Siddharth Sabata

siddharth@sabata.net | (408) 805-0692 | siddsabata.com

Education

Carnegie Mellon University — Pittsburgh, PA
Master of Science, Quantitative Biology and Bioinformatics — Dec 2025
Honors Thesis: *Learning Tumor Evolution with Diffusion Models*
Advisor: Dr. Russell Schwartz

University of California, Santa Barbara — Santa Barbara, CA
Bachelor of Science, Statistics and Data Science — Jun 2024
Advisor: Dr. Jea-Hyun Park

Publications

Sabata, S. and Schwartz, R. "Learning Tumor Evolution with Diffusion Models." *BioRxiv* (preprint, expected December 2025)

Sabata, S. and Iyer, P. "Multi-omic Latent Space Prediction of Gout in UK Biobank Participants." *BioRxiv* (preprint, expected December 2025)

Sabata, S. et al. "Addressing Background Genomic and Environmental Effects on Health Through Accelerated Computing and Machine Learning: Results from the 2025 Hackathon at Carnegie Mellon University." *BioHackrXiv*, 4 June 2025.

Research Experience

Graduate Research Assistant — Pittsburgh, PA
CMU Computational Biology Department, Schwartz Lab — Aug 2024 – Present
Learning Tumor Evolution with Diffusion Models

- Identified a computational bottleneck in MCMC-based tumor evolution inference, motivating a shift toward diffusion models for faster, scalable inference.
- Developed end-to-end synthetic data and preprocessing pipelines (GCP, Docker, Slurm, Python) and designed graph-based mutation-aware tumor trees.
- Adapted and trained discrete graph diffusion model (DiGress) for conditional tumor phylogeny generation, demonstrating feasibility of diffusion-based clonal deconvolution.

Mase-phi HPC

- Scaled biomarker inference workflow into an HPC-compatible pipeline for graph-driven longitudinal cancer monitoring (100+ patients).
- Parallelized MCMC inference across distributed nodes, achieving >10x throughput improvements.
- Built a robust concurrent data-processing framework with YAML orchestration, redundancy checks, and structured logging.
- Mentored undergraduate researcher on HPC workflows, ML fundamentals, and project sustainability.

Research Assistant — Santa Barbara, CA
UCSB Department of Mathematics — Jun 2023 – Jun 2024

- Applied sparse modeling techniques with NumPy and scikit-learn to estimate PDE parameters, enabling analysis of phase separation and heat transport.

Research Assistant — Santa Barbara, CA
UCSB Department of Molecular, Cellular, and Developmental Biology — Mar 2021 – Jun 2022

- Developed Python bioinformatics tools (BioPython) for CRISPR Cas9 guide-RNA design and gene data extraction, improving molecular research workflow.

Work Experience

Machine Learning Research Intern — (Remote) New York, NY
Sedona Health — Aug 2025 – Present

- Built adaptive health scoring algorithms for precision-health platform integrating biomarkers, blood work, and wearable data.
- Initiated collaborations to develop multimodal time-series transformer model integrating disease history for enhanced injury prediction.

Computational Biology Intern (Remote) New York, NY
Envisagenics Jun 2025 – Present

- Fine-tuned Gaussian Mixture Model (GMM) for RNA transcript grading using Microsoft Fabric and Plotly; achieved 71% reduction in false positives.
- Enhanced metric collection via custom GMMs in scikit-learn and PySpark, with visual analytics through Plotly dashboards.
- Built reusable benchmarking platform evaluating 6000+ GMMs, reducing analysis time by 3x.

Projects

Population-Specific Multiomics Graph Analysis of ACE Protein Expression Mar 2025
Machine Learning & AI Approaches to Multimodal Problems in Computational Biology Hackathon

- Led development of scalable multimodal graph framework integrating pQTL and genome annotations.
- Transformed NetworkX prototype into PyG pipeline in 48 hours; won *Most Innovative Project* and lead-authored a BioHackrXiv paper.

Medical Reasoning with Distilled Models Jan 2025 – May 2025
Language Technologies Institute, CMU

- Fine-tuned DeepSeek-R1-Distill-Llama-8B for medical reasoning, showing supervised fine-tuning enhances diversity of solutions.
- Built HPC fine-tuning pipeline using QLoRA with automated accuracy and pass@k evaluation.

Relevant Courses

Machine Learning: Deep Learning, Statistical Machine Learning, Language Model Inference
Mathematics: Discrete Mathematics, Linear Algebra
Statistics: Stochastic Processes, Regression Analysis, Time Series, Bayesian Inference

Technical Skills

Programming: Python, Bash, SQL
Libraries: NumPy, scikit-learn, PyTorch, PyTorch Geometric (PyG), Hugging Face
Toolkits: PySpark, Git, Docker, Slurm, Microsoft Fabric

Presentations

Ray & Stephanie Lane Computational Biology Department Annual Retreat, CMU, Pittsburgh, PA, Mar 2025.
S. Sabata, R. Schwartz. "Modeling Tumor Clonal Evolution through Longitudinal ctDNA Profiling" (poster).

UCSB Mathematics Directed Reading Program Poster Presentation, UC Santa Barbara, Jun 2023.
S. Sabata. "An Introduction to Hamiltonian Monte Carlo Methods" (poster).

11th Annual Southern California Systems Biology Symposium, UCLA, Apr 2022.
S. Sabata, C. Qiu, N. Jones, S. Pecchia, M. Wilson. "Computational Techniques for Large-Scale CRISPR-guide RNA Molecule Generation" (poster).