

Pengfei Zhang

📍 727 E Tyler St, Tempe, AZ 85281 USA ✉ pzhang84@asu.edu ☎ +1(480) 853-4514

🔗 pzhang84.github.io

Summary

Self-motivated Computer Science Ph.D. Candidate specializing in **Generative AI**, **Large Language Models (LLMs)**, and **AI-driven protein design**. Strong background in developing foundation models for biological sequences, optimizing reinforcement learning pipelines, and building scalable predictive frameworks for molecular interaction and functional sequence modeling. Demonstrated ability to translate cutting-edge ML research into practical tools for computational biology and drug discovery. Core strengths include:

- **Generative Protein Modeling**: Adapted GPT-style LLMs for conditional protein sequence design.
- **Reinforcement Learning**: Reframed RLAIIF to expose reward model flaws and improve alignment.
- **Representation Learning**: Built context-aware embeddings to boost prediction on complex bio data.

Education

Arizona State University <i>Ph.D. in Computer Science, GPA: 3.9/4.0</i>	<i>08/2019 – 04/2026 (expected)</i>
Arizona State University <i>M.S. in Computer Science</i>	<i>08/2019 – 12/2022</i>
Changchun University of Science and Technology <i>B.E. in Optoelectronic Engineering (with honors)</i>	<i>09/2015 – 06/2019</i>

Selected Projects

Protein Language Models and Generative Models <i>3 first-author papers delivered</i>	<i>05/2023–Present</i>
--	------------------------

- Adapted generative model training to an **in-context learning** approach, enabling **conditional protein sequence generation** conditioned on novel target binding sequences and few-shot prompting.
- Developed an automated **chain-style, self-contemplating prompting** method that eliminated the need for human-fed few-shot examples without compromising model performance.
- Fine-tuned and deployed protein language models with reinforcement learning from AI feedback (RLAIIF) to attack and defend biological prediction models, increasing model robustness **fivefold**.
- Designed a modular **reward hacking mitigation framework** for conditional protein generation, incorporating reward smoothing, ensemble rewards, and distributional regularization to preserve sequence plausibility during RL.

Data-Efficient Machine Learning for Biological Prediction <i>1 first-author papers delivered</i>	<i>12/2022–04/2023</i>
--	------------------------

- Designed and deployed active learning frameworks, reducing future annotation overhead by **50%** through entropy-based query strategies that optimized data selection for unlabeled datasets.
- Repurposed active learning to identify **40% redundant labeled samples**, enhancing model efficiency by focusing training on more informative data, reducing computational cost and model training time.

Representation Learning for Protein Sequences <i>2 first-author papers and 1 US patent delivered</i>	<i>05/2021–02/2022</i>
--	------------------------

- Developed **context-aware embedding models** for amino acid and protein sequences using bidirectional LSTM and self-attention, optimized for immune receptors.
- Benchmarked against traditional embedding methods like BLOSUM, word2vec, and Bert, achieving superior performance in predictive accuracy and cluster quality across supervised and unsupervised tasks.
- Enhanced sequence binding predictions by over 20% in AUC scores and reduced data requirements by 90%.

Technical Skills

- **Programming Languages:** Python, R, SQL, Java, C++, Bash
- **Machine Learning Frameworks:** TensorFlow, PyTorch, Keras, Scikit-Learn, Hugging Face Transformers
- **Data Analysis:** NumPy, Pandas, Seaborn, Matplotlib
- **Cloud Platforms:** AWS SageMaker, Google Cloud Platform
- **Version Control:** Git, GitHub, GitLab

Selected Publications

LLMs and Generative Models

- **P. Zhang**, S. Bang, and H. Lee, “Self-Contemplating In-Context Learning Enhances T Cell Receptor Generation for Novel Epitopes”, accepted in *Machine Learning in Computational Biology (MLCB)* 2025.
- **P. Zhang**, H. Mei, S. Bang and H. Lee, “Iterative Attack-and-Defend Framework for Improving TCR-Epitope Binding Prediction Models”, in *Intelligent Systems For Molecular Biology (ISMB/ECCB)* 2025.
- **P. Zhang**, F. Guan, X. He, H. Mei, G. Grama, S. Bang, and H. Lee, “Mitigating Goodhart’s Law in Epitope-Conditioned TCR Generation Using Plug-and-Play Reward Designs,” submitted to *International Conference on Research in Computational Molecular Biology (RECOMB)* 2025.

Representation Learning

- **P. Zhang**, S. Bang, M. Cai and H. Lee, “Context-Aware Amino Acid Embedding Advances Analysis of TCR-Epitope Interactions”, in *eLife* 2024.
- **P. Zhang**, S. Bang and H. Lee, “PiTE: TCR-epitope Binding Affinity Prediction Pipeline using Transformer-based Sequence Encoder”, in *Pacific Symposium on Biocomputing (PSB)* 2022.

Data-Efficient Machine Learning

- **P. Zhang**, S. Bang and H. Lee, “Active Learning Framework for Cost-Effective TCR-Epitope Binding Affinity Prediction”, in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 2023.
- M. Cai, S. Bang, **P. Zhang** and H. Lee, “ATM-TCR: TCR-Epitope Binding Affinity Prediction Using a Multi-Head Self-Attention Model”, in *Frontiers in Immunology* 2022.

Patents

- H. Lee, **P. Zhang**, M. Cai, and S. Bang, “Systems and methods for a bidirectional long short-term memory embedding model for T-cell receptor analysis”, *US Patent US20240339173*, filed on April 10, 2024.

Awards

- *ASU Outstanding Research Award*, 2024.
- *ASU SCAI Doctoral Fellowship*, 2024.
- *PSB 23 Travel Award*, 2023.
- *ASU GPSA Travel Grant Individual Award*, 2022.
- *ASU Biodesign Travel Award*, 2022.
- *ASU SCAI Travel Award*, 2022.