ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

NCCR Catalysis

Automating Biology with ML: Guiding Generative Modelling for Improved Protein Design

Vignesh Ram Somnath Mojmír Mutný Andreas Krause

June 28, 2024

1 Motivation

Rational protein design is currently at the centerpoint of interest in academia as well as in industry. With the development of AlphaFold2 for approximate protein structure prediction (Jumper et al., 2021), there is renewed interest in designing novel proteins. The hope is to use these new proteins for theraupetic purposes (antibodies), biotechnology, diagnostics, cell reprogramming or chemical manufacturing.

LAS GROUP (Prof. Krause group) is part of large collaborative effort called *NCCR Catalysis* that aims to research and develop new catalysists for greener chemistry. We focus on enzymes, which are specific protein macromolecules, that facilitate chemical reactions by lowering energetic constraints for the reaction happen. The biochemistry field has seen big improvement by including machine learning in the enzyme design pipeline Patsch and Buller (2023); Vornholt et al. (2024).

2 Problem Statement

The overall problem is to elucidate the relationship between the sequence and enzymatic activity. This dependence is extremely elusive even for experts and even in the presence of protein structure. Our goal is come up with the most accurate description of this mapping given the experimental data we have at hand. In many instances, the data is limited to produce a highly-accurate predictive model. In these case we resort to active learning and propose to gather a new dataset that leads to further improvements of the predictive model. Instead of doing this randomly we propose highly informative DNA libraries using novel machine learning techniques.

A core challenge encountered in these pipeline is the fact that the space of possible proteins is as vast as space of potential DNA sequences encoding them. Many of these proteins are neither biologically plausible nor functionally viable. Only a small but unknown subset contains the relevant design space. Searching over these vast spaces is a big challenge where generative modelling can help and focus only on the relevant parts.

In this project the goal is implement and train a generative model for protein sequences using different generative modelling approaches, namely:

- Energy-based models
- Variational Auto-Encoder(s) (VAE) (Kouba et al., 2023)
- diffusion and flow models for discrete data (Alamdari et al., 2023; Stark et al., 2024)

Each of these have their benefits and use we would like to explore and compare. In our lab and through collaborations, we have access to unique large scale datasets of labeled and unlabeled datasets that would allow us to learn large generative models of the protein landscape.

Using this protein generative model, we would like to inform search space in an active learning pipeline. Given a condition on the enzyme, we would like to generate a *batch* of sequences that are then tested in the wet-lab. The process on informing is often referred to a guidance where a particular part of the support of the generative model is preferred over the others.

2.1 Challenges and Learning outcomes

- Training of state-of-art generative models for sequences akin to one working with language data
- Guidance of generative models by properly assigned weighting of the generative process
- Proper regularization of generative models to facilitate innovation
- Coupling the generative model to active learning pipelines

3 Background

We are seeking a student passionate about spatial data and willing to learn about experiment design in the context of point processes. This thesis is well suited for data science, statistics, computer science, or bioinformatics study program. You will be working closely with Vignesh Ram Somnath and Dr. Mojmir Mutny. For further questions please contact us at `vsomnath@inf.ethz.ch` or `mojmir.mutny@inf.ethz.ch`.

References

Alamdari, S., Thakkar, N., van den Berg, R., Lu, A. X., Fusi, N., Amini, A. P., and Yang, K. K. (2023). Protein generation with evolutionary diffusion: sequence is all you need. *bioRxiv*, pages 2023–09.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589.

Kouba, P., Kohout, P., Haddadi, F., Bushuiev, A., Samusevich, R., Sedlar, J., Damborsky, J., Pluskal, T., Sivic, J., and Mazurenko, S. (2023). Machine Learning-Guided Protein Engineering. *ACS Catalysis*, 13(21):13863–13895.

Patsch, D. and Buller, R. (2023). Improving Enzyme Fitness with Machine Learning. *Chimia*, 77(3):116–116.

Stark, H., Jing, B., Wang, C., Corso, G., Berger, B., Barzilay, R., and Jaakkola, T. (2024). Dirichlet flow matching with applications to dna sequence design. *arXiv preprint arXiv:2402.05841*.

Vornholt, T., Mutný, M., Schmidt, G. W., Schellhaas, C., Tachibana, R., Panke, S., Ward, T. R., Krause, A., and Jeschek, M. (2024). Enhanced Sequence-Activity Mapping and Evolution of Artificial Metalloenzymes by Active Learning. *bioRxiv*.