4950 N MARINE DR
Chicago IL 60611

yangyang.li@northwestern.edu
+1-507-606-8822

Yangyang Li

yangyangli.top | github.com/cauliyang | linkedin.com/in/yangyang-liz5

Self-introduction

I am a fourth-year Ph.D. candidate specializing in the convergence of deep learning, algorithmic optimization, and computational biology. My research leverages state-of-the-art machine learning architectures and distributed computing systems to decode complex biological phenomena, with a particular focus on genomic sequence analysis. I have developed multiple high-impact tools including a transformer-based genomic language model and a high-performance sequence alignment library, demonstrating my ability to bridge theoretical advances with practical implementations. Having led projects that achieved substantial improvements in accuracy and computational efficiency, I combine strong technical expertise with collaborative problem-solving skills. I am eager to apply my experience in machine learning, algorithm development, and large-scale data analysis to tackle challenging problems in a research-driven environment that values innovation and technical excellence.

Research Experience and Project

Northwestern University Chicago, US

Ph.D. in Computational Biology and Bioinformatics *June 2022 – June 2026*

- Developed and implemented a transformer-based genomic language model featuring hyena operators for long-context processing, achieving 80% improvement in chimera artifact detection accuracy for Nanopore Direct RNA Sequencing data through advanced deep learning techniques
- Architected and optimized PxBLAT, a high-performance Python library that interfaces with BLAT, implementing parallel processing and memory optimization techniques to achieve 20% faster genomic sequence alignments while maintaining accuracy
- Designed and implemented a novel graph-based machine learning algorithm that significantly improved non-linear transcript identification in long-read sequencing data, incorporating attention mechanisms and achieving state-of-the-art accuracy in complex RNA structure analysis
- Created Aurora, a web application for intuitive graph algorithm visualization, reducing algorithm comprehension time and enhancing collaboration among bioinformatics researchers

University of Minnesota Minneapolis, US

Ph.D. in Bioinformatics and Computational Biology *Sep. 2020 – June 2022*

- Innovated a novel graph algorithm that improved detection of non-linear structural variations in transcriptomes, enabling more comprehensive analysis of complex RNA structures and alternative splicing events
- Developed and deployed a transformer-based deep learning model to predict gene fusion-structural variation causality, achieving 85% prediction accuracy and reducing analysis time by 60% compared to traditional methods
- Spearheaded comprehensive benchmarking of 10+ state-of-the-art alternative splicing detection tools, resulting in data-driven tool selection guidelines that improved detection accuracy by 40% across diverse genomic datasets
- Courses (Grade): Advanced Machine Learning (A), Introduction to Data Mining (A), Adv. Algs. & Data (B)

China Agricultural University Beijing, CN

Master in Crop Bioinformatics *Sep. 2018 – June 2020*

- Developed and implemented a machine learning pipeline using ensemble methods and feature selection algorithms to analyze 1,400 maize genomics datasets, identifying novel genetic markers that improved yield prediction accuracy by 45% and enabled targeted crop enhancement strategies
- Spearheaded large-scale genome-wide association study (GWAS) across 450 maize populations, implementing custom statistical models and machine learning approaches to map genetic variations to phenotypic traits, resulting in identification of 25 novel genetic loci associated with improved crop yield

Work Experience

Northwestern University IT Organization Chicago, US
Student Consultant in Data Science, Statistics, and Visualization *June 2024 – now*

- Engineered a social network analysis pipeline using NetworkX to analyze complex collaboration patterns across 1000+ researchers, creating interactive visualizations
- Architected and implemented a high-performance data summarization tool in Rust, incorporating parallel processing and optimized memory management to process around 10M records daily, reducing analysis time by 75% and enabling real-time decision-making capabilities
- Designed and delivered a comprehensive data visualization workshop series using Python's Seaborn and Matplotlib libraries, training more than 50 researchers in advanced visualization techniques and statistical plotting, resulting in improved data presentation across the department
- Designed and implemented an LSTM-based deep learning model for gender classification from video frames, incorporating CNN feature extraction and temporal modeling to achieve 92% accuracy, demonstrating expertise in computer vision and sequence modeling

Education

Northwestern University Chicago, US
Ph.D in Bioinformatics. GPA: 3.7 *June 2022 – June 2025*

University of Minnesota Minneapolis, US
Ph.D. in Bioinformatics and Computational Biology. GPA: 3.68 *Sep. 2020 – June 2022*

China Agricultural University Beijing, CN
Master in Crop Bioinformatics. GPA 3.14 *Sep. 2018 – June 2020*

Northeast Agricultural University Harbin, CN
Bachelor of Arts in Agricultural Engineering. GPA 3.04 *Sep. 2014 – June 2018*

Technical Skills

Languages and Frameworks: C++, Python, Rust, Pytorch, Jax, Candle, GGML
Development Stack: Neovim, GDB, Git, Numpy, Pandas, Matplotlib, Docker, GitHub Action, CMake, HTML, GCC, Clang, Linux, LaTeX
Specializations: Algorithm Development, Concurrency Programming, Data Analysis and Visualization, Natural Language Processing

Grants and Honors

- First place of Computation and Data Exchange (CoDEx) Interactive Visualization Challenge (2024)

Conference talk

- Computation and Data Exchange (CoDEx) Visualization Challenge
- Workshop: Data Visualization with Seaborn

Publications

Li, Yangyang, Wang, T.-Y., Guo, Q., Ren, Y., Lu, X., Cao, Q., & Yang, R. (2024). A genomic language model for chimera artifact detection in nanopore direct rna sequencing. *bioRxiv*. doi: 10.1101/2024.10.23.619929

Li, Yangyang, & Yang, R. (2024, 12). PxBLAT: an efficient python binding library for BLAT. *BMC Bioinf.*, *25*(1), 1–8. doi: 10.1186/s12859-024-05844-0

Fry, J., **Li, Yangyang**, & Yang, R. (2022, 09). ScanExitronLR: characterization and quantification of exitron splicing events in long-read RNA-seq data. *Bioinformatics*. doi: 10.1093/bioinformatics/btac626