Understanding Language Model Scaling on Protein Fitness Prediction

Chao Hou[1,#], Di Liu[2], Aziz Zafar[2], Yufeng Shen[1,2,3,4,#]

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

4 Program for Mathematical Genomics, Columbia University Irving Medical Center, NY 10032

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

Abstract

Protein language models and models incorporating structure or homologous sequences predict sequence likelihoods *p(sequence)* that reflect the protein fitness landscape and are widely used for mutation effect prediction and protein design. It is widely believed in deep learning field that larger models perform better across tasks. However, for fitness prediction, language model performance declines beyond a certain size, raising concerns about their scalability. Here, we showed that model size, training dataset, and stochastic elements can bias the predicted *p(sequence)* away from real fitness. Model performance on fitness prediction depends on how well *p(sequence)* matches evolutionary patterns in homologs, which is best achieved at a moderate *p(sequence)* level for most proteins. At extreme predicted wild-type sequence likelihoods, models predict uniformly low or high likelihoods for nearly all mutations, failing to reflect the real fitness landscape. Notably, larger models tend to predict proteins with higher *p(sequence)*, which may exceed the moderate range and thus reduce performance. Our findings clarify the scaling behavior of protein models on fitness prediction and provide practical guidelines for selecting and improving them.

Key works: self-supervised deep training, sequence likelihood, protein fitness landscape, mutation effect.

Introduction

Understanding protein fitness—how mutations affect protein function, abundance, activity, and interaction—is a central challenge in biology. It is crucial for elucidating the mechanisms underlying diseases, advancing precision medicine, guiding viral surveillance, and advancing protein design and engineering. While deep mutational scanning[1] (DMS) experiment has been applied to hundreds of proteins, it's time-consuming, labor-intensive, and limited to molecular effects that are easy to assay. To complement experimental efforts, supervised machine learning methods have been developed by training on curated mutation datasets[2–4]. However, these datasets are limited in size and biased toward functionally important genes[2], restricting the models' generalizability and robustness[5]. As a result, there is an urgent need for predictive models capable of unbiased exploration of the protein fitness landscape.

In recent years, self-supervised models have been developed for zero-shot fitness prediction by estimating sequence likelihood *p(sequence)* (Throughout this manuscript, "sequence likelihood" and *p(sequence)* refer to the likelihood of wild-type sequence unless otherwise noted for mutant sequences.): the probability of a protein sequence under the learned evolutionary distribution from natural proteins. self-supervised models often match or even surpass the performance of supervised models[6]. Representative models include protein language models (pLMs) such as ESM2[7]; multi-sequence alignment (MSA)-based models like MSA-Transformer[8] and EVE[9]; inverse folding models such as ESM-IF1[10], which predict sequence from structure; and hybrid models like ESM3[11], which integrate sequence, structure, and other information. Based on their training data, these methods fall into two main categories: **general models** trained on massive datasets of tens to hundreds of millions of proteins from diverse protein families, and **family-specific models** trained on MSA of individual protein family (e.g., EVE, which requires training a separate

model for each protein). MSA-Transformer integrates both family-specific information and general information across diverse protein families by being trained on millions of MSAs[8].

These models are trained to maximize the likelihood of training protein sequences using strategies such as masked or next residue prediction[12], conditioned on sequence, structure, or MSA (**Figure 1A**). While some models incorporate additional training objectives[11], only sequence prediction is used to predict fitness[13]. Fitness of a mutant is estimated by comparing the predicted likelihoods of the mutant and wild-type sequences[6,13] (**Figure 1A**), usually by calculating the log-likelihood ratio (LLR; also called log-odds ratio, log-probability ratio). In this framework, an LLR close to zero means that the mutant is nearly as fit as the wild-type, implying a neutral mutation effect, whereas a strongly negative LLR indicates the mutant is much less fit and potentially deleterious. For models trained via mask prediction, computing the *p(sequence)* is intractable. Instead, pseudo-likelihood is typically used, in which each residue is masked individually and predicted in turn. The LLRs calculated for mutations at the masked position are then used to estimate their fitness[6,13].

Among these models, pLMs gain particular attention for their strong performance and minimal input requirement—they only require sequence as input, without needing MSA or structure. In deep learning field, scaling up models is wildly used to achieve better performance on downstream tasks[14]. pLM scaling has proven effective for residue prediction and structure modeling[7]; however, for fitness prediction, model performance declines beyond a certain size[13,15,16]. This raises a key question in model training: why does scaling up pLMs not consistently improve performance on fitness prediction? While pLMs are typically released in multiple sizes, enabling direct investigation of the scaling behavior, other models are usually released in one size, making it unclear whether the scaling trend applies to them. Beyond scaling, several important questions remain: Under what conditions are these models most effective? And how should one choose the most appropriate model for a given protein?

In this study, we found that various factors unrelated to fitness—including model size, training dataset, and stochastic elements—can influence general model-predicted likelihoods, making them uninformative for fitness prediction in extreme cases. By analyzing diverse general models on a large fitness benchmark[13], we found that the performance of general models depends on how well predicted *p(sequence)* aligns with evolutionary patterns in homologs. Notably, they perform best at a moderate *p(sequence)* level for most proteins, which explain the scaling behavior of pLMs as medium-sized models like ESM2-650M predict more proteins with moderate *p(sequence)*. Our findings clarify how likelihood-based models behave in fitness prediction and lay the groundwork for developing next-generation fitness predictors.

Results

General model-predicted *p(sequence)* is influenced by factors unrelated to protein fitness.

As general models trained on large datasets function as black boxes, we examined whether factors unrelated to fitness can influence the predicted *p(sequence)*. We first investigated two key factors in scaling up models: the size of the training dataset and the number of trainable parameters. To do this, we analyzed predictions from six ESM2 models, ranging from 8 million to 15 billion parameters, on 154 fitness measurements from deep mutational scanning (DMS) experiments in the ProteinGYM[13] benchmark (see Methods for data filtering). ProteinGYM includes proteins from diverse taxa and fitness assays spanning activity, binding, stability, organismal fitness, and expression.

Larger datasets include more protein families and more homologs within each family. Since dissimilar proteins have little influence on the predicted likelihood of a protein[17], we focused on the role of its homologs. We identified UniRef50 (the ESM2 training dataset) homologs with over 20% sequence identity and 80% coverage to proteins in the ProteinGYM benchmark. We first analyzed the number of homologs and found that proteins with more homologs tend to have higher predicted likelihoods (**Figure 1B**, quantified using negative log-likelihood (NLL, *–log p(sequence)*)). However, proteins with high predicted likelihoods do not necessarily have many UniRef50 homologs (**Figure 1B**). We then examined the sequence identity of the closest homologs and observed that proteins with high predicted likelihoods often have highly similar homologs in UniRef50. But the presence of such homologs does not guarantee high

predicted likelihoods (**Figure S1A**). Overall, homologs in the training set can influence model-predicted likelihoods, but the relationship is complex.

We then analyzed model size. These models are typically trained using the NLL of masked or next token as the loss function. As larger models achieve lower training loss, they tend to predict higher sequence likelihoods. We observed this trend for the proteins in the ProteinGYM benchmark (**Figure 1C**). Furthermore, we examined the magnitude of likelihood increase as model size scaled up and observed substantial variability: some proteins showed little or no increase, while others exhibited large gains (**Figure 1D**). Notably, the magnitude of likelihood increase between larger and smaller models are not associated with the likelihoods from smaller models, nor with the number or similarity of homologs in UniRef50 (**Figure 1D, S1B**).

These results suggest that, beyond the homologs in training dataset and model size, additional factors influence model-predicted likelihoods. One such factor is the stochastic elements inherent in model training, such as parameter initialization, data shuffling, and masked residue sampling. To directly assess the impact of these stochastic elements, we analyzed five ESM1v[18] models, which share the same architecture and training dataset but were trained with different random seeds. By comparing the predicted sequence likelihoods from five ESM1v models for each ProteinGYM proteins, we found that approximately 10% of proteins show NLL differences greater than 0.5 among five models (**Figure S2**), and the maximum observed difference is 1.2: the protein Mafg (UniProt ID: O54790) has a predicted sequence NLL of 1.1 in ESM1v_5 but 2.3 in ESM1v_2. Thus, stochastic elements during training can lead models to converge on different local minima, resulting in variability in the predicted sequence likelihoods for some proteins. Overall, model size, training dataset, stochastic elements, and other unknown factors complicate the interpretation of general model-predicted likelihoods.

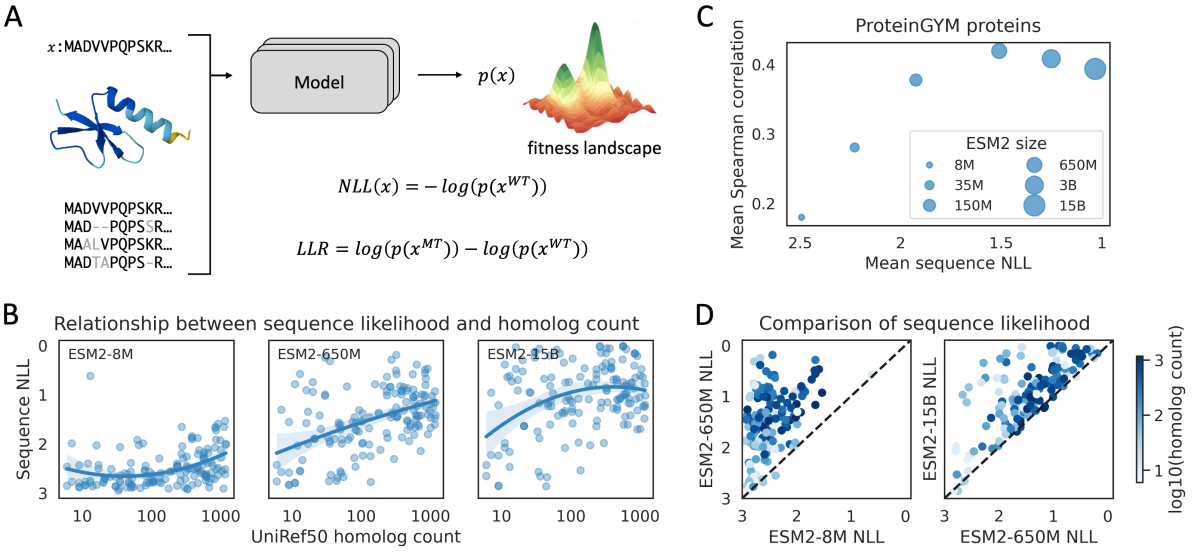

Figure 1. Model-predict sequence likelihood is influenced by factors unrelated to protein fitness.

A, Overview of the calculation of sequence likelihood and log-likelihood ratio (LLR). Models are trained to estimate sequence likelihood $p(x)$ using information from sequence, structure, and homologs, negative log-likelihood (NLL) is usually used as the training loss. For mask-prediction-based models, pseudo-likelihood is used. LLR is calculated based on the entire sequence for generative models, and masked residues for mask-prediction-based models. **B**, Relationship between predicted sequence likelihood and the number of homologs. Each point represents a protein; the x-axes show the number of UniRef50 homologs on the log scale. Homologs are defined as those with ⩾20% sequence identity and ⩾80% coverage. **C**, The mean performance on fitness prediction, and predicted sequence likelihood of ESM2 models on proteins in the ProteinGYM benchmark. Point size indicates ESM2 model size. **D**, Comparison of sequence likelihoods predicted by different ESM2 models. Each point represents a protein; colors indicate the log-scaled number of homologs.

Magnitude of predicted p(sequence) affects fitness estimation by influencing LLR values.

As model-predicted sequence likelihood is affected by many factors, the LLR value, directly tied to sequence likelihood (LLR is the NLL difference between wild-type and mutant sequences, **Figure 1A**) and used to estimate fitness, is also affected. This may impact fitness prediction performance. To investigate this, we focused on model size, as other factors are difficult to control given the models we have access to. Larger models, achieving lower training loss, tend to assign higher probabilities to wild-type residues and lower probabilities to others (since the total probability per site sum to one), resulting in LLRs with larger magnitudes. In extreme cases, a non-informative model that assigns equal probability to all 20 amino acids yields LLRs of zero for all mutations, while an overconfident model that assigns a probability of one to the wild-type and zero to all others produces LLRs of negative infinity. Although certain proteins may predominantly harbor neutral or deleterious mutations, the collapsed LLR distributions in these two extreme scenarios are uninformative for most proteins.

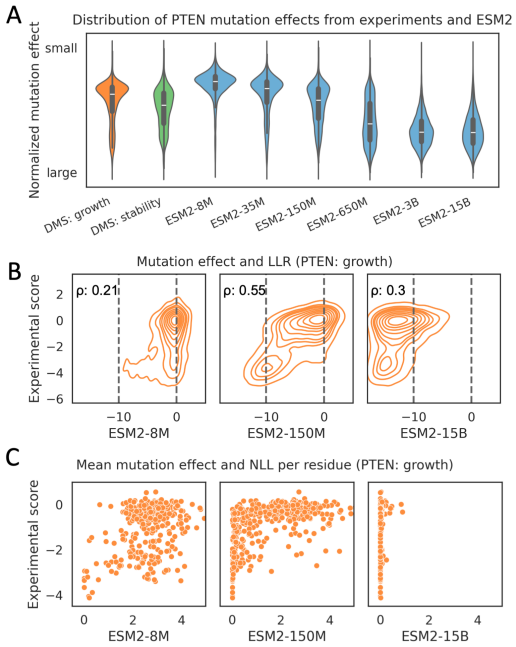

Figure 2. Distributions of experimental and predicted PTEN mutation fitness.

A, Distributions of normalized mutation effects from experiments and ESM2 predictions. Experimental effects and ESM2 LLRs are normalized to the range of 0–1 for visualization. **B**, Relationship between ESM2-predicted LLRs and experimental effects for PTEN mutations. The x-axis represents predicted LLR, the y-axis represents experimental fitness from ProteinGYM. ρ: Spearman correlation. **C**, Relationship between ESM2-predicted probability per residue (quantified using NLL) and mean experimental effects for mutations at each residue (residues with at least 10 mutations are shown). The x-axis represents predicted NLL per residue (*-log p(residue)*), the y-axis represents mean experimental fitness from ProteinGYM. Each point represents a residue.

To illustrate this, we examined predictions of six ESM2 models on PTEN, one of the most extensively studied proteins, for which DMS experiments of both cell growth[19] and protein stability[20] are available. Across both DMS experiments, the distribution of mutation effects is clearly bimodal, with approximately 20% of mutations exhibiting deleterious effects (**Figure 2A**). However, the distributions of ESM2 predicted LLR vary substantially with model size. Smaller models predict LLRs clustered near zero, while larger models predict strongly negative LLRs to most mutations (**Figure 2B, S3A**), mirroring the two extreme scenarios we described above. Notably, the medium-sized model ESM2-150M reproduces the bimodal distribution observed in experiments (**Figure 2B, S3A**) and achieves the best performance, with Spearman correlations of 0.55 for growth and 0.46 for stability. In contrast, both the smallest (ESM2-8M) and largest (ESM2-15B) models yield correlations below 0.3 in both DMS experiments (**Figure 2B, S3A**).

We further investigated this relationship at the residue level by comparing the mean mutation effect per residue with the predicted probability of the wild-type residue. Residues assigned high probability by the model tend to have strongly negative LLRs for mutations, indicating that such residues are mutation sensitive. ESM2-15B assigns high probability approaching one to nearly all residues, while ESM2-8M assigns high probability to very few residues—both failing to capture the experimentally observed distribution of mutation-sensitive residues (**Figure 2C, S3B**). In contrast, ESM2-150M, predicts residue probability that better reflect mean mutation effects (**Figure 2C, S3B**), indicating a better understanding of residue-level importance. We observed similar results in other proteins that also exhibit a rise-then-fall performance trend with increasing ESM2 model size (**Figure S3C-H**).

General model performance on fitness prediction peaks at a moderate level of sequence likelihood.

The above analyses highlight a critical caveat when using general models for fitness prediction: their predicted likelihoods can be influenced by unrelated factors and may not reliably reflect real fitness. To systematically investigate the relationship between model-predicted sequence likelihood and performance on fitness prediction, we evaluated models on 154 DMS experiments from the ProteinGYM benchmark[13] (see Methods for dataset filtering).

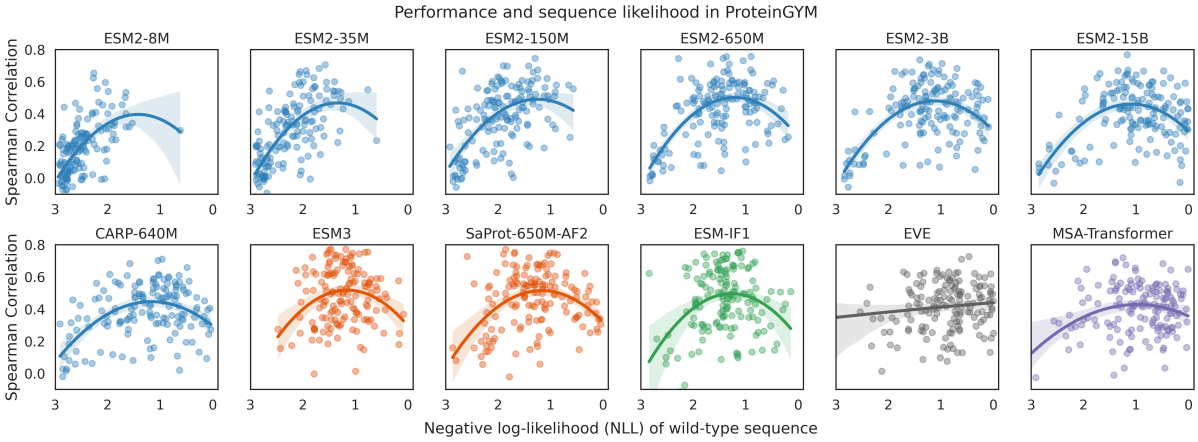

Figure 3. Relationship between fitness prediction performance and model-predicted wild-type sequence likelihood.

The y-axes show the Spearman correlation between model predicted LLRs and experimental effects, the x-axes represent the negative log-likelihood (NLL) of wild-type sequences. Only residues with mutations were included in the NLL calculation. Each point represents an experiment from ProteinGYM (154 experiments after filtering). The curves show second-order polynomial regressions, with shaded areas indicating the 95% confidence intervals, all regression analyses were performed using the Python package Seaborn. Colors indicate model types: blue: protein language models; orange: hybrid models, green: inverse-folding models, grey: family-specific models trained on MSA of individual protein family, and purple: MSA-Transformer that integrate both general and family-specific information.

We first evaluated six ESM2 models. Although larger models consistently predict higher sequence likelihood (**Figure 1C-D**), higher sequence likelihood does not always translate to better performance. Notably, the ESM2-650M model outperforms the larger 3B and 15B models (**Figure 1C**). By comparing wild-type sequence likelihoods with performances of 154 experiments, we observed a bell-shaped relationship, with performance peaking at a moderate likelihood level—corresponding to an wild-type sequence NLL of approximately 1.2 (**Figure 3,** $p(sequence) \approx 0.3$). Notably, within the optimal likelihood range, all ESM2 models, regardless of size, perform comparably (**Figure 3**). This indicates that the level of sequence likelihood, rather than model size, is the primary determinant of model performance on fitness prediction. While more homologs in the ESM2 training set can lead to higher predicted sequence likelihoods (**Figure 1B**), we found that more homologs improve fitness prediction performance (**Figure**

S1C), consistent with our previous results on another DMS datasets[21]. Thus, the performance drop at high likelihood is not caused by homolog overrepresentation in the training set.

We further evaluated a broad range of general models that predict fitness using sequence likelihood. These included Transformer encoder–based mask language models ESM1v[18] and ESMC[22], which share similar architecture with ESM2 but differ in their training datasets; the convolution-based model CARP[23]; and the Transformer decoder–based generative model RITA[24]. We also evaluated hybrid models ESM3[11] and SaProt[25], as well as the inverse folding model ESM-IF1[10]. Despite differences in architecture, input modalities, and training strategies, all these general models exhibit the bell-shaped relationship between performance and wild-type sequence likelihood (**Figure 3, S4**). Remarkably, peak performances occur at the similar level of sequence likelihood (**Figure 3, S4**).

Beyond general models, we also analyzed family-specific models, including a site-independent model[26] that considers each position independently, EVmutation[26] that accounts for co-evolution, and autoencoder-based models DeepSequence[27] and EVE[9]. These models are trained independently on MSA of each protein family with few parameters, making them less susceptible to the unrelated factors that affect general models. Notably, these family-specific models do not exhibit the bell-shaped relationship between performance and sequence likelihood (**Figure 3, S4**). For MSA-Transformer[8] which integrates both general and family-specific information, the bell-shaped trend is present but less pronounced (**Figure 3**).

The bell-shaped relationship arises from model's varying ability to capture context information.

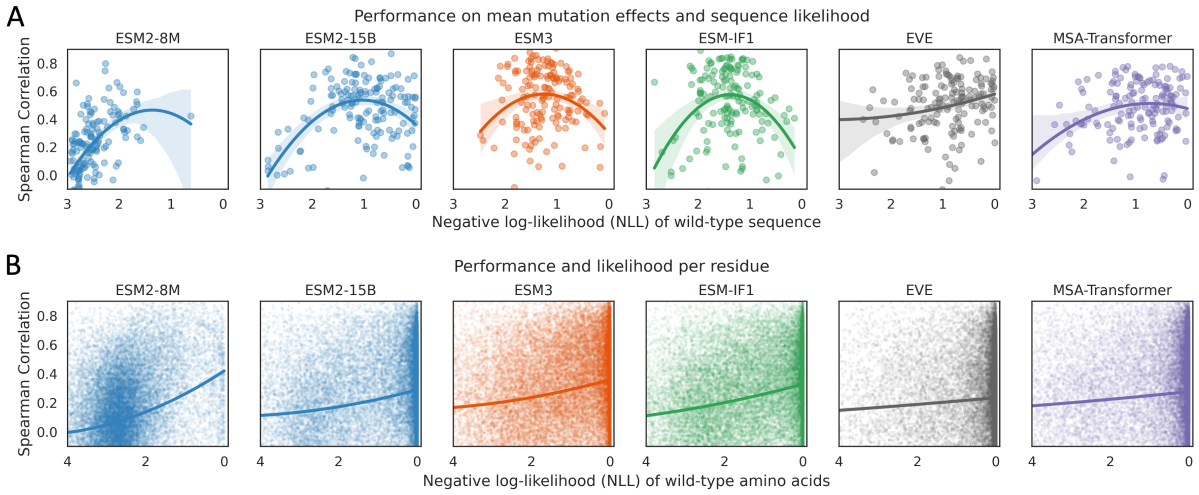

Figure 4. Model performance on mean mutation effects and mutation effects per residue.

A, The y-axes show the Spearman correlation between mean LLRs and mean experimental effects per residue in each protein, reflecting model understanding of context. The x-axes represent the NLL of wild-type sequences. Only residues with mutations were included in the NLL calculation. Each point represents an experiment. **B**, The y-axes show the Spearman correlation between LLRs and experimental effects of all 19 mutations per residue, reflecting model understanding of substitution specificity. The x-axes represent the negative log predicted probability of each residue. Each point represents a residue. The curves show second-order polynomial regressions, with shaded areas representing the 95% confidence intervals.

Fitness prediction requires models to capture both context information (i.e., the sequence and structural context that determine mutation sensitivity) and substitution specificity (i.e., how well different amino acids fit a given context). To investigate the origin of the bell-shaped relationship between performance and sequence likelihood, we disentangled two components. Context understanding was quantified by comparing the mean LLR and the mean experimental effect for mutations on each residue within a protein.

Substitution specificity was assessed by correlating LLRs with experimental effects of 19 mutations at each residue.

By comparing model performance on the two components with predicted likelihood, we found that general models exhibit the bell-shaped relationship between context understanding (mean mutation effect prediction performance) and wild-type sequence likelihood, with performance peaking at a similar likelihood range (**Figure 4A, S5A**). Extreme cases are exemplified by our results for PTEN and other proteins (**Figure 2C, S3**). In contrast, per-residue performance increases monotonically with predicted likelihood (i.e., the predicted probability of each residue; **Figure 4B, S5B**). Family-specific models, however, do not display the bell-shaped trend for either component (**Figure 4, S5**). The performances of all models on two PTEN DMS experiments are shown in Table S1. Notably, all models show substantially stronger performance in understanding context than substitution specificity on 154 experiments (**Figure 4, S5**). None of the models evaluated in ProteinGYM achieves mean Spearman correlation above 0.3 for substitution specificity. These findings indicate that the predictive power of current models primarily stems from their ability to capture protein context, which also underlies the bell-shaped relationship observed between fitness prediction performance and sequence likelihood in general models.

Comparing model-predicted likelihood to biophysical context and mutation sensitivity.

To describe the context of each residue more quantitatively, we considered both the biophysical aspect and mutation sensitivity. Biophysical context was measured using relative solvent accessibility (RSA), which is closely associated with mutation effects[28]: mutations at buried sites tend to destabilize the protein and lead to loss of function, whereas mutations at surface residues typically have minimal effects on stability, with only a small fraction involved in functions. For mutation sensitivity, although the mean experimental effect is informative as used in previous analyses, it is not comparable across DMS experiments. Therefore, we manually examined the distribution of experimental effects. Among the 154 experiments, 122 display bimodal distributions (as exampled in **Figure S6**). For these experiments, we manually defined the threshold between the two modes to distinguish deleterious and neutral mutations (not the threshold provided in ProteinGYM). Mutation sensitivity at each site was then quantified by counting the number of deleterious mutations among the 19 possible mutations. We observed a strong inverse relationship between RSA and the number of deleterious mutations (**Figure S7**), supporting the connection between protein structure and mutation sensitivity (functional importance).

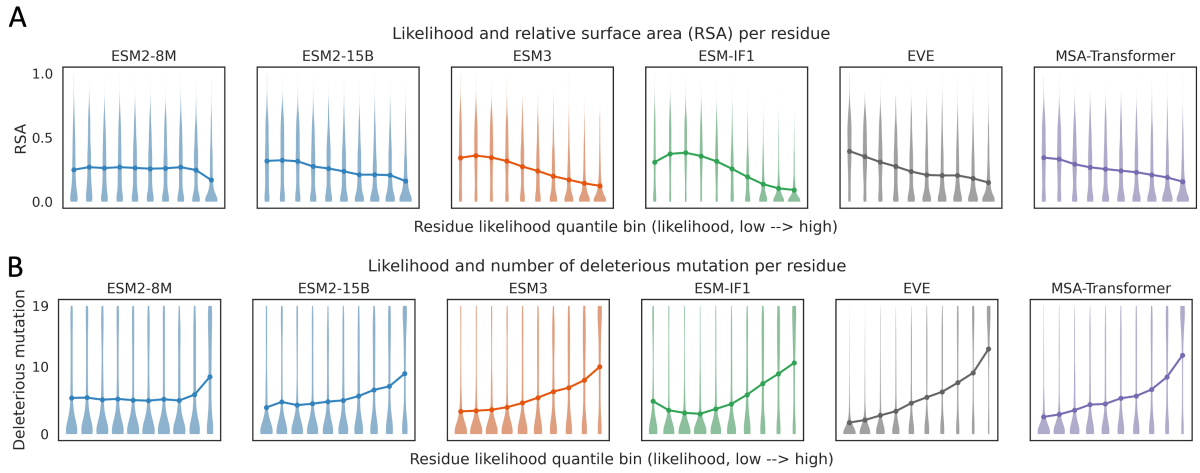

Figure 5. Comparison of model-predicted residue likelihood with the biophysical context and mutation sensitivity.

Residues are grouped into ten quantile bins based on model-predicted probability. For each bin, the distribution and mean value are shown for: **A**, relative solvent accessibility (RSA; range 0–1), and **B**, the number of deleterious mutations per residue (range 0–19). Only experiments with bimodal distribution are shown, only residues with 19 mutations are shown.

By comparing model-predicted per-residue likelihoods with the biophysical context and mutation sensitivity, we found that residues with high predicted likelihoods tend to have lower RSA and more deleterious mutations (**Figure 5**). This demonstrates that both general and family-specific models can capture biophysical and functional context to some extent. Biophysical context is better captured by structure-informed models such as ESM3 and ESM-IF1 at high likelihoods (**Figure 5A**), owing to their direct use of structural input. Mutation sensitivity, however, is better captured by family-specific models (**Figure 5B, S8**). Among the top 10% of high-likelihood residues, the mean number of deleterious mutations is 8.5 for ESM2-8M, 9.0 for ESM2-15B, 9.9 for ESM3, and 10.5 for ESM-IF1, compared to 12.6 for EVE and 11.7 for MSA-Transformer. Conversely, among the bottom 10% of low-likelihood residues, the mean number of deleterious mutations is 5.3 for ESM2-8M, 3.9 for ESM2-15B, 3.3 for ESM3, and 4.9 for ESM-IF1, compared to 1.7 for EVE and 2.5 for MSA-Transformer (**Figure 5B**). These results indicate that general models' predicted high-likelihood residues do not reliably correspond to functionally important (mutation sensitive) sites, while residues predicted with low likelihood are not all unimportant.

The performance of general models depends on how well predicted *p(sequence)* matches evolutionary patterns in homologs.

Finally, we set out to explain the bell-shaped relationship between sequence likelihood and fitness prediction performance. Our previous results showed that predicted likelihoods derived from evolutionary patterns in homologs, as captured by family-specific models, better reflect protein fitness (**Figure 5B**). Although general models are also trained to capture evolutionary information, their predicted sequence likelihoods show poor correlation with those of family-specific models (**Figure S9A**), likely because they are influenced by various unrelated factors. We hypothesized that the bell-shaped curve arises because general models best capture evolutionary information when the predicted sequence likelihood falls within the optimal range in the bell-shaped curve.

To test this, we compared ESM2-15B and EVE, both learn evolutionary patterns from sequence only. Their predicted sequence likelihoods are uncorrelated but show only small differences for many proteins (**Figure 6A**): 72 out of 154 experiments have sequence NLL differences below 0.5. For these 72 experiments, ESM2-15B achieves a mean Spearman correlation of 0.47 (**Figure S10A**)—higher than its overall performance of 0.39 across all 154 experiments, and slightly outperforming EVE's mean of 0.45 on these 72 experiments. Notably, among these 72 experiments, 19 have ESM2-15B sequence NLL values below 0.5, a range corresponds to decreased performance on the bell-shaped curve, yet ESM2-15B still achieves a mean Spearman correlation of 0.45 (**Figure S10B**). In contrast, for experiments with ESM2-15B sequence NLL values below 0.5 but EVE and ESM2-15B NLL differences larger than 0.5, ESM2-15B's mean Spearman correlation is 0.27 (**Figure S10B**). Interestingly, proteins with low NLLs from both models tend to have a higher proportion of deleterious mutations than those with higher EVE NLLs (**Figure S10C**), suggesting that general model-predicted sequence likelihood should reflect proportion of deleterious mutations for accurate fitness prediction.

Overall, we observed that the better a general model's predicted likelihood aligns with the evolutionary pattern in homologs (quantified using EVE), the better its performance on fitness prediction (**Figure 6B**). General models capture evolutionary pattern best at the moderate sequence likelihoods (**Figure 6C**), explaining the bell-shaped relationship. These analyses use EVE NLLs, other family-specific models can also be used since their predicted NLLs are highly correlated (**Figure S9**). While the NLL difference used in Figure 6B can be influenced by the absolute scale of likelihoods, we observed the same trend when using Pearson correlation between the predicted likelihoods of residues within each protein (**Figure S11**), confirming the robustness of this relationship. These results are more pronounced for ESM2 compared to structure-informed models (**Figure 6B, C**), because both pLMs and family-specific models rely solely on sequence data. In contrast, structure-informed models can additionally leverage structural data[29], capturing evolutionary patterns that may differ from those derived from sequence alone. For example,

remote homologs sharing the same structural fold but with low sequence identity can still contribute meaningful information to structure-informed models.

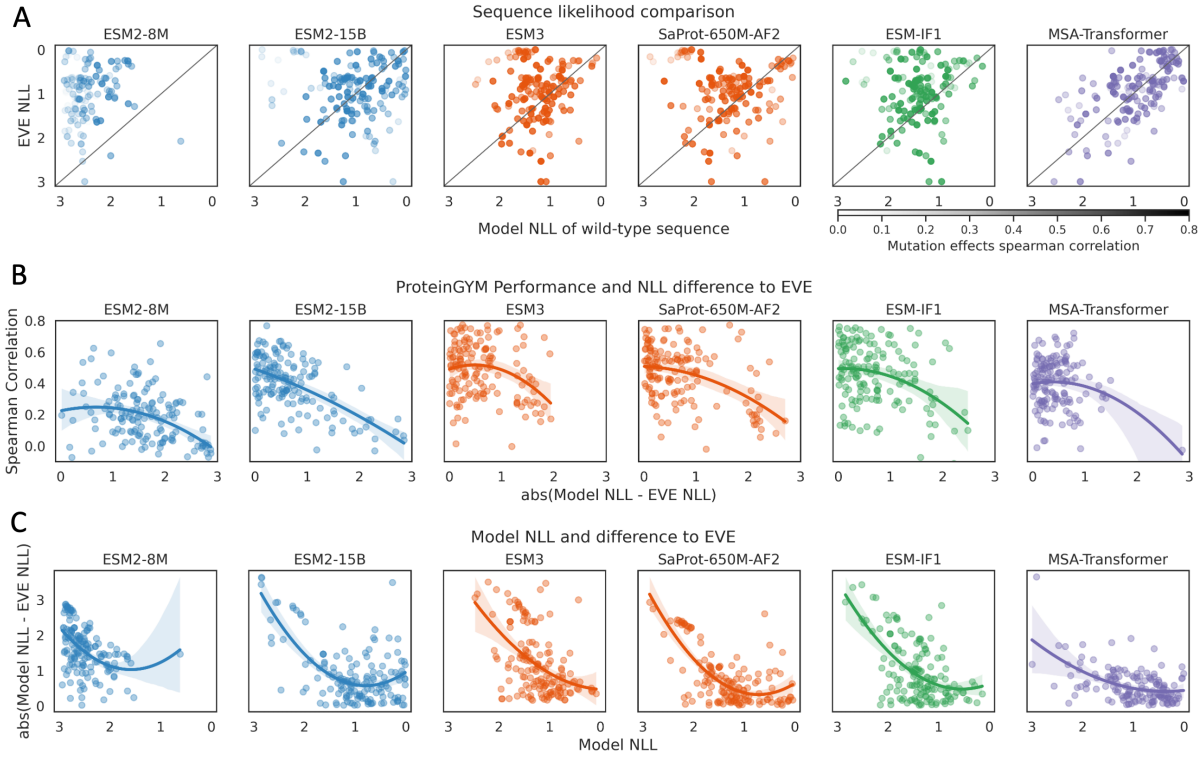

Figure 6. The performance of general models depends on how well they capture evolutionary information.

A, Comparison of sequence NLL from different models to EVE. Each point represents an experiment; darker color indicates better performance on fitness prediction. **B**, Relationship between model performance and NLL difference to EVE. The y-axes show model performance. The x-axes show the absolute difference between the model NLL and EVE NLL. **C**, Relationship between sequence likelihood and difference to EVE. The y-axes show the absolute difference between model NLL and EVE NLL. The x-axes show the model NLL. Each point represents an experiment. The curves represent second-order polynomial regressions, with the shaded areas indicating the 95% confidence intervals.

Discussion

Predicting protein fitness is one of the most important applications of pLMs and other deep learning models in biology. Our work explains why these models can predict fitness, why larger models do not always perform better, and under what conditions they should or should not be used for this task. Overall, we found that the better a model capture evolutionary information, the better its performance on fitness prediction. However, unlike family-specific models that directly capture evolutionary patterns in homologs, general models are influenced by unrelated factors which can decouple predicted likelihoods from true evolutionary patterns. This disconnect leads to poor performance on fitness prediction, particularly at extreme sequence likelihood, which leads to the bell-shaped relationship between model performance and predicted likelihood.

Besides the factors we mentioned in the result section, model-predicted sequence likelihood is also influenced by model architecture and training strategy. For example, models that use MSAs as input tend to predict higher likelihoods, as it's easier to predict sequence with homologous sequences as input. Similarly, high-capacity models produce higher likelihoods due to their increased learning power. This trend is evident in ESM2 models of different sizes (**Figure 1C, D**), as well as in comparisons between EVmutation and the site-independent model (**Figure S4**). Although both use the same framework,

EVmutation assigns higher likelihoods because it has greater capacity by capturing co-evolution between residues, while the site-independent model treats each residue independently. Training strategies also play a role: higher weights on the regularization loss can suppress overfitting, reducing likelihoods of training sequences. Moreover, sequence patterns can also affect predicted likelihoods[30]. Therefore, predicted likelihoods should be interpreted with caution. In our evaluation, most analyses were conducted for different proteins within each model, so differences in model architecture and training strategy do not affect our results. For comparisons between general models and EVE, we additionally used Pearson correlation (**Figure S11**), which captures similarity in likelihood patterns independent of absolute values. Interestingly, while sequence-level likelihoods are uncorrelated between general and family-specific models (**Figure S9A**), we found that within individual proteins, per-residue likelihoods are correlated (**Figure S9B**). This suggests that although general models may not accurately reflect the proportion of mutation-sensitive residues, they still can distinguish important from unimportant sites to some extent.

Weinstein et al.[31] studied the benefits of model misspecification when applying models to predict fitness and attributed the poorer performance of larger pLMs to their improved density estimation. Here, we show that likelihoods predicted by general models (including pLMs) do not always reliably reflect true density in homologs; thus, the poor performance of larger pLMs cannot be simply attributed to improved density estimation. Gordon et al.[32] reported that the preference (measured by likelihood) for a given protein sequence imbued during pretraining is predictive of pLMs' fitness prediction capabilities. However, their conclusion does not explain why pLMs still perform well for some proteins with high predicted likelihoods (**Figure S10B**).

Understanding model-predicted likelihood for different data is a fundamental question in deep learning and the key to understand how models generate their outputs. Interestingly, a bell-shaped relationship has also been observed between LLM sentence likelihoods and human quality judgments[33]. Thus, an important caveat for both LLMs and protein models is that data points predicted with high likelihoods may not be real or biologically meaningful data. Prior work has shown that overparameterized models trained on small datasets tend to memorize data: assigning high likelihoods to training data. In contrast, smaller models trained on large datasets tend to generalize by learning shared patterns[34]. For applying protein models to mutation effect prediction, certain level of generalization, where the model can integrate information from homologs, is more desirable. In our analysis, current models appear to operate in the intermediate (determined by the ratio of training tokens to model parameters): they memorize some proteins / residues and generalize to others. Which proteins / residues a model chooses to memorize versus generalize remains an open question.

Our results offer practical guidance for applying general models to predict fitness, we recommend first verifying whether the predicted likelihoods align with evolutionary patterns in homologs, which can be calculated using simple family-specific models. For training next generation fitness predictors, we recommend incorporating evolutionary patterns during training. This can be achieved by estimating evolutionary patterns from homologs beforehand and encouraging alignment between predicted likelihoods and evolutionary patterns within the training objective. While our manuscript focuses on protein mutation effects, the issue we found is general and may extend to DNA/RNA language models and other applications, such as protein design.

Author contributions

Y.S. and C.H. conceived the study. C.H. performed the analyses and drafted the manuscript. 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

Mutation dataset, structure, and sequence analysis

Experimental mutation effects, model predictions, and protein structures were downloaded from the ProteinGYM website (https://proteingym.org/) and its GitHub repository in May 2025. From the 217 mutational scanning datasets available, we included only residues with all 19 possible substitutions and proteins containing at least 20 such residues, resulting in a dataset of 486,932 single-residue substitution mutations across 154 experiments.

Protein solvent-accessible surface areas (SASA) were calculated using the MDTraj package (version: 1.10.2). Relative surface areas were computed by dividing the SASA of each residue by the maximum SASA observed for that amino acid across all proteins.

For homolog detection, MMseqs2 (version 16.747c6) was used to search against UniRef50 (Release 2025_03) with the following parameters: *--min-seq-id 0.2 -c 0.8 --max-accept 1000*.

Sequence likelihood and log-likelihood ratio

ProteinGYM provides mutant sequence likelihoods for generative models and log-likelihood ratios (LLRs) for mask-prediction-based models and family-specific models we evaluated. We used their predictions instead of running all models ourselves. Some models incorporate additional information alongside sequence likelihood, such as surface area; we excluded these models from our analysis. For models with both single model and ensemble predictions, single model predictions were used. The site independent model in ProteinGYM is trained using the EVmutation framework and considers each site independently.

For the generative models we evaluated, we followed the procedures and codes provided by ProteinGYM to calculate the per-residue likelihood of the wild-type sequence. For the generative language model RITA-xl, we used the average likelihood from both the forward and reverse sequences as recommended. For ESM-IF1, only the forward pass was used. For other models that provide LLRs in ProteinGYM, the per-residue likelihood (probability) of the wild-type amino acid was calculated using the following equations, provided that LLRs for all 19 possible mutations at a site were available.

$$p(wt) + \sum p(mt_i) = 1$$

$$1/p(wt) = 1 + \sum p(mt_i)/p(wt) = 1 + \sum e^{LLR_i}$$

$$-log\big(p(wt)\big) = log(1 + \sum e^{LLR_i})$$

Here, $p(wt)$ represents the model-predicted probability of the wild-type residue, and $p(mt_i)$ represents the probability of mutation i. The probabilities of the wild-type residue and all 19 possible mutations sum to one. For calculating sequence likelihoods using five ESM1v models, the sudo-likelihood was used, where a single residue is masked at a time.

The LLR and likelihood values for mask-prediction-based methods were computed using the mask-one-residue approach, a widely adopted strategy for applying these models to fitness prediction. Alternatively, models like ESM1,2 can also be applied without masking, using only the wild-type sequence, which requires just one forward pass per protein. The sequence likelihood computed without masking is highly correlated with that from the mask-one-residue approach[17], and the LLR scores derived from both

methods show nearly identical performance in predicting mutation effects[6]. Therefore, our conclusions also apply when using pLMs with the single-pass approach.

Among general models, the bell-shaped relationship is less pronounced for the generative model RITA (Figure S4). In particular, its peak performance at the optimal likelihood remains relatively low compared to other models. One possible explanation is that autoregressive generative models are less suitable for mutation effect prediction, as they only use information from preceding residues. In ProteinGYM benchmark, they tend to underperform compared to masked language models, which consider bidirectional information.

Reference

1. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801–807 (2014).
2. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* (2023) doi:10.1126/science.adg7492.
3. Zhang, H., Xu, M. S., Fan, X., Chung, W. K. & Shen, Y. Predicting functional effect of missense variants using graph attention neural networks. *Nat Mach Intell* **4**, 1017–1028 (2022).
4. Gao, H. *et al.* The landscape of tolerated genetic variation in humans and primates. *Science* **380**, eabn8153 (2023).
5. Livesey, B. J. & Marsh, J. A. Using deep mutational scanning to benchmark variant effect predictors and identify disease mutations. *Mol. Syst. Biol.* **16**, e9380 (2020).
6. 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).
7. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
8. Rao, R. *et al.* MSA Transformer. (2021) doi:10.1101/2021.02.12.430858.
9. Frazer, J. *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021).
10. Hsu, C. *et al.* Learning inverse folding from millions of predicted structures. *bioRxiv* 2022.04.10.487779 (2022) doi:10.1101/2022.04.10.487779.
11. Hayes, T. *et al.* Simulating 500 million years of evolution with a language model. *Science* **387**, 850–858 (2025).
12. Xiong, J. *et al.* Guide your favorite protein sequence generative model. Preprint at https://doi.org/10.48550/arXiv.2505.04823 (2025).
13. Notin, P. *et al.* ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design. *Adv. Neural Inf. Process. Syst.* **36**, 64331–64379 (2023).
14. Kaplan, J. *et al.* Scaling Laws for Neural Language Models. Preprint at https://doi.org/10.48550/arXiv.2001.08361 (2020).
15. Bhatnagar, A. *et al.* Scaling unlocks broader generation and deeper functional understanding of proteins. 2025.04.15.649055 Preprint at https://doi.org/10.1101/2025.04.15.649055 (2025).
16. Notin, P. Have We Hit the Scaling Wall for Protein Language Models? *Pascal Notin* https://pascalnotin.substack.com/p/have-we-hit-the-scaling-wall-for (2025).
17. Gordon, C., Lu, A. X. & Abbeel, P. Protein Language Model Fitness Is a Matter of Preference. *bioRxiv* 2024.10.03.616542 (2024) doi:10.1101/2024.10.03.616542.
18. Meier, J. *et al.* Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv* 2021.07.09.450648 (2021) doi:10.1101/2021.07.09.450648.
19. Mighell, T. L., Evans-Dutson, S. & O'Roak, B. J. A Saturation Mutagenesis Approach to Understanding PTEN Lipid Phosphatase Activity and Genotype-Phenotype Relationships. *Am. J. Hum. Genet.* **102**, 943–955 (2018).
20. Matreyek, K. A., Stephany, J. J., Ahler, E. & Fowler, D. M. Integrating thousands of PTEN variant activity and abundance measurements reveals variant subgroups and new dominant negatives in cancers. *Genome Med.* **13**, 165 (2021).
21. Hou, C., Zhao, H. & Shen, Y. Learning Biophysical Dynamics with Protein Language Models. *bioRxiv* 2024.10.11.617911 (2025) doi:10.1101/2024.10.11.617911.
22. Evolutionary Scale · ESM Cambrian: Revealing the mysteries of proteins with unsupervised learning. https://www.evolutionaryscale.ai/blog/esm-cambrian.
23. Yang, K. K., Fusi, N. & Lu, A. X. Convolutions are competitive with transformers for protein sequence pretraining. *Cell Syst.* **15**, 286-294.e2 (2024).
24. Hesslow, D., Zanichelli, N., Notin, P., Poli, I. & Marks, D. RITA: a Study on Scaling Up Generative Protein Sequence Models. Preprint at https://doi.org/10.48550/arXiv.2205.05789 (2022).

25. Su, J. *et al.* SaProt: Protein Language Modeling with Structure-aware Vocabulary. 2023.10.01.560349 Preprint at https://doi.org/10.1101/2023.10.01.560349 (2024).
26. Hopf, T. A. *et al.* Mutation effects predicted from sequence co-variation. *Nat Biotechnol* **35**, 128–135 (2017).
27. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nat Methods* **15**, 816–822 (2018).
28. Tsishyn, M., Hermans, P., Rooman, M. & Pucci, F. Residue conservation and solvent accessibility are (almost) all you need for predicting mutational effects in proteins. *Bioinformatics* **41**, btaf322 (2025).
29. Mahajan, A. What could Alphafold 4 look like? (Sergey Ovchinnikov, Ep #3). https://www.owlposting.com/p/what-could-alphafold-4-look-like (2025).
30. Kantroo, P., Wagner, G. P. & Machta, B. B. In-Context Learning can distort the relationship between sequence likelihoods and biological fitness. Preprint at https://doi.org/10.48550/arXiv.2504.17068 (2025).
31. Weinstein, E. N., Amin, A. N., Frazer, J. & Marks, D. S. Non-identifiability and the Blessings of Misspecification in Models of Molecular Fitness. 2022.01.29.478324 Preprint at https://doi.org/10.1101/2022.01.29.478324 (2023).
32. Gordon, C. W., Lu, A. X. & Abbeel, P. Protein Language Model Fitness is a Matter of Preference. in (2024).
33. Zhang, H., Duckworth, D., Ippolito, D. & Neelakantan, A. Trading Off Diversity and Quality in Natural Language Generation. Preprint at https://doi.org/10.48550/arXiv.2004.10450 (2020).
34. Morris, J. X. *et al.* How much do language models memorize? Preprint at https://doi.org/10.48550/arXiv.2505.24832 (2025).