

Yupeng Li

📞 706-206-9324 • ✉ liyupeng111@gmail.com

PROFILE

- **Experience in Pharmaceuticals:** 8+ years of industry experience in broad pharmaceutical landscape, spanning from early-stage drug discovery to clinical trials and real-world evidence
- **Innovation in Data Science:** Consistently pioneered and developed innovative statistical and machine learning (ML) algorithms to distill insights from biological and healthcare datasets
- **Multifaceted Expertise:** Profound expertise in data science (statistics, ML, AI, LLM, NLP), engineering (informatics, database, web-development, programming in Python, R, SQL) and biology (genetics, epidemiology, neuroscience, immunology, etc.)
- **Strategic Leadership:** Demonstrated leadership in shaping strategic directions, steering project management, fostering cross-functional collaborations, and guiding high-performance teams to success

EXPERIENCE

EncodeBox

Founder & Consultant

Remote

September, 2023 - Present

- Offer business consulting services specializing in data science and AI applications in pharmaceutical R&D
- Develop a unified analytics platform across multiple data sources, solving the segmentation issues due to the different data structures from different data sources
- Built an AI-powered informatics system for a top pharma client, which leveraged GPT-4 API for efficient, timely, and accurate disease surveillance
- Established the data infrastructure and machine learning capabilities of AI-based mRNA and LNP design for a biotech startup client

RVAC Medicines (an mRNA therapeutics startup)

Group Lead, Data Science

Waltham, Massachusetts

May, 2022 – August, 2023

- Directed a team of three to utilize AI for RNA and protein design, build AWS-based data infrastructure, and foster collaborative efforts for target discovery using omics data
- Trained and patented a large language model (transformer) for RNA sequences, leading to novel 5' UTR designs that markedly enhance protein production by 30% more than the commercial benchmark
- Optimized a deep learning model (CNN) for full-length internal ribosome entry site (IRES) prediction, achieving standout 0.89 AUC and surpassing other complex architectures
- Launched an integrative web platform utilizing advanced algorithms and user-friendly UI to timely monitor, predict, and analyze emerging SARS-CoV-2 variants from large genomic and epidemiological data

Merck & Co.

Associate Director, Center for Observational and Real-world Evidence

Boston, Massachusetts

April, 2020 – May, 2022

- Proactively addressed business needs by innovating workflows and enhancing analytical efficiency for real-world data (RWD), exemplified by fostering cross-functional collaborations that bolstered early drug discovery and designing a Looker data dashboard for streamlined multi-source data analysis
- Built a first-of-its-kind AI algorithm (transformer/NLP) for querying structured claims data directly through

natural language, reducing data retrieval time to <3 minutes with >99% accuracy

- o Developed a large language model (transformer) for claims data that enhanced the accuracy and generalizability for diverse patient outcome predictions
- o Initiated multiple efforts in the OHDSI community to systematically assess and improve the quality issues in OMOP common data model to enhance real-world data quality

Eli Lilly and Company

Research Scientist, Neuroscience Discovery

Indianapolis, Indiana

December, 2015 – April, 2020

- o Led the development of Parkinson's disease digital biomarkers that leveraged deep learning (CNN) to process smartphone sensor data and revealed significant treatment effects in a clinical trial that were undetected by conventional endpoints
- o Innovated a Bayesian network algorithm to build lupus and Alzheimer's gene networks from transcriptomic data, pinpointing potential therapeutic targets
- o Identified and validated several genetic biomarkers (SNP) associated with treatment response, paving the way for precision therapeutics in Alzheimer's disease and migraine
- o Streamlined GWAS, imputation, methylation, and GWAS/eQTL integration processes by building automated and parallelized pipelines with R, Python, and Perl, yielding substantial efficiency gains and saving millions in outsourcing costs

The University of Georgia

Research Assistant

Athens, Georgia

July, 2011 – September, 2015

- o Developed a probabilistic graphical model for gene network construction using transcriptomic data, leveraging prior knowledge to overcome data constraints and enhance precision and reliability
- o Engineered a gene network construction platform using Javascript, PHP, Markdown, and Github, hosted on AWS cloud, offering intuitive crowdsourcing and visualization
- o Offered expert consulting in statistics and bioinformatics analysis to both team members and external collaborators

Purdue University

Visiting Scholar

West Lafayette, Indiana

September, 2010 – July, 2011

- o Substantial experience with next-generation sequencing (NGS) data analysis: de novo genome assembly, sequence alignment, resequencing, SNP calling, RNA-Seq, genotyping-by-sequencing (GBS), etc.

EDUCATION

Massive Open Online Courses (Coursera, Udemy, edX, DataCamp, etc.)

Certificates in Data Science, Biology, and Healthcare

Remote

2011 – Present

The University of Georgia

PhD in Plant Breeding, Genetics and Genomics (Bioinformatics)

Athens, Georgia

2011 – 2015

The University of Georgia

MS in Statistics

Athens, Georgia

2011 – 2015

Northwest A&F University

BS in Horticulture

Yangling, China

2005 – 2009

PEER-REVIEWED PUBLICATIONS

- Chu YY, Yu D, **Li YP**, et al: A 5' UTR language model for decoding untranslated regions of mRNA and function predictions. *Nature Machine Intelligence* 2024, 6: 449–460 **AI** **LLM** **Bioinformatics**
- **Li YP**, et al: AI-assisted chart review to understand disease flares in systemic lupus erythematosus. Poster at *ISPOR* 2024, Atlanta, GA **AI** **LLM** **NLP** **RWD**
- **Li YP**: Prediction of full-length internal ribosome entry sites (IRES) using deep learning. Poster at *Fifth Annual RNA Therapeutics: From Concept to Clinic* 2023, Worcester, MA **AI** **Bioinformatics**
- **Li YP**, Huang Y, Zhang J: VIVID - An integrated system to closely monitor and predict emerging and high-risk SARS-CoV-2 variants. Poster at *The 22nd China Biological Products Annual Conference* 2023, Zhuhai, China **Web-Dev** **ML** **Bioinformatics**
- Wei S, Mobley M, Tao R, **Li YP**, et al: mRNA encoded antibodies improve biodistribution and efficacy of checkpoint inhibitors for liver cancer. *Journal for ImmunoTherapy of Cancer* 2023, 11(1), A1-A1731 **Bioinformatics**
- Abeysinghe R*, Black A*, Kaduk D*, **Li YP***, et al: Towards quality improvement of vaccine concept mappings in the OMOP vocabulary with a semi-automated method. *Journal of Biomedical Informatics* 2022, 134:104162 (* co-first authors) **RWD** **Informatics**
- **Li YP***, Dong W*, Ru BS, et al: Generic medical concept embedding and time decay for diverse patient outcome prediction tasks. *iScience* 2022, 25(9): 104880 **AI** **LLM** **RWD**
- Black A, **Li YP**, Kaduk D, et al: Constructing vaccine vocabulary hierarchy using formal concept analysis. Poster at *OHDSI Global Symposium* 2022, Bethesda, MD **RWD** **Informatics**
- Calvo MR*, **Li YP***, Mehrizghi T, et al: Machine learning-assisted query and information retrieval system on real-world data. Poster at *OHDSI Global Symposium* 2021, Remote **AI** **NLP** **RWD** **Informatics**
- Kaduk D, Black A, **Li YP***, et al: Evaluation of vaccine concept mappings in OMOP vocabulary: a real-world database study. Poster at *OHDSI Global Symposium* 2021, Remote **RWD**
- **Li YP**, Black A, Baltus GA, et al: Quality assessment of vaccine concepts in OMOP common data model. Poster at *OHDSI Global Symposium* 2020, Remote **RWD**
- **Li YP**, Higgs R, Hoffman R, et al: A Bayesian gene network reveals insight into the JAK-STAT pathway in systemic lupus erythematosus. *PLOS ONE* 2019, 14(12): e0225651 **Statistics** **ML** **Bioinformatics**
- **Li YP**, Guan YF, et al: Use digital sensor and deep learning to evaluate motor performance in the D1PAM phase 1B Parkinson's disease clinical trial. Poster at *International Congress of Parkinson's Disease and Movement Disorders* 2019, Nice, France **AI** **Statistics**
- Wang J, C Battioui C, **Li YP**, et al: Treatment monitoring using objective and frequent digital testing in the D1PAM (LY3154207) phase 1B Parkinson's disease clinical trial. Poster at *International Congress of Parkinson's Disease and Movement Disorders* 2019, Nice, France **Informatics** **Statistics**
- Wang H, **Li YP**, Ryder JW, et al: Genome-wide RNAseq study of the molecular mechanisms underlying microglia activation in response to pathological tau perturbation in rTg4510 Tau transgenic animal model. *Molecular Neurodegeneration* 2018, 13:65 **Bioinformatics** **NGS**
- **Li YP**, Liu YS: Gene co-expression analysis using non-negative matrix factorization in late-onset Alzheimer's disease. Poster at *Systems biology: networks* 2017, Cold spring harbor, NY **Statistics** **ML** **Bioinformatics**

- o **Li YP**, Jackson SA: Crowdsourcing the nodulation gene network discovery. *BMC Bioinformatics* 2016, 17(1):223 [Bioinformatics](#) [Web-Dev](#)
- o **Li YP**, Jackson SA: Nodulation gene networks in legumes. Presentation at *International Plant & Animal Genome XXIV Conference* 2016, San Diego, CA [Bioinformatics](#)
- o Gao DY, **Li YP**, Abernathy B, Jackson SA: Landscape and evolutionary dynamics of terminal repeat retro-transposons in miniature (TRIMs) in 48 whole plant genomes. *Genome biology* 2016, 17:7 [Bioinformatics](#)
- o **Li YP**, Pearl SA, Jackson SA: Gene networks in plant biology: approaches in reconstruction and analysis. *Trends in Plant Science* 2015, 20(10):664-675 [Statistics](#) [ML](#) [Bioinformatics](#)
- o **Li YP**, Jackson SA: Gene network reconstruction by integration of biological prior knowledge. *G3: Genes / Genomes / Genetics* 2015, 5(6): 1075-1079 [Statistics](#) [ML](#) [Bioinformatics](#)
- o Ferguson BJ, Li DX, Hastwell AH, Reid DE, **Li YP**, et al: The soybean (*Glycine max*) nodulation-suppressive CLE peptide, GmRIC1, functions interspecifically in common white bean (*Phaseolus vulgaris*), but not in a supernodulating line mutated in the receptor PvNARK. *Plant Biotechnology Journal* 2014, 12(8):1085-1097 [Bioinformatics](#)
- o Iwata A, Tek AL, Richard MMS, Abernathy B, Fonseca A, Schmutz J, Chen NWG, Thareau V, Magdelenat G, **Li YP**, et al: Identification and characterization of functional centromeres of the common bean. *Plant Journal* 2013, 76(1):47-60 [Bioinformatics](#)
- o Thudi M, **Li YP**, Jackson SA, et al: Current state-of-art of sequencing technologies for plant genomics research. *Briefings in Functional Genomics* 2012, 11(1):3-11 [Bioinformatics](#) [NGS](#)
- o Varshney RK, Chen WB, **Li YP**, et al: Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. *Nature Biotechnology* 2012, 30(1):83-89 [Bioinformatics](#) [NGS](#)
- o Zhai JX, Jeong DH, De Paoli E, Park S, Rosen BD, **Li YP**, et al: MicroRNAs as master regulators of the plant NB-LRR defense gene family via the production of phased, trans-acting siRNAs. *Genes & Development* 2011, 25(23):2540-2553 [Bioinformatics](#)

SELECTED BLOGS

EncodeBox.Beehiiv.com

- o Leveraging large language models for real-world evidence generation. 2023 [AI](#) [LLM](#) [NLP](#) [RWD](#)
- o Are attention and convolution all you need for RNA modeling? 2023 [AI](#) [Bioinformatics](#)
- o Small-molecule drug discovery in the age of AI. 2023 [AI](#)
- o To fine-tune or not to fine-tune? 2023 [AI](#) [LLM](#) [RWD](#)
- o AlphaFold is expanding beyond proteins. 2023 [AI](#) [Bioinformatics](#)
- o What's next after AlphaFold2 on protein structure prediction? 2023 [AI](#) [Bioinformatics](#)

EncodeBox.Medium.com

- o A silver medal solution to the NFL Big Data Bowl kaggle competition. 2020 [AI](#) [ML](#)
- o Autoencoder in biology - review and perspectives. 2019 [AI](#) [Bioinformatics](#)
- o Apply deep learning to transcriptome-based supervised learning. 2019 [AI](#) [Bioinformatics](#)