Tanaya Datar

Curriculum Vitae

Email | Website | LinkedIn | Research Articles | Github

Education

University of Toronto, St. George — Toronto, ON
Honours BSc in Bioinformatics & Computational Biology, Minor in Computer Science & Immunology — *Sept 2022 – Apr 2026*

- **Relevant Coursework:** Algorithm Design, Analysis & Complexity; Data Structures and Algorithms; Software Tools & Systems Programming; Computational Genomics and Bioinformatics; Computational Systems Biology

Machine Learning Specialization — DeepLearning AI, Stanford
Built & deployed supervised/unsupervised machine learning models in Python (NumPy, scikit-learn, TensorFlow) — *Jun 2025 – Oct 2025*

Research Experience

Bioinformatics Research Associate – Pai Lab — Sept 2025 – Present
Ontario Institute for Cancer Research (OICR) — *Toronto, ON*

- Engineered and maintained end-to-end R pipelines for single-cell sequencing data across 7+ datasets (1.1M+ cells), including neurodevelopmental data from the mouse cerebellum and medulloblastoma tumour scRNA-seq data, building custom QC filters, batch-effect correction, and dimensionality-reduction steps to turn raw sequencing output into analysis-ready data
- Filtered low-quality reads/cells, corrected batch effects, and normalized expression counts; applied statistical and differential expression analysis to surface transcriptomic signatures tied to neural cell-lineage identity, tumor biology, and clinical outcomes
- Curated sequencing datasets from public repositories and collaborators on OICR's HPC cluster, scheduling and resource-managing (memory, CPU, walltime) high-throughput compute jobs; prepared datasets and figures for external collaborators contributing toward peer-reviewed publications

Computational Biology Intern – Goeva & Agrawal Labs — May 2025 – Sept 2025
Dept. of Molecular Genetics | Dept. of Ecology and Evolutionary Biology, University of Toronto — *Toronto, ON*

- Built reproducible pipelines converting raw sequencing data (FASTQ) through alignment (STAR) and quantification (SAMtools, DESeq2) into standardized, analysis-ready formats using R and Bash scripting on Linux/UNIX systems
- Processed bulk RNA-seq data across 246 samples and 35,000+ transcripts – applying QC filters, alignment, and normalization to produce high-integrity datasets for downstream statistical analysis
- Validated the performance and reliability of bioinformatics algorithms through statistical testing, assessing robustness and sensitivity of pipeline outputs on large genomic datasets

Bioinformatics Research Associate – Provart Lab — Sept 2024 – Apr 2025
Dept. of Cell & Systems Biology, University of Toronto — *Toronto, ON*

- Designed a SQL database and developed data integration algorithms in Python and R to merge 86 genome-wide datasets and multiple public databases into a unified, query-able gene regulatory network
- Built and maintained a version-controlled bioinformatics software tool (HTML, JavaScript) for processing and distributing transcription factor network data, integrating with AIV2/ePlant databases

Projects

SeedBench-Bio | *Python, Prompt Engineering, LLM Evaluation* — GitHub

- Built an evaluation benchmark to separate prompt design from raw model capability in AI-assisted bioinformatics code review, seeding real RNA-seq, single-cell, and ML disease-prediction pipelines with hand-curated errors as ground truth

- Benchmarked 5 prompting strategies across code review and scientific claim validation tasks; found that anchoring prompts to prior "expert" agreement weakened reasoning depth and evidence-checking, while plain prompts without added context performed as strongly as the baseline

EEG Seizure Detection Pipeline | *Python, Machine Learning, Signal Processing* GitHub

- Built a seizure-detection pipeline on 39K+ labeled EEG windows (5 CHB-MIT patients, 1.6% class imbalance), engineering 330 time/frequency/entropy/Hjorth features and benchmarking Logistic Regression, Random Forest, and XGBoost; a 110-feature entropy+Hjorth subset raised F1 from 0.539 to 0.613
- Identified patient-level data leakage in naive random-split evaluation and implemented leave-one-patient-out cross-validation, revealing a 0.42–0.56 F1 gap vs. honest evaluation; closed part of it via patient normalization and tuning, raising real-world F1 from 0.334 to 0.489

Medicare Genomic Testing | *SQL, Power BI, Excel, Python* GitHub

- Built an end-to-end analytics pipeline in SQL Server, Excel, and Power BI analyzing 7 years (2018–2024) of CMS Medicare claims alongside CDC cancer incidence data to study geographic access to genomic testing
- Normalized testing rates per 100,000 beneficiaries and corrected for lab-billing distortions in the raw data, uncovering a 10x geographic reporting anomaly in Wisconsin and flagging testing access gaps in high-cancer-burden states such as Kentucky, Ohio, and Nebraska

BMP7 Signaling in Medulloblastoma | *R, Bioconductor, Cytoscape* GitHub

- Recreated and extended the analysis behind Ohata et al. (2025) on BMP7-driven oncogenic signaling in medulloblastoma: curated, cleaned, and normalized bulk RNA-seq data (GSE229150) in a Dockerized analysis environment and performed differential gene expression analysis
- Conducted thresholded and non-thresholded gene set enrichment analysis (GSEA) on DE results and built an enrichment map in Cytoscape to visualize and interpret dysregulated pathways

ExprCompareR | *R, Shiny, tidyverse, ggplot2* GitHub

- Sole developer of an open-source R package and Shiny app automating cross-tissue RNA and protein expression integration using GTEx and Human Protein Atlas data.
- Built a reproducible statistical workflow calculating Spearman rank correlations and flagging RNA/protein outlier genes across cancer versus normal tissue.

Publications

Co-Authored Manuscript – Dr. Vijay Ramaswamy Lab – Developmental Cell *(In Press)* October 2026

- Applied statistical modeling to single-cell gene expression data across tumor datasets to characterize developmental origins of ependymomas and medulloblastomas.

Co-Authored Manuscript – Dr. Nicholas Provart Lab – Genetics *(In Press)* January 2027

- Methods-focused manuscript documenting computational tools and pipelines for integrating and visualizing plant regulatory interaction networks.

Posters & Presentations

Building a single-cell atlas of the mouse cerebellum to uncover developmental origins of Group 3 and Group 4 medulloblastoma Aug 2026 | Presentation

Medical Biophysics Department, UofT; Ontario Student Neuroscience Conference

- Integrated single-cell sequencing datasets across developmental stages to study tumor lineage relationships and clinical behavior

Awards & Recognition

1st Place – Life Sciences Case Competition | *WISE 14th Annual Conference*

- Competed against university teams nationwide to develop a data-driven disease outbreak prediction framework for the Zika virus; proposed an ML-based anomaly detection pipeline using FFT signal processing to identify disease-carrying mosquitoes from wingbeat frequency signatures

Clinical Research Knowledge

Authored articles on clinical trial processes, covering clinical trial management and product development, Good Clinical Practice (GCP), regulatory inspections and quality audits, Risk-Based Monitoring (RBM) methodologies, and electronic Trial Master File (eTMF) management. Clinical Trial Management — AI in Clinical Trials & GCP

Leadership & Community Engagement

Mentor – Cell & Systems Biology Student Union (CSBSU) Sept 2025 – Apr 2026
University of Toronto *Toronto, ON*

- Coached undergraduate mentees in Cell & Systems Biology program through regular one-on-one mentoring sessions on course selection, research opportunities, and career development; fostered academic and professional growth by building mentee independence and decision-making skills

Outreach and Partnerships Coordinator – Biotechnology & Computer Science Association Sept 2025 – Apr 2026
University of Toronto *Toronto, ON*

- Developed and maintained relationships with guest speakers, faculty, researchers, and industry professionals to support BACSA's initiatives; coordinated judges and panelists for annual Biotech & Computer Science Conference
- Facilitated collaborations between students and industry, providing hands-on experience, networking opportunities, and exposure to cutting-edge research in biotech and AI

Tutor – Saturday Program, University of Toronto Jan 2025 – April 2025
University of Toronto *Toronto, ON*

- Tutored Math and Chemistry to a Grade 10 high school student through personalized one-on-one sessions, providing academic support and guidance on higher education and career opportunities; adapted teaching style to individual learning needs

Volunteer Note Taker – Accessibility Services Sept 2025 – Apr 2026
University of Toronto *Toronto, ON*

- Compiled and provided comprehensive lecture notes for students with accessibility needs in courses like Data Structures & Algorithms, Immunology, and Biochemistry; demonstrated technical accuracy and clear scientific communication

Technical Skills

Bioinformatics & Genomics: Single-cell RNA-seq analysis (Seurat, Scanpy), Bulk RNA-seq & differential expression analysis, sequence alignment & genomic pipelines, STAR, SAMtools, DESeq2, MultiQC, NCBI & public genomic databases
Machine Learning & AI: Supervised/unsupervised learning, feature engineering & model evaluation, dimensionality reduction (PCA, UMAP), classification & clustering, scikit-learn, XGBoost, TensorFlow, LLM evaluation
Data Analysis: Data cleaning, transformation & mining, SQL & relational data analysis, Python (pandas, NumPy), Excel & Power BI, statistical hypothesis testing & data visualization
Programming & Software: Python 3, R, Java, C, JavaScript, HTML/CSS, Bash/Shell scripting, Git/GitHub, Docker, NextFlow, Linux/UNIX, Jupyter Notebook & VS Code
Collaboration & Communication: Cross-functional collaboration with PIs and wet-lab scientists, manuscript & presentation preparation, scientific literature review, reproducible documentation