CAMELLIA (XINYUE) RUI

Greater Los Angeles Area | 6085048266 | xruiapp@gmail.com | LinkedIn

EDUCATION

University of Southern California **Aug 2022 - May 2027**

Ph.D., Biostatistics

- **GPA:** 3.8/4
- **Achievements:** Guest lecturer on AI-assisted coding & best practices in bios dept
- **Coursework:** Multi-Modal Deep Generative Modeling, Representation Learning, Variational Autoencoder, LLM agents, Machine Learning, Deep Learning, Transformer, Data Analysis, Statistical Inference (TA), Mathematical Statistics, Probability, Advanced Statistical Computing (TA), Statistical Analysis of High-Dimensional Data

University of Southern California **Aug 2019 - May 2022**

B.A., Mathematics

- **GPA:** 3.6/4

SELECTED PROJECTS

PhD_Agent: AI-powered research assistant **Sep 2025 - Present**

- Built an LLM agent with Claude Code SDK, integrating GitHub MCP to summarize weekly commits and track project progress
- Integrated with Slack MCP and Zotero MCP to retrieve research papers, store references, and generate automatically structured meeting agendas contextualized by the prior week's work
- Extended the agent with a Retrieval-Augmented Generation (RAG) system using LangChain, Chroma, and OpenAI API to answer domain-specific questions from papers
- Developed an AI-powered conference session recommender using RAG and vector similarity search to automatically filter and rank 500+ abstracts against personalized research interests, achieving 92.3% precision (F1: 0.96) validated through systematic human evaluation
- Designed an end-to-end agent pipeline combining multi-platform MCP APIs, RAG workflows, and conversational interfaces for a research assistant that serves PhD researchers

EXPERIENCE

Genentech, Inc. **May 2025 - Aug 2025**

ML Research Intern | AI for Biology *South San Francisco, CA*

- Conducted research and developed a deep learning Variational Autoencoder (VAE) model to model gene regulatory networks in a team of four using PyTorch
- Engineered a model prototype from scratch using JAX and identified identifiability and misparameterization issues in the existing codebase
- Reduced overall model loss from 6.7 × 10^−2 to 1 × 10^−7 and improved inference accuracy by 19.7%
- Successfully implemented the knockoff procedure within the VAE model to denoise real biological signals while controlling the false discovery rate (FDR) under 10%
- Managed reproducible code through GitLab using Merge Request-based Model Context Protocol (MCP), integrating open collaboration and clear communication across the team

University of Southern California **Mar 2024 - Present**

Research Assistant - SCFM

Developed a machine learning method SCFM that identifies gene-to-disease associations on the largest-scale single-cell RNA-seq data (4.1GB), utilizing coordinate ascent variational inference

- Achieved an average of 32% improvement in sensitivity and discovered an average of 15% more genetic variants when benchmarking against the existing method through extensive simulations
- Built a new Python package implementing SCFM framework with JAX, leveraging big data technologies and HPC clusters to achieve ultra-fast computing speed with an average inference time 15x faster than the existing method (1.3s vs 20s)
- Enabled robustness on calibration and model misspecification over 4000+ simulation scenarios and benchmarked the method against baseline and other published models
- Accepted as the first-author abstract to a top-tier conference American Society of Human Genetics, demonstrating strong communication and publication skills

University of Southern California **Mar 2024 - Present**

Research Assistant - PerturbVI

Developed a machine learning method PerturbVI that discovered gene regulatory networks with CRISPR perturbation data and single-cell RNA-seq data using Variational Inference and Jax in a team of three

- Simulated model misspecification of latent variables using Python and improved 6.5% sensitivity compared to existing methods
- Enabled ultra-fast inference speed with an average convergence time of 70x faster on the largest scale perturbation matrix (310,385 x 8563) than the existing method
- Optimized core algorithms, improving statistical inference by reducing computation time of local false sign rate by 4x and significantly accelerating large-scale genetic analyses
- Collaborated with team members to enhance model initialization, decreasing compiling time from 3.5 minutes to 1 minute and improving overall productivity

TECHNICAL SKILLS

- **Core Competencies**: Multi-Modal Deep Generative Modeling, Representation Learning, Variational Autoencoder, LLM agents, Machine Learning, Deep Learning, Transformer, Data Analysis, Statistical Inference, Mathematical Statistics, Probability, Advanced Statistical Computing, Statistical Analysis of High-Dimensional Data, Natural Language Processing, Big Data Technologies
- **Programming Languages**: Python, R, Bash, SQL, SAS
- **Libraries & Frameworks**: JAX, PyTorch, Scikit-learn, Numpy, Pandas, SciPy, Hatch, Keras
- **Tools & Platforms**: Git, GitHub, GitLab, ssh, Linux, HPC, LaTeX, Chroma, OpenAI SDK, Claude Code SDK, Cursor, MCP, Cline, Gemini, LangChain
- **Professional Skills**: Research Experience, Communicator, Publications

PUBLICATIONS

- scFM: An efficient statistical fine-mapping approach for eQTLs using large-scale single-cell data.*1st author(in Prep), ASHG 2024 Abstract*
- peturbVI: A Scalable Latent Factor Model to Infer Genetic Regulatory Modules through CRISPR Perturbation Data.*2nd author(in Prep), BOG talk, ASHG 2024 Abstract*
- Estimating heritability explained by local ancestry and evaluating stratification bias in admixture mapping from summary statistics.-*2nd author, American Journal of Human Genetics*
- A global view of disparity in imputation resources for conducting genetic studies in diverse populations.*2nd author, American Journal of Human Genetics*