Learning Protein Fitness Landscapes with Multimodal Stability Priors

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

1 School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA, USA

Introduction & Motivation

Protein fitness prediction is central to ML-guided protein engineering, yet most DMS datasets are **low-N** (often <100 labeled variants per assay).

Pre-trained protein language models (pLMs) capture rich evolutionary knowledge but risk overfitting when fine-tuned directly on sparse fitness labels.

ConFit (Zhao et al., 2024) addressed this with contrastive fine-tuning that calibrates pLM likelihoods to observed rankings.

SPURS (Li and Luo, 2025) a multimodal sequence-structure model can predict mutation-induced ΔΔG at scale.

PsiFit extends ConFit by injecting a **multimodal stability prior** from SPURS, a sequence + structure foundation model predicting mutation-induced ΔΔG directly into the some last layer scores as a stability bias

Problem Formulation

Given: Small labeled set D = $\{(x^{(k)}, y^{(k)})\}_{k=1}^{N}$ ($N \ll 100$)

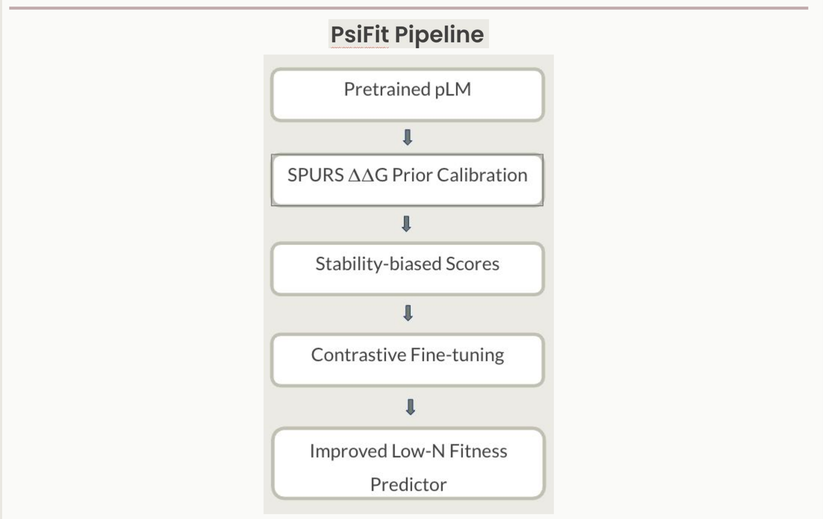
Goal: Learn a predictor that correctly ranks unseen variants by experimental fitness.

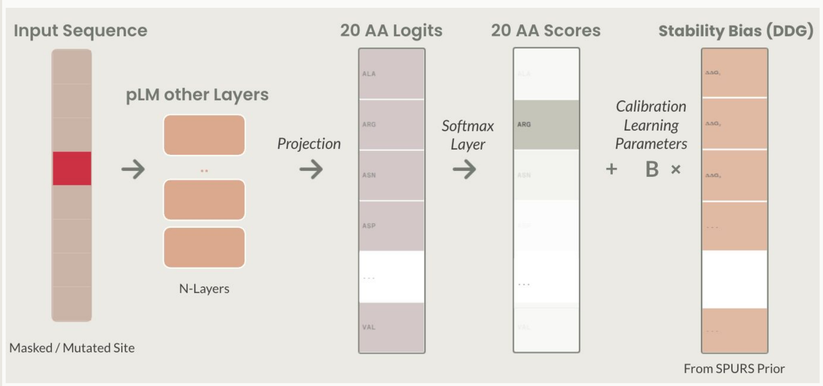
pLM Fitness Proxy (Zero-shot):

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

This score preserves the pLM output space while serving as a scalar proxy for variant fitness.
Direct fine-tuning on sparse labels risks overfitting: contrastive approaches (ConFit) improve ranking calibration.

PsiFit Architecture





SPURS Prior Calibration

PsiFit injects stability information directly into the
- final-layer scores
- second-last-layer logits

using a
- position-specific
- context-specific

stability bias:

$$X'_{\theta,\phi}(x) = X_\theta(x) + \mathrm{MLP}_\phi\big(G(x)\big)$$

Xθ: Original pLM Scores Matrix.
G(x): Predicted ΔΔG values from SPURS.
MLPϕ: Learned calibration module for stability effects.

Bradley–Terry Contrastive Loss (adapted from ConFit)

For every pair of labeled variants where y(i) > y(j), encourage the model to assign higher predicted score to the better variant:

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

This converts N labeled examples into O(N²) pairwise comparisons, greatly improving data efficiency in the low-N regime. Temperature τ controls the sharpness of the ranking signal.

KL Regularization & Implementation

To prevent the adapted model from drifting too far from the rich pre-trained distribution, add a KL term that penalizes divergence from the original pLM:

$$\mathcal{L}_{\mathrm{KL}} = \sum_{x \in \mathcal{B}} \sum_{i \in M(x)} \mathrm{D}_{\mathrm{KL}}\Big(p_{\theta_0}(\cdot \mid x_{-M}) \,\|\, p_{\theta,\phi}(\cdot \mid x_{-M})\Big)$$

Final objective: L = L_{BT} + λ_{KL} L_{KL}. Fine-tuning uses parameter-efficient LoRA adapters on the pLM backbone.

Results

Baselines

SPURS Augmented DeepSequence and Augmented DeepSequence (state-of-the-art supervised low-N method combining sequence features + evolutionary density).

Datasets

114 matched deep mutational scanning datasets spanning Expression, Activity, Binding, and Organismal Fitness (stability-measuring assays excluded to test generalization of the stability prior).

Low-N supervised regime

N = 240 labeled variants for training; performance measured by Spearman correlation on held-out variants (primary ranking metric

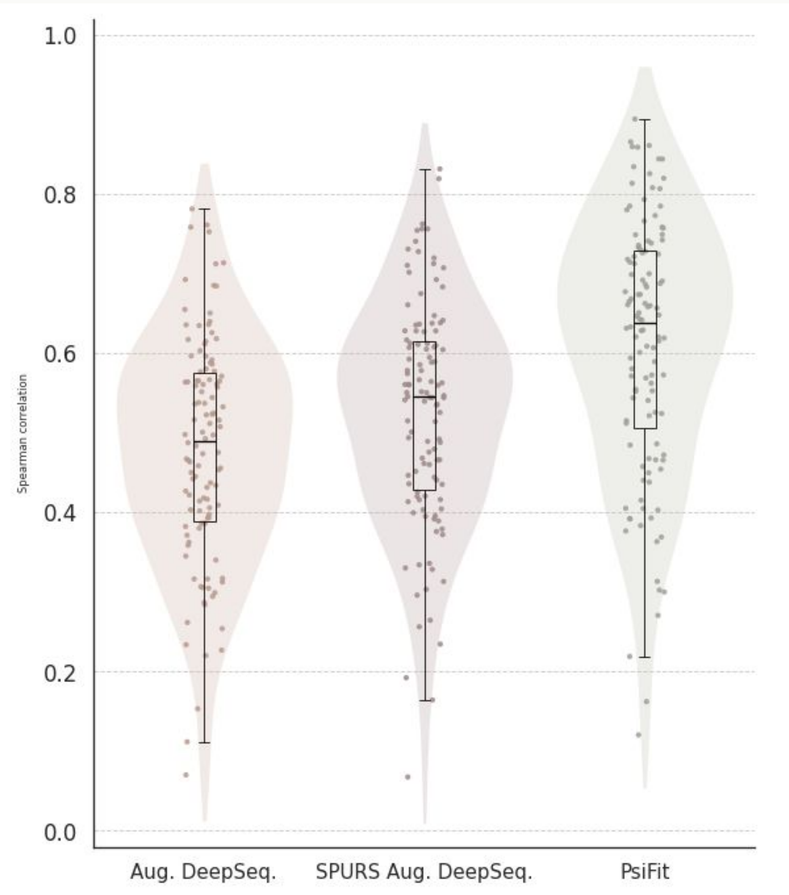


Figure 1. Spearman correlation between predicted and experimental fitness scores across 116 protein datasets. Each dot represents one dataset. Aug. DeepSeq.: DeepSequence-augmented ESM-1v; SPURS Aug. DeepSeq.: SPURS-corrected variant; PsiFit: Position-specific calibration (shot = 240).

Win rates (all 114 datasets):
PsiFit > SPURS Aug. Seq. : 103/114 = 90.4%
PsiFit > Aug. Seq. : 105/114 = 92.1%

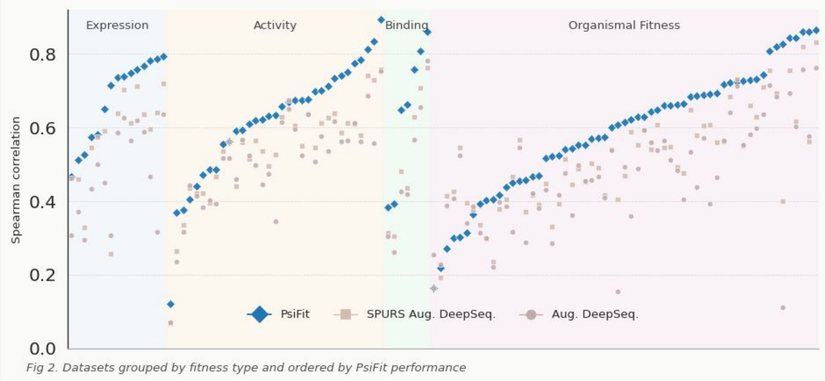


Fig 2. Datasets grouped by fitness type and ordered by PsiFit performance

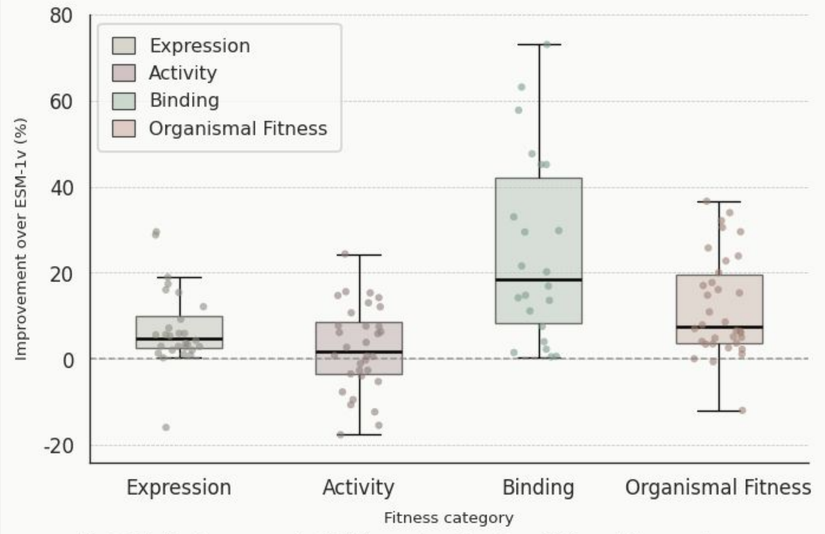


Fig 3. Relative Improvement of PsiFit over Aug. DeepSeq within each fitness category

Protein fitness reflects multiple biophysical constraints (stability, binding, catalysis, expression, cellular selection).

The stability prior helps most when it is a dominant factor or when zero-shot leaves clear room for supervised calibration.

PsiFit maintains competitive performance across all fitness categories, showing that the stability prior is not limited to stability-measuring assays.

Limitation and Future Improvement

- Stability is only one determinant of fitness; relevance varies by assay and protein family.
- SPURS predictions may contain systematic errors for proteins with uncertain structures or complex conformational effects.
- Current formulation focuses on additive single-site stability priors; higher-order epistasis can matter for multi-mutant prediction.
- Future work

extend to epistatic stability predictions
incorporate protein dynamics
binding-site annotations
experimental uncertainty estimates.

Conclusion

- Multimodal stability priors from SPURS can be effectively integrated into contrastive fine-tuning of protein language models.
- Low-N prediction performance across diverse fitness readouts while preserving the data efficiency is improved.
- This approach offers a general strategy for leveraging mechanistic biophysical signals when labeled experimental data are limited.

- Zhang et al. (2026). Learning Protein Fitness Landscapes with Multimodal Stability Priors. ICML Workshop FM4LS.
- Li & Luo (2025). Generalizable and scalable protein stability prediction with rewired protein generative models. bioRxiv.
- Zhao et al. (2024). Contrastive fitness learning: Reprogramming protein language models for low-n learning. RECOMB.
- Notin et al. (2023). ProteinGym: Large-scale benchmarks for protein fitness prediction and design. NeurIPS.
- Meier et al. (2021). Language models enable zero-shot prediction of mutation effects. NeurIPS.
- Hsu et al. (2022). Learning protein fitness models from evolutionary and assay-labeled data. Nat. Biotechnol.

Supported by IDEaS project GR00034047, Georgia Institute of Technology

Shannon Zhang • shannonz@gatech.edu