CS7643 Project Proposal: *Protein Engineering with Deep Learning using Evolutionary Scale Modeling (ESM)*

Amy Defnet, Chris Lindeman, Josh Meehl, Thet Hein Tun

June 10, 2024

1 Project Summary

In this project, we will use deep learning techniques for protein engineering tasks. Since 2018, there have been rapid advances in machine learning for biochemistry, such as AlphaFold2 for protein structure estimation [1].

Protein engineering is the modification of amino acid sequences to enhance various properties of a protein, including its thermal stability or enzymatic activity [2] It has conventionally relied on two approaches: *rational design* or *directed evolution*. In rational design, an expert, based on scientific understanding, models the protein and decides which amino acids to modify. This method is slow and error-prone. In directed evolution, many sequences are randomly mutated and characterized. The best performing variants are then selected and used as a starting point for the next round of mutations. The process is iterated until the desired properties are obtained [3]. This method is expensive and requires high-throughput processing of a large number of modification variations, many of which result in non-functional proteins.

Recently deep learning models have been applied to this problem of protein engineering. A family of models called *protein language models* (PLMs) have used representation learning approaches from natural language processing (NLP) to take amino acid sequence inputs and produce numeric representations. Our project will experiment with current state-of-the-art representation learning models applied to this domain.

2 Approach

We will deploy a family of transformer-based PLMs, specifically Evolutionary Scale Modeling (ESM), from the Facebook AI Research group (FAIR). Currently, FAIR has released ESM-1 [4] and ESM-2 models [5]. The code for the models is available here. We will use these trained models and code as a foundation for our work, and modify and augment them for our own experiments. A few examples of planned experiments are included below.

2.1 Representation Learning Architectures

The ESM models provide several architectures and scales to test on model performance. We will explore the following architectures:

- ESM-1 has three variants, ranging from 43 million to 670 million parameters, with 6 to 34 attention heads, respectively.
- ESM-1v for variant tasks has 5 essembled variants of ESM-1.
- ESM-2 has six variants, ranging from 35 million to 15 billion parameters, with 6 to 48 attention heads, respectively.

2.2 Zero-shot Learning

We will examine the effectiveness of zero-shot learning on the ESM models. The ESM-1v is an ensemble of five ESM-1 models trained to perform zero-shot learning tasks [6]. We will apply the zero-shot methods used with the ESM-1v models to the ESM-2 models and test for performance gains.

2.3 Few-shot Learning

We will design a head on the representation model to perform learned tasks. The goal is to find the fewest number of training examples to achieve a performance target. Various architectures of the head will be explored, such as a linear model and multi-level perceptrons.

2.4 Objective Functions

For protein engineering, the objective is to find top performing variants. We will explore objective functions that best identify top performers. Some example loss functions to be tested are:

- Mean square error on functional performance
- Cross-entropy loss on labels of "superior variants"
- Margin ranking loss on relative performance rank

2.5 Input Optimization

In protein engineering, the *fitness landscape* is the functional relationship between a protein's sequence and its functional value. We will explore methods for maximizing the fitness landscape, using the following optimization and search methods.

- Monte Carlo Markov chains (simulated annealing)
- Bayesian Optimization
- Optimization of the embedding space

2.6 Twin Neural Network

Often protein engineering is looking at a few point mutations in a sequence. These subtle changes in functional values can be challenging for supervised learning. We will look into methods that look at the difference between two inputs, such as twin neural networks [7].

3 Related Work

Various models have successful implementations of protein variant selection using protein language models (PLM). UniRep was the first LSTM model to show success, with results validated by engineering physical versions of top-performing protein sequences in the lab [8]. A later study demonstrated few-shot learning with UniRep and produced successful protein variants in wet lab validation [9]. Successive transformer-based models, such as TAPE, ESM-1b, ProtTrans, and ESM-2, have shown improved performance and have achieved state-of-the-art results [4], [5], [10], [11].

Several PLMs employed other biological context to improve performance. For example, ECNet used *multi-sequence alignments* (MSA) of related proteins to teach the model evolutionary context [12]. LM-GVP used graph neural networks have been adapted to allow language models to learn structural information of a protein [13].

ESMFold, a recent breakthrough with PLMs, uses a single sequence to predict the structure of a given protein [5]. This is in contrast to AlphaFold2 [1], which, in addition the sequence of the target protein, uses a MSA, residue pair representation, and template structures. ESMFold has been extended to generate novel proteins with natural language inputs from a user [14], [15].

Generative models, such as Generative adversarial networks (GAN), have been shown to be successful in generating novel proteins [16]. Encoder-decode models have been also been used to optimize protein variants in the latent space embedding [17].

4 Datasets

For this project, we will use data from *deep mutational scans* (DMS), which characterize the impact of changes on wild-type (WT) protein sequences. The dataset is composed of 44 DMS studies, [18]–[21], including about 25 proteins for organisms such as virus, bacteria, jellyfish, and human. The number of variants per study ranges from 1,000 to 500,000. The functional data has been log-normalized relative to the WT value.

References

[1] J. Jumper, R. Evans, A. Pritzel, *et al.*, "Highly accurate protein structure prediction with AlphaFold," eng, *Nature*, vol. 596, no. 7873, pp. 583–589, Jul. 2021, ISSN: 1476-4687. DOI: 10.1038/s41586-021-03819-2.

[2] K. M. Poluri and K. Gulati, *Protein Engineering Techniques* (SpringerBriefs in Applied Sciences and Technology). Singapore: Springer, 2017, ISBN: 978-981-10-2731-4 978-981-10-2732-1. DOI: 10.1007/978-981-10-2732-1. [Online]. Available: http://link.springer.com/10.1007/978-981-10-2732-1 (visited on 09/20/2022).

[3] F. H. Arnold, "Design by Directed Evolution," *Accounts of Chemical Research*, vol. 31, no. 3, pp. 125–131, Mar. 1998, Publisher: American Chemical Society, ISSN: 0001-4842. DOI: 10.1021/ar960017f. [Online]. Available: https://doi.org/10.1021/ar960017f (visited on 03/09/2023).

[4] A. Rives, J. Meier, T. Sercu, *et al.*, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," eng, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 118, no. 15, e2016239118, Apr. 2021, ISSN: 1091-6490. DOI: 10.1073/pnas.2016239118.

[5] Z. Lin, H. Akin, R. Rao, *et al.*, *Evolutionary-scale prediction of atomic level protein structure with a language model*, en, Pages: 2022.07.20.500902 Section: New Results, Oct. 2022. DOI: 10.1101/2022.07.20.500902. [Online]. Available: https://www.biorxiv.org/content/10.1101/2022.07.20.500902v2 (visited on 11/13/2022).

[6] J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, and A. Rives, *Language models enable zero-shot prediction of the effects of mutations on protein function*, en, Pages: 2021.07.09.450648 Section: New Results, Jul. 2021. DOI: 10.1101/2021.07.09.450648. [Online]. Available: https://www.biorxiv.org/content/10.1101/2021.07.09.450648v1 (visited on 11/03/2022).

[7] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a "Siamese" time delay neural network," in *Proceedings of the 6th International Conference on Neural Information Processing Systems*, ser. NIPS'93, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., Nov. 1993, pp. 737–744. (visited on 03/09/2023).

[8] E. C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, and G. M. Church, "Unified rational protein engineering with sequence-based deep representation learning," en, *Nature Methods*, vol. 16, no. 12, pp. 1315–1322, Oct. 2019, Number: 12 Publisher: Nature Publishing Group, ISSN: 1548-7105. DOI: 10.1038/s41592-019-0598-1. [Online]. Available: https://www.nature.com/articles/s41592-019-0598-1 (visited on 08/16/2022).

[9] S. Biswas, G. Khimulya, E. C. Alley, K. M. Esvelt, and G. M. Church, "Low-N protein engineering with data-efficient deep learning," eng, *Nature Methods*, vol. 18, no. 4, pp. 389–396, Apr. 2021, ISSN: 1548-7105. DOI: 10.1038/s41592-021-01100-y. (visited on 08/01/2022).

[10] R. Rao, N. Bhattacharya, N. Thomas, *et al.*, "Evaluating Protein Transfer Learning with TAPE," eng, *Advances in Neural Information Processing Systems*, vol. 32, pp. 9689–9701, Dec. 2019, ISSN: 1049-5258. DOI: 10.1101/676825.

[11] A. Elnaggar, M. Heinzinger, C. Dallago, *et al.*, *ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Learning*, en, Pages: 2020.07.12.199554 Section: New Results, May 2021. DOI: 10.1101/2020.07.12.199554. [Online]. Available: https://www.biorxiv.org/content/10.1101/2020.07.12.199554v3 (visited on 08/12/2022).

[12] Y. Luo, G. Jiang, T. Yu, *et al.*, "ECNet is an evolutionary context-integrated deep learning framework for protein engineering," en, *Nature Communications*, vol. 12, no. 1, p. 5743, Sep. 2021, Number: 1 Publisher: Nature Publishing Group, ISSN: 2041-1723. DOI: 10.1038/s41467-021-25976-8. [Online]. Available: https://www.nature.com/articles/s41467-021-25976-8 (visited on 08/16/2022).

[13] Z. Wang, S. A. Combs, R. Brand, *et al.*, "LM-GVP: An extensible sequence and structure informed deep learning framework for protein property prediction," en, *Scientific Reports*, vol. 12, no. 1, p. 6832, Apr. 2022, Number: 1 Publisher: Nature Publishing Group, ISSN: 2045-2322. DOI: 10.1038/s41598-022-10775-y. [Online]. Available: https://www.nature.com/articles/s41598-022-10775-y (visited on 08/12/2022).

[14] R. Verkuil, O. Kabeli, Y. Du, *et al.*, *Language models generalize beyond natural proteins*, en, Pages: 2022.12.21.521521 Section: New Results, Dec. 2022. DOI: 10.1101/2022.12.21.521521. [Online]. Available: https://www.biorxiv.org/content/10.1101/2022.12.21.521521v1 (visited on 01/07/2023).

[15] B. Hie, S. Candido, Z. Lin, *et al.*, *A high-level programming language for generative protein design*, en, Pages: 2022.12.21.521526 Section: New Results, Dec. 2022. DOI: 10.1101/2022.12.21.521526. [Online]. Available: https://www.biorxiv.org/content/10.1101/2022.12.21.521526v1 (visited on 01/07/2023).

[16] D. Repecka, V. Jauniskis, L. Karpus, *et al.*, "Expanding functional protein sequence spaces using generative adversarial networks," en, *Nature Machine Intelligence*, vol. 3, no. 4, pp. 324–333, Apr. 2021, Number: 4 Publisher: Nature Publishing Group, ISSN: 2522-5839. DOI: 10.1038/s42256-021-00310-5. [Online]. Available: https://www.nature.com/articles/s42256-021-00310-5 (visited on 08/01/2022).

[17] E. Castro, A. Godavarthi, J. Rubinfien, K. B. Givechian, D. Bhaskar, and S. Krishnaswamy, *ReLSO: A Transformer-based Model for Latent Space Optimization and Generation of Proteins*, arXiv:2201.09948 [cs], May 2022. DOI: 10.48550/arXiv.2201.09948. [Online]. Available: http://arxiv.org/abs/2201.09948 (visited on 08/17/2022).

[18] A. J. Riesselman, J. B. Ingraham, and D. S. Marks, "Deep generative models of genetic variation capture the effects of mutations," en, *Nature Methods*, vol. 15, no. 10, pp. 816–822, Sep. 2018, Number: 10 Publisher: Nature Publishing Group, ISSN: 1548-7105. DOI: 10.1038/s41592-018-0138-4. [Online]. Available: https://www.nature.com/articles/s41592-018-0138-4 (visited on 08/25/2022).

[19] K. S. Sarkisyan, D. A. Bolotin, M. V. Meer, *et al.*, "Local fitness landscape of the green fluorescent protein," en, *Nature*, vol. 533, no. 7603, pp. 397–401, May 2016, Number: 7603 Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: 10.1038/nature17995. [Online]. Available: https://www.nature.com/articles/nature17995 (visited on 08/11/2022).

[20] E. Firnberg, J. W. Labonte, J. J. Gray, and M. Ostermeier, "A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape," *Molecular Biology and Evolution*, vol. 31, no. 6, pp. 1581–1592, Feb. 2014, ISSN: 0737-4038. DOI: 10.1093/molbev/msu081. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4032126/ (visited on 10/20/2022).

[21] C. E. Gonzalez and M. Ostermeier, "Pervasive Pairwise Intragenic Epistasis among Sequential Mutations in TEM-1 -Lactamase," eng, *Journal of Molecular Biology*, vol. 431, no. 10, pp. 1981–1992, May 2019, ISSN: 1089-8638. DOI: 10.1016/j.jmb.2019.03.020.