
StabilityArc: Decoding Protein Sequence Embeddings into Generalizable Stability Landscapes

Aaron L. Feller

The University of Texas at Austin
Interdisciplinary Life Sciences
aaron.feller@utexas.edu

Andrew D. Ellington

The University of Texas at Austin
Interdisciplinary Life Sciences
andy.ellington@austin.utexas.edu

Claus O. Wilke

The University of Texas at Austin
Integrative Biology
wilke@austin.utexas.edu

Abstract

Every protein has a unique stability landscape, but the physical consequences of mutation are governed by recurring biochemical constraints. We test whether a shared decoder, trained on measurements from diverse proteins, can interpret these constraints in an unseen target, enabling cross-protein transfer for initial experimental round prescreening. We present STABILITYARC, which maps frozen ESMC-600M residue representations through a shared RoPE transformer to an $L \times 20$ matrix of substitution effects; a symmetric, contact-aware residual aids in predicting epistasis in simultaneous substitutions. In 66 strict leave-one-protein-out evaluations covering 134,794 ProteinGym variants, STABILITYARC achieves 0.7134 Spearman correlation, exceeding the strongest zero-shot baseline, ProSST-2048 (0.6526), by 0.0608. We further explore the utility of this method by providing the score as a prior for Kermut, achieving Spearman correlation of 0.8280 across three supervised split schemes, improving on Kermut’s reported 0.8167.

Introduction

The majority of proteins exist in a carefully folded organization that is both unique and necessary to confer the specific function of that protein [Martin et al., 1998]. When engineering a protein, our goal is to modify the sequence of that protein to enhance protein function, while avoiding disruption of the often fragile form that is required for activity [Listov et al., 2024]. Often, we do not have available data, and the evaluation of that protein is very low throughput, requiring high cost and time. However, there are many proteins that have been evaluated for stability in high throughput manner. Recent large-scale measurements from deep mutational scanning (DMS) have been used to map mutations to their effect on protein’s stability [Fowler and Fields, 2014, Riesselman et al., 2018, Tsuboyama et al., 2023].

A stability landscape is highly local to a sequence and its folded state: the same substitution can be benign in one context and damaging in another [Karagöl et al., 2025]. Yet the physical mechanisms that penalize mutations, including packing frustration, unsatisfied charge, and solvent exposure, recur universally [Otwinowski, 2018]. If a pre-trained embedding provides the structural and biochemical prior for the wild-type environment, a shared decoder, trained across many proteins, may be able to map amino acid representations to the impact of substitutions on stability shifts. The evaluation should be simple: can a single decoder trained across multiple protein stability landscapes transfer to an out-of-distribution protein?

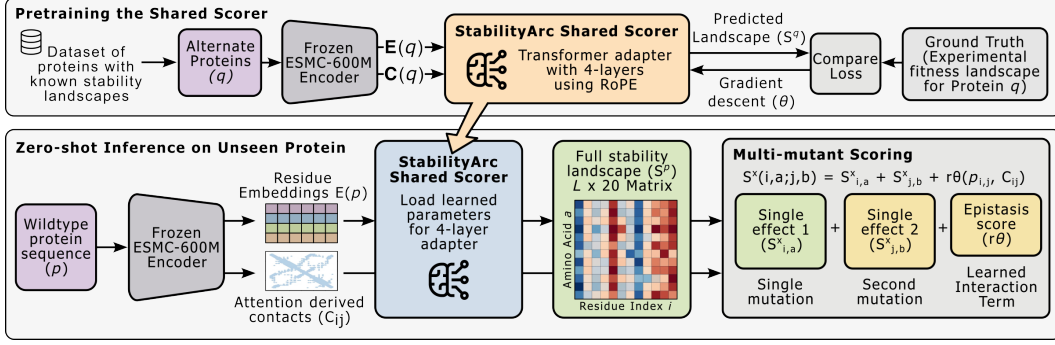


Figure 1: Cross-protein learning and deployment of STABILITYARC. During training (top), alternate proteins (q) are passed through a frozen ESMC-600M encoder to train the shared 4-layer RoPE transformer adapter against ground-truth experimental fitness landscapes. All labels from the target protein are strictly withheld. At deployment (bottom), the target wild-type sequence (p) is encoded to provide structural representations and contact maps, which the learned adapter decodes into a complete $L \times 20$ single-substitution stability landscape. For simultaneous substitutions, a multi-mutant scoring module (right) combines these single effects with a learned epistasis interaction term using attention-derived contacts.

STABILITYARC (Figure 1) explores this potential; frozen ESMC-600M embeddings and attention-derived contacts provide a label-free description of the target wild type [Candido et al., 2026, Rao et al., 2021]. A shared four-layer transformer decodes that description into effects for all 20 substitutions at every site, and a symmetric contact-aware residual accounts for epistasis in double mutants. During held-out evaluation, every DMS label from the target protein is excluded; the target supplies sequence-derived features only.

We evaluate the resulting landscape at two points in an engineering cycle. Before target measurements, we use STABILITYARC to rank variants under strict protein holdout. After measurements arrive, the score from STABILITYARC is used as a prior in Kermut [Groth et al., 2024], a Gaussian-process (GP) model that revises the landscape using target-assay data, structure, and variant embeddings. Ultimately, we demonstrate that a robust cross-protein prior can both succeed at predicting mutation effects in a blind manner and can seamlessly integrate with target-specific optimization once experimental data becomes available.

Method

Shared wild-type decoder. Let $x = (x_1, \dots, x_L)$ be a wild type sequence and let $E(x) \in \mathbb{R}^{L \times d}$ be its frozen ESMC-600M residue embeddings. We train a learned projection using four pre-norm transformer layers [Vaswani et al., 2017] with rotary position embeddings [Su et al., 2024] and eight attention heads to produce residue states $H \in \mathbb{R}^{L \times 256}$. A linear head applied to the embedding at a given position gives:

$$S^x = f_\theta(E(x)) \in \mathbb{R}^{L \times 20}, \quad (1)$$

where $S^x_{i,a}$ is the predicted effect of substituting amino acid a at site i . Parameters θ are shared across proteins.

Multiple mutations with contact-aware composition. To predict the effect of a double mutation, the model sums the individual substitution scores (S_1 and S_2) and adds a learned interaction term (r):

$$\hat{y} = S_1 + S_2 + r. \quad (2)$$

The term r (further explained in Appendix A.3) calculates how the two mutations influence each other based on their sequence distance and their likelihood of physical contact in the folded structure. Following prior work showing that protein language models capture structural information [Rao et al., 2021], we extract these contact probabilities directly from the frozen ESMC attention maps rather than relying on experimental coordinates.

Table 1: Macro Spearman on 66 proteins assayed for stability. All comparator model values are from the ProteinGym score export; overall score concatenates singles and doubles within each assay.

Method	Singles	Doubles	Overall
ESM3	0.6460	0.3262	0.6406
VenusREM	0.6370	0.3803	0.6503
ProSST-4096	0.6243	0.4100	0.6522
ProSST-2048	0.6385	0.3801	0.6526
ESM-C-600M (frozen prior)	0.4887	0.2355	0.5277
STABILITYARC (ours)	0.6946	0.4119	0.7134

Learning objective. Training combines Huber loss with a pairwise ranking loss. For each ranking-loss evaluation, we select a batch of 1,024 random pairs of variants and train their predicted ordering to match their measured ordering. Single-mutant labels are used directly. Double-mutant labels are centered and scaled within each assay using the median and median absolute deviation of that assay’s single-mutant labels.

Supervised prior. The Gaussian process in Kermut builds its baseline prediction by taking a predicted score as a prior and adjusting it with a simple linear equation [Groth et al., 2024]. It pairs this with a covariance function that combines structural mutation similarity and an RBF kernel over ESM-2 variant embeddings. For a target protein $p \notin P$, we replace only its mean input:

$$m(v) = b + \beta s_{\text{STABILITYARC}}^p(v), \quad (3)$$

where $s_{\text{STABILITYARC}}^p$ is the predicted stability for the target protein, generated by the model trained on the remaining dataset. The GP still learns from the four allowed training folds of the target assay, separating cross-protein transfer in the prior from within-assay posterior adaptation.

Experimental Setup

Data and hold-one-protein-out protocol. We use 66 ProteinGym stability assays containing 134,794 variants: 68,855 singles and 65,939 doubles. We train 66 outer models, each excluding all labels for one protein. The held-out wild-type sequence and frozen contacts are available at inference. Singles and doubles are concatenated within each assay, one Spearman correlation is computed, and the unweighted mean over assays is reported. Every model is trained for 40 epochs with AdamW, learning rate 10^{-4} , weight decay 10^{-3} , and up to 40,000 measured singles and doubles per training assay.

Comparisons. For cross-protein comparison, we include methods with complete score coverage for the same variants in all 66 assays and report pooled, single-only, and double-only macro Spearman. The inputs for various models differ: e.g. ProSST uses AlphaFold2-derived structure tokens, whereas STABILITYARC uses ESMC-600M embeddings from the wild-type sequence [Candido et al., 2026]. For supervised adaptation, we follow ProteinGym’s [Notin et al., 2024] official five-fold random, modulo, and contiguous splits, yielding $66 \times 3 \times 5 = 990$ GP fits. We calculate each assay’s Spearman ρ across all available measurements.

Results

Cross-protein transfer. For prediction of protein stability, each assay is scored by a model trained without that protein, so the result measures cross-protein transfer rather than interpolation from its labels. In this schema, STABILITYARC reaches 0.7134 all-variant macro Spearman (Table 1), 0.0608 above the leading model, ProSST-2048. STABILITYARC also leads the single- and double-mutant ranking individually, where again, ProSST-2048 previously held the strongest overall baseline.

In a second cross-protein evaluation on the dataset T2837 [Ouyang-Zhang et al., 2023], STABILITYARC improves recall and accuracy (Supplementary Table 3) but shows decreased Spearman, AUROC, and precision when compared to StabilityOracle [Diaz et al., 2024].

Table 2: Supervised stability benchmark. Kermut values are reported by Groth et al. [2024].

Method	Random	Modulo	Contig.	Mean
Kermut (reported)	0.9046	0.7871	0.7583	0.8167
Kermut + STABILITYARC	0.9094	0.7984	0.7763	0.8280
Gain	+0.0048	+0.0113	+0.0180	+0.0113

Supervised adaptation. Replacing only Kermut’s prior input improves every official split (Table 2). The overall mean rises from the reported 0.8167 to 0.8280. The gain is smallest under random splitting and largest under contiguous splitting, consistent with a transferable prior helping most across structured gaps in sequence position. Kermut still contributes target-assay labels, structural features, variant embeddings, and posterior uncertainty; the result shows that the protein-held-out prediction from STABILITYARC remains informative beyond the other incorporated embeddings.

Discussion

Sequence-based stability prediction. Historically, computational stability prediction has relied heavily on structure-based energy functions such as Rosetta [Alford et al., 2017] or FoldX [Schymkowitz et al., 2005] as well as newer machine learning methods [Diaz et al., 2024]. While these methods are highly effective, they are inherently limited to targets with high-resolution coordinates or highly accurate structural models. STABILITYARC bypasses this bottleneck by operating purely on sequence. Although structure-free approaches face distinct biophysical challenges, our results demonstrate highly competitive performance without requiring input coordinates. This capability underscores the depth of physical and evolutionary information captured directly by the underlying ESMC embeddings.

Decoding universal stability constraints. The central challenge of zero-shot protein engineering is that every wild-type stability landscape is unique [Tsuboyama et al., 2023]. STABILITYARC demonstrates that while stability landscapes vary, they can be accurately predicted by applying a single decoding rule to rich evolutionary representations. By training a shared transformer on diverse cross-protein measurements, we force the model to capture physical regularities that persist across distinct folds and lengths. This approach yields a significant improvement over purely evolutionary baselines in zero-shot settings [Notin et al., 2022] and provides a highly effective prior for supervised adaptation. The success of this architecture confirms that empirical stability data contains a universal, task-specific signal absent from structural tokens alone.

Isolating the stability signal. While evolutionary likelihood provides a strong zero-shot baseline, it inherently conflates stability with other selective pressures, such as binding or catalytic activity [Meier et al., 2021]. By contrast, STABILITYARC isolates the physical constraints of folding. The performance gap between our model and ProSST [Li et al., 2024] highlights the limits of relying solely on structural tokens or evolutionary plausibility. Because the target sequence is entirely excluded from training, the decoder cannot memorize specific mutational pathways. It must instead construct a purely biophysical landscape by interpreting the wild-type coordinates provided by ESMC.

Limitations. Our evaluation relies on retrospective benchmarking, one random seed per outer fold, and is evaluated only on single and double substitutions due to data availability. While strict protein-level holdout prevents DMS label leakage, it is impossible to exclude the target sequences from appearing in the unsupervised ESMC pretraining corpus. Finally, because protein language models perform best on families well-represented in their pretraining data, our model may face challenges predicting stability for singletons or orphan proteins.

Conclusion

Our contribution is a unified representation evaluated across two distinct information regimes. Before any experimental budget is spent, the shared decoder produces a complete wild-type-relative landscape under strict protein holdout, establishing a new state of the art for matched zero-shot stability

prediction. As empirical measurements arrive, this exact same landscape accelerates target-specific optimization, surpassing reported Kermut benchmarks without altering base kernels or data splits. By extracting generalized rules of protein stability, STABILITYARC translates these patterns into actionable priors for any new protein engineering campaign.

Reproducibility

The code, data, and models to reproduce our results, and evaluate your protein of interest are all available at <https://github.com/AaronFeller/StabilityArc>. We also provide a huggingface model that can be found at <https://huggingface.co/aaronfeller/StabilityArc>.

Acknowledgements

A.L.F was supported by the Charles W. Smith Jr. Graduate Fellowship in Artificial Intelligence and Machine Learning for Biosciences. A.D.E. is supported by the Welch Foundation (F-1654). C.O.W. acknowledges support from the Blumberg Centennial Professorship in Molecular Evolution.

Author contributions

A.L.F contributed to the study conception, design, and execution, as well as writing of the manuscript. All authors contributed to editing and revising the manuscript.

Declaration of interests

C.O.W. is a shareholder of and consultant for Proteolyze Therapeutics, Inc. The other authors declare no competing interests.

References

- Rebecca F Alford, Andrew Leaver-Fay, Jeliasko R Jeliaskov, Matthew J O’Meara, Frank P DiMaio, Hahnbeom Park, Maxim V Shapovalov, P Douglas Renfrew, Vikram K Mulligan, Kalli Kappel, et al. The rosetta all-atom energy function for macromolecular modeling and design. *Journal of chemical theory and computation*, 13(6):3031, 2017.
- Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, et al. Language modeling materializes a world model of protein biology. *bioRxiv*, June 2026. doi: 10.64898/2026.06.03.729735. URL <https://www.biorxiv.org/content/10.64898/2026.06.03.729735v1.full>.
- Daniel J Diaz, Chengyue Gong, Jeffrey Ouyang-Zhang, James M Loy, Jordan Wells, David Yang, Andrew D Ellington, Alexandros G Dimakis, and Adam R Klivans. Stability oracle: a structure-based graph-transformer framework for identifying stabilizing mutations. *Nature Communications*, 15(1):6170, 2024.
- David M. Fowler and Stanley Fields. Deep mutational scanning: a new style of protein science. *Nature Methods*, 11(8):801–807, 2014.
- Peter Mørch Groth, Mads Herbert Kerrn, Lars Olsen, Jesper Salomon, and Wouter Boomsma. Kermut: Composite kernel regression for protein variant effects. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0929.
- Taner Karagöl, Alper Karagöl, and Shuguang Zhang. Co-evolution of alpha-helical transmembrane protein residues: large-scale variant profiling and complete mutational landscape of 2277 known pdb entries representing 504 unique human protein sequences. *Journal of Molecular Evolution*, 93(5):581–599, 2025.
- Mingchen Li, Yang Tan, Xinzhu Ma, Bozitao Zhong, Huiqun Yu, Ziyi Zhou, Wanli Ouyang, Bingxin Zhou, Pan Tan, and Liang Hong. ProSST: Protein language modeling with quantized structure and disentangled attention. In *Advances in Neural Information Processing Systems*, volume 37, 2024.
- Dina Listov, Casper A Goverde, Bruno E Correia, and Sarel Jacob Fleishman. Opportunities and challenges in design and optimization of protein function. *Nature Reviews Molecular Cell Biology*, 25(8):639–653, 2024.
- Andrew CR Martin, Christine A Orengo, E Gail Hutchinson, Susan Jones, Maria Karmirantzou, Roman A Laskowski, John BO Mitchell, Chiara Taroni, and Janet M Thornton. Protein folds and functions. *Structure*, 6(7):875–884, 1998.
- Joshua Meier et al. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34, 2021.
- Pascal Notin, Mafalda Dias, Jonathan Frazer, Javier Marchena-Hurtado, Aidan N Gomez, Debora Marks, and Yarin Gal. Tranception: protein fitness prediction with autoregressive transformers and inference-time retrieval. In *International Conference on Machine Learning*, pages 16990–17017. PMLR, 2022.
- Pascal Notin et al. Proteingym: Large-scale benchmarks for protein fitness prediction and design. *Advances in Neural Information Processing Systems*, 36, 2024.
- Jakub Otwinowski. Biophysical inference of epistasis and the effects of mutations on protein stability and function. *Molecular Biology and Evolution*, 35(10):2345–2354, 2018.
- Jeffrey Ouyang-Zhang, Daniel Diaz, Adam Klivans, and Philipp Krähenbühl. Predicting a protein’s stability under a million mutations. *Advances in Neural Information Processing Systems*, 36: 76229–76247, 2023.
- Roshan M. Rao et al. Transformer protein language models are unsupervised structure learners. In *International Conference on Learning Representations*, 2021.
- Adam J. Riesselman, John B. Ingraham, and Debora S. Marks. Deep generative models of genetic variation capture the effects of mutations. *Nature Methods*, 15(10):816–822, 2018.

- Joost Schymkowitz, Joost Borg, Francois Stricher, Robyn Nys, Frédéric Rousseau, and Luis Serrano. The foldx web server: an online force field. *Nucleic Acids Research*, 33:W382–W388, 2005. doi: 10.1093/nar/gki387.
- Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.
- Kotaro Tsuboyama, Justas Dauparas, Jue Chen, Elodie Laine, et al. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature*, 620:434–444, 2023.
- Ashish Vaswani et al. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

A Appendix

A.1 Comparison to StabilityOracle on T2837

We compared STABILITYARC to StabilityOracle on a matched subset of the T2837 dataset, comprising 1,884 shared mutations across 19 targets. Using a standard destabilization threshold of $\Delta\Delta G \geq 1$ kcal/mol, STABILITYARC achieves higher recall (0.527 vs. 0.376) and accuracy (0.705 vs. 0.669). Because failing to identify a destabilizing mutation can compromise downstream engineering, this improved recall offers a distinct safety advantage during conservative pre-screening. However, StabilityOracle retains a stronger pooled Spearman correlation (0.718 vs. 0.606) and AUROC (0.852 vs. 0.804), making it highly effective for continuous ranking of candidate mutations.

These results are not an artifact of class imbalance, as the matched cohort is evenly split (962 destabilizing, 922 non-destabilizing).

Table 3: Paired pooled T2837 comparison on 1,884 shared raw PDB-chain mutations from 19 targets with more than 30 shared mutations. Δ is STABILITYARC minus StabilityOracle; both use $\Delta\Delta G \geq 1$ kcal/mol as the positive class.

Method	Spearman	AUROC	Recall	Precision	Accuracy
StabilityOracle (supplied)	0.718	0.852	0.376	0.938	0.669
STABILITYARC (ours)	0.606	0.804	0.527	0.835	0.705
Δ (STABILITYARC – StabilityOracle)	−0.112	−0.048	+0.151	−0.103	+0.037

A.2 Mutation-Order Decomposition

The main text reports the all-variant metric because it evaluates the complete measured landscape. For completeness, the strict held-out predictions achieve 0.6946 macro Spearman on singles and 0.4119 on doubles; the latter is averaged over assays with a defined double-mutant correlation.

Table 4: Zero-shot STABILITYARC performance by mutation order.

Metric	Singles	Doubles	All variants
Macro Spearman	0.6946	0.4119	0.7134

A.3 Multiple mutations with contact-aware composition.

For a double mutant (i, a) and (j, b) , first form mutation-aware site vectors:

$$u_i = h_i + e_a, \quad u_j = h_j + e_b \quad (4)$$

where $h_i, h_j \in \mathbb{R}^{256}$ are the decoder’s contextual site states and $e_a, e_b \in \mathbb{R}^{256}$ are learned embeddings of the substituted amino acids.

The residual’s input is then:

$$z_{ij} = \left[u_i + u_j, |u_i - u_j|, u_i \odot u_j, \frac{|i - j|}{L - 1}, C_{ij} \right] \quad (5)$$

This comprises 770 values in total (3×256 residue-pair vectors + 2 scalar features):

- $u_i + u_j$: what the two edited residues have in common.
- $|u_i - u_j|$: how they differ, without depending on edit order.
- $u_i \odot u_j$: feature interactions between the two residue contexts.
- $\frac{|i - j|}{L - 1}$: normalized sequence separation (scalar).
- C_{ij} : the frozen ESMC attention-derived contact feature (scalar).

The residual function r_θ processes this 770-dimensional input through a two-layer network to output a single scalar epistasis correction:

$$r_\theta(z_{ij}) = \text{Linear}_{256 \rightarrow 1}(\text{GELU}(\text{Linear}_{770 \rightarrow 256}(z_{ij}))) \quad (6)$$

The output modifies the additive score:

$$\hat{y} = S_{i,a}^x + S_{j,b}^x + r_\theta(z_{ij}) \quad (7)$$

So if $r_\theta = 0$, the model treats the double mutant as additive. A nonzero residual captures epistasis: whether the combined effect is more or less harmful than the sum of the two single substitutions.

It is symmetric by construction: swapping (i, a) and (j, b) leaves every input feature unchanged.

B Broader Impact

Improved mutation-effect prediction could prioritize proteins for experimental characterization and help engineers avoid destabilizing substitutions. It could also encourage overconfidence in unvalidated variants or facilitate optimization of biologically active proteins. Predictions should therefore be treated as ranked hypotheses, accompanied by uncertainty and assay-domain information, and validated experimentally before clinical or other high-consequence use.

C NeurIPS Paper Checklist

Claims. The paper distinguishes protein-held-out evaluation from within-assay supervised evaluation and identifies the Kermut comparison as one against reported benchmark values.

Reproducibility. Dataset coverage, folds, model hyperparameters, metrics, and baseline scope are specified. The main text points to a github repository and huggingface-hosted model.

Compute. The 66 held-out models were trained for 40 epochs on 4x NVIDIA RTX 6000 Ada Generation; saved artifacts span 3.4 wall-clock hours, or approximately 13.6 GPU-hours. Exporting priors and fitting the 198 Kermut assay/split jobs added about 0.4 wall-clock hours, for an estimated total of 15 GPU-hours. Energy was not metered; assuming 250–400 W per active GPU, this corresponds to roughly 4–6 kWh. We froze ESMC, reused cached embeddings and contact maps, used a four-layer 256-dimensional decoder, and capped each assay at 40,000 measured variants. We did not run a broad hyperparameter sweep.