Phone：+971 585358674 Email：jiayou.zhang@mbzuai.ac.ae

Jiayou Zhang

Research Interests: *AI for drug discovery, protein structure prediction, bio–foundation models, generative modeling.*

Machine learning researcher specializing in protein structure prediction and bio–foundation models. Experienced in large-scale model training, generative modeling, and AI-driven drug discovery, with research contributions at CMU, BioMap, and GenBio AI.

Education

Mohamed bin Zayed University of Artificial Intelligence (MBZUAI) Ph.D. in Machine Learning 2022 – 2026
Tsinghua University B.Eng. in Computer Science and Technology 2018 – 2022

Experience

GenBio AI – R&D Intern 2024 – Now

1. Deployed, accelerated, and optimized protein structure prediction models, leveraging PyTorch, Protenix, MMSeqs2, and related tools to deliver efficient and scalable model performance.
2. Led the development of a protein structure tokenizer (AIDO.StructureTokenizer) and integrated it into a 160B-parameter protein language model (AIDO.Protein) using the Megatron-LM framework. The work was featured at the NeurIPS MLSB Workshop, and the model achieved SOTA results on MSA-free structure prediction tasks, strengthening the company's influence in bio FMs.
3. Improved the MSA data pipeline by tuning Jackhmmer, MMSeqs2, and protein structure models (AlphaFold3, Protenix, Boltz) across diverse datasets; built MSA subsampling and denoising tools to enhance antibody–antigen structure prediction.
4. Contributed to end-to-end productization, refactoring and optimizing tokenizer and training components, and collaborating across teams to integrate and validate internal models. The project has been open-sourced on HuggingFace.

BioMap – R&D Intern 2023 – 2024

1. Developed a diffusion-based method for generating diverse antibody conformations, enabling more accurate and flexible modeling of antibody CDR regions.
2. Designed and implemented a template-driven antibody–antigen docking pipeline that achieved SOTA performance on multiple benchmarks, strengthening the company's technical edge in antibody drug discovery and adopted by partner pharma companies.
3. Cleaned and curated antibody–antigen PDB datasets for training and evaluation by using visualization tools and domain knowledge.
4. Optimized the docking pipeline through modular refactoring and performance improvements.

Projects

Protein Structure Tokenizer: Led the design and integration of a protein structure tokenizer into a 16B-parameter protein language model, achieving SOTA performance on MSA-free structure prediction tasks.
Project link: https://huggingface.co/genbio-ai/AIDO.Protein2StructureToken-16B
MSA Optimization & PDB Data Curation: Evaluated multiple MSA search tools and pretrained models; developed an MSA subsampling tool and used visualization workflows to effectively remove noisy data and improve prediction quality.
Antibody Conformation Generation & Docking: Built a diffusion-based framework for generating diverse antibody conformations and designed a template-driven antibody–antigen docking pipeline that achieved SOTA benchmark performance and was adopted by partner pharma companies.
Open-Sourcing & Productionization: Contributed extensively to core code development and optimization, completed multi-module integration, and drove the open-source release of projects on Hugging Face.

Other Experience

CMU	SAILING Lab	Visiting Researcher	2024
MBZUAI	SAILING Lab	Research Assistant	2022
University of Washington	Wang Lab	Research Assistant	2021
Tsinghua University	THUNLP Lab	Research Assistant	2020

Publications

- Zou, S., **Zhang, J.**, Zhao, B., Li, H., Song, L. (2026). *Accurate RNA 3D Structure Prediction via Language Model-Augmented AlphaFold 3.* **ICLR (under review).**
- Hu, J., **Zhang, J.**, Cui, S., Zhang, K., Chen, G. (2026). *MixAR: Mixture Autoregressive Image Generation.* **CVPR (under review).**
- **Zhang, J.**, Shen, Y., Chen, G., Song, L., Xing, E. P. (2025). *Dimensional Collapse in VQVAEs: Evidence and Remedies.* **NeurIPS**.
- **Zhang, J.***, Meynard-Piganeau, B. *, Gong, J., Cheng, X., Luo, Y., Ly, H., Song, L.†, Xing, E. P. (2024). *Balancing Locality and Reconstruction in Protein Structure Tokenizer.* **NeurIPS MLSB Workshop**.
- Wang, X., Li, C., Wang, Z., Bai, F., Luo, H., **Zhang, J.**, Jojic, N., Xing, E. P., Hu, Z. (2024). *PromptAgent: Strategic Planning with LLMs Enables Expert-level Prompt Optimization.* **ICLR**.
- Xu, H.*, **Zhang, J.***, Wang, Z.*, Zhang, S., Bhalerao, M., Liu, Y., Zhu, D., Wang, S.† (2023). *GraphPrompt: Graph-Based Prompt Templates for Biomedical Synonym Prediction.* **AAAI**.

Skills

Protein Structure Prediction Bio Foundation Model PyTorch Megatron-LM Hugging Face
AlphaFold OpenFold Protenix Boltz MMSeqs2 Jackhmmer Hhblits PyMol PDB
Python C/C++ Docker Distributed Training