

Education

Massachusetts Institute of Technology

Sep. 2025 - Present

PH.D. IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE, EXPECTED JUN. 2030

- GPA: **5.0/5.0**
- Relevant Coursework: Deep Learning (**A+**), AI for Protein Biology (A), Advances in Computer Vision (A)

Tsinghua University

Sep. 2021 - Jul. 2025

B.ENG. IN COMPUTER SCIENCE AND ENGINEERING

- GPA: **3.94/4.0**

Technical Skills

Languages & Frameworks Python, C/C++, PyTorch, Hugging Face Transformers, DeepSpeed
Infrastructure Distributed model training, CUDA, Slurm, Docker, Git

Publications

- Bo Chen*, **Zhilei Bei***, Xingyi Cheng, Pan Li, Jie Tang, Le Song. *MSAGPT: Neural Prompting for Protein Structure Prediction via MSA Generative Pre-Training*. *NeurIPS* (2024). * Equal contribution.
- Bo Chen*, Xingyi Cheng*, Pan Li, Yangli-ao Geng, Jing Gong, Shen Li, **Zhilei Bei**, et al. *xTrimoPGLM: Unified 100B-Scale Pre-Trained Transformer for Deciphering the Language of Proteins*. *Nature Methods* (2025). * Equal contribution.
- Ajinkya Deshpande, **Zhilei Bei**, Jian Ma, Spencer Krieger. *MIMYR: Generative Modeling of Missing Tissue in Spatial Transcriptomics*. *Gen2 Workshop at ICLR* (2026).
- Boyan Wang, Yangliao Geng, Xingyi Cheng, Bo Chen, **Zhilei Bei**, et al. *ProtGO: Universal Protein Function Prediction Utilizing Multi-Modal Gene Ontology Knowledge*. *Bioinformatics* (2025).

Research Experience

Massachusetts Institute of Technology

Ph.D. Researcher, Hwang Lab

Sep. 2025 - Present

- **Interpretable Learning of Bacterial Regulatory Networks from Genome Sequence** (Hwang Lab)
 - Developing an interpretable machine learning framework to predict gene expression from microbial genome.
 - Recovering transcriptional regulatory interactions and outperforming established bioinformatics baselines.
- **Context-aware Protein Subcellular Localization Prediction** (rotation: Kaiming He, Le Cong)
 - Constructed a benchmark for protein subcellular localization classification and inpainting across 37 cell lines.
 - Improved AUPRC by 22% over a sequence-only baseline on unseen cell lines by incorporating immunofluorescence microscopy images and transcriptomic profiles as cellular context.
- **Protein Metal Binding Site Prediction for Metalloprotein Discovery** (rotation: Yunha Hwang)
 - Developed a sequence-based predictor of metal-binding residues for efficient metalloprotein screening.
 - Enabled zero-shot generalization to unseen metal ions by learning a joint protein-metal embedding space.
- **Few-shot Conditional Generation for Protein Design** (rotation: Alex Rives)
 - Developed a masked diffusion language model that conditions on few-shot examples at inference time, achieving up to a 90% success rate in generating sequences with target EC classification across diverse enzyme families.
 - Incorporated classifier-free guidance into sampling, doubling the Spearman correlation between model-predicted residue probabilities and experimentally measured fitness relative to unguided sampling.

Zhipu AI

Student Researcher

Feb. 2025 - Jun. 2025

- **Reinforcement Learning for Self-Improvement in Machine Learning Agents**
 - Designed agent scaffolding for long-horizon ML engineering, with a memory module accumulating context.
 - Applied RL over single-step actions to train on high-information trajectories: medal rate 12% → 14% on MLE-Bench.

Carnegie Mellon University

Research Assistant, Advisor: Prof. Jian Ma

Jun. 2024 - Sep. 2025

- **Predicting Missing Tissue Regions in Spatial Transcriptomics** (Mimyr, *ICLR Gen2 Workshop* 2026)
 - Paired a VQVAE over expression profiles with a ViT that inpaints spatial transcriptomics for missing tissue regions.
 - Improved gene expression Spearman correlation by 23% over neighborhood lookup baseline on held-out regions.

- **Protein Language Foundation Models** (xTrimopGLM, *Nature Methods* 2025)
 - Fine-tuned 100B-parameter models and discovered 4 types of pre-training scaling for different downstream tasks.
 - Raised model generation performance by up to 3 fold via supervised fine-tuning and reinforcement self-training.
- **Protein Multiple Sequence Alignment (MSA) Generation** (MSAGPT, *NeurIPS* 2024)
 - Developed an autoregressive language model to generate synthetic MSAs for proteins in low-homology regimes, improving AlphaFold2 structure-prediction TM-score from 0.44 to 0.51 on orphan proteins in CASP.
 - Used AlphaFold2 structure-prediction accuracy as feedback for RL, further improving TM-score by 0.06.

- **Pocket-aware Ligand Generation with Diffusion Models**
 - Guided pocket-conditioned small molecule ligand diffusion with a GNN affinity predictor.

- **Disease Target and Drug Discovery Using Community Detection on Knowledge Graphs**
 - Ranked therapeutic targets and drugs for breast cancer from knowledge-graph embeddings + community detection.

- **Blocking the Binding of Bacterial Lipid Droplets to Phage DNA**
 - Identified a blocker of bacterial lipid droplet-phage DNA binding *in vitro*.

Honors & Awards

2025 **Sunlin (1966) and Priscilla Chou Fellowship**, MIT EECS Graduate Student Fellowship

2024 **Spark Scientific and Technological Innovation Fellowship**, Top 1% of Tsinghua University

2024 **Zhong Shimo Scholarship**, Top 1% of Computer Science Department, Tsinghua University

2024 **National Scholarship**, Top 3% of Tsinghua University

2023 **Comprehensive Scholarship**, Top 10% of Tsinghua University

2022 **Comprehensive Scholarship**, Top 10% of Tsinghua University

2020 **Finalist**, Regeneron International Science and Engineering Fair (ISEF)