# Phi Bya

hungbya@gmail.com | phibya.com | linkedin.com/in/phibya | github.com/phibya | 470-708-2689

**SUMMARY:** Computer scientist with 6+ years of experience in data science and analysis pipeline development specializing in bioinformatics. Expert in integrating multi-modal data (genomic, text, sequence, network, image) using statistical modeling and deep learning. Proficient in large-scale data engineering, full-stack development, and HPC management.

# TECHNICAL SKILLS

- Programming: Python, R, JavaScript, C++, HTML/CSS
- Data & ML frameworks: PyTorch, Pandas, Polars, scikit-learn, dplyr, ggplot, letsplot
- Statistical & ML methods: hypothesis testing, experimental design, regression analysis, random forest, causal forest, pathway analysis, survival analysis, deep neural networks, matrix factorization, collaborative filtering, clustering, classification, ensemble methods, dimensionality reduction, representation learning.
- Databases: PostgreSQL, MySQL, MongoDB, Elasticsearch
- Infrastructure: Docker, Singularity, Slurm, Google Cloud Platform.
- Web Technologies: React.js, D3.js, Cytoscape.js, Node.js, REST APIs, GraphQL

## PROFESSIONAL EXPERIENCE

Research Assistant Professor | Auburn University | Auburn, ALMay 2024 – Present

- Lead data science initiatives for 5 PhD students and cross-functional research teams
- Architected and deployed HPC infrastructure (768 CPU cores, 16 GPUs, 6TB RAM) processing 50TB+ genomic data
- Reduced researcher onboarding time by 80% (1 week to 2 days) through automated environment setup and comprehensive documentation

Graduate Research Assistant | Auburn University | Auburn, AL August 2023 – May 2024

- Built end-to-end data analysis and machine learning pipeline in Python and R for single-cell genomic analysis processing 1M+ cells (<u>https://cytoanalyst.com</u>)
- Optimized dimension reduction algorithms (PCA, t-SNE, UMAP) for single-cell data with millions of cells
- Developed interactive visualization dashboard using React.js and D3.js serving 100+ concurrent users
- Applied deep learning models for cell classification and trajectory inference with 95% accuracy

Graduate Research/Teaching Assistant | University of Nevada, Reno | Reno, NV January 2018 – August 2023

- Taught statistical methods and R programming to 100+ students across diverse backgrounds (biology, biochemistry & computer science).
- Designed and maintained protein knowledge databases with 3M+ records using PostgreSQL and MongoDB; Analyzed and classified protein sequence into different protein families; Identified representative sequence using submodular optimization
- Built NLP pipeline using Python to crawl and index 200M+ scientific articles in Elasticsearch; Developed entity recognition models for detecting gene-disease-organism relationships.

Software Engineer | YouthDev Co., Ltd | Ho Chi Minh City, Vietnam May 2016 – December 2017

- Engineered real-time trending detection system processing 40M+ daily content items
- Implemented velocity-based scoring algorithm for trending detection reducing latency by 60%
- Built recommendation engine using collaborative filtering serving 100K+ users

Research and Development Engineer | YouNet Group | Ho Chi Minh City, Vietnam April 2014 – May 2016

- Applied collaborative filtering and matrix factorization to analyze 50M+ social media profiles
- Developed sentiment analysis models using Naive Bayes achieving 87% accuracy
- Built user attribute prediction system using Random Forest with 82% precision

## **EDUCATION**

<ul><li>Ph.D., Computer Science, Auburn University</li><li>M.Sc., Computer Science, University of Nevada, Reno</li><li>B.Eng., Information Systems, UIT, Vietnam National University</li></ul>	May 2024 May 2022 May 2016
--	----------------------------------

# **KEY PROJECTS**

## Cancer Subtype Discovery using Multi-Modal Data Integration

- Developed novel clustering algorithms using randomize transformation and consensus networks to integrate 4 data types (gene, protein, methylation, clinical) for patient stratification
- Published R package (PINSPlus) with 30,000+ downloads on CRAN
- Improved survival prediction accuracy by 25% using Cox proportional hazards models

# Novel Pathway Analysis Pipeline with Web Interface

- Developed innovative pathway analysis method using signal deconvolution to decompose noisy gene expression into interpretable components, and applying variance stabilizing transformation and regularized regression to detect disease-impacted biological processes with higher sensitivity
- Deployed web application supporting 500+ organisms with 200+ monthly active users
- Reduced analysis time from hours to minutes while improving detection accuracy by 30%
- Available at: <u>https://ipsa.tinnguyen-lab.com</u>

# **DeconBenchmark – Cellular Deconvolution Framework**

- Standardized and containerized 50+ deconvolution methods into an unified R package framework for reproducible research and comparison analysis.
- Benchmarked methods including matrix factorization, neural networks, and Bayesian approaches
- Open-source: <u>https://github.com/tinnlab/DeconBenchmark</u>

# SELECTED PUBLICATIONS

Full publication list (30+ papers): https://scholar.google.com/citations?user=yw6kjSYAAAAJ

- Phi Bya, Ha Nguyen, Duc Tran, Sorin Draghici, and Tin Nguyen. "Fourteen years of cellular deconvolution: methodology, applications, technical evaluation and outstanding challenges." *Nucleic Acids Research* 52, no. 9 (2024): 4761-4783. <u>https://doi.org/10.1093/nar/gkae267</u>
- Duc Tran, Phi Bya, Bang Tran, Carlo La Vecchia, Hung N. Luu, and Tin Nguyen. "Fast and precise singlecell data analysis using a hierarchical autoencoder." Nature communications 12, no. 1 (2021): 1029. <u>https://doi.org/10.1038/s41467-021-21312-2</u>
- Phi Bya, Duc Tran, Bang Tran, Monikrishna Roy, Adam Cassell, Sergiu Dascalu, Sorin Draghici, and Tin Nguyen. "SMRT: Randomized data transformation for cancer subtyping and big data analysis." Frontiers in oncology 11 (2021): 725133. <u>https://doi.org/10.3389/fonc.2021.725133</u>
- Phi Bya, Sushil J. Louis, and Tin Nguyen. "MGKA: A genetic algorithm-based clustering technique for genomic data." In 2019 IEEE Congress on Evolutionary Computation (CEC), pp. 103-110. IEEE, 2019. https://doi.org/10.1109/CEC.2019.8790225