Xusheng Ai

97 Heritage Riverwood Dr, Central, SC | xusheng.ai369@gmail.com | (864)207-9788 | GitHub

Professional Summary

Machine Learning Research Scientist Ph.D. Candidate specializing in Generative AI and Foundation Models for Genomics . Expertise in architecting Transformer-based diffusion models (GemDiff) to simulate realistic biological states , utilizing Multi-GPU/Slurm infrastructure for high-throughput inference . Proven track record of developing modular reasoning workflows for single-cell epigenomics at MD Anderson , aligning with the lab's mission to advance genomic language modeling and therapeutic design.

Education

Clemson University, Computer Engineering   Clemson, SC
Doctor of Philosophy, advisor: Prof. Melissa C. Smith, GPA: 3.81/4.0   *Exp. May 2026*

University of the Pacific, Engineering Science   Stockton, CA
Master of Science, GPA: 3.94/4.0   *July 2021*

Changsha University of Science and Technology, Automation Engineering   Changsha, China
Bachelor of Engineering   *Jun 2018*

Skills

Machine Learning: Diffusion Model, Transformer, VAEs, GANs, PyTorch (DDP), Multi-GPU Training, HPC/Slurm, Docker.
Computational Biology: Single-cell Multi-omics (scRNA-seq, scATAC-seq),High-Dimensional Data Integration, Trajectory Inference, Differential Expression and Pathway Enrichment
Computing Languages: Python (Pandas, Scikit-Learn) , PyTorch (DDP) , R (Advanced),Multi-GPU Training , HPC/Slurm Workload Management , Docker, Kubernetes , Linux Environments , Cloud Orchestration (AWS/Azure)

Professional Experience

Applied Research Intern, MD Anderson Cancer Center – Houston, TX   *May 2025 – Nov 2025*
Modeling Biological Dynamics & Cross-Functional Collaboration

- Designed Biological Reasoning Workflows to integrate heterogeneous single-cell epigenomics data (H3K4me3, RNAPII), enabling the logical deduction of cellular lineage commitments from static snapshots.
- Implemented Metacell Aggregation algorithms to resolve data sparsity issues in "messy" high-dimensional datasets, ensuring robust downstream analysis.
- Validated High-Confidence Insights by collaborating with biologists to benchmark pipeline outputs against ground truth, identifying 5,065 significant target genes with 96% cross-validation accuracy.

Research Experience

Graduate Research Assistant, Clemson University – Clemson, SC   *Jan 2022 – Present*
Latent Representation Learning & Trajectory Inference   *Aug 2023 – Present*

- Developed a Hierarchical VAE Backend to extract interpretable latent features from noisy biological time-series data, improving signal-to-noise ratio in developmental datasets.
- Modeled Latent Space Dynamics to infer biological progression (pseudotime), successfully addressing computational ambiguity in complex developmental trajectories.

GemDiff: Generative Foundation Model for Genomics   *Jan 2022 – Aug 2023*

- Architected GemDiff, a Transformer-based Diffusion model capable to simulate realistic tumor gene expression profiles, enabling the synthesis of biological states under varying perturbations.
- Engineered High-Performance Inference Pipelines using PyTorch (DDP) and Slurm on multi-GPU clusters, optimizing the generation of thousands of patent-ready synthetic samples.
- Developed a Benchmarking Suite to validate model performance, implementing modules for manifold projection (UMAP), distribution matching, and functional enrichment to ensure biological fidelity.

Backend Systems Researcher, University of the Pacific – Stockton, CA   *Aug 2019 – July 2021*
Cloud Resource Optimization (A2Cloud-H)

- Optimized Scientific Workflows by developing "A2Cloud-H," a hybrid resource recommendation system that orchestrates complex analysis jobs across AWS environments.

Selected Publications

- X. Ai, et al., "Gemdiff: A diffusion workflow bridges between normal and tumor gene expression states," **Briefings in Bioinformatics**, 2025. (*Tier 1 Journal, IF: ~9.5*)
- X. Ai, et al., "Generative adversarial networks applied to gene expression analysis," **Comp. and Sys. Oncology**, 2023.