

Shizhao (Joshua) Yang

+1 206 688 9210 | shizhao.yang.gr@dartmouth.edu | <https://sz-yang.github.io/JoshuaYang.github.io/>

Research Interests: Develop statistical and machine (deep) learning methods to solve problems in computational genomics and system biology, with a focus on developmental study, transcriptional regulation, spatial transcriptomics, and multi-omics analysis (at the single-cell level).

EDUCATION

Dartmouth College, Hanover, NH Aug 2025- present
Ph.D. in Quantitative Biomedical Science (1st Year)

University of Washington, Seattle, WA Sep 2023-Mar 2025
M.S. in Biostatistics (Modeling and Methods pathway)

- Overall GPA: 3.9/4.0
- Core Courses: Statistical Inference, Nonparametric Regression, Theory of Linear Models, Stochastic Modeling

New York University Shanghai Sep 2019- May 2023
B.S. in Data Science (Concentration: Genomics; Minor: Mathematics)

- Core Courses: Math Modeling, Numerical Analysis, Ordinary Differential Equation, Introduction to Optimization and Mathematical Programming; Machine Learning, Database, Data Structure; Bioinformatics and Genetics, Transcriptomics
- Awards: Dean's List for Academic Year (2021-2022, 2022-2023)

New York University, New York, NY Jan-Dec 2022

- Study away program, GPA: 3.96/4.0

PUBLICATIONS

- **Yang, S. J.,** Wang, Y., & Lin, K. Z. (2024). LCL: Contrastive learning for lineage barcoded scRNA-seq data. *bioRxiv*. <https://doi.org/10.1101/2024.10.28.620670> (to be published in Proceedings of Machine Learning Research (PMLR) 2026)

POSTER PRESENTATIONS

- **Yang, S. J.,** Wang, Y., & Lin, K. Z. (2026). LCL: Lineage-aware contrastive learning for scRNA-seq data. [Poster presentation]. *Forty-Third International Conference on Machine Learning (ICML 2026)*, Seoul, Korea, July 2025.
- **Yang, S. J.,** Wang, Y., & Lin, K. Z. (2024). [Contrastive learning for lineage barcoded scRNA-seq data](#). [Poster presentation]. *Machine Learning in Computational Biology (MLCB) Conference*, Seattle, WA, September 2024.
- Dong, Q., Wu, Y., Boskovic, A., **Yang, J.,** Yu, Z., Carvajal Velez, L., Gashi, N., Kauv, S., Bao, M., Jiang, J., Li, R., & Wakefield, J. (2024). Small Area Estimation using MICS Data – A Methodological Approach. [Poster presentation]. *Children's Lives: International Conference on Children and Their Families Using the Multiple Indicator Cluster Surveys (MICS)*, Bristol, United Kingdom, September 2024. (delivered by Jon Wakefield)

RESEARCH EXPERIENCE

Contrastive Learning for Lineage Barcoded scRNA-seq Data Seattle, WA, Nov 2023-April 2026
Main Contributor *Supervisor: Kevin Z. Lin, University of Washington; Yixin Wang, University of Michigan*

- Designed and implemented a contrastive deep learning algorithm to learn low-dimensional embeddings for single-cell data, facilitating the identification of cell fate-determining signatures
- Implemented the Calinski-Harabasz (CH) Index, KNN classification, and UMAP visualization and also built a Multi-label neural network classifier to comprehensively evaluate the performance of our contrastive learning model in capturing lineage-specific signals and compared with other published methods.
- Applied local autocorrelation analysis to investigate genes correlated with the model's embeddings, aiding in the identification of gene pathways involved in cell fate determination.

Small Area Estimation using MICS (Multiple Indicator Cluster Surveys) Data Seattle, WA, Jan 2024 – April 2026
Researcher *Supervisors: Jon Wakefield, University of Washington*

- Implemented small area estimation techniques to derive subnational estimates from MICS survey data.

Shizhao (Joshua) Yang

+1 206 688 9210 | shizhao.yang.gr@dartmouth.edu | <https://sz-yang.github.io/JoshuaYang.github.io/>

- Conducted spatial analysis on large-scale survey datasets to evaluate critical child health indicators, including neonatal mortality rate (NMR), vaccination coverage, and stunting, across diverse regions such as Sierra Leone, Gambia, and Laos.

Python-based RNA-seq Analysis Algorithm using Negative Binomial GLM

New York, USA, June-Dec 2022

Research Assistant

Supervisor: Manpreet Katari, New York University

- Implemented negative binomial GLMs in Python for differential gene expression (DGE) analysis.
- Implemented backtracking line search and IRLS for efficient dispersion and coefficient estimation and applied the Wald Test to the estimated log fold changes.
- Exploited Python's multiprocessing capacity to shrink the overall runtime from 3 hours to about 3 minutes.
- Tested the algorithm on several different datasets and compared the results with actual and estimated values generated by Deseq2.

SRPMS (Student Research Program in Molecular Science)

Investigation of Horizontal Transfer in Metagenomics

Shanghai, China, June 2021-Jan 2022

Research Assistant

Supervisor: Gang Fang, NYU Shanghai, NYU

- Implemented RNA-seq Analysis on Next-Generation Sequencing data of human gut microbiome, including *de novo* genome assembly, mapping, gene-calling, and annotation using HPC (high-performance computer) and related pipelines.
- Implemented the Louvain Method to generate pseudo ortholog communities by comparing the similarity between each gene.
- Calculated TPM (transcripts per million) and our own index of gene persistence and identified correlations between TPM and our index.

Refined SIR Model and its Application in 2022 NYC Influenza A Prediction

Sep 2022-Dec 2022

Intro to Math Modeling

New York, USA

- Implemented ODE methods (including fixed point models and their stability, phase plane, and herd immunity) to fit the modified SIR model to account for vaccination.
- Simulated six years of NYC influenza data under the SIRV model and estimated the transmission rate, removal rate, and reproduction number (R_0) using the Quasi-Newton method on Python.
- Compared past seasons' estimated coefficients and used them to predict the infection peak in NYC of year 2022

LEADERSHIP & MENTORSHIP EXPERIENCE

- ASA DataFest 2024 Mentor, Department of Statistics, University of Washington Mar 2024
- Teaching Assistant, BIOST 310 Biostatistics for the Health Sciences, University of Washington Jan-Mar 2025

STANDARDIZED TESTS

- GRE: 330+3.5 (Verbal 160, Quantitative 170, Analytical Writing 3.5) Aug 2022
- TOEFL: 111 (Reading 30, Listening 30, Speaking 24, Writing 27) Mar 2025

LANGUAGES & PROFESSIONAL SKILLS

- Chinese (native), English (fluent)
- Programming Languages: Python (Pytorch), R, Linux (Shell), SQL, MongoDB, HTML, MATLAB, Javascript
- Analytics Skills: Statistical modeling, Numerical analysis, Machine Learning, Deep Learning
- Bioinformatics Applications: SPAdes, Bowtie2, Hisat2, Samtools, diamond, Blast, Limma, Deseq2