ARNAV GUPTA

San Francisco, CA ☎ 412-583-8615 ✉ arnav1211@gmail.com
🎓 Google Scholar linkedin.com/in/arnav1211/ github.com/ArnavGuptaa

Experience

Gordian Biotechnology **June 2023 – Present**

Data Scientist 1, Computational Biology, March 2025 - Present *San Francisco, CA*

- Designed and productionized a perturbation-assignment algorithm that improved signal-to-noise ratio by 10% in pooled CRISPR screens across multiple tissues (*bioRxiv pre-print*)
- Deployed and benchmarked ambient RNA removal pipelines (SoupX, DecontX, CellBender) improving batch integration, differential expression analysis, and perturbation-effect detection in experiments with >1M scRNA and snRNA-seq cells
- Mentored 2 engineers on ML and single-cell analysis best practices. Instituted code review, versioned pipeline standards, and reproducible workflows

Senior Data Analyst, Human Genomics *San Francisco, CA*

- Fine-tuned Geneformer, a transformer-based single-cell foundation model for disease-vs-normal classification and applied it to in-silico perturbation prediction, ranking gene knockouts and overexpressions by predicted effect on single-cell embedding states to nominate screening targets
- Architected end-to-end scRNA-seq pipelines from QC, to downstream processing and dashboard for >1M cells/run on GCP eliminating manual bottlenecks and reducing processing time from days to hours
- Partnered with wet-lab scientists to translate perturbation-screen outputs into experimental designs, nominating targets that advanced to validation and steering screening-program priorities

Helomics **May 2022 – Aug 2022**

Computational Biology Intern *Pittsburgh, PA*

- Trained TensorFlow autoencoders for multi-omic latent representation learning from cancer patient data, surpassing production model accuracy by 10% across held-out cohorts
- Built Cromwell/WDL nested cross-validation pipelines on AWS for hyperparameter search, enforcing data-leakage-free model selection across cancer patient cohorts

Projects

SpatialChat: Multi-Agent RAG System for Spatial Transcriptomics | *LangGraph, ChromaDB, Streamlit* **Mar. 2026**

- Designed a supervisor-routed multi-agent system in LangGraph with Pydantic-validated tool-calling agents for dataset discovery, spatial statistics, and gene expression analysis
- Implemented ChromaDB-backed RAG with semantic retrieval over spatial transcriptomics metadata and gene expression data
- Integrated LangSmith tracing for agent observability and built a Streamlit conversational interface for interactive analysis
- *Github:* https://github.com/ArnavGuptaa/spatialchat

Multi-Omics Cancer Classification with Deep Representation Learning **Sep. 2022 – Dec 2022**

- Trained PyTorch autoencoders for unsupervised multi-omic feature extraction (gene expression + CNV), achieving 95% classification accuracy on lung cancer subtypes
- Systematically compared mono-omic vs. multi-omic representations across SVM and Random Forest classifiers demonstrating that multi-omic embeddings outperform single-source features
- *Github:* https://github.com/ArnavGuptaa/02718-Comp-Medicine-Project

Education

Carnegie Mellon University **Sep. 2021 – May 2023**

M.S. Computational Biology, GPA 4.0 *Pittsburgh, PA*

- Courses: Machine Learning, Neural Networks and Deep Learning, Advanced Data Structures and Algorithms, Bioinformatics Data Integration, Mathematics and Statistics, Quantitative Genetics

Birla Institute of Technology and Science Pilani, Goa Campus **Aug. 2015 – May 2020**

B.E. Electrical and Electronics Engineering, GPA 9.29/10.00 *Goa, India*

M.Sc. Biological Sciences, GPA 9.29/10.00

- Courses: Optimization, Linear Algebra, Probability and Statistics, Computer Programming, Signal Processing

Technical Skills

Languages: Python, C++, SQL, R, Bash
AI/ML: PyTorch, TensorFlow, Scikit-Learn, LangChain, LangGraph, Hugging Face Transformers, RAG, ChromaDB
Agentic AI: Multi-Agent Systems, Function Calling, Tool Orchestration, Prompt Engineering, LLMs
Bioinformatics: scRNA-seq, snRNA-seq, Spatial Transcriptomics, CRISPR Screens, Scanpy, Squidpy, AnnData, STAR, bedtools
Infrastructure: AWS, GCP, Docker, Singularity, Cromwell/WDL, Git, CI/CD, HPC