GREEN FLUORESCENT PROTEIN ENGINEERING WITH A BIOPHYSICS-BASED PROTEIN LANGUAGE MODEL

Sam Gelman & Bryce Johnson
Department of Computer Sciences, University of Wisconsin-Madison
Morgridge Institute for Research
{sgelman2,bcjohnson7}@wisc.edu

Chase Freschlin & Sameer D'Costa
Department of Biochemistry, University of Wisconsin-Madison
{freschlin,dcosta2}@wisc.edu

Anthony Gitter *
Department of Computer Sciences, University of Wisconsin-Madison
Morgridge Institute for Research
Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison
gitter@biostat.wisc.edu

Philip A. Romero *
Department of Biochemistry, University of Wisconsin-Madison
promero2@wisc.edu

Deep neural networks and language models are revolutionizing protein modeling and design (Bepler & Berger, 2021), but these models struggle in low data settings and when generalizing beyond their training data. Although prior neural networks have proven capable in learning complex evolutionary or sequence-structure-function relationships from large datasets, they largely ignore the vast accumulated knowledge of protein biophysics, limiting their ability to perform the strong generalization required for protein engineering. We introduce Mutational Effect Transfer Learning (METL), a specialized protein language model for predicting quantitative protein function that bridges the gap between traditional biophysics-based and machine learning approaches. METL incorporates synthetic data from molecular simulations as a means to augment experimental data with biophysical information. Molecular modeling can generate large datasets revealing mappings from amino acid sequences to protein structure and properties. Pretraining protein language models on this data can impart fundamental biophysical knowledge that can be connected with experimental observations.

METL operates in three steps: synthetic data generation, synthetic data pretraining, and experimental data finetuning. First, we generate synthetic pretraining data via molecular modeling with Rosetta (Alford et al., 2017) to model the structures of millions of protein sequence variants. For each modeled structure, we extract 55 biophysical attributes including molecular surface areas, solvation energies, van der Waals interactions, and hydrogen bonding. Second, we pretrain a transformer encoder (Vaswani et al., 2017) to learn relationships between amino acid sequences and these biophysical attributes and to form an internal representation of protein sequences based on their underlying biophysics. The transformer utilizes a protein structure-based relative positional embedding (Shaw et al., 2018) that considers the three-dimensional distances between residues. Finally, we finetune the pretrained transformer encoder on limited experimental sequence-function data to produce a model that integrates prior biophysical knowledge with experimental data. Finetuned models predict specific quantitative functions assayed in experimental datasets such as binding, thermostability, and expression. The complete METL model and evaluations are presented in Gelman et al. (2024).

To demonstrate METL's ability to guide protein engineering with limited training data, we applied it to design green fluorescent protein (GFP) sequence variants in complex scenarios (Gelman et al., 2024). We used a version of METL, referred to as METL-Local, pretrained on 20M synthetic GFP variants and their corresponding Rosetta scores. This model was finetuned to predict GFP brightness

*These authors contributed equally to this work.

using only 64 variants randomly sampled from an experimental dataset (Sarkisyan et al., 2016). We tested two design scenarios, *Observed* and *Unobserved*, whereby designed variants were constrained to either include or exclude amino acid substitutions found in the training set, respectively. The designed variants contained 5 or 10 amino acid substitutions from wild-type GFP, which is more than the average 3.9 substitutions in the training data.

Within these strict experimental constraints, we used METL-Local, simulated annealing, and clustering to design 20 GFP sequences that were not part of the original dataset and experimentally validated the resulting GFP variants to measure their relative brightness (Fig. 1). Of the 20 designed sequences, 16 exhibited fluorescence. The hit rate was 40% (2/5) in the most restrictive design scenario, Unobserved with 10 amino acid substitutions, and it increased to 80% (4/5) with 5 substitutions. In the Observed design scenario, the hit rate was 100% (10/10). While several of the designed sequences matched or exceeded the relative brightness of the best training set variant, none surpassed that of the wild type. Despite this, six variants exhibited greater GFP fluorescence than the wild type. We theorize improved stability compensates for reduced brightness such that the total GFP fluorescence for these variants is greater than wild type GFP.

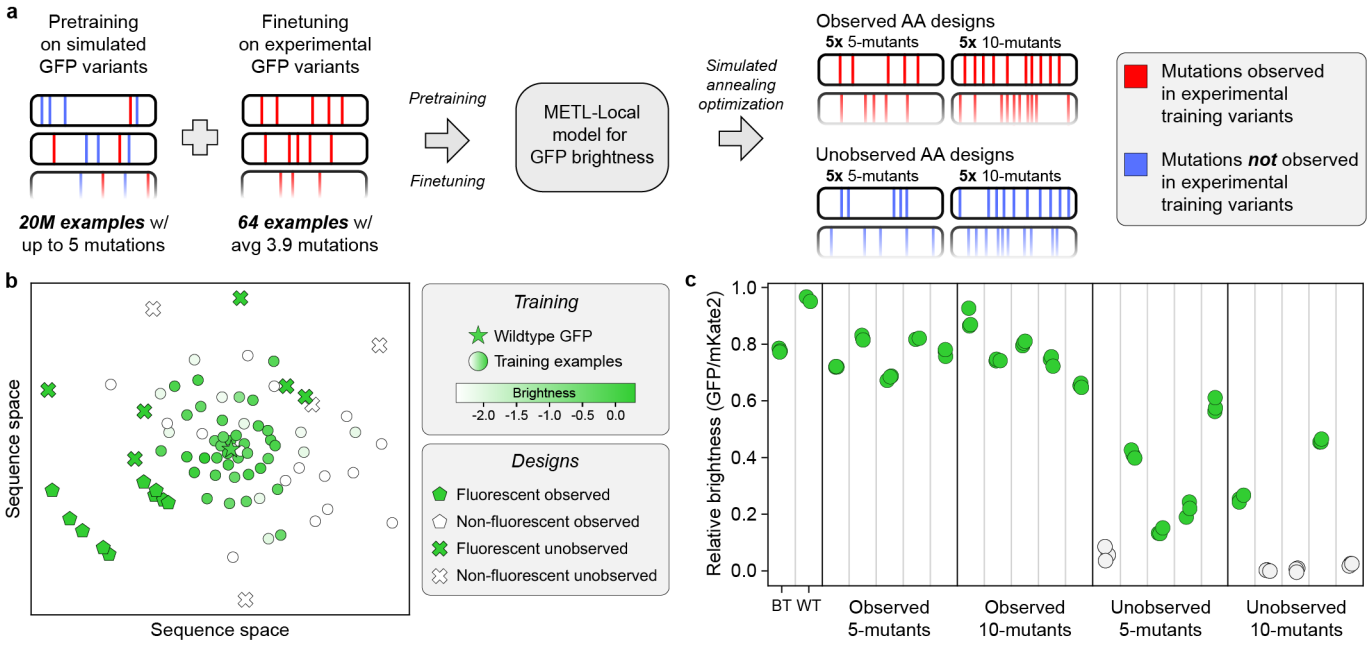

Figure 1: **Engineering GFP variants in challenging design scenarios with METL. (a)** Overview of the GFP design experiment. **(b)** Multidimensional scaling sequence space visualization of the wild type GFP sequence, the 64 GFP training sequences, and the 20 designed proteins. Training set sequences are colored on a gradient according to their experimental brightness score. Designed sequences are colored according to whether they exhibited fluorescence. **(c)** Experimentally characterized brightness (multiple replicates) of the designed sequences, the best training set sequence (BT), and the wild-type sequence (WT).

METL fits within the broader trend of combining simulations and machine learning (Cranmer et al., 2020), and it represents a significant step toward effectively integrating biophysics insights with machine learning-based protein fitness prediction. The METL framework pretrains protein language models on molecular simulations, capturing underlying signals present in the simulated data. METL can pretrain on general stability terms or more specific function-related scores, offering the potential to model protein functions that can be simulated but are not highly evolutionarily constrained. As the field of biophysics and molecular simulation continues to evolve, METL stands to benefit from faster and more accurate simulations. Biophysics-based pretraining can help overcome key challenges in protein engineering, such as prioritizing protein variants for experimental analysis with limited training data. Consequently, METL emerges as a promising tool for protein engineering with a distinct approach from the many existing methods rooted in evolutionary information.

REFERENCES

Rebecca F. Alford et al. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation*, 13(6):3031–3048, June 2017. ISSN 1549-9618. doi: 10.1021/acs.jctc.7b00125.

Tristan Bepler and Bonnie Berger. Learning the protein language: Evolution, structure, and function. *Cell Systems*, 12(6):654–669.e3, June 2021. ISSN 2405-4712. doi: 10.1016/j.cels.2021.05.017.

Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, December 2020. doi: 10.1073/pnas.1912789117.

Sam Gelman, Bryce Johnson, Chase Freschlin, Sameer D'Costa, Anthony Gitter, and Philip A. Romero. Biophysics-based protein language models for protein engineering. *bioRxiv*, pp. 2024.03.15.585128, March 2024. doi: 10.1101/2024.03.15.585128.

Karen S. Sarkisyan et al. Local fitness landscape of the green fluorescent protein. *Nature*, 533 (7603):397–401, May 2016. ISSN 1476-4687. doi: 10.1038/nature17995.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-Attention with Relative Position Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 464–468, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2074.

Ashish Vaswani et al. Attention Is All You Need. *arXiv:1706.03762 [cs]*, December 2017. doi: 10.48550/arXiv.1706.03762.