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 **Computational Biology**. Proven track record of delivering innovative ML solutions, including fine-tuning GPT-based models, developing protein language models, and optimizing data-efficient learning frameworks. Expertise includes:

- **Generative AI**: Fine-tuning GPT-based LLMs for **novel immune sequence design**.
- **Reinforcement Learning**: Implementing RLAIF to improve reward model robustness and performance.
- **Protein Language Models**: Leveraging transformer architectures for protein sequence analysis and design.
- **Representation Learning**: Developing context-aware embeddings for **imuunological sequnece data**.

Education

Arizona State University *08/2019 – 12/2025*
Ph.D. in Computer Science, GPA: 3.9/4.0 *(expected)*

Arizona State University *08/2019 – 12/2022*
M.S. in Computer Science

Changchun University of Science and Technology *09/2015 – 06/2019*
B.E. in Optoelectronic Engineering (with honors)

Selected Projects

Protein Language Models and Generative Models *05/2023–Present*
2 first-author papers delivered

- Adapted generative model training to an **in-context learning** approach, enabling immune T cell receptor design 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 (RLAIF) to attack and defend biological prediction models, increasing model robustness **fivefold**.
- Developed an **antibody language model** utilizing multiple sequence alignment and axial attentions to accurately recover missing residues in immune B cell receptor profiling.

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

- Designed **active learning** frameworks using innovative entropy-based query strategies to optimize data selection.
- Reduced **50% annotation costs** for unlabeled pairs and minimized **40% data redundancy** among annotated pairs.

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

- 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 **93%**.

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.

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.