Yeonji Ji, Ph.D.

Cliffside Park, NJ | fortunate.yj@gmail.com | 929-810-5531 | LinkedIn | GitHub | Portfolio

Summary

Data Scientist with a Ph.D. in Biochemistry, experienced in computational modeling, Python workflows, and large-scale data analysis. Expanding expertise in machine learning and statistical modeling through applied projects, leveraging research discipline to develop predictive models and reproducible workflows.

Technical Skills

Programming:
Python (pandas, scikit-learn, PyTorch, matplotlib, etc.), Bash, Git, Familiar with SQL and C++

Data Analysis & Modeling:
Supervised/unsupervised learning, Feature Engineering, Ensembles, Deep Learning

Statistical Analysis:
Experiment Design, Model Evaluation, Statistical Mechanics

Data Management:
Workflow Automation, Data Cleaning & Validation, HPC (Linux, Slurm)

Specialized Expertise:
Computater-Aided Drug Discovery

Collaboration & Communication:
Mentoring, Teaching

Education

Ph.D. in Biochemistry
The Graduate Center, CUNY *2018 – 2024*
Dissertation Title:
"Incorporating Solvation Thermodynamic Mapping in Computer-Aided Drug Design"

B.S. in Chemistry
Kyung Hee University, Seoul *2012 – 2017*

Certifications

Python for Machine Learning and Data Science Masterclass (Udemy)
Supervised/unsupervised ML, PCA, model evaluation

Deep Learning Specialization (Coursera)
Neural networks (CNN, RNN, LSTM, etc.), optimization, project structuring

Honors & Grants

CUNY DSRG (2023)
Penny J. Gilmer Grant, OpenEye (2023)
CUNY Science Scholarship (2018–2024)
Superiority Scholarship, KHU (2014–2015)

Machine Learning Projects

Movie Recommendations: MF to Hybrid Ranking (GitHub) *2025*
- Built a hybrid recommendation system with matrix factorization and content features, evaluated with ranking metrics.
- Designed pipelines to convert user behavior data into predictive signals for personalization.

Amazon Review Sentiment Classification (GitHub) *2025*
- Processed 3M+ reviews and applied NLP models (Logistic Regression, Naive Bayes, SVM, XGBoost).
- Identified drivers of user sentiment to inform marketing and product strategies.

Teleco Customer Churn Prediction (GitHub) *2025*
- Developed churn prediction models using Logistic Regression, Random Forest, and XGBoost.
- Applied feature engineering and SMOTE to handle class imbalance and evaluated against KPIs.

Research & Data Projects

Binding Site Prediction from Simulation Data *(Publication)* *2023 – 2025*
- Analyzed large-scale molecular simulation data (time-series, 3D spatial) with statistical models, building reproducible workflows for scalable insights.
- Published and presented findings, contributing actionable results to the research community.

Water Data-driven Pharmacophore Modeling *(In Process)* *2021 – 2024*
- Automated workflows to integrate hydration datasets into scalable pharmacophore models.
- Benchmarked predictive performance across compound libraries, improving efficiency of screening pipelines.

COVID-19 Solvation Mapping Repository *(Publication)* *2019 – 2020*
- Contributed datasets and code to an open-source repository, resulting in peer-reviewed publication.
- Collaborated with researchers to inform rapid public health decisions.

Professional Experience

Postdoctoral Researcher *Lehman College, CUNY* *2024 – Present*
- Building Python pipelines for reproducible workflows while mentoring graduate students and collaborating across teams.

Adjunct Lecturer *CUNY Research Foundation* *2019 – 2024*
- Taught labs to 100+ students, simplifying technical concepts and guiding data analysis to strengthen communication skills.

Cosmetic Chemist Intern *Englewood Lab, NJ* *2015 – 2016*
- Conducted experiments and analyzed formulation data to optimize product performance.
- Collaborated with senior scientists to align results with business and quality standards.