ZAN CHEN

Phone: (+86) 182-7023-8123 ◇ Email: cz990816@gmail.com

Homepage

Google Scholar ◇ Github ◇ LinkedIn

RESEARCH INTERESTS

My research interests lie in the transformative potential of **AI for Science**.

Currently, I am working on leveraging **natural language processing (NLP)** techniques to decode biomedical information. My projects include developing **large language models (LLMs)** capable of aligning human language with protein sequences and designing multi-agent frameworks driven by LLMs to facilitate intricate tasks in biology, such as protein engineering.

Looking ahead, I aspire to delve deeper into critical areas of AI research, including multimodality, trustworthiness, interpretability, usability, and so on. By advancing these facets, I aim to contribute to the broader development of AI for Science, enabling more robust and accessible tools to tackle scientific challenges across disciplines.

EDUCATION

M.Sc. in Data Analytics Engineering ***May 2023***
Northeastern University (NEU), *Boston, MA.*
GPA: 3.758/4.0

B.Sc. in Computer Science and Technology (Artificial Intelligence) ***June 2021***
Shanghai Normal University (SHNU), *Shanghai, China.*
GPA: 3.25/4.0

PUBLICATIONS AND PREPRINTS

Accepted

[1] Y. Shen*, **Zan Chen***, M. Mamalakis*, *et al.*, "A fine-tuning dataset and benchmark for large language models for protein understanding," in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2024)*, 2024.

[2] Y. Shen*, **Zan Chen***, M. Mamalakis*, *et al.*, "**TourSynbio**: A multi-modal large model and agent framework to bridge text and protein sequences for protein engineering," in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2024)*, 2024.

[3] Y. Liu*, **Zan Chen***, Y. G. Wang, and Y. Shen, "**AutoProteinEngine**: A large language model driven agent for multimodal automl in protein engineering," in *The 31st International Conference on Computational Linguistics(COLING 2025)*, 2024.

Under Review

[4] E. Yi-Ge, **Zan Chen**, J. Chen, Y. G. Wang, and Y. Shen, "**RA-Agent**: A human-centered large language model for enhanced academic research support," 2024.

WORK EXPERIENCE

AI Algorithm Research Scientist Aug 2023 - Present
Supervisors: Dr. Shen, Prof. Wang Toursun Synbio

- ***Multi-modal Protein Large Language Model Development*** [1]

Led the development of **TourSynbio**, a groundbreaking large language model designed to understand both protein sequences and human language without the use of external protein encoders.

- Developed **ProteinLMDataset**, a pretraining dataset containing 17.46 billion tokens, integrating pure protein sequence data, human language descriptions of proteins, and scientific text. This dataset has set new standards in the self-supervised training of protein models.
- Innovated in instruction fine-tuning by creating a dataset with 893K instructions, significantly enhancing the model's ability to perform protein-related tasks such as mutation analysis, protein folding, and visualization.
- Designed and generated a specialized benchmark for the protein field, **ProteinLMBench**, and validated that Toursynbio achieved SOTA performance with an accuracy of 62.18%, surpassing all other open-source and commercial models at the time, including GPT-4.

· ***Protein Engineering Agents Development*** [2], [3]

Co-created **TourSynbio-Agent**, an LLM Multi-Agent framework that seamlessly integrates deep learning models for protein engineering tasks. This framework facilitates complex tasks like mutation prediction, inverse folding, and enzyme catalysis, significantly streamlining workflows in experimental biology.

- Developed **AutoProteinEngine (AutoPE)**, the first multimodal AutoML framework for protein engineering, enabling users to interact with the system via natural language, simplifying tasks like model selection and hyperparameter tuning for both protein sequences and protein graphs.
- Applied the model in two successful wet-lab case studies, resulting in higher mutation accuracy, enhanced enzyme activity, and faster experimental turnaround, thereby improving efficiency in biocatalytic processes and enzyme modifications.

· ***Research Agents Development*** [4]

Led the development of **RA-Agent**, a human-centered research assistant designed to improve academic productivity by automating key research tasks.

- Developed an advanced LLM-based system that integrates document summarization, translation, text refinement, and innovation review, significantly improving research efficiency.
- Implemented retrieval-augmented generation (RAG) and domain-specific recognition, enhancing the accuracy and fluency of summarization and text generation tasks.

PROFESSIONAL SERVICE

Workshop Reviewer:

MICCAI 2024 2nd International Workshop on Foundation Models for General Medical AI

Workshop Organizer:

The 1st International Workshop on Foundation Models for Bioinformatics and Biomedicine (BIBM 2024)

SKILLS/HOBBIES

Programming Languages	Python, R, C/C++, MATLAB,
Machine Learning Tools	Pytorch, Pandas, Numpy, Tensorflow, Sklearn
Hobbies	Photography, Anime