

ZIRUI WANG

Syracuse, NY • (517) 721-9616 • wangzr926@gmail.com • [LinkedIn](#) • [GitHub](#) • [Portfolio](#)

EDUCATION

Syracuse University , Master of Science in Computer Science, GPA: 3.9/4.0	2022 – 2025
Michigan State University , Bachelor of Science in Chemical Engineering, GPA: 3.5/4.0	2015 – 2019

TECHNICAL SKILLS

Programming Languages: Bash/Shell, Python, C++, C, Java, JavaScript, HTML, CSS, SQL, NoSQL, R, Haskell, MATLAB
Technologies: Git, Docker, Kubernetes, Jenkins, AWS, BigQuery, Snowflake, Firebase, Airflow, Spark, Vertex AI, SageMaker
Frameworks: Flask, Django, FastAPI, Streamlit, Spring Boot, React.js, Next.js, Node.js, Tailwind CSS, LangChain, LlamaIndex

EXPERIENCE

Data Scientist Intern Regeneron Pharmaceuticals	May – Aug 2024 remote
<ul style="list-style-type: none">Collaborated with data engineering teams to automate the retrieval and processing of 40k+ publications using Airflow and Docker, resulting in a scalable ETL pipeline that stores model-ready data in AWS S3.Fine-tuned a Hugging Face LLM with Low-Rank Adaptation in TensorFlow to summarize clinical studies, enhancing patient outcome extraction and reducing manual review time by 25%.Conducted document- and entity-level sentiment analysis on 100+ studies using NLTK Named Entity Recognition and a roBERTa model, generating actionable insights on patient experiences and treatment efficacy.	
Software Engineer Co-op Regeneron Pharmaceuticals	May – Dec 2023 Tarrytown, NY
<ul style="list-style-type: none">Developed a Flask application with a RESTful API, Bootstrap UI, and MongoDB/PostgreSQL persistence to parse, conform, monitor, and query legacy data, improving data usability and accessibility by over 80%.Applied domain-driven design and implemented an event-driven architecture using RabbitMQ message broker, reducing technical debt by 30% and enhancing system scalability and resilience.Implemented unit and integration tests using pytest, increasing test coverage by 50% and reducing pre-release defects by 30%.Collaborated with cross-functional teams using Bitbucket for version control alongside Docker, Kubernetes, and Jenkins for CI/CD processes, decreasing deployment time by 40% and improving delivery speed by 30%.Leveraged Jira and Confluence to facilitate Agile Scrum ceremonies, driving continuous improvement through backlog refinement, sprint planning, and reviews, resulting in a 30% reduction in blockers and faster delivery cycles.	
Machine Learning Engineer Institute for Quantitative Health Science and Engineering	Mar 2020 – Jul 2022 East Lansing, MI
<ul style="list-style-type: none">Developed a BART transformer model and a Variational Autoencoder in PyTorch to generate novel protein sequences, achieving a 24% improvement in stability and a 300% increase in sequence library size.Built a predictive model in Scikit-learn by integrating Hierarchical Clustering with a Random Forest classifier to streamline molecule screening, increasing lead yield 4-fold.Led the development of data pipelines using PySpark, NumPy, and Pandas to collect, clean, normalize, and encode biological data, reducing preprocessing time by 40% and improving data accuracy.Communicated results to stakeholders through visualizations and reports, applying PCA and statistical tests to highlight findings.	

PROJECTS

Developer Compensation Estimation	
<ul style="list-style-type: none">Implemented an ETL pipeline to process and store Stack Overflow survey data in a NEON cloud data warehouse, and performed exploratory analysis, anomaly detection, imputation, and feature engineering to prepare data for modeling.Trained and optimized Random Forest regressor, XGBoost, and Ridge Regression models with cross-validation, evaluated via MAE, RMSE, and R², and deployed model to Vertex AI as a production-ready cloud endpoint.	
Movie Recommendation System	
<ul style="list-style-type: none">Built a user profiling pipeline with Latent Dirichlet Allocation to extract latent viewing topics from watch history and behavioral metadata, generating interpretable user preference distributions.Implemented user- and item-based collaborative filtering to recommend movies, blending profile-driven similarity with explicit ratings for improved personalization and cold-start handling.	
Life Science Research Agent	
<ul style="list-style-type: none">Designed and implemented an AI agent integrating Google Gemini with a Model Context Protocol server exposing tool calls to arXiv, ClinicalTrials, OpenFDA, and PDB using Anthropic SDK, enabling tool-augmented reasoning across data sources.Connected FastAPI backend to a Streamlit chatbot UI using asynchronous context management and robust error handling to enable real-time AI interactions.	
PUBLICATIONS	
<ul style="list-style-type: none"><u>Generative Models for Protein Sequence Modeling: Recent Advances and Future Directions.</u> <i>Briefings in Bioinformatics</i>.<u>Phytochemical drug discovery for COVID-19 using high-resolution computational docking and machine learning assisted binder prediction.</u> <i>Journal of biomolecular structure & dynamics</i>.	