BINGJUN LI

+1 (857) 424-8545 ⋄ bingjun.li@uconn.edu ⋄ Vernon, CT, 06066
https://www.bingjunli.com

SUMMARY

My research work focuses on analysis tools on **genomics data** (single-cell and spatial transcriptomics), **Multiomics Integration** and **Alignment** on both bulk-cell and single-cell sequencing data, and utilizing **graph neural network (GNN)** to accommodate the structure of sequencing data and incorporate prior knowledge.

EDUCATION

University of Connecticut, Storrs, CT 08.2019 - Present
Ph.D in Computer Science & Engineering
Ph.D in Statistics (Transferred after 1 year)

The George Washington University, Washington, DC 08.2015 - 06.2017
Master of Science in Statistics

Boston University, Boston, MA 01.2011 - 06.2014
Bachelor of Science in Mathematics

PUBLICATION

Li, B., Karami, M., Junayed, M.S., & Nabavi, S. (2024, August) Multi-modal Spatial Clustering for Spatial Transcriptomics Utilizing High-resolution Histology Images. (Under Review)

Li, B., & Nabavi, S. (2024, January). A Multimodal Graph Neural Network Framework for Cancer Molecular Subtype Classification. *BMC bioinformatics* 25.1 (2024): 27.

Li, B., & Nabavi, S. (2023, December). scGEMOC, A Graph Embedded Contrastive Learning Single-cell Multiomics Clustering Model. *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2023.

Li, B., & Nabavi, S. (2023, September). Contrastive Learning in Single-cell Multiomics Clustering. In *Proceedings of the 14th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*.

Li, B.*, Bai, J.*, & Nabavi, S. (2022, August). Semi-supervised classification of disease prognosis using CR images with clinical data structured graph. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* (pp. 1–9). (* indicates equal contributions)

Li, B., Wang, T., & Nabavi, S. (2021, August). Cancer molecular subtype classification by graph convolutional networks on multi-omics data. In *Proceedings of the 12th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (pp. 1-9).

Weiner, S., **Li, B.**, & Nabavi, S. (2024). Improved allele-specific single-cell copy number estimation in low-coverage DNA-sequencing. *Bioinformatics*, 40(8), btae506.

Wang, T., **Li, B.**, & Nabavi, S. (2021, December). Single-cell RNA sequencing data clustering using graph convolutional networks. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 2163-2170). IEEE.

RESEARCH EXPERIENCE

Multimodal Spatial Clustering for Spatial Transcriptomics with High-resolution Histology Images
09.2023-08.2024

- Spatial transcriptomics (ST) sequences genomic data by spots with spatial information and corresponding high-resolution histology images, and one of the major challenges in ST data analysis is clustering spatial spots.
- Led the development of stMMC, a first deep learning model that integrates genomic data with histology image features using a contrastive learning mechanism for accurate spatial clustering in ST data.

· Tested stMMC against four state-of-art models in two datasets and achieves 6% performance gain on average.

Multimodal GNN Framework for Cancer Molecular Subtype Classification 09.2021-12.2023

· Developed an end-to-end multiomics Graph Neural Network (GNN) framework for accurate and robust classification of cancer molecular subtypes based on multimodal genomics data.
· Utilized heterogeneous multi-layer graphs combining both inter-omics and intra-omic connections, leveraging the established biological knowledge for improved classification accuracy.
· Conducted a comparison study between Graph Attention Network (GAT) and Graph Convolution Network (GCN) as the backbone model, showing the superior performance of the proposed model over four baseline models in terms of accuracy, F1 score, precision, and recall.

scGEMOC, Graph Embedded Contrastive Learning Multiomics Clustering Model 09.2022-08.2023

· Single-cell multiomics data are multimodal, high-dimensional and sparse data (over 90%) with nonuniform noise.
· Proposed scGEMOC, a scalable self-learning method for single-cell multiomics cell clustering that uses contrastive learning to effectively align modality features, and prioritize critical modality in the feature fusion.
· Embedded the graph consists of the gene-gene interaction and gene-ATAC relationship as a pseudo omic.
· Achieved improved clustering performance compared to five baseline models on three public datasets in terms of clustering accuracy, adjusted rand index (ARI), and normalized mutual information (NMI).

Semi-supervised classification of disease prognosis using CR images with clinical data structured graph 06.2021-08.2022

· Developed a semi-supervised classification model, CGMNN, that integrates chest computed radiography (CR) images with clinical data to predict the COVID-19 patients' ICU demand at hospital admission stage.
· Utilized a graph Markov neural network that extracts not only node features but also the local label distribution to make a better prediction.
· Achieved superior performance compared to baseline models with an accuracy of 0.82, sensitivity of 0.82, precision of 0.81, and an F1 score of 0.76, using a dataset of 1,342 patients.

Cancer molecular subtype classification by graph convolutional networks on multi-omics data 09.2020-08.2021

· Proposed an end-to-end deep learning model that integrates multi-omics data with prior biological knowledge using a GCN to classify pan-cancer molecular subtypes with server imbalance.
· Designed a parallel network structure to extract localized features through a GCN with a gene-gene interaction network, and global features through a fully connected neural network, then concatenating both for classification..
· Demonstrated superior performance over both deep learning and conventional machine learning models with prediction accuracy of 0.87, precision of 0.87, recall, of 0.87 and F1 score of 0.87.

WORK EXPERIENCE

Boehringer Ingelheim, Ridgefield, CT 08.2023-01.2024
Data Scientist Internship

· Developed machine learning algorithms for predicting patients' prognostic diagnosis using demographic information, medical history, and screening eye exams, which improve future clinical screening efficiency by 50%.
· Designed new statistical data analysis pipelines that are more suitable for severely imbalanced, small-sized datasets, which improves the performance of different downstream analysis by 30-50%.
· Created machine learning-based time-to-event predictive methods, revealing insights between patients' attributes and prognostic outcomes by comparing significant attributes from prognostic prediction and time-to-event models.

University of Connecticut, Storrs, CT 09.2020 - Present
Research Assistant

· Two wokring projects: 1) a diffusion-based model for copy number variation (signals in genome) segmentation and inference; 2) a contrastive learning guided cell clustering using histology images and genome data (like text).

- Created a scalable self-learning multimodal model for single-cell multiomics cell clustering that embeds the heterogeneous graph using GNN and uses contrastive learning for modality alignment and better interpretability.
- Proposed a semi-supervised prognostic prediction model for COVID-19 patients, integrating X-ray images and clinical data via a graph Markov neural network that extracts both node features and local label distribution.
- Developed a deep learning method with a parallel structure of GNN and a fully-connected neural network to extract both local and global features of genome for cell classification on imbalanced bulk-cell multiomics data.

EdLab, Teachers College, Columbia University, New York, NY 05.2019 - 08.2019
Data Engineer

- Built a OCR processing pipeline as a cloud service in Python, significantly reducing memory usage by 80%.
- Developed web-based visualization tools for the management team: one to understand AWS expenses, cutting costs by 13%; and another to show the library item network, improving staff working efficiency by 10%.

Utofun, New York, NY 08.2017 - 05.2019
Data Analyst

- Designed ETL processes for millions of real estate listing data and created automated visual report generator.
- Developed standardized procedures in R to select optimal time series model for market behavior prediction.

HONOR

Cigna Fellowship, University of Connecticut 2020, 2021
Doctoral Travel Award, University of Connecticut 2022
Predoctoral Fellowship, University of Connecticut 2022, 2023, 2024

PROFESSIONAL SERVICE

Session Chair
IEEE BIBM 2023
Program Chair
IEEE BIBM 2024
ICLR Tiny Paper 2022
Reviewer
AISTATS
ICLR Tiny Paper
ICML AdvML
EAIKDD
Bioinformatics
NeurIPS
Frontier in Oncology
IEEE BIBM
BMC Bioinformatics