

Hao XUE

hx222@cornell.edu | [linkedin](#) | [code](#) | 6172303339

EDUCATION

Cornell University PhD in Computational Biology	2021 – 2026 GPA: 3.988
Harvard University SM in Biostatistics	2019 – 2021 GPA: 3.91
Xi'an Jiaotong-Liverpool University (XJTLU) BSc in Applied Mathematics (First Class with Honors)	2015 – 2019 GPA: 3.89

SUMMARY

7+ years of interdisciplinary research applying **machine learning** and **AI systems** to biomedical informatics, including developing image-omics **multimodal** foundation models, **Natural Language Processing** algorithms for extracting temporal information from Electronic Health Records (EHRs), **statistical methods** for multi-omics time-series pattern recognition, and **Bayesian hierarchical models** for multi-omics integration.

WORK EXPERIENCE

Vision-omics Multimodal Foundation Model AI Engineer Intern, Data Storage Product Line	May 2025 – Aug. 2025 Huawei Technologies Co., Ltd., Shanghai, China
- Construct a novel multimodal foundation model to bridge histopathology with spatial transcriptomics. - Finetune a pathology image foundation model (ViT) and a single-cell foundation model (transformers) with LoRA to learn representations using contrastive learning distributively. - Implement downstream tasks, including histology image-transcriptomics retrieval, domain clustering, tissue annotation, and gene expression prediction by histology images to evaluate the capability of the model. - Collaborate with an interdisciplinary team (pathologists, geneticists, engineers) to prepare research papers and transition research prototypes to clinical applications (SpaRuiPath).	

RESEARCH EXPERIENCES

Directional Embedding for EHRs (Published) Supervisors: Prof. Junwei Lu, Prof. Tianxi Cai	July 2020 – Jan. 2024 Independent Study at Harvard
- Generalize word2vec (skip-gram), a famous NLP algorithm, to generate word embeddings that contain temporal order information for medical entities from EHRs in Python. - Apply an alternative minimizing method using Nesterov's Accelerated Gradient Descent to accelerate the optimization of the corresponding objective function. - Use the resulting word embeddings to construct temporally directional knowledge graphs and benchmark with transformer-based methods, including SapBERT, BiomedBERT, and CODER.	
Integrative Analysis of Time-Course Multi-Omics Data (under Review) Supervisors: Prof. Sumanta Basu, Prof. Andrew G. Clark	Dec. 2021 – Dec. 2023 Research Assistant at Cornell
- Build an R-package, MINT-DE (Multiomics INtegration of Timecourse with Differential Expression), which can select time-series that are both differentially expressed and temporally correlated across multi-omics by using Edgington's method to summarise p-values from related statistical tests. - Benchmark MINT-DE with existing statistical methods (e.g. sparse partial least square) in protein-protein interaction network statistics and Gene Ontology enrichment analysis. - Present at speed and poster presentations at Joint Statistical Meetings, Portland, OR, USA, 2024.	
Analysis of Mitochondrial Proteomics and Metabolomics (in Progress) Supervisors: Prof. Sumanta Basu, Prof. Andrew G. Clark	Oct. 2023 – Present Research Assistant at Cornell
- Analyze proteomics and metabolomics lung cancer mouse model with different mutated oncogenes - Use Gene Set Analysis package to find enriched pathway modules.	
Integrating Multiplatform Genomic Data (Published) Supervisor: Prof. Sounak Chakraborty, Dr. Tanujit Dey	July 2020 – Jan. 2024 Independent Study at Harvard

- Use Expectation–maximization algorithm based Bayesian variable selection method to group genes by their effect to DNA methylation.
- Employ group Bayesian Lasso to select biomarkers by adding normal-exponential-gamma priors so that grouping information could be embedded in penalties.
- Combine both models to detect genes associated with glioblastoma survival by integrating methylation, gene expression and clinical with both simulated data and real data obtained from TCGA dataset using accelerated failure time model.

Predict mRNA Methylation from Gene Expression (Published)

Sept. 2018 – Sept. 2019

Supervisor: Prof. Jia Meng & Prof. Jionglong Su

Final Year Project at XJTU

- Predict mRNA methylation status (MeRIP-Seq) from gene expression data (RNA-Seq) with machine learning methods in R, including elastic net penalized logistic regression, Support Vector Machine and Random Forest.

TEACHING ASSISTANTSHIP

ECON 3110: Probability Models and Inference for the Social Sciences	Spring 2023, Spring 2024, Cornell
STSCI 6780: Bayesian Statistics and Data Analysis	Spring 2025, Cornell
STSCI 4060: Python Programming and Its Application in Statistics	Fall 2025, Cornell
MATH 1920: Multivariable Calculus for Engineers	Spring 2026, Cornell

AWARDS

Meritorious Winner of Interdisciplinary Contest in Modeling, <i>Consortium for Mathematics and its Applications</i>	2018
Academic Excellence Award (top 5%), <i>XJTU</i>	2018
Outstanding Winner of the Tenth Mathematical Contest in Modeling of XJTU (top 2), <i>XJTU</i>	2017
Mathematics Cluster Winner in Summer Undergraduate Research Fellow poster competition (top 1), <i>XJTU</i>	2016

REVIEW

Diabetes, Obesity and Metabolism, IEEE Journal of Biomedical and Health Informatics, Aging and Disease, Bioinformatics, BMC Bioinformatics, and Journal of Big Data.

SKILLS

Technical:

R: 7+ years research experience, familiar with implementing statistical algorithms and analyzing biological data,
 Python: 7+ years research experience, familiar with deep learning platforms (pyTorch, Tensorflow, Keras, Transformers)
 and data science toolkits (Pandas, Numpy, scikit-learn, Matplotlib), both theoretical understanding and practical
 experience of frequently used deep learning models (Transformers, VAE, GNN, CNN, RNN, etc.),
 MATLAB, JAVA, Linux, L^AT_EX, Microsoft Office (Word, Excel, PowerPoint)

Languages: Mandarin (native), English (proficient)

PUBLICATIONS&MANUSCRIPT IN PREP.

Xue H., Ruan M., Yan S., et al. SpaRuiPath: A Multimodal Foundation Model for Aligning Histology and Spatial Transcriptomics in Breast Cancer. *Manuscript in prep.*
<https://www.overleaf.com/read/chrkjyrdvjpn#b89af9>.

Wen J.*, **Xue H.***, Rush E., et al. DOME: Directional Medical Embedding Vectors from Electronic Health Records. *Journal of Biomedical Informatics*, 2025. <https://doi.org/10.1016/j.jbi.2024.104768>. (*=equal contribution)

Xue H., Delbare S. Y. N., Wells M. T., et al. Integrative Analysis of Differentially Expressed Genes in Time-Course Multi-Omics Data with MINT-DE. *Major revision, BMC Bioinformatics*, 2024.
<https://doi.org/10.21203/rs.3.rs-3806701/v2>.

Xue H., Chakraborty S., Dey T., Bayesian shrinkage models for integration and analysis of multiplatform high-dimensional genomics data. *Statistical Analysis and Data Mining*, 2024.
<https://doi.org/10.1002/sam.11682>.

Xue H., Wei Z., Chen K., et al. Prediction of RNA Methylation Status From Gene Expression Data Using Classification and Regression Methods. *Evolutionary Bioinformatics*, 2020. <https://journals.sagepub.com/doi/full/10.1177/1176934320915707>.

Liu Z., Lin H., Li X., **Xue H.**, et al. The network structural entropy for single-cell RNA sequencing data during aging. *Briefings in Bioinformatics*, 2024. <https://doi.org/10.1093/bib/bbae698>.

Cho M., Dey T., **Xue H.**, et al., BayesMultiomics: An R Package for Bayesian Shrinkage Models for Integration and Analysis of Multi-platform High-dimensional Genomics Data, *Statistical Analysis and Data Mining*, 2025. <http://dx.doi.org/10.1002/sam.70009>.