Yifan Lin

Seattle, WA, 98125 | +1 (858)-888-1236 | yifan121100@gmail.com | https://yifanlinz.github.io/ |Pronouns: she/her/hers

Education

University of Washington (UW), Seattle, WA **Sep 2024 – Present**

- **Master of Science**, Biostatistics • **GPA**: 3.720
- **Relevant Courses**: Statistical Learning, Statistical Methods for Omics Data, Gene Sequence Analysis, Theory of Statistical Inference, Biostatistics Methods, Longitudinal Data Analysis, Survival Data Analysis.
- **Research-Oriented Course Projects**: Amyloid Beta Proximity and Microglia Activation Dysregulation in Alzheimer's Disease, scRNA-seq Sequencing Matrix Normalization Methods, Mammography Accuracy Investigation for Breast Cancer.

University of California San Diego (UCSD), La Jolla, CA **Sep 2018 – Jun 2022**

- **Bachelor of Science**, Bioengineering: Bioinformatics • **GPA**: 3.725
- **Honors:** Provost Honors, Warren College Honor Society
- **Relevant Courses**: Calculus, Linear Algebra, Advanced Probability and Statistics, Machine Learning, Design and Analysis of Algorithm, Modeling and Computation in Bioengineering, Genetics, Genomics, Molecular Biology, Organismic and Evolutionary Biology, Next-Generation Sequencing (NGS) Data Analysis, Biological Databases.
- **Research-Oriented Course Projects**: SIR Model for Assessing SARS-CoV-2 Omicron Variant Infection, PacBio HiFi Sequence Assembly and Alignment using Dynamic Programming, Multi-omics NGS Data Analysis.

Publications & Conferences

[1] **Y. Lin**, K. Lin. **Quantifying Confounding Effect of Co-pathologies in Neurodegenerative Diseases with *sensGAN***. *Machine Learning in Computational Biology 2025*. 2025 Sept 10 – 11. New York, NY, U.S. Add links to personal website

[2] Z. Li, **Y. Lin**, K. Lin. ***MIRAGE*: Manifold-Informed Gene-module Extraction for Disentangling Simultaneous Dynamics in Single-Cell RNA-seq**. *Machine Learning in Computational Biology 2025*. 2025 Sept 10 – 11. New York, NY, U.S. Add links to personal website

[3] G. Di Caro, ET. Lam, **Y. Lin**, *et.al*. **Demonstration of the Prognostic Value of Distinct Morphological CTC Subtypes in Metastatic Breast Cancer Patients Undergoing Late Lines of Treatment** (Submitted to *The American Association for Cancer Research* in 2024 Oct)

[4] **Y. Lin***, B. Yuan*, D. Murali, *et.al*. **Regulation of Zinc Fingers and Resulting Functional Directions in HuVECs Under Shear Stress** (Submitted to *Proceedings of the National Academy of Sciences of the United States of America* in 2023 Oct)

[5] **Y. Lin**, E. Johnston. **Detection of multiple protein-protein interactions in parallel in vivo using designed synthetic oligonucleotides**. Bioengineering Day 2022. 2022 April 22. San Diego, CA, U.S. Add links to personal website

Relevant Experience

Independent Study, Lin Lab, UW **Oct 2024 – Present**

Advised by **Prof. Kevin Z. Lin**, I proposed ***sensGAN***, a generative adversarial network guided by sensitivity analysis, to model unmeasured confounding in omics research of complex diseases. In Alzheimer's disease (AD) pathogenesis, for example, microglial cell states are biologically critical but often confounded by age and environmental exposures, making it difficult to isolate AD-specific effects. By constraining the predictive gains of latent confounder to be no greater than the strongest observed covariate, *sensGAN* facilitates the **identification of resilient and consistent biomarkers in complex biological systems**.

- Developed predictive gain metrics based on deviance-based partial R^2 (DBPR) to quantify strengths of latent confounding to predictions of disease status and gene expression, which retain statistical interpretability grounded in generalized linear models (GLMs).
- Translated GLMs into neural network (NN) predictors of disease status and gene expression, preserving GLM links and interpretability while enabling nonlinear flexibility; coupled with the generator network, this framework identified the most powerful latent confounder.
- Designed and implemented sens*GAN* that learned a series of interpretable confounders, whose predictive gains were constrained by that of the most powerful one, revealing varying robustness of differentially expressed genes (DEGs) to unmeasured confounding.
- Validated *sensGAN* on the Lupus PBMC pseudobulk RNA-seq dataset, where the method successfully recovered latent confounders correlated with withheld clinical covariates after confounding effect calibration through predictive gains.

- Applied *sensGAN* to the Seattle Alzheimer's Disease Brain Cell Atlas (SEA-AD) pseudobulk RNA-seq dataset, revealing (i) AD-associated genes that remained significant even under maximal confounding were strongly enriched for microglial mobility, and (ii) genes nullified by the latent confounder were enriched for pathways linked to Lewy body dementia and bipolar disorder, unrelated to core AD pathology.

Independent Study, Jayadev Lab, UW **Mar 2025 – Present**

Advised by **Prof. Suman Jayadev**, I applied fastTopics, a probabilistic topic model, to **characterize microglial functional heterogeneity on a continuous spectrum**. This approach represents each cell as a mixture of functional modules, enabling the discovery of continuous microglial states. In both unsorted brain samples and microglia-specific datasets, the model recovered biologically coherent programs aligned with known cell types.

- Implemented the model, evaluated diagnostic and convergence criteria, and tuned hyperparameters for stability and interpretability.
- Validated the applicability of the framework through a proof-of-concept analysis on unsorted single-cell RNA-seq data, identifying biologically meaningful functional modules and DEGs corresponding to various cell types (e.g. AQP4 for astrocytes, CD45 for microglia).
- Revealed six biologically meaningful topics representing microglial functional modules (e.g. phagocytosis), aligned with previously reported microglial clusters, and are currently assessing their associations with clinical phenotypes such as Alzheimer's Disease Neuropathologic Change (ADNC) and Endolysomoal network polygenetic risk score (ePRS).

Data Analyst, Epic Sciences, Inc. **Aug 2022 – June 2024**

I contributed to the **Bioinformatics components of comprehensive liquid biopsy** tests to characterize **breast cancer** based on NGS data of circulating tumor DNA (ctDNA) and circulating tumor cells (CTCs).

- Constructed a data analysis pipeline to clinically validate the assay and the computational pipeline, enabling the comprehensive liquid biopsy test to detect clinically relevant ctDNA variants with 96%+ accuracy.
- Evaluated the application of circular binary segmentation (CBS) algorithm in gene copy number estimation with simulated genomes of different complexities, demonstrating method limitations in highly localized ploidy regions of metastatic breast cancer patients.
- Adapted the *XGBoost* cancer classification model to the novel breast cancer subtype classification model based on ctDNA variants and imaging features, leveraging one-hot encoding, imputation, and feature engineering for model performance improvements of 19%.
- Established data engineering frameworks with automated pipelines to transfer, analyze, and visualize large-scale data, optimizing data towards usability in diverse contexts of genomics assay development, translational research, and commercialization.

Research Associate, Epic Sciences, Inc. **Mar 2023 – Oct 2023**

I worked with translational research scientists from Epic Sciences, Inc. and Baylor University Medical Center to study the **survival of metastatic breast cancer (mBC) patients** undergoing late lines of treatment with distinct morphological CTC subtypes. Specifically, I established a data compilation pipeline and refined the survival models with the compiled data, improving analysis power to understand roles of CTCs in survival. The manuscript was submitted to *AACR*.

- Refined CTC survival analysis methods by data censoring, discovering a significant association between small-cell/neuroendocrine-cell CTCs and survival, which proved the feasibility of using CTCs for mBC patient prognostics.
- Developed a data compilation pipeline and automated treatment annotation with the ChatGPT API, aggregating clinical data and genomics biomarkers of multiple sources, correcting 30+ typing errors, 27 manual annotation mistakes, and improving annotation accuracy by 23%, resolving the long-standing data inconsistency issues.

Independent Study, Subramaniam Lab, UCSD Bioengineering **July 2020 – Oct 2022**

Mentored by **Prof. Shankar Subramaniam**, I co-led bioinformatics analyses and lab experiments to understand the **role of Zinc finger (ZNFs) transcription factors in the Endothelial-Mesenchymal Transition (EMT)** using human umbilical vein endothelial cells (HuVECs).

- Developed the bioinformatics analysis pipeline to analyze the time series pseudobulk RNA-seq data identifying differentially expressed genes (DEGs), enriched functional pathways, and the cell-state trajectories in the baseline-anchored model.
- Led the time-series regulatory network analysis based on functional enrichment results, with a specific focus on ZNF genes, identifying ZNF-driven modules involved in morphological remodeling and cytoskeletal organization, highlighting EGR1 as a key master regulator.
- Discovered the differential clonal cell states in the time series scRNA-seq data analysis to refine the regulatory networks proposed based on RNA-seq data. Co-led manuscript writing for a publication submitted to *PNAS.*

Teaching Experience/Service

Tutor, Statistics Tutor & Study Center, UW **Sep 2024 – Present**

- Support students in understanding statistical concepts, completing coursework, and developing R programming skills.
- Tutor students from diverse backgrounds, ranging from public health, science, and social science, in an accessible, context-specific way.

Equity, Diversity, and Inclusion (EDI) Committee Member, Biostatistics Department, UW **Sep 2024 – Present**

- Co-led "Be Here Now" community gatherings for students, staff, and faculty to foster inclusive dialogue, mutual support, and shared reflection on academic and personal challenges.

Teaching Assistant, BENG 1 (Introduction to Bioengineering), UCSD **Jan 2021 – Mar 2021**

- Designed activities and assignments for over 300 undergraduate students to introduce the topics of Bioseparations and 3D image processing.
- Directly mentored 30+ students to aid them in understanding 8 major Bioengineering technologies and topics.
- Liaised with the course's lead professor to initiate the design of a new module. Introducing Bioinformatics to future student cohorts.

Team Lead, Biomedical Engineering Society, UCSD **Jun 2019 – Jun 2022**

- Connected 50+ students with faculty PIs through structured lab-matching programs to expand undergraduate research participation.
- Promoted departmental Bioengineering town halls to improve access, representation, and inclusion for underrepresented and international students.
- Organized and led technical and professional development workshops, including programming fundamentals, data analysis, and elevator pitching.
- Actively contributed to community outreach programs that communicated core concepts in biomedical sciences to the public, especially younger generations.

Summary

My experiences working with large-scale biomedical datasets have shown that while modern technologies generate rich measurements, our analytical frameworks struggle to disentangle technical artifacts, environmental factors, and true biology. This challenge motivates me to develop Machine Learning methods that are both statistically rigorous and biologically interpretable. Through my Master's thesis, I designed *sensGAN*, a generative adversarial network integrating sensitivity analysis, to investigate latent confounding in complex diseases such as AD. The project has further strengthened my commitment to creating principled computational tools that reveal mechanistic insight from heterogeneous biological data. These experiences strongly motivated my pursuit of a PhD to build interpretable, translational ML models for biomedical big data.

Relevant Skills

- ***Statistical & Machine Learning***: Neural networks, generalized linear models, regularization methods (LASSO, ridge), dimension reduction, clustering, high-dimensional inference, causal inference, simulation studies.
- ***Biostatistics & Biomedical Data Science and Engineering***: Survival and longitudinal analysis, negative binomial and count data modeling, differential expression analysis, next-generation sequencing (NGS) data analysis, single-cell and multi-omics integration, pipeline development and automation, reproducible research workflows.
- ***Computational Tools & Engineering***: Data engineering (PySpark, Databricks), distributed computing, workflow automation, cloud computing (AWS), version control (Git), Linux environments.
- ***Programming technologies:*** Python, R, C, C++, Bash, Java, JavaScript.

References

- Kevin Z. Lin, PhD
  Assistant Professor, Department of Biostatistics
  University of Washington
  Email: kzlin@uw.edu
  Graduate Research Advisor
- Suman Jayadev, MD
  Associate Professor, Department of Neurology
  University of Washington
  Email: sumie@uw.edu
  Research Collaborator, Jayadev Lab
- Ernest T. Lam, PhD
  Senior Director, AI/ML
  Mirador Therapeutics
  Email: ernest.t.lam@gmail.com
  Former Supervisor at Epic Sciences
- Shankar Subramaniam, PhD
  Distinguished Professor, Department of Bioengineering
  University of California, San Diego
  Email: shsubramaniam@ucsd.edu
  Undergraduate Research Advisor