Understanding Protein Language Model Scaling on Mutation Effect Prediction

Chao Hou[1,*], Di Liu[2], Aziz Zafar[2], Yufeng Shen[1,2,3,*]

1 Department of Systems Biology, Columbia University Irving Medical Center, New York, NY 10032

2 Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY 10032

3 JP Sulzberger Columbia Genome Center, Columbia University, New York, NY 10032

* Corresponding author: ch3849@cumc.columbia.edu (C.H.), ys2411@cumc.columbia.edu (Y.S.)

Abstract

Protein language models (pLMs) can predict mutation effects by computing log-likelihood ratios between mutant and wild-type amino acids, but larger models do not always perform better. We found that the performance of ESM2 peaks when the predicted perplexity for a given protein falls within the range of 3–6. Models that yield excessively high or low perplexity tend to predict uniformly near-zero or large negative log-likelihood ratios for all mutations on the protein, limiting their ability to discriminate between deleterious and neutral mutations. Larger models often assign uniformly high probabilities across all positions, reducing specificity for functionally important residues. We also demonstrated how the evolutionary information implicitly captured by pLMs can be linked with the conservation patterns observed in homologous sequences. Our findings highlight the importance of perplexity in mutation effect prediction and suggest a direction for developing pLMs optimized for this application.

Main

Protein language models (pLMs), such as ESM[1], are typically trained to predict the probability of masked or next residues on large protein sequence datasets. These models effectively capture conserved motifs and co-evolutionary residue pairs from their training data[2, 3], and have been widely applied to a range of crucial biological questions[1, 4, 5]. One of the most important applications of pLMs is mutation effect prediction[6, 7]—a critical challenge in human genetics and protein engineering. pLMs predict the log-likelihood ratio (LLR; Figure 1A) between the wild-type and mutant amino acids conditioned on the sequence context, enabling accurate zero-shot prediction of mutation effects, and outperforming many methods specifically designed for this application[7].

In the field of language modeling, larger models generally achieve lower validation perplexity (defined as the exponential of the negative log-likelihood of wild-type amino acids; Figure 1A) and better performances on downstream tasks[1]. However, in the application of zero-shot mutation effect prediction, several studies[6, 8] showed that larger pLMs often underperform compared to medium-sized models such as ESM2-650M. In this study, we sought to understand this phenomenon by evaluating ESM2 models on ProteinGYM[6] and the mega-scale protein stability dataset[9] (see Methods for details). Across both datasets, larger ESM2 models consistently achieve lower perplexity (i.e., better prediction of the masked residue, one residue is masked at a time; Figure 1B-C). However, this does not always result in better prediction of mutation effects, as measured by Spearman correlation (Figure 1B–C). Notably, the medium-sized ESM2-650M model outperforms the larger ESM2-3B and ESM2-15B models on both datasets (Figure 1B–C).

Results in Figure 1B–C suggest there may exists an optimal perplexity range for ESM2 models to predict mutation effects. To test this hypothesis, we examined the relationship between prediction performance and perplexity for each protein across ESM2 models. We found that performance peaks when the model yield a perplexity of 3-6 for a given protein (Figure 1D, S1A). For proteins with larger perplexity values (to the left of the grey zone in Figure 1D, S1A), performance improves as perplexity decreases. Conversely, for proteins with lower perplexity values (to the right of the grey zone), performance declines as perplexity continues to decrease. This rise-then-fall trend is consistent with a previous study[10]. ESM2-predicted perplexity for a protein is correlated with the number of homologs

the protein has in the training set (Figure S2). Small models do not show a performance decline, as they yield relatively high perplexities across all proteins due to their limited learning capacity. The performance decline at low perplexity is mainly observed in larger models, particularly for proteins with sufficient homologs in the training set (Figure S3). Proteins with limited number of homologs may not reach the low perplexity range, even with the largest models (Figure S3).

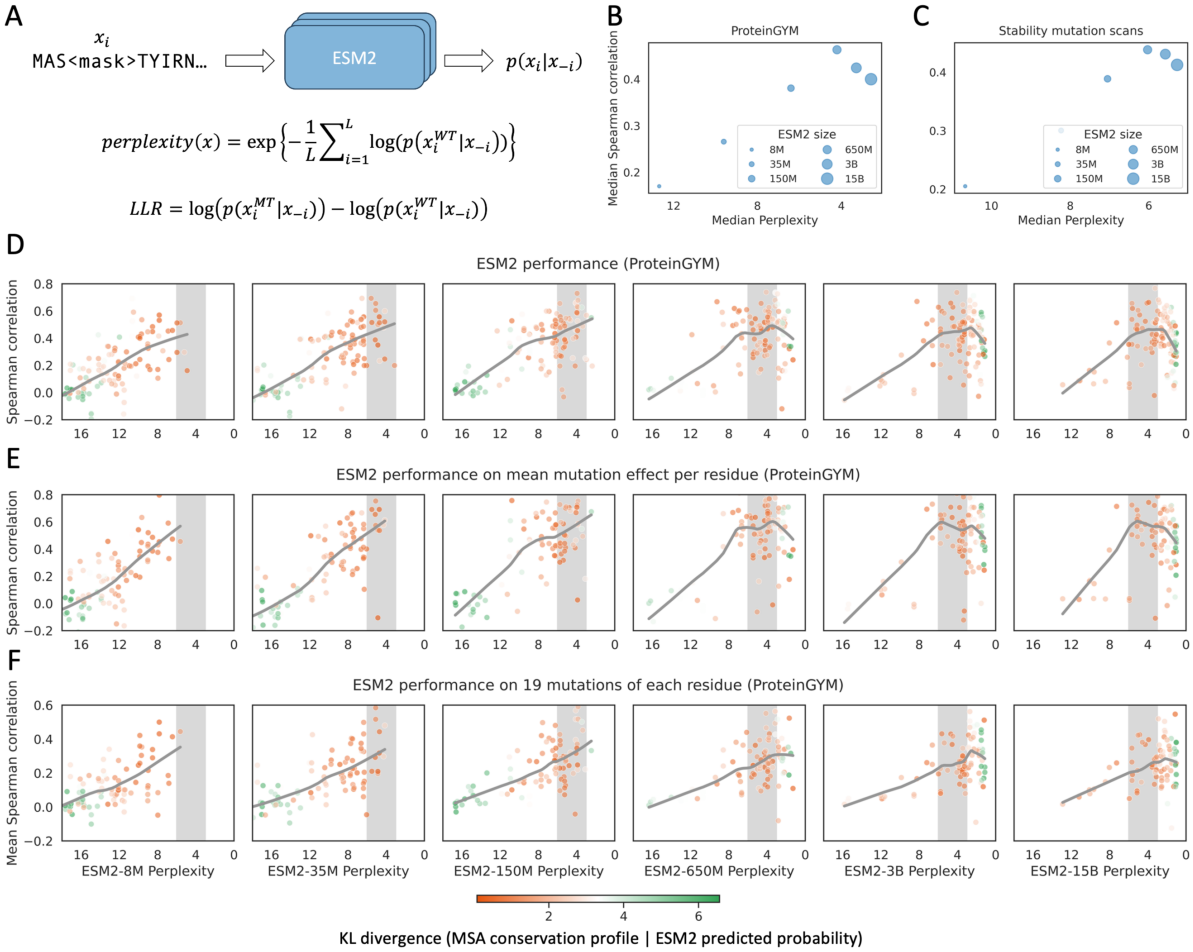

Figure 1. The relationship between ESM2 perplexity and zero-shot performance on mutation effects.

A, Overview of using ESM2 models to predict mutation effects, with equations for computing perplexity and log-likelihood ratio (LLR); x denotes the protein sequence and x_{-i} is the sequence with residue i masked. **B–C**, Median perplexity and zero-shot Spearman correlation across different ESM2 model sizes on two datasets, with marker size indicating model size. **D**, Spearman correlation between LLRs and experimentally measured mutation effects across all mutations in each protein. **E**, Spearman correlation between average LLRs and average mutation effects per residue, reflecting the model's understanding of sequence or structural context. **F**, Mean Spearman correlation between LLRs and effects of all 19 possible substitutions per residue, reflecting substitution specificity. **D-F**, Each point represents a protein, with color indicating the KL divergence between the empirical distribution observed in homologous sequences and the amino acid probability predicted by ESM2. Lower KL divergence (orange color) reflects higher similarity between ESM2 predictions and the conservation pattern in multi-sequence alignments (MSAs). The grey line indicates locally weighted regression, shaded gray region denotes the perplexity range of 3–6. Note that the x-axis is reversed, with higher perplexity on the left and lower perplexity on the right.

Prediction of mutation effects relies on the model's ability to understand both the sequence/structure context and the differences between wild-type and mutant amino acids. To dissect which of these aspects contributes to the rise-then-fall trend observed in Figure 1D and S1A, we compared the mean LLR and mean mutation effect per residue (reflecting context understanding), as well as the LLRs and mutation effects for the 19 substitutions at each residue (reflecting substitution specificity). We found that models yielding high perplexity on a given protein perform poorly on both aspects (Figure 1E–F, S1B–C). In contrast, models yielding low perplexity for a given protein primarily exhibit reduced performance in context understanding, with no significant change in their ability to distinguish amino

acid substitutions compared to those within the optimal perplexity range (Figure 1E–F, S1B–C). This pattern is further illustrated by a detailed comparison of per-residue performance within individual proteins (Figure S4). Larger models tend to assign uniformly high probabilities (i.e., lower perplexity) across all positions, leading to reduced specificity for functionally important residues (Figure S4).

To further investigate why ESM2 achieve best performance at the perplexity range, we visualized the LLR distributions for four proteins that show rise-then-fall trend in performance as ESM2 models size increase. For these proteins, the small model ESM2-8M yields relatively high perplexities of 10 or greater, resulting in LLR values close to zero for all mutations—including those with large experimentally measured effects (Figure 2). In contrast, the large model ESM2-15B yields very low perplexities below 2, predicting large negative LLRs for nearly all mutations—even those with minimal experimentally measured effects (Figure 2). In both cases, performance suffers due to a collapse in the dynamic range of LLR values. The medium-sized models, yield perplexities in the optimal range for these proteins and produce more distinguishable LLR predictions for deleterious versus neutral mutations, resulting in better agreement with experimental data (Figure 2). This observation aligns with intuition: a model with no predictive ability that assigns equal probability (i.e., 0.05) to all 20 amino acids (perplexity = 20) yields LLRs of zero for all mutations. Conversely, a powerful model that assigns a probability of 1 to the wild-type amino acid and 0 to all others (perplexity = 0) produces LLRs of negative infinity for all mutations. In both extremes, the model fails to differentiate between deleterious and neutral mutations.

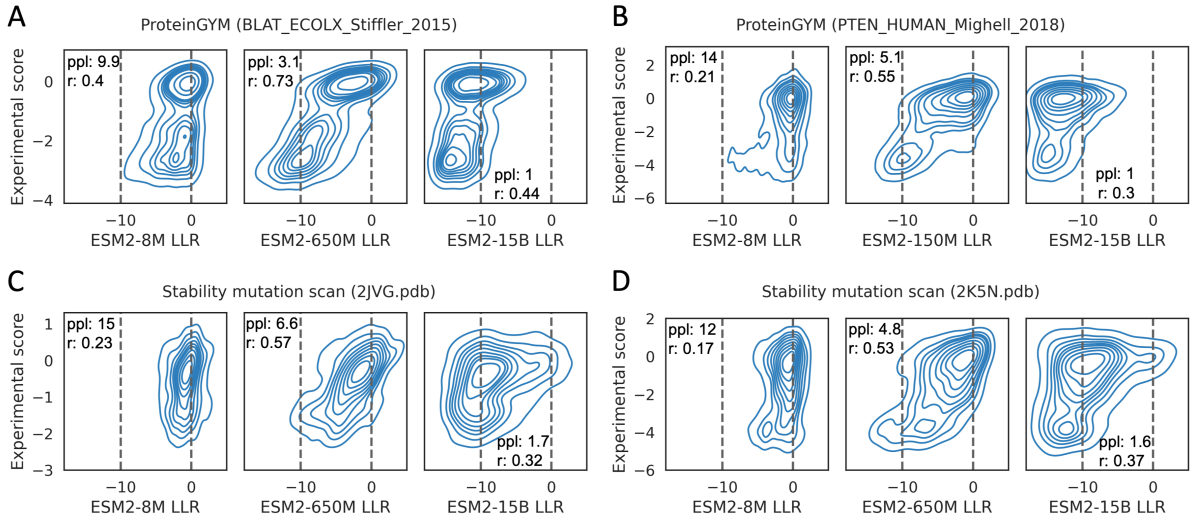

Figure 2. Relationship between ESM2-predicted LLRs and experimentally measured mutation effects for representative proteins.

2JVG.pdb (C3-binding domain 4 of *S. aureus* Sbi), 2K5N.pdb (N-terminal domain of ECA1580 from *E. carotovora*); BLAT_ECOLX_Stiffler_2015 (TEM-1 β-lactamase mutation effects on *E. coli* growth with ampicillin); PTEN_HUMAN_Mighell_2018 (PTEN mutation effects in a humanized yeast growth model). ppl: perplexity; r: Spearman correlation. Two dash lines indicate LLR of -10 and 0.

As pLMs implicitly learn evolutionary signals in the training data, proteins with more homologs in the training data tend to exhibit lower perplexity across all model sizes (Figure S2). The success of pLMs in mutation effect prediction largely stems from the strong correlation between evolutionary conservation and mutation effect: mutations frequently observed among homologous sequences tend to be neutral, while rare or unseen substitutions tend to be deleterious. In the field of homolog search, sequence identity thresholds of 20% and 30% are commonly used to define homologous relationships[11]. Interestingly, these thresholds correspond to probabilities of the wild-type amino acid of 0.2–0.3, which translate to perplexity of 3.3–5. This range closely matches the optimal perplexity range we observed (Figure 1D, S1A). Furthermore, we found that for proteins with predicted perplexities within the optimal range, the ESM2-predicted amino acid probabilities closely resemble the amino acid frequencies observed in homologous sequences (Figure 1D-F). In contrast, this similarity diminishes for proteins with either very high or very low perplexity (Figure 1D-F). Thus, our

results connect the implicit conservation captured by pLMs with explicit conservation in homologs defined by sequence identity.

Our results suggest that mutation effect prediction may not be an ideal benchmark for evaluating pLMs, which are trained to minimize perplexity. Our results also offer practical guidance for applying pLMs in mutation effect prediction. If a pLM predicts a given protein with very low perplexity, users should switch to smaller models that yield perplexities in the optimal range. Conversely, if the perplexity is too high, users can fine-tune the pLM using homologous sequences of the target protein[10] to reduce perplexity to the optimal range, or opt for other approaches like MSA-Transformer[12]. We propose a strategy for training pLMs specifically designed to predict mutation effects. During training, the pLM should first compute the perplexity of each protein in a batch. If a protein has very low perplexity, it should be excluded from training for that epoch. This approach prevents overfitting to proteins with abundant homologs and reallocates learning capacity to underrepresented proteins. By doing so, the pLM can achieve a balanced perplexity distribution centered around the optimal range across proteins, thereby improving performance on mutation effect prediction.

Our results underscore critical caveats for pLMs. First, pLMs are trained to predict wild-type amino acids, not to explicitly model conservation profiles. pLMs with a medium number of parameters have restricted learning capacity, and thus tend to minimize training loss by memorizing conservation patterns among homologs in the training set (Figure 1D-F), enabling them to perform well on mutation effect prediction. In contrast, larger pLMs can overfit by memorizing wild-type residues, assigning low probabilities to all alternative amino acids, and consequently losing the ability to distinguish between deleterious and neutral mutations. Secondly, pLMs are usually trained on natural proteins, the evolutionary information captured in the training set may not generalize to proteins outside the distribution. As shown in Figure S5, viral and designed proteins exhibit consistently lower performance than other proteins, even when their perplexities are comparable. These proteins are not optimized for the fitness of an independently living organism—designed proteins are often optimized for specific biophysical objectives, and viral proteins are optimized for their ability to infect and replicate within hosts. In such cases, models trained on biophysical properties—such as ESMDance or SeqDance[13]—may provide more suitable alternatives.

Author contributions

Y.S. and C.H. conceived the study. C.H. performed the evaluations and drafted the manuscript. D.L. and A.Z. assisted with the evaluations. Y.S. and C.H. interpreted the results. All authors revised the manuscript.

Competing interests

The authors declare no competing interests.

Acknowledgment

This work was supported by NIH grants R35GM149527 and Simons Foundation SFARI #1019623.

Methods

Dataset

For protein folding stability, the dataset “Tsuboyama2023_Dataset2_Dataset3_20230416.csv” was downloaded from https://zenodo.org/records/7992926. Only single-residue substitution mutations with available ddG_ML values were included in the analysis. For multiple measurements of the same mutation, the mean ddG_ML value was used. The final dataset comprised approximately 379,000 single-residue substitutions across 412 proteins, including both de novo designed and naturally occurring proteins.

ProteinGYM data were downloaded from the official website (https://proteingym.org/) and GitHub repository in June 2024. From the original set of 217 proteins, we applied the following filters: (1) excluded 64 proteins that overlapped with the folding stability dataset, (2) excluded 3 proteins with mutation regions longer than 1,024 residues, and (3) excluded 39 proteins with an average mutation scan depth (defined as the number of single-residue substitution mutations divided by the length of the mutation region) less than 10. The resulting dataset included approximately 583,000 single-residue substitutions across 114 proteins, encompassing mutation scans of activity, binding, and expression across diverse taxa.

Homologous sequences were identified using MMseqs2 with the following parameters: *-s 7 -a 1 --max-seqs 10000*. Homologs were defined as sequences with at least 20% identity and 50% coverage relative to the query protein. The UniRef50 database used for the search was downloaded in February 2025.

For the per-residue performance analysis, only residues with all 19 possible substitution mutations, and proteins containing at least 20 such residues were considered.

ESM2 Inference Procedure

Following the procedure used in ProteinGYM, each residue in the protein was sequentially masked, and ESM2 was used to predict the probabilities of 20 standard amino acids at the masked position. For a protein of length L, ESM2 was run for L times to compute the log-likelihoods of both the wild-type and mutated amino acids at each site. For inference with ESM2-15B, half-precision (float16) was used, as the model cannot be loaded onto a single 48 GB GPU in full precision. The ESM2 models were trained with a maximum input length of 1,024 residues. For proteins shorter than 1,024 residues, the full sequences were used. For longer proteins, segments of 1,024-residue centered on the mutation regions were used.

KL Divergence Between ESM2-Predicted and MSA-Observed Amino Acid Distributions

For proteins in the ProteinGYM benchmark, multiple sequence alignment (MSA) profiles were obtained from the ProteinGYM website. For each residue, we computed the Kullback–Leibler (KL) divergence between the amino acid distribution observed in the MSA and the distribution predicted by ESM2, defined as:

$$D_{\mathrm{KL}}\left(f^{\mathrm{MSA}} \parallel p^{\mathrm{ESM2}}\right) = \sum_{a \in \{20 \text{ amino acids}\}} f_a^{\mathrm{MSA}} \log\left(\frac{f_a^{\mathrm{MSA}}}{p_a^{\mathrm{ESM2}}}\right)$$

Here, f_a^{MSA} represents the empirical frequency of amino acid a in the MSA, and p_a^{ESM2} is the predicted probability of amino acid a by ESM2. KL divergence was calculated at each residue and then averaged across all residues within a protein to obtain a protein-level score (Figure S5).

Reference

1. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123-1130 (2023).
2. Zhang, Z. et al. Protein language models learn evolutionary statistics of interacting sequence motifs. *Proc Natl Acad Sci U S A* **121**, e2406285121 (2024).
3. Rao, R., Meier, J., Sercu, T., Ovchinnikov, S. & Rives, A. Transformer protein language models are unsupervised structure learners. *bioRxiv*, 2020.2012.2015.422761 (2020).
4. Chang, Y. et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology* **15**, 1-45 (2024).
5. Kulmanov, M. et al. Protein function prediction as approximate semantic entailment. *Nature Machine Intelligence* **6**, 220-228 (2024).

6. Notin, P. et al. ProteinGym: Large-Scale Benchmarks for Protein Design and Fitness Prediction. *bioRxiv* (2023).
7. Brandes, N., Goldman, G., Wang, C.H., Ye, C.J. & Ntranos, V. Genome-wide prediction of disease variant effects with a deep protein language model. *Nat Genet* **55**, 1512-1522 (2023).
8. Bhatnagar, A. et al. Scaling unlocks broader generation and deeper functional understanding of proteins. *bioRxiv*, 2025.2004.2015.649055 (2025).
9. Tsuboyama, K. et al. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature* **620**, 434-444 (2023).
10. Gordon, C., Lu, A.X. & Abbeel, P. Protein Language Model Fitness Is a Matter of Preference. *bioRxiv*, 2024.2010.2003.616542 (2024).
11. Pearson, W.R. An introduction to sequence similarity ("homology") searching. *Curr Protoc Bioinformatics* **Chapter 3**, 3.1.1-3.1.8 (2013).
12. Rao, R. et al. MSA Transformer. (2021).
13. Hou, C., Zhao, H. & Shen, Y. Learning Biophysical Dynamics with Protein Language Models. *bioRxiv*, 2024.2010.2011.617911 (2025).

Supplementary Information

Understanding Protein Language Model Scaling on Mutation Effect Prediction

Figure S1-5

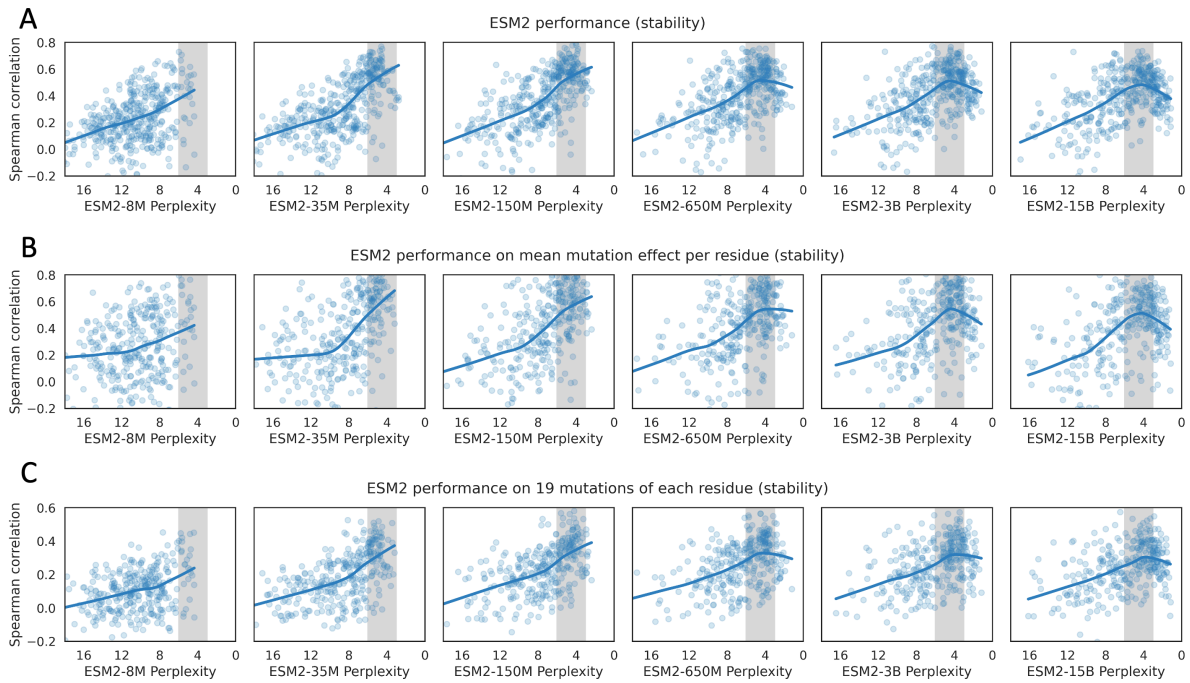

Figure S1. The relationship between ESM2 perplexity and performance on stability mutation scans.

A, Spearman correlation between predicted log-likelihood ratios (LLRs) and experimentally measured mutation effects across all mutations in each protein. **B**, Spearman correlation between average LLRs and average mutation effects per residue, reflecting the model's understanding of sequence or structural context. **C**, Mean Spearman correlation between LLRs and effects of all 19 possible substitutions per residue, reflecting substitution specificity. Each point represents a protein, blue line indicates locally weighted regression, shaded gray region denotes the perplexity range of 3–6. Note that the x-axis is reversed, with higher perplexity on the left and lower perplexity on the right.

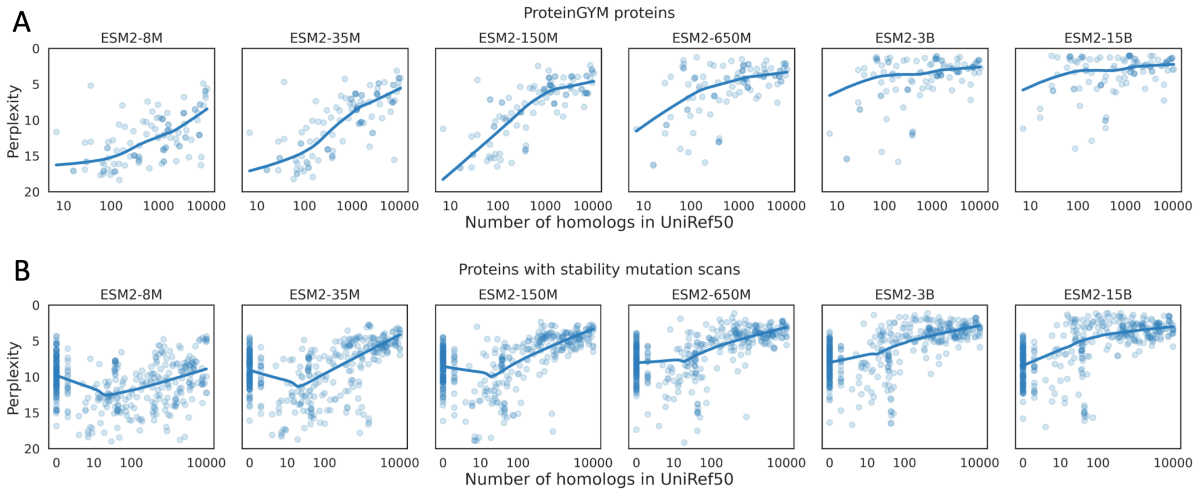

Figure S2. **The relationship between number of homologs in UniRef50 and perplexity.**
Each dot is a protein, blue line represents locally weighted regression. Homologs are defined as at least 20% sequence identity and 50% coverage. The x-axis shows the log-scaled number of homologs.

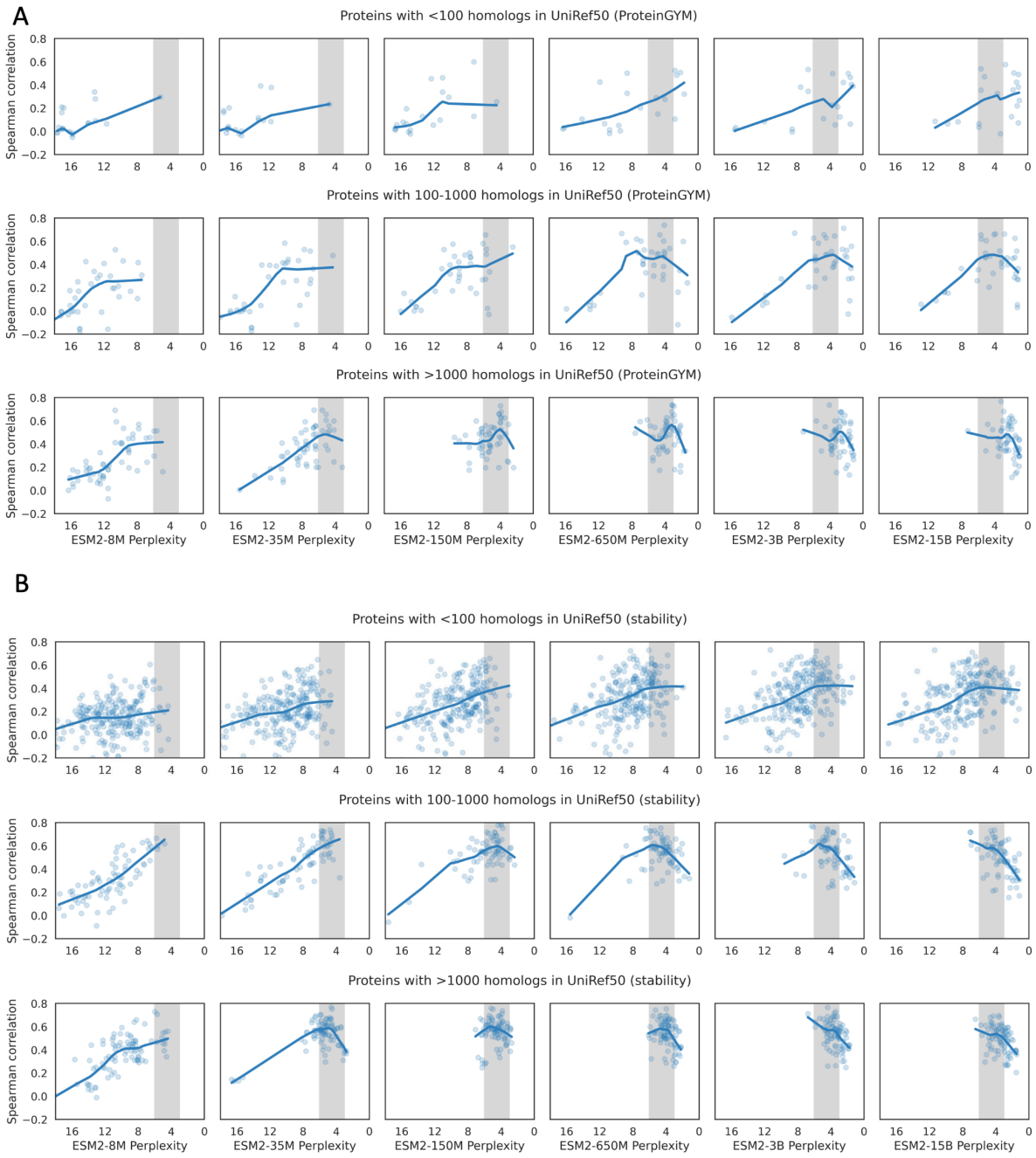

Figure S3. The relationship between ESM2 perplexity and performance on mutation effects.
Spearman correlation between LLRs and experimentally measured mutation effects across all mutations in each protein. Each point represents a protein, blue line indicates locally weighted regression, shaded gray region denotes the perplexity range of 3–6. Note that the x-axis is reversed, with higher perplexity on the left and lower perplexity on the right. Homologs are defined as at least 20% sequence identity and 50% coverage.

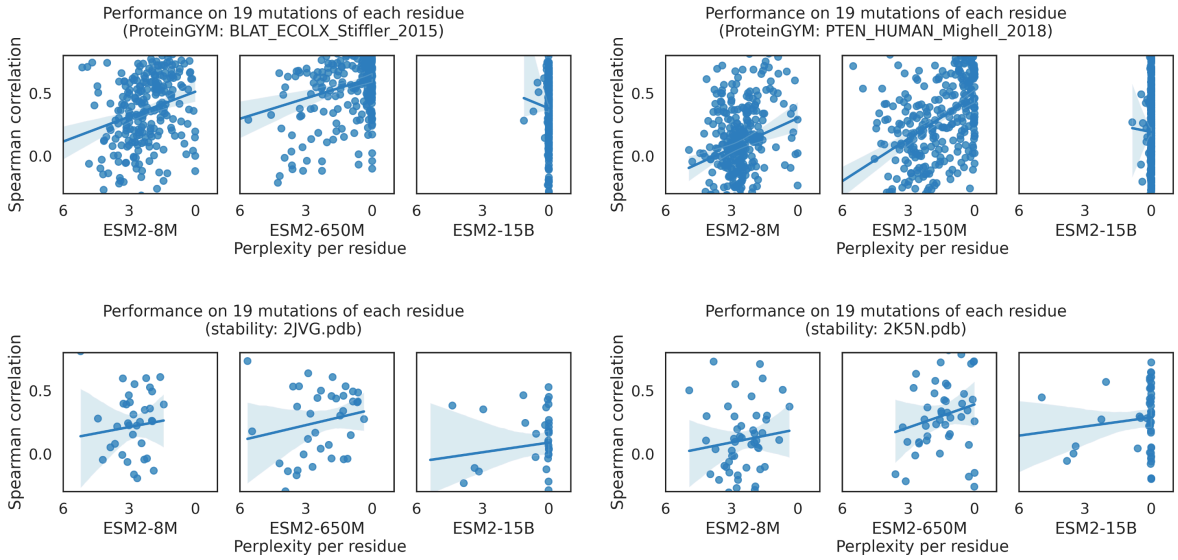

Figure S4. The relationship between ESM2 perplexity and performance on mutation effect per residue.
Each dot is a residue with 19 mutations, the blue line represents linear regression, with shaded regions represent 95% confidence interval.

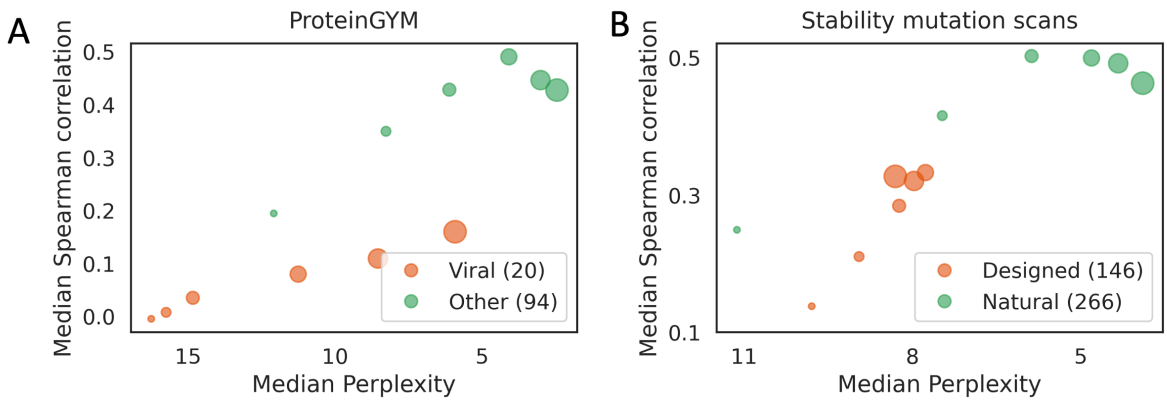

Figure S5. The relationship between ESM2 perplexity and performance on mutation effect.
Marker size indicates ESM2 model size (same scale to Figure 1B-C in main text).