Department of Biostatistics and Medical Informatics Seminar

Wenjin Jim Zheng, PhD

Professor, Department of Bioinformatics and Systems Medicine, McWilliams School of Biomedical Informatics at UTHealth Houston

Friday, April 11, 2025

12:00-1:00 pm

Biotech Center Auditorium *or* via Zoom:

https://uwmadison.zoom.us/j/99879638765?pwd=wbtqxoucEFIlPVCVc9SFbvKB1Av7Xk.1

Passcode: 343271

Leveraging LLMs and AI Agent Networks for Community based Gene Set and Cell Type Annotation

Abstract: Single-cell RNA sequencing has transformed our ability to identify diverse cell types and their transcriptomic signatures. However, annotating these signatures—especially those involving poorly characterized genes—remains a major challenge. Traditional gene set analysis methods, such as Gene Set Enrichment Analysis (GSEA), rely heavily on well-curated annotations and often underperform in such contexts. Large Language Models (LLMs) offer a promising alternative but struggle to represent complex biological knowledge within structured ontologies. To address this, we present a novel approach that integrates free-text descriptions with ontology labels for more accurate and robust gene set annotation. Our method outperforms state-of-the-art tools, correctly annotating over 68% of gene sets within the top five predictions. By incorporating retrieval-augmented generation (RAG), we developed a robust agentic workflow that refines predictions using relevant PubMed literature to reduce hallucinations and enhance interpretability. Using this workflow, we annotated 5,322 brain cell clusters from the complete mouse brain cell atlas generated by the BRAIN Initiative Cell Census Network, creating a valuable resource to support community-driven cell type annotation efforts.

Bio: W. Jim Zheng, PhD, MS, is a professor at McWilliams School of Biomedical Informatics at UTHealth Houston. His research focuses on integrating, modeling, visualizing, and mining eukaryotic genome data for translational medicine, alongside developing novel approaches for biomedical knowledge representation. Dr. Zheng's innovations include Genome3D, the first 3D genome visualization platform; Ontology Fingerprints, a pioneering method for distributed representation of genes by Gene Ontology terms derived from literature; and early applications of deep learning to predict effective drug combinations. His group has developed a range of deep learning models and, together with colleagues, earned national recognition, placing 2nd in both the 2021 Large Scale Track of the DrugProt BioCreative VII and 2022 NIH/NCATS LitCoin NLP Challenge. Recently, his team leveraged AlphaFold to analyze structural impacts of alternative splicing, laying the groundwork for AI-driven biomedical knowledgebases. Dr. Zheng also directs the Data Science and Informatics Core for Cancer Research (DSICCR) and the Bioinformatics and High-Performance Computing Service Center. His translational research targets cancer, Alzheimer's, and chronic disease, supported by NIH, DoD, and CPRIT. He also serves on editorial boards of two bioinformatics journals.

School of Medicine and Public Health
UNIVERSITY OF WISCONSIN–MADISON