

YANGYANG LI

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

RESEARCH SUMMARY

I am a fifth-year Ph.D. candidate in computational biology specializing in deep learning, large language models, and large-scale biological data analysis. My research advances precision medicine by developing novel AI architectures that extract clinically relevant knowledge from multimodal biological data, including genomic sequences, transcriptomic profiles, and complex molecular structures. I developed two genomic language models (DeepChopper for direct RNA sequencing and ChimeraLM for single-cell genomics) that achieved up to 91% artifact reduction through long-context modeling using Hyena operators and attention-based architectures. Beyond model development, I design end-to-end machine learning systems, including custom data formats, high-performance Rust-based CLI tools, and interactive web visualization platforms. With extensive experience in PyTorch, distributed training, and processing millions of biological records, I translate foundation-model research into scalable, production-ready tools for real-world healthcare applications.

EDUCATION

Northwestern University	Chicago, US
<i>Ph.D. in Computational Biology</i>	<i>June 2022 – Present (expected Winter 2026)</i>
University of Minnesota	Minneapolis, US
<i>Ph.D. in Bioinformatics and Computational Biology (transferred to Northwestern)</i>	<i>Sep. 2020 – June 2022</i>
China Agricultural University	Beijing, CN
<i>Master in Crop Bioinformatics</i>	<i>Sep. 2018 – June 2020</i>
Northeast Agricultural University	Harbin, CN
<i>Bachelor of Arts in Agricultural Engineering</i>	<i>Sep. 2014 – June 2018</i>

RESEARCH EXPERIENCE AND PROJECTS

Northwestern University	Chicago, US
<i>Ph.D. in Computational Biology and Bioinformatics</i>	<i>June 2022 – Present</i>
• Developed DeepChopper, a Hyena operator-based genomic language model for detecting chimeric artifacts in nanopore direct RNA sequencing data, achieving 91% artifact reduction and improving transcript-level accuracy for precision oncology applications using PyTorch and distributed computing on HPC clusters	
• Developed ChimeraLM, a deep learning model leveraging Hyena operators for chimeric artifact detection in single-cell genomics data from clinical tumor samples, achieving 90% artifact reduction while preserving 80% of ground truth structural variations critical for cancer diagnosis and treatment selection	
• Designed a novel transcript segment graph (TSG) algorithm integrating graph-based representations to detect full-length transcripts with non-linear structures in long-read RNA sequencing data, enabling improved isoform quantification for personalized medicine applications	
• Designed and built a complete multi-modal data processing ecosystem for transcriptomic analysis: created a TSG file format specification for representing complex RNA structures, implemented high-performance Rust CLI tools processing millions of sequencing records, and developed Aurora, an interactive web visualization platform for collaborative biomedical research	
• Engineered PxBLAT, a production-ready Python library for genomic sequence alignment with time and memory optimization, achieving linear speed improvement while processing large-scale patient genomic data for variant calling and clinical interpretation	
University of Minnesota	Minneapolis, US
<i>Ph.D. in Bioinformatics and Computational Biology</i>	<i>Sep. 2020 – June 2022</i>
• Pioneered graph-based algorithms for detecting non-colinear transcript structures, establishing foundational methods for identifying disease-associated alternative splicing events and gene fusions in cancer genomics	
• Developed a transformer-based deep learning model in PyTorch to predict gene fusion-structural variation causality from genomic data, achieving 85% prediction accuracy, enabling rapid clinical decision support for precision oncology	
• Conducted comprehensive benchmarking of state-of-the-art alternative splicing detection tools using patient cohort data from The Cancer Genome Atlas (TCGA), producing evidence-based guidelines that improved detection accuracy for clinical bioinformatics workflows	

PUBLICATIONS

- Li, Yangyang**, Guo, Q., & Yang, R. (2026). ChimeraLM detects amplification artifacts for accurate structural variant calling in long-read single-cell sequencing. *Under Review*.
- Li, Yangyang**, Wang, T.-Y., Guo, Q., Ren, Y., Lu, X., Cao, Q., & Yang, R. (2026, 01). Genomic language model mitigates chimera artifacts in nanopore direct RNA sequencing. *Nat Commun.* doi: 10.1038/s41467-026-68571-5
- Guo, Q., **Li, Yangyang**, Wang, T.-Y., Ramakrishnan, A., & Yang, R. (2025). OctopusV and TentacleSV: a one-stop toolkit for multi-sample, cross-platform structural variant comparison and analysis. *Bioinformatics*, btaf599. doi: 10.1093/bioinformatics/btaf599
- 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). ScanExitonLR: characterization and quantification of exitron splicing events in long-read RNA-seq data. *Bioinformatics*. doi: 10.1093/bioinformatics/btac626

WORK EXPERIENCE

Northwestern University IT Organization

Chicago, US

Consultant in Data Science and Visualization

June 2024 – Present

- Developed an LSTM-based deep learning model for classification tasks with video data, achieving 69% accuracy and demonstrating ability to rapidly prototype AI solutions for diverse real-world applications
- Integrated social network data with institutional metadata using NetworkX, creating visualizations that identified collaboration patterns across researchers
- Built a high-performance data processing system in Rust with parallel processing capabilities, handling millions of records from diverse data sources with 70% reduction in processing time, enabling real-time analytics
- Delivered comprehensive workshops training researchers in statistical analysis, pytest for software testing, and advanced visualization techniques with seaborn, establishing best practices in research
- Developed a scalable PDF processing system extracting key terms from 10,000+ documents with parallel processing, automated error handling, failure recovery, and progress tracking, leveraging DuckDB for efficient data storage and downstream analytics

TECHNICAL SKILLS

- Machine Learning & AI:** Deep learning (PyTorch, transformers, LSTMs), genomic language models, Hyena operators, attention mechanisms, CNNs, graph neural networks
- Biomedical Data:** Multi-omics analysis (genomics, transcriptomics), single-cell sequencing, nanopore sequencing, electronic health record integration, GWAS, structural variation detection
- NLP & Sequence Modeling:** Transformer architectures, long-context modeling, sequence alignment, tokenization strategies, feature extraction
- Systems & Engineering:** Rust, Python, C++, distributed computing (SLURM/HPC), parallel processing, custom file format design, CLI development
- Data Visualization:** Interactive web applications (JavaScript, Shiny, Plotly), NetworkX, Seaborn, Matplotlib, graph visualization