Yucheng Wang

yuw260@pitt.edu | 1826A Wesley W. Posvar Hall, Pittsburgh, PA

Education

University of Pittsburgh — Expected June 2027
Ph.D., Statistics — *Pittsburgh, PA*

Carnegie Mellon University — Aug. 2021 – May 2022
M.S. in Statistical Practice, GPA: 4.05/4.00 — *Pittsburgh, PA*

Southern University of Science and Technology — Aug. 2017 – June 2021
B.S. in Statistics, GPA: 3.76/4.00 — *Shenzhen, Guangdong*

Work & Internship Experience

Bioinformatics Software Engineer Intern — May. 2025 – Aug. 2025
SignatureDx — *Pittsburgh, PA*

- Collaborated with AstraZeneca to build a **machine-learning–driven**, cell-free DNA deconvolution framework for monitoring organ-transplant status and immune rejection.
- Developed a **Bayesian tree-based probabilistic model** to construct methylation signatures that differentiate highly similar cell types, with model parameters estimated via an novel **tree-based EM algorithm**, achieving 30% lower RMSE than existing methods.
- Designed a likelihood-based fragment-level deconvolution algorithm that accounts for CpG-site correlations and integrates genotype information, employed a **pseudo-likelihood approach** to efficiently compute likelihoods of DNA fragments.
- Achieved a 0.2% limit of detection for deconvolution of highly similar cell types, demonstrating advanced **statistical estimation**, **feature engineering**, and model sensitivity analysis.

Algorithm Engineer Intern — May. 2020 – Sep. 2020
JD.com, Label R & D Dept. — *Beijing*

- Applied MySQL, HIVE, and PySpark to build large-scale data pipelines and feed-style recommendation algorithms in a **big-data environment**.
- Implemented TF-IDF–based content filtering to model item text features and enhance candidate generation using **natural language processing (NLP)** techniques.
- Developed user–item **collaborative filtering** models and matrix-factorization approaches (SVD) for personalized recommendations.
- Built a **LightGBM (XGBoost) gradient-boosted decision tree** classification model with feature selection for JD.com's first drug-procurement app's recommendation system.
- Executed candidate generation and ranking, performed **online A/B testing**, tuned hyperparameters, raised recommendation accuracy by 15%.
- Monitored 6.18 festival traffic and designed fallback strategies to maintain system stability under heavy load.

Skills and Courses

Programming: R, Python, C++, Java, MySQL, PySpark, SAS, SLURM, Linux/Unix Shell.
Related courses: Data Structure and Algorithm, Computational Statistics, Machine Learning, Multivariate Statistical Analysis, Generalized Linear Model, Applied Linear Model, Survival Analysis, Statistical Genomics and High Dimensional Inference, Mathematical Analysis, Linear algebra, Measure Theory, Real Analysis.

Research and Projects

Research Project (Under review at Nature Communications) — Nov. 2023 – Aug. 2025
A Novel Cell-free DNA Analysis Framework cf-TREBLE — *Pittsburgh, PA*

- Proposed cf-TREBLE, a rigorous **Bayesian statistical** framework for large-scale methylation-data analysis and noninvasive disease diagnostics.
- Developed a novel **Bayesian tree-based hierarchical model** to estimate cell-type–specific methylation signatures, identifying more informative marker CpG sites than existing methods.
- Applied the BLEND deconvolution algorithm with individualized priors to model **inter-subject biological variability** and generate personalized methylation signatures.

- Reimplemented all core methods in **C++** with optimized memory management and multi-threading, improving runtime by 70% over the R prototype.
- Achieved a 30% reduction in RMSE for estimating cell-type proportions across 47 cell types; demonstrated high sensitivity for low-abundance cell types and an AUC of 0.83 for adenomyosis diagnosis, while identifying key cell-type biomarkers for endometriosis.

Research Project Mar. 2024 – Present

Fragment-based Deconvolution Algorithm *Pittsburgh, PA*

- Developed a likelihood-based fragment-level cfDNA deconvolution method capturing CpG correlations via a **spatial probit model** with a Matérn kernel.
- Estimated genome-wide correlation structures across 200+ tissue references and multiple disease cohorts, incorporating complex fragment- and subject-level dependencies.
- Implemented efficient **Gibbs sampling** with conjugate Bayesian probit regression using unified skew-normal distributions, enabling parallel computation on **high-performance clusters** and reducing runtime by over 40%.
- Validated the model on large-scale simulated and clinical cfDNA datasets, demonstrating improved sensitivity in detecting low-abundance cell types and strong potential for non-invasive diagnostic prediction.

Collaborative Project with Allegheny County Health Department Feb. 2022 – May 2022

Automated Weather Forecasting Pipeline *Pittsburgh, PA*

- Built a automated web-scraping and ETL pipeline for continuous weather-data acquisition and cleaning.
- Developed predictive weather-forecasting models using **Random Forest, XGBoost, and Time Series**, integrating ARIMA terms to improve forecast accuracy.
- Applied cross-validation and calibration against official forecasts, achieving high predictive agreement.
- Deployed a fully automated workflow with GitHub Actions for data scraping, training, and daily reporting.

Research Project Aug. 2023 – Dec. 2024

Cell-free DNA Methylation Analysis for Endometriosis (Minor Revision) *Pittsburgh, PA*

- Analyzed cfDNA methylation profiles from 637 endometriosis cases and 347 controls, including matched lesion–endometrium pairs.
- Conducted **factor analysis** using reference-free deconvolution (**BCconf**) to identify latent factors associated with endometriosis; a key factor (Factor 11) reflected immune and endothelial cell variation ($p_{adj} < 10^{-4}$).
- Performed large-scale multiple testing with **FDR control** to detect shifts in epithelial, stromal, immune, and vascular cell populations (FDR<0.05), identifying potential **biomarkers** for minimally invasive diagnosis.

Collaborative Research Project with Pitt School of Medicine Aug. 2023 – Jan. 2024

Admixture and Survival Analysis of Two Stages of Heart Failure *Pittsburgh, PA*

- Estimated African ancestry proportions in 505 patients via **admixture modeling** and evaluated their association with heart-failure progression.
- Performed detailed **survival analysis** to assess the impact of ancestry, genetic loci, and treatment effects on heart-failure outcomes.
- Conducted a **Genome-Wide Association Study (GWAS)** to identify genetic markers linked to treatment efficacy and mortality risk, contributing to precision-medicine approaches.

Undergraduate Thesis Jan. 2021 – June 2021

A Solution of the Discrete Linear Inverse Problem *Shenzhen, Guangdong*

- Connected the discrete linear inverse problem to **maximum-likelihood estimation** across multiple distributions.
- Replaced the KL-divergence in Vardi & Lee (1993) with the log-likelihood of Poisson and other distributions, solving the problem via **EM** and **MM** algorithms.
- Provided a more computationally efficient approach with successful application to **image restoration**.

Honors & Awards

- Andrew W. Mellon Predoctoral Fellowship 2025-2026
- Excellent Student Scholarship 2^{nd} Prize 2020
- Excellent Student Scholarship 3^{rd} Prize 2019
- Excellent Student Scholarship 3^{rd} Prize 2018
- SUSTech Programming Contest 3^{rd} Prize 2018