# 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 large-scale computational drug discovery, Python workflows, and reproducible data pipelines. Complemented this research background with applied machine learning projects in NLP, recommendation, and churn prediction, leveraging SQL, ETL, and KPI-based model validation for scalable decision support.

## Technical Skills

**Programming Infrastructure:**
Python (pandas, scikit-learn, PyTorch, matplotlib), SQL, Bash, Git, C++, ETL pipelines, Workflow automation, HPC (Linux, Slurm)

**Data Analysis & Modeling:**
Supervised/unsupervised learning, Feature Engineering, Ensembles, NLP, Recommender Systems, Deep learning, Statistical & mathematical modeling

**Statistical Analysis:**
Experiment Design, Hypothesis testing, Model Evaluation

**Domain Expertise:**
Computater-Aided Drug Discovery, Molecular Simulation

**Collaboration & Communication:**
Stakeholder engagement, Mentoring, Teaching

## Education

**Ph.D. in Biochemistry**
The Graduate Center, CUNY *2018 – 2024*

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

## Certifications

**Python for Machine Learning and Data Science Masterclass (Udemy)**
Covered supervised/unsupervised learning, PCA, model evaluation, and applied methods to build ML pipelines.

**Deep Learning Specialization (Coursera)**
Trained and optimized neural networks (CNN, RNN, LSTM), learning best practices for model structuring and deployment.

## 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 data pipelines to transform user logs into predictive features for personalization.
- Developed a hybrid recommender (MF + LightGBM/XGBoost), evaluated with ranking metrics (Precision@K, Recall@K, NDCG).

**Amazon Review Sentiment Classification (GitHub)** *2025*

- Processed millions of reviews with TF-IDF pipelines and ML models (LogReg, NB, SVM, XGBoost).
- Evaluated against Accuracy, Precision, Recall, F1, AUC, highlighting linear models for sparse text.

**Teleco Customer Churn Prediction (GitHub)** *2025*

- Built churn models (LogReg, Random Forest, XGBoost) with SMOTE/weights for imbalance.
- Improved churn detection against business KPIs, generating insights to support retention strategies and decision-making.

## 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*

- Designed automated data integration pipelines to incorporate hydration datasets into pharmacophore models.
- Benchmarked predictive performance across compound libraries, improving efficiency of screening pipelines.

**COVID-19 Solvation Mapping Repository** *(Publication)* *2019 – 2020*

- Contributed datasets and reproducible code to an open-source repository, resulting in peer-reviewed publication.
- Collaborated with team 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*

- Performed formulation experiments and data analysis to optimize product performance.
- Supported senior scientists in ensuring alignment with business and quality standards.