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 that incorporate structure or homologous sequences, estimate sequence likelihoods *p(sequence)* that reflect the protein fitness landscape and are commonly used in mutation effect prediction and protein design. It is widely believed in deep learning field that larger model performs 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 their application and future development.

Keywords: 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 estimation 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 distribution from natural proteins. Self-supervised models often match or even surpass the performance of supervised models[6,7]. Representative models include protein language models (pLMs) such as ESM2[8]; multi-sequence alignment (MSA)-based models like MSA-Transformer[9] and EVE[6]; 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[9].

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 (see Methods for details)[13]. Fitness of a mutant is estimated by comparing the predicted likelihoods of the mutant and wild-type sequences[7,13] (**Figure 1A**), usually by calculating the log-likelihood ratio (LLR). 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 and LLR calculated from marginal approaches are used (see Methods for details)[7,13].

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 a common strategy to improve performance on downstream tasks[14]. pLM scaling has proven effective for residue prediction and structure modeling[8]; 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 to choose the most appropriate model for a given protein?

In this study, we systematically investigated the relationship between sequence likelihood and the performance of fitness prediction across diverse models. 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 a large fitness benchmark[13], we found that the performance of general models depends on how well their 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 predict fitness 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 training 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, version2021_04) homologs with over 20% sequence identity and 80% coverage to proteins in the ProteinGYM benchmark. By analyzing the number of homologs, we 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 is not associated with the likelihoods from the smaller model, nor with the number or similarity of homologs in the training set (**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 could be the stochastic elements introduced during model training, including parameter initialization, data shuffling, and masked residue sampling. To directly assess the impact of these stochastic elements, we analyzed five ESM1v[18] models, which have 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. This result suggests that stochastic elements 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.

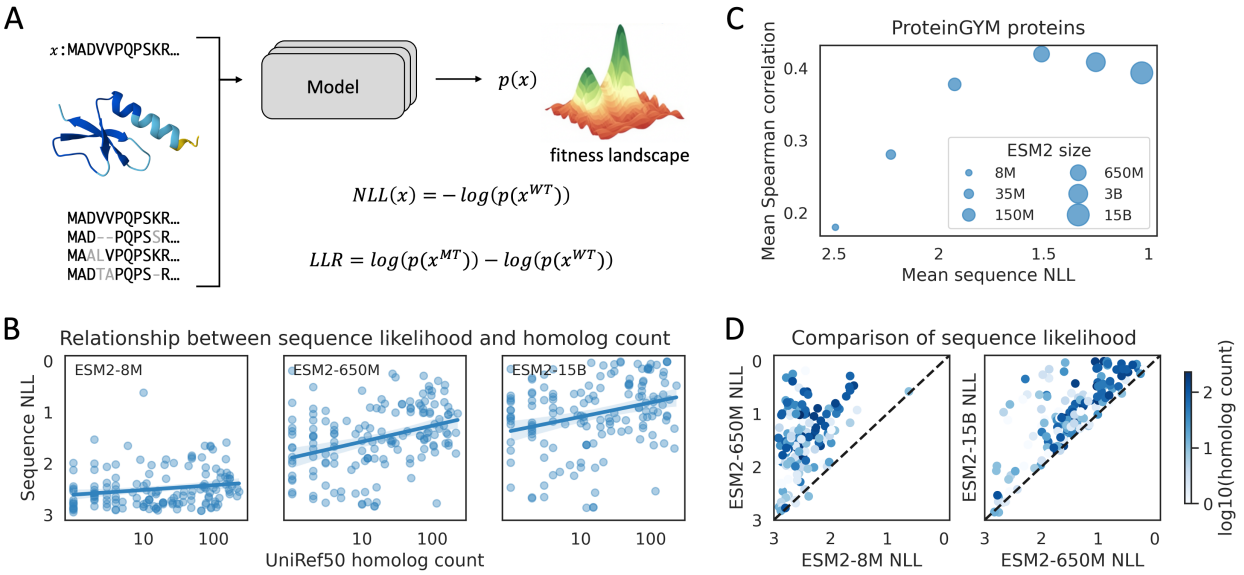

Figure 1. Model-predict sequence likelihood is influenced by factors unrelated to protein fitness.

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 in the log scale. Homologs are defined as those with ⩾20% sequence identity and ⩾80% coverage. The curves represent linear regressions, with the shaded areas indicating the 95% confidence intervals. **C**, The mean performance on fitness prediction, and predicted sequence likelihood of ESM2 models on 154 DMS experiments 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 probability to the wild-type amino acid and lower probabilities to the others (since the total probability per site sum to one), resulting in LLRs with larger magnitudes. In extreme cases, a non-informative model that predicts 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.

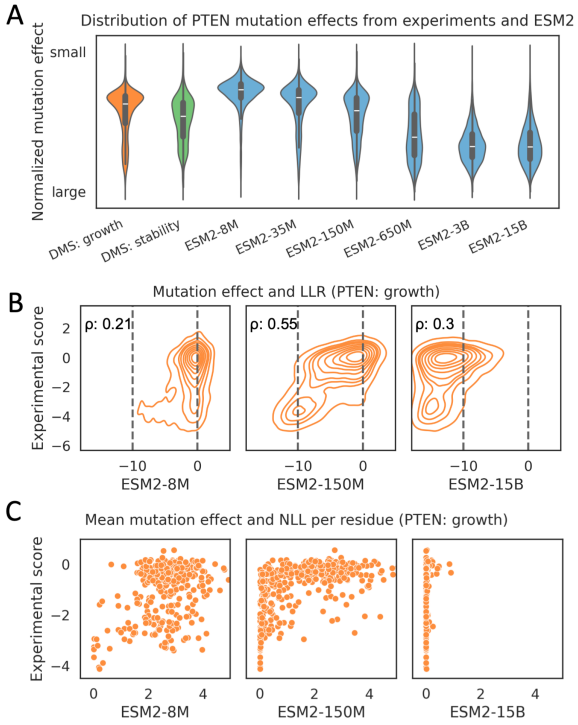

Figure 2. Distributions of experimental and predicted PTEN mutation fitness.

A, Distributions of normalized mutation effects from DMS 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, 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 (ESM2-150M outperforms ESM2-650M on PTEN), 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).

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 a 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 the performance drop at high likelihood is not caused by homolog overrepresentation in the training set (**Figure S1C**).

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] (see Methods for details). 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 are achieved at the similar level of sequence likelihood (**Figure 3, S4**).

Beyond general models trained on diverse protein families, we also analyzed family-specific models. These included simple frequency-based site-independent models that treat each position independently with or without homologous sequence weighting (see Methods for details), and the variational autoencoder-based model EVE[6], which captures inter-residue dependencies. These models are trained independently on MSA of each protein family with no or much less 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[9] which

integrates both general and family-specific information, the bell-shaped trend is present but less pronounced (**Figure 3**).

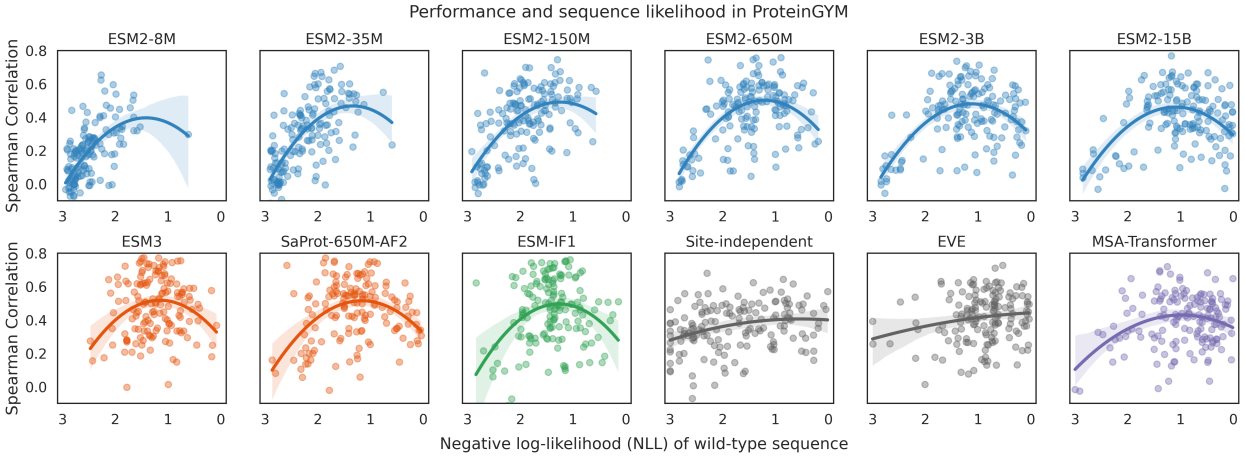

Figure 3. Relationship between fitness prediction performance and model-predicted wild-type sequence likelihood.

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.

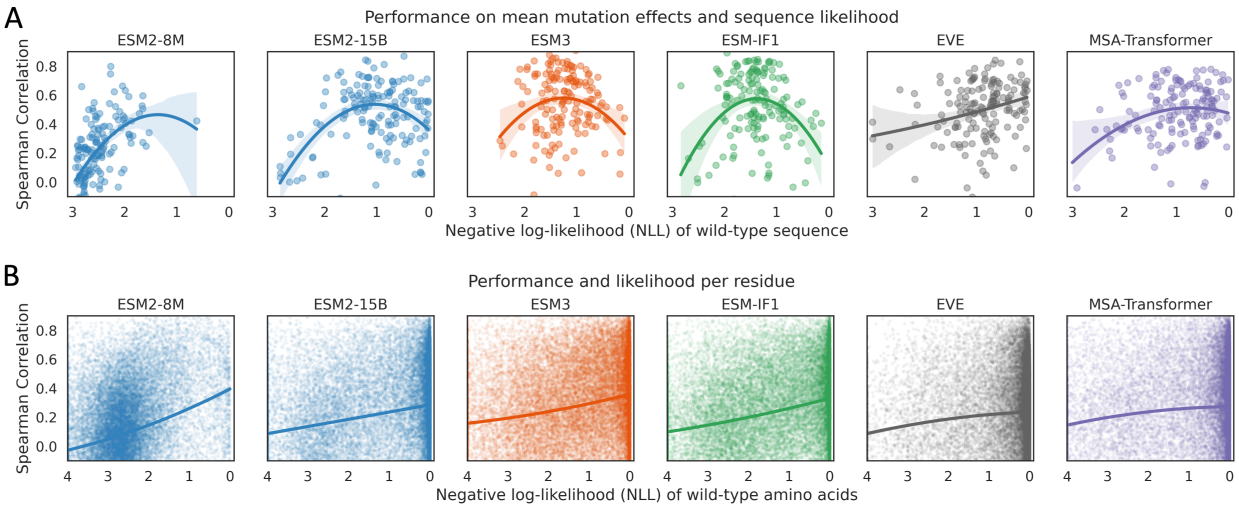

Figure 4. Model performance on mean mutation effects and mutation effects per residue.

A, The y-axes show the Spearman correlation between mean LLRs and mean experimental effects per residue in each protein, reflecting model understanding of protein 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.

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

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**). 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 fitness 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[26]: 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. 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).

By comparing model-predicted per-residue likelihoods with the biophysical context and mutation sensitivity, we found that residues with high predicted likelihoods by all methods 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 12.8 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 1.6 for EVE and 2.5 for MSA-Transformer (**Figure 5B**). Furthermore, at the protein level, we observed that proteins with high EVE sequence likelihood tend to harbor more deleterious mutations (**Figure S9A**)—a trend not observed for ESM2 (**Figure S9B**). These results indicate that, compared with family-specific models, general model predicted likelihood do not reliably reflect functionally importance (mutation-sensitivity).

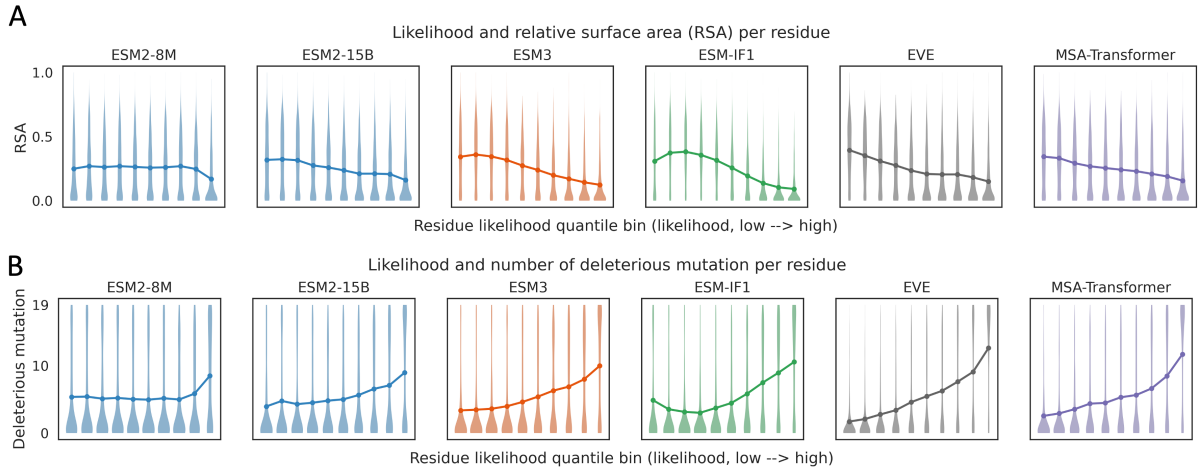

Figure 5. Comparison of model-predicted residue likelihood with the biophysical context and mutation sensitivity.

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 of mutation effect are shown, only residues with 19 mutations are shown.

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. We began by comparing the predicted likelihoods from different methods. Family-specific model predicted likelihoods directly reflect evolutionary patterns in homologous sequences: a high residue-level likelihood indicates a conservated residue, while a high protein-level likelihood indicates a larger fraction of conserved residues. Even though the family-specific models we evaluated employ different architectures, their predicted likelihoods are highly correlated at both the residue and protein levels (**Figure 6A-B**). General models, in contrast, implicitly learn evolutionary information from large-scale datasets. Likelihoods predicted by different general models are also correlated at both the residue and protein levels (**Figure 6A-B**). However, when comparing general models to family-specific models, correlations are observed only at the residue level, with protein-level likelihoods showing no correlation (**Figure 6A-B**). Because likelihoods derived from homologs by family-specific models better reflect protein fitness (**Figure 5B, S9A**), We therefore hypothesized that the bell-shaped curve arises because general models capture evolutionary information most effectively when their predicted sequence likelihood falls within a moderate range.

To test this, we compared ESM2 and EVE, both of which learn evolutionary patterns from sequence alone. We used EVE likelihoods to represent evolutionary patterns in homologous sequences and quantified how well ESM2 captures evolutionary patterns by computing the Pearson correlation between residue-level likelihoods predicted by ESM2 and EVE. We found that the better ESM2 captured evolutionary information, the better it performed in fitness prediction (**Figure 6C**). Across 154 DMS experiments, ESM2-15B achieves a mean Spearman correlation of 0.47 for the 69 datasets where it captured evolutionary information well (NLL Pearson correlation with EVE > 0.6), but only 0.23 for the 34 datasets where it failed to do so (NLL Pearson correlation with EVE < 0.3) (**Figure S9C**). Notably, even for proteins with sequence NLL below 0.5, a range corresponding to decreased performance on the bell-shaped curve, ESM2-15B still achieves a mean Spearman correlation of 0.43 for the eight datasets where it captured evolutionary information well (**Figure S9D**). Notably, ESM2 captures evolutionary information most strongly at moderate sequence likelihoods (**Figure 6D**), which corresponds to the range where its fitness prediction performance is optimal (**Figure 3**).

Overall, the ability of ESM2 to predict fitness relies on how well its predicted likelihoods align with evolutionary patterns observed in homologs. While we used EVE likelihoods to represent evolutionary patterns, other family-specific models can also be used as their predictions are highly correlated (**Figure 6A–B**). We also compared other general models to EVE and found that ESM3 and ESM-IF1 behaved differently (**Figure S10B**). Unlike pLMs, these structure-informed models can capture evolutionary patterns that are different from those derived from sequence alone. For example, remote homologs that share the same structural fold but have low sequence identity can still provide meaningful information to structure-informed models as their input structure are similar, but may not contribute meaningfully to sequence-only models.

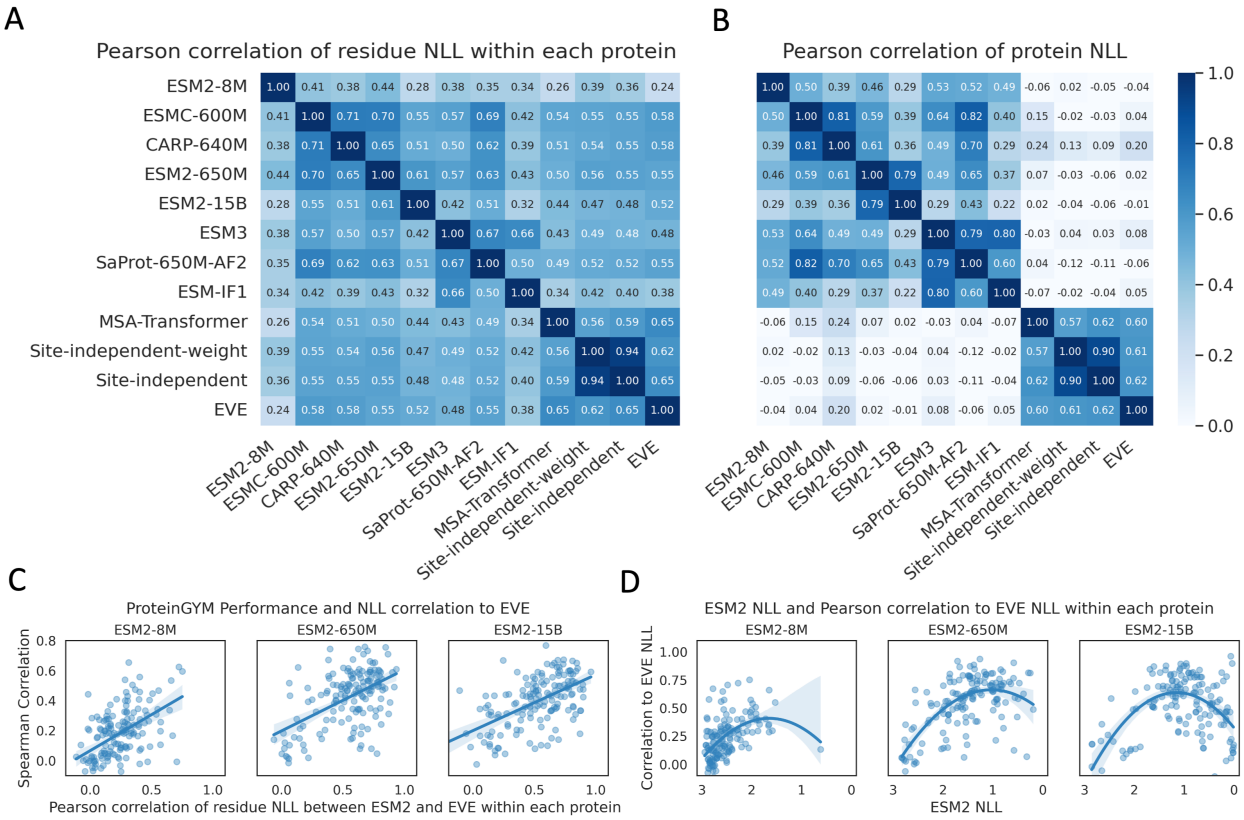

Figure 6. The performance of general models depends on how well they capture evolutionary information.

A, Pearson correlation of per-residue NLL within each protein, the mean correlation across proteins in 154 experiments were shown. **B**, Pearson correlation of whole-sequence NLL, which is calculated as the mean of per-residue NLL within the protein. **C**, Relationship between ESM2 fitness prediction performance and correlation to EVE NLL. The y-axes show fitness prediction performance. The x-axes show the Pearson correlation between ESM2 NLL and EVE NLL per residue within each protein. Each point represents an experiment. The curves represent linear regressions, with the shaded areas indicating the 95% confidence intervals. **D**, Relationship between ESM2 sequence likelihood and correlation to EVE NLL. The y-axes show the Pearson correlation between ESM2 NLL and EVE NLL per residue within each protein. The x-axes show the ESM2 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 general model captures evolutionary information, the better its performance on fitness prediction. However, unlike family-specific models that directly utilize 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. 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[27]. Therefore, predicted likelihoods should be interpreted with caution. In our evaluation, most analyses were conducted within each model, so differences in model architecture and training strategy do not affect our results. For comparisons between models, we used Pearson correlation which captures similarity in likelihood patterns independent of absolute values.

Weinstein et al.[28] studied the benefits of model misspecification in fitness prediction and attributed the decreased performance of larger pLMs to their improved density estimation. Here, we show that likelihoods predicted by pLMs do not always reliably reflect the true density of homologs (**Figure 6B**); thus, the reduced performance of larger pLMs cannot be solely explained by improved density estimation. Gordon et al.[29] reported that the preference (measured by likelihood) for a given protein sequence established during pretraining is predictive of pLM fitness prediction performance. However, they did not provide a mechanistic explanation, nor did they explain why pLMs still perform well for certain proteins with high predicted likelihoods (**Figure S9D**). Yu et al.[30] found that model prediction entropy is related to viral protein fitness prediction, with low entropy (corresponding to high sequence likelihood) associated with better performance. Gurev et al.[31] reported that larger pLMs perform better on viral protein fitness prediction. While their conclusions on viral proteins may appear to conflict with our findings, we note that viral proteins are underrepresented in protein datasets and are therefore assigned relatively low likelihoods by general models. As a result, their sequence likelihoods have not reached the optimal levels observed for well-represented proteins, thus the expected performance decline with increasing likelihood or model size is not observed (**Figure S11**). This phenomenon also extends to other proteins that are poorly learned by general models, such as de novo designed proteins (**Figure S11**).

Understanding model-predicted likelihood is a fundamental question in language models 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[32]. 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[33]. 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 study 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.

Supplementary information

Supplementary Figures 1-11

Supplementary Tables:

1. 154 DMS experiments evaluated in this study, including manually defined cutoffs for classifying deleterious and neutral mutations.
2. Performance on 154 DMS experiments across different methods.
3. Mean mutation effect prediction performance.
4. Performance per residue.
5. Predicted likelihood per residue from different methods.

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

Dataset, protein structure, and sequence analysis

In this study, we focused exclusively on single-residue substitutions. Other mutation types—including multiple-residue substitutions, insertions, deletions, and truncations—were not considered for the following reasons: (1) current model performance on these mutation types is markedly lower compared to single substitutions[13]; (2) several of the evaluated methods cannot be directly applied to these mutation types; and (3) while some studies have extended models to these mutation types, the predictions are derived from prediction of single-residue substitutions[7] (e.g., the effect of a multiple-residue substitution is estimated as the sum of the corresponding single-substitution effects, and the effect of a truncation as the maximum single-substitution effect in the truncated region).

Experimental mutation effects, model predictions, Multiple sequence alignments (MSAs), and protein structures were downloaded from the ProteinGYM website (https://proteingym.org/) and its GitHub repository (https://github.com/OATML-Markslab/ProteinGym) in May 2025. From the 217 mutational scanning datasets available, only residues with experimental effects of all 19 possible substitutions were included. Restricting to proteins with at least 20 such residues resulted in a dataset comprising 486,932 single-residue substitution mutations across 154 experiments. The thresholds for classifying deleterious and neutral mutations provided in ProteinGYM were not used, as they are defined based on the effect distribution of both single and multiple substitutions and appeared unreasonable for some proteins in our visualizations. Instead, we manually defined thresholds using only single substitutions. Another DMS dataset[34] of protein folding stability ("Tsuboyama2023_Dataset2_Dataset3_20230416.csv") was also analyzed.

ProteinGYM provides predictions from a wide range of methods, only likelihood-based methods were evaluated here. For models offering both single-model and ensemble predictions, the single-model predictions were used.

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 (version 2021_04, which was used to train ESM2) with the following parameters: *--min-seq-id 0.2 -c 0.8 --max-accept 1000*.

Applying masked language models for protein fitness

Masked language models (MLMs) evaluated in the study include ESM2, ESM1v, ESMC, CARP, ESM3, SaProt, and MSA-Transformer. Among them, ESM2, ESM1v, ESMC, and CARP are sequence-only MLMs; ESM3 and SaProt are sequence and structure hybrid MLMs; MSA-Transformer is an MSA-based MLM. For MLM, computing the exact sequence likelihood $p(\boldsymbol{x})$ is intractable. Instead, pseudo-likelihood is used, where each residue is masked individually and predicted in turn. For a protein of length L with sequence $\boldsymbol{x} = (x_1, \ldots, x_L)$ (x_i denotes the wild-type amino acid, which is also explicitly written as $x_{i,wt}$ when it needs to be distinguished from mutations in the following sections.), structure $\boldsymbol{s}$ (structure token of the protein is also a list of tokens with length L), and a set of homologs $\boldsymbol{h}$, the pseudo-likelihood of sequence-only MLM is defined as:

$$log\left(\hat{p}_{sequence\ MLM}(\boldsymbol{x})\right) = \frac{1}{L}\sum_{i=1}^{L} log\left(p\left(x_i \mid \boldsymbol{x}_{\backslash i}\right)\right) \tag{1}$$

Where $\boldsymbol{x}_{\backslash i}$ denotes the sequence with position i masked.

ESM3 and SaProt were trained with both sequence and structure tokens. ESM3 predicts the probabilities of sequence and structure tokens separately, whereas SaProt predicts them jointly. Specifically, SaProt predicts the probabilities of 400 combinations of sequence–structure tokens, corresponding to 20 amino acids × 20 FoldSeek[35] structure tokens. During fitness inference, all structure tokens were provided, while the residue at the mutation site was masked. For SaProt, the predicted probability for each amino acid was obtained by summing over 20 sequence–structure combinations with that amino acid. The pseudo-likelihood of sequence and structure hybrid MLM is defined as:

$$log\left(\hat{p}_{hybrid\ MLM}(\boldsymbol{x})\right) = \frac{1}{L}\sum_{i=1}^{L} log\left(p\left(x_i \mid \boldsymbol{x}_{\backslash i}, \boldsymbol{s}\right)\right) \tag{2}$$

MSA-Transformer leverages homologous sequences to provide evolutionary context. The choice of homologs can be made by random sampling, selecting the most similar, or selecting the most dissimilar sequences[9]. In this study, we followed the protocol used in ProteinGYM, where homologs were sampled according to sequence-similarity–based weights (also used for EVE). The pseudo-likelihood of MSA-Transformer is defined as:

$$log\left(\hat{p}_{MSA-Transformer}(\boldsymbol{x})\right) = \frac{1}{L}\sum_{i=1}^{L} log\left(p\left(x_i \mid \boldsymbol{x}_{\backslash i}, \boldsymbol{h}\right)\right) \tag{3}$$

For fitness prediction, the log-likelihood ratio (LLR) between the mutated sequence $\boldsymbol{x_{mt}}$ and the wild-type sequence $\boldsymbol{x_{wt}}$ can be computed using the pseudo-likelihood of their full sequences:

$$LLR_{pseudo-likelihood} = log\left(\hat{p}(\boldsymbol{x_{mt}})\right) - log\left(\hat{p}(\boldsymbol{x_{wt}})\right) \tag{4}$$

However, this approach is computationally intensive, since for each mutated sequence of length L, computing the pseudo-likelihood requires L forward passes through the model, making it impractical for large-scale fitness inference. Instead, simplified approximations are used, including the masked marginal

and wild-type marginal approaches, both of which consider only the predicted probabilities at the mutation site rather than the likelihood of the entire sequence. The underlying hypothesis is that the difference in predicted probabilities at a single residue is proportional to the difference in overall sequence likelihood.

For the masked marginal approach, the residue at the mutation site is masked, and the model directly predicts the probabilities of the wild-type and mutant amino acids, the LLR is then calculated as:

$$LLR_{masked-marginal} = log\left(p\left(x_{i,mt}|\boldsymbol{x}_{\backslash i}\right)\right) - log\left(p\left(x_{i,wt}|\boldsymbol{x}_{\backslash i}\right)\right) \tag{5}$$

Previous studies have shown that for MLMs trained with the standard masking scheme (e.g., ESM2, where 15% of tokens are randomly selected during pre-training, with 80% replaced by a mask token, 10% kept unchanged, and 10% replaced by a random token), predictions from the unmasked wild-type sequence can be used directly for fitness inference[7]. In this case, the LLR is calculated as:

$$LLR_{wild-type-marginal} = log\left(p\left(x_{i,mt}|\boldsymbol{x}\right)\right) - log\left(p\left(x_{i,wt}|\boldsymbol{x}\right)\right) \tag{6}$$

The computational cost differs substantially among these methods. For a protein of length L, scoring all possible single-residue substitutions (19 per position) requires running the model L×*19*×L times with $LLR_{pseudo-likelihood}$, L times with $LLR_{masked-marginal}$, and only one time with $LLR_{wild-type-marginal}$. In this study, we used masked marginal approach for all MLM methods, which is the most used methods to estimate fitness from MLM. The sequence likelihood computed without masking is highly correlated with that from the masked approach[17], and the LLR scores derived from both approaches show nearly identical performance in predicting mutation effects[7]. Therefore, our conclusions also apply when using ESM2 with the wild-type marginal approach.

For all models, the predicted probability of all possible tokens in each position sum to one:

$$p\left(x_{i,wt}\right) + \sum_{a!=wt} p(x_{i,a}) = 1 \tag{7}$$

Where a represent amino acids different from the wild-type. ProteinGYM provides LLRs for the MLMs we evaluated, but not the underlying sequence likelihoods. However, given the definition of LLRs (Eq. 5) together with Eq. 7, the predicted residue probabilities can be inferred using the following equations:

$$1\Big/p\left(x_{i,wt}\right) = 1 + \sum_{a!=wt} p\left(x_{i,a}\right)\Big/p\left(x_{i,wt}\right) = 1 + \sum_{a!=wt} e^{LLR_a} \tag{8}$$

$$log\left(p\left(x_{i,wt}\right)\right) = -log\left(1 + \sum_{a!=wt} e^{LLR_a}\right) \tag{9}$$

Eq. 8 and 9 were used to estimate sequence likelihoods for positions with LLRs of all 19 possible mutations in the ProteinGYM dataset, eliminating the need to run these models. Sequence likelihoods of five ESM1v models and ESM2 prediction for the DMS dataset[34] of protein folding stability were computed by us.

Applying Autoregressive models for protein fitness

We evaluated two autoregressive generative models: the protein language model RITA-xl and the inverse-folding model ESM-IF1. Both models were trained in an autoregressive manner: RITA on both forward and reverse sequences, and ESM-IF1 on forward sequences only (N-terminal to C-terminal). For autoregressive models, the sequence likelihood can be computed explicitly as:

$$log\left(p(\boldsymbol{x})\right) = \frac{1}{L}\sum_{i=1}^{L} log\left(p\left(x_i \mid \boldsymbol{x}_{<i}\right)\right) \tag{10}$$

Where $\boldsymbol{x}_{<i}$ denotes the residues before position i. Specifically, for RITA, the likelihood of a sequence were computed as the mean of predictions from the forward and reverse sequences:

$$log\big(p_{RITA}(\boldsymbol{x})\big) = \frac{1}{2L}\sum_{i=1}^{L}[log\big(p(x_i \mid \boldsymbol{x}_{<i})\big) + log\big(p(x_i \mid \boldsymbol{x}_{>i})\big)] \tag{11}$$

For ESM-IF1, the likelihood of a sequence was computed as:

$$log\big(p_{ESM-IF1}(\boldsymbol{x})\big) = \frac{1}{L}\sum_{i=1}^{L} log\big(p(x_i \mid \boldsymbol{x}_{<i}, \boldsymbol{s})\big) \tag{12}$$

For fitness prediction using autoregressive models, the sequence likelihood is computed directly for each mutated sequence. Since the wild-type sequence likelihood is constant across mutations, calculating LLR is unnecessary for performance evaluation. ProteinGYM provides negative log likelihood (NLL) for each mutated sequence rather than LLRs. The sequence likelihoods for wild-type proteins were computed using Eq. 11 and 12.

The bell-shaped relationship is less pronounced for RITA compared to other general models (Figure S4). In the evaluation from Gordon et al.[29], the bell-shaped relationship for generative model ProGen2[36] is also less pronounced. In particular, their peak performance at the optimal likelihood remains relatively low compared to other models. This may reflect an intrinsic limitation of autoregressive pLMs for fitness prediction, as they cannot simultaneously incorporate information from both upstream and downstream residues.

We also investigated whether autoregressive pLMs could be applied using marginal approaches, which can reduce the times of model running from 2× the number of sequences to only two (forward and reverse sequences). In this setting, the probabilities of both the wild-type and mutant amino acids at a given position are predicted using upstream and/or downstream residues as context. Using RITA-xl, we evaluated four strategies: (1) computing the LLR from the forward sequence only, (2) computing the LLR from the reverse sequence only, (3) averaging the LLRs obtained from the forward and reverse sequences, and (4) averaging the predicted probabilities from the forward and reverse sequences before computing the LLR. On the ProteinGYM dataset, these four strategies achieved mean Spearman correlations of 0.28, 0.28, 0.31, and 0.29, respectively, all substantially lower than the mean Spearman correlation of 0.37 obtained when using the full sequence likelihood. Thus, in the Results section, only the results obtained using the full sequence likelihood for generative models were reported.

Site-independent model

The site-independent model was used as a baseline family-specific model. MSAs provided by ProteinGYM were used, and homologous sequences with more than 50% gaps were removed. For each site, the probabilities of the 20 amino acids and the gap character were calculated as:

$$p(x_i = a) = \frac{\sum_h w_h \delta(x_i^h, a) + \epsilon}{\sum_h w_h + 21\epsilon} \tag{13}$$

where w_h denotes the weight of homologous sequence h, which was either calculated using the same strategy as in EVE or set as one for all sequences. δ is the Kronecker delta. ϵ is the pseudo count, for the unweighted model, ϵ was set to 1; for the weighted model, ϵ was set to the minimum positive value in the frequency table. Fitness was estimated using LLR of mutation and wild-type amino acids at each position. Benchmarking on the ProteinGYM mutations evaluated in this study, the site-independent (unweighted) and the site-independent-weight model achieved mean Spearman correlations of 0.36 and 0.35. Despite their simplicity, they outperform the site-independent model trained with EVmutation[37] (mean Spearman correlation 0.33).

Family-specific model EVE

EVE models were trained separately for each MSA using the code provided in the ProteinGYM GitHub repository, taking approximately a week on ten A6000 GPUs. Model was trained with the random seed of 0 and *threshold_focus_cols_frac_gaps* = 1 (all positions were modeled). As a variational autoencoder, EVE does not provide exact sequence likelihoods. Instead, the evidence lower bound (ELBO) was used as a proxy, calculated as:

$$ELBO = \mathbb{E}_{q(z|\boldsymbol{x})}\left[log\left(p(\boldsymbol{x} \mid z)\right)\right] - KL[q(z \mid \boldsymbol{x}) \mid\mid p(z)] \tag{14}$$

Where z is the latent variable, $q(z \mid \boldsymbol{x})$ is the approximate posterior from the encoder, $p(\boldsymbol{x} \mid z)$ is the decoder likelihood, and $KL[\cdot||\cdot]$ denotes the Kullback–Leibler divergence (KLD). The ELBO was computed for both wild-type and mutated sequences by averaging over 20,000 samples of z to estimate fitness. The decoder conditional probability $p(\boldsymbol{x} \mid z)$ was used to approximate the probability of each amino acid at each site, averaged over the 20,000 samples using wild-type sequence as input.

Using EVE to estimate fitness is computationally intensive, as it requires training a separate model with millions of parameters for each MSA, and during inference, the decoder must be run tens of thousands of times per mutated sequence to obtain a stable ELBO estimation. To reduce this burden, we implemented a simplified approach, which we termed EVE-marginal. In this approach, conditional probabilities $p(\boldsymbol{x} \mid z)$ are obtained from the wild-type sequence only, and fitness is estimated like the MLM wt-marginal approach. EVE-marginal achieves a mean Spearman correlation of 0.41 on the ProteinGYM mutations we evaluated, comparable to the full EVE model (0.41), while being substantially more computationally efficient.

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. Frazer, J. *et al.* Disease variant prediction with deep generative models of evolutionary data. *Nature* **599**, 91–95 (2021).
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. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
9. Rao, R. *et al.* MSA Transformer. (2021) doi:10.1101/2021.02.12.430858.
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. 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).
27. 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).
28. 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).
29. Gordon, C. W., Lu, A. X. & Abbeel, P. Protein Language Model Fitness is a Matter of Preference. in (2024).
30. Yu, Y., Jiang, F., Zhong, B., Hong, L. & Li, M. Entropy-driven zero-shot deep learning model selection for viral proteins. *Phys. Rev. Res.* **7**, 013229 (2025).
31. Gurev, S., Youssef, N., Jain, N. & Marks, D. S. Variant effect prediction with reliability estimation across priority viruses. 2025.08.04.668549 Preprint at https://doi.org/10.1101/2025.08.04.668549 (2025).
32. 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).
33. Morris, J. X. *et al.* How much do language models memorize? Preprint at https://doi.org/10.48550/arXiv.2505.24832 (2025).
34. Tsuboyama, K. *et al.* Mega-scale experimental analysis of protein folding stability in biology and design. *Nature* **620**, 434–444 (2023).
35. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* (2023) doi:10.1038/s41587-023-01773-0.
36. Nijkamp, E., Ruffolo, J. A., Weinstein, E. N., Naik, N. & Madani, A. ProGen2: Exploring the boundaries of protein language models. *Cell Syst.* **14**, 968-978.e3 (2023).
37. Hopf, T. A. *et al.* Mutation effects predicted from sequence co-variation. *Nat Biotechnol* **35**, 128–135 (2017).