Hongzhi Wen

Email: wenhongz@msu.edu
Phone: +1-5175823840
Homepage: www.cse.msu.edu/~wenhongz

Education

- **Michigan State University** 09/2021 - 12/2024
  PhD in Computer Science and Engineering
- **Peking University** 09/2016 - 07/2021
  Bachelor of Arabic, School of Foreign Languages
  Bachelor of Computer Science and Technology (Minor), EECS

Research Interests

- Reasoning Ability of Large Language Models
- AI for Science, Foundation Models for Single-Cell Analysis

Work Experience

- **Artificial Intelligence Unit, CartaBio Inc** 08/2024 – Present
  Head of AI Analytics Boston, MA (*worked remote from MI before 12/2024 and WA after 01/2025*)
  – Lead a group of bioinformatics engineers and computer scientists for interdisciplinary collaboration, data collection and analysis, and AI model development.
  – Develop a proprietary single-cell foundation model (CellBERT) for the company. Integrated T cell datasets and predicted genetic perturbation effect using the foundation model.
  – Construct a modern database and a whole pipeline for collecting, cleaning and storing single-cell multi-omics data.
- **Data Science and Engineering Lab, Michigan State University** 09/2021 – 08/2024
  Research Assistant East Lansing, MI
  – Built multimodal foundation models for single-cell biology based on graph transformers.
  – Developed large-scale graph neural networks for node classification and single-cell analysis.
- **Artificial Intelligence Team, CartaBio Inc** 01/2024 – 04/2024
  Research Scientist Intern Boston, MA (*worked remote from MI*)
  – Developed T cell foundation model for immunology research.
  – Integrated T cell datasets and predicted genetic perturbation effect.
- **Search Assistant Team, Amazon** 08/2023 – 12/2023
  Applied Scientist Intern Palo Alto, CA (*worked remote from MI*)
  – Built and evaluated large language models for complex reasoning.
  – Fine-tuned large language models for better chain-of-thoughts prompting.
  – Collaborated with other scientists on improving the safety of language models and published a paper in AI conference as a co-author.
- **Search Assistant Team, Amazon** 06/2023 – 08/2023
  Applied Scientist Intern Palo Alto, CA
  – Built and evaluated large language models for complex reasoning.
  – Fine-tuned large language models for better chain-of-thoughts prompting.
- **Machine Learning Team, TigerGraph Company** 05/2022 – 08/2022
  Machine Learning Engineer Intern Redwood City, CA (*worked remote from MI*)
  – Proposed and conducted a research project that applies graph neural networks to single-cell data imputation, meanwhile demonstrating the advantage of TigerGraph database and cloud service in the bio-industry.
  – Participated in Machine Learning Workbench release project. Explored tensorboard on-cloud service, and wrote tutorials and documentation.
- **Intelligent Scheduling Group, Songguo Company** 09/2020 – 07/2021
  Machine Learning Algorithm Intern Beijing, China
  – Participated in the company's core project and leaded multi-department collaboration. Implemented machine learning algorithms to improve vehicle scheduling efficiency.
  – Processed big data with PySpark and Hadoop, predicted quantities of orders using decision tree models such as XGBoost and LGBM, and deep learning models such as LSTM and Graph Neural Networks.
- **Natural Language Processing Group, Didi Company** 11/2019 - 05/2020
  Algorithm Intern Beijing, China
  – Developed NLP models for POI optimization and machine translation.
  – Assisted senior engineers in data annotation, data analysis, and exploration of existing methods.
- **Text Mining Group, Institute of Computer Science and Technology, Peking University** 11/2017 – 09/2019
  Research Intern Beijing, China
  – Conducted research on the topic of "Universal Semantic Representation Based on Deepbank and HPSG".
  – Built an end-to-end neural model based on LSTM and attention mechanism to parse English sentences to their semantic representation.

Competitions

- **NeurIPS OGB-LSC 2022 competition, node-level track** (2nd place) 2022
- **Kaggle Single-cell Multimodal Integration competition** (top 2%) 2022
- **NeurIPS Multimodal Single-Cell Data Integration competition, predict modality track** (1st place) 2021
- **National Computer Design Competition for College Students** (5th in 300 groups) 2017

Open-Source Projects

- **DANCE [Link]** 06/2022 - Present

A python toolkit to support deep learning models for analyzing single-cell omics data at scale

  – The first deep learning library and the first comprehensive benchmark for single-cell analysis.
  – 7k+ downloads, 300+ stars

- **CellPLM [Link]** 05/2023 - Present

A multitask foundation model for single-cell transcriptomics analysis.

  – The first foundation model for single-cell analysis that leverages spatial positional information with a transformer. State-of-the-art performance on various tasks.
  – 5k+ downloads, 70+ stars

Publications

•**Revisiting the graph reasoning ability of large language models: Case studies in translation, connectivity and shortest path**
Xinnan Dai, Qihao Wen, Yifei Shen, **Hongzhi Wen**, Dongsheng Li, Jiliang Tang, Caihua Shan. arXiv preprint 2024

•**Learning on Graphs with Large Language Models (LLMs): A Deep Dive into Model Robustness**
Kai Guo, Zewen Liu, Zhikai Chen, **Hongzhi Wen**, Wei Jin, Jiliang Tang, Yi Chang. arXiv preprint 2024

•**MARVEL: Microenvironment Annotation by Supervised Graph Contrastive Learning**
Yan Cui, **Hongzhi Wen**, Robert Yang, Xi Luo, Hui Liu, Yuying Xie. bioRxiv preprint 2024

•**Content Knowledge Identification with Multi-agent Large Language Models (LLMs)**
Kaiqi Yang, Yucheng Chu, Taylor Darwin, Ahreum Han, Hang Li, **Hongzhi Wen**, Yasemin Copur-Gencturk, Jiliang Tang, Hui Liu. AIED 2024

•**IterAlign: Iterative Constitutional Alignment of Large Language Models**
Xiusi Chen, **Hongzhi Wen**, Sreyashi Nag, Chen Luo, Qingyu Yin, Ruirui Li, Zheng Li, Wei Wang. NAACL 2024

•**CellPLM: Pre-training of Cell Language Model Beyond Single Cells**
Hongzhi Wen*, Wenzhuo Tang*, Xinnan Dai, Jiayuan Ding, Wei Jin, Yuying Xie, Jiliang Tang. ICLR 2024

•**Are Large Language Models (LLMs) Good Social Predictors?**
Kaiqi Yang, Hang Li, **Hongzhi Wen**, Tai-Quan Peng, Jiliang Tang, Hui Liu. arXiv preprint 2024

•**Copyright Protection in Generative AI: A Technical Perspective**
Jie Ren, Han Xu, Pengfei He, Yingqian Cui, Shenglai Zeng, Jiankun Zhang, **Hongzhi Wen**, Jiayuan Ding, Hui Liu, Yi Chang, Jiliang Tang. arXiv preprint 2024

•**Investigating Out-of-Distribution Generalization of GNNs: An Architecture Perspective**
Kai Guo, **Hongzhi Wen**, Wei Jin, Yaming Guo, Jiliang Tang, Yi Chang. SIGKDD 2024

•**Label-free Node Classification on Graphs with Large Language Models (LLMs)**
Zhikai Chen, Haitao Mao, **Hongzhi Wen**, Haoyu Han, Wei Jin, Haiyang Zhang, Hui Liu, Jiliang Tang. ICLR 2024

•**Exploring the potential of large language models (llms) in learning on graphs**
Zhikai Chen, Haitao Mao, Hang Li, Wei Jin, **Hongzhi Wen**, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, Wenqi Fan, Hui Liu, Jiliang Tang. ACM SIGKDD Explorations Newsletter 2024

•**SpatialCTD: a large-scale TME spatial transcriptomic dataset to evaluate cell type deconvolution for immuno-oncology**
Jiayuan Ding, Lingxiao Li, Qiaolin Lu, Julian Venegas, Yixin Wang, Lidan Wu, Wei Jin, **Hongzhi Wen**, Renming Liu, Wenzhuo Tang, Xinnan Dai, Zhaoheng Li, Wangyang Zuo, Yi Chang, Yu Leo Lei, Lulu Shang, Patrick Danaher, Yuying Xie, Jiliang Tang. Journal of Computational Biology 2024

• **DANCE: A Deep Learning Library and Benchmark for Single-Cell Analysis**
Jiayuan Ding, Renming Liu, **Hongzhi Wen**, Wenzhuo Tang, Zhaoheng Li, Julian Venegas, Runze Su, Dylan Molho, Wei Jin, Yixin Wang, Qiaolin Lu, Lingxiao Li, Wangyang Zuo, Yi Chang, Yuying Xie, Jiliang Tang. Genome Biology 2024

• **Deep Learning in Single-Cell Analysis**
Dylan Molho*, Jiayuan Ding*, Zhaoheng Li, **Hongzhi Wen**, Wenzhuo Tang, Yixin Wang, Julian Venegas, Wei Jin, Renming Liu, Runze Su, Patrick Danaher, Robert Yang, Yu Leo Lei, Yuying Xie, Jiliang Tang. TIST 2024

•**Amazon-M2: A Multilingual Multi-locale Shopping Session Dataset for Recommendation and Text Generation**
Wei Jin, Haitao Mao, Zheng Li, Haoming Jiang, Chen Luo, **Hongzhi Wen**, Haoyu Han, Hanqing Lu, Zhengyang Wang, Ruirui Li, Zhen Li, Monica Xiao Cheng, Rahul Goutam, Haiyang Zhang, Karthik Subbian, Suhang Wang, Yizhou Sun, Jiliang Tang, Bing Yin, Xianfeng Tang. NeurIPS 2023 Datasets & Benchmarks

•**A General Single-Cell Analysis Framework via Conditional Diffusion Generative Models**
Wenzhuo Tang, Renming Liu, **Hongzhi Wen**, Xinnan Dai, Jiayuan Ding, Hang Li, Wenqi Fan, Yuying Xie, Jiliang Tang. arXiv preprint 2023

• **Single Cells Are Spatial Tokens: Transformers for Spatial Transcriptomic Data Imputation**
Hongzhi Wen, Wenzhuo Tang, Wei Jin, Jiayuan Ding, Renming Liu, Feng Shi, Yuying Xie, Jiliang Tang. arXiv preprint 2023

•**MEM-GAN: A Pseudo Membrane Generator for Single-cell Imaging in Fluorescent Microscopy**
Yixin Wang, Jiayuan Ding, Lidan Wu, Aster Wardhani, Patrick Danaher, Qiaolin Lu, **Hongzhi Wen**, Wenzhuo Tang, Yi Chang, Yu Leo Lei, Jiliang Tang, Yuying Xie. bioRxiv preprint 2023

• **Single-Cell Multimodal Prediction via Transformers**
Wenzhuo Tang*, **Hongzhi Wen***, Renming Liu*, Jiayuan Ding, Wei Jin, Yuying Xie, Hui Liu, Jiliang Tang. CIKM 2023

• **Bi-channel Masked Graph Autoencoders for Spatially Resolved Single-cell Transcriptomics Data Imputation**
Hongzhi Wen, Wei Jin, Jiayuan Ding, Christopher Xu, Yuying Xie, Jiliang Tang. NeurIPS 2022 AI for Science workshop

• **Graph Neural Networks for Multimodal Single-Cell Data Integration**
Hongzhi Wen, Jiayuan Ding, Wei Jin, Yiqi Wang, Yuying Xie, Jiliang Tang. SIGKDD 2022

Services

Reviewer

- International Conference on Learning Representations (ICLR) 2024
- Conference on Neural Information Processing Systems (NeurIPS) 2023
- IEEE International Conference on Data Mining (ICDM) 2023
- ACM Conference on Recommender Systems (RecSys) 2023
- SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) 2023
- The Web Conference (WWW) 2023
- SIAM International Conference on Data Mining (SDM) 2024
- Transactions on Knowledge Discovery and Data Engineering (TKDE) 2023
- Transactions on Information Systems (TOIS) 2023
- AAAI Conference on Artificial Intelligence (AAAI) 2023, 2024

Voluntary

- SIGKDD Conference on Knowledge Discovery and Data Mining (KDD) 2022

Presentations

Invited Talks

- "Exploring the Potential of Large Language Models (LLMs) in Learning on Graphs" 2023
  - Presentation at Sydney University
- "Single Cells Are Spatial Tokens: Towards Cell Language Models" 2023
  - Presentation at NanoString Technologies
- "Graph Neural Networks for Single-Cell Analysis" 2022
  - Presentation at 10X Technologies, Inc
- "Advancing Single-cell Multi-omics Data Integration with Graph Representation Learning" 2022
  - Presentation at Emory Graph Mining Lab, Emory University

Conference Oral Presentations

- "Single Cells Are Biological Tokens: Towards Cell Language Models" 2023
  - Presentation at GLSIAM Symposium
- "Inverse APPNP: Solution for OGB-LSC 2022 MAG240M" 2022
  - Presentation at NeurIPS OGB Large-scale Challenge (OGB-LSC) workshop
- "Team DANCE Solution: Graph-based CITE-seq Prediction" 2022
  - Presentation at NeurIPS Multimodal Single-Cell Integration workshop
- "Advancing Single-cell Multi-omics Data Integration with Graph Representation Learning" 2022
  - Presentation at MLCB conference
- "Graph Neural Networks for Multimodal Single-Cell Data Integration" 2022
  - Presentation at SIGKDD conference
- "Graph Neural Networks for Single Cell Analysis" 2021, 2022
  - Presentation at NeurIPS Multimodal Single-Cell Data Integration workshop

Teaching Experience

- Teaching Assistant for CSE 482 Big Data Analysis 2022
  - Duties include online discussions, office hours, and grading.