Last Updated on August 23, 2024

Daiheng Zhang

(929) 496-8375 dz2266@nyu.edu

RESEARCH INTEREST

I am recently interested in AI for protein and lify science, especially in generative modeling and protein language model.

EDUCATION

New York University, MSc in Center for Data Science *Sep 2021 - May 2023*

- MSc Coursework: DS-GA 1003 Machine Learning, DS-GA 1011 Natural Language Processing with Representation Learning, DS-GA 1012 Natural Language Understanding, DS-GA 1013 Mathematical Tools for Data Science

Ohio State University, B.S. in Math *Aug 2019 - May 2021*

- Dean's List for 3 semesters
- Undergraduate Coursework: Ordinary differential equation, Differential Equations and Their Applications, Probability, Financial Economics

Zhejiang University, B.S. in Pharmacy *Aug 2015 - June 2019*

RESEARCH EXPERIENCES

AI Lab Research Intern, ByteDance, Advisor: Dr. Quanquan Gu *Aug 2024 - Now*

Project 1: Pre-train multi-modal protein language model to unify representation and generation.

DL scientist Intern, MoleculeMind (AI for Protein Startup, led by Dr. Jinbo Xu), *April 2024 - July 2024*

- Project1: Build the model for enzyme kinetics prediction with protein language models
- Project2: Design the mutation effect predictor on melting temperature.

Research Assistant, The University of Texas at Austin, Advisor: Dr. Qiang Liu *May 2023 - current*

- Project 1: Flow Model For Structure Based Drug Design
- Apply an ODE-based rectified flow framework on ligand generation task, get comparable results on main metrics to other diffusion methods and provide more flexibility. This work is in progress.
- Project 2: EC number classification predictor
- Use semi-supervised method to achieve SOTA performance on the enzyme commission number datasets and shows robustness to sequence modifications. This work is accepted to ICML 2024 Workshop ML4LMS.

Research Assistant, Westlake University, Advisor: Stan Z. Li *Mar 2023 - June 2023*

- Project: A Systematic Survey of Molecular Pre-trained Models.
- Evaluate the true capabilities of various molecular pre-trained models under a unified experimental setting.

Research Assistant, New York University, Advisor: Dr. Michael Picheny *Sep 2022 - Dec 2022*

- Project: Improve Speech Recognition Performance with Unpaired Audio and Text Data. This paper has already been acceptted by INTERSPEECH 2023.
- Utilized advanced pseudo-labeling and language modeling to leverage large quantities of unlabeled data, resulting in significant improvements in Word Error Rate (WER) for the MALACH corpus.

PUBLICATIONS

[1] M. Picheny, Q. Yang, D. Zhang, and L. Zhang. The MALACH Corpus: Results with End-to-End Architectures and Pretraining, accepted to INTERSPEECH, 2023.

[2] D. Zhang, C. Gong, Q. Liu, Flow Model for Structure Based Drug Design, submitted to ICML 2024.

[3] C. Gong, D. Zhang, J. Zhang, A. Klivans, Q. Liu, Enzyme Classification via Semi-Supervised Functional Residue Learning, accepted to ICML 2024 Workshop ML4LMS

PROGRAMMING SKILLS

Proficient: Python, PyTorch, MySQL

Intermediate: Bash/CMD, Matlab, Java

My journey in the realm of AI and biology has been both challenging and enlightening. With a solid academic foundation in Pharmacy and Math majors and lots of machine learning project experience, my ambition is to develop machine learning algorithms, with a specific interest in generative modeling and protein language models, to unravel complex problems in biology.

Generative modeling for bio/chem The effectiveness of generative modeling has been very promising in recent years. Collaborating with **Dr. Qiang Liu** at the University of Texas at Austin, I worked on a Structure Based Drug Design task. The main method was the **Rectified Flow** [1] proposed by Dr. Qiang Liu, which is an ODE-based probability flow model. This method, previously explored in the field of computer vision, faced unique challenges when adapted to the molecule domain due to differences in data formats. In order to solve this, I proposed to use additional cost function motivated by optimal transport to further improve model performance and finally achieve competing performance with the SOTA diffusion-based method. The primary work has been submitted to the **ICML 2024** once. Recently, I have been modifying this model with previous rebuttal feedback and plan to resubmit it to **ICLR 2025**.

Protein language model My understanding of protein language models originates from their powerful representation of learning capabilities and the support they provide for downstream tasks. After the previous project was submitted, I continued to work on a second project in **Dr. Qiang Liu**'s group. Here, we focused on the problem of predicting enzyme EC number classifications. The main motivations here were: (1) The previous contrastive learning baseline had a long-tail problem, resulting in infrequent samples not being well-trained; and (2) The baseline used average pooling of ESM model embeddings, but different amino acids contribute differently to enzyme function. To address this issue, we proposed a semi-supervised approach that combined mCSA data with evolutionary priors to reweigh the protein residues. Our model achieved a $\sim$30% and $\sim$15% improvement in F1 score with two well-established benchmarks compared to the next best model. This work has already been accepted to the **ICML 2024 ML4MS worksop** as "Enzyme Classification via Semi-Supervised Functional Residue Learning," and a full version was also reviewed in **NeurIPS 2024**.

I believe AI for science is not just about research papers. I'm very curious about how the things I've studied are being applied in the industry. Driven by this curiosity, I undertook a research internship at MoleculeMind, an AI for protein startup led by **Dr. Jinbo Xu**. The first research project I undertook was enzyme kinetics prediction. Our goal was to predict how mutations in protein sequences could result in favorable reactions with the substrate and then use this trained model to guide the next steps in enzyme design. The protein training set used was from the deep mutational scanning dataset. My algorithm mainly involved using ESM2/ProtT5 to extract protein embeddings, then using SMILES transformer/Uni-Mol to extract substrate embeddings, and finally, making the two embeddings interact to predict kcat/km. My results showed a slight improvement over the previous baseline, but that is still not good enough. This may be due to not utilizing structure-related features, or it might be that end-to-end prediction for this task is inherently challenging (as no existing model has fully succeeded in this task yet).

After finishing my exploration on the previous project, I tested the performance of some deep learning-based protein function predictors for future use. While testing DDG predictors, I noticed that another protein fitness-related property, DTm (melting temperature), was rarely predicted. Additionally, I was motivated to fine-tune and evaluate the performance of several state-of-the-art protein language models, primarily ESM2, ESM3, and SaProt, as well as OpenFold, which is structure-based, but can also be used to extract embeddings. I found that the model based on OpenFold performed the best, indicating that structural features are important in representation. Additionally, I improved Spearman's correlation from the current SOTA of 0.47 to 0.51. I will be submitting this project to the **NeurIPS 2024 MLSB workshop**.

After finishing my work at MoleculeMind, I realized the importance of the representational capabilities of protein language models. Motivated by ESM3, I also recognized the potential to unify representation and generation to create a multimodal protein language model. This will be the focus of my upcoming work at ByteDance AI Lab, led by **Dr. Quanquan Gu**. I believe this pre-trained protein language model will be a promising project.

Other projects At NYU, I collaborated with **Dr. Michael Pienchy** on speech recognition for the MALACH corpus. We employed semi-supervised learning techniques with a fine-tuned wav2vec2 system, which led to a substantial reduction in the Word Error Rate (WER) to 13.5%. This work was recognized and accepted for presentation at **INTERSPEECH 2023** [2]. This also marked the beginning of my machine learning projects.

I also collaborated with **Dr. Stan Z. Li** from Westlake University to conduct fair and comprehensive evaluation on small molecule pretrain models. My role involved extending the empirical experiments on current SOTA pretrain models based on the previously published paper and providing an in-depth analysis of this field. Currently, the work is in its final stages and is slated for submission to the **TPAMI** journal in the near future.

Research Plan My interest remains firmly rooted in AI for Protein and Life Science. My interdisciplinary background is an advantage in applying algorithms more effectively in practical scenarios. I aim to engage in cross-disciplinary exchanges, contribute to core and foundation problems in this area, and also design specific algorithms to apply to real-world scenarios.

Why School? I would be thrilled to pursue a PhD at the Columbia University CS department. Based on my background and interests, I hope to collaborate with professor **Mohammed AlQuraishi**, his previous works in machine learning for drug discovery, especially openfold, are very important reason for my choice of Columbia University. I also hope to collaborate with professor **David Blei** for his expert in Variational Inference and professor **Shih-Fu Chang** for his expert in multimedia. Looking ahead, I aspire to further my career in academia. I believe that my background and skills make me a suitable candidate for Columbia University, where I am committed to continuous learning and contributing my expertise.

References

[1] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

[2] Michael Picheny, Qin Yang, **Daiheng Zhang**, and Lining Zhang. The malach corpus: Results with end-to-end architectures and pretraining. *INTERSPEECH 2023*, 2023.