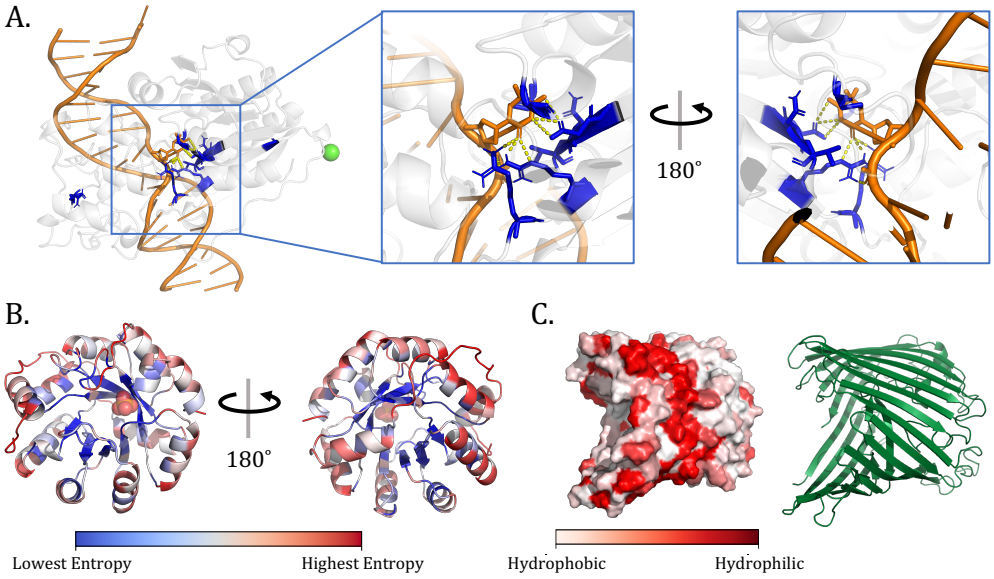

Figure 4: ESM-1v accurately captures functional properties. **(A)** DNA methylase HaeIII (pdbid: 1DCT [42]) with the top 10 positions with lowest entropy under ESM-1v predictions highlighted in blue. Each position in the DNA methylase HaeIII Uniprot sequence is masked and the entropy of the prediction is recorded. The Uniprot sequence is then aligned to the PDB sequence. We highlight in blue the top 10 aligned positions with lowest entropy on the structure. Fig. 18 contains additional examples. **(B)** TIM Barrel (pdbid: 1IGS [19]) with residues colored by entropy. The model's predictions for residues on the surface have highest entropy (red) while those on the core have lower entropy (blue), i.e. the model is more confident in the amino acid identity at positions in the core; and more flexible with those at the surface. Results are quantified across the trRosetta dataset in (Fig. 17). Notably, residues on the alpha helices show a clear gradient from high (red) to medium (white) to low (blue) entropy as residues transition from surface-facing to core-facing. **(C)** Sucrose-specific Porin (pdbid: 1A0T [14]), a transmembrane protein. The model predicts a hydrophobic band where the protein is embedded in the membrane.

of conservation at the site. The lowest entropy positions cluster at binding sites. Fig. 14 compares the distribution of the model's entropy between binding sites and non-binding sites. A significant difference is observed between the entropy assignment to binding and non-binding site residues. Fig. 4 visualizes the side chains of the 10 lowest entropy residues as predicted by the model on the crystal structure of DNA methyltransferase M.HaeIII interacting with its DNA substrate. In the crystal structure a cytosine of the substrate is inserted into the active site of the enzyme. The low entropy residues cluster in the active site and interact with the cytosine.

The model probabilities also correspond to the structure. Fig. 15 compares the entropy assigned to sites that are buried in the core of the protein vs. exposed. The model assigns significantly lower entropy to sites that are in the core of the protein, consistent with the idea that tight packing in the core places greater constraints on the selection of residues. Fig. 4B visualizes the entropy assigned by the model to each position overlayed in the structure of Indole-3-glycerolphosphate Synthase a TIM barrel protein. Higher entropy is assigned to residues having outward facing side chains on the alpha helices, while lower entropy is assigned to the inward facing positions. Fig. 17 compares the probability assigned to hydrophobic, polar, and charged amino acids for buried sites vs non-buried sites. The model prefers hydrophobic residues in the core and hydrophilic residues on the surface. The model probabilities closely match the empirical probabilities and those from the PSSM. Fig. 4C visualizes probability assigned to hydrophobic amino acids on the structure of Sucrose-specific Porin a transmembrane protein. The model predicts a hydrophobic band in the center where the protein embeds in the membrane.

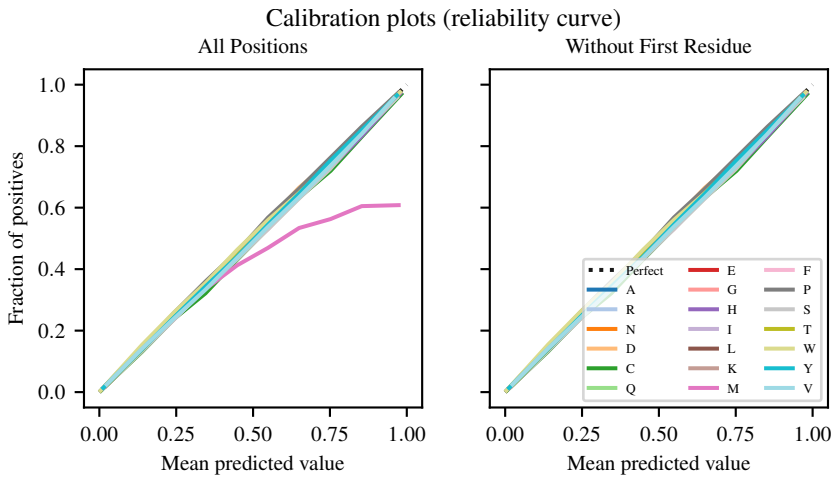

Figure 5: Calibration plot for ESM-1v predictions on each of the 20 naturally occurring amino acids on the trRosetta dataset. The multi-class classification was converted into a set of 20 one-versus-all classifications for the purposes of this analysis. Left and right plots show calibration of all positions and positions excluding the first residue, respectively. Since full sequences always start with Methionine, the model overwhelmingly predicts it in the first position. When evaluating the model on subsequences, such as those in the trRosetta dataset, this causes a miscalibration at the first residue. Including the first residue, the model as an average calibration error (ACE) of 0.011 in the first case and 0.006 in the second.

Calibration We evaluate model calibration using 15008 sequences with length < 1024 from the trRosetta [56] dataset. ESM-1v probabilities for each amino acid at each position were calculated with masked marginal probability (Appendix A). We find that the model is generally well calibrated for all amino acids except Methionine. ESM-1v always predicts Methionine as the first position in the sequence since full protein sequences always start with it, so care must be used when applying the model to a subsequence. The standard expected calibration error (ECE) performs poorly for highly imbalanced data [33]. Following Neumann et al. [33] and Nixon et al. [34] we adapt average calibration error for the multi-class setting as follows

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{B_k^+}\sum_{b=1}^{B_k^+}|\texttt{acc}(b,k)-\texttt{conf}(b,k)| \tag{2}$$

where K is the number of classes, B_k^+ is the number of non-empty bins for class k, and $\texttt{acc}$ and $\texttt{conf}$ are the accuracy and confidence for bin b and class k. When excepting the first residue, the model achieves an ACE of 0.006.

Additionally, we explore the relationship between conservation (entropy of the PSSM) and the model's predicted entropy. We find that these are well correlated (Pearson $r = 0.44$), suggesting the model is able to identify conserved positions.

6 Related Work

6.1 Protein language models

In the past few years, a number of groups have developed language models for protein sequences [3, 7, 11, 29, 30, 39, 40, 45]. These models have been used for many tasks, including supervised low-N function prediction [9, 45], remote homology detection [39, 45], and protein generation [30]. The approach to the tasks typically involves transfer learning, where a pretrained language model is fine-tuned for a particular problem. Vig et al. [54] and Rao et al. [41] found that transformer attention

corresponds to known biological properties such as structure and binding sites and can be used to predict contacts.

6.2 Mutation effect prediction

Supervised and unsupervised methods have been developed for prediction of mutational effects. Supervised methods train models using experimental measurements or labels from databases of clinical variants. Standard machine learning tools including linear regression, random forests, and support-vector machines can be used [57]. Models have been designed specifically for proteins, using feature engineering such as Envision [17] and PolyPhen-2 [2], ensemble methods such as Revel, [22], MPC [47], CADD [24], and M-CAP [23], language models such as UniRep [3, 9] and ESM [45], and other representation learning approaches [51, 59].

Unsupervised mutation effect predictors work by inferring the likelihood of a mutant from the evolutionary landscape of the original protein. A density model fit to related sequences is used for scoring. SIFT [48] is a first order approach using a position-specific-scoring-matrix. EVMutation [20] extends this to a second-order approach by training a Potts model on the MSA. DeepSequence [44] includes higher-order interactions by training a VAE on the MSA instead, using the ELBO to score mutations. Riesselman et al. [43] proposes using an autoregressive model that doesn't require the sequences to be aligned.

Hsu et al. [21] show that unsupervised mutational effect predictors can be extended to perform supervised predictions, with better unsupervised predictors generally resulting in better supervised predictors. This suggests improving unsupervised prediction can drive progress in both settings.

7 Discussion

In biology advances in language modeling at scale are bringing the goal of a general purpose model for proteins closer to realization. This line of work aspires to a model that learns to read and write biology in its native language, that can be directly applied across a range of protein understanding and design tasks. For scalability, learning from sequences is important: while there are no central databases of high-throughput functional measurements, and few compilations exist, billions of sequences are available to learn from in sequence databases [35, 49]. Sequences give an unparalleled view into the vast diversity and complexity of molecular parts invented by nature through billions of years of evolution.

Unsupervised structure [26, 53, 6] and function [13, 20] learning methods first effectively realized the idea that biological properties could be read directly from sequences without supervision from experimental measurements. However these methods are not general purpose in the sense that a specialized model must be trained for every protein for which a prediction is to be made. We show that the same performance can be realized by a general purpose model that has been trained across many diverse protein families. Similar to observations on the learning of tertiary protein structure in large language models [45, 41], we find that increasing the scale of models leads to improvements in function learning. The understanding of mutational landscapes in the models correlates with the molecular basis of function in proteins, capturing binding sites, and amino acid preferences that are determined by the folded structure.

Zero-shot transfer is an interesting capability of large scale language models, and represents a major point of departure from the unsupervised learning methods that are the basis for current state-of-the-art inference of protein structure and function. The capability for zero-shot transfer implies that a model can be trained once and then applied to perform inference for many tasks. It is also a window into deeper questions about the forms of generalization that are possible in learning from sequences. Reading structural and functional design principles from sequences is a necessary capability for writing new biologically active sequences. Generalization in the zero-shot setting suggests the potential for large language models to capture knowledge that can be transferred to generating new functional proteins.

References

[1] Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, Chen M, and Sutskever I. Zero-shot text-to-image generation.

[2] Ivan A. Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E. Ramensky, Anna Gerasimova, Peer Bork, Alexey S. Kondrashov, and Shamil R. Sunyaev. A method and server for predicting damaging missense mutations, apr 2010. ISSN 15487091.

[3] Ethan C. Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M. Church. Unified rational protein engineering with sequence-only deep representation learning. *Nature Methods*, 12:1315–1322, 3 2019. ISSN 15487105. doi: 10.1101/589333. URL https://www.biorxiv.org/content/10.1101/589333v1.

[4] D. Altschuh, A.M. Lesk, A.C. Bloomer, and A. Klug. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *Journal of Molecular Biology*, 193(4):693–707, 1987. ISSN 0022-2836. doi: 10.1016/0022-2836(87)90352-4.

[5] D Altschuh, T Vernet, P Berti, D Moras, and K Nagai. Coordinated amino acid changes in homologous protein families. *Protein Engineering, Design and Selection*, 2(3):193–199, 1988.

[6] Sivaraman Balakrishnan, Hetunandan Kamisetty, Jaime G. Carbonell, Su-In Lee, and Christopher James Langmead. Learning generative models for protein fold families. *Proteins: Structure, Function, and Bioinformatics*, 79(4):1061–1078, 4 2011. ISSN 08873585. doi: 10.1002/prot.22934. URL http://doi.wiley.com/10.1002/prot.22934.

[7] Tristan Bepler and Bonnie Berger. Learning protein sequence embeddings using information from structure, 2 2019. URL http://arxiv.org/abs/1902.08661https://arxiv.org/abs/1902.08661.

[8] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[9] Surojit Biswas, Grigory Khimulya, Ethan C Alley, Kevin M Esvelt, and George M Church. Low-N protein engineering with data-efficient deep learning. *bioRxiv*, page 2020.01.23.917682, 2020. doi: 10.1101/2020.01.23.917682.

[10] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165.

[11] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing. *bioRxiv*, 7 2020. URL http://arxiv.org/abs/2007.06225.

[12] Charles J Epstein, Robert F Goldberger, and Christian B Anfinsen. The genetic control of tertiary protein structure: studies with model systems. In *Cold Spring Harbor symposia on quantitative biology*, volume 28, pages 439–449. Cold Spring Harbor Laboratory Press, 1963.

[13] Matteo Figliuzzi, Hervé Jacquier, Alexander Schug, Oliver Tenaillon, and Martin Weigt. Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase tem-1. *Molecular biology and evolution*, 33(1):268–280, 2016.

[14] Doris Forst, Wolfram Welte, Thomas Wacker, and Kay Diederichs. Structure of the sucrose-specific porin scry from salmonella typhimurium and its complex with sucrose. *Nature structural biology*, 5(1):37–46, 1998.

[15] Douglas M Fowler and Stanley Fields. Deep mutational scanning: a new style of protein science. *Nature methods*, 11(8):801, 2014.

[16] Ulrike Göbel, Chris Sander, Reinhard Schneider, and Alfonso Valencia. Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Genetics*, 18(4):309–317, 4 1994. ISSN 0887-3585. doi: 10.1002/prot.340180402. URL http://www.ncbi.nlm.nih.gov/pubmed/8208723http://doi.wiley.com/10.1002/prot.340180402.

[17] Vanessa E. Gray, Ronald J. Hause, Jens Luebeck, Jay Shendure, and Douglas M. Fowler. Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data. *Cell Systems*, 6(1):116–124.e3, jan 2018. ISSN 24054720. doi: 10.1016/j.cels.2017.11.003.

[18] Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. Scaling laws for autoregressive generative modeling. *CoRR*, abs/2010.14701, 2020. URL https://arxiv.org/abs/2010.14701.

[19] Michael Hennig, Beatrice Darimont, Reinhard Sterner, Kasper Kirschner, and Johan N Jansonius. 2.0 å structure of indole-3-glycerol phosphate synthase from the hyperthermophile sulfolobus solfataricus: possible determinants of protein stability. *Structure*, 3(12):1295–1306, 1995.

[20] Thomas A. Hopf, John B. Ingraham, Frank J. Poelwijk, Charlotta P.I. Schärfe, Michael Springer, Chris Sander, and Debora S. Marks. Mutation effects predicted from sequence co-variation. *Nature Biotechnology*, 35(2):128–135, feb 2017. ISSN 15461696. doi: 10.1038/nbt.3769.

[21] Chloe Hsu, Hunter Nisonoff, Clara Fannjiang, and Jennifer Listgarten. Combining evolutionary and assay-labelled data for protein fitness prediction. *bioRxiv*, page 2021.03.28.437402, mar 2021. doi: 10.1101/2021.03.28.437402.

[22] Nilah M Ioannidis, Joseph H Rothstein, Vikas Pejaver, Sumit Middha, Shannon K McDonnell, Saurabh Baheti, Anthony Musolf, Qing Li, Emily Holzinger, Danielle Karyadi, et al. Revel: an ensemble method for predicting the pathogenicity of rare missense variants. *The American Journal of Human Genetics*, 99(4):877–885, 2016.

[23] Karthik A Jagadeesh, Aaron M Wenger, Mark J Berger, Harendra Guturu, Peter D Stenson, David N Cooper, Jonathan A Bernstein, and Gill Bejerano. M-cap eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nature genetics*, 48(12): 1581, 2016.

[24] Martin Kircher, Daniela M Witten, Preti Jain, Brian J O'Roak, Gregory M Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*, 46(3):310–315, 2014.

[25] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958. IEEE, 2009.

[26] Alan S. Lapedes, Bertrand G. Giraud, LonChang Liu, and Gary D. Stormo. Correlated Mutations in Models of Protein Sequences: Phylogenetic and Structural Effects. *Lecture Notes-Monograph Series*, 33:236–256, 1999. doi: 10.2307/4356049. URL http://www.jstor.org/stable/4356049.

[27] Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. Zero-data learning of new tasks. In *AAAI*, volume 1, page 3, 2008.

[28] Andrew Leaver-Fay, Matthew J O'meara, Mike Tyka, Ron Jacak, Yifan Song, Elizabeth H Kellogg, James Thompson, Ian W Davis, Roland A Pache, Sergey Lyskov, et al. Scientific benchmarks for guiding macromolecular energy function improvement. *Methods in enzymology*, 523:109–143, 2013.

[29] Amy X Lu, Haoran Zhang, Marzyeh Ghassemi, and Alan Moses. Self-Supervised Contrastive Learning of Protein Representations By Mutual Information Maximization. *bioRxiv*, page 2020.09.04.283929, 9 2020. doi: 10.1101/2020.09.04.283929. URL https://doi.org/10.1101/2020.09.04.283929.

[30] Ali Madani, Bryan McCann, Nikhil Naik, Nitish Shirish Keskar, Namrata Anand, Raphael R. Eguchi, Po-Ssu Huang, and Richard Socher. ProGen: Language Modeling for Protein Generation. *bioRxiv*, 3 2020. URL http://arxiv.org/abs/2004.03497.

[31] Milot Mirdita, Lars Von Den Driesch, Clovis Galiez, Maria J. Martin, Johannes Soding, and Martin Steinegger. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Research*, 45(D1):D170–D176, 1 2017. ISSN 13624962. doi: 10.1093/nar/gkw1081.

[32] Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S. Marks, Chris Sander, Riccardo Zecchina, José N. Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America*, 108(49):E1293–E1301, 12 2011. ISSN 00278424. doi: 10.1073/pnas.1111471108.

[33] Lukas Neumann, Andrew Zisserman, and Andrea Vedaldi. Relaxed Softmax: Efficient Confidence Auto-Calibration for Safe Pedestrian Detection. Technical report, oct 2018.

[34] Jeremy Nixon, Mike Dusenberry, Ghassen Jerfel, Timothy Nguyen, Jeremiah Liu, Linchuan Zhang, and Dustin Tran. Measuring Calibration in Deep Learning. apr 2019. URL http://arxiv.org/abs/1904.01685.

[35] Henrik Nordberg, Michael Cantor, Serge Dusheyko, Susan Hua, Alexander Poliakov, Igor Shabalov, Tatyana Smirnova, Igor V Grigoriev, and Inna Dubchak. The genome portal of the department of energy joint genome institute: 2014 updates. *Nucleic acids research*, 42(D1): D26–D31, 2014.

[36] Didier Nurizzo, Steven C Shewry, Michael H Perlin, Scott A Brown, Jaydev N Dholakia, Roy L Fuchs, Taru Deva, Edward N Baker, and Clyde A Smith. The crystal structure of aminoglycoside-3-phosphotransferase-iia, an enzyme responsible for antibiotic resistance. *Journal of molecular biology*, 327(2):491–506, 2003.

[37] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[39] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S. Song. Evaluating Protein Transfer Learning with TAPE. In *Neural Information Processing Systems*. Cold Spring Harbor Laboratory, 6 2019. doi: 10.1101/676825. URL https://doi.org/10.1101/676825http://arxiv.org/abs/1906.08230.

[40] Roshan Rao, Jason Liu, Robert Verkuil, Joshua Meier, John F. Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. MSA Transformer. *bioRxiv*, page 2021.02.12.430858, feb 2021. doi: 10.1101/2021.02.12.430858.

[41] Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer protein language models are unsupervised structure learners. *ICLR*, page 2020.12.15.422761, 12 2021. doi: 10.1101/2020.12.15.422761.

[42] KM Reinish, L Chen, GL Verdine, and WN Lipscomb. The crystal structure of haeiii methyltransferase covalently complexed to dna: an extrahelical cytosine and rearranged base pairing. *Cell*, 82(1):143–153, 1995.

[43] Adam Riesselman, Jung-Eun Shin, Aaron Kollasch, Conor McMahon, Elana Simon, Chris Sander, Aashish Manglik, Andrew Kruse, and Debora Marks. Accelerating Protein Design Using Autoregressive Generative Models. *bioRxiv*, page 757252, 2019. doi: 10.1101/757252. URL https://www.biorxiv.org/content/10.1101/757252v1.

[44] 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, 10 2018. ISSN 15487105. doi: 10.1038/s41592-018-0138-4.

[45] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences. *bioRxiv*, page 622803, 4 2020. doi: 10.1101/622803. URL https://www.biorxiv.org/content/10.1101/622803v3.

[46] Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. Pseudolikelihood reranking with masked language models. *CoRR*, abs/1910.14659, 2019. URL http://arxiv.org/abs/1910.14659.

[47] Kaitlin E Samocha, Jack A Kosmicki, Konrad J Karczewski, Anne H O'Donnell-Luria, Emma Pierce-Hoffman, Daniel G MacArthur, Benjamin M Neale, and Mark J Daly. Regional missense constraint improves variant deleteriousness prediction. *BioRxiv*, page 148353, 2017.

[48] Ngak Leng Sim, Prateek Kumar, Jing Hu, Steven Henikoff, Georg Schneider, and Pauline C. Ng. SIFT web server: Predicting effects of amino acid substitutions on proteins. *Nucleic Acids Research*, 40(W1):W452, jul 2012. ISSN 03051048. doi: 10.1093/nar/gks539.

[49] M. Steinegger, M. Mirdita, and J. Söding. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat Methods*, 16(7):603–606, 07 2019.

[50] Martin Steinegger and Johannes Söding. Clustering huge protein sequence sets in linear time. *Nature communications*, 9(1):1–8, 2018.

[51] Laksshman Sundaram, Hong Gao, Samskruthi Reddy Padigepati, Jeremy F McRae, Yanjun Li, Jack A Kosmicki, Nondas Fritzilas, Jörg Hakenberg, Anindita Dutta, John Shon, et al. Predicting the clinical impact of human mutation with deep neural networks. *Nature genetics*, 50(8):1161–1170, 2018.

[52] Baris E. Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H. Wu. UniRef: Comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, 23(10):1282–1288, 5 2007. ISSN 13674803. doi: 10.1093/bioinformatics/btm098. URL http://www.uniprot.org.

[53] John Thomas, Naren Ramakrishnan, and Chris Bailey-Kellogg. Graphical models of residue coupling in protein families, 4 2008. ISSN 15455963. URL https://pubmed.ncbi.nlm.nih.gov/18451428/.

[54] Jesse Vig, Ali Madani, Lav R. Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. BERTology Meets Biology: Interpreting Attention in Protein Language Models. *bioRxiv*, page 2020.06.26.174417, 6 2020. doi: 10.1101/2020.06.26.174417. URL http://arxiv.org/abs/2006.15222.

[55] Fan Yang, Larry G Moss, and George N Phillips. The molecular structure of green fluorescent protein. *Nature biotechnology*, 14(10):1246–1251, 1996.

[56] Jianyi Yang, Ivan Anishchenko, Hahnbeom Park, Zhenling Peng, Sergey Ovchinnikov, David Baker, and John Harvard. Improved protein structure prediction using predicted inter-residue orientations. *bioRxiv*, page 846279, 2019. doi: 10.1101/846279. URL https://www.biorxiv.org/content/10.1101/846279v1.

[57] Kevin K. Yang, Zachary Wu, and Frances H. Arnold. Machine-learning-guided directed evolution for protein engineering, aug 2019. ISSN 15487105.

[58] C Yanofsky, V Horn, and D Thorpe. Protein Structure Relationships Revealed By Mutational Analysis. *Science (New York, N.Y.)*, 146(3651):1593–4, 12 1964. ISSN 0036-8075. URL http://www.ncbi.nlm.nih.gov/pubmed/14224506.

[59] Haicang Zhang, Michelle S Xu, Wendy K Chung, and Yufeng Shen. Predicting functional effect of missense variants using graph attention neural networks. *bioRxiv*, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We note the compute costs associated with our work in a section in Appendix F. Otherwise, basic methods in computational biology have a long time horizon to concrete social impact.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We note the compute costs associated with the research in the Discussion section.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We are releasing model weights and code for zero-shot prediction of protein fitness landscapes with ESM-1v and MSA Transformer.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We tune hyperparameters on a validation set. Through a series of experiments described in the figures, we explain how we reach our final hyperparameters.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Depending on the figure, error bars are with respect to random seeds, 20 bootstraps, or variance across proteins in the validation set.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? We bring attention to the total amount of compute used in Fig. 6, in the discussion, and in a separate compute section in the appendix.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] See Appendix E.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We are releasing model weights and code.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We use open-source artifacts only.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]