XINYUE(CAMELLIA) RUI

Greater Los Angeles Area| xruiapp@gmail.com | LinkedIn | (608) 504-8266

EDUCATION

Ph.D. Biostatistics - University of Southern California *August 2022 - May 2026 (3.8/4)*
Focus: Machine Learning in Statistical Genetics

M.S. Biostatistics - University of Southern California *August 2021 - May 2022 (3.8/4)*

B.A. Mathematics - University of Southern California *August 2019 - May 2022 (3.6/4)*

Courses:
Machine Learning, Deep Learning, Large Language Models, LLM Agent, Reinforcement Learning, Transformers, Data Analysis, Statistical Inference (TA), Mathematical Statistics, Probability, Advanced Statistical Computing (TA), Statistical Analysis of High-Dimensional Data

TECHNICAL SKILLS

Languages	Python, R, Bash, SQL, SAS
Libraries	Jax, PyTorch, Numpy, Pandas, SciPy, Scikit-learn, Keras, LangChain, OpenAI
Tools	Git, GitHub, ssh, Linux, HPC, LaTeX, AWS, Chroma

EXPERIENCE

University of Southern California
Research Assistant - SCFM Aug 2022 – Present

- 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 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

Research Assistant - PerturbVI Mar 2024 – Present

- 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

Research Assistant - Worldwide Imputation Analysis May 2020 – May 2022
University of Southern California

- Built a statistical analysis pipeline (**linear regression**) using **Python** and R and conducted the experiments for accessing genotype imputation quality over **123** populations

- Discovered that the imputation quality fell short **6.5%–42%** in imputation R square among minority populations compared to European controls
- Raised minority awareness by presenting research results during the undergraduate poster session and awarded Provost Research Fellowship twice (in fall 2020 and fall 2021)
- Findings were published in the top tier journal (AJHG, IF=12.6)

Southern California Clinical and Translational Science Institute
Statistical Consultant Aug 2023 - Dec 2023

- Worked with clients as a part-time statistical consultant for a fatherhood intervention research project of 443 men of African and Hispanic group from LA area
- Performed extensive exploratory data analysis and feature engineering on the study cohort
- Developed a generalized linear model and conducted data analysis and visualization in R for fatherhood intervention effectiveness
- **Effectively communicated** data-driven insights through detailed reports and presentations, facilitating productive weekly client meetings
- Consistently received positive client feedback for clarity and precision, while exceeding supervisor expectations through strong problem-solving abilities and attention to detail

SELECTED PROJECTS

Research Assistant Chatbot for Statistical Genetics Nov 2024 - Present

Developed a Retrieval-Augmented Generation **(RAG)** system using LangChain, Chroma, and the OpenAI API to extract and synthesize information from local literature in statistical genetics

- Built a conversational interface to provide enhanced, accurate answers to complex questions in statistical genetics, streamlining research workflows
- Optimized the retrieval pipeline for domain-specific data, ensuring relevance and accuracy of results through vector similarity search and fine-tuned prompts

Grocery Sales Forecasting in Ecuador Aug 2022 - Dec 2022
Ensemble Forecasting Project

- Developed a **machine learning** approach using ensemble methods to forecast product sales utilizing Scikit-learn and NumPy
- The ensembled regressor combined extra tree & **random forest** regression, ridge & support vector machine regression and achieved an RMSLE of 0.41

PUBLICATIONS

scFM: An efficient statistical fine-mapping approach for eQTLs using large-scale single-cell data May 2024
1st author(in Prep), ASHG 2024 Abstract

peturbVI: A Scalable Latent Factor Model to Infer Genetic Regulatory Modules through CRISPR Perturbation Data Mar 2024
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 Dec 2023
2nd author, American Journal of Human Genetics

A global view of disparity in imputation resources for conducting genetic studies in diverse populations Mar 2023
2nd author, American Journal of Human Genetics