Learning Protein Fitness Landscapes with Multimodal Stability Priors

Shannon Zhang [1] **Yunan Luo** [1]

Abstract

Predicting how mutations change protein fitness is central to protein engineering and variant interpretation, yet most experimentally measured fitness landscapes contain only limited labeled variants. We present PsiFit, a stability-informed framework that adapts protein language models for low-N fitness prediction by injecting mutation-induced stability changes predicted by a multimodal sequence-structure foundation model. By integrating biophysical stability priors into contrastive fine-tuning, PsiFit aims to improve data efficiency, reduce overfitting, and provide a general strategy for learning protein fitness landscapes from sparse assays.

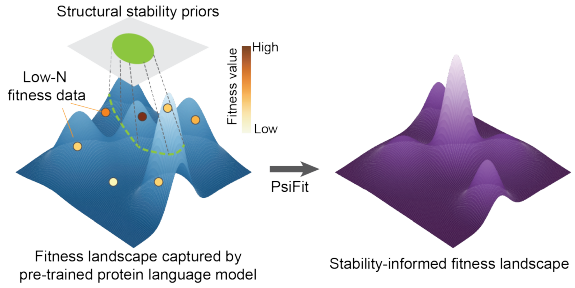

Figure 1. **Overview of PsiFit.** A pre-trained protein language model captures an initial sequence–fitness landscape, but sparse low-N fitness measurements may be insufficient to fully calibrate this landscape for a specific protein fitness. PsiFit incorporates a structural stability prior, derived from predicted mutation-induced stability changes, to guide contrastive fine-tuning and produce a stability-informed fitness landscape for improved low-N protein fitness prediction.

1. Introduction

Protein fitness prediction asks how one or more amino acid substitutions alter a protein property of interest, such as binding affinity, expression, catalytic activity, stability, or organismal fitness (Yang et al., 2019; Wittmann et al., 2021; Freschlin et al., 2022; Kouba et al., 2023). This problem is central to machine learning-guided protein engineering, where the goal is to search an exponentially large sequence space for variants with improved or novel functions. Despite the growing availability of deep mutational scanning (DMS) datasets (Notin et al., 2023), most practical engineering campaigns remain data-limited: experimental assays are costly, protein-specific, and often yield only tens to hundreds of labeled variants (Biswas et al., 2021). This low-N regime creates a mismatch between the small size of supervised fitness data and the large capacity of modern protein foundation models (Zhao et al., 2024).

Protein language models (pLMs) have emerged as powerful foundation models for protein biology. Pre-trained on large-scale natural protein sequences, they capture evolutionary and biophysical constraints that transfer to downstream tasks, including zero-shot mutation effect prediction (Meier et al., 2021; Lin et al., 2023; Nijkamp et al., 2023). However, directly fine-tuning a large pLM on sparse fitness data can lead to catastrophic forgetting and overfitting. A recent contrastive fitness learning framework (Zhao et al., 2024) addressed this challenge by reprogramming pLMs through contrastive fine-tuning: rather than fitting absolute fitness values, it calibrates pLM-derived likelihood scores to preserve the experimentally observed ranking between variants. This strategy improves low-N fitness prediction while maintaining the pre-trained knowledge encoded in the pLM.

In this work, we propose PsiFit (Fig. 1), a stability-informed extension of contrastive fitness learning. The key biological motivation is that proteins generally need to maintain sufficient structural stability to perform their functions. A mutation that strongly destabilizes the folded state is therefore likely to reduce many downstream fitness readouts, even when the assay measures a property other than stability. To incorporate this prior, PsiFit uses mutation-induced stability changes predicted by SPURS (Li & Luo, 2025), a multimodal protein foundation model that rewires a sequence-based pLM with a structure-based inverse folding model to predict $\Delta\Delta G$ values for amino acid substitutions. PsiFit

[1]School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA, USA. Correspondence to: Yunan Luo <yunan@gatech.edu>.

Proceedings of the ICML 2026 3rd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences, Seoul, Korea. 2026.

injects these stability predictions directly into the final-layer logits of a pLM and learns a position-specific stability bias during contrastive fine-tuning. In this way, the model integrates sequence-derived evolutionary information, structure-informed stability priors, and sparse experimental fitness labels within a unified learning framework.

2. Methods

2.1. Problem formulation

Let $x^{\mathrm{WT}} = (x_1^{\mathrm{WT}}, \ldots, x_L^{\mathrm{WT}})$ denote a wild-type protein sequence of length L, where each residue belongs to the amino acid alphabet $\mathcal{A}$ with $|\mathcal{A}| = 20$. A mutant sequence x differs from x^{WT} at a set of mutated sites

$$\mathcal{M}(x) = \{i : x_i \neq x_i^{\mathrm{WT}}\}. \tag{1}$$

We are given a small labeled fitness dataset

$$\mathcal{D} = \{(x^{(k)}, y^{(k)})\}_{k=1}^{N}, \tag{2}$$

where $y^{(k)}$ is an experimentally measured fitness value. In the low-N setting, N may be as small as < 100 variants. The goal is to learn a protein-specific predictor that ranks unseen variants by fitness.

2.2. Protein language model fitness score

PsiFit builds on a masked pLM f_θ initialized from pre-trained weights θ_0. Given a protein sequence context, the pLM outputs a logit matrix

$$X_\theta(x) \in \mathbb{R}^{L \times |\mathcal{A}|}, \tag{3}$$

where $X_{\theta,i,a}$ denotes the logit for amino acid a at position i. The corresponding conditional distribution is

$$p_\theta(a \mid x_{-i}) = \frac{\exp(X_{\theta,i,a})}{\sum_{b \in \mathcal{A}} \exp(X_{\theta,i,b})}. \tag{4}$$

Following prior pLM-based mutation effect prediction (Meier et al., 2021), the fitness proxy of a mutant is computed as the log-likelihood ratio between mutant and wild-type residues at the mutated sites:

$$\hat{y}_\theta(x) = \sum_{i \in \mathcal{M}(x)} \left[\log p_\theta(x_i \mid x_{-\mathcal{M}}) - \log p_\theta(x_i^{\mathrm{WT}} \mid x_{-\mathcal{M}})\right]. \tag{5}$$

Here, $x_{-\mathcal{M}}$ denotes the sequence context with mutated positions masked. This score preserves the pLM pre-training output space while providing a scalar proxy for variant fitness.

2.3. Multimodal stability prior from SPURS

To incorporate biophysical stability information, PsiFit uses SPURS (Li & Luo, 2025) to predict mutation-induced stability changes. SPURS integrates a sequence-based pLM and a structure-based inverse folding model through neural rewiring. Specifically, it uses ProteinMPNN (Dauparas et al., 2022) to encode structural features from the wild-type protein backbone and injects these features into ESM-derived sequence representations (Lin et al., 2023) through lightweight adapter modules. This design fuses evolutionary constraints from large-scale sequence pre-training with geometric constraints from protein structure.

For a wild-type sequence and structure, SPURS predicts a stability matrix

$$G \in \mathbb{R}^{L \times |\mathcal{A}|}, \tag{6}$$

where $G_{i,a}$ denotes the predicted $\Delta\Delta G$ for substituting the wild-type residue at position i with amino acid a. Positive $\Delta\Delta G$ values indicate destabilizing mutations, whereas negative values indicate stabilizing mutations. Importantly, SPURS produces $\Delta\Delta G$ predictions for all possible single substitutions in one forward pass, making it efficient for site-saturation scoring.

2.4. Stability-biased pLM logits

To incorporate stability information, PsiFit applies a small multilayer perceptron (MLP) to transform the predicted stability landscape into a learned logit-level correction:

$$B_\phi(x) = \mathrm{MLP}_\phi\big(G(x)\big), \tag{7}$$

where ϕ denotes the trainable parameters of the calibration module. The stability-biased pLM logits are then defined as

$$X'_{\theta,\phi}(x) = X_\theta(x) + B_\phi(x). \tag{8}$$

The resulting stability-aware amino acid distribution is

$$p_{\theta,\phi}(a \mid x_{-i}) = \mathrm{softmax}\big(X'_{\theta,\phi,i,:}(x)\big)_a. \tag{9}$$

This formulation allows the model to learn how stability information should reshape the pLM likelihood landscape during fine-tuning. Because the relationship between stability and measured fitness can vary across proteins, positions, and assay types, the MLP-based calibration module provides a flexible mechanism for reweighting stability effects rather than imposing a fixed stability penalty. The calibrated logits are then used in the contrastive fine-tuning objective to align stability-aware pLM scores with experimentally observed fitness rankings.

Using the stability-biased distribution, PsiFit computes the mutant fitness proxy as

$$\hat{y}_{\theta,\alpha}(x) = \sum_{i \in \mathcal{M}(x)} \left[\log p_{\theta,\alpha}(x_i \mid x_{-\mathcal{M}}) - \log p_{\theta,\alpha}(x_i^{\mathrm{WT}} \mid x_{-\mathcal{M}})\right]. \tag{10}$$

2.5. Contrastive fine-tuning

PsiFit therefore follows a contrastive learning objective adapted from ConFit (Zhao et al., 2024). For two labeled variants $(x^{(i)}, y^{(i)})$ and $(x^{(j)}, y^{(j)})$ with $y^{(i)} > y^{(j)}$, the model is encouraged to assign a higher predicted score to $x^{(i)}$:

$$\mathcal{L}_{\mathrm{BT}} = \sum_{y^{(i)} > y^{(j)}} \log \left[1 + \exp\left(- \frac{\hat{y}_{\theta,\alpha}(x^{(i)}) - \hat{y}_{\theta,\alpha}(x^{(j)})}{\tau} \right) \right], \tag{11}$$

where τ is a temperature hyperparameter. This Bradley–Terry-style loss converts N labeled variants into $O(N^2)$ pairwise comparisons, improving data efficiency under low-N supervision.

To prevent the adapted model from drifting too far from the pre-trained pLM distribution, we add a KL regularization term:

$$\mathcal{L}_{\mathrm{KL}} = \sum_{x \in \mathcal{B}} \sum_{i \in \mathcal{M}(x)} D_{\mathrm{KL}} \left(p_{\theta_0}(\cdot \mid x_{-\mathcal{M}}) \,\|\, p_{\theta,\alpha}(\cdot \mid x_{-\mathcal{M}}) \right), \tag{12}$$

where $\mathcal{B}$ is a mini-batch and p_{θ_0} is the original pre-trained pLM distribution. The final training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{BT}} + \lambda_{\mathrm{KL}} \mathcal{L}_{\mathrm{KL}} \tag{13}$$

In implementation, we fine-tune the pLM using parameter-efficient adaptation, such as LoRA (Hu et al., 2022).

3. Results

3.1. Benchmark setting

We evaluated PsiFit on matched protein fitness prediction benchmarks where results were available for all compared methods. The evaluation used a low-N supervised setting with $N = 96$ labeled variants for model training. We compared PsiFit with two representative sequence-model baselines: ESM-1v, a leading unsupervised zero-shot pLM predictor (Meier et al., 2021), and Augmented DeepSequence, a state-of-the-art supervised low-N fitness prediction method that combines sequence features with evolutionary density scores (Hsu et al., 2022). We conducted the evaluation on the ProteinGym benchmark (Notin et al., 2023), which includes over 200 deep mutational scanning (DMS) datasets spanning diverse protein fitness measurements, such as catalytic activity, binding affinity, expression, stability, and organismal fitness. Following SPURS (Li & Luo, 2025), we excluded DMS datasets measuring stability to test whether PsiFit generalizes beyond the stability prior incorporated in the model. For each dataset, performance was measured by Spearman correlation between model predictions and experimentally measured fitness values on held-out variants. This metric directly evaluates whether a model can correctly rank candidate variants, which is the primary use case in ML-guided protein engineering.

3.2. Overall prediction performance

Across 114 matched non-stability DMS datasets, PsiFit achieved the best overall performance among the three evaluated methods (Fig. 2a). PsiFit reached a mean Spearman correlation of 0.509 and a median correlation of 0.527, compared with 0.433 mean and 0.461 median for ESM-1v, and 0.479 mean and 0.490 median for Augmented DeepSequence. Thus, PsiFit improved the average Spearman correlation by 0.075 over ESM-1v and by 0.029 over Augmented DeepSequence. In paired dataset-level comparisons, PsiFit outperformed ESM-1v on 84 of 114 datasets and outperformed Augmented DeepSequence on 79 of 114 datasets. PsiFit also achieved the best performance among the three methods on 65 datasets, compared with 25 datasets for Augmented DeepSequence and 24 datasets for ESM-1v. Paired Wilcoxon signed-rank tests confirmed that the improvements are statistically significant against both baselines (PsiFit vs. ESM-1v: $p < 0.001$; PsiFit vs. Augmented DeepSequence: $p = 2.32 \times 10^{-5}$.

The improvement over ESM-1v suggests that incorporating limited experimental supervision and stability-informed calibration can improve upon zero-shot evolutionary likelihood scoring. The improvement over Augmented DeepSequence is more modest but important, as Augmented DeepSequence is already a strong supervised low-N baseline that uses both sequence features and evolutionary density information. These results indicate that stability-informed pLM adaptation provides additional predictive value beyond commonly used sequence and evolutionary features.

3.3. Fitness-category-stratified performance

We next stratified the results by the coarse fitness categories defined in ProteinGym (Notin et al., 2023) (Expression, Activity, Binding, and Organismal Fitness) (Fig. 2b). Within each category, datasets were sorted by PsiFit performance to visualize how the three methods behave across the full range of prediction difficulty. PsiFit maintains competitive performance across all four categories, indicating that the stability-informed prior is not limited to assays that directly measure stability. Its advantage is most apparent on datasets where the zero-shot ESM-1v baseline performs moderately but leaves room for supervised adaptation; in these cases, PsiFit often improves over ESM-1v while matching or exceeding Augmented DeepSequence. At the same time, the relative ordering among methods varies across datasets, reflecting the biological heterogeneity of fitness measurements and suggesting that some assays are dominated by evolutionary constraints or assay-specific signals not fully explained by stability.

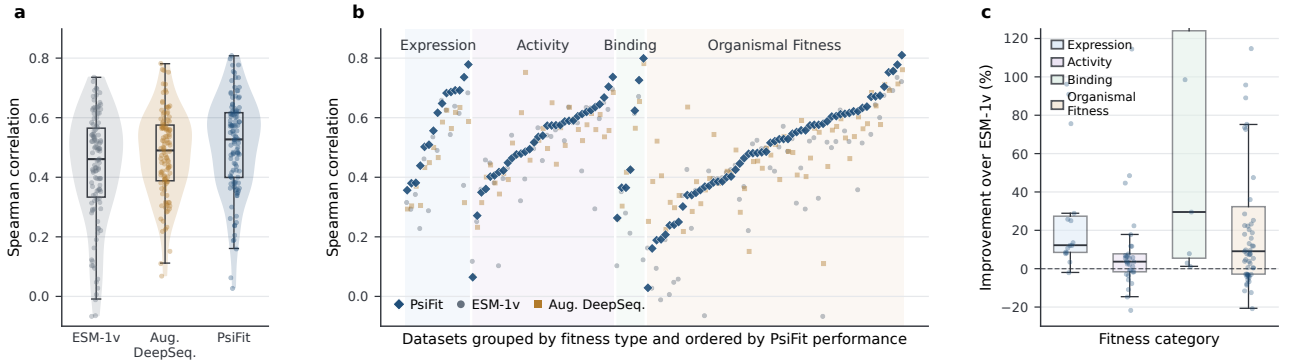

Figure 2. **Comparison of PsiFit with ESM-1v and Augmented DeepSequence across matched protein fitness prediction benchmarks**. Performance is measured by Spearman correlation between model predictions and experimentally measured fitness values. **(a)** Overall distribution of Spearman correlations across all matched datasets. Violin plots summarize the full distribution, transparent box plots show the median and interquartile range, and overlaid points represent individual datasets. **(b)** Dataset-level comparison grouped by fitness category. Datasets are first grouped by coarse fitness type and then sorted within each group by PsiFit performance. Shaded background regions denote fitness categories, and each point shows the Spearman correlation of one method on one dataset. **(c)** Relative improvement of PsiFit over ESM-1v within each fitness category.

To further quantify the improvement over zero-shot pLM prediction, we computed the relative improvement of PsiFit over ESM-1v for each dataset. Fig. 2c summarizes these improvements by fitness category. Across all matched datasets, the median relative improvement over ESM-1v was 7.8%, supporting the overall benefit of stability-informed adaptation. The category-level view shows positive median improvement across the major fitness categories, with particularly visible gains in Expression and Binding datasets and a broader range of improvements in Organismal Fitness datasets. The distribution also shows that PsiFit's benefit is heterogeneous: some datasets exhibit large gains, whereas others show little or negative improvement relative to ESM-1v. This pattern is expected because protein fitness reflects multiple biophysical and functional constraints, including folding stability, molecular binding, catalytic mechanism, expression, and cellular selection. Overall, these results support stability prediction as a broadly useful inductive bias for low-N protein fitness learning, while also highlighting that its contribution depends on the biological property measured by each assay.

4. Discussion

PsiFit introduces a stability-informed strategy for adapting protein language models to sparse fitness data. Its central idea is to use $\Delta\Delta G$ predictions from a multimodal sequence-structure foundation model as a biophysical prior that reshapes pLM likelihoods during contrastive fine-tuning. This approach preserves the data efficiency of ConFit (Zhao et al., 2024) while adding an explicit mechanistic constraint: mutations that strongly disrupt structural stability should be penalized unless the fitness data suggest otherwise.

More broadly, PsiFit illustrates how multimodal foundation models can support protein engineering beyond direct prediction. SPURS integrates sequence and structure modalities to produce scalable stability predictions; PsiFit then uses these predictions as a transferable prior for learning diverse fitness readouts. This separation between a general biophysical prior and a task-specific adaptation objective may be useful for other biological foundation model applications, where labeled data are scarce but mechanistic auxiliary signals are available.

The current framework has several limitations. First, stability is only one determinant of protein fitness, and its relevance varies by assay and protein family. Second, SPURS (Li & Luo, 2025) predictions may contain systematic errors for proteins with uncertain structures or for mutations involving complex conformational effects. Third, the present formulation focuses primarily on additive single-site stability priors, although higher-order epistasis can be important for multi-mutant fitness prediction. Future work will extend PsiFit to incorporate epistatic stability predictions and additional modalities, such as protein dynamics, binding-site annotations, or experimental uncertainty.

Impact Statement

This work aims to improve data-efficient protein fitness prediction for machine learning-guided protein engineering and variant interpretation. More accurate low-N models may reduce experimental burden by helping prioritize candidate variants for laboratory screening. As with other predictive models for biological design, downstream use should include experimental validation and careful consideration of application-specific risks.

References

Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M., and Church, G. M. Low-n protein engineering with data-efficient deep learning. *Nature methods*, 18(4):389–396, 2021.

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I., Courbet, A., de Haas, R. J., Bethel, N., et al. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378 (6615):49–56, 2022.

Freschlin, C. R., Fahlberg, S. A., and Romero, P. A. Machine learning to navigate fitness landscapes for protein engineering. *Current opinion in biotechnology*, 75: 102713, 2022.

Hsu, C., Nisonoff, H., Fannjiang, C., and Listgarten, J. Learning protein fitness models from evolutionary and assay-labeled data. *Nature biotechnology*, 40(7):1114–1122, 2022.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Kouba, P., Kohout, P., Haddadi, F., Bushuiev, A., Samusevich, R., Sedlar, J., Damborsky, J., Pluskal, T., Sivic, J., and Mazurenko, S. Machine learning-guided protein engineering. *ACS catalysis*, 13(21):13863–13895, 2023.

Li, Z. and Luo, Y. Generalizable and scalable protein stability prediction with rewired protein generative models. *Nature Communications*, 2025.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023.

Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., and Rives, A. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in neural information processing systems*, 34:29287–29303, 2021.

Nijkamp, E., Ruffolo, J. A., Weinstein, E. N., Naik, N., and Madani, A. Progen2: exploring the boundaries of protein language models. *Cell systems*, 14(11):968–978, 2023.

Notin, P., Kollasch, A., Ritter, D., Van Niekerk, L., Paul, S., Spinner, H., Rollins, N., Shaw, A., Orenbuch, R., Weitzman, R., et al. Proteingym: Large-scale benchmarks for protein fitness prediction and design. *Advances in neural information processing systems*, 36:64331–64379, 2023.

Wittmann, B. J., Johnston, K. E., Wu, Z., and Arnold, F. H. Advances in machine learning for directed evolution. *Current opinion in structural biology*, 69:11–18, 2021.

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

Zhao, J., Zhang, C., and Luo, Y. Contrastive fitness learning: Reprogramming protein language models for low-n learning of protein fitness landscape. In *International Conference on Research in Computational Molecular Biology*, pp. 470–474. Springer, 2024.