The 14th International Conference on Intelligent Biology and Medicine (ICIBM 2026)

August 2–5, 2026
Buffalo, New York, USA

Hosted by:
The International Association for Intelligent Biology and Medicine (IAIBM)
Jacobs School of Medicine and Biomedical Sciences, University at Buffalo

Table of Contents

Welcome to ICIBM 2026!

On behalf of all our conference committees and organizers, we are thrilled to welcome you to the 2026 International Conference on Intelligent Biology and Medicine (ICIBM 2026). ICIBM is the official conference of The International Association for Intelligent Biology and Medicine (IAIBM, http://iaibm.org/), a non-profit organization dedicated to advancing intelligent biology and medical science through member collaboration, education, and global networking.

As we step into 2026, the fields of bioinformatics, systems biology, and intelligent computing continue to experience rapid advancements, profoundly influencing scientific research and medical innovations. Building on the successes of previous years, ICIBM 2026 is designed to be a platform for interdisciplinary research, dynamic discussions, educational growth, and collaborative opportunities across these evolving fields.

This year, we are excited to present an exceptional line-up of keynote speakers, including Drs. Gary Bader, James Cimino, Mingyao Li, and Ting Wang. Additionally, we are honored to feature four eminent scholar speakers: Drs. Han Liang, Yun Li, Rong Xu, and Chongzhi Zang. These distinguished researchers are global leaders in their respective fields, and we are privileged to have them share their insights at ICIBM 2026. The conference will also include seventeen workshops and one tutorial, along with presentations from faculty members, postdoctoral fellows, Ph.D. students, and trainee-level awardees, selected from outstanding manuscripts and abstracts. These presentations will highlight the innovative technologies and approaches that define our interdisciplinary fields.

We anticipate that this year's program will be invaluable to advancing research, education, and innovation, and we hope you share our enthusiasm for the opportunities ICIBM 2026 will offer. We extend our deepest gratitude to our current sponsors: Admera Health, Singleron Biotechnologies, Pixelgen Technologies, the Computational and Structural Biotechnology Journal, Biomedical Engineering Frontiers, and 10x Genomics.

Finally, our heartfelt thanks go to all members of the ICIBM 2026 committees and our volunteers for their dedication and hard work. Their commitment to making ICIBM 2026 a success is a testament to the strength and resilience of our community.

We hope that the program we have prepared will be thought-provoking, foster collaboration and innovation, and provide an enjoyable experience for all attendees. Thank you for joining us at ICIBM 2026. We look forward to your active participation in all that the conference has to offer!

Sincerely,

Michael Buck	**Hongbo Liu**	**Wenyao Xu**	**Qianqian Song**	**Yijun Sun**	**Zhongming Zhao**
Program Chair	**Program Chair**	**Program Chair**	**General Chair**	**General Chair**	**General Chair**
Professor	Assistant Professor	Professor	Assistant Professor	Professor	Professor
University at Buffalo	University of Rochester	University at Buffalo	University of Florida	University at Buffalo	UTHealth Houston

ACKNOWLEDGEMENTS

General Chairs
Qianqian Song, University of Florida
Yijun Sun, University at Buffalo
Zhongming Zhao, UTHealth Houston

Program Committee
Jonathan Bard, University at Buffalo
Michael Buck, University at Buffalo
James Cai, Texas A&M University
Sapuni Chandrasena, The Ohio State University Wexner Medical Center
Xiao Chang, Children's Hospital of Philadelphia
Jianlin Cheng, University of Missouri Columbia
Yan Cui, University of Tennessee Health Science Center
Lei Du, Northwestern Polytechnical University
Pijush Kanti Dutta Pramanik, National Institute of Technology Durgapur
Shiaofen Fang, Indiana University Indianapolis
Hao Feng, University of Texas Health Science Center at Houston
Mingchen Gao, University at Buffalo
Yixing Han, National Institutes of Health
Md Rejuan Haque, The Ohio State University
Zhe He, Florida State University
Ruifeng Hu, University of Texas Health Science Center at Houston
Weichun Huang, U.S. Environmental Protection Agency
Tao Huang, Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences
Frank Huang, Cincinnati Children's Hospital Medical Center
Zuotian Li, Purdue University
Lei Li, Shenzhen Bay Laboratory
Fuhai Li, The Methodist Hospital, Weill Medical College of Cornell University
Lu Li, University at Buffalo
Aimin Li, Xi'an University of Technology
Shuo Li, University of California, Los Angeles
Li Liao, University of Delaware
Hongbo Liu, University of Rochester
Tao Liu, Roswell Park Comprehensive Cancer Center
Qian Liu, University of Nevada, Las Vegas
Li Liu, Arizona State University
Tianle Ma, Oakland University
Xiaokui Mo, The Ohio State University
Maciej Pietrzak, The Ohio State University
Hong Qin, Old Dominion University
Gang Qu, University of Texas Health Science Center at Houston
Jianhua Ruan, University of Texas at San Antonio
Rama Shankar, Michigan State University
Li Shen, University of Pennsylvania
Mindy Shi, Temple University
Daling Shi, University of Florida

Qianqian Song, University of Florida
Wenyu Song, Brigham and Women's Hospital
Yijun Sun, University at Buffalo
Jiao Sun, University of Central Florida
Fengzhu Sun, University of Southern California
Haixu Tang, Indiana University Bloomington
Manabu Torii, Kaiser Permanente
Alper Uzun, Brown University
Ece Uzun, Brown University
Jun Wan, Indiana University
Shibiao Wan, University of Nebraska Medical Center
Junbai Wang, Radium Hospital
Qing Wang, University of Florida
Kai Wang, University of Pennsylvania
Yufeng Wang, University of Texas at San Antonio
Daifeng Wang, University of Wisconsin - Madison
Jiayin Wang, Xi'an Jiaotong University
Yingying Wei, The Chinese University of Hong Kong
Ka-Chun Wong, City University of Hong Kong
Huanmei Wu, Temple University
Jinchuan Xing, Rutgers University
Wenyao Xu, University at Buffalo
Min Xu, Carnegie Mellon University
Jianhua Xuan, Virginia Tech
Yu Xue, Huazhong University of Science and Technology
Huihuang Yan, Mayo Clinic
Xiaohui Yao, Harbin Engineering University
Rui Yin, University of Florida
Lianbo Yu, The Ohio State University
Shiyang Zhang, Columbia University
Wei Zhang, University of Central Florida
Shaojie Zhang, University of Central Florida
Zhongming Zhao, University of Texas Health Science Center at Houston
Min Zhao, University of the Sunshine Coast
Cuncong Zhong, University of Kansas
Yunyun Zhou, Fox Chase Cancer Center
Qianqian Zhu, Roswell Park Comprehensive Cancer Center

Steering Committee
Peter Elkin, SUNY Buffalo
Marc Halterman, SUNY Buffalo
Song Liu, Roswell Park Cancer Center
Siwei Lyu, SUNY Buffalo
Mattia Prosperi, University of Florida
Jinjun Xiong, UTSA

Publication Committee
Dawei Li, Texas Tech University
Li Liu, Arizona State University
Jun Wan, Indiana University

Workshop Committee
Yuanyuan Fu, University of Hawaii at Manoa
Leng Han, Indiana University
Hongying Sun, University of Rochester
Xun Wu, Columbia University Irving Medical Center
Qiyun Zhu, Arizona State University

Tutorial Committee
Riyue Bao, University of Pittsburgh
Maximilian Haeussler, UCSC Genome Browser
Daofeng Li, Washington University in St. Louis
Qiyun Zhu, Arizona State University

Award Committee
Kai Wang, University of Pennsylvania
Huihuang Yan, Mayo Clinic

Future Scientist in AI Committee
Jingwen Yan, Indiana University
Chi Zhang, Oregon Health & Science University

SCHEDULE

International Conference on Intelligent Biology and Medicine
ICIBM 2026 | August 2–5, 2026

Jacobs School of Medicine and Biomedical Sciences, University at Buffalo
955 Main Street, Buffalo, NY 14203-1121

Sunday, August 2nd, 2026

8:00 AM – 5:00 PM	**Registration**		
9:00 AM – 9:30 AM	**Opening Remarks** (Room 2220A&B)		
	Concurrent Sessions/Workshops		
Room 2220A&B		Room 2213A	Room 2213B
Spatial Single-Cell and Multi-Omics Chair: Yuehua Cui		**AI and Computational Modeling for Biomedical Discovery and Precision Health** Chairs: Yijie Wang, Xun Wu	**Genomic Modeling and Biomolecular Structure** Chair: Shenglin Mei
9:30 AM – 9:50 AM	**Atlas-based analysis of high-resolution spatial transcriptomics** H. Robert Frost Dartmouth College	**Structured and scalable AI for high-dimensional biological data: beyond lasso under correlation** Yijie Wang Indiana University Bloomington	**Architecting for Discovery: Scaling AI-Ready Multimodal Data Infrastructure for Precision Oncology** Stephan Schürer University of Miami
9:50 AM – 10:10 AM	**Tools for comparative spatial transcriptomics** Saurabh Sinha, Georgia Institute of Technology	**Explainable AI models for decoding proteins and microbiome** Yuzhen Ye Indiana University Bloomington	**TFBindFormer: A cross-attention transformer for transcription factor–DNA binding prediction** Ping Liu University of Missouri
10:10 AM – 10:30 AM	**Categorization of 34 computational methods to detect spatially variable genes from spatially resolved transcriptomics data** Guanao Yan Michigan State University	**Reinforcement post-training aligns flow-matching perturbation models with DEG recovery** Xuhong Zhang Indiana University Bloomington	**Comprehensive benchmarking of long-read fusion-detection methods leads to improved fusion discovery** **Pre-recording** Fei Zhang Hong Kong Polytechnic University
10:30 AM – 10:50 AM ***Coffee Break***			
10:50 AM – 11:10 AM	**CySCI: Regularized multitasking graphical attention for integrative single-cell analysis** Xinlei Wang University of Texas, Arlington	**Advancing dietary assessment with computer vision** Jianpeng He Indiana University Bloomington	**KoGNER: A novel framework for knowledge graph distillation on biomedical named entity recognition** Heming Zhang Washington University

11:10 AM – 11:30 AM	**Identification of high-risk cells in single-cell spatially resolved transcriptomics data using transfer learning and spatial smoothing** Debolina Chatterjee University of Mississippi	**Detecting precise adverse drug events from real-world data** Pengyue Zhang Indiana University School of Medicine	**Multiple versus pairwise sequence alignments for protein phylogenetics using foundation models** Rohan Alibutud Temple University
11:30 AM – 11:50 AM	**Decoding pediatric scleroderma across scales using spatial omics and histology** Wei Chen University of Pittsburgh	**Unbiased in vitro and in vivo CRISPR screens identify novel regulatory genes in macrophage efferocytosis** Xun Wu Columbia University	**ChromDiffusion: Refining chromatin structure characterization through diffusion models** Tomoya Kanno Temple University
11:50 AM – 12:10 PM	**Decipher non-coding SNPs by high-resolution reconstruction of single-cell spatial genome architectures in 3D space** Jianrong Wang Michigan State University	**Aging Synergizes with Hypercholesterolemia to Drive Smooth Muscle Cell Transdifferentiation** Mingqiao Li Tulane University	**A structural antibody benchmark of AlphaFold3 reveals hallucinated epitopes and a bias for orderness** Arnav Solanki UTHealth Houston
12:10 AM – 12:30 PM	**Multimodal AI for tumor microenvironment characterization and therapeutic prediction** Qianqian Song University of Florida	**Model-guided reconstruction of single-cell secretion dynamics during live CAR-T/tumor-cell interactions** Yujing Song New York University	
12:30 PM – 2:00 PM	***Lunch Break*** (Room 1220)		

CONCURRENT SESSIONS/WORKSHOPS

	Room 2220A&B	Room 2213A	Room 2213B	Room 1220
	Eminent Scholar *Dr. Han Liang* **Computational and AI models to decode disease mechanisms based on chromatin and multi-omics integration** Chair: Jianrong Wang	**AI and Multi-Modal Integration for Disease Characterization** Chair: Shibiao Wan	**AI and Statistical Methods for Genomics and Multi-Omics** Chairs: Wanding Zhou, Gang Peng	**From Data to Models to Agents: Using LLMs to Accelerate Clinical Development and Personalize Patient Care** Chairs: Huanmei Wu, Yuzhou Chen
2:00 PM – 2:20 PM	***Eminent Scholar: Comprehensive Characterization of N6-methyladenosine RNA modification in Cancer*** *Han Liang* *MD Anderson Cancer Center*	**Agentic-AI for hypothesis generation for aging and drug repurposing** Kyoung Jae Won Cedars-Sinai Medical Center	**Advancing the Interpretation of Human Genetic Variation Using Generative AI** Xinghua (Mindy) Shi Temple University	**Dual-stream cross-modal alignment for clinical trial outcome prediction** Yuzhou Chen University of California, Riverside

2:20 PM – 2:40 PM	**Combining natural genetic variations with single molecule footprinting to learn about the rules of chromatin regulation** Xin He University of Chicago	**An optimal-transport based multi-modal integration for categorizing T-cell acute lymphoblastic leukemia** Shibiao Wan University of Nebraska Medical Center	**Development of Epigenetic Clocks from Oxford Nanopore Methylation Data in Human Caudate Tissue** Gang Peng Indiana University	**SepsisGuard-ICU: an evidence-grounded multi-agent system for safe and explainable sepsis decision support in the ICU** Lei Liu Fudan University
2:40 PM – 3:00 PM	**Robust causal inference from imperfect CRISPR interventions in Perturb-seq: Heterogeneous Perturbation Strength and Off-target Bias** Sunduz Keles University of Wisconsin-Madison	**ShinyFeatures: a Shiny App for prioritizing features for machine learning models** Timothy Shaw H Lee Moffitt Cancer Center	**The Cancer Proteome Atlas: From Functional Proteomics to Next-Generation Analytics** Jun Li University of North Carolina at Chapel Hill	**Developing social work language models for SYNC-Smart AI assistant** Phillip McCarllion Temple University
3:00 PM – 3:20 PM	**Learning single-cell multiomic trajectories using neural ODEs** Christina Leslie Memorial Sloan Kettering Cancer Center	**TENET+: A tool for reconstructing gene networks by integrating single cell expression and chromatin accessibility data** Junil Kim Soongsil University	**Unified Statistical Inference Beyond Gene Expression in Single-Cell RNA Sequencing** Guoshuai Cai University of Florida	**BehaviorGround: a fair benchmark for cognitive decline detection under subjective-complaint contamination** Sambit Mukherjee University of South Carolina
3:20 PM – 3:40 PM ***Coffee Break***				
3:40 PM – 4:00 PM	**From ontology fingerprints to BRAINCELL-AID: literature-driven gene function and brain cell annotation** Wenjin Zheng UTHealth Houston	**Advancing prenatal congenital heart disease screening through artificial intelligence** Jieqiong Wang University of Nebraska Medical Center	**Can We Trust Protein Search in the Era of Large Language Models?** Yang Lu University of Wisconsin	**Leveraging synthetic text and image data to improve the robustness and generalizability of biomedical AI models** Kai Wang Children's Hospital of Philadelphia
4:400 PM – 4:20 PM	**Interpreting deep learning models in genomics via concept attribution** Shaun Mahony Pennsylvania State University	**Machine learning-based small-RNA liquid biopsy enables accurate classification of benign, malignant, and metastatic breast diseases** Xuhang Liu Roswell Park Cancer Center	**Transposable Element–Driven Viral Mimicry Constrains Multistep Pancreatic Cancer Progression** Siyu Sun University of Rochester	**From EHR phenotyping to proactive maternal care: Trustworthy AI for early risk detection and personalized decision support** Huanmei Wu Temple University

4:20 PM – 4:40 PM	**Spatially-informed reference-free deconvolution for spatial transcriptomics with SpatialCD** Yuehua Cui Michigan State University	**Interpretable spatial programs from multi-sample tumor spatial transcriptomics** Yi Zhang Duke University	**Single-molecule RNA modification mapping by AI analysis of nanopore sequencing** Yunxia Wang Boston Children's Hospital	**AI in biomedical research-- hypothesis generation, drug target and activity prediction** Jie Zhang Indiana University
4:40 PM – 5:00 PM	**spatiAlytica: Spatially intelligent multimodal agentic system for decoding tumor ecosystems** Yufei Huang University of Pittsburgh	**Computational techniques for cell-level inference of gene effects on human traits** Jinzhuang Dou The University of Alabama at Birmingham	**Recurrent Methylation Pattern Encoding for Interpretable Analysis of Sparse DNA Methylomes** Wanding Zhou Children's Hospital of Philadelphia	**Prompt-tuned open-source large language model for structured extraction of social determinants of health from clinical notes** Yan Zhuang, Indiana University
5:00 PM – 5:20 PM	**Single-molecule RNA fate modeling by deep learning** Dean Li Boston Children's Hospital	**A rigorous leak-free microbiome pipeline revealing batch effects as the dominant challenge in a multi-cohort immunotherapy response prediction** Amey Garg Santa Clara High School 5:00 PM – 5:10 PM		**Identifying subgroups in patient with low back pain for personalized treatment using machine learning** Leming Zhou University of Pittsburg

Monday, August 3rd, 2026

8:00 AM – 5:00 PM	**Registration**			
8:20 AM – 9:00 AM	**Keynote Lecture: The Human Pangenome Project** **Ting Wang, PhD** (Room 2220A&B)			
CONCURRENT SESSIONS/WORKSHOPS				
Room 2220A		Room 2213A	Room 2213B	Room 2220B
Eminent Scholar *Dr. Yun Li* **Precision Medicine and Cancer Systems** Chair: Rachael Hageman Blair		**Computational Multi-Omics Approaches to Dissect Regulatory Variation and Disease Mechanisms** Chair: Lei Li	**Bridging Methodological Innovation and Real-World Applications in Genetic Studies** Chair: Zikun Yang	**AI and Multi-omics for Modeling Aging and Disease** Chair: Sheng Li

9:00 AM – 9:20 AM	***Eminent Scholar: Multi-omics Understanding of Complex Human Diseases*** *Dr. Yun Li*	**Prediction of rapid fibrosis progression in metabolic dysfunction–associated steatotic liver disease with interpretable machine learning models** Tahmina Sultana Priya Virginia Tech	**fastGxE: Powering genome-wide detection of genotype-environment interactions in biobank studies** Xiang Zhou Yale University	**Transcriptional heterogeneity predicts and shapes clonal selection in aging hematopoiesis** Sheng Li University of Southern California
9:20 AM – 9:40 AM	**Gain and loss of DNA methylation during tumor evolution** Deyana Tabatabaei Temple University	**Atlas of cell type-specific post-transcriptional genetic impacts in immune-related diseases** Lei Li Shenzhen Bay Laboratory	**Distributional genetic effects reveal context-dependent molecular regulation in human brain aging and Alzheimer's disease** Tianying Wang Colorado State University	**Liquid biopsy biomarkers for cancer detection and disease monitoring** Bodour Salhia University of Southern California
9:40 AM – 10:00 AM	**Deep learning for early hepatocellular carcinoma risk prediction in patients with chronic liver diseases and diabetes: insights from retrospective cohort analysis** Jinlian Wang UT Austin	**Multi-omics integration predicts 17 disease incidences in the UK biobank** Quan Sun University of Pennsylvania	**Powerful statistical frameworks for genome-wide survival association analysis** Shiyang Ma Shanghai Jiao Tong University	**A transcriptomics-native foundation model for universal cell representation and virtual cell synthesis** Jichun Xie Duke University
10:00 AM – 10:20 AM	**An immune cell signature for sepsis with multiple organ dysfunction syndrome** Rama Shankar Michigan State University	**Accurate estimation and correction of open chromatin biases in CUT&Tag data** Sheng'en Shawn Hu University of Virginia School of Medicine	**Interplay of Polygenic and Monogenic Kidney Disease** Atlas Khan Columbia University	**RNA stability as a key link between genetic variation and disease: insights from multi-omics and deep learning** Xinshu Grace Xiao University of California
10:20 AM – 10:40 AM ***Coffee Break***				
10:40 AM – 11:00 AM	**GenePriori: a leakage-aware multi-agent framework for bias-robust gene prioritization with translational oncology relevance** Yuxuan Jiang Southwest Medical University	**A novel admixed population fine-mapping method: mitigating false discovery rate through local genetic ancestry modeling** Kai Yuan Indiana University School of Medicine	**Integrating genomic and proteomic data improves complex trait prediction in diverse populations.** Haoyu Zhang National Cancer Institute	**Precisely integrate single-cell data to decipher aging signatures in the monkey brain under calorie restriction** Chao Zhang Boston University
11:00 AM – 11:20 AM	**Identification of intrinsic pancreatic cancer expression subtypes from**	**Associated cell types and influential genes identified through integration of genome-wide association**	**Multi-ancestry modeling leveraging genetic diversity reveals heterogeneous architecture**	**Integrative AI-driven multi-omics identifies NR2F1 as a key driver of lineage plasticity and**

	patient-derived xenograft samples Orry Huang University of Texas Health Science Center, San Antonio	**studies and single-cell transcriptomics for circulating fatty acid traits** Elijah Sterling University of Georgia	**in primary open-angle glaucoma** Jessica Cooke Bailey East Carolina University	**metastatic progression in prostate cancer** Ping Mu Yale University
11:20 AM – 11:40 AM	**Oncoordinate captures microenvironmental state dynamics and identifies aggressive niches in tumors** Vignesh Venkat Rutgers Cancer Institute & Icahn School of Medicine at Mount Sinai	**Single-cell multiomics reveals malignant plasticity and immune suppression in high-risk pediatric cancers** Wenbao Yu, Temple University	**Signatures of recent positive selection refine novel Primary open-angle glaucoma risk loci** Timothy Ciesielski Case Western Reserve University	**Quantifiable blood TCR repertoire components associated with immune aging** Bo Li University of Pennsylvania
11:40 AM – 12:00 PM	**A density-based spatial functional modularity framework for characterizing tumor functional niches in HER2-positive breast cancer** Fan Jhen Liu National Taiwan University		**Tractor-PGS uses local ancestry to improve prediction accuracy in admixed populations** Yi-Sian Lin Baylor College of Medicine	**Cell-free DNA fragmentomics and methylation as AI-enabled multi-omics biomarkers across cancer and neuroinflammatory disease** Yaping Liu Northwestern University
12:00 PM – 1:00 PM	***Lunch Break*** (Room 1220)			
CONCURRENT SESSIONS/WORKSHOPS				
Room 2220A		Room 2213A	Room 2213B	Room 2220B
Invited Speaker *Dr. Min Xu* **Genome Browser Tutorial** Chair: Daofeng Li		**Biological and Biomedical Ontology Workshop in the AI Era** Chairs: Oliver He, Alexander Diehl	**Cancer Genomics** Chairs: Siyuan Zheng, Xiaojing Wang	**Deep Learning and CRISPR Screening** Chair: Xueqiu Lin
1:00 PM – 1:20 PM	**From Pixels to Biological Reasoning: Semantic, Hierarchical, and Graph-Structured Learning for Interpretable Biomedical AI** Min Xu Mohamed bin Zayed University of Artificial Intelligence & CMU	**Representing dental care-related fear and anxiety** Bill Duncan University of Florida	**An embedding-based framework for statistical testing of LLM-inferred gene-set functions** Yu-Chiao Chiu University of Pittsburgh **Creative applications of AI in biomedical research**	**Decoding Non-Coding Regulatory Mutations in Cancer Using FFPE-CUTAC and Interpretable Deep Learning** Xueqiu (Chu) Lin Fred Hutchinson Cancer Center

1:20 PM – 1:40 PM	**Browser tutorial session introduction** Daofeng Li Washington University 1:20PM-1:30PM	**Pain in both general and specific ontological representation** Alexander Diehl University at Buffalo	Yan Guo University of Miami **Proteogenomics guides tumor systems biology: opportunities and challenges** Chen Huang University of Alabama at Birmingham **Phenylalanine reprograms TLS architecture and plasma cell fate in MASH-HCC** Chunru Lin The University of Texas MD Anderson Cancer Center **HOXA9 maintains cancer-specific 3D genome organization in AML** Qili Shi UT Health San Antonio **From unresolved variants to actionable biomarkers: defining when long-read sequencing is required** Xiaotu Ma, St. Jude Children's Research Hospital **AI-empowered virtual immunopeptidomics of neoantigen immunogenicity** Yuhao Tan The Children's Hospital of Philadelphia / University of Pennsylvania **Genomic alterations across racial/ethnic groups and cancer outcome disparities** Xiaojing Wang UT Health San Antonio	**Harnessing tandem repeats to understand the genetic basis of human diseases** Ya (Allen) Cui University of California
1:40 PM – 2:00 PM	**UCSC Genome Browser: Introduction + 5min QA** **Virtual** Robert Kuhn, UCSC Genomics Institute 1:30PM-2:00PM	**Recent developments in social ontology** Bill Hogan Medical College of Wisconsin		**Structured Modeling of Allelic-specific Gene Expression** **Zoom** William H. Majoros Duke University
2:00 PM – 2:20 PM	**UCSC Brain Explorer + 5min QA** **Virtual** Mary Goldman UCSC Genomics Institute 2:00PM-2:30PM	**Introducing the Injury Ontology** Barry Smith University at Buffalo		**Characterizing the genetic architecture and molecular mechanism of circulating fatty acids** Kaixiong (Calvin) Ye University of Georgia
2:20 PM – 2:40 PM	**WashU Epigenome Browser: Introduction** Daofeng Li Washington University 2:30PM-2:40PM	**The 2026 Cell Line Ontology Updating and CellCards Web Display** Jie Zheng University of Michigan Medical School		**How to Align Unpaired Single-Cell Omics: Benchmarks, Practical Guidelines, and Impacts on Gene Regulatory Modeling** Zhana Duren Indiana University School of Medicine
2:40 PM – 3:00 PM	**Develop visualization portal (example: BICAN & IGVF)** Wenjin Zhang Washington University 2:40PM-2:45PM	**A decade of progress in automated quality assurance of biomedical ontologies** **Zoom** Licong Cui University of Texas Health Science Center at Houston		**Perturbation of allele specific chromatin variants in human colon organoids links gene regulation to non-coding variants associated with digestive disease** Siming Zhao Dartmouth College
3:00 PM – 3:20 PM	**HPRC Epigenome Browser** Tianjie Liu Washington University 2:45PM-2:50PM	**Automating Intervention Discovery from Scientific Literature: A Progressive Ontology Prompting and Dual-LLM Framework** Jinjun Xiong University at Buffalo		**Improving Disease Risk Prediction with Functional Genomics Annotations Using LLMs** Wanwen Zeng Stanford University

3:20 PM – 3:40 PM	**5min QA for WashU Epigenome Browser** 2:50PM-2:55PM **5min Break** 2:55PM-3:00PM **Hands-on for WashU Epigenome Browser** Chad Seng, Washington University 3:00PM-3:40PM **Bring laptop**	**From Biomedical Ontologies to Knowledge Discovery: Ontology-Driven Applications in Ignet 2.0 and Vignet** Junguk Hur University of North Dakota	**Genome-scale discovery of RNA-centered regulation in cancer drug response and immunotherapy** Da Yang University of Pittsburgh	**Interpret human genetic variants via single-cell multimodal omics and AI/ML models** Yang Li Washington University in St. Louis
3:40 PM – 4:00 PM ***Coffee Break***				
	Clinical Imaging, Multimodal AI, and Healthcare Deployment Chair: Shaolei Teng	**Technology Session**	**Special panel: AI in Research and Scientific Writing** Chair: Wei Zhang	**Flash Talks** Chair: Tong Wang
4:00 PM – 4:20 PM	**Matching speech models to ADRD subtypes: A mechanistic benchmark** Xiaoyu Zhang SUNY at Buffalo	**Empowering genomic discovery: Accelerating regulatory biology through multi-omic solutions** Yun Zhao, Zonghui Peng Admera Health (30 minutes) 4:00 PM – 4:30 PM	**TBD** Zhandong Liu	**CO-PICO/PECO: Condition-based occupational PICO/PECO and its applications for LLM extraction and ontological modeling of structured occupational safety and health** Hasin Rehana, University of North Dakota 4:00 PM – 4:10 PM **CIRCA identifies cell- and niche-level interactions with amyloid-β plaques using causal representation learning** Tyler Lovelace UTHealth Houston 4:10 PM – 4:20 PM
4:20 PM – 4:40 PM	**A framework for dynamically training and adapting deep reinforcement learning models to different, low-compute, and continuously changing**	**Unlocking Archived Pathology Specimens: Robust Single-Cell Transcriptomics from FFPE Tissue** Jing Zhou Singleron Biotechnologies (20 minutes)	**From Prompt to Publication: Applying AI Agents to Bioinformatics Research** Wei Zhang University of Central Florida	**Machine learning augmented mechanistic modeling for identifying predictive tumor biomarkers** Alexander Silalahi MD Anderson Cancer Center

	radiology deployment environments **Pre-recording** Vishwa S. Parekh Rice University	4:30 PM – 4:50 PM		4:20 PM – 4:30 PM
				Multimodal protein function inference via vision-language models and structural similarity representations, Renzhi Cao Pacific Lutheran University 4:30 PM – 4:40 PM
4:40 PM – 5:00 PM	**Multimodal deep learning for Alzheimer's disease prediction with Improved training stability and flexibility** Jianan Liu Southeast University	**Protein interactomics by proximity networks of single cells** **Zoom** Michael Förster Pixelgen (30 minutes) 4:50 PM – 5:20 PM	**Agentic Loop Engineering for Bioinformatics Research: Data Gathering, Algorithm Development, and Paper Writing** Tejas Sandip Mahajan University of Central Florida	**scDCBC: Model-based scRNA-seq deep clustering with batch correction** Xiaohe Chen, Florida State University 4:40 PM – 4:50 PM
				A systematic investigation of molecular encoding methods for drug property predictions across neural network and transformer encoder-based model Shan-Ju Yeh, National Tsing Hua University 5:00 PM – 5:10 PM
5:00 PM – 5:20 PM	**RA-CMF: Region-adaptive conditional MeanFlow for CT image reconstruction** Md Shifatul Ahsan Apurba University of Alabama at Birmingham	**Use of Stereo-seq to dissect cell type spatial differences in human brain organoids and mouse eyeball** **Zoom** Elaine Lim MGI US LLC (20 minutes) 5:20 PM – 5:40 PM	**5:00-6:00pm** **Panelists**	**Beyond PCA: Multiple correspondence analysis for robust cell-type annotation and trajectory inference** Tsai-Jung Lu, National Taiwan University UT San Antonio 5:10 PM – 5:20 PM
5:20 PM – 5:40 PM	**WoundWise: A mobile deep learning solution for accurate pressure ulcer staging** Daniel Change University of Nevada, Las Vegas	**Built for What's Next: How 10x's Flex Apex, Xenium, and Atera Platforms Are Powering the AI Era of Biology** Aleksandar Janjic 10x Genomics		

5:40 PM – 6:00 PM	**ICA-based analysis of progressive multi-scale structural–functional fusion for Alzheimer's disease** Yanteng Zhang TReNDS Center	(20 minutes) 5:40 PM – 6:00 PM		
6:00 PM – 7:30 PM	**Poster Session** (Atrium)			

Tuesday, August 4th, 2026

8:00 AM – 5:00 PM	**Registration**			
8:20 AM – 9:00 AM	**Keynote Lecture: Mapping the Multiscale Human** **Gary Bader, PhD** (Room 2220A&B)			
CONCURRENT SESSIONS/WORKSHOPS				
Room 2220A		Room 2213A	Room 2213B	Room 2220B
Invited Speaker *Dr. Zemin Zhang* **AI-Driven Multi-Omics for Disease Mechanisms and Therapeutic Discovery** Chairs: Hongying Sun, Kaifu Chen		**From Metagenomes to Mechanisms: Functional Microbiome Discovery** Chairs: Qiyun Zhu, Xiaofang Jiang	**AI and Data Science in Health Informatics** Chairs: Yu Huang, Yan Zhuang	**Integrative omics to understand and predict disease progression** Chairs: Jingwen Yan, Sha Cao
9:00 AM – 9:20 AM	***Invited Speaker:*** **Decoding the tumor microenvironment by pan-cancer single cell analysis** Zemin Zhang Peking University	**Spatiotemporal microbiome dynamics: from early disease prediction to single-cell functional analysis and sorting** Shi Huang University of Hong Kong	**Agentic AI for complex clinical informatics workflows** Xing He Indiana University School of Medicine	**Continuous Modeling of Disease Progression in Alzheimer's Disease and Related Dementias** Lu Zhang Indiana University Indianapolis
9:20 AM – 9:40 AM	**Longitudinal epigenome-wide associations and pre-chemotherapy prediction of post-chemotherapy inflammation in patients with breast cancer undergoing chemotherapy** Hongying Sun University of Rochester	**Decoding the gut microbiome's functional dark matter via comparative genomics** Xiaofang Jiang National Institutes of Health (NIH)	**Multi-modal AI using neuroimaging genomics to study Alzheimer's and dementia** Da Ma Wake Forest University	**Stability-aware integrative clustering for characterizing disease progression from multi-modal data** Rachael Hageman Blair New York University at Buffalo

9:40 AM – 10:00 AM	**Quantum computing for single-cell biology: networks, dynamics, and generative models** James J. Cai Texas A&M University	**Microbiome-Informed Therapeutic Peptide Discovery** Shanlin Ke Ohio State University	**Immune digital twin and vascular disease** Juilee Thakar University of Rochester	**Continuous staging of amyloid progression in Alzheimer's disease with enhanced sensitivity to early changes** Jingwen Yan Indiana University
10:00 AM – 10:20 AM	**Cross-Trait PRS and data-driven subtyping to uncover obesity heterogeneity and enable precision risk prediction** Shulan Tian Mayo Clinic	**Microbiome-based early screening of intestinal tumors driven by intelligent sensing and privacy protection** Xiaoquan Su Qingdao University	**Unlocking clinical development innovation with artificial intelligence and real-world evidence** Chuan Gao Johnson & Johnson Innovative Medicine	**AD-LLaVA-3D: Forecasting Alzheimer's Disease Progression with a Unified Multimodal Video-Language Framework** Sha Cao Oregon Health & Science University
10:20 AM – 10:40 AM ***Coffee Break***				
10:40 AM – 11:00 AM	**Agentic AI for atlas analysis of cell-cell communications** Kaifu Chen Harvard Medical School	**Multi-dimensional virome profiling from bulk metagenomes reveals non-redundant signatures of gut dysbiosis** Zheng Sun Harvard University	**Agentic AI for automated RNA structural motif analysis and biomedical discovery** Jie Hou Michigan State University	**Staging and Pseudotime Inference of Alzheimer's Disease Progression Using Multimodal Imaging Data** Zixuan Wen University of Pennsylvania
11:00 AM – 11:20 AM	**Cellular and molecular dynamics of smooth muscle cells during aortic disease development** Yanming Li Maine Health Institute for Research	**Predicting metabolite response to dietary intervention using deep learning** Tong Wang Purdue University	**LLM-Agent-Based clinical trial cohort building: an experimental evaluation using MIMIC-IV** Hao Liu Montclair State University	**From 2D Snapshots to 3D Molecular Landscapes: Reconstructing Volumetric Spatial Transcriptomics with AI** Wei Zhang University of Central Florida
11:20 AM – 11:40 AM	**Computational biology in early drug discovery: turning complex data into biological insights** Yu Sun Biogen	**Calculating differential genome coverages among metagenomic sources** Yuhan Weng University of California San Diego	**A scalable multi-omics representation framework for disease modeling across proteomic, single-cell, and spatial transcriptomic data** Daisy Shen Moffitt Cancer Center	**How Much Biological Knowledge Is Encoded in LLM-Derived Gene Embeddings?** Jun Li Notre Dame University

11:40 AM – 12:00 PM	**Machine learning-based prediction of gene expression using RNA motifs and structural properties** Mingyi Zhu Yale School of Medicine	**Enabling AI-driven discovery in complex biological systems with scikit-bio** Qiyun Zhu Arizona State University	**Cross-Hospital privacy-preserving distributed AI** Tsung-Ting Kuo Yale University	**Toward integrative modeling of drug response in acute myeloid leukemia** Olga Nikolova Oregon Health & Science University
12:00 PM – 1:00 PM	***Lunch Break*** (Room 1220)			
1:00 PM – 1:40 PM	**Keynote Lecture: A Research Agenda for Clinical Informatics in the Post-AI, Post-EHR Vendor World** **James Cimino, MD (Room 2220A&B)**			
CONCURRENT SESSIONS/WORKSHOPS				
Room 2220A		Room 2213A	Room 2213B	Room 2220B
Eminent Scholar *Dr. Chongzhi Zang* **Multi-Omics, Microbiome, and Metabolomics** Chair: Hao Liu		**Single-Cell, Spatial, and Regulatory Networks** Chair: Lu Li	**Flash Talk** Chair: Yufang Jin	**Future Scientist Session** Chair: Chi Zhang, Jingwen Yan, Arhan Patel
1:40 PM – 2:00 PM	***Eminent Scholar: Chromatin and Gene Regulation through Transcriptional Condensates*** *Dr. Chongzhi Zang*	**Single-cell analysis of intracellular transport and expression of cell surface proteins** Rekha Mudappathi Arizona State University	**Nanoparticle-Induced Thermogenic Adipocytes Improve Endothelial Tube Formation and Network** Parker Olkowski UT San Antonio	Derek Ai Park Tudor High School
			Electroencephalography (EEG) for Remote Vehicle Control Applications Seth Johnston UT at San Antonio	Christopher Barker Pacific Lutheran University
2:00 PM – 2:20 PM	**Multi-modality foundation model to link pathology with omics** Guangyu Wang Houston Methodist, Weill Cornell Medicine School	**Knowledge grounded multi-agent orchestration for interpretable functional summarization of single-cell transcriptomics** Jinlian Wang UTHealth Houston	**Hardware-Oriented Programming for an Autonomous Robot** Angela Zhang Basis San Antonio Shavano Campus	Joshua Bie Havard Westlake School
			Bridging Computer Vision and Food Policy: A YOLOv11 Produce Freshness Detector Aligned with California AB 660	Alyssa Chen Carrollwood Day High School Tim Chu

			Date Label Vocabulary Aston Liu The Kinkaid School, Houston	Forest Hills Central High School
2:20 PM – 2:40 PM	**Analyzing soil microbial diversity in organically managed fields amended with dredged material in northwest ohio** Shikshya Gautam Bowling Green State University	**miR-CellMap: a graph-transformer framework for single-cell microRNA-gene regulatory analysis** Jane Ohia University of Nebraska-Lincoln	**DUET-Knockoffs: FDR-Controlled Cross-Platform Feature Selection from Paired 16S and Shotgun Microbiome Count Data** Shiyuan Wang UT at San Antonio	Bianca Ehling The Ohio State University
			An AI-Powered Single-Cell and Spatial Omics Ecosystem for Neurodegenerative Disease Weidong Wu Ohio State University	Katelyn Hur Red River High School Jason Li Westwood High School
2:40 PM – 3:00 PM	**DUET-Knockoffs: fdr-controlled cross-platform feature selection from paired 16S and shotgun microbiome count data** Yiqiao Zhu The University of Texas at San Antonio & University of Michigan	**Cell segmentation using spatial transcriptomic data from the multi-objective optimization perspective** Aleksandar Necakov, Brock University	**BindFuse: A Multimodal Graph Learning Framework for Protein-Ligand Binding Affinity Prediction** Bilal Ahmad University at Buffalo	Isabella Jiayi Li Westlake High School Addison Liu Park Tudor School
			Integrative Mendelian Randomization and Protein-Protein Interaction Network Analysis Links Multi-Organ Aging to Complex Diseases Chunxuan Ma Cleveland Clinic	Emily Liu The Bishop Strachan School
3:00 PM – 3:20 PM ***Coffee Break***				
3:20 PM – 3:40 PM	**Multiomic integration reveals regulatory function of GWAS variants within transposable elements in brain aging** Sreejato Chatterjee, University of Rochester	**MolGene-E: inverse molecular design to modulate single cell transcriptomics** Rahul Ohlan The City University of New York	**STITCH: A Platform-Agnostic Pipeline for Barcoding-Based Spatial Transcriptomics** Leo Liu Mayo Clinic	Sunny Lowel Brock University Alex Mao Winston Churchill High School
			Strategies for Robust, Accurate, and Generalizable	Emma K. Molnar Waterford School Christina Teng

			Benchmarking of Drug Discovery Platforms Melissa Van Norden University at Buffalo	Thomas S. Wootton High School Alexander Wu Arizona State University
3:40 PM – 4:00 PM	**metaSRNA: a unified framework for reprocessing bacterial data from microrna sequencing protocols and drafting bacterial small RNA catalogs** Mengyuan Zhou University of Nebraska Lincoln	**Dysregulated intercellular communication and signaling pathways across brain regions in Alzheimer's disease** Hani Alostaz UTHealth Houston	**Scalable and Reproducible Software for Microbiome Data Analysis** Andres Benavides Arevalo Universidad Industrial de Santander **MetaAge: A Highly Accurate Ensemble Machine Learning Model for Biological Age Estimation** Liguo Wang Mayo Clinic	
4:00 PM – 4:20 PM	**SPLISUM: spectral prediction and library searching for untargeted metabolomics** Chhavi Thakur Indiana University, Bloomington	**EVTS: a topology-aware metric for evaluating RNA velocity embeddings in single-cell transcriptomics** Xiuquan Wang Tougaloo College	**Building a Psychoplastogenic Signature: A Comparative Evaluation of Compound-Based and Target-Based Approaches in the CANDO Platform** Justin Shinkle University at Buffalo **Benchmarking Variation Detection Methods with Linear Reference Genomes and Pangenomes in Grapevines** Brooke Atamanyk Brock University	Austin Xu Williamsville East High School Shaoyu Xu Shanghai World Foreign Language Academy Adrian Z. Yang Lawrence E. Elkins High School Henry Yang Sage Hill School
4:20 PM – 4:40 PM	**Celltaminate: a bioinformatic pipeline for rapid detection of pathogenic microorganisms** Ahmad Salman Sirajee Rutgers University, New Brunswick	**STEP: Spatial Transcriptomics Embedding Procedure for Multi-scale Biological Heterogeneities Revelation in Multiple Samples** Xiaojiang Xu Tulane University		Alina Yu Germantown Friends School

Flash Talks Chair: Jun Wan		**Flash Talks** Chair: Zhaohui Xu, Xiaojiang Xu	**Flash Talks** Chair: Shengyu Li	
4:40 PM – 4:50 PM	**Cross-species gene set and cell type annotation through large language models and agentic AI with literature-grounded validation** Wenbo Chen UTHealth Houston	**SomAtt: a foundation model for the tumor genome** Avinash D Sahu University of New Mexico	**Improved drug discovery through accurate prediction of adverse drug reactions within the CANDO platform** Vida Bodaghi-Namileh University at Buffalo	Daniel Zhang Canyon Crest Academy
4:50 PM – 5:00 PM	**Drug-drug interaction prediction using proteomic scale signatures within the CANDO platform** Lucille Tomin University at Buffalo	**AI-based Characterization of Alzheimer's Disease Phenotypes from population scale single cell data** Chenfeng He University of Wisconsin-Madison	**Prevalence and Genetic Determinants of Chronic Diseases in a University Student Population: A Cross-Sectional Analysis with International Comparisons** Afagh Bapirzadeh Islamic Azad University **Pre-recording**	Simon Zhang Lexington High School
5:00 PM – 5:10 PM	**L2b-Probiotics: A Fusion Genetic Language model for probiotics identification** Caihao Sun The University of Hong Kong	**FACET: criterion-anchored semantic phenotype matching with a guarded LLM for DSM-5-TR face validity of psychiatric mouse models** An Le SUNY Upstate Medical University	**AI in Fast Detection of Diseases** Zarrin Minuchehr National Research Institute of Genetic Engineering and Biotechnology, Tehran-Karaj Highway **Pre-recording**	Eva Zhao Middlesex school
5:10 PM – 5:20 PM	**Single-Cell viral detection from unmapped scRNA-seq reads** Mohammadamin Mahmanzar University of Hawaii at Manoa			Katherine Zhou Choate Rosemary Hall
				Shashvat Hariharan Linganore High School

Wednesday, August 5th, 2026

8:00 AM – 5:00 PM	**Registration**
8:30 AM – 9:10 AM	**Keynote Lecture: Digital Tissue Twins: Shaping the Future of Precision Medicine** **Mingyao Li, PhD** (Room 2220A&B)
9:10 AM –	**Award Ceremony** (Room 2220A&B)

9:50 AM			
CONCURRENT SESSIONS/WORKSHOPS			
Room 2220A&B		Room 2213A	Room 2213B
Computational Drug Discovery and Therapeutic Modeling Chair: Zackary Falls		**Multi-omics, AI, and the Future of Precision Health** Chair: Yuanyuan Fu	**Emerging BioHealth and Environmental Frontiers** Chair: Jane Ohia
9:50 AM – 10:10 AM	**KGPredict-Com: a knowledge graph model for drug combination prediction** Zhenxiang Gao Case Western Reserve University	**From Genetic Susceptibility to Modifiable Biology: Multi-Omic and Dietary Integration Toward Precision Alzheimer's Disease Prevention** Yuxi Liu Harvard T.H. Chan School of Public Health	**Weakly-supervised topology-preserving 3D mitochondria segmentation** Elliott Seo Stony Brook University,
10:10 AM – 10:30 AM	**Sparsity, structure, and interpretability in biologically informed neural networks for drug response prediction** Jason Chia Purdue University & Argonne National Laboratory	**Evaluation of statistical differential analysis methods for identification of senescent cells using single-cell transcriptomics** Dongmei Li University of Rochester	**Learning-augmented resource allocation for intelligent healthcare scheduling** Fei Li George Mason University
10:30 AM – 10:50 AM	**A DALEX-based explainable ai framework for lymphography prediction** **Zoom** Tapas Si University of Engineering & Management, Jaipur	**TRACED: A Graph-Based Algorithm to Infer Single-cell Time-lagged Gene Regulation** Yang Yu Boston Children's Hospital	**Hyperspectral deep learning for compositional classification of municipal solid waste: an environmental health informatics application** Jason Li Westwood High School
10:50 AM – 11:10 AM ***Coffee Break***			
11:10 AM – 11:30 AM	**TransGen: de novo enzyme design via catalytic-conditioned sequence generation** **Zoom** Hongming Cai South China University of Technology	**Risk-stratified hepato-pancreato-biliary cancer screening with self-evolving trustworthy AI** **Zoom** He Ren The Affiliated Hospital of Qingdao University	**XLCenter: crosslink-induced truncation site identification reveals sequence and structural specificities of RNA binding protein motifs** Yuping Zhang University of Connecticut
11:30 AM – 11:50 AM	**DrugPTM-Bench: a large-scale dataset for predictive modeling of drug-induced cell type-specific protein post-translational modifications** **Zoom** Amitesh Badkul	**Genetic heterogeneity underlying breast cancer subtypes: genetic epidemiologic investigations to disentangle disease complexity** Xiang Shu	**Machine learning-derived subtypes of autism reveal distinct genetic contributions from common and rare variants** **Zoom** Chang Shu University of Southern California

	City University of New York	Memorial Sloan Kettering Cancer Center	
11:50 AM – 12:10 AM	**From generalist to specialist: evolution of PS2 α-integrins and implications for drug targeting** Shixi Sun Arizona State University	**Leveraging gene-environment interactions to understand colorectal cancer risk** Mulong Du Nanjing Medical University	
12:10 AM – 12:30 PM	**A reproducible hybrid QSPR-machine learning framework integrating graph-theoretic and cheminformatics descriptors for anticancer drug property prediction** **Zoom** H Muhammad Fraz COMSATS University Islamabad & Riphah International University	**Beyond exposure: computational calibration and ancestry-aware genetics of skin carotenoid spectroscopy for precision nutrition** Yixing Han National Institutes of Health (NIH)	
12:30 PM – 12:40 PM	**Closing Remarks** (Room 2220A&B)		

KEYNOTE SPEAKERS

Keynote Speaker
Gary Bader, Ph.D.
Tuesday, August 4, 2026
8:20 AM – 9:00 AM
Room: 2220A&B

Dr. Gary Bader is a Professor at the Donnelly Centre for Cellular and Biomolecular Research and in the Departments of Molecular Genetics and Computer Science at the University of Toronto, where he holds the Ontario Research Chair in Biomarkers of Disease. An internationally recognized leader in computational biology, Dr. Bader develops computational methods for understanding biological systems at cellular and tissue levels. His laboratory integrates molecular-interaction networks, biological pathways, and bulk and single-cell "omics" data to construct mechanistic models of normal and disease phenotypes, with applications in development, cancer, regenerative wound healing, and precision medicine. His group has made major contributions to biological network and pathway analysis and to the interpretation and visualization of genomic data. Dr. Bader earned his B.Sc. in Biochemistry from McGill University and completed his Ph.D. at the University of Toronto, where he developed the Biomolecular Interaction Network Database. He subsequently conducted postdoctoral research with Chris Sander at Memorial Sloan Kettering Cancer Center. His current research advances an ecosystem-based understanding of tissue function by connecting intracellular pathways, cell–cell interactions, and tissue organization to uncover disease mechanisms and identify potential therapeutic targets.

Title: Mapping the multiscale human

Abstract: Generative models have the potential to help us understand how genomes encode phenotypes and biological functions. We consider the human genome as analogous to a 'natural' generative model encoding the full complexity of an individual, from cellular architecture to physiological functions. According to this perspective, generative models can provide a unified framework to model the human body across scales and to capture factors determining health and disease. This talk will cover example models of parts of the human body across spatial scales and time.

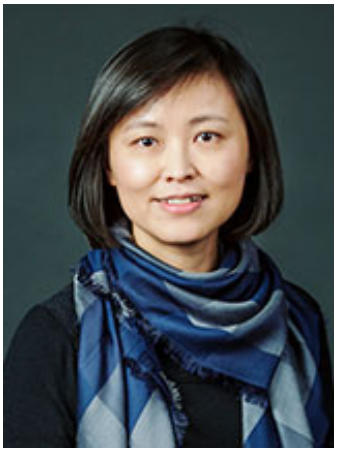
Keynote Speaker
James Cimino, M.D.
Tuesday, August 4, 2026
1:00 PM – 1:40 PM
Room: 2220A&B

Dr. James J. Cimino is a Distinguished Professor of Medicine and Chair of the Department of Biomedical Informatics and Data Science at the University of Alabama at Birmingham's Marnix E. Heersink School of Medicine. A board-certified practicing internist and internationally recognized leader in biomedical informatics, Dr. Cimino has made foundational contributions to medical concept representation, controlled clinical terminologies, clinical decision support, electronic health records, and clinical research informatics. He earned his undergraduate degree from Brown University and his M.D. from New York Medical College, completed an internal medicine residency at St. Vincent's Hospital in New York, and pursued fellowship training in medical informatics at Massachusetts General Hospital and Harvard. Before joining UAB in 2015 as the inaugural director of the Informatics Institute, he spent two decades on the faculty of Columbia University and seven years leading the Laboratory for Informatics Development at the NIH Clinical Center and National Library of Medicine. His current work focuses on improving electronic health records and developing information systems that support clinicians, patients, and clinical and translational researchers. Dr. Cimino is a member of the National Academy of Medicine, a Fellow of the American College of Medical Informatics and the American College of Physicians, and a recipient of the American Medical Informatics Association's Morris F. Collen Award for Excellence.

Title: A Research Agenda for Clinical Informatics in the Post-AI, Post-EHR Vendor World

Abstract: The remarkable success of artificial intelligence and the maturation of commercial electronic health record (EHR) platforms have fundamentally changed the landscape of clinical informatics research. Problems that once defined the field are increasingly addressed by artificial intelligence (AI) methods or incorporated into commercial EHR products, challenging investigators to identify where the next generation of transformative research will emerge. At the same time, growing emphasis on learning health systems and equitable healthcare has expanded expectations for informatics beyond supporting individual clinical decisions to improving health outcomes across patients, populations, and healthcare systems.

This presentation proposes a research agenda for clinical informatics in this post-AI, post-EHR vendor era. The agenda emerged from an analysis of 240 publications selected by American Medical Informatics Association (AMIA) Working Group members for the annual Biomedical Informatics Year in Review. Rather than viewing artificial intelligence, learning health systems, and health equity as separate domains, the analysis revealed that the most compelling advances occur at their intersection. These observations led to a conceptual framework that integrates these themes and identifies high-impact research opportunities that remain largely unaddressed by current AI technologies and commercial EHR platforms. Using representative papers from the Year in Review, the presentation will illustrate this framework and argue that the future of clinical informatics lies not in competing with AI or EHR vendors, but in developing methods that enable health systems to learn continuously, deliver equitable care, and translate knowledge into measurable improvements in health.



Keynote Speaker
Mingyao Li, Ph.D.
Wednesday, August 5, 2026
8:00 AM – 8:40 AM
Room: 2220A&B

Dr. Mingyao Li is a Professor of Biostatistics in the Department of Biostatistics, Epidemiology and Informatics at the University of Pennsylvania's Perelman School of Medicine. She serves as Chair of the Graduate Program in Biostatistics and Director of the Statistical Center for Single-Cell and Spatial Genomics, and she holds a secondary appointment in the Department of Statistics at the Wharton School. Dr. Li earned her B.S. and M.S. degrees in Mathematics from Nankai University and her Ph.D. in Biostatistics from the University of Michigan before joining the University of Pennsylvania faculty in 2006. Her research lies at the intersection of statistical genetics and genomics, bioinformatics, computational biology, and machine learning. A central focus of her work is developing statistical and computational methods for single-cell and spatial-omics data to characterize cellular heterogeneity, gene-expression diversity, cell-state transitions, and communication among cells in human tissues. She collaborates extensively with biomedical investigators studying cardiometabolic diseases, age-related macular degeneration, Alzheimer's disease, chronic kidney disease, type 1 diabetes, and cancer. Through this work, Dr. Li seeks to translate complex genomic data into biological insights that can guide cell-specific functional studies, disease modeling, biomarker discovery, and precision therapeutic development.

Title: Digital tissue twins: shaping the future of precision medicine

Abstract: Spatial omics technologies are reshaping how we study the molecular organization of tissues, offering unprecedented opportunities to map biology in its native spatial context. However, key challenges remain in resolution, scalability, and accessibility. In this talk, I will describe how my lab is developing AI-powered approaches that integrate spatial omics with histopathology to overcome these barriers. I will highlight computational strategies that make spatial omics experiments smarter, faster, and more informative, from enhancing spatial resolution beyond current experimental limits, to extending analyses across large and complex tissues, to integrating diverse molecular measurements within a unified framework. Together, these approaches enable the creation of "digital tissue twins," computational representations of tissues that capture both molecular and structural organization. Finally, I will discuss our efforts to design AI systems that are transparent and trustworthy, ensuring that these tools can support robust biological discovery and clinical translation. These advances aim to unlock the full potential of spatial omics and enable new insights in the era of precision medicine.

Keynote Speaker
Ting Wang, Ph.D.
Monday, August 3, 2026
8:20 AM – 9:00 AM
Room: 2220A&B

Dr. Ting Wang is the inaugural Sanford C. and Karen P. Loewentheil Distinguished Professor of Medicine and Head of Genetics Department at Washington University School of Medicine in St. Louis. Dr. Wang studies the genetic and epigenetic impact of transposable element (TE) on gene regulation. His group is known for defining the widespread contribution of TEs to the evolution of gene regulatory networks as well as to the 3D genome architecture, and for revealing that epigenetic dysregulation of TEs is a major mechanism driving oncogenesis. His lab is home to the WashU Epigenome Browser, utilized by investigators around the world to access hundreds of thousands of genomic datasets generated by large Consortia including the NIH Roadmap Epigenome Project, ENCODE, 4D Nucleome, and TaRGET. Dr. Wang currently directs the NIEHS Environmental Epigenomics Data Center, the Human Pangenome Reference Consortium, the IGVF Data Administrative and Coordination Center, the SMaHT Network Organization Center and Genome Characterization Center, and the Multi-Omics for Health and Diseases Data Production Center.

Title: The Human Pangenome Project

Abstract:

The Human Pangenome Project is creating the next generation of human genome references. By moving beyond a single linear genome and embracing the full spectrum of human genetic diversity, the Human Pangenome Reference provides a graph-based, telomere-to-telomere framework for representing genomic variation at unprecedented resolution. Together with a rapidly expanding ecosystem of computational methods and resources, the project is laying the foundation for a new era of genomic analysis. Like the Human Genome Project before it, the Human Pangenome Project has the potential to transform biological discovery and precision medicine for decades to come.

In this lecture, I will discuss the current status of the Human Pangenome Reference Consortium (HPRC), with an emphasis on its second release, which includes 466 near-complete T2T haplotype assemblies spanning globally diverse populations. I will highlight emerging applications of pangenome-enabled analyses and present examples demonstrating how these resources improve variant discovery, genomic interpretation, and our understanding of human biology. Finally, I will discuss future directions and opportunities for accelerating pangenome adoption across the genomics community.

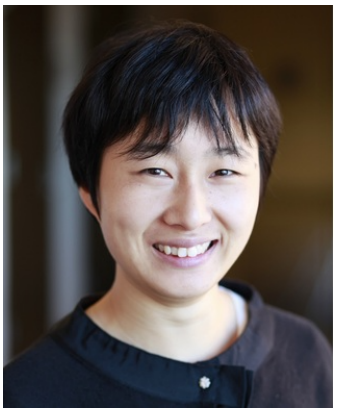
EMINENT SCHOLAR TALKS

Eminent Scholar Award Speaker
Han Liang, Ph.D.
Sunday, August 2, 2026
2:00 PM – 2:20 PM
Room: 2220A&B

Dr. Han Liang is the Barnhart Family Distinguished Professor in Targeted Therapies and Interim Chair of the Department of Bioinformatics and Computational Biology at The University of Texas MD Anderson Cancer Center, with a joint appointment in Systems Biology. He earned his B.S. in Chemistry from Peking University and his Ph.D. in Quantitative and Computational Biology from Princeton University, followed by postdoctoral training in evolutionary and computational genomics at the University of Chicago. Since joining MD Anderson in 2009, Dr. Liang has established an internationally recognized research program focused on bioinformatics tool development, integrative cancer-genomics analysis, RNA regulation and modification, artificial intelligence, and cancer systems biology. His laboratory has developed several widely used computational platforms—including TCPA, TANRIC, FASMIC, and DrBioRight—that enable researchers to analyze complex multi-omics data and translate large-scale cancer datasets into biologically and clinically meaningful discoveries. Dr. Liang has held leadership roles in major international cancer-genomics initiatives, including The Cancer Genome Atlas PanCancer Atlas, the International Cancer Genome Consortium Pan-Cancer Analysis of Whole Genomes project, and the National Cancer Institute Genomic Data Commons. His contributions have been recognized through honors including the AACR Team Science Award, the MD Anderson R. Lee Clark Prize, and the AACR Award for Outstanding Achievement in Basic Cancer Research. He is an elected Fellow of the American Association for the Advancement of Science and the American Institute for Medical and Biological Engineering.

Title: Comprehensive Characterization of N6-methyladenosine RNA modification in Cancer

Abstract: N6-methyladenosine (m6A) is the most abundant internal mRNA modification, yet its landscape, regulatory architecture, and functional impact on human cancers remain incompletely understood. Here we generate a comprehensive pan-cancer atlas of m6A methylation by profiling 226 tumor samples across 23 cancer types in The Cancer Genome Atlas (TCGA), quantifying m6A levels at 15,812 high-confidence sites. We show that global m6A patterns segregate largely dictated by tissue lineage. Systematic analysis of somatic mutations reveals widespread and diverse effects on m6A deposition, including disruption, weakening, or creation of DRACH motifs. Multi-omics integration further identifies 15–20% of protein-coding genes whose m6A levels consistently modulate mRNA abundance or protein expression across cancer types. We also characterize the protein-expression landscape of 15 m6A regulators across 7,482 TCGA samples and uncover frequent dysregulations arising through diverse genetic and epigenetic mechanisms. Finally, a deep-learning model integrating cis-sequence and expression features with trans-regulatory context predicts a substantial fraction of m6A sites with high accuracy, enabling inference of m6A variation in large patient cohorts. Together, this study establishes a foundational resource for decoding m6A-mediated gene regulation in tumors and lays a scalable framework for translating epitranscriptomic variation into biological understanding and therapeutic opportunity.



Eminent Scholar Award Speaker
Yun Li, Ph.D.
Monday, August 3, 2026
9:00 AM – 9:20 AM
Room: 2220A&B

Dr. Yun Li is a Professor in the Departments of Genetics and Biostatistics at the University of North Carolina at Chapel Hill and a core faculty member in UNC's Bioinformatics and Computational Biology program. She earned her Ph.D. in Biostatistics from the University of Michigan in 2009 and joined the UNC faculty that same year. Dr. Li's research focuses on developing statistical methods and computational tools to identify the genetic and genomic factors underlying complex human diseases and traits. She is particularly recognized for developing the MaCH and MaCH-Admix genotype-imputation methods, which have become widely used tools in human genetic studies. Her current work encompasses haplotype phasing, genotype imputation, local-ancestry analysis, polygenic risk prediction, multi-omics integration, and the study of three-dimensional chromatin organization. A major objective of her research is to improve the accuracy and clinical usefulness of genetic analyses across diverse ancestral populations, helping address disparities in genomic research and precision medicine. Dr. Li also collaborates extensively on studies of blood-cell traits, neuropsychiatric conditions, Alzheimer's disease, and other complex disorders. Through the integration of biostatistics, population genetics, and computational genomics, her research provides new insights into disease biology and supports the development of more inclusive approaches to risk assessment, prevention, and personalized treatment.

Title: Multi-omics Understanding of Complex Human Diseases

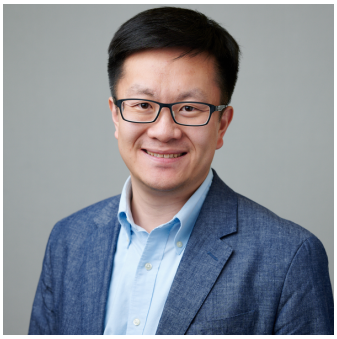
Abstract: Traditional clinical predictors, while valuable, often fail to capture the full complexity underlying disease risk. Multi-omics data offer complementary, deeper molecular perspectives on human health, and their integration holds significant promise for improving disease risk prediction, refining prognosis, and enabling more personalized care. However, large-scale studies systematically evaluating the added predictive value of combining multiple omics layers, and methods that fully leverage the richness of such data, remain limited. To address this, we first evaluated whether adding metabolomics and proteomics data to traditional clinical predictors improves risk prediction for 17 common incident diseases. Using 159 NMR-based metabolites and 2,923 Olink proteins from 23,776 UK Biobank participants, we fit Cox proportional hazard models and assessed performance via Harrell's C-index. Omics data significantly improved prediction across all 17 diseases compared to clinical predictors alone, with proteomics-only models generally outperforming metabolomics-only models, and key contributing proteins, including established markers like KLK3 for prostate cancer and novel ones such as PRG3 for skin cancer, were identified. Building on this, we have developed OMEGA (Omics Multi-modality Embedding via Graphical & Articulated data), a supervised deep learning method that integrates tabular, text, and image representations of omics data, fused via a mixture-of-experts model. Applied to the same 17 endpoints, OMEGA outperformed mixOmics and MOGONET, with mean ΔC-index gains of 0.043 (p=7.7e-15) and 0.103 (p=2.9e-37), respectively, and the largest improvements for breast cancer, glaucoma, and prostate cancer (mean ΔC = 0.168, 0.095, 0.067). Together, these results demonstrate that multi-omics integration, particularly through multimodal representations, meaningfully enhances disease risk prediction.

Eminent Scholar Award Speaker
Zemin Zhang, Ph.D.
Tuesday, August 4, 2026
9:00 AM – 9:20 AM
Room: 2220A&B

Dr. Zemin Zhang is a leading researcher in cancer genomics, bioinformatics, and tumor immunology. He earned his bachelor's degree in Genetics from Nankai University, his Ph.D. in Biochemistry and Molecular Biology from Pennsylvania State University and completed his postdoctoral training at the University of California, San Francisco. Before joining Peking University, Dr. Zhang spent more than 16 years at Genentech/Roche, where he led the cancer genomics and bioinformatics group and contributed to the discovery of novel anticancer targets, therapeutic biomarkers, and genomic strategies for drug development. His research focuses on computational cancer biology, single-cell genomics, tumor heterogeneity, and the complex interactions among cancer, immune, and stromal cells within the tumor microenvironment. He is particularly recognized for pioneering large-scale studies of tumor microenvironment heterogeneity and for generating detailed immune-cell atlases across multiple cancer types. Through the integration of genomic technologies, computational methods, and translational cancer research, his work has provided important insights into cancer development, treatment response, biomarker discovery, and immunotherapy. Dr. Zhang serves on the editorial boards of several leading journals, including Cell, Cancer Cell, Cancer Discovery, and Cell Research, and has been recognized as both a CUSBEA Scholar and a Cheung Kong Scholar.

Title: Decoding the tumor microenvironment by pan-cancer single cell analysis

Abstract: TBD



Eminent Scholar Award Speaker
Chongzhi Zang, Ph.D.
Tuesday, August 4, 2026
1:40 PM – 2:00 PM
Room: 2220A&B

Dr. Chongzhi Zang is a Professor of Genome Sciences at the University of Virginia, with faculty appointments in Biomedical Engineering and Biochemistry and Molecular Genetics and serves as the inaugural Director of Computational Genomics at the UVA Comprehensive Cancer Center. A computational biologist with expertise in cancer epigenetics and chromatin biology, Dr. Zang develops quantitative models and computational methods for analyzing high-throughput genomic and epigenomic data. His laboratory investigates how chromatin structure and transcriptional regulation control gene activity in normal mammalian cells and how disruption of these processes contributes to cancer and other human diseases. His research integrates bulk, single-cell, and spatial genomics technologies to identify regulatory mechanisms, characterize cellular heterogeneity, and discover biomarkers and therapeutic vulnerabilities that can support precision medicine. Dr. Zang earned his B.S. in Physics from Peking University and his Ph.D. in Physics from George Washington University, where his doctoral research focused on biological physics and computational epigenomics in collaboration with the National Institutes of Health. Before joining UVA in 2016, he completed postdoctoral training in computational biology at Harvard University and Dana-Farber Cancer Institute. His interdisciplinary work brings together physics, data science, genomics, and cancer biology to extract fundamental biological principles from complex datasets and translate them into improved approaches for disease diagnosis and treatment.

Title: Chromatin and gene regulation through transcriptional condensates

Abstract: Transcriptional regulation by chromatin-associated biomolecules is central to gene expression control in eukaryotic cells and is dysregulated in many diseases. Transcriptional condensates, a special class of nuclear biomolecular condensates, have emerged as important regulatory hubs in gene activation and are increasingly implicated in cancer and immune dysfunction. However, computational genomics approaches for systematically characterizing their regulatory components and activities remain underdeveloped. In this talk, I will present our integrative computational framework for investigating transcription factor (TF) organization at super-enhancers, candidate sites of transcriptional condensate formation. Across human cell types, we found that TFs with highly clustered genomic binding patterns are preferentially associated with super-enhancers and show stronger liquid-liquid phase-separation properties, supporting a mechanistic link between super-enhancers and transcriptional condensates. In multiple cancer types, densely clustered TF-binding sites are enriched at cancer super-enhancers, exhibit increased chromatin activity, and are strongly associated with clinical outcomes. We further applied this framework to identify TFs associated with BRD4-dependent, pH-sensitive transcriptional condensates in primary mouse macrophages, and used super-resolution microscopy to quantify estrogen-induced chromatin condensate formation in human breast cancer cells. Together, these findings establish a computational strategy for linking TF-binding organization to transcriptional condensate-associated regulatory programs in immunity and cancer.

WORKSHOPS/SESSIONS

Workshop – Spatial Single-Cell and Multi-Omics
August 2nd
9:30 AM – 12:30 PM
Room: 2220A&B

Chair: Yuehua Cui

Title: Atlas-based analysis of high-resolution spatial transcriptomics

Author list: H. Robert Frost[1]

Detailed Affiliations:

[1]Department of Biomedical Data Science, Geisel School of Medicine Dartmouth, Dartmouth College.

Abstract: Advances in spatial transcriptomics (ST) have transformed tissue biology, with high-resolution assays (e.g., 10x Visium HD, Xenium) now profiling thousands of genes across millions of spatial locations. However, resolution growth (e.g., ~5k 55 μm spots in Visium vs. ~10M 2 μm spots in Visium HD) has created a severe gap between data size and the scalability of analysis pipelines. Current strategies address this by reducing both locations (via sketching/aggregation) and genes (via feature selection) before PCA-based dimensionality reduction. Downstream tasks like clustering are then performed on the reduced embedding. While feasible, this approach risks discarding biologically important signals, since genes excluded before dimensionality reduction cannot be recovered. Even with these approximations, nonlinear analyses using deep learning remain impractical for high-resolution ST data.

A key opportunity lies in the rapid growth of single-cell atlases such as the Human Cell Atlas, CELLXGENE and scBaseCamp. These resources already contain >100M human scRNA-seq profiles, with projects like the CZI Billion Cell initiative poised to expand by an order of magnitude. While widely used for label transfer and foundation models (e.g., scGPT, STACK), their potential for improving computational efficiency in ST pipelines remains underexplored. We are developing atlas-based dimensionality reduction approaches that eliminate the need for feature selection while enabling efficient gene-level analyses through reduced rank reconstruction. This talk will highlight three strategies: 1) Atlas-guided linear dimensionality reduction: Using low-dimensional embeddings from atlas scRNA-seq data (e.g., Tabula Muris, Tabula Sapiens) to initialize truncated SVD (e.g., IRLBA) reduces iteration counts by nearly two orders of magnitude. This enables full-gene linear dimensionality reduction of high-resolution ST data without feature selection, and potentially without sketching. 2) Atlas-guided nonlinear dimensionality reduction: A novel SVD–autoencoder fusion model pretrained on atlas data and fine-tuned on ST achieves computationally feasible nonlinear embedding and reconstruction without feature selection. 3) Foundation model–based dimensionality reduction: Transformer-based FMs trained on tens of millions of scRNA-seq profiles encode deep transcriptional knowledge but are typically too large to integrate efficiently. By using FM gene embedding matrices to initialize truncated SVD, we achieve the speed of linear methods while generating embeddings that capture richer biological variation.

Together, these approaches leverage atlases and foundation models to overcome computational bottlenecks in high-resolution ST, enabling scalable analyses without compromising biological resolution.

Keywords: Spatial transcriptomics, Single-cell atlases, Atlas-guided dimensionality reduction

Title: Tools for comparative spatial transcriptomics

Author list: Saurabh Sinha[1]

Detailed Affiliations:

[1]Wallace H. Coulter Department of Biomedical Engineering, H. Milton Stewart School of Industrial & Systems Engineering, Georgia Institute of Technology.

Abstract: As Spatial Transcriptomics (ST) technologies mature, researchers will start collecting data from many samples (as they do with non-spatial transcriptomics today) to study how spatial gene expression varies with phenotypes. This question is important because spatial expression patterns reflect location-dependent gene regulation and intercellular communication that shape tissue function and its changes in diseases. Yet there is a

dearth of tools that can compare ST samples from different groups (e.g., case vs. control) to identify genes whose spatial expression pattern changes between them. Such "differential spatial expression" (DSE) of genes is distinct from the concept of "differential expression", requiring specialized ST-focused tools. While ST analytics is intensely researched, existing methods mostly analyze within-sample variation of gene expression, and multi-sample analysis tools focus on integration or alignment of samples rather than detecting gene-level changes between conditions. Statistically rigorous, unbiased DSE detection marks a major gap in ST analytics today.

We have developed SpaceExpress, a general computational framework for identifying genes whose spatial expression differs significantly between conditions (even if the mean expression does not). Inter-condition comparison of spatial expression is challenging because the underlying spatial "canvas" can vary between samples. SpaceExpress addresses this by learning a "natural" spatial coordinate system that describes spatial organization of the multiple analyzed samples. This provides a general and unbiased solution to the problem, not relying on prior knowledge of tissue organization nor making restrictive assumptions about inter-sample registration. SpaceExpress uses a self-supervised neural network to extract each cell's natural spatial coordinates, reflecting biological organization features such as regions and axes. It then trains a spline model of each gene's expression variation in the learnt coordinate space, and tests if this variation depends on the sample's condition. It thus quantifies differential spatial expression (DSE) of a gene with rigorous error control (FDR). It can also accommodate multiple replicates per group and handle cell types as confounders. Using it on whole brain ST data from two castes of honeybees, we discovered the gene hr38 to be DSE between soldier bees and forager bees, an inter-condition change that is not captured by current approaches but has been independently linked to aggressive response. We also compared visual cortex ST samples from normally reared and dark-reared mice, revealing spatial transcriptional patterns related to sensory experience. Finally, we analyzed two groups of ST samples from the mouse hypothalamic preoptic area, highlighting spatial expression differences associated with aggressive behavior.

Keywords: Spatial transcriptomics, Differential spatial expression, Multi-sample analysis, Self-supervised learning

Title: Categorization of 34 computational methods to detect spatially variable genes from spatially resolved transcriptomics data

Author list: Guanao Yan[1]

Detailed Affiliations:

[1]Department of Computational Mathematics, Science and Engineering, Michigan State University.

Abstract: In the analysis of spatially resolved transcriptomics data, detecting spatially variable genes (SVGs) is crucial. Numerous computational methods exist, but varying SVG definitions and methodologies lead to incomparable results. We review 34 state-of-the-art methods, classifying SVGs into three categories: overall, cell-type-specific, and spatial-domain-marker SVGs. Our review explains the intuitions underlying these methods, summarizes their applications, and categorizes the hypothesis tests they use in the trade-off between generality and specificity for SVG detection. We discuss challenges in SVG detection and propose future directions for improvement. Our review offers insights for method developers and users, advocating for category-specific benchmarking.

Keywords: Spatial transcriptomics, SVG detection methods, Hypothesis testing, Benchmarking

Title: CySCI: Regularized multitasking graphical attention for integrative single-cell analysis

Author list: Xinlei Wang[1], Yuqiu Yang[2], and Xinlei Wang[3]

Detailed Affiliations:

[1]Davidson College, [2]University of Texas Southwestern Medical Center, [3]University of Texas at Arlington.

Abstract: CySCI is a unified computational framework for integrative single-cell analysis that jointly models heterogeneous modalities, including CyTOF and scRNA-seq data, using a regularized multitasking graphical attention network. By leveraging cross-modal biological priors and shared latent structures, CySCI simultaneously performs dimension reduction, cell typing, and differential analysis within a probabilistic framework, enabling robust inference without requiring shared cells. The approach improves integration accuracy, enhances interpretability, and provides a scalable solution for dissecting complex cellular heterogeneity in multi-omic single-cell studies.

Keywords: Graph attention network, Cross-modal learning, Cellular heterogeneity

Title: Identification of high-risk cells in single-cell spatially resolved transcriptomics data using transfer learning and spatial smoothing

Author list: Debolina Chatterjee[1], Justin L. Couetil[2], Ziyu Liu[3], Kun Huang[2], Chao Chen[4], Jie Zhang[2], Michael A. Kalwat[5], and Travis S. Johnson[2]

Detailed Affiliations:
[1]University of Mississippi Medical Center, MS, USA, [2]Indiana University School of Medicine, IN, USA, [3]Purdue University, IN, USA, [4]Stony Brook University, NY, USA, [5]Indiana Biosciences Research Institute, IN, USA.

Abstract: Examination of high-risk cells & spatial regions in tissue samples using spatially resolved transcriptomics (SRT) provides crucial insights into disease mechanisms. Existing methods identify disease-associated cell types or clusters but cannot directly associate disease attributes with individual cells. Our previously published framework, Diagnostic Evidence Gauge of Single-cells and spatial transcriptomics, leverages latent representations of gene expression and domain adaptation to transfer disease attributes from patients to individual cells but was originally developed for single-cell RNA sequencing data. We introduce spatial smoothing techniques—Sliding Window Smoothing (SWiS) and Field of View Smoothing (FoVS)—to extend our framework to diverse single-cell spatially resolved transcriptomics (scSRT) platforms. These include 10x Genomics Xenium and NanoString CosMx across diseases like hepatocellular carcinoma (LIHC) and type II diabetes (T2D). These methods identified high-risk cell types in LIHC and revealed significant upregulation of T2D-associated genes in high-risk cells using DEGAS+FoVS and DEGAS+SWiS, respectively. Collectively, these findings establish DEGAS with spatial smoothing as a powerful, generalizable framework for translational scSRT research with broad diagnostic and therapeutic potential.

Keywords: Disease-risk mapping, Spatial smoothing, Domain adaptation, Translational diagnostics

Title: Decoding pediatric scleroderma across scales using spatial omics and histology

Author list: Wei Chen[1,2]

Detailed Affiliations:
[1]Department of Pediatrics, University of Pittsburgh, [2]UPMC Children's Hospital of Pittsburgh, Pittsburgh, PA, USA

Abstract: Recent advances in spatial transcriptomics (ST) have enabled large-scale, spatially resolved molecular profiling of tissues and created new opportunities for training data-driven models in computational pathology. However, a central challenge remains how to effectively integrate multi-scale visual evidence for accurate cell-level characterization in histology images, particularly in non-cancer inflammatory diseases. In this work, we first leverage our previously developed STARS framework to generate high-quality training data by integrating histology images with spatial transcriptomics across multiple resolutions. STARS reconstructs biologically meaningful single-cell gene expression profiles and enables reliable cell-type labelling at scale. To enhance clinical relevance, we further incorporate expert pathologist annotations using a hybrid supervision strategy that combines molecularly informed labels with human expertise. Building on this foundation, we develop a novel deep learning framework for cell type classification that explicitly models how pathologists interpret histology by dynamically integrating information across multiple fields of view. Rather than relying on a fixed image context, our approach retrieves and aggregates evidence from cell-intrinsic morphology, local microenvironment, and broader tissue architecture in a data-driven manner, with adaptive weighting across scales. We apply this framework to our newly generated spatially resolved skin dataset in scleroderma, a rare autoimmune disease characterized by inflammation and fibrosis. Our method demonstrates improved identification of key cell populations, including fibroblasts, immune infiltrates, and vascular cells, and better captures context-dependent patterns such as inflammatory niches and fibrotic remodeling. Compared to strong baselines, we observe consistent gains in classification performance, particularly for morphologically ambiguous and context-dependent cell types. These results highlight the importance of multi-scale evidence integration for understanding complex skin pathology and provide a data resource for more accurate and interpretable computational tools in scleroderma and related diseases.

Keywords: Multi-scale histology analysis, Cell type classification, Computational pathology, Scleroderma

Title: Decipher non-coding SNPs by high-resolution reconstruction of single-cell spatial genome architectures in 3D space

Author list: Jianrong Wang[1]

Detailed Affiliations:

[1]Department of Computational Mathematics, Science and Engineering, Michigan State University.

Abstract: Spatial chromosomal conformations in 3D space play pivotal roles in modulating interactions of long-range transcription regulation and demonstrate high levels of cell-to-cell heterogeneity. However, current single-cell Hi-C experiments are highly limited to just a few cellular-contexts with extremely high rates of missing contacts (>99.9%), making high-resolution characterization of detailed genome folding challenging. Excessive missing data of single-cell chromatin contacts also makes the identification of long-range cis-regulatory links infeasible for the majority of genes. Here we developed a family of computational models (Tensor-FLAMINGO and its variants), which are able to reconstruct 10kb-resolution spatial configurations of the human and mouse genomes (i.e. the highest resolution to date) across diverse cell types and at single-cell specific levels. Tensor-FLAMINGO consistently demonstrates superior reconstruction accuracy and strong robustness against high rates of missing contacts. Integrative analysis of reconstructed single-cell 3D genome structures with context-specific multi-omics data and genetic association studies (such as eQTLs and GWAS SNPs) revealed a series of novel discoveries, including cell-specific spatial hubs of multi-way interactions, interplay between epigenetic landscapes and 3D folding, long-range eQTLs (>900kb), and significant single-cell specific long-range regulatory interactions. The completion of missing chromatin contacts further enables the characterization of regulatory networks for the majority of genes, which is not possible based on the highly sparse experimental data. Beyond 1D genome annotations and 2D contact map analyses, our predictive and analytical framework of single-cell 3D conformation promotes systems-level insights into the spatial coordination of regulatory programs and its functional impacts, across diverse cellular-contexts and at unprecedented resolution.

Keywords: Single-cell 3D genome, Chromatin conformation reconstruction, Tensor-FLAMINGO, Long-range gene regulation, Multi-omics integration

Title: Multimodal AI for tumor microenvironment characterization and therapeutic prediction

Author list: Qianqian Song[1]

Detailed Affiliations:

[1]Department of Health Outcomes and Biomedical Informatics, University of Florida College of Medicine, Gainesville, FL, USA.

Abstract: Understanding how the tumor microenvironment (TME) shapes therapeutic response remains a central challenge in precision oncology. Recent advances in spatial omics, single-cell technologies, and high-content imaging have enabled unprecedented characterization of tumor ecosystems; however, integrating these diverse and heterogeneous data sources into actionable insights remains a major bottleneck. In this talk, we present AI-driven tools for modeling the TME and predicting drug response through multimodal data integration. Our approach leverages advances in machine learning, including graph-based models and foundation models, to capture complex cellular interactions and spatial organization within the tumor microenvironment. By modeling molecular, spatial, and phenotypic data, these methods enable the identification of functional tumor-stroma niches and uncover regulatory signals that influence treatment response. This integrative perspective provides new opportunities to systematically characterize heterogeneity and to link microenvironmental context with therapeutic outcomes. We further discuss how AI-driven representations can be used to model drug effects across biological systems, supporting applications in drug discovery, response prediction, and therapeutic prioritization. By connecting cellular phenotypes with molecular and spatial features, these approaches facilitate a more comprehensive understanding of drug mechanisms and resistance patterns.

Workshop – AI and Computational Modeling for Biomedical Discovery and Precision Health
August 2nd
9:30 AM – 12:30 PM
Room: 2213A

Chairs: Yijie Wang, Xun Wu

Title: Structured and scalable AI for high-dimensional biological data: beyond lasso under correlation

Author list: Yijie Wang[1]

Detailed Affiliations:
[1]Computer Science Department, IU Bloomington, Bloomington, IN, USA.

Abstract: High-dimensional biological data pose fundamental challenges for artificial intelligence due to strong feature correlations, limited sample sizes, and the need to balance predictive and explanatory power. Traditional sparsity-inducing methods, such as Lasso, often fail under high correlation, leading to unstable feature selection and reduced interpretability. In this work, we present a scalable optimization framework based on Boolean convex programming to address these limitations. Our approach reformulates subset selection using a tight convex relaxation with theoretical guarantees, enabling efficient optimization via projected quasi-Newton methods. We further extend this framework to structured sparsity settings, including group, overlapping, and graph-based regularization. Empirical evaluations on genome-wide association studies (GWAS) demonstrate improved scalability, robustness to correlation, and enhanced feature selection accuracy compared to existing methods. This work provides a unified pathway from classical sparsity models to modern structured AI methods for biological data analysis.

Keywords: Lasso, Correlation, Sparse learning

Title: Explainable AI models for decoding proteins and microbiome

Author list: Yuzhen Ye[1]

Detailed Affiliations:
[1]Computer Science Department, IU Bloomington, Bloomington, IN, USA.

Abstract: Starting with MinPath, which we developed in 2009 to reconstruct biological pathways encoded by microbial communities, our work has focused on creating computational approaches to enable the analysis of multi-omics microbiome data and uncover the structure and function of microbial communities. More recently, we have been developing knowledge-guided AI models designed to be both interpretable and generalizable. In this talk, I will first highlight several tools we have developed for microbiome research. I will then present two examples of our work on explainable AI models. The first focuses on MicroKPNN and MicroKPNN-MT, designed for microbiome-based host phenotype prediction. These models incorporate prior knowledge, including microbial community structure, directly into their neural network architectures. The second example introduces DCTdomain, a domain-based embedding approach that captures the contextual information of protein domains, the fundamental structural and functional units of proteins. We show that these domain-level embeddings enable rapid and accurate detection of protein similarity, and can be effectively leveraged for viral taxonomic assignment, a challenging task due to the high divergence of viruses.

Keywords: Microbiome, Explainable AI, Protein Domain Embeddings

Title: Reinforcement post-training aligns flow-matching perturbation models with DEG recovery

Author list: Xuhong Zhang[1]

Detailed Affiliations:
[1]Computer Science Department, IU Bloomington, Bloomington, IN, USA.

Abstract: Understanding how gene perturbations reshape cellular programs across contexts is central to functional genomics and underpins therapeutic discovery and precision medicine. Flow-matching generative models can represent perturbation as a distributional transformation from control to perturbed single-cell states and often achieve strong overall accuracy under global metrics such as R^2 on mean expression. However, R^2 is dominated by

non-differential genes and can mask failures on the biologically actionable task of recovering differentially expressed genes (DEGs). Recent work argues that AUPRC better reflects DEG identification quality, but AUPRC is non-differentiable and therefore difficult to optimize with standard gradient-based training.

We propose a post-training alignment strategy that directly targets DEG-centric objectives by casting DEG recovery as a reward optimization problem. Specifically, we integrate an online REINFORCE estimator with a leave-one-out (RLOO) baseline into the flow-matching sampling procedure and use AUPRC as the reward, yielding stable, critic-free optimization that improves DEG recovery without modifying the base training objective. Across multi-factor perturbation settings, our reinforcement post-training improves AUPRC for condition-specific DEGs while preserving global fidelity (e.g., R^2) and distributional similarity, making in-silico perturbation prediction more useful for mechanistic interpretation and target prioritization.

Keywords: Single-cell Perturbation Prediction, Flow Matching, Reinforcement Learning

Title: Advancing dietary assessment with computer vision

Author list: Jianpeng He[1]

Detailed Affiliations:

[1]Computer Science Department, IU Bloomington, Bloomington, IN, USA.

Abstract: Accurate dietary assessment is essential for understanding and managing chronic diseases, including obesity and diabetes. However, traditional methods such as 24-hour recall and manual logging are burdensome and prone to bias. Image-based dietary assessment has emerged as a promising alternative, enabled by advances in computer vision and the widespread use of smartphones and wearable devices. By analyzing food images captured in daily life, AI systems can estimate food type, portion size, and nutritional content in a scalable and objective manner. Despite this progress, real-world visual food data remain highly challenging due to variations in lighting, viewpoint, occlusion, mixed dishes, and cross-cultural diversity. This presentation will discuss key challenges and recent advances in computer vision for dietary assessment towards building robust, real-world deployable digital nutrition monitoring systems.

Keywords: Computer Vision, Dietary Assessment, Digital Nutrition

Title: Detecting precise adverse drug events from real-world data

Author list: Penyue Zhang[1]

Detailed Affiliations:

[1]Computer Science Department, IU Bloomington, Bloomington, IN, USA.

Abstract: Adverse drug event (ADE) is a significant public health burden. Many ADEs cannot be identified in premarketing clinical trials due to limitations in patient heterogeneity. As real-world data (RWD) include diverse subpopulations, RWD offers a powerful resource for identifying precise ADEs in subpopulations. First, a trajectory-informed model (TIM) will be presented. TIM can use drug label-supervised learning to improve the performance of ADE detection and can identify adverse drug combinations and their corresponding risk factors. Particularly, TIM can identify drug combinations that remain at a low risk in the general population but possess a higher risk in certain subpopulations. Second, a sensitive and timing-aware model (STEM) will be presented. STEM can identify drug-drug interactions (DDIs) by accounting for the temporal sequence of drug administration. For doing so, STEM has better probability to detect adverse drug combinations in the new initiator subpopulation and can identify DDIs with time-dependent mechanisms of interaction. TIM and STEM have identified new ADEs in RWD analyses and have better performance metrics in simulation studies. In conclusion, extension of RWD-based ADE data mining with respect to risk factors and pattern of drug exposure could identify new and precise ADEs.

Keywords: Adverse Drug Events, Real-World Data, Drug-Drug Interactions

Title: Unbiased in vitro and in vivo CRISPR screens identify novel regulatory genes in macrophage efferocytosis

Author list: Xun Wu[1]

Detailed Affiliations:

[1]Department of Medicine, Division of Cardiology, Columbia University Irving Medical Center, New York, NY, USA.

Abstract: Efficient efferocytosis by macrophages is essential for tissue homeostasis, inflammation resolution, and tissue repair, yet the regulators that control this process in vivo remain poorly defined. To address the challenge of applying CRISPR screening to macrophages in vivo, we developed an integrated strategy that links unbiased target discovery to enhanced macrophage efferocytosis. Using lentiviral delivery of sgRNAs into bone marrow cells from Cas9 transgenic mice, we established a genome-wide CRISPR screening platform in primary macrophages to identify genes regulating efferocytosis in vitro. This system provides a foundation for defining the molecular pathways that control apoptotic cell recognition, engulfment, and processing. To translate these findings in vivo, we developed apoptotic-cell delivery models in mice that enable quantification of macrophage-mediated clearance in the spleen and capture bona fide efferocytosis in vivo. We also incorporated myocardial infarction models, in which a high burden of dying cardiomyocytes provides a clinically relevant setting to study efferocytosis in the injured heart. Using these models, we are testing dying-cell clearance in vivo while defining how macrophages survive and function within the inflammatory cardiac environment. In parallel, we are evaluating macrophage-editing strategies in vitro and in vivo, including pericardial delivery approaches, to test the feasibility of genome editing for macrophage-based repair. Together, these studies define a translational pipeline from unbiased CRISPR screening to in vivo efferocytosis models and macrophage genome editing. Our long-term goal is to identify actionable regulators of efferocytosis and reprogram macrophages to enhance dying-cell clearance in vivo, particularly in the injured heart and in vascular inflammation.

Keywords: CRISPR Screen, Macrophage Efferocytosis, Genome Editing

Title: Aging Synergizes with Hypercholesterolemia to Drive Smooth Muscle Cell Transdifferentiation

Author list: Mingqiao Li[1], Sergiy Sukhanov, Yusuke Higashi, Patrice Delafontaine

Detailed Affiliations:

[1]Department of Medicine, Division of Cardiology, Tulane University School of Medicine, New Orleans, LA, USA.

Abstract: While hyperlipidemia plays a central role in plaque initiation and progression, severe human atherosclerosis typically occurs in aged individuals with concomitant dyslipidemia, suggesting that aging independently contributes to disease advancement. However, the cellular and molecular mechanisms through which aging modulates atherosclerotic plaque composition remain poorly understood. To address this, we performed single-cell RNA sequencing (scRNA-seq) on aortas from young and aged hypercholesterolemic mice. Our analysis revealed that aged hypercholesterolemic mice exhibit distinct cellular heterogeneity and gene expression profiles compared with their young counterparts, with the most pronounced differences observed in smooth muscle cell (SMC) and macrophage populations. Notably, a macrophage-like SMC population and an additional macrophage cluster were present exclusively in aged hypercholesterolemic mice — absent in both aged normocholesterolemic and young hypercholesterolemic groups — indicating a synergistic effect of aging and hypercholesterolemia in promoting macrophage-like SMC transdifferentiation. Furthermore, SMCs in aged normocholesterolemic mice exhibited features of dedifferentiation, including downregulation of contractile markers, yet did not acquire a fibroblast-like phenotype, suggesting that hypercholesterolemia is required for this specific phenotypic switch. Differential gene expression and pathway enrichment analyses demonstrated that SMCs in aged hypercholesterolemic mice upregulate pathways associated with cellular senescence, DNA damage response, and myeloid-related phenotypes, whereas SMCs in young hypercholesterolemic mice are enriched for extracellular matrix deposition and calcification-related pathways. Together, these findings reveal that aging reprograms SMC fate decisions in the context of hypercholesterolemia, shifting the balance from a fibrotic, calcifying phenotype toward a senescent, pro-inflammatory state, with important implications for understanding age-dependent atherosclerosis progression.

Keywords: Aging, Hypercholesterolemia, Atherosclerosis

Title: Model-guided reconstruction of single-cell secretion dynamics during live CAR-T/tumor-cell interactions

Author list: Yujing Song[1], Yifei Fang[2], Ruohan Wang[3], Haoran Zhou[1], Christopher Buglino[1], Rongjin Zhao[1], Ying Ma[3,4], Weiqiang Chen[1,2], Katsuo Kurabayashi[1,5]

Detailed Affiliations:

[1]Department of Mechanical and Aerospace Engineering, New York University, Brooklyn, NY 11201, USA, [2]Department of Biomedical Engineering, New York University, Brooklyn, NY 11201, USA, [3]Department of Biostatistics, Brown University, Providence, RI 02903, USA, [4]Center for Computational Molecular Biology, Brown University, Providence, RI 02903, USA, [5]Department of Chemical and Biomolecular Engineering, New York University, Brooklyn, NY 11201, USA

Abstract: T Live-cell imaging has become an essential tool for studying cellular and multicellular interaction systems, enabling direct observation of cell migration, target engagement, morphology changes and cell killing. However, conventional live-cell imaging remains largely blind to the secreted molecular signals that accompany these dynamic behaviors. This limitation is particularly important for immune-cell interactions, where cell function is shaped not only by physical contact and killing outcome, but also by the timing, location and magnitude of cytotoxic and inflammatory molecule release.

Here we present digital FluoroSpot amplification for spatial-temporal imaging (dFAST), a live-cell-linked molecular imaging and modeling framework for reconstructing single-cell secretion dynamics during CAR-T/tumor-cell interactions. dFAST combines time-lapse live-cell imaging, AI-assisted cell recognition and trajectory tracking, endpoint in situ multiplex molecular footprinting, and model-guided reconstruction of secretion dynamics. Rather than relying on real-time soluble detection reagents during cell culture, dFAST records cell movement and interaction histories during minimally perturbed live imaging, then captures secreted proteins as spatial molecular footprints for high-sensitivity endpoint imaging. A calibrated kernel-convolution model uses the recorded cell trajectories and endpoint molecular patterns to infer single-cell secretion dynamics, q(t), thereby converting static molecular footprints into dynamic secretion histories.

We applied this framework to CD19 CAR-T cells interacting with antigen-positive tumor cells as a model multicellular immune system. By integrating live-cell trajectories with reconstructed molecular secretion profiles, dFAST resolved heterogeneous functional states across individual CAR-T cells, including differences in migration, target engagement, killing outcome, cytotoxic secretion and communication-associated cytokine release. The reconstructed features further enabled quantitative comparison of secretion phenotypes, including cytotoxic-biased, communication-biased and polyfunctional secretion patterns, and supported systems-level analysis of how live interaction behavior is associated with molecular output. In additional perturbation and functional-state comparisons, dFAST captured changes in CAR-T interaction and secretion modules associated with altered activation or dysfunction.

Together, dFAST provides a model-guided approach for linking live-cell image analysis with molecular secretion reconstruction in single cells. This framework offers a generalizable strategy for studying cellular interaction systems in which dynamic behavior and secreted molecular communication must be interpreted together, with potential applications in immunotherapy profiling, systems biology and AI-enabled biomedical image analysis.

Keywords: Single-cell Analysis; Live-cell Imaging; Systems Biology; CAR-T Cells; Secretion Dynamics; Biomedical Image Analysis

Workshop – Genomic Modeling and Biomolecular Structure
August 2nd
9:30 AM – 12:30 PM
Room: 2213B

Chair: Shenglin Mei

Title: Architecting for Discovery: Scaling AI-Ready Multimodal Data Infrastructure for Precision Oncology

Author list: Stephan Schürer[1,2,3]
Detailed Affiliations:
[1]Department of Pharmacology, Miller School of Medicine, University of Miami, [2]Sylvester Comprehensive Cancer Center, Miller School of Medicine, University of Miami, [3]Frost Institute for Data Science & Computing, University of Miami
Abstract:

Rapid advances in artificial intelligence are shifting the principal constraint on biomedical discovery from access to computational intelligence toward access to high-quality, harmonized, and well-governed data. Cancer centers generate increasingly diverse clinical, molecular, imaging, pathology, patient-reported, and experimental data, yet these resources often remain distributed across disconnected systems and are difficult to reuse at scale. This presentation describes an informatics strategy for transforming these fragmented data assets into an AI-ready ecosystem supporting precision oncology and translational research.

At the University of Miami Sylvester Comprehensive Cancer Center, the Sylvester Data Portal provides a secure, cloud-based infrastructure for multimodal data storage, FAIR data management, automated processing, interactive exploration, and controlled collaboration. Automated pipelines integrate clinico-genomic data from multiple molecular profiling partners with tumor registry, pathology, and other clinical information. The resulting real-world database includes more than 35,000 molecularly profiled samples across major cancer types and supports privacy-tiered access for cohort discovery, clinical-trial feasibility assessment, evaluation of treatment patterns, and investigation of molecularly defined patient populations.

The platform also connects research and operational workflows, including automated processing and quality control of sequencing data, standardized metadata capture, and access to raw and processed datasets across a range of computational expertise. Its underlying infrastructure is being extended through AACR Project GENIE, the statewide Florida CARES/PAC3R network, and deployment at Nicklaus Children's Hospital.

Together, these efforts establish a scalable foundation for federated collaboration and the future integration of large language models and agentic workflows for natural-language data access, automated curation, biomarker and vulnerability discovery, and translational therapeutic research.

Keywords: TBD

Title: TFBindFormer: A Cross-Attention Transformer for Transcription Factor–DNA Binding Prediction
Author list: Ping Liu[1,*], Lyuwei Wang,[2] Shreya Basnet[1,2] and Jianlin Cheng[1,2]
Detailed Affiliations:
[1]Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, 65211, MO, U.S.A., [2]NextGen Precision Health, University of Missouri, Columbia, 65211, MO, U.S.A. *Corresponding author
Abstract: Transcription factors (TFs) are central regulators of gene expression, and their selective recognition of genomic DNA underlies various biological processes. Experimental profiling of TF–DNA interactions using chromatin immunoprecipitation followed by sequencing (ChIP-seq) provides high-resolution maps of in vivo TF-DNA binding but remains costly, labor-intensive, and inherently low-throughput, limiting their scalability across different transcription factors, cell types, and regulatory conditions. Computational modeling therefore plays an essential role in inferring TF–DNA interactions at genome scale. However, most existing computational models rely solely on DNA sequence and chromatin features to predict TF–DNA binding, neglecting TF-specific protein information. This omission limits their ability to capture protein-dependent binding specificity. Here, we present TFBindFormer, a hybrid cross-attention transformer that explicitly integrates genomic DNA features with TFspecific representations derived from protein sequences and structures. By modeling protein-conditioned, positionspecific TF–DNA interactions, TFBindFormer enables direct learning of molecular determinants underlying DNA recognition. Evaluated across hundreds of cell-types–specific TFs and hundreds of millions of genome-wide DNA bins, TFBindFormer consistently outperforms DNA-only baselines, achieving substantial gains in both area under precision-recall curve (AUPRC) and area under receiver operating characteristic curve (AUROC). Together, these results demonstrate that integrating TF and DNA features via cross-attention enables TFBindFormer to serve as an effective and scalable framework for large-scale TF–DNA binding prediction.

Keywords: transcription factor, gene regulation, transcription factor-DNA binding, transformer, cross-attention
Abbreviations TF: Transcription Factor, AUROC: area under receiver operating characteristic curve, AUPRC: area under precision-recall curve, MHA: multi-head attention

Title: Comprehensive Benchmarking of Long-read Fusion-detection Methods Leads to Improved Fusion Discovery

Author list: Ke Ma[1,2†], Weixu Wang [1,2†], Kang Li [1,2†], Fei Zhang [1,2†], Shulan Qiu [1,2*], Weixiong Zhang[1,2,3*]
Detailed Affiliations:
[1]Hong Kong Jockey Club STEM Laboratory for Genomics and AI in Healthcare, [2]Department of Health Technology and Informatics, [3]Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University, Hong Kong, China. † Equal contribution. *Corresponding author
Abstract: Motivation: Long-read RNA sequencing enables direct observation of full-length fusion transcripts and their isoforms, addressing key limitations of short-read–based fusion detection. Several computational tools have been developed so far for long-read fusion discovery. Due to the diversity of fusion transcripts, these methods introduced heterogeneous algorithmic assumptions, filtering strategies, and output definitions, making it challenging to determine the best approach for diverse fusion-detection applications. The difficulty is exacerbated by two competing long-read sequencing platforms, Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio). Therefore, a systematic, quantitative comparison across sequencing platforms, data types, and experimental conditions is urgently required. Results: We benchmarked eight long-read gene fusion detection tools using a unified evaluation framework over simulated datasets, publicly available longread transcriptomes, and an in-house direct RNA sequencing dataset. Simulations revealed that fusion detection performance is strongly influenced by sequencing coverage, read length, and base-level accuracy, with no single method consistently outperforming the others across all conditions examined. On real datasets, the performance of the methods varied substantially across platforms (ONT versus PacBio), library construction strategies (PCR-amplified cDNA, direct cDNA, and direct RNA), and analytical tasks, including fusion discovery and isoform characterization. Our consensus analysis highlighted distinct trade-offs between sensitivity and precision, demonstrating that method-specific strengths are largely driven by their underlying modeling principles and data characteristics. To address methodological discrepancies, we developed LongFuse, an evidence-guided post hoc integration approach that improved the balance between sensitivity and stringency across complementary callers. Together, this study provides a rigorous benchmark for selecting and evaluating fusion-detection methods in long-read transcriptomics. Availability and Implementation: All benchmarking workflows, simulation scripts, containerized tool implementations, and standardized evaluation code are publicly available at https://github.com/GenomicMedicine/LongReadsFusionBenchmarking.
Keywords: Gene Fusion, Long-read Sequencing, Benchmarking, Bioinformatics, Transcriptomics

Title: KoGNER: A Novel Framework for Knowledge Graph Distillation on Biomedical Named Entity Recognition

Author list: Heming Zhang[1*], Wenyu Li[2,3*], Di Huang[1,2*], Hao Li[1], Yinjie Tang[3], Yixin Chen[2], Philip Payne[1], Fuhai Li[1,†]
Detailed Affiliations:
[1]Institute for Informatics, Data Science and Biostatistics, Washington University in St. Louis, [2]Department of Computer Science and Engineering, Washington University in St. Louis, [3]Energy, Environmental and Chemical Engineering, Washington University in St. Louis. †Correspondence, *Equal contribution as co-first authors
Abstract: Named Entity Recognition (NER) is a fundamental task in Natural Language Processing (NLP) that plays a crucial role in information extraction, question answering, and knowledge-based systems. Traditional deep learning-based NER models often struggle with domain-specific generalization and suffer from data sparsity issues. In this work, we introduce Knowledge Graph distilled for Named Entity Recognition (KoGNER), a novel approach that integrates Knowledge Graph (KG) distillation into NER models to enhance entity recognition performance. Our framework leverages structured knowledge representations from KGs to enrich contextual embeddings, thereby improving entity classification and reducing ambiguity in entity detection. KoGNER employs a two-step process: (1) Knowledge Distillation, where external knowledge sources are distilled into a lightweight representation for seamless integration with NER models, and (2) Entity- Aware Augmentation, which integrates contextual

embeddings that have been enriched with knowledge graph information directly into GNN, thereby improving the model's ability to understand and represent entity relationships. Experimental results on benchmark datasets demonstrate that KoGNER achieves state-ofthe-art performance, outperforming finetuned NER models and LLMs by a significant margin. These findings suggest that leveraging knowledge graphs as auxiliary information can significantly improve NER accuracy, making KoGNER a promising direction for future research in knowledge-aware NLP.

Title: Multiple versus pairwise sequence alignments for protein phylogenetics using foundation models

Author list: Rohan Alibutud[1,2] and Sudhir Kumar[1,2*]

Detailed Affiliations:
[1]Institute for Genomics and Evolutionary Medicine, Temple University, PA 19122, USA, [2]Department of Biology, Temple University, Philadelphia, PA 19122, USA. [*]Corresponding author

Abstract: Phylogenetic inference is a common task in molecular and evolutionary biology and has conventionally required a multiple sequence alignment (MSA), a statistical model of amino acid substitutions, and an optimality principle. Recently, global models of amino acid substitutions have been inferred from millions of MSAs using transformer-based deep learning, resulting in protein foundation models (pFMs), also known as protein language models (PLMs). Training pFMs on MSAs hypothetically enables them to encode residue dependencies and the phylogenetic structure of the MSA collection. In contrast, pFMs trained on individual sequences lack access to such phylogenetic structure. Here, we assess the phylogeny inference gains offered by the use of MSA for training pFMs by comparing the relative accuracies of phylogenies inferred using two types of pFMs: one trained on a large collection of MSAs (msat-pFM, [1]) and the other trained using a collection of single sequences (esm-pFM). For msat-pFM analysis, we inferred neighbor-joining trees using pairwise distances estimated directly from the sequence attention matrices. For esm-pFM [2], pairwise distances were obtained using the correlation of attentions of homologous residues, where pairwise sequence alignments (PSA) were used to establish residue homologies. Surprisingly, MSA phylogenies inferred using the msat-pFM were less accurate than esm-pFMs. This pattern was seen across datasets spanning both small and large numbers of species and proteins. Also, PSA phylogenies obtained using residue attentions from early ESM-PFM layers were much more accurate. These results suggest that the multiple sequence alignment step, which is obligatory to establish residue homologies across multiple sequences, may not add information when using evolutionary distances based on attentions in pFMs.

Keywords: phylogenetics, deep learning, foundation models, multiple sequence alignment

Title: ChromDiffusion: Refining Chromatin Structure Characterization Through Diffusion Models

Author list: Yang Zhao[1], Tomoya Kanno[1], Chong Li[1] and Xinghua Shi[1,*]

Detailed Affiliations:
[1]Department of Computer and Information Sciences, Temple University, Philadelphia, PA, USA. [*]Corresponding author

Abstract: High-resolution Hi-C experiments require significant sequencing depth, limiting the acquisition of dense chromatin contact maps. Consequently, low-coverage Hi-C data results in sparse and noisy matrices, with focal chromatin loops being especially difficult to recover compared with broader domain-scale structures such as topologically associated domains (TADs). Existing deep learning methods that directly reconstruct high-resolution maps can recover global contact patterns but may oversmooth loop-scale interactions or introduce structural artifacts. To address these limitations, we propose ChromDiffusion, an annotation-free Hi-C refinement framework based on residual diffusion. Instead of directly reconstructing the contact matrix or retraining a super-resolution backbone, ChromDiffusion attaches to a frozen pretrained predictor and progressively denoises the residual interaction patterns between the backbone output and the high-resolution target. This residual diffusion approach enables the model to repair fine-scale loop-band signals while largely preserving domain-scale chromatin structure. We evaluated ChromDiffusion on the GM12878 Hi-C benchmark dataset across four representative Hi-C super-resolution backbones. Experimental results show that ChromDiffusion increases the recovery of true-positive chromatin loops under low-coverage conditions while maintaining high TAD detection accuracy, although the precision cost depends on the backbone and refinement configuration. These results suggest that ChromDiffusion can serve as a model-independent, caller-annotation-free refinement module for improving loop-scale signal recovery in Hi-C super-resolution analysis.

Keywords: Chromatin Structure, 3D Genome, Diffusion Models, Generative Modeling, Hi-C data, Super-resolution, Computational Genomics

Title: A Structural Antibody Benchmark of AlphaFold3 reveals Hallucinated Epitopes and a Bias for Orderness

Author list: Arnav Solanki,[1] Neha Shree Maurya,[1] Meaghan Ramlakhan,[1] Rongbin Li[1], Wenbo Chen[1], Zhuhao Wu[2] and Wenjin Jim Zheng [1,*]

Detailed Affiliations:
[1]McWilliams School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, Texas, United States, [2]Appel Alzheimer's Disease Research Institute, Feil Family Brain and Mind Research Institute, Weill Cornell Medicine, New York City, New York, United States. *Corresponding author

Abstract: AlphaFold3 has shown promise as a tool for predicting antibody-antigen binding, yet its performance across large datasets has not been fully characterized. In this study, 3401 experimentally validated antibody-antigen complexes were sourced from the Structural Antibody Database and screened alongside 23798 negative controls to benchmark AlphaFold3's binding prediction capabilities. Confidence metrics including Predicted Aligned Error and Interface Predicted Template Modeling score were used to achieving a maximum recall of 53% at 100 inference seeds. Several factors were found to influence prediction accuracy: a notable bias was observed toward antibodies derived from X-ray crystallography structures versus those from electron microscopy, and positive prediction rates were found to decrease with increasing target protein size and surface area. In contrast, neither the amino acid composition or lengths of the complementarity determining regions, nor training data leakage were found to introduce significant bias. An innate false positive rate of approximately 3% was identified, with AF3 shown to hallucinate plausible binding interfaces across the surface of decoy targets while avoiding disordered regions. Epitope mapping using DockQ, epitope shift, and antibody displacement revealed that approximately 34% of false negatives retained the correct epitope location despite poor structural alignment, suggesting that conformation refinement tools could recover additional true binding predictions. These findings provide a comprehensive characterization of AlphaFold3's strengths and limitations for antibody screening in computational drug discovery.

Keywords: AlphaFold3, Antibody, AI, Disorder, Epitope

Workshop – Computational and AI models to decode disease mechanisms based on chromatin and multi-omics integration

August 2nd
2:00 – 5:20 PM
Room: 2220A&B

Chair: Jianrong Wang

Title: Combining natural genetic variations with single molecule footprinting to learn about the rules of chromatin regulation

Author list: Kaixuan Luo[2]*, Ayelen Lizarraga[1]*, Xiaotong Sun[2], Diana Vera Cruz[1], Xin He[2], Sebastian Pott[1]

Detailed Affiliations:
[1]Department of Medicine, University of Chicago, Chicago, IL, US, [2]Department of Human Genetics, University of Chicago, Chicago, IL, USA. *Contributed equally

Abstract: Gene regulation is a highly dynamic process mediated through interactions of transcription factors (TFs), nucleosomes, and DNA methylation (DNAm). The key challenge of the field is to unravel the rules of these molecular interactions. Single-molecule footprinting (SMF) emerges as a promising technology to address this challenge. SMF combines DNA methyltransferase treatment and direct long-read sequencing to simultaneously capture chromatin accessibility, DNAm, and TF binding across intact 10-20 kb molecules. Given the complexity of gene regulation, however, extract causal and mechanistic rules from SMF data is difficult. We reason that natural genetic variations in DNA sequences provide large-scale perturbations of gene regulatory processes, creating an

excellent opportunity to learn about chromatin regulation. For example, a SNP disrupting the motif of a TF would provide information of the effect of this TF on the nearby chromatin state.

We generated SMF data in 30 human lymphoblastoid cell lines. Using these data, we detected 6,000 cis-regulatory elements that show differential accessibility between different alleles of genetic variants, denoted as allele-specific fiber-seq inferred regulatory elements (AS-FIRE). Many of the variants underlying AS-FIRE disrupt motifs of TFs controlling chromatin accessibility such as CTCF. These variants often have broad effects on nucleosome positioning and phasing. We also detected 2,000 TF footprints that show allelic difference, denoted as AS-footprints. Notably, half of the AS-footprints are missed by AS-FIREs, and these genetic variants often exhibit more local effects on nearby chromatin states. Motif analysis in AS-FIRE and AS-footprints revealed overlapping but different sets of motifs, e.g. PU.1-IRF composition motif only in AS-FIRE and KLF motif only in AS-footprints. The shared motifs include CTCF, IRF, and BATF, suggesting that the effects of these motifs may depend on sequence context. Altogether, these results highlight that different TFs have variable chromatin effects, and the effects may be sequence-context dependent.

Our results demonstrate that combining natural genetic variations with SMF is a powerful strategy of learning about gene regulation.

Keywords: Single-molecule footprinting, Allele-specific chromatin regulation, Transcription factor binding, Genetic variation, Nucleosome positioning

Title: Robust causal inference from imperfect CRISPR interventions in Perturb-seq: Heterogeneous Perturbation Strength and Off-target Bias

Author list: Zhongxuan Sun[1], Hyunseung Kang[2], Sündüz Keleş[1,2]

Detailed Affiliations:

[1]Department of Biostatistics & Medical Informatics, [2]Department of Statistics, University of Wisconsin-Madison.

Abstract: Perturb-seq maps gene function and regulatory relationships by coupling pooled CRISPR perturbations with single-cell transcriptomic readouts. Yet causal interpretation requires more than linking guide RNA (gRNA) assignment to downstream expression. Two factors remain insufficiently modeled. CRISPRi/a often produces partial interventions whose target effects vary across genes and guides, and gRNAs can induce expression changes through unintended binding rather than the assigned target. These violations of clean-intervention assumptions can bias causal estimates and make technical artifacts appear as biological regulation.

We present two complementary approaches for extracting reliable causal information from Perturb-seq by modeling imperfect CRISPR interventions. ADAPRE addresses heterogeneity in CRISPRi perturbation strength across perturbed genes. Because knockdown efficiency varies with guide design and target context, strongly perturbed genes can be mistaken for strong regulators and appear as spurious hubs in inferred gene regulatory networks. ADAPRE mitigates this bias using an adaptive instrumental-variable framework. Perturbation assignments are treated as instruments for latent target-gene expression, total effects are estimated from UMI counts through a Poisson-lognormal model, and sparse direct effects are recovered by adaptive penalized inverse regression. The penalty is indexed by empirical knockdown strength, reducing signal driven by differences in target-gene repression across perturbations. In simulations and genome-scale K562 and endothelial Perturb-seq datasets, ADAPRE improves graph recovery, reduces strength-dependent degree bias, maintains stability and cross-dataset reproducibility, and yields stronger enrichment against orthogonal reference sets from transcription factor binding, protein interactions, and protein complexes.

Our second approach addresses off-target bias in perturbation effect estimation. We reinterpret multiple gRNAs assigned to the same target as candidate instruments for the target-gene perturbation. Existing estimators use naïve procedures such as pooling cells across guides, averaging guide-level effects, or selecting the strongest guide-level signal, implicitly requiring all retained gRNAs to satisfy the exclusion restriction. Off-target binding violates this assumption in an outcome-specific manner, since a guide may be invalid for one response gene but valid for others. We instead use robust estimators that remain valid under the weaker majority-valid-guide assumption, aggregating guide-level effects while limiting the influence of invalid guides. Across CRISPRa and CRISPRi datasets, this approach reduces sequence-predicted off-target bias and target-level clustering driven by off-target signal without using off-target annotations during estimation, while preserving validated regulatory programs. In a semi-synthetic contamination benchmark, it avoids spurious regulatory signals when valid focal guides constitute a majority.

Together, these approaches move Perturb-seq analysis beyond idealized intervention assumptions toward causal estimators that explicitly account for technical distortions.

Keywords: Perturb-seq; causal modeling; instrumental variables; robust estimation; off-target bias; gene regulatory networks

Title: Learning single-cell multiomic trajectories using neural ODEs

Author list: Christina Leslie[1]

Detailed Affiliations:

[1]Memorial Sloan Kettering Cancer Center.

Abstract: We will present a new ML/AI approach for learning cell dynamics from single-cell multiome (scRNA+ATAC-seq) data using a generative neural ordinary differential equation (ODE) model called DynaVelo. Here, the neural ODE learns a functional form of the dynamics — corresponding to learned RNA velocity and transcription factor (TF) motif velocity vector fields — from single-cell multiome data. We apply DynaVelo to learn the complex in vivo dynamics of wildtype and mutant germinal center B cells. We show that Jacobian analysis or in silico perturbations can recover dynamic regulatory networks governing cell state transitions in the germinal center. We also show how to predict the impact of genetic loss-of-function mutations on cell dynamics, and how to predict TF perturbations that rescue loss-of-function phenotypes.

Keywords: DynaVelo, Neural ordinary differential equations, Single-cell multiome, Cell-state dynamics, Transcription factor perturbation

Title: From ontology fingerprints to BRAINCELL-AID: literature-driven gene function and brain cell annotation

Author list: Wenjin Zheng[1]

Detailed Affiliations:

[1]McWilliams School of Biomedical Informatics, University of Texas Health Science Center at Houston.

Abstract: Biomedical literature contains a vast and rapidly expanding body of knowledge, making it both a valuable resource and a challenge to systematically extract information relevant to gene function and cellular identity. Our early works introduced Ontology Fingerprints, a framework that represents genes as interpretable embeddings derived from Gene Ontology terms enriched in literature. This approach captures gene functions through statistical associations while preserving interpretability, enabling applications such as gene prioritization and pathway discovery through a self-attention like mechanism. Building on this foundation, we extend literature-driven annotation to the cellular level using modern deep learning and large language models. We developed BRAINCELL-AID, an agentic AI framework that integrates fine-tuned LLMs, literature mining, and retrieval-augmented generation to annotate brain cell marker gene sets and cell types at scale. BRAINCELL-AID provides literature-grounded annotations for more than 20,000 gene sets and over 5,300 brain cell types, demonstrating strong cross-species performance and enabling biologically meaningful hypothesis generation. Together, this work establishes a continuum from knowledge-based embeddings to agentic AI systems, highlighting the role of literature-integrated representations in advancing gene function characterization, pathway discovery, and large-scale brain cell annotation.

Keywords: Ontology Fingerprints, Agentic AI, Biomedical literature mining, Brain cell annotation, Retrieval-augmented generation

Title: Interpreting deep learning models in genomics via concept attribution

Author list: Jianyu Yang[2], Shaun Mahony[1,2]

Detailed Affiliations:

[1]Center for Eukaryotic Gene Regulation, [2]Department of Biochemistry and Molecular Biology, The Pennsylvania State University

Abstract: Interpreting genomics deep learning models remains challenging. Existing feature attribution methods are largely restricted to one-hot DNA inputs and therefore cannot assess the influence of more general genomic features such as chromatin states or genomic repeats. Concept attribution methods offer an input-agnostic global

interpretation framework, yet they have not been systematically applied to interpret neural network applications in genomics.

I will present the first application of concept attribution to interpret genomics deep learning models by adapting the Testing with Concept Activation Vectors (TCAV) method. I will show that incorporating a PCA-based decorrelation transformation to address correlated and redundant embedding features improves over the original TCAV, resulting in the Testing with PCA-projected Concept Activation Vectors (TPCAV) approach.

I will show that TPCAV provides comparable motif interpretation to TF-MoDISco on one-hot encoded DNA-based transcription factor binding prediction models. TPCAV also enables robust interpretive analysis of more general biological concepts such as repetitive elements, genome annotations, and chromatin states. It also generalizes to interpreting important features in tokenized foundation models as well as models incorporating chromatin signals as inputs. I will further show that TPCAV can identify representative regions associated with specific concepts, motivating downstream investigation of distinct regulatory mechanisms. TPCAV can be applied to a wide range of model architectures and genomics applications and provides a flexible and robust complement to existing model interpretation techniques.

Keywords: Genomics deep learning, Model interpretability, Chromatin states

Title: Spatially-informed reference-free deconvolution for spatial transcriptomics with SpatialCD

Author list: Phuong Vo[1] and Yuehua Cui[1]

Detailed Affiliations:

[1]Department of Statistics & Probability Michigan State University.

Abstract: Cell-type deconvolution has been instrumental for the analysis of spatial transcriptomics (ST) data to reveal underlying tissue heterogeneity. While reference-based methods have been widely explored, practical limitations, particularly the need for matched scRNA-seq datasets, highlight the value of robust reference-free methods. Existing reference-free approaches, such as STdeconvolve, overlook spatial information, despite the well-established observation that spatially adjacent spots often share similar cellular compositions. Motivated by this, we propose SpatialCD, a Spatially-informed reference-free Cell-type Deconvolution method that extends Latent Dirichlet Allocation (LDA) model with spatial regularization to encourage neighboring spots to share similar cell-type compositions. SpatialCD produces improved estimates of cell-type proportions and gene expression profiles. Across simulated and real datasets, including MERFISH-derived simulations, mouse olfactory bulb (MOB), 10x Visium, and DBiT-seq data, SpatialCD consistently outperforms existing reference-free methods (e.g., STdeconvolve and SpiceMix), recovering more accurate transcriptional patterns and revealing biologically coherent spatial organization. This work advances statistical tools for spatial transcriptomics and enriches the methodological toolkit for complex spatial gene expression analysis.

Keywords: TBD

Title: spatiAlytica: Spatially intelligent multimodal agentic system for decoding tumor ecosystems

Author list: Arun Das, Kexun Zhang, Jifeng Song, Meiru Han, Angela Chen, Wen Meng, Hugh Galloway, Po-Yuan Chen, Sumin Jo, Zhentao Liu, Md Musaddaqul Hasib, Adam Officer, Harsh Sinha, Yu-Chiao Chiu, Shou-Jiang Gao, Lei Li, Yufei Huang

Detailed Affiliations:

University of Pittsburgh.

Abstract: Spatial single-cell omics is transforming cancer research by revealing how tumor, immune, and stromal cells organize within tissue, but analysis remains challenging because it requires programming expertise, iterative visualization, and spatial reasoning across complex data layers. This talk presents spatiAlytica, a visually interactive, viewer-centric multimodal agentic system that enables biologists to explore spatial omics data through natural language while directly interacting with tissue visualizations. spatiAlytica integrates viewer-state awareness, agentic memory, biological concept-to-data mapping, code generation and debugging, Spatial VQA, and grounded interpretation to support iterative, hypothesis-driven analysis. I will highlight its benchmarked gains over existing agentic systems, then focus on a Kaposi's sarcoma case study showing how spatiAlytica recapitulates known tumor niche remodeling and reveals progressive CD8+ T-cell dysfunction during disease progression.

Keywords: Spatial single-cell omics, Agentic AI, Spatial visual question answering, Tumor microenvironment

Title: Single-molecule RNA fate modeling by deep learning

Author list: Dean Li[1,2], Jiayi Chen[1,2], Yunxia Wang[1,2], Rongbin Zheng[1,2], Kaifu Chen[1,2,3]

Detailed Affiliations:

[1]Basic and Translational Research Division, Department of Cardiology, Boston Children's Hospital, Boston, MA 02115, USA, [2]Department of Pediatrics, Harvard Medical School, Boston, MA 02115, USA, [3]Lead contact: Kaifu.chen@childrens.harvard.edu

Abstract: RNA, as the nexus of the central dogma, shapes nearly every facet of biology. However, quantitative assessment of RNA fate, including its stability, splicing, and translation, remains challenging, particularly at single-molecule resolution. Here, we present Rondo, a transformer-based deep learning framework that leverages Nanopore RNA-sequencing to model single-molecule RNA fate. Compared with existing leading models, Rondo achieves superior performance in predicting diverse RNA functions, including RNA stability, translation efficiency, and expression levels. Moreover, Rondo learns biologically meaningful RNA features, revealing that regulatory elements such as miRNA target sites contribute to RNA fate prediction. Using an in silico perturbation strategy, we further dissect the causal effects of RNA features on RNA fate. Finally, Rondo accurately predicts the functional consequences of cancer-associated variants. Collectively, Rondo enables quantitative single-molecule modeling of RNA fate, facilitating the investigation of RNA regulatory mechanisms and disease-associated variants while paving the way for RNA-based therapeutic design.

Keywords: Deep learning; Language model; RNA fate; RNA sequence elements

Workshop – AI and Multi-Modal Integration for Disease Characterization

August 2nd
2:00 – 5:20 PM
Room: 2213A

Chair: Shibiao Wan

Title: Agentic-AI for hypothesis generation for aging and drug repurposing

Author list: Byounggook Cho, Junghyun Jung, Kyoung Jae Won

Detailed Affiliations:

Department of Computational Medicine, Cedars-Sinai Medical Center, West Hollywood, CA 90024, USA

Abstract: Elucidating the mechanisms of aging is hindered by its stochastic, multi-scale nature and pervasive cellular heterogeneity. These challenges are further compounded by the vast and rapidly expanding volume of biomedical literature, as well as the complexity of genome-wide datasets. To address these limitations, we present PersonaAI, an interactive agentic AI framework designed to function as a digital co-scientist. By integrating literature-based reasoning with autonomous in silico validation, PersonaAI leverages retrieval-augmented generation (RAG) to synthesize insights from over 560,000 aging-related publications and generate mechanistic hypotheses. These hypotheses are subsequently evaluated by autonomous agents using single-cell RNA sequencing (scRNA-seq) data.

Using a temporal cutoff strategy restricted to pre-2020 literature, we demonstrate that PersonaAI can generate hypotheses that are independently validated by post-2021 discoveries, highlighting its capacity for inference beyond simple information retrieval. In practical applications, the system identified senescent Cirbp^{+} hepatocytes as a liver-intrinsic aging program and uncovered a middle-aged, male-specific decline in adipose stem and progenitor cells, driven by vascular niche deterioration and disrupted VEGF–VEGFR signaling.

These findings establish PersonaAI as a scalable platform that augments human intuition with autonomous, data-driven validation, offering a generalizable approach to accelerating discovery in aging biology. Furthermore, we extended this framework to investigate Alzheimer's disease (AD) by integrating the scRNA-seq agent with electronic health record (EHR) data. This multi-modal integration enables PersonaAI to identify cell-type-specific signatures in the AD brain, map these signatures to candidate therapeutic compounds, and validate their clinical relevance using large-scale EHR cohorts.

Keywords: PersonaAI, Aging, Drug repurposing, Agentic AI

Title: An optimal-transport based multi-modal integration for categorizing T-cell acute lymphoblastic leukemia

Author list: Lusheng Li, Jieqiong Wang, Shibiao Wan

Detailed Affiliations:
Department of Genetics, Cell Biology and Anatomy, University of Nebraska Medical Center, Omaha, NE, USA

Abstract: As an aggressive pediatric malignancy, T-cell acute lymphoblastic leukemia (T-ALL) hasn't been fully characterized due to its phenotypic complexity and molecular heterogeneity. Identifying T-ALL subtypes is essential for downstream risk stratification and tailored treatment design. Conventional wet-lab approaches for T-ALL characterization, such as immunophenotyping, cytogenetic analysis, fluorescence in situ hybridization (FISH), and targeted molecular assays, are often labor-intensive, time-consuming, and costly. Recently, there are numerous artificial intelligence (AI) and machine learning (ML) approaches that have been proposed to characterize T-ALL. However, they are usually based on single-omics (e.g., transcriptomics, or genomics) data, which might not capture cross-omics interaction to unravel the mechanism of pediatric cancer. To address these concerns, we develop a novel multi-modal learning framework based on optimal transport (OT) that integrates both transcriptomics and genomics data for accurate and cost-effective T-ALL categorization. Specifically, we first processed single nucleotide variations (SNVs) and gene expression profiles, respectively, through attention modules to capture the most informative features. Then, we leveraged an optimal transport (OT) method to align and integrate heterogeneous omics modalities, capturing complementary biological information across data types. Based on this, the OT–derived cost matrix learned the cross-modal interdependencies by quantifying the cost of aligning one modality to the other. In this way, our framework generated a shared latent representation that integrated both omics data, enabling a more comprehensive and coherent representation of each patient sample. Experimental results demonstrated that our proposed approach achieved a substantially higher performance than state-of-the-art ensemble learning models and attention-based models. We believe our multi-modal integration framework will provide a new way to significantly improve downstream T-ALL risk assessment and personalized treatment design.

Keywords: Optimal transport, T-ALL, Cancer subtyping, Multi-omics integration, Transcriptomics, Genomics

Title: ShinyFeatures: a Shiny App for prioritizing features for machine learning models

Author list: Timothy Shaw, Trang Ta, Alyssa Obermayer, Ken Dao, Yi Luo

Detailed Affiliations:
H Lee Moffitt Cancer Center.

Abstract: Feature selection is a critical step in developing interpretable and accurate cancer outcome predictive models from high-dimensional demographic, physical, and biological data. We developed ShinyFeatures, an interactive, user-guided framework, for feature prioritization and predictive modeling from tabular datasets structured as a feature-by-sample matrix. The platform integrates multiple complementary feature-selection strategies, including Pearson correlation, Random Forest–based importance, Markov blanket methods, and least absolute shrinkage and selection operator (LASSO) regularization, with missing values handled via multivariate imputation by chained equations. ShinyFeatures supports training–testing data partitioning, identification of overlapping top-ranked features across methods, and evaluation of multiple machine-learning models, including logistic regression, random forests, and extreme gradient boosting (XGBoost), using metrics such as area under the receiver operating characteristic (ROC) curve, Matthews correlation coefficient, precision, recall, and F1 score. We demonstrate the utility of ShinyFeatures using gene-expression data from the Cancer Cell Line Encyclopedia, where integrative feature selection consistently prioritized biologically meaningful predictors, including expression of Y-chromosome loci associated with biological sex. ShinyFeatures provides a transparent and extensible framework for reproducible feature selection and model evaluation. The shiny app can be accessed https://trangtausf.shinyapps.io/shiny_features/ and source code https://github.com/ttrang87/feature_selection.

Keywords: Feature selection, Shiny web interface, Genomics

Title: TENET+: A tool for reconstructing gene networks by integrating single cell expression and chromatin accessibility data

Author list: Junil Kim[1]

Detailed Affiliations:

[1]School of Systems Biomedical Science, Soongsil University, 369 Sangdo-Ro, Dongjak-Gu, Seoul 06978, Republic of Korea.

Abstract: Gene expression is regulated by multiple epigenetic factors including chromatin structure. Single cell chromatin accessibility data combined with transcription factor (TF) binding motif information enhances our understanding of gene expression regulation. However, enhancer-driven gene regulation remains limited due to the lack of promoter-enhancer interaction data. Here, we developed TENET+, a tool for reconstructing enhancer-driven gene regulatory networks by integrating single cell transcriptome and chromatin accessibility data. By leverating causal relationship inference based on pseudo-time-ordered gene expression and chromatin accessibility, TENET+ predicts interaction between TF expression, target gene expression, and open chromatin regions including long-range enhancer. Applying TENET+ to a paired scRNAseq and scATACseq dataset of human peripheral blood mononuclear cells, we identified critical regulators and their epigenetic regulations. Interestingly, TENET+ exhibited a superior performance to predict promoter-enhancer interactions provided by cell type-specific Hi-C datasets compared with other motif-based GRN reconstruction tools.

Keywords: Single cell multi-omics, Gene regulatory networks, Transfer entropy, Epigenomics causality

Title: Advancing prenatal congenital heart disease screening through artificial intelligence

Author list: Mohamed Azzam, Ziyang Xu, Ruobing Liu, Esther C. Ugwueke, Shibiao Wan, Jason Christensen, Ling Li, Jieqiong Wang

Detailed Affiliations:

Department of Neurological Sciences, University of Nebraska Medical Center, Omaha, NE

Abstract: Prenatal detection of congenital heart disease (CHD) remains a significant clinical challenge. Screening is often time-consuming, highly dependent on clinician expertise, and prone to missed diagnoses due to the complexity of fetal cardiac anatomy and variability in ultrasound image quality. In routine practice, sonographers must interpret multiple standard cardiac views, each providing partial and complementary information. This process can be especially difficult in settings with limited resources or less experienced operators. Artificial intelligence (AI) offers a promising opportunity to improve the efficiency and reliability of prenatal CHD screening. In this presentation, we introduce a novel AI framework that integrates multi-view fetal ultrasound data, enabling the model to capture complementary anatomical information across different cardiac views. By jointly learning from multiple perspectives, the framework aims to better reflect how clinicians synthesize information during real-world diagnosis. However, developing such an AI system is not straightforward. Several key challenges must be addressed. First, annotated medical imaging data are often limited. Second, fetal ultrasound data are heterogeneous. Third, each case may contain a variable number of views or frames, resulting in inconsistent input length for model training. Finally, the “black-box” nature of many AI models raises concerns about interpretability and clinical trust. In this talk, we will present our strategies to address these challenges. We will discuss approaches for improving the annotation number of ultrasound data. We will also describe a unified framework that can handle heterogeneous data and variable-length inputs in a flexible manner. Importantly, we will demonstrate how our model provides interpretable outputs, such as highlighting clinically relevant regions and summarizing decision logic, to support clinician understanding and confidence. Beyond detection, we will explore the potential of our framework for CHD subtyping, which may further assist clinical decision-making and patient management. Overall, our approach aims to reduce clinician burden, improve diagnostic consistency, and enable earlier and more reliable detection of CHD, ultimately contributing to better maternal and fetal outcomes.

Keywords: Congenital heart disease, Multi-instance learning, Ultrasound imaging

Title: Machine learning-based small-RNA liquid biopsy enables accurate classification of benign, malignant, and metastatic breast diseases

Author list: Xuhang Liu[1,3], Elizabeth Comen[2], Wenbin Mei[3], Sohail F. Tavazoie[2,3]

Detailed Affiliations:

[1]Department of Pharmacology & Therapeutics, Roswell Park Comprehensive Cancer Center, Buffalo, NY 14203, [2]Department of Medicine, Memorial Sloan-Kettering Cancer Center, New York, NY 10065, USA, [3]Laboratory of Systems Cancer Biology, The Rockefeller University, New York, NY 10065, USA.

Abstract: Mammographic screening is the primary modality for breast cancer detection but suffers from high false-positive rates that drive unnecessary and costly biopsies, causing psychological stress, and delaying the average time to diagnosis for patients with actual malignancies. We hypothesized that circulating small RNAs may provide a minimally invasive molecular complement to imaging for both risk stratification and perhaps systemic cancer staging. We prospectively collected serum from 320 women at MSKCC who were undergoing breast cancer screening or therapy and generated circulating small RNA profiles. Using these data, we trained machine-learning models to classify women as having benign breast disease, localized breast cancer, or metastatic breast cancer. Models based on serum small RNA signatures accurately distinguished benign, malignant, and metastatic diseases. Small nuclear RNA fragments (snRFs), an underexplored class of small RNAs, emerged as major contributors to model performance. A liquid biopsy based on ML-integrated small RNA and snRF signatures could be deployed alongside mammography to reduce unnecessary biopsies, accelerate diagnosis in women with malignancy, and noninvasively identify patients at high risk for underlying metastatic disease. These findings identify snRF/miRNA-informed liquid biopsy as a promising strategy for improving diagnostic precision and detection of metastases in breast cancer.

Keywords: Machine learning, Breast cancer, Metastasis, Small RNA, MicroRNA, snRF

Title: Interpretable spatial programs from multi-sample tumor spatial transcriptomics

Author list: Zhujun Yao, Palak Jolly, Yi Zhang

Detailed Affiliations:

[1]Department of Neurosurgery, Department of Biostatistics and Bioinformatics, Duke University.

Abstract: Spatial transcriptomics has transformed our ability to map gene expression in situ, yet most current analytic frameworks remain limited to single samples or focuses on domain-segmentation tasks on organized tissue architecture, which overlook continuous gradients and recurrent microenvironmental states in complex tissues such as tumors. We here develop MetaSpatial, a computational framework to identify reproducible spatial programs, termed meta-components (MeCs), across multiple high-resolution spatial transcriptomic datasets. MetaSpatial detects reproducible interpretable gene programs based on matrix factorization and detects reproducible spatial niches with distinct gene signatures. Applied to high-definition spatial transcriptomics of breast cancer samples, MetaSpatial recovered 47 recurrent MeCs consisting of immune, fibroblast, and tumor niches, including immune hubs, neuroendocrine-like tumor regions, and metabolically active fibroblasts that were not captured by domain clustering based on spatial gene expression profiles. In glioblastoma Xenium data, MetaSpatial revealed brain tumor spatial programs and active inflammation hotspots preferentially observed in recurrent tumors, with cell-cell communication enriching cancer-microenvironmental signaling. MetaSpatial provides a scalable approach for detecting focal or gradient-like cell states shared across heterogeneous spatial transcriptomic samples that is adaptable to multiple ST platforms, enabling the discovery of frequently occurring microenvironmental programs from multiple samples of tumor.

Keywords: Spatial transcriptomics, Machine learning, Cell state, Gene program, Tumor microenvironment

Title: Computational techniques for cell-level inference of gene effects on human traits

Author list: Xin Liang[1], Ruofan Mao[1], Jinzhuang Dou[1]

Detailed Affiliations:

[1]Biomedical Informatics and Data Science, The University of Alabama at Birmingham, AL, USA.

Abstract: Genome-wide association studies (GWAS) have identified thousands of loci associated with complex traits and diseases, yet mapping these associations to underlying biological mechanisms remains a fundamental challenge. Existing approaches are limited by the predominance of non-coding variants and the lack of cellular

resolution in bulk functional genomics data. While single-cell sequencing (SCS) enables the identification of trait-relevant cell types and states, current methods largely focus on where genetic effects act, without modeling how these effects propagate through molecular networks. Here, we introduce a computational framework for cell-level inference of genetic effects that integrates GWAS signals with gene regulatory networks derived from single-cell data. Our method decomposes gene-level effects into direct components informed by GWAS and indirect components mediated through inferred regulatory interactions. We formulate this integration as a diffusion process over cell-state–resolved gene networks, enabling the propagation of genetic effects across genes, gene programs, and cellular contexts. This diffusion-based formulation captures dynamic relationships between genetic variation and downstream molecular programs, allowing us to model how genetic effects are transmitted through regulatory networks within specific cell populations. Compared to existing approaches that focus on cell-type enrichment or static annotations, our method provides a unified framework for mechanistic inference at single-cell resolution. Our results demonstrate that modeling genetic effects as network diffusion significantly improves the identification of trait-relevant gene programs and reveals context-specific pathways underlying complex traits and diseases. This work establishes a general strategy for bridging statistical genetic associations with dynamic cellular mechanisms.

Keywords: Cell-type–specific genetic mechanism, Single cell, Causal inference, GWAS

Title: A Rigorous Leak-Free Microbiome Pipeline Revealing Batch Effects as the Dominant Challenge in a Multi-Cohort Immunotherapy Response Prediction

Author list: Amey Garg[1]

Detailed Affiliations:

[1] Santa Clara High School, Santa Clara, CA, USA.

Abstract: One of the largest challenges in oncology includes predicting patient response to immune checkpoint inhibitors (ICI). Although gut microbiome composition has been associated with ICI efficacy, current research is limited by either small cohort, data leakage in cross-validation, and inefficient multi-cohort integration. This study evaluates and presents a rigorous, leak-free bioinformatics pipeline for gut-microbiome-based ICI response prediction models, as well as a systematic analysis of batch effects that arise when integrating independent sequencing cohorts. Shotgun metagenomic sequencing data from 118 melanoma patients treated with anti-PD-1 therapy were processed across 3 distinct cohorts (Frankel et al.2017, n=39, HiSeq; n=40, NovaSeq; Matson et al.2018, n=39, NextSeq), spanning BioProjects PRJNA397906 and PRJNA399742. Raw paired-end FASTQ reads were processed using fastp for quality control, while Kraken2 (k2_standard_08gb) was used for genus-level taxonomic classification. Human contamination was removed post-classification, and 113 Candidatus genus name parsing artifacts were corrected. Centred log-ratio (CLR) transformations were applied to the composition abundances. Given feature selections (prevalence filter, variance filter, univariate ranking) were performed strictly within each leave-one-out cross-validation (LOOCV) fold to prevent data leakage. ElasticNet, L1 logistic regression, and Random Forest were evaluated per fold, with significance assessed via N=100 label-permutation tests. PERMANOVA on Aitchison distances revealed that batch effects explained 5.9-7.7% of microbiome variance ($p<0.001$) across all three taxonomic levels (phylum, genus, species), while response explained only 0.7-2.7% (all $p>0.05$), an 8.4-11.3 x batch to signal ratio. Response variance declined monotonically as cohorts were added (2.7% -1.1%-0.7%), while batch variance remained dominant. Cross-cohort holdout validation across all three cohort combinations (3 leave-one-cohort-out, 6 sources to source) produced no significant transfer (best AUC=0.550, $p=0.338$). A six-method batch correction comparison (mean-centering, location-scale, quantile mapping, percentile normalization, cohort-as covariate, uncorrected) applied per-fold on n=118 showed no method-recovered predictive signal. Nested LOOCV with hyperparameter tuning across ElasticNet, Random Forest, and XGBoost yielded AUCs of 0.338, 0.548, and 0.617 (permutation $p=0.99$, 0.24, 0.10), confirming robustness to model class and hyperparameter selection. FDR-corrected cross-cohort feature analysis identified sign inversion in 65% of genera, mechanistically explaining cross-cohort transfer failure. Power analysis established that ~80 patients per protocol are required for 80% statistical power, a threshold no published cohort meets. This pipeline and framework provide infrastructure for future microbiome-based ICI prediction studies.

Keywords: gut microbiome; immunotherapy; batch effects; machine learning; metagenomics; bioinformatics

Workshop – AI and Statistical Methods for Genomics and Multi-Omics
August 2nd
2:00 – 5:20 PM
Room: 2213B

Chairs: Wanding Zhou, Gang Peng

Title: Advancing the Interpretation of Human Genetic Variation Using Generative AI

Author list: Xinghua (Mindy) Shi[1,2]

Detailed Affiliations:
[1]Department of Computer & Information Sciences, Temple University, Philadelphia, PA, USA, [2]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA, USA.

Abstract: T Interpreting the functional impact of human genetic variation is fundamental to advancing human genetics, genome biology, and precision medicine. Although tens of millions of genetic variants have been identified through large-scale sequencing efforts, the molecular and phenotypic consequences of many of these variants, particularly structural variants, remain poorly understood. In this talk, I will present recent methods and resources for improving the characterization and interpretation of human genetic variation, with a particular focus on structural variation. I will discuss approaches to identifying functionally relevant variants, linking genetic variation to molecular traits and human phenotypes, and assessing the impact of structural variants on three-dimensional (3D) genome organization and gene regulation. I will then introduce state-of-the-art generative artificial intelligence (AI) approaches for modeling complex genomic data, highlighting how foundation models and generative learning frameworks can address key challenges in analyzing both linear (2D) genome sequences and higher-order (3D) genome architecture. Together, these advances provide new opportunities to improve variant interpretation, uncover the mechanisms by which genetic variation influences genome function, and accelerate the translation of genomic discoveries into precision medicine.

Keywords: Human Genetic Variation, Structural Variation, 3D Genome Structure, Generative AI, Foundation Models

Title: Development of Epigenetic Clocks from Oxford Nanopore Methylation Data in Human Caudate Tissue

Author list: Steven Brooks[1], Melissa Robins[1], Yunlong Liu[1], and Gang Peng[1]

Detailed Affiliations:
[1]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, 46202, USA.

Abstract: Background: Epigenetic clocks have emerged as valuable tools in many areas of biomedical research. Most existing epigenetic clocks were developed using array-based DNA methylation data. Oxford Nanopore Technologies (ONT) long-read sequencing has recently been applied to DNA methylation profiling. However, few studies have evaluated the performance of array-based epigenetic clocks on ONT-derived methylation data, and robust epigenetic clocks specifically developed for ONT data are still lacking.

Methods: DNA methylation in human post-mortem caudate tissue from 144 donors was measured using ONT sequencing. Several widely used epigenetic clocks originally developed from array data were applied to ONT-derived methylation profiles. Their performance was assessed under different read-depth cutoffs and different approaches for estimating methylation proportions. We then developed new ONT-based epigenetic clocks using traditional clock-building strategies on multiple methylation-related metrics, including total modification, 5mC proportion, 5hmC proportion, hFlux (5hmC / [5hmC + 5mC]), and mFlux (5mC / [5mC + C]). Clock performance was evaluated using nested cross-validation. Given the sparsity and relatively low abundance of 5hmC, we also constructed epigenetic clocks based on methylation estimates aggregated within predefined genomic windows.

Results: Array-based epigenetic clocks developed from blood samples showed only weak correlations with chronological age when applied to ONT data (correlation coefficients < 0.5). In contrast, Shireby's clock, which was developed from cortex samples, showed a stronger correlation with chronological age ($r \approx 0.73$), suggesting that array-based clocks may be transferable to ONT data when developed in biologically similar tissues. Among the ONT-based clocks, the model based on 5mC proportion achieved the best performance, with a correlation coefficient of 0.83, whereas models based on other metrics showed correlations ranging from 0.56 to 0.69. When

donors in the first and last age quartiles were compared using logistic regression, all metrics demonstrated good classification performance, with AUCs ranging from 0.71 to 0.92. These findings suggest that, at limited read depth, ONT data may not reliably detect age-related changes in 5hmC given its low abundance (typically below 20%) over small age differences but can capture broader age-related differences when age groups are more distinct. We further estimated average methylation levels within 10 kb and 100 kb genomic windows to build window-based clocks. This strategy improved performance for clocks based on 5hmC proportion and hFlux.

Conclusion: These results demonstrate the promise of ONT for epigenetic clock development and suggest that ONT-specific, tissue-matched, and window-based approaches may improve biological age estimation from long-read methylation data.

Keywords: Epigenetic clock, ONT, 5-hydroxymethylcytosine, Aging, Brain tissue

Title: The Cancer Proteome Atlas: From Functional Proteomics to Next-Generation Analytics

Author list: Jun Li[1,2,3,4], Wei Liu[5], Kamalika Mojumdar[5], Yitao Tang[1,5], Yining Zhao[5], Liudeng Zhang[5], Yiling Lu[6], Rehan Akbani[5], Gordon B. Mills[7], Han Liang[5,8]

Detailed Affiliations:

[1]School of Data and Information Sciences, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, [2]UNC Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, [3]Curriculum in Bioinformatics and Computational Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, [4]Computational Medicine Program, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, [5]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA, [6]Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA, [7]Knight Cancer Institute and Cell, Developmental and Cancer Biology, Oregon Health and Science University, Portland, OR, USA, [8]Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

Abstract: Functional proteomics provides critical insights into cancer biology by directly quantifying protein abundance and signaling pathway activity, facilitating the discovery of novel biomarkers, therapeutic targets, and disease mechanisms. We have developed The Cancer Proteome Atlas (TCPA), a comprehensive cancer functional proteomics resource based on reverse phase protein arrays (RPPA), comprising profiles from nearly 8,000 patient tumors from The Cancer Genome Atlas (TCGA) and ~900 cancer cell lines from the Cancer Cell Line Encyclopedia (CCLE). The resource includes nearly 500 clinically relevant protein markers spanning all major cancer hallmark pathways and has enabled numerous biological and translational discoveries across diverse cancer types. To further enhance the accessibility and analytical capabilities of this resource, we developed DrBioRight, an AI-powered bioinformatics platform that leverages state-of-the-art large language models to enable intuitive exploration of cancer proteomics and multi-omics data. Through natural language interactions, researchers can perform advanced analyses, generate publication-quality visualizations, interpret biological findings, and interactively explore complex molecular datasets without requiring programming expertise. By combining a large-scale functional proteomics resource with next-generation AI-enabled analytics, TCPA and DrBioRight provide an integrated framework that accelerates biological discovery and advances precision cancer research.

Keywords: Functional Proteomics, Integrative Multi-Omics Analysis, Large Language Models, Precision Oncology

Title: Unified Statistical Inference Beyond Gene Expression in Single-Cell RNA Sequencing

Author list: Guoshuai Cai[1,2,3], Dayuan Wang[2,3]

Detailed Affiliations:

[1]Department of Surgery, College of Medicine, University of Florida, Gainesville, FL, 32610, USA, [2]Department of Biostatistics, College of Public Health and Health Promotions & College of Medicine, University of Florida, Gainesville, FL, 32610, USA, [3]Sepsis and Critical Illness Research Center, University of Florida, Gainesville, FL, 32610, USA

Abstract: Single-cell RNA sequencing (scRNA-seq) profiles transcriptomes at single-cell resolution, enabling diverse downstream analyses, including cell composition, intercellular communication, differentiation dynamics, and drug–cell targeting. While these analyses have generated biologically meaningful insights, they also reveal a

broader methodological challenge: many scRNA-seq–derived features extend beyond raw gene expression and lack a unified statistical framework for rigorous inference that accounts for biological variation, complex experimental designs, and heterogeneous outcome structures. To address this gap, this study develops a comprehensive statistical inference framework for derived single-cell features, enabling principled hypothesis testing across a broad range of biologically meaningful summaries while maintaining compatibility with established scRNA-seq analytical pipelines. The first component establishes guidelines for upstream scRNA-seq preprocessing, with particular emphasis on how preprocessing decisions propagate to downstream statistical inference. It prioritizes preserving biological replication and maintaining consistent sample-level metadata throughout the analytical workflow. The second component provides structured, inference-oriented pipelines for generating derived features from scRNA-seq data. The third component develops a unified inferential strategy for the resulting standardized feature matrices. The central principle is to preserve biological replication while explicitly accounting for the correlation structure induced by the study design. The fourth component focuses on the visualization, interpretation, and summarization of inferential results. Together, this framework provides a unified statistical foundation for rigorous inference across diverse scRNA-seq–derived features, facilitating more convenience and reliable biological discovery from single-cell studies.

Keywords: Single-cell RNA sequencing (scRNA-seq), Analysis framework, Statistical inference, Complex study design

Title: Can We Trust Protein Search in the Era of Large Language Models?

Author list: Lingfeng Shi, Han Zhou, Yang Young Lu*

Detailed Affiliations:

Department of Biomedical Engineering and Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI, USA. *Corresponding author

Abstract: Fast protein structure search tools such as Foldseek and Reseek are rapidly becoming indispensable for homology detection, annotation transfer, and structural database mining. Their E-values appear to provide an interpretable statistical assessment of structural similarity, analogous to BLAST in sequence search. Here we show that these E-values can be seriously miscalibrated, with Foldseek tending to be too conservative and Reseek at times too liberal. We instead estimate significance by generating a small query-specific null sample through shuffling the per-residue structural representation used by the search engine and comparing the observed score against the resulting random optimal scores. Across the settings we examined, the resulting p-values were never liberal, substantially improving calibration for Foldseek and correcting anti-conservative behavior for Reseek. Our results suggest that native E-values in modern structure search should be interpreted with caution and that this framework offers a practical route to statistically reliable, error-controlled structure similarity discovery.

Keywords: protein structure search, structural similarity, remote homology detection, structural bioinformatics

Title: Transposable Element–Driven Viral Mimicry Constrains Multistep Pancreatic Cancer Progression

Author list: Siyu Sun[1,*,#], Julie K Bray[2,*], Kimal Rajakshe[2,+], Benson Chellakkan Selvanesan[2,+], Yuhui Song[3], Linda T. Nieman[3], Changsoo Kwak[2], Fredrik Thege[2], Meelad Dawlaty[4], Amit Verma[4], Alexander Solovyov[1], David T. Ting[3], Benjamin D. Greenbaum[1,5;#], Anirban Maitra[2,6#]

Detailed Affiliations:

[1]Halvorson Computational Oncology, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA, [2]Sheikh Ahmed Center for Pancreatic Cancer Research, University of Texas MD Anderson Cancer Center, Houston, TX, [3]Mass General Brigham Cancer Institute, Boston, MA 02114, USA, [4]Albert Einstein College of Medicine, New York City, NY, [5]Physiology, Biophysics & Systems Biology, Weill Cornell Medicine, Weill Cornell Medical College, New York, NY, USA, [6]Current address; Perlmutter Cancer Center, New York University Langone Health, New York, NY 10016. *,+ denote equal contribution, # Correspondence

Abstract: Transposable elements (TEs) comprise nearly half of the mammalian genome and have long been considered genomic "dark matter." Increasing evidence, however, suggests that their aberrant activation in cancer

can generate endogenous double-stranded RNAs that engage innate immune pathways through a process known as viral mimicry. While this response has been implicated in therapeutic sensitivity, its role in shaping the natural evolution of cancer remains poorly understood.

In this seminar, I will present our work investigating how transposable element expression influences pancreatic cancer progression. By integrating multi-regional patient cohorts, genetically engineered mouse models, functional genomics, and single-cell transcriptomics, we identify evolutionarily young SINE elements as potent drivers of innate immune activation. We show that TE-derived viral mimicry imposes an early barrier to tumor progression, creating selective pressure for cancer cells to suppress immunogenic repeats through multiple adaptive mechanisms, including epigenetic silencing and context dependent regulatory programs. These findings support a model in which the dynamic interplay between transposable elements and innate immunity constitutes a bottleneck that constrains multistep pancreatic tumor evolution.

Together, this work highlights transposable elements as active participants in cancer evolution rather than passive genomic bystanders and suggests that manipulating repeat-mediated viral mimicry may offer new opportunities for cancer immunotherapy and early intervention.

Keywords: Transposable elements, viral mimicry, cancer evolution, innate immunity

Title: Single-molecule RNA modification mapping by AI analysis of nanopore sequencing

Author list: Yunxia Wang[1,2], Yanqiang Li[1,2], Lili Zhang[1,2], Kaifu Chen[1,2]

Detailed Affiliations:

[1]Department of Cardiology, Boston Children's Hospital, [2]Department of Pediatrics, Harvard Medical School, [3]Lead contact: Kaifu.chen@childrens.harvard.edu

Abstract: 2'-O-methylation (Nm) is well known at some sites in noncoding RNAs and has recently emerged as an abundant modification widespread in mRNAs. However, accurate de novo site detection, stoichiometric quantification, and single-molecule mapping of Nm across the transcriptome remain major challenges. Here, we develop NanoNmD, an AI model that decodes Nanopore ionic current signals to achieve single-molecule Nm mapping by Deep learning. NanoNmD generalizes robustly to sequence contexts and RNA types unseen during model training, enabling transcriptome-wide detection beyond constraints of training data sparsity. Applying NanoNmD to human, Drosophila, and yeast datasets further demonstrated cross-species generalization. Finally, NanoNmD was benchmarked on both earlier ONT RNA002 and the latest ONT RNA004 sequencing chemistries, outperforming prior open-source models as well as the commercial model Dorado in both site mapping and stoichiometric quantification. Together, NanoNmD provides a powerful computational framework for decoding the dynamic regulatory dimensions of RNA Nm modification across diverse biological contexts.

Keywords: RNA modification, 2'-O-methylation, Nanopore direct RNA sequencing, Deep learning, single-molecule, transcriptome-wide

Title: Recurrent Methylation Pattern Encoding for Interpretable Analysis of Sparse DNA Methylomes

Author list: Hongxiang Fu, Hao Xu, Chin Nien Lee, Cameron Cloud, Yanxiang Deng, Wanding Zhou

Detailed Affiliations:

Center for Computational and Genomic Medicine, The Children's Hospital of Philadelphia, Philadelphia, PA, 19104, USA, Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA

Abstract: DNA methylation is a foundational epigenetic mark widely used for both biological discovery and clinical diagnostics, yet its analysis is uniquely difficult because methylomes are far higher-dimensional and far sparser than other omics measurements. Single-cell and spatial assays typically recover only a small fraction of the genome's roughly 30 million CpGs, and conventional window- or region-based methods lose power under such extreme sparsity. We address this by learning global co-methylation structure rather than local context. By enumerating genome-wide CpG sets that vary together, we identify Most Recurrent Methylation Patterns (MRMPs), which capture long-range, punctate correlations reflecting focal regulatory activity that region-based approaches overlook. A relatively small number of MRMPs account for the majority of genomic CpGs and reveal higher-order organizing principles of epigenetic programming, providing a compact and interpretable basis for modeling methylation variation. Building on this representation, we developed a fast machine learning framework in which

simple additive MRMP encoders yield low-dimensional embeddings analogous to those of foundation models, but without costly pretraining. This encoding supports rapid cell-type annotation, mini-bulk deconvolution of spatial methylation maps, whole-genome upscaling, and cell embeddings with improved lineage resolution. Decoder networks reconstruct cell-type-specific methylomes from as little as roughly 0.1% CpG coverage with high accuracy, including for cell types and culture states absent from training. The framework remains reliable at sparsity levels where existing tools fail, and it operates consistently across sequencing and array platforms, offering a scalable and unified approach for resolving cell identity, tracing cancer cell of origin, and rescuing sparse methylation profiles.

Keywords: DNA methylation; epigenetics; single-cell omics; spatial omics; methylome imputation; representation learning; cell-type deconvolution

Title: TBD

Author list: TBD

Detailed Affiliations:
[1]TBD

Abstract: TBD

Keywords: TBD

Workshop – From Data to Models to Agents: Using LLMs to Accelerate Clinical Development and Personalize Patient Care August 2nd 2:00 – 5:20 PM Room: 1220

Chairs: Huanmei Wu, Yuzhou Chen

Title: Dual-stream cross-modal alignment for clinical trial outcome prediction

Author list: Xinyue Wang[1], Chenguang Yang[2], Huanmei Wu[3], Thomas L. Martin[4], Yulia Gel[4], Yuzhou Chen[2]
Detailed Affiliations:
[1]UC Berkeley, [2]UC Riverside, [3]Temple University, [4]Virginia Tech

Abstract: Multimodal representation learning seeks to capture both shared and complementary semantics across heterogeneous data sources. Yet, effectively integrating diverse modalities remains a fundamental challenge. This challenge is particularly pronounced in prediction of clinical trial outcomes, where reliable inference must synthesize complex multi-source information. While clinical trials remain the gold standard for evaluating the safety and efficacy of therapeutics, accurately predicting their outcomes is inherently challenging due to the complex interplay of heterogeneous factors, including drug properties, disease characteristics, and patient eligibility criteria. Hence, there is a critical need for predictive models that can effectively integrate diverse data modalities to anticipate trial success or failure early in the development process. In this work, we propose an innovative Dual-stream Cross-modal Alignment (DCA) framework that jointly models structured clinical records, temporal patient trajectories, and unstructured eligibility text to improve clinical trial outcome prediction. Our extensive experiments on longitudinal amyotrophic lateral sclerosis (ALS) and stroke patient data demonstrate that our DCA consistently outperforms state-of-the-art methods. These results underscore the effectiveness of DCA in enabling robust cross-modal reasoning, achieving superior alignment while preserving modality-specific signals, and establishing a principled foundation for multimodal learning in clinical decision-making.

Keywords: Cross-modal alignment, Clinical trial outcome prediction, Representation learning

Title: SepsisGuard-ICU: An Evidence-Grounded Multi-Agent System for Safe and Explainable Sepsis Decision Support in the ICU

Author list: Yanwen Wang[1], BinBin Li[2], Fan Zhong[1], Ming Zhong[2], Lei Liu[1,*]

Detailed Affiliations:
[1]Intelligent Medicine Institute, Fudan University, 138 Yixueyuan Road, Xuhui District, Shanghai, 200032, China,
[2]Department of Critical Care Medicine, Zhongshan Hospital, Fudan University, Shanghai, 200032, China.
[*]Corresponding author

Abstract: Sepsis is a life-threatening organ dysfunction driven by a dysregulated host response to infection. In the intensive care unit (ICU), effective management requires continuous interpretation of heterogeneous, rapidly evolving data—vital signs, laboratory tests, blood gases, microbiology, imaging, organ support, and clinical notes. Existing decision support systems largely rely on static scores, isolated prediction models, or non-explainable alerts, offering limited support for dynamic reassessment, evidence traceability, missing-information detection, safety review, or clinician-in-the-loop decision-making. We present SepsisGuard-ICU, an end-to-end multi-agent clinical decision support system for ICU sepsis monitoring, early warning, structured assessment, and explainable recommendation generation. The system integrates guideline consensus documents, guideline-derived Q&A pairs, physician-curated QA data, de-identified real-world ICU sepsis cases, and MIMIC-IV sepsis data (structured + textual). SepsisGuard-ICU features a two-level architecture: (1) a lightweight warning layer for deterministic scoring, critical-value detection, trend monitoring, and rule-based risk triggering. (2) a deep reasoning layer combines retrieval-augmented generation with five clinical agents (infection, hemodynamics, respiration, kidney function, and medication safety) and three workflow agents (coordination, safety gating, final summarization). Unlike conventional alert-oriented systems, SepsisGuard-ICU links every output to patient-specific findings, guideline-derived evidence, missing variables, uncertainty statements, safety flags, and physician-confirmation requirements. Deterministic tools handle scoring, rule execution, and evidence retrieval, while large language models are constrained to contextual synthesis, conflict recognition, and explainable reasoning. A visual ICU dashboard supports patient timeline review, risk cards, scoring trends, physiological/lab trajectories, agent-level outputs, evidence panels, audit logs, and clinician feedback. SepsisGuard-ICU transforms large language models from passive question-answering tools into evidence-grounded, safety-controlled, and clinically auditable decision support systems. The platform is designed for retrospective case replay, expert review, offline safety assessment, and future prospective validation of AI-assisted sepsis care in the ICU—directly addressing the conference theme of AI-driven clinical transformation and personalized care.

Keywords: SepsisGuard-ICU, Multi-agent clinical decision support, Retrieval-augmented generation, Explainable AI, ICU sepsis monitoring

Title: Developing social work language models for SYNC-Smart AI assistant

Author list: Phillip McCarllion[1,2]

Detailed Affiliations:
[1]School of Social Work, [2]College of Public Health, Temple University.

Abstract: As the aging population expands, older adults increasingly strive to maintain independence while relying on caregivers for daily support when they experience chronic conditions such as Parkinson's disease. In diseases such as Parkinson's much time is taken up with arranging appointments, checking on permitted insurance expenditures, responding to changing information needs as the disease progresses and arranging appointments, transportation and receipt of supplies. There is little time for social workers and others to address critical psychosocial needs of the Person with Parkinson's and the caregiver. This work presents preliminary efforts to develop an intelligent robotic companion powered by agentic large language models (LLMs), designed to assist with routine tasks, medical appointments, and medication reminders for social workers. Central to design are principles of interpretability, reproducibility, and trustworthiness, achieved through the integration of the n8n framework. Based on social work practice principles and interviews with social workers, caregivers and care recipients, we have structured common questions typically asked by social workers and responses from caregivers and care recipients, organizing them into a decision tree. This decision tree forms the foundation of the chatbot, enabling conversations that are both structured, adaptable, and with the flexibility to respond contextually to caregiver and care recipient inputs. The SYNC-smart system is built using n8n, a workflow automation platform,

to integrate AI capabilities with patient care processes for efficient multi-agent management, organized into four functional categories: 1) a master agent that classifies user input and routes it to downstream agents; 2) an information collection agent; 3) a daily care recipient support agent guided by expert-defined decision trees that promote individual well-being and address ongoing needs, while securely storing data in local knowledge bases accessible only to authorized users; and 4) task-oriented agents that leverage external tools or internal knowledge to perform specific tasks or deliver targeted responses.

A scalable and dependable solution that has the potential to enhance safety, autonomy and care for older adults with Parkinson's and for their caregivers. Next steps will include testing of feasibility and usability, development of implementation manuals and small and larger scale trials with outcome measurement targeting knowledge, psychosocial concerns and well-being and cost effectiveness/utility of use.

Keywords: Synthetic data, generative AI, phenotype extraction, rare disease

Title: BehaviorGround: a fair benchmark for cognitive decline detection under subjective-complaint contamination

Author list: Sambit Mukherjee[1], Dezhi Wu[1*], and Chih-Hsiang Jason Yang[2]

Detailed Affiliations:
[1]HI3 TECH LAB, University of South Carolina, Columbia, SC 29208, USA, [2]Arnold School of Public Health, University of South Carolina, Columbia, SC 29208, USA. [*]Corresponding author

Abstract: Digital phenotyping offers a promising pathway for detecting early cognitive decline by replacing infrequent clinic-based memory assessments with continuous, low-burden measurements from smartwatches and brief smartphone-based tasks in daily life. Recent studies have reported high accuracy in predicting Subjective Cognitive Decline (SCD), an early stage of the Alzheimer's disease continuum. However, we show that much of this apparent success may be inflated by a previously underexamined failure mode: subjective-complaint contamination. In many existing models, both the outcome labels and the most predictive input features are derived from self-reported complaints, including perceived memory difficulties, mental sharpness, fatigue, anxiety, and related subjective states. As a result, models may learn to predict one self-report from another rather than identifying behavioral signals of cognitive change.

To address this problem, we introduce BehaviorGround, a benchmark and modeling framework designed to evaluate digital phenotyping models against stricter, more clinically meaningful standards. BehaviorGround restricts predictive evidence to objective signals, including wearable-derived measures of walking, sleep, and daily activity, as well as cognitive-game performance metrics such as reaction times, response variability, and error rates. Subjective experiences, including perceived memory difficulty, fatigue, and anxiety, are retained as clinically relevant contextual variables but are excluded as direct predictive inputs. The framework further incorporates an adversarial scrubbing layer to reduce residual self-report-like information that may leak into objective features, and it adopts participant-level train-test separation to prevent the same individual's data from appearing in both training and testing sets.

When we re-evaluated recent popular models under the BehaviorGround protocol, their performance drops substantially, indicating that previously reported accuracy depended heavily on leaked subjective complaints and participant-level shortcuts. Although BehaviorGround yields more modest and predictive performance, its estimates provide a more credible measure of whether behavioral data alone can support early detection of cognitive decline. More broadly, this work offers a generalizable guardrail against proxy leakage in high-stakes predictive modeling with implications for healthcare, finance, insurance, hiring, and fraud detection. Our BehaviorGround framework provides the field with a cleaner benchmark, a diagnosis of why current progress may be overstated, and a more trustworthy foundation of digital biomarkers of cognitive decline.

Keywords: Digital phenotyping, cognitive decline, data leakage, wearable sensors, adversarial debiasing, benchmarking

Title: Leveraging synthetic text and image data to improve the robustness and generalizability of biomedical AI models

Author list: Kai Wang[1,2]

Detailed Affiliations:

[1]Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, [2]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

Abstract: The rapid emergence of generative artificial intelligence (AI) is transforming how biomedical data can be created, shared, and utilized for clinical AI model development. In healthcare settings where access to large-scale, diverse, and well-annotated patient datasets remains constrained by privacy, rarity, cost, and regulatory limitations, AI-generated synthetic clinical text and image data offer a promising strategy to augment real-world datasets for both machine-learning and benchmarking applications.

This presentation will highlight two applications of synthetic biomedical data. First, we will discuss the generation of synthetic clinical notes using large language models to improve AI-based recognition of clinical phenotypes from unstructured electronic health records, particularly for rare diseases and rare phenotypes. Second, we will examine the use of synthetic patient photo generation to improve AI models for facial phenotype recognition on rare genetic disorders across diverse ancestry and age groups, addressing a limitation of current prediction models trained on small and demographically biased datasets.

The talk will further discuss approaches for evaluating synthetic data fidelity, fairness, and downstream clinical utility, as well as challenges related to hallucination, bias propagation and privacy protection. Together, these applications demonstrate how synthetic data can help overcome clinical data scarcity and improve the robustness, generalizability, and equity of next-generation healthcare AI systems.

Keywords: TBD

Title: From EHR phenotyping to proactive maternal care: Trustworthy AI for early risk detection and personalized decision support

Author list: [1]Mukesh Patel, [2]Carissa Pekny, [1]Huanmei Wu

Detailed Affiliations:

[1]Barnett Colleg of Public Health, Temple University, Philadelphia, PA, [2]Reproductive Endocrinology and Infertility, University of California, San Diego, CA

Abstract: Artificial intelligence can shift maternal care from retrospective risk documentation and episodic intervention toward earlier detection, longitudinal monitoring, and personalized decision support. Realizing this potential requires more than predictive accuracy. Trustworthy clinical AI must be grounded in reproducible electronic health record (EHR) phenotyping, clinically meaningful outcomes, temporally reliable data, calibrated predictions, and transparent tools that support clinical judgment and shared decision-making. Using severe maternal morbidity (SMM) as a motivating use case, this presentation describes an end-to-end framework for responsible AI-enabled maternal care. Retrospective cohorts of delivery hospitalizations were constructed from MIMIC-IV and Healthcare Cost and Utilization Project data from five states, with SMM identified using 21 CDC-defined indicators. Analyses characterize SMM prevalence, maternal risk factors, clinical phenotypes, co-occurring complications, healthcare utilization, and outcomes. Building on this phenotyping foundation, we are developing interpretable AI models that integrate longitudinal demographic, clinical, laboratory, medication, procedural, and utilization data to identify high-risk patients earlier. A complementary visualization and decision-support platform will present patient histories, laboratory trajectories, prior complications, changing risk estimates, explanations, and uncertainty in clinically interpretable formats. Expected outcomes include a validated, reusable SMM phenotyping pipeline, calibrated early-risk models, and an interactive dashboard that improves recognition of clinically meaningful temporal changes. Initial findings reveal interconnected hemorrhagic, renal, infectious, cardiovascular, and respiratory complications. SMM is also associated with higher frequencies of preeclampsia, chronic hypertension, pre-existing diabetes, preterm and cesarean delivery, prolonged hospitalization, and intensive clinical interventions. Sub-cohort modeling will assess demographic variation, geospatial patterns, and temporal trends to strengthen generalizability and support personalized care planning. Model evaluation will address discrimination, calibration, uncertainty, subgroup fairness, clinical lead time, and temporal, geographic, and external validation. By integrating

rigorous phenotyping, predictive modeling, and workflow-oriented decision support, this work provides a reproducible pathway toward proactive, personalized, and responsibly deployed AI in maternal health.

Keywords: Severe maternal morbidity; EHR phenotyping; Trustworthy AI; Personalized clinical decision support

Title: AI in biomedical research-- hypothesis generation, drug target and activity prediction

Author list: Jie Zhang[1]

Detailed Affiliations:

[1]Indiana University.

Abstract: The biomedical research field has witnessed a dramatic increase of AI involvement in almost all aspects, including literature analysis, disease mechanistic hypothesis generation, drug target identification and molecular activity prediction. This talk will go over several studies in our lab that involve AI approach in above mentioned areas. Specifically, I will talk about AI in small molecule activity prediction, in patient subtyping, as well as the development of an agentic AI platform ARSA (the Alzheimer Research Scientist Agent), an agentic AI framework designed to support evidence-integrated hypothesis generation and target discovery in the Alzheimer's disease research.

Keywords: Artificial intelligence; Drug discovery; Alzheimer's disease

Title: Prompt-tuned open-source large language model for structured extraction of social determinants of health from clinical notes

Author list: Cheok Long Tang[1], Yu Huang[1,2], Jiang Bian[1,2], Hao Liu[1,3], Yan Zhuang[1]

Detailed Affiliations:

[1]Department of Biomedical Engineering and Informatics, Luddy School of Informatics, Computing, and Engineering, Indiana University Indianapolis, Indianapolis, IN, [2]Department of Biostatistics and Health Data Science, School of Medicine, Indiana University, Indianapolis, IN, [3]School of Computing, College of Science and Mathematics, Montclair State University, Montclair, NJ

Abstract: Social determinants of health (SDoH) are critical for patient stratification, risk prediction, population health management, and personalized care, yet much of this information remains embedded in narrative electronic health record notes. Existing NLP approaches often rely on static rules, institution-specific lexicons, or fine-tuned models that produce flat labels and are difficult to update as definitions and documentation practices evolve. We developed a modular, logic-guided prompting framework that uses an open-source large language model to transform narrative clinical documentation into structured, explainable SDoH variables without model retraining.

The framework uses factor-specific prompts, explicit inclusion and exclusion rules, logic-gate checks, structured outputs, and evidence sentences to capture not only whether an SDoH factor is present, but also clinically meaningful attributes such as status, type, and temporality. Using 80 clinical notes from the Indiana Network for Patient Care, we evaluated seven SDoH factors: alcohol use, tobacco use, drug use, marital status, sleep, family support, and sexual activity. The model achieved macro-precision, macro-recall, and macro-F1 of 0.71, 0.93, and 0.79 at the factor level.

These findings suggest that prompt-based open-source LLMs can support scalable, explainable, and adaptable extraction of social-risk information from real-world clinical documentation. By converting hidden narrative EHR information into structured and evidence-supported variables, this approach provides a practical foundation for downstream clinical AI applications, including patient stratification, risk prediction, and personalized care.

Keywords: Social determinants of health, large language models, clinical notes, information extraction, prompt tuning, EHR

Title: Identifying subgroups in patient with low back pain for personalized treatment using machine learning

Author list: Leming Zhou[1], William Anderst[2], Kevin M. Bell[3], Carol M. Greco[4,5], Gina P. McKernan[6], Charity G. Patterson[5], Sara R. Piva[5], Michael J. Schneider[7], Nam V. Vo[2], Ajay Wasan[8], Gwendolyn A. Sowa[2,3,6]

Detailed Affiliations:

[1]Department of Health Information Management, School of Health and Rehabilitation Sciences, [2]Department of Orthopaedic Surgery, School of Medicine, [3]Department of Bioengineering, Swanson School of Engineering,

[4]Department of Psychiatry, School of Medicine, [5]Department of Physical Therapy, School of Health and Rehabilitation Sciences, [6]Department of Physical Medicine and Rehabilitation, School of Medicine, [7]Doctor of Chiropractice Program, School of Health and Rehabilitation Sciences, [8]Department of Anesthesiology and Perioperative Medicine, School of Medicine; University of Pittsburgh, Pittsburgh, PA, USA
Abstract: Chronic low back pain (cLBP) is clinically heterogeneous, and identifying reproducible patient subgroups may support more targeted, mechanism-informed treatment. Unsupervised clustering of baseline data offers a data-driven route to such subgroups. The objective of this study is to identify and characterize cLBP phenotypes through unsupervised clustering of multidomain baseline data, and to rigorously evaluate the stability and robustness of the resulting subgroups. We analyzed baseline data from 1007 participants in the University of Pittsburgh Low Back Pain Biological, Biomechanical, and Behavioral (LB3P) project. Candidate features spanned multiple domains: demographics, comorbidities, patient-reported outcomes, physical and biomechanical examination, clinical and biological biomarkers, and electronic-medical-records. After standardized preprocessing (missing data imputation and data transformation), we ranked features by the unsupervised Laplacian Score and retained the 50 most informative features. We used K-Prototypes (mixed-type) and K-Means (numeric) for clustering, each evaluated at k = 3-5 across 15 random initializations to assess clustering reliability. The primary solution was selected by integrating across-seed reliability, agreement between the reduced and full feature sets, and cross-algorithm concordance. Clusters were compared across all baseline features. The Laplacian top-50 feature set reproduced the full-feature K-Prototypes clustering (ARI = 0.91), and K-Prototypes and K-Means agreed substantially on the top-50 features (ARI = 0.98), indicating features beyond top 50 contributed little to the partition of patient subgroups. K-Means on the top-50 features at k = 4 was selected as the primary solution and showed high across-seed reliability (mean ARI = 0.96). The 4 phenotypes differed significantly in 95 features, most prominently in measures such as pain intensity, psychosocial burden, and physical function. In conclusion, Unsupervised clustering of multidomain baseline data identified 4 reproducible cLBP phenotypes that were robust to algorithm choice and feature-set size. These data-driven subgroups can guide the personalized treatment selection.
Keywords: Low back pain, personalized treatment, statistics, machine learning, Artificial Intelligence

Workshop – Precision Medicine and Cancer Systems
August 3rd
9:00 AM – 12:00 PM
Room: 2220A&B

Chair: Rachael Hageman Blair

Title: Gain and Loss of DNA Methylation During Tumor Evolution

Author list: Deyana Tabatabaei[1,2], Sudhir Kumar[1,2], and Sayaka Miura[1,2,3*]

Detailed Affiliations:
[1]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, Pennsylvania, USA, [2]Department of Biology, Temple University, Philadelphia, Pennsylvania, USA, [3]Department of Biology, University of Mississippi, University, Mississippi, USA. [*]Corresponding author
Abstract: Advances in sequencing and array technologies have enabled detailed characterization of genetic and epigenetic heterogeneity within tumors. However, the evolutionary dynamics of genetic and epigenetic alterations during cancer initiation and progression remain poorly understood. To investigate these dynamics, we reconstructed tumor phylogenies using single-nucleotide variants (SNVs) identified from multiple tumor samples obtained from patients with colorectal cancer (CRC) and hepatocellular carcinoma (HCC). We then mapped somatic mutations and gains and losses of DNA methylation onto each patient's tumor phylogeny to infer the timing and lineage specificity of these alterations during tumor evolution. Our analyses revealed that gains and losses of DNA methylation occurred more frequently than somatic mutations across tumor evolution. In some patients, methylation changes were also more prevalent along later branches of the tumor phylogeny, whereas genetic mutations were more often observed earlier in tumor development. In addition, some genes accumulated both driver mutations and driver-associated methylation changes at different evolutionary stages and along different lineages, generating

heterogeneity in the number and timing of driver events even within the same gene. Together, these findings suggest that epigenetic alterations accumulate extensively during tumor progression and may play an important role in the functional diversification of cancer cell populations.

Keywords: DNA methylation, somatic evolution, tumor evolution, tumor phylogeny

Title: Deep Learning for Early Hepatocellular Carcinoma Risk Prediction in Patients with Chronic Liver Diseases and Diabetes: Insights from Retrospective Cohort Analysis

Author list: Jinlian Wang[1], Hui Li[2], Changqing Ju[2], Hongfang Liu[1]

Detailed Affiliations:
[1]UTHealth Houston School of Biomedical Informatics, Houston, TX, USA, [2]McGovern Medical School UTHealth, Houston, TX, USA.

Abstract: Objective: To develop and evaluate a deep learning model for predicting 1-year survival in hepatocellular carcinoma (HCC) patients, focusing on the prognostic interplay between chronic liver diseases (e.g., cirrhosis, viral hepatitis) and diabetes as key comorbidities, aligned with the European Association for the Study of the Liver (EASL) Clinical Practice Guidelines (2018). Methods: Analysis was performed on the UCI HCC Survival Dataset (n=165 patients, 49 features selected per EASL-EORTC guidelines). Preprocessing included imputation of ~10% missing values using means for quantitative variables. Exploratory data analysis (EDA) assessed correlations (Pearson's r) and associations (chi-square) among diabetes, cirrhosis, viral markers, and survival. A logistic regression baseline and multilayer perceptron neural network (MLP; 3 layers: input=16 → 32 → 16 → 1, ReLU/sigmoid activations, PyTorch) were trained on 16 features (e.g., diabetes, AST, platelets), with an 80/20 train/test split. Performance was evaluated via accuracy and AUC-ROC, with 95% confidence intervals via bootstrap resampling. Results: EDA revealed diabetes prevalence of ~38% and a negative correlation with survival ($r = -0.11$, $\chi^2 = 4.12$, $p = 0.04$), alongside decompensated liver markers like AST ($r = -0.19$). The MLP achieved 70% accuracy and AUC- ROC of 0.73 (95% CI: 0.56-0.90), outperforming logistic regression (61% accuracy, AUC-ROC 0.73). Diabetes reduced survival odds by ~30% (logistic coefficient -0.35), consistent with its additive risk in chronic liver contexts. Discussion: Results corroborate EASL guidelines' emphasis on diabetes as an exacerbating factor in HCC progression via metabolic syndrome and NAFLD, with the MLP's non-linear modeling offering superior capture of interactions (e.g., diabetes and AST elevation). Limitations include small sample size and retrospective design, potentially biasing toward European cohorts; strengths lie in guideline alignment and interpretability for clinical translation. Conclusion: This DL framework provides a robust tool for HCC risk stratification in high-risk diabetic/chronic liver patients, operationalizing EASL recommendations to advance precision oncology and early intervention. Future validation with larger cohorts and cfDNA integration could enhance predictive utility.

Keywords: Hepatocellular carcinoma; deep learning; diabetes; chronic liver disease; EASL guidelines; survival prediction

Title: An immune cell signature for sepsis with multiple organ dysfunction syndrome

Author list: Rama Shankar[1], Austin W. Goodyke[3], Shubham Koirala[1], Shreya Paithankar[1], Ruoqiao Chen[2], Aastha Guragain[1], Nicholas L. Hartog[3], Dave Chesla[3], Surender Rajasekaran[1,3,*], and Bin Chen[1,2,4,*]

Detailed Affiliations:
[1]Department of Pediatrics and Human Development, College of Human Medicine, Michigan State University, Grand Rapids, MI 49503, USA. [2]Department of Pharmacology and Toxicology, College of Human Medicine, Michigan State University, Grand Rapids, MI 49503, USA. [3]Office of Research, Corewell Health, Grand Rapids, MI 49503, USA. [4]Department of Computer Science and Engineering, College of Engineering, Michigan State University, Grand Rapids, MI 49503, USA. [*]Corresponding author

Abstract: Sepsis is a leading cause of mortality worldwide, and progression to multiple organ dysfunction syndrome (MODS) represents its most severe clinical manifestation. However, the cellular mechanisms underlying immune dysregulation in sepsis-associated MODS remain incompletely defined. Here, we generated a single-cell RNA–sequencing (scRNA-seq) dataset of immune cells from pediatric sepsis patients at the onset of MODS and matched controls. We identified 22 immune cell populations, with a marked expansion of neutrophils in MODS. Among immune populations, T cells and neutrophils exhibited the most pronounced transcriptional remodeling. Integration with independent adult datasets revealed seven transcriptionally distinct neutrophil subsets, highlighting

neutrophil heterogeneity during critical illness. One subset, Neu_3, characterized by elevated expression of CD74, ITGB7, SPINT2, TMEM9B, and IL13RA1, displayed proliferative and activation signatures but was consistently reduced in pediatric and adult MODS and inversely correlated with Sequential Organ Failure Assessment (SOFA) scores. Neu_3 levels were higher in survivors and restored in recovered patients. Additionally, the Neu_3 gene signature demonstrated robust diagnostic performance in distinguishing sepsis from ICU controls and systemic inflammatory response syndrome (SIRS) patients. These findings identify Neu_3 as a neutrophil subset linked to disease severity and immune recovery, with potential diagnostic relevance in sepsis.

Title: GenePriori: a leakage-aware multi-agent framework for bias-robust gene prioritization with translational oncology relevance

Author list: Yuxuan Jiang[1,2], Xuancheng Zhou[2], Lai Jiang[2], Gang Huang[2], Yuheng Gu[2], Shiyang Yin[2], Lexin Wang[1,4], Vedbar Khadka[5], Hua Yang[5], Shaoqiu Chen[5], Zhuokun Feng[5], Jiayu Xu[6], Ke Xu[1,3,*], Xiaosong Li[1,4,*], Hao Chi[1,5,*], Youping Deng[5,*]

Detailed Affiliations:
[1]Western Institute of Digital-Intelligent Medicine, Chongqing 401329, China, [2]Southwest Medical University, Luzhou 646000, China, [3]Department of Oncology, Chongqing General Hospital, Chongqing University, Chongqing 401147, China, [4]Clinical Molecular Medicine Testing Center, Key Laboratory of Clinical Laboratory Diagnostics (Chinese Ministry of Education), The First Affiliated Hospital of Chongqing Medical University, Chongqing 400016, China, [5]Department of Quantitative Health Sciences, John A. Burns School of Medicine, University of Hawaii at Manoa, Honolulu, Hawaii, USA, [6]School of Computer Science and Informatics, Cardiff University, Cardiff, CF24 4AG, United Kingdom. [*]Corresponding author

Abstract: Background: Reliable gene prioritization is central to translational oncology, where candidate targets and biomarkers often inform downstream studies of drug response and resistance. However, evaluation can be distorted when literature-derived evidence is used alongside target annotations, creating circularity. Methods: We developed GenePriori, a claim-driven multi-agent framework in which four specialized large language model agents propose and refine parameterized hypotheses through structured debate, with a Governor module enforcing statistical acceptance criteria and a Leakage Auditor preserving holdout integrity. We evaluated the framework with dual-holdout 5- fold cross-validation across 19 diseases, including seven solid tumors, and quantified evidence type contributions and disease-level heterogeneity. Results: The consensus strategy achieved the best performance under drug-target holdout, improving mean area under the precision–recall curve (AUPRC) from 0.0127 to 0.0192 (+50.9%), whereas global RWR performed best under genetic-association holdout, improving AUPRC from 0.0684 to 0.0794 (+16.1%). Literature evidence alone produced a +185% AUPRC gain under drug holdout but only +37% under genetics holdout, indicating substantial circularity. Network propagation accounted for more than 60% of total performance gain, and oncology relevant diseases showed meaningful heterogeneity in response. Conclusions: GenePriori provides a leakage-aware and interpretable framework for gene prioritization that can support target and biomarker nomination in translational oncology, while illustrating how rigorous holdout design can reduce overinterpretation of literature-driven signals.

Keywords: gene prioritization; translational oncology; drug resistance; multi-agent debate; network propagation; large language models

Title: Identification of intrinsic pancreatic cancer expression subtypes from patient-derived xenograft samples

Author list: Orry D. Huang[1], Siyuan Zheng[1,2†]

Detailed Affiliations:
[1]Greehey Children's Cancer Research Institute, University of Texas Health Science Center, San Antonio, TX, USA, [2]Department of Population Health Sciences, University of Texas Health Science Center, San Antonio, TX, USA. [†]Corresponding author

Abstract: Expression-based subtyping of pancreatic cancer samples is challenging due to extensive stromal and immune cell infiltration, which confounds tumor-derived gene expression signals. Here we explore the use of patient-derived xenograft (PDX) models to identify tumor-intrinsic expression subtypes of pancreatic cancer. In PDX tumors, human stromal and immune cells are replaced by mouse stromal and immune cells, and mouse-derived sequencing reads can be computationally removed to obtain a nearly pure human tumor transcriptome. Using transcriptomic data from 94 pancreatic cancer PDX samples, we identified three distinct expression subtypes. These

subtypes were subsequently validated in 147 patient samples from The Cancer Genome Atlas (TCGA). In the TCGA cohort, subtype 1 showed significant better survival than the other subtypes. Notably, subtype 2, which had the poorest survival among the subtypes, demonstrated much more frequent progressive disease following primary therapy compared with the other subtypes. The subtypes also exhibited distinct patterns of copy number aberrations. Together, these results demonstrate that PDX models can serve as a valuable resource for identifying tumor-intrinsic expression subtypes in cancers with high levels of non-tumor cell infiltration.

Title: Oncoordinate captures microenvironmental state dynamics and identifies aggressive niches in tumors

Author list: Vignesh V. Venkat[1,2,*] & Subhajyo[4] De[1,*]

Detailed Affiliations:

[1]Department of Pathology & Laboratory Medicine, Rutgers Cancer Ins4tute, New Brunswick, NJ, 08901, USA. Presently: Icahn School of Medicine at Mount Sinai, New York, NY 10029. [*]Corresponding author

Abstract: Cancer is characterized by uncontrolled malignant growth and dysregulated multicellular crosstalk within a complex, heterogeneous tissue ecosystem. However, the specific spatial and functional properties that transform a local tumor niche into an aggressive, invasive environment remain poorly defined. Elucidating the co-evolutionary dynamics within this perturbed microenvironment is essential for understanding how synchronized cellular behaviors drive malignancy. We introduce Oncoordinate, an interpretable, sequential, attention-based deep learning framework that integrates atlas-scale single-cell and spatial transcriptomes to jointly model the malignancy potentials across heterogeneous cell lineages, and project them in their spatial contexts to identify aggressive tumor niches. Applied to lung cancer atlases, Oncoordinate traces the trajectories of epithelial, fibroblast, and immune populations along their malignancy-associated potentials, and recovers coupled epithelial–fibroblast, immune activation, and stromal remodeling programs. Within individual tumor samples, such variations converge within specialized spatial niches characterized by coordinated MYC activation, epithelial–mesenchymal transition, and extracellular matrix remodeling. Oncoordinate thus links multi-compartment malignant progression to its spatial niches, providing a general framework for ecosystem-level biomarker discovery and spatially informed therapeutic targeting in solid tumors.

Title: A Density-Based Spatial Functional Modularity Framework for Characterizing Tumor Functional Niches in HER2-Positive Breast Cancer

Author list: Fan-Jhen Liu[1#], Kuo-Yen Huang[1,2,3#], Yu-Ting Kang[4], Li-Wen Lee[4], Yidong Chen[5,6*], Tzu- Hung Hsiao[4,7,8*]

Detailed Affiliations:

[1]Program for Precision Health and Intelligent Medicine, Graduate School of Advanced Technology, National Taiwan University, Taipei 106319 , Taiwan, [2]National Taiwan University Yong Lin Institute of Health, National Taiwan University, Taipei 106319, Taiwan, [3]Department of Clinical Laboratory Sciences and Medical Biotechnology, College of Medicine, National Taiwan University, Taipei 106319, Taiwan, [4]Department of Medical Research, Taichung Veterans General Hospital, Taichung 407219, Taiwan, [5]Greehey Children's Cancer Research Institute, San Antonio, TX 78229, USA, [6]Department of Population Health Sciences, University of Texas San Antonio, San Antonio, TX 78229, USA, [7]Department of Public Health, Fu Jen Catholic University, New Taipei 24205, Taiwan, [8]Institute of Genomics and Bioinformatics, National Chung Hsing University, Taichung 402202, Taiwan. [#]Co-first author: Yu-Ting Kang and Kuo-Yen Huang. [*]Corresponding author

Abstract: Spatial transcriptomics offers deep insights into cellular functional localization and communication by mapping gene expression to spatial locations. However, systematically characterizing spatially organized functional states remains challenging, as conventional analyses often rely on individual gene expression and fail to capture weak, spatially fragmented, or context-dependent signals. Here, we introduce KDESH (KDE-based Spatial Hotspot Detection), a novel density-based analytical framework that shifts the focus from gene-level variation toward identifying spatially organized functional pathways. By integrating pathway activity scoring, kernel density estimation, and permutation-based statistical testing, this framework provides a robust and statistically grounded approach to model the spatial distribution of biological functions. KDESH enables the detection of regions with locally enriched activity (spatial hotspots) and quantifies co-localization between pathways. Applying this framework to high-resolution (8 μm) HER2-positive breast cancer data, we identified dormancy-like epithelial cells that exhibit significant spatial aggregation. Our analysis revealed that tumor dormancy is not a uniform state but is

spatially structured and functionally heterogeneous, manifesting as hypoxia-associated (HDH) and normoxia-associated (NDH) dormancy hotspots. This framework provides a new perspective on mapping spatially structured functional states and advances our understanding of tumor ecosystem organization.

Workshop – Computational Multi-Omics Approaches to Dissect Regulatory Variation and Disease Mechanisms August 3rd 9:00 AM – 12:00 PM Room: 2213A

Chair: Lei Li

Title: Prediction of Rapid Fibrosis Progression in Metabolic Dysfunction–Associated Steatotic Liver Disease with Interpretable Machine Learning Models

Author list: Tahmina Sultana Priya[1,2,3], Huihuang Yan[3], Konstantinos N. Lazaridis[4], Eric W. Klee[3], Danfeng (Daphne) Yao[1,2], Shulan Tian[3]

Detailed Affiliations:
[1]Department of Computer Science, Virginia Tech, Blacksburg, VA, USA. [2]Sanghani Center for Artificial Intelligence and Data Analytics, Virginia Tech, Blacksburg, VA, USA. [3]Division of Computational Biology, Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, USA. [4]Division of Gastroenterology and Hepatology, Department of Internal Medicine, Mayo Clinic, Rochester, MN, USA.

Abstract

Introduction: Metabolic dysfunction-associated steatotic liver disease (MASLD) is characterized by substantial clinical and biological heterogeneity, leading to highly variable rate of fibrosis progression. Accurately identifying patients at risk of rapid progression remains a critical unmet clinical need. Although existing non-invasive tests can effectively stage current fibrosis, they have limited ability to predict future disease risk and often fail to identify patients who have minimal baseline fibrosis but later experience rapid progression.

Methods: We developed and validated a partitioned machine learning (ML) framework using two independent Mayo Clinic cohorts, the Mayo Clinic Biobank and the Tapestry Study, comprising a total of 13,227 patients with MASLD. The framework integrates longitudinal laboratory measurements, diagnosis code trajectories, and genetic information from electronic health records to predict the risk of rapid fibrosis progression. We first applied latent class analysis with clinical variables to identify clinically meaningful subgroups. Within each subgroup, we performed multiple feature selection and trained supervised ML models, including a stacking ensemble. To enhance clinical interpretability, we derived the partitioned Rapid Fibrosis Progression Score (pRFPS), an interpretable point-based risk score, using SHAP (SHapley Additive exPlanations)-based feature attribution via the AutoScore framework.

Results: Unsupervised phenotypic partitioning identified two reproducible subgroups with distinct metabolic and hepatic characteristics. The cardiometabolic-dominant subgroup was characterized by higher BMI, poorer glycemic control, and elevated triglycerides, whereas the liver-centric subgroup exhibited relatively milder metabolic dysfunction but more pronounced hepatic injury, a stronger genetic predisposition, and higher rates of hepatocellular carcinoma and liver transplantation. Subgroup-specific ML models consistently outperformed non-partitioned population-level models, achieving up to 4% increase in accuracy and a maximum accuracy of 85%. The pRFPS effectively stratified patients into clinically meaningful risk groups. The high-risk individuals had significantly higher rates of hepatocellular carcinoma, liver transplantation, and other adverse liver-related outcomes. Notably, the framework identified 23%–25% of rapid progressors who would have been classified as low risk by conventional clinical tools.

Conclusion: This study demonstrates that, with routinely collected clinical and laboratory data, our partitioned, subtype-aware ML framework could reliably predict the risk of rapid fibrosis progression in MASLD. By explicitly accounting for disease heterogeneity, our approach improves predictive performance and uncovers subtype-specific drivers of progression that are not captured by conventional fibrosis staging systems. The pRFPS provides a clinically actionable tool for risk stratification and early identification of high-risk patients with MASLD. This work offers a generalized framework for precision medicine in MASLD and other heterogeneous chronic diseases.

Keywords: MASLD; Rapid Fibrosis Progression; Machine Learning; Explainable AI; Subgrouping

Title: Atlas of cell type-specific post-transcriptional genetic impacts in immune-related diseases

Author list: Ting Zhang[1#], Hui Chen[1,2#], Shuxin Chen[1#], Ruofan Ding[1], Dinglin Luo[1], Xudong Zou[1], Kewei Xiong[1], Yinuo Wang[1], Xing Li[1], Shibin Hu[3], Jian Yang[4], Gao Wang[5], Stephen Montgomery[6], Qin Li[7*], Lei Li[1*]

Detailed Affiliations:

[1]Institute of Systems and Physical Biology, Shenzhen Bay Laboratory, Shenzhen, 518055, China, [2] Department of Physiological Science, School of Medicine, Universitat de Barcelona (UB), Barcelona, 08907, Spain, [3]Institute of Bio-Architecture and Bio-Interactions (IBABI), Shenzhen Medical Academy of Research and Translation (SMART), Shenzhen, 518107, China, [4]School of Life Sciences, Westlake University, Hangzhou, 310024, China,[5]Department of Neurology, The Gertrude H. Sergievsky Center, Columbia University, New York, NY 10032, USA, [6]Departments of Pathology and Genetics, Stanford School of Medicine, Stanford, CA 94305, USA, [7]Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA. [#]Co-first author, [*]Corresponding author

Abstract: Genetic variants contribute to immune disease risk through complex, cell type-specific regulatory mechanisms, yet post-transcriptional processes including alternative polyadenylation (APA) and RNA editing remain poorly characterized. Using single-cell RNA-sequencing of over 6.06 million peripheral blood mononuclear cells from 2,022 individuals, we mapped 32,763 single-cell quantitative trait loci (sc-xQTLs) across 31 immune cell types and diverse stimulation conditions. Compared to sc-eQTLs, single-cell APA and RNA editing QTLs exhibited even stronger heritability enrichment for immune-related diseases across various immune cell types. Furthermore, our sc-xQTLs colocalize with 62.88% of immune-related disease risk loci, significantly higher than bulk-xQTLs (25.90%). We linked 217 putatively causal genes underlying immune-related disease risk to APA or RNA editing changes, with the majority (70.04%) acting independently of expression. Further integration of protein QTLs with sc-xQTLs and disease GWAS reveals 32 disease causal genes, 19 of which were known druggable targets. By focusing on the DLD enzyme, a critical flavoprotein found in the mitochondria and associated with ulcerative colitis susceptibility, our in vitro experiments confirmed that the minor allele of sc-xQTL rs4518 significantly shortened the 3′UTR length of DLD, leading to decreased protein abundance. Overall, our comprehensive sc-xQTLs atlas transforms our understanding of regulatory mechanisms underlying immune-related disease by revealing the critical and previously underappreciated role of post-transcriptional regulation in immune cell function and disease susceptibility.

Keywords: Single-cell xQTLs, Post-transcriptional regulation

Title: A lifespan regulatory atlas of the East Asian brain resolves the developmental origins of psychiatric traits

Author list: Cong Han[1], Yu Chen[1], RujiaDai[1,2], Qiuman Liang[1], Yongcong Li[1], Minghui Guo[1], Chunyu Liu[1,2*], Chao Chen[1*]

Detailed Affiliations:

[1]Center for Medical Genetics, School of Life Sciences, Central South University, Changsha, China, [2]Department of Psychiatry, SUNY Upstate Medical University, Syracuse, NY, USA. [*]Corresponding author

Abstract:

Background: Large-scale GWAS studies have identified hundreds of risk loci of neuropsychiatric disorders including schizophrenia (SCZ), yet their biological underlying remains unclear. Genetic regulation of gene expression during fetal brain development may contribute to the susceptibility to neuropsychiatric disorders. Previous studies have focused on early- and mid-gestation brain development in European populations, lacking data from late-gestation and East Asian populations. In this study, we aim to construct expression quantitative trait loci (eQTLs) in East Asian under different contexts, including developmental stages and cell types, to elucidate the GWAS signals observed in East Asian populations and their potential roles in the etiology of disorders.

Methods: The study includes 669 brain samples from the EAS population across three developmental stages: 170 fetal samples (Stage 1), 132 childhood and adult samples (Stage 2, age ≤ 65 years), and 367 elderly samples (Stage 3, age > 65 years). Following whole-genome sequencing (10×) and RNA sequencing, we mapped stage-stratified eQTLs and genotype–stage interaction eQTLs. We utilized stratified LD score regression (sLDSC) to partition SCZ heritability across stages and applied cell-type deconvolution to identify the cellular drivers of these heritability shifts. Finally, we performed transcriptome-wide association studies (TWAS) and colocalization to fine-map GWAS signals.

Results: We constructed a lifespan eQTL atlas of the EAS brain, identifying 12,368 eGenes (FDR<0.05). Stage-specific eGenes showed higher enrichment for psychiatric disorder (AD, BD, and SCZ) risk genes compared to shared eGenes (OR > 2, FDR < 0.05). Furthermore, sLDSC revealed a temporal decay in the heritability enrichment of SCZ across the lifespan. Specifically, SCZ heritability showed the highest enrichment (14.9-fold) during the fetal stage and attenuated to 10.9-fold in the elderly stage. Cell-type deconvolution demonstrated that tissue-level heritability attenuation is driven by the depletion of early-developmental cell populations, specifically glial progenitor cells. Genotype-stage interaction analysis identified 1,103 dynamic eQTLs, characterizing the temporal shifts of genetic regulatory effects. By integrating eQTLs with GWAS summary statistics (PPH4 > 0.8), we fine-mapped SCZ GWAS to 66 candidate risk genes. These include 46 stage-unique genes (11 in Stage 1, 6 in Stage 2, and 29 in Stage 3) and 20 genes shared across multiple stages.
Conclusion: We constructed the first lifespan brain eQTL atlas for the East Asian population, addressing the ancestry bias in functional genomics. This study delineates distinct developmental risk windows for psychiatric disorders and provides a resource to prioritize causal disease genes.
Keywords: eQTL atlas, East Asian population

Title: Functional reconciliation of human gut microbes through comparative pangenome analysis
Author list: WeiHua Chen[1,2]
Detailed Affiliations:
[1]Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Key Laboratory of Bioinformatics and Molecular-Imaging, Center for Artificial Intelligence Biology, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China, [2]School of Biological Science, Jining Medical University, Rizhao 272111, China
Abstract: High-throughput metagenomics and culturomics have expanded our knowledge of human gut microbial diversity, yet aligning functional potential with updated taxonomic lineages remains a critical challenge. Misalignment between taxonomy and function complicates mechanistic insights and cross-study comparisons.
Here, we address this issue by analyzing over 240,000 high-quality gut microbial genomes, clustered into more than 4,000 species-level genome bins. Comparison with type strains and existing classifications revealed that over 15% of species underwent taxonomic merging or splitting, highlighting substantial inconsistencies in current lineage assignments.
We performed comparative pangenome analysis to assess functional alignment across these revised species. Strikingly, we observed widespread mismatches between taxonomic reclassifications and functional repertoires: functions previously attributed to single species were often distributed across multiple new lineages, while merged species retained distinct functional profiles. These discrepancies were particularly evident in key microbial metabolites and gene clusters, including short-chain fatty acids, bile acid metabolism, and aromatic amino acid pathways.
Our results demonstrate the necessity of integrated taxonomic and functional refinement. Reconciling genomic lineages with their functional potential clarifies redundancy and specialization in the gut microbiome, improving the reliability of mechanistic studies and informing strain selection for experimental and therapeutic applications.
In conclusion, functional reconciliation via comparative pangenome analysis provides a framework to bridge the gap between taxonomy and function, ensuring that gut microbiome research accurately reflects both evolutionary and metabolic realities.
Keywords: Gut microbiome, Comparative pangenomics, Taxonomic reclassification, Functional potential

Title: Multi-omics integration predicts 17 disease incidences in the UK biobank
Author list: Quan Sun[1]
Detailed Affiliations:
Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA
Abstract
Background: Traditional clinical predictors for disease risks have limitations in capturing underlying disease complexity. Multi-omics technologies, such as metabolomics and proteomics, offer deeper molecular perspectives that could enhance risk prediction, but large-scale studies integrating the two omics are scarce.

Methods: In this study, we investigated 17 incident diseases across 23,776 UK Biobank participants with complete baseline omics data for 159 NMR-based metabolites and 2,923 Olink affinity-based proteins. We considered three omics models (metabolomics-only, proteomics-only and combined-omics models) coupled with three sets of baseline clinical covariates, with the minimal set including only age and sex and the most inclusive set having 39 lab measurements. We adjusted covariates using Cox proportional hazard (CoxPH) models to obtain martingale residuals and employed mixOmics, a multi-omics integration model for prediction. We evaluated performance by adding predicted residuals back to CoxPH models, using Harrell's C-index (C), with both inner (5-fold, stratified for case balance) and outer (repeated with 5 random seeds) cross-validation.
Results: We found that adding omics data (either single-omics or combined-omics) significantly improved risk prediction for all 17 diseases compared to models with clinical predictors alone (mean ΔC=0.063, 0.034, and 0.030 over 3 baseline sets, p-value < 2e-4). Proteomics-only models generally demonstrated superior predictive performance over metabolomics-only models for 14 of the 17 disease endpoints (e.g. for asthma, proteomics consistently yields ΔC=0.060-0.081, p<3e-8). In addition, combined-omics models significantly outperform both single-omics models for 11/17 traits in at least one baseline. However, such advantages diminished with more comprehensive baseline predictors, where proteomics-only models showed superiority for progressively more traits, though ΔC are minimal (0.003-0.007). Glaucoma uniquely demonstrated greater power with metabolomics-only models (p<5e-4). Furthermore, we identified key driver proteins, such as NT-proBNP/BNP for atrial fibrillation (AFib), including potential novel ones (e.g., PRG3 for skin cancer). These driver proteins also showed significant drug target enrichments and associations with socioeconomic factors.
Conclusions: Our results demonstrate that integration of Olink proteomics and Nightingale metabolomics improves risk prediction for a spectrum of common diseases beyond established clinical factors, highlighting the clinical utility of multi-omics for risk prediction, providing molecular insights into disease mechanisms, and potentially guiding future therapeutic development.
Keywords: Multi-omics integration, Disease risk prediction, Proteomics and metabolomics, UK Biobank

Title: Accurate estimation and correction of open chromatin biases in CUT&Tag data
Author list: Sheng'en Shawn Hu[1]
Detailed Affiliations:
[1]Center for Public Health Genomics, University of Virginia School of Medicine, Charlottesville, VA, USA.
Abstract: Precise profiling of epigenomes is essential for better understanding chromatin biology and gene regulation. Cleavage Under Targets & Tagmentation (CUT&Tag) is an easy and low-cost epigenomic profiling technique that can be performed on a low number of cells and at the single-cell level. CUT&Tag has been increasingly used in the community, generating a large number of datasets in various biological systems as a valuable resource. CUT&Tag experiments use the hyperactive transposase Tn5 for DNA tagmentation. The preference of Tn5 transposase toward accessible chromatin can influence the distribution of CUT&Tag sequence reads in the genome and cause open chromatin biases, confounding downstream data analysis. The high sparsity of single-cell CUT&Tag data makes the open chromatin biases more substantial than in bulk CUT&Tag data. Here, we present PATTY (Propensity Analyzer for Tn5 Transposase Yielded bias), a comprehensive computational method to mitigate the open chromatin bias in CUT&Tag data at both bulk and single-cell levels, by utilizing accompanying ATAC-seq data. By integrating existing transcriptomic and epigenomic data using machine learning and integrative modeling, we demonstrate that PATTY results in accurate and robust detection of occupancy sites for both active and repressive histone modifications, including H3K27ac, H3K27me3, and H3K9me3, with experimental validation. We further develop a single-cell CUT&Tag analysis framework using PATTY and show improved cell clustering from corrected single-cell CUT&Tag data compared to using raw data without bias correction. Beyond bulk and single-cell CUT&Tag, PATTY sets a foundation for further development of bias correction methods for improving data analysis for all Tn5 transposase involved high-throughput assays.
Keywords: CUT&Tag, Tn5 transposase bias, Epigenomic profiling, Single-cell analysis

Title: A novel admixed population fine-mapping method: mitigating false discovery rate through local genetic ancestry modeling
Author list: Kai Yuan[1,2,3,4,5], Tian Ge[4,6,7], Hailiang Huang[3,4,5]

Detailed Affiliations:
[1]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN 46202, USA, [2]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN 46202, USA, [3]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA,[4]Stanley Center for Psychiatric Research, the Broad Institute of MIT and Harvard, Cambridge, MA, USA, [5]Department of Medicine, Harvard Medical School, Boston, MA, USA, [6]Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA, [7]Center for Precision Psychiatry, Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA
Abstract: Genome-wide association studies (GWAS) have identified numerous loci associated with complex traits and diseases; however, these loci often contain hundreds of correlated variants, only a subset of which are truly causal. Statistical fine-mapping aims to refine these loci into smaller credible sets of likely causal variants. Existing fine-mapping methods have achieved substantial success through integrating multi-population genetic data and functional annotations but generally assume homogeneous ancestry across the genome within each dataset. Consequently, admixed populations have been largely excluded from fine-mapping analyses, limiting inclusivity and potentially introducing bias.

To address this limitation, we developed a novel fine-mapping framework tailored for admixed populations by incorporating local ancestry information. Our method extends the mathematical framework of SuSiEx, a recently developed multi-ancestry fine-mapping approach that models homogeneous ancestral populations independently. In contrast, our approach integrates cross-ancestry linkage disequilibrium (LD) information within admixed individuals, enabling the modeling of genetic correlations across local ancestral tracts and improving fine-mapping precision. Similar to SuSiEx, the model assumes shared causal variants across ancestries while permitting ancestry-specific effect sizes.

We evaluated the method through simulations of 100,000 admixed individuals mimicking African American demographic history, consisting of 80% African and 20% European ancestry across ten generations of admixture. Performance was assessed under varying heritability levels, local genetic correlations, and numbers of causal variants per locus. Haplotype phasing and local ancestry inference were performed using SHAPEIT5 and FLARE, while ancestry-specific association statistics were estimated using SAIGE-TRACTOR.

Simulation results demonstrated that our method outperformed existing single-population fine-mapping approaches by achieving greater statistical power, lower false positive rates, smaller credible sets, and higher posterior inclusion probabilities (PIPs). As a proof-of-concept, application to Crohn's disease cohorts of African American and Latino individuals identified a credible set containing three highly linked missense variants in CARD6 not reported in previous studies. These variants exhibit substantial ancestry-specific allele frequency differences, and our method effectively accounted for local ancestry effects to capture these signals.

Our work provides a robust framework for fine-mapping in admixed populations and expands opportunities to uncover disease-associated variants and biological mechanisms in historically underrepresented populations.
Keywords: GWAS; Statistical Fine-Mapping; Admixed Populations; Local Ancestry; Linkage Disequilibrium; Crohn's Disease

Title: Associated Cell Types and Influential Genes Identified through Integration of Genome-Wide Association Studies and Single-cell Transcriptomics for Circulating Fatty Acid Traits

Author list: Elijah Sterling[1,2], Saurav Choudhary[3], Kaixiong Ye[1,3]
Detailed Affiliations:
[1]Department of Genetics, University of Georgia, Athens, GA, USA, [2]Regenerative Bioscience Center, University of Georgia, Athens, GA, USA, [3]Institute of Bioinformatics, University of Georgia, Athens, GA, USA.
Abstract: Background: Genome-Wide Association Studies (GWAS) identify single-nucleotide polymorphisms (SNPs) associated with traits. Fatty acid traits are of particular interest because they contribute to diverse biological processes, including membrane composition, cytokine signaling, and liver disease. A challenge in GWAS is that most association signals reside in non-coding regions of the human genome. Recent advances in multi-omic technologies enable the functional annotation of GWAS signals and the discovery of underlying biological mechanisms. Methods: We used a multi-method approach to identify cells, tissues, and genes associated with fatty acid traits. Our lab previously published a GWAS of 19 circulating fatty acid (cFA) traits in more than 200,000 European-ancestry individuals from the UK Biobank. In this study, we integrate the GWAS summary statistics with

single-cell transcriptomic data from a human liver cell atlas spanning postnatal to older adult stages (n > 100,000 cells) and spatial transcriptomics data from a human embryo dataset representing Carnegie stage 8 (n > 35,000 spots). We applied several bioinformatic tools to identify associated cell types (scDRS v1.0.3, seismicGWAS v1.0.0), enriched tissues (gsMap v1.73.5), and influential genes (seismicGWAS v1.0.0). Results: We found that centrilobular and periportal hepatocytes were significantly associated with 19 cFA traits across multiple cell-type association methods (FDR < 0.05). We also identified significant enrichment in the endoderm for cFA traits ($p < 0.05$). In addition, we identified influential genes contributing to the signal in centrilobular hepatocytes (FDR < 0.05) with three genes being identified as influential across all 19 cFA traits (MLXIPL, APOB, and SLC22A1). KEGG enrichment analysis further highlighted significant pathways including complement and coagulation cascades, bile secretion, and lipid and atherosclerosis ($p < 0.05$). Conclusions: In summary, our findings suggest that two specific hepatocyte subpopulations are associated with cFA traits. We further identified the endoderm as an enriched developmental tissue, supporting a potential role for early lineage specification in fatty acid biology. In addition, we elicited three influential genes relevant for liver function that are present across all CFA traits. These results also implicate specific genes and pathways in mediating the effect of observed associated variants on the fatty acid traits.

Keywords: Fatty acids, Hepatocytes, Genome-Wide Association Studies (GWAS)

Title: Single-cell multiomics reveals malignant plasticity and immune suppression in high-risk pediatric cancers

Author list: Wenbao Yu[1,2]

Detailed Affiliations:

[1]Department of Cancer and Cellular Biology, Lewis Katz School of Medicine, Temple University, [2]Fox Chase Cancer Center

Abstract: High-risk pediatric cancers remain a major clinical challenge due to their remarkable capacity for therapy resistance and relapse. Deep characterization of the tumor heterogeneity and the tumor microenvironment is a critical step toward addressing this challenge. Recent advances in single-cell multiomics have enabled unprecedented resolution for this purpose. In this talk, I will present findings from single-cell multiomic studies of two devastating pediatric cancers: KMT2A-rearranged (KMT2A-r) infant acute lymphoblastic leukemia (ALL) and high-risk neuroblastoma.

First, in KMT2A-rearranged infant acute lymphoblastic leukemia, we discovered that leukemic cells from infants younger than six months harbor markedly increased lineage plasticity, with steroid response pathways selectively downregulated in the most immature blast populations. Critically, we identified a hematopoietic stem and progenitor-like (HSPC-like) population in younger patients that forms an immunosuppressive signaling circuit with cytotoxic lymphocytes, offering a mechanistic explanation for chemotherapy and immune evasion in this age group. Second, in high-risk pediatric neuroblastoma, we resolved six distinct malignant cell states spanning from conventional adrenergic to mesenchymal states, correlating with distinct prognoses. Chemotherapy drove a significant expansion of mesenchymal malignant cells and reshaping of the tumor immune microenvironment, most strikingly an increase in pro-tumorigenic, immunosuppressive, and pro-angiogenic tumor-associated macrophages. Using robust cell-cell interaction analysis validated by spatial proteomics (CODEX), an immunocompetent murine model and in vitro cell co-culture, we identified the HBEGF–ERBB4 paracrine axis between macrophages and malignant subsets as a key driver of tumor growth through ERK signaling activation.

Together, these studies demonstrate how single-cell multiomics can decode the molecular architecture of treatment resistance and immune evasion in pediatric cancer and point toward new therapeutic vulnerabilities at the intersection of tumor cell plasticity and the immune microenvironment.

Keywords: Pediatric cancer, Single-cell multiomics, Therapy resistance, Tumor microenvironment

Workshop – Bridging Methodological Innovation and Real-World Applications in Genetic Studies August 3rd
9:00 AM – 12:00 PM
Room: 2213B

Chair: Zikun Yang

Title: fastGxE: Powering genome-wide detection of genotype-environment interactions in biobank studies

Author list: Xiang Zhou[1]

Detailed Affiliations:
[1]Department of Statistics and Data Science, Yale University, New Haven, CT , USA.

Abstract: Traditional genome-wide association studies (GWAS) have primarily focused on detecting main genotype effects, often overlooking genotype-environment interactions (GxE), which are essential for understanding context-specific genetic effects and refining disease etiology. Here, we present fastGxE, a scalable and effective genome-wide GxE method designed to identify genetic variants that interact with environmental factors to influence traits of interest. fastGxE controls for polygenic effects, polygenic interaction effects, and noise heterogeneity; it remains robust to the number of environmental factors involved in GxE interactions; and it ensures scalability for genome-wide GxE analysis in large biobank studies, achieving speed improvements of 32.98-126.49 times over existing approaches. We illustrate the benefits of fastGxE through extensive simulations and an in-depth analysis of 32 physical traits and 67 blood biomarkers from the UK Biobank. In real data applications, fastGxE identifies seven genomic loci associated with physical traits, including four novel ones, and 18 genomic loci associated with blood biomarkers, 12 of which are novel. The new discoveries highlight the dynamic interplay between genetics and the environment, uncovering potentially clinically significant pathways that could inform personalized interventions and treatment strategies.

Keywords: GxE, GWAS, UKB

Title: Distributional genetic effects reveal context-dependent molecular regulation in human brain aging and Alzheimer's disease

Author list: Anjing Liu[1], Roulan Jiang[2], Ruixi Li[1], Xuewei Cao[1,3], Zining Qi[1,4], Ru Feng[1,3], Hao Sun[1,5], Masashi Fujita[6,16], Natacha Comandante-Lou[6,16], Chirag M Lakhani[7], Jenny Empawi[8], David A. Knowles[9,10,11], Xiaoling Zhang[8,12], Kushal K. Dey[3,12,13], Philip De Jager[6,15,16], David Bennett[17], The Alzheimer's Disease Functional Genomics Consortium, **Tianying Wang**[18,*], Gao Wang[1,16,19,*]

Detailed Affiliations:
[1]Center for Statistical Genetics, The Gertrude H. Sergievsky Center, Columbia University, New York, NY, USA, [2]Department of Statistics and Data Science, Tsinghua University, Beijing, China, [3]Computational and System Biology, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, NY, USA, [4]Division of Genetic Medicine, Department of Medicine, The University of Chicago, Chicago, IL, USA, [5]Graduate School of Biomedical Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA, [6]Center for Translational & Computational Neuroimmunology, Columbia University, New York, NY, USA,[7]New York Genome Center, New York, NY, USA, [8]Department of Biomedical Genetics, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA,[9]Department of Computer Science, Columbia University, New York, NY, USA, [10]Department of Systems Biology, Columbia University, New York, NY, USA, [11]Data Science Institute, Columbia University, New York, NY, USA, [12]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA, [13]Physiology, Biophysics and System Biology, Weill Cornell Medicine, New York, NY, USA, [14]Gerstner Sloan Kettering Graduate School of Biomedical Sciences, New York, NY, USA, [15]Taub Institute for Research on Alzheimer's Disease and the Aging Brain, Columbia University, New York, NY, USA [16]Department of Neurology, Columbia University, New York, NY, USA, [17]Rush Alzheimer's Disease Center and Department of Neurological Sciences, Rush University Medical Center, Chicago, IL, USA, [18]Department of Statistics, Colorado State University, Fort Collins, CO, USA, [19]Department of Biostatistics, Columbia University, New York, NY, USA. [*]Corresponding author

Abstract: Molecular QTL studies quantify whether genetic variants affect molecular traits, but non-linear effects including distributional patterns, variance, and interactions provide mechanistic insights beyond mean-level

associations. Methods for detecting distributional effects have been developed for eQTL analysis, yet applications have focused on method demonstrations rather than large-scale biological discovery. We comprehensively mapped quantile, variance, and interaction QTLs across 34 data-set from 22 molecular contexts in >2,300 human brain donors, revealing that 48.7% of quantile QTLs (qQTLs) exhibit context-dependent regulation invisible to linear models, with enrichment at phenotypic extremes and in cell-type-specific regulatory elements, chromatin accessibility regions, and long-range chromosomal contacts. qQTL variants explained additional trait heritability beyond linear QTLs for brain-related traits. At Alzheimer's disease (AD) risk loci, qQTL analysis revealed complex regulatory architecture including variance effects at PITRM1, lower-quantile-specific effects at TMEM106B partially explained by APOE ε4 interactions and coordinated epigenetic regulation at loci harboring CHRNE/SCIMP/RABEP1. Quantile-based transcriptome-wide association studies identified 34 AD risk genes and additional aging-related genes beyond standard TWAS, with enrichment in immune regulation and telomere maintenance pathways where distributional effects may reflect threshold-dependent mechanisms. Our non-linear QTL atlas and qTWAS resource enable characterization of context-dependent regulatory effects in complex disease genetics.

Keywords: Non-linear QTLs, Quantile QTL analysis, Alzheimer's disease genetics, Context-dependent gene regulation

Title: Powerful statistical frameworks for genome-wide survival association analysis

Author list: Shiyang Ma[1], Tao Zhang[2], Lin Hu[2], Jingtong Wang[3]

Detailed Affiliations:

[1]Shanghai Jiao Tong University, China, [2]Department of Epidemiology and Health Statistics, West China School of Public Health and West China, China, [3]Department of Biostatistics, University of North Carolina at Chapel Hill, NC.

Abstract: In genome-wide survival association analysis, time-to-event traits offer the advantage of capturing both diagnosis status and timing, facilitating the identification of genetic variants associated with age of onset, disease progression, and lifespan. However, existing methods primarily focus on quantitative and binary traits, which do not leverage censored time information and have limitations in detecting rare variants associated with disease progression. To address these challenges, we propose a set of powerful statistical frameworks for genome-wide survival analysis, including SurvSTAAR and Cox-MK. SurvSTAAR provides a scalable pipeline for rare variant analysis of TTE traits, accounting for sample relatedness, population structure, and functional annotations. Cox-MK integrates model-X knockoff inference with saddlepoint approximation to achieve SNP-level false discovery rate (FDR) control. Extensive simulations and applications to UK Biobank data demonstrate that our methods improve statistical power while maintaining well-calibrated FDR control compared to existing approaches. Notably, our frameworks identify novel associations for Alzheimer's disease, as well as additional candidate genes for asthma and ischemic heart disease.

Keywords: model-X knockoff, genome-wide survival association analysis, linkage disequilibrium, false discovery rate

Title: Interplay of Polygenic and Monogenic Kidney Disease

Author list: Atlas Khan[1]

Detailed Affiliations:

[1]Division of Nephrology, Department of Medicine, Columbia University Irving Medical Center, New York, NY, USA.

Abstract: Chronic kidney disease (CKD) arises from a complex interplay of monogenic, polygenic, and environmental risk factors. Although monogenic kidney disorders such as autosomal dominant polycystic kidney disease (ADPKD), COL4A-associated nephropathy (COL4A-AN), and UMOD-related tubulointerstitial kidney disease exhibit variable penetrance and expressivity, the contribution of polygenic background to this heterogeneity remains incompletely understood.

We integrated SNP array, exome/genome sequencing, and electronic health record data from the UK Biobank, All of Us, and MyCode cohorts to investigate how genome-wide polygenic scores (GPS) modify CKD risk in monogenic kidney disease. Across ADPKD, COL4A-AN, and UMOD T62P variant carriers, GPS robustly

stratified CKD risk. Among ADPKD carriers, individuals in the highest GPS tertile exhibited up to a 54-fold increased risk of CKD, whereas those in the lowest tertile had only modestly elevated risk. Similarly, COL4A-AN carriers in the highest GPS tertile showed significantly increased CKD risk, while those in the lowest tertile had risk comparable to the general population. In UMOD T62P carriers, GPS strongly modified age-dependent CKD risk, with a significant gene–polygenic interaction.

Together, these findings establish polygenic background as a major modifier of penetrance and disease severity across monogenic kidney disorders and demonstrate that integrating polygenic risk with rare variant status substantially improves CKD risk stratification, with important implications for precision nephrology.

Keywords: Chronic kidney disease; polygenic risk score; monogenic kidney disease; genetic modifiers; precision nephrology; UMOD; ADPKD

Title: Integrating genomic and proteomic data improves complex trait prediction in diverse populations.

Author list: Haoyu Zhang[1]

Detailed Affiliations:

[1]National Cancer Institute.

Abstract:

Introduction

Polygenic risk score (PRS) is widely used to quantify the genetic risk for complex traits. Meanwhile, the growing availability of large-scale proteomics data has enabled new opportunities to include proteomic biomarkers for disease and trait prediction. However, the development and evaluation of proteomic risk scores (ProRS), particularly in combination with PRS, remains limited. In particular, the predictive utility and transferability of these integrated scores across diverse populations and a range of complex traits have not been comprehensively assessed.

Methods

We developed an ensemble modeling framework to assess total and mediated prediction effects from genomic and proteomic data. After imputing missing proteomic data and adjusting for batch effects, ProRS models were constructed using super learning, an ensemble of statistical and machine learning methods. ProRS models were developed using a training-validation-testing random split of 36,903 unrelated UK Biobank participants with plasma proteomic data across four ancestry groups: European (EUR), African (AFR), Admixed American (AMR) and South Asian (SAS). For PRS, we used cutting-edge PRSs from the PGS catalog excluding UK Biobank participants. The combined PRS+ProRS model was evaluated across five binary and six continuous traits using the same testing dataset. We assessed prediction performance and the extent to which ProRS mediated PRS effects. For binary traits, we also compared prediction accuracy based on whether biomarkers were collected before or after disease diagnosis.

Results

The combined PRS+ProRS significantly improved across traits and ancestries. On average, PRS alone yielded R2 of 3.60% for continuous traits and AUC of 60.7% for binary traits. With PRS+ProRS, average R2 increased to 45.3%, and AUC rose to 73.5%. Relative to PRS-only models, R2 improved by an average of 4233%, and AUC by 22.7%. Notable gain included prostate cancer in SAS (AUC: 47.4% -> 79.6%), asthma in EUR (AUC: 67.7% -> 80.5%), type 2 diabetes in EUR (AUC: 63.3% -> 84.4%) and breast cancer in non-EUR (AUC: 67.0% -> 75.0%). Among cases diagnosed within 10 years of biomarker collection, prediction was stronger when proteins were measured after rather than before diagnosis (AUC 83.2% vs. 71.8%). Mediation analysis showed that 1.05%-23.3% of the PRS effect was mediated through ProRS for binary traits, and 26.6%-82.1% for continuous traits in EUR.

Summary

Our modeling framework provides a comprehensive approach integrating genomic and proteomic data in complex trait prediction, offering accurate and interpretable results across diverse populations.

Keywords: Polygenic risk scores, Proteomic risk scores, Multi-omics prediction, Cross-ancestry transferability

Title: Multi-ancestry modeling leveraging genetic diversity reveals heterogeneous architecture in primary open-angle glaucoma

Author list: Tyler G. Kinzy[1], Osahon J. Asowata[1], Lauren A. Cruz[2], Timothy Ciesielski[2], Ayobami Alimi[1], Puja Mehta[13], Briana McIntosh[1], Andrea R. Waksmunski[2], Inas Aboobakar[1], Karen Mruk[1], Jacquelaine Bartlett[2], Bryan R. Gorman[3], Yuki Bradford[4], Rebecca Salowe[4], Shefali S. Verma[4], Joan O'Brien[4], Linda Zangwill[5], Radha Ayyagari[5], Robert N. Weinreb[5], Michael Hauser[6], Chen Jiang[7], Hélène Choquet[7], Kelsey Stuart[4], Sjoerd Driessen[9], Pieter Bonnemaijer[9], Pirro Hysi[10], Christopher Hammond[10], Christopher Halladay[11], Cari Nealon[12], Wen-Chih Wu[11], Anthony P. Khawaja[8], Lou R. Pasquale[13], Janey L. Wiggs[13], Dana C. Crawford[2], Ayellet V. Segrè[13], Yutao Liu[14], Neal S. Peachey[12], Sudha K. Iyengar[12], Million Veteran Program[15], Scott M. Williams[2], Jessica N. Cooke Bailey[1]

Detailed Affiliations:
[1]East Carolina University, Greenville, NC, USA, [2]Case Western Reserve University, Cleveland, OH, USA,[3]VA Boston Healthcare System, Boston, MA, USA, [4]University of Pennsylvania, Philadelphia, PA, USA, [5]University of California, San Diego, La Jolla, CA, USA, [6]Duke University, Durham, NC, USA, [7]Kaiser Permanente Northern California, Oakland, CA, USA, [8]University College London, London, UK, [9]Erasmus Medical School, Rotterdam, the Netherlands, [10]King's College London, London, UK, [11]Providence VA Medical Center, Providence, RI, USA, [12]VA Northeast Ohio Healthcare System, Cleveland, OH, USA, [13]Harvard University, Boston, MA, USA, [14]Augusta University, Augusta, GA, USA, [15]Million Veteran Program, U.S. Department of Veterans Affairs

Abstract: Genetic diversity encodes structured variation in allele frequency, linkage disequilibrium (LD), and demographic history that can be leveraged to improve statistical inference in genome-wide association studies (GWAS). When appropriately modeled, cross-ancestry variation enhances locus discovery, increases fine-mapping resolution, and enables formal quantification of effect heterogeneity across populations. We applied this framework to primary open-angle glaucoma (POAG), a highly heritable complex disease, using the Million Veteran Program (MVP), an ancestrally diverse biobank linked to electronic health records. We conducted ancestry-stratified GWAS using scalable mixed-model association testing (SAIGE), followed by cross-ancestry meta-regression to accommodate allelic effect heterogeneity. LD-aware fine-mapping was performed to refine credible sets across diverse LD structures. To interrogate ancestry-specific signal architecture at a local level, we applied tract-level ancestry decomposition (TRACTOR), enabling partitioning of association effects by local haplotype background. Independent African-ancestry cohorts were incorporated through inverse-variance meta-analysis to strengthen inference.

Across analyses, we identified both shared and population-enriched association signals, with several loci demonstrating statistically significant effect size heterogeneity across ancestry groups. SNP-based heritability estimates were similar across ancestry groups, while cross-population genetic correlation was high but not complete, consistent with partially shared genetic effects. Critically, cross-ancestry modeling improved credible set resolution relative to single-ancestry analyses, illustrating how differences in LD structure sharpen localization of candidate causal variants. Local ancestry modeling suggested that observed heterogeneity at selected loci reflects ancestry-specific haplotype structure rather than global allele frequency differences alone.

These findings highlight the value of incorporating diverse populations in genetic studies. Cross-ancestry analysis improves discovery, reveals population-specific effects, and strengthens biological interpretation of complex disease risk.

Keywords: Cross-ancestry GWAS, Primary open-angle glaucoma, Fine-mapping, Local ancestry analysis

Title: Signatures of recent positive selection refine novel Primary open-angle glaucoma risk loci

Author list: Timothy Ciesielski[1], Himiede Sesay[1], Tyler G. Kinzy[2], Osahon J. Asowata[2], Lauren A. Cruz[1], Ayobami Alimi[2], Puja Mehta[13], Briana McIntosh[2], Andrea R. Waksmunski[1], Inas Aboobakar[2], Karen Mruk[2], Jacquelaine Bartlett[1], Bryan R. Gorman[3], Yuki Bradford[4], Rebecca Salowe[4], Shefali S. Verma[4], Joan O'Brien[4], Linda Zangwill[5], Radha Ayyagari[5], Robert N. Weinreb[5], Michael Hauser[6], Chen Jiang[7], Hélène Choquet[7], Kelsey Stuart[4], Sjoerd Driessen[9], Pieter Bonnemaijer[9], Pirro Hysi[10], Christopher Hammond[10], Christopher Halladay[11], Cari Nealon[12], Wen-Chih Wu[11], Anthony P. Khawaja[8], Lou R. Pasquale[13], Janey L. Wiggs[13], Dana C. Crawford[1], Ayellet V. Segrè[13], Yutao Liu[14], Neal S. Peachey[12], Sudha K. Iyengar[12], Million Veteran Program[15], Jessica N. Cooke Bailey[2], Scott M. Williams[1]

Detailed Affiliations:
[1]Case Western Reserve University, Cleveland, OH, USA, [2]East Carolina University, Greenville, NC, USA,[3]VA Boston Healthcare System, Boston, MA, USA, [4]University of Pennsylvania, Philadelphia, PA, USA, [5]University of California, San Diego, La Jolla, CA, USA, [6]Duke University, Durham, NC, USA, [7]Kaiser Permanente Northern California, Oakland, CA, USA, [8]University College London, London, UK, [9]Erasmus Medical School, Rotterdam, the Netherlands, [10]King's College London, London, UK, [11]Providence VA Medical Center, Providence, RI, USA [12]VA Northeast Ohio Healthcare System, Cleveland, OH, USA, [13]Harvard University, Boston, MA, USA, [14]Augusta University, Augusta, GA, USA, [15]Million Veteran Program, U.S. Department of Veterans Affairs

Abstract: Natural selection leaves measurable signatures in the human genome when genetic variants influence reproductive fitness. Because selection operates on phenotypic consequences, genomic regions demonstrating recent positive selection are unlikely to be functionally neutral. This evolutionary principle may be particularly informative for late-onset diseases, where associated variants are unlikely to have been selected for the disease itself but may persist due to pleiotropic effects on traits expressed earlier in life.

Primary open-angle glaucoma (POAG), a highly heritable optic neuropathy that typically manifests after peak reproductive age, provides a compelling model to evaluate this framework. Although genome-wide association studies (GWAS) have identified numerous POAG loci, distinguishing causal variants from correlated markers remains a central challenge. We hypothesized that signatures of recent positive selection could provide additional functional evidence to refine association peaks under models of pleiotropy and antagonistic selection.

We evaluated newly identified POAG-associated loci from multi-ancestry GWAS meta-analyses for evidence of recent positive selection using haplotype-based metrics. Whole-genome sequence data from 2,504 individuals across 26 global populations in the 1000 Genomes Project were analyzed using the nSL statistic, which detects extended haplotype homozygosity while accounting for local recombination rate heterogeneity. For each novel locus, we assessed enrichment of extreme |nSL| scores within pre-specified genomic windows and quantified the relationship between GWAS association strength and selection intensity using generalized estimating equations to account for linkage disequilibrium structure.

Two loci demonstrated corroborated evidence of recent positive selection in at least one population. At a chromosome 9 locus identified in African-ancestry meta-analysis, selection signals in multiple European populations overlapped the primary association peak and extended across an LD block encompassing the terminal exon of MLLT3. At a chromosome 15 locus identified in cross-ancestry meta-analysis, selection signals were observed in East Asian populations and localized within an LD block spanning early exons of CHSY1. Within both regions, variants in strongest linkage disequilibrium with the lead GWAS SNP exhibited among the largest |nSL| scores, and selection intensity was significantly associated with GWAS signal strength.

These findings support a model in which variants contributing to late-onset disease risk may reside within haplotypes shaped by recent positive selection acting on pleiotropic traits. Integrating evolutionary signatures with GWAS results provides a biologically grounded strategy to prioritize functional variants and refine causal inference at complex disease loci.

Keywords: Positive selection, Primary open-angle glaucoma, Multi-ancestry GWAS, Evolutionary genomics

Title: Tractor-PGS uses local ancestry to improve prediction accuracy in admixed populations

Author list: Yi-Sian (Helen) Lin[1*], Nirav N Shah[1], Linfeng Hu[2,3,4], Wei Zhou[2,3,4], Elizabeth G Atkinson[1,5]
Detailed Affiliations:
[1]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA, [2]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA, [3]Center for Genomic Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA,[4]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA,[5]Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, Houston, TX, USA

Abstract: Polygenic scores (PGS) aggregate weighted genomic dosage into a score that estimates an individual's genetic liability to a trait or disease. However, previous large-scale genomic studies have been predominantly conducted in European populations, leading to poor transferability in prediction accuracy when the target cohort is non-European, which can exacerbate health disparities. Prediction in individuals with admixed genomes presents further challenges, as their chromosomes consist of components from multiple ancestries and ancestry patterns vary

across the genome. Mainstream PGS methodologies do not account for such genetic admixture, and as such admixed populations are often excluded from analyses or experience reduced PGS accuracy.

Here, we present the Tractor-PGS, a novel method designed specifically for prediction in admixed individuals. Tractor-PGS integrates the Tractor framework, which leverages local ancestry information from both the discovery and target cohorts to provide ancestry-specific effect size estimates and dosages for polygenic prediction. In this way, variant weights are tuned in a highly personalized manner to the local ancestry context of each participant, improving their fit and increasing predictive performance. We tested our framework in extensive simulations of 2-way African/European individuals, varying parameters of trait polygenicity, causal variant effect size and distribution across ancestries, and minor allele frequencies. We find that Tractor-PGS consistently outperformed the basic clumping and thresholding method, especially when effect size heterogeneity was large between component ancestries and when causal variants were more common in the ancestry with a small global proportion. When re-weighting the estimated effect sizes by the average local ancestry proportion of the target cohort, Tractor-PGS showed superior performance even when causal variants were common in the ancestry with a large global proportion. Compared with GAUDI, another local ancestry-informed PGS method, Tractor-PGS exhibited better prediction accuracy in almost all settings. Current efforts are underway to implement Tractor-PGS on admixed empirical data from the All of Us Research Workbench and UK Biobank, benchmarking against both ancestry-informed and traditional methods.

In summary, this study introduces a new method that utilizes local ancestry information for PGS in admixed populations via Tractor framework, enabling fast computation and better prediction accuracy. With ancestry-informed summary statistics to be released publicly, this method is expected to improve polygenic prediction for underrepresented admixed individuals and advance precision health.

Keywords: Tractor-PGS, Admixed populations, Local ancestry, Polygenic score transferability

Workshop – AI and Multi-omics for Modeling Aging and Disease
August 3rd
9:00 AM – 12:00 PM
Room: 1220

Chair: Sheng Li

Title: Transcriptional heterogeneity predicts and shapes clonal selection in aging hematopoiesis

Author list: Sheng Li[1,2,3]

Detailed Affiliations:

[1]Department of Cancer Biology, Keck School of Medicine, University of Southern California, Los Angeles, CA, [2]Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA,[3]Department of Quantitative and Computational Biology, Dornsife College of Letters, Arts and Sciences, University of Southern California, Los Angeles, CA.

Abstract: Clonal expansion of hematopoietic stem cells (HSCs) drives clonal hematopoiesis and elevates myeloid malignancy risk, yet why some clones dominate while similar neighbors do not remain unresolved and is poorly predicted by mean gene expression. Identifying expansion-prone clones from their early molecular state would enable pre-clinical risk stratification. We hypothesized that clonal fate is encoded in cell-to-cell variability among sister cells rather than their mean state, and tested this in HSCs, where heritable lineage barcoding links early molecular state to long-term functional output.

We coupled lineage tracing with single-cell RNA-sequencing across heterochronic and homochronic transplantation, generating matched early-state and long-term-output measurements for thousands of clones. We developed scCloneVar, a computational framework that estimates mean-adjusted gene-expression variance at clonal resolution, identifies differentially variable genes (DVG), and trains a predictive model that uses these variance features to forecast long-term clonal self-renewal from early single-cell transcriptomes.

Applied to aging HSCs, scCloneVar revealed expanded transcriptional heterogeneity concentrated in stem-cell regulatory programs. Clonal architecture remained preserved after transplantation, indicating that aging reshapes clonal behavior through selection rather than uniform expansion. Heterochronic transplantation showed that

identical sister clones exhibit age-dependent shifts in self-renewal, preferentially expanding in age-matched environments.

The scCloneVar predictive model forecast long-term clonal self-renewal from early transcriptional state, with DVG features carrying predictive power beyond mean expression. Variability-associated programs were conserved between mouse and human HSCs, established by middle age in humans, and enriched for pathways linked to clonal hematopoiesis and myeloid malignancy risk, defining a candidate signature for early identification of expansion-prone clones.

By showing that clonal fate is partially decodable from early single-cell heterogeneity, this work establishes variance-based feature engineering as a predictive modality in single-cell omics, with direct relevance to clonal hematopoiesis surveillance, pre-malignant risk stratification, and predictive modeling of cellular selection in aging and disease.

Keywords: Clonal hematopoiesis, Single-cell gene expression variability, Hematopoietic stem cells, Clonal fate prediction

Title: Liquid biopsy biomarkers for cancer detection and disease monitoring

Author list: Bodour Salhia

Detailed Affiliations:

Chair (Interim), Department of Cancer Biology, Royce and Mary Trotter Chair in Cancer Research, Co-Leader Epigenetics Regulation in Cancer, Norris Comprehensive Cancer Center, Director Preclinical Models Shared Resource, Norris Comprehensive Cancer Center, Keck School of Medicine of USC, Senior Member of the National Academy of Inventors

Abstract: Advances in liquid biopsy technologies are enabling new opportunities for cancer detection, risk assessment, and disease monitoring through the integration of multiomic biomarkers and machine learning approaches. This presentation highlights the development of two plasma-based methylation assays, OvaPrint™ and MammaTrace™, designed to address critical unmet needs in ovarian and breast cancer management. OvaPrint™ is a cfDNA methylation-based assay developed for molecular risk assessment in women presenting with adnexal masses. Current diagnostic approaches lack sufficient specificity to reliably distinguish benign from malignant masses prior to surgery, contributing to delays in diagnosis and suboptimal referral patterns. By combining genome scale methylation profiling with machine learning derived classifiers, OvaPrint demonstrates high specificity and strong positive predictive value. MammaTrace™ is a tumor naïve methylation based minimal residual disease assay designed for longitudinal surveillance in breast cancer. Using plasma derived methylation signatures and machine learning models, MammaTrace aims to detect minimal residual disease providing significant lead times prior to recurrence. Emerging multiomic strategies integrating methylation, proteomics, and spatial transcriptomics will also be discussed as approaches to better understand tumor biology. Together, these studies demonstrate how artificial intelligence enabled biomarker discovery may advance precision oncology across the continuum of cancer care, from risk stratification and early detection to recurrence monitoring.

Keywords: Liquid biopsy, cfDNA methylation, Machine learning, Precision oncology

Title: A transcriptomics-native foundation model for universal cell representation and virtual cell synthesis

Author list: Xiaohui Jiang[1], Jichun Xie[1,2,3]

Detailed Affiliations:

[1]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA,[2]Department of Mathematics, Duke University, Durham, NC, USA,[3]Department of Computer Science, Duke University, Durham, NC, USA.

Abstract: Current single-cell foundation models rely on language-model architectures that ignore transcriptomic data distributions, often underperforming specialized methods. We introduce xVERSE, a transcriptomics-native foundation model coupling batch-invariant representation learning with the probabilistic generation of expression profiles. xVERSE outperforms the leading foundation and batch-effect correction methods in representation learning by 17.9% and 11.4%, respectively, successfully preserving biological heterogeneity while diminishing batch effects. Furthermore, xVERSE surpasses the second-best spatial imputation method by 34.3% and uniquely synthesizes virtual cells indistinguishable from biological data (AUROC ≈ 0.5). As a powerful data-augmentation engine, xVERSE utilizes these high-fidelity virtual cells to enable accurate clustering and marker detection in tiny

datasets—resolving rare cell types with as few as four cells—while improving the generalizability of cross-modality predictions across diverse pathological states. These results establish xVERSE as a transformative framework unlocking analytical capabilities beyond conventional models.

Keywords: Single-cell foundation model, Transcriptomics-native learning, Batch-effect correction, Virtual cell generation

Title: RNA stability as a key link between genetic variation and disease: insights from multi-omics and deep learning

Author list: Xinshu Grace Xiao[1]

Detailed Affiliations:

[1]University of California Los Angeles.

Abstract: Messenger RNA stability is a critical determinant of gene expression, yet its contribution to complex traits and disease remains largely underexplored relative to transcriptional regulation. Genetic variants residing within mRNA sequences have the potential to alter stability-modulating elements, thereby influencing gene abundance through mechanisms that conventional expression-focused analyses may overlook. Here, I will present our computational and experimental frameworks for systematically identifying genetic variants that regulate mRNA stability and for connecting these variants to disease phenotypes. Using metabolic labeling data across multiple human cell lines, we developed RNAtracker, a workflow that distinguishes allele-specific RNA stability events from transcriptional regulation, uncovering thousands of stability-regulating variants enriched in functional regulatory regions and immune-related disease pathways. Complementing this, our massively parallel screening approach, MapUTR, enables high-throughput functional characterization of 3′ UTR variants affecting post-transcriptional mRNA abundance, including rare variants and somatic mutations in cancer driver genes. Building on these foundations, I will discuss our ongoing efforts to apply deep learning models that integrate sequence features and omics data to predict RNA stability and identify stability-mediated mechanisms underlying disease. Together, these approaches highlight RNA stability as an underappreciated yet important axis linking genetic variation to disease and demonstrate how AI-driven strategies can advance our understanding of this regulatory layer.

Keywords: mRNA stability, Genetic regulatory variants, Post-transcriptional regulation, Deep learning genomics

Title: Precisely integrate single-cell data to decipher aging signatures in the monkey brain under calorie restriction

Author list: Christina Dimovasili[1], Jou-Hsuan Roxie Lee[2], Simon Liang Lu[2], Ana T Vitantonio[1], Dong Xu[3], Douglas L Rosene[1], JuexinWang[4], Chao Zhang[2]

Detailed Affiliations:

[1]Department of Anatomy & Neurobiology, Boston University Chobanian and Avedisian School of Medicine, Boston, MA, USA, [2]Department of Medicine, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA, [3]Department of Electrical Engineering and Computer Science, and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO, USA, [4]Department of Biomedical Engineering and Informatics, Luddy School of Informatics, Computing, and Engineering, Indiana University Indianapolis, IN, USA.

Abstract: During brain aging, terminally differentiated neuroglia exhibits metabolic dysfunction and increased oxidative damage, compromising their functional capacity. These cellular and molecular changes impair the maintenance of myelin sheath integrity, contributing to age-related white matter degeneration. Calorie restriction (CR) is a well-established intervention that slows biological aging and may mitigate metabolic alterations, thereby preserving glial function.

Here, we generate a unique single-nucleus transcriptomic dataset of the aging rhesus monkey brain under lifelong 30% calorie restriction. However, integrating single-cell transcriptomic datasets across laboratories, conditions, and technologies remains challenging due to cell-type-specific batch effects and heterogeneous biological composition. Existing methods often fail to disentangle true biological variation from technical noise, leading to misalignment of distinct cell populations or incomplete batch correction. To address this, we present CellDiffusion, an unsupervised graph-based model that propagates cellular features through a diffusion process to enhance meaningful cell–cell relationships during integration. By learning a cell–cell attention network, CellDiffusion adaptively identifies corresponding populations across batches. This approach performs cell-specific corrections,

generating embeddings that align cells across datasets while preserving intrinsic cell identities and developmental trajectories.

Using this framework, we find that oligodendrocytes from CR subjects exhibit increased expression of myelin-related genes and enrichment in glycolytic and fatty acid biosynthetic pathways. A subpopulation of oligodendrocytes in CR animals upregulates the cell adhesion gene NLGN1 and shows closer spatial association with axons. Microglia from CR subjects display upregulation of amino acid and peptide metabolism pathways, along with a reduced myelin debris signature. Together, these results reveal cell-type-specific transcriptional reprogramming in response to long-term calorie restriction and highlight potential protective mechanisms against myelin pathology in the aging primate brain.

Keywords: Brain aging, Calorie restriction, Single-nucleus transcriptomics, CellDiffusion

Title: Integrative AI-driven multi-omics identifies NR2F1 as a key driver of lineage plasticity and metastatic progression in prostate cancer

Author list: Siyuan Cheng[1, 3*], Julisa Gonzalez[1,2*], Yaru Xu[1, 3], Chuanli Zhou[4], Xiaoling Li[1,3], Quanhui Xu[1,3], Yuyin Jiang[1,3], Nick Johnson[4], Choushi Wang[4], Carla Rodriguez Tirado[2], Xiang Li[4], Lauren Metang[4], Amanda Leonita[1,3], Melanie Fraidenburg[1,3], Isabella De Pasquale[2,5], Igor Lopes Dos Santos[5], Fei Peng[5], Yunguan Wang[6], Ravikanth Maddipatti[9], Srinivas Malladi[10], Su Deng[1, 3], Maralice Conacci-Sorrell[2,8], Ping Mu[1,2, 3†]

Detailed Affiliations:

[1]Department of Urology, Yale School of Medicine, Yale University, New Haven, Connecticut, 06520, USA, [2]Cancer Biology Graduate Program, University of Texas Southwestern Medical Center, Dallas, Texas 752390, USA, [3]Yale Cancer Center, Yale School of Medicine, Yale University, New Haven, Connecticut, 06520, USA, [4]Department of Molecular Biology, University of Texas Southwestern Medical Center, Dallas, Texas 752390, USA, [5]Department of Cell Biology, University of Texas Southwestern Medical Center, Dallas, Texas 752390, USA, [6]Division of Pediatric Gastroenterology, Hepatology and Nutrition, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA, [8]Department of Cellular Biology, University of Texas Southwestern Medical Center, Dallas, Texas 752390, USA, [9]Department of Internal Medicine, UT Southwestern Medical Center, Dallas, Texas 752390, USA, [10]Department of Pathology, UT Southwestern Medical Center, Dallas, Texas 752390, USA. [*]Co-first first author, [†] Corresponding author

Abstract: Prostate cancer progression and metastasis remain major therapeutic challenges, particularly in the context of androgen receptor (AR)–independent disease. Using an advanced artificial intelligence–driven multi-omics framework that integrates spatial, single-cell, and transcriptomic data from large patient cohorts and prostate cancer cell line models, we identified the transcription factor NR2F1 as a central regulator of lineage plasticity and resistance to AR-targeted therapies. This framework incorporates four synergistic analytical modules, including patient-specific resistance mapping, signaling network inference, drug response prediction, and virtual cell simulation, enabling systematic identification of dynamic regulatory hierarchies and therapeutically actionable vulnerabilities at single-cell resolution. Through AI-guided prioritization followed by CRISPR-based secondary screening, NR2F1 emerged as a potent driver of metastatic and invasive behavior in prostate cancer cells, with ectopic expression significantly enhancing proliferation, migration, invasion, tumor growth, and dissemination to distant organs, including the lungs, liver, and spinal cord. Consistent with AI-predicted regulatory programs, NR2F1 induced transcriptional changes associated with metastatic progression, including activation of the CXCL12 signaling axis, and both genetic and pharmacologic inhibition of the NR2F1–CXCL12–CXCR4/7 axis markedly attenuated metastatic phenotypes. Collectively, this integrative AI-driven analysis identifies NR2F1 as a key driver of prostate cancer plasticity and metastasis and demonstrates a generalizable framework for predictive, functionally validated discovery of clinically relevant drivers and therapeutic vulnerabilities in advanced cancer.

Keywords: Prostate cancer metastasis, NR2F1, Lineage plasticity, AI-driven multi-omics

Title: Quantifiable blood TCR repertoire components associated with immune aging

Author list: Jing Hu[1,2], Mingyao Pan[1], Brett Reid[3], Shelley Tworoger[3,4], Bo Li[1,2*]

Detailed Affiliations:

[1]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, PA, USA, [2]Department of Pathology and Laboratory Medicine, Children's Hospital of Philadelphia, PA, USA, [3]Department of Cancer Epidemiology, Moffitt Cancer Center, Tampa, FL,USA, [4]Knight Cancer Institute and

Division of Oncological Sciences, Oregon Health and Science University, Portland, OR, USA. *Corresponding author

Abstract: T cell senescence alters the homeostasis of distinct T cell populations and results in decayed adaptive immune protection in older individuals, but a link between aging and dynamic T cell clone changes has not been made. Here, using a newly developed computational framework, Repertoire Functional Units (RFU), we investigate over 6500 publicly available TCR repertoire sequencing samples from multiple human cohorts and identify age-associated RFUs consistently across different cohorts. Quantification of RFU reduction with aging reveals accelerated loss under immunosuppressive conditions. Systematic analysis of age-associated RFUs in clinical samples manifests a potential link between these RFUs and improved clinical outcomes, such as lower ICU admission and reduced risk of complications, during acute viral infections. Finally, patients receiving bone marrow transplantation show a secondary expansion of the age-associated clones upon stem cell transfer from younger donors. Together, our results suggest the existence of a 'TCR clock' that could reflect the immune functions in aging populations.

Keywords: T cell aging, TCR repertoire, Repertoire Functional Units, Immune senescence

Title: Cell-free DNA fragmentomics and methylation as AI-enabled multi-omics biomarkers across cancer and neuroinflammatory disease

Author list: Yaping Liu[1,2] Ravi Bandaru[1,2], Jocelyn Liang[1,2], Hailu Fu[1,2], Haizi Zheng[1,2], Li Wang[1,2], Kevin Huang[3], Wen Zhu[4], Lili Zhang[4], Shuchi Gulati[5], Benjamin H. Hinrichs[6], Mingxiang Teng[7], Bing Zhang[8], Marsha Kocherginsky[1], Dechen Lin[9], David A Hildeman[8], Francis P. Worden[10], Matthew O. Old[11], Neal E. Dunlap[12], John M. Kaczmar[13], Maura L. Gillison[14], Dalia El-Gamal[5], Trisha M. Wise-Draper[5], Zongqi Xia[4], Farrah J. Mateen[1]

Detailed Affiliations:

[1]Feinberg School of Medicine, Northwestern University, Chicago, IL, USA, [2]Robert H. Lurie Comprehensive Cancer Center of Northwestern University, Chicago, IL, USA, [3]Genentech Inc., South San Francisco, CA, USA, [4]Department of Neurology, University of Pittsburgh, Pittsburgh, PA, USA, [5]University of Cincinnati Cancer Center, Cincinnati, OH, USA, [6]University of Cincinnati, Cincinnati, OH, USA, [7]H. Lee Moffitt Cancer Center, Tampa, FL, USA, [8]Cincinnati Children's Hospital Medical Center, OH, USA, [9]University of Southern California, Los Angeles, CA, [10]University of Michigan, Ann Arbor, MI, USA, [11]The Ohio State University, Columbus, OH, USA, [12]University of Louisville, Louisville, KY, USA, [13]Medical University of South Carolina, Charleston, SC, USA, [14]The University of Texas MD Anderson Cancer Center, Houston, TX, USA

Abstract: Circulating cell-free DNA (cfDNA) in plasma harbors rich multi-omic signatures, including fragmentation patterns and DNA methylation, which reflect the cells and tissues contributing to cellular turnover in health and disease. Coupled with modern machine learning, these signals offer a powerful, minimally invasive window into disease state and progression across diverse conditions. In this talk, I will present our recent work applying cfDNA multi-omics to two clinically distinct challenges: predicting immunotherapy response in head and neck squamous cell carcinoma (HNSCC) and monitoring disease progression in multiple sclerosis (MS).

First, using a prospective multi-institutional phase II clinical trial (NCT02641093), we performed whole-genome sequencing on 185 plasma cfDNA samples longitudinally collected from 68 patients with locally advanced, surgically resectable HNSCC undergoing neoadjuvant and adjuvant pembrolizumab. We developed the regional motif diversity score (rMDS), a novel fragmentomic metric quantifying the entropy of cfDNA 5′ end motifs across genomic regions. Unsupervised analysis using rMDS robustly distinguished immunotherapy responders from non-responders, outperforming established fragmentomic metrics and copy number alterations while remaining independent of technical confounders. Regions with the most dynamic rMDS changes were enriched at telomere-proximal loci, suggesting a novel link between telomere biology and cfDNA fragmentation. A machine-learning classifier based on rMDS achieved AUC of 0.89-0.99 across validation settings, outperformed PD-L1 expression and tumor fraction, and stratified patients by disease-free survival (hazard ratio 2.67, 95% CI 1.03–6.92, log-rank $p = 0.035$).

Second, we performed low-coverage whole-genome bisulfite sequencing (WGBS) on 75 plasma cfDNA samples from a well-characterized real-world prospective MS clinic cohort with longitudinal disability outcomes. We identified thousands of differentially methylated CpGs and hundreds of differentially methylated regions (DMRs) that distinguished MS from controls, separated MS subtypes, and stratified disability severity, with DMRs enriched in immunologically and neurologically relevant cis-regulatory elements. Models built on DMRs and on inferred

tissue-of-origin patterns reached AUCs of 0.67-0.82, outperforming neurofilament light chain (NfL) and glial fibrillary acidic protein (GFAP) benchmarks in the same cohort. A prognostic methylation risk score further predicted disability progression within a 3-year window (AUC = 0.74).

Together, these studies illustrate how AI-driven analysis of cfDNA fragmentomic and methylation signatures can yield clinically actionable, minimally invasive biomarkers across cancer and neuroinflammatory disease, supporting a unified multi-omics framework for modeling disease in humans.

Keywords: Cell-free DNA multi-omics, Fragmentomics and DNA methylation, Machine learning biomarkers, Disease monitoring

Workshop – Genome Browser Tutorial
August 3rd
1:00 – 3:40 PM
Room: 2220A&B

Chair: Daofeng Li

Title: Browser tutorial session introduction

Author list: Daofeng Li[1]

Detailed Affiliations:
[1]Washington University.

Abstract: TBD

Keywords: TBD

Title: UCSC Genome Browser: Introduction + 5min QA Virtual

Author list: Robert Kuhn,[1]

Detailed Affiliations:
[1]UCSC Genomics Institute.

Abstract: TBD

Keywords: TBD

Title: UCSC Brain Explorer + 5min QA Virtual

Author list: Mary Goldman[1]

Detailed Affiliations:
[1]UCSC Genomics Institute.

Abstract: TBD

Keywords: TBD

Title: WashU Epigenome Browser: Introduction

Author list: Daofeng Li[1]

Detailed Affiliations:
[1]Washington University.

Abstract: TBD

Keywords: TBD

Title: Develop visualization portal (example: BICAN & IGVF)
Author list: Wenjin Zhang[1]
Detailed Affiliations:
[1]Washington University.
Abstract: TBD
Keywords: TBD

Title: HPRC Epigenome Browser
Author list: Tianjie Liu[1]
Detailed Affiliations:
[1]Washington University.
Abstract: TBD
Keywords: TBD

Title: 5min QA for WashU Epigenome Browser
Author list: TBD
Detailed Affiliations:
[1]TBD
Abstract: TBD
Keywords: TBD

Title: Hands-on for WashU Epigenome Browser
Author list: Chad Seng,[1]
Detailed Affiliations:
[1]Washington University.
Abstract: TBD
Keywords: TBD

Workshop – Biological and Biomedical Ontology Workshop in the AI Era
August 3rd
1:00 – 3:40 PM
Room: 2213A

Chairs: Oliver He, Alexander Diehl

Title: Representing dental care-related fear and anxiety
Author list: Bill Duncan[1]
Detailed Affiliations:
[1]University of Florida.
Abstract: Dental care-related fear, anxiety, and phobia (DFA) broadly refer to distressing cognitive, physiological, and behavioral responses related to receiving professional dental care. DFA is dimensional, ranging from mild discomfort to phobia, and high levels of DFA can be one manifestation of generalized anxiety disorder, or with a history of trauma during dental care, can be conceptualized as post-traumatic stress disorder. Furthermore, the patterns of utilization by those who are highly fearful or anxious (e.g., avoiding preventive care often until pain or dysfunction makes it necessary) place significant strain on the care delivery system in terms of delay of preventive and timely care, as well as missed appointments (e.g., late cancellations). Although dental phobia has been (imperfectly) categorized by the Diagnostic and Statistical Manual of Mental Disorders (DSM) and the International Classification of Diseases (ICD), a standard nosology for DFA is not presently available. To address these

shortcomings in DFA terminology and classification, we are developing the Ontology of Dental care-related Fear, Anxiety, and/or Phobia (ODFA).

Keywords: Dental care-related anxiety, Dental phobia, Ontology development, Diagnostic classification

Title: Pain in both general and specific ontological representation

Author list: Alexander Diehl[1]

Detailed Affiliations:

[1]University at Buffalo.

Abstract: We have recently developed the Pain Ontology and pain-related terms in the Oral Health and Disease Ontology (OHD), for representing types, dimensions, and locations of pain experiences. This work includes both high level classes that provide a top-down approach to representing pain related entities and bottom up representations of specific localized pain. In particular, we have transformed the International Classification of Orofacial Pain terminology into an ontology module as part of the OHD. Our ICOP classes provide ontologically correct definitions and include logical relations to external ontology terms for anatomy, biological processes, and dispositions, so that data annotated with these terms can be queried and integrated along multiple axes. We are currently working on a similar ontology-conversion for the International Classification of Headache Disorders (ICHD), which involves careful discernment between headache processes, phenotypes, and clinical findings, as well as associated disorders and diseases. The work on both ICOP and ICHD have provided key insights for the upper Pain Ontology representations.

Keywords: Pain Ontology, Orofacial pain classification, Headache disorders, Biomedical ontologies

Title: Recent developments in social ontology

Author list: Bill Hogan[1]

Detailed Affiliations:

[1]Medical College of Wisconsin.

Abstract: Our ongoing work in social ontology includes representing social modifiers of behavior change interventions. This work is being done under the auspices of the Advancing Prevention Research in Cancer through Ontology Tools (APRICOT) Project, jointly funded by the National Cancer Institute and the National Institutes of Health Office of Data Science Strategy. A key goal of the work is to harmonize the Ontology for Modeling and Representation of Social Entities (OMRSE) with the Behavior Change Intervention Ontology (BCIO). We have been working on modeling social categories, social identity, money and financial entities, ownership, education, and facilities. The APRICOT project is one of four funded projects in the ACCELERATE-BASSO (Behavioral And Social Science through Ontology, BASSO) network, which is organized through a separately funded coordination center. Besides APRICOT, work on a separate rehabilitation ontology has necessitated adding to OMRSE additional representations of communication, including verbal and gestural communication in the setting of therapy. OMRSE has also harmonized occupations with the Occupation Ontology. Future goals include representing additional entities of relevance to the most frequently used determinants of health screeners.

Keywords: Social ontology, Behavior change interventions, OMRSE, Ontology harmonization

Title: Introducing the Injury Ontology

Author list: Barry Smith[1]

Detailed Affiliations:

[1]University at Buffalo.

Abstract: 'Injury' is a term widely used not only in common speech but also in medical, public health, occupational safety, sports science, military, forensic, legal and other domains. The definitions of 'injury' across these disciplines vary widely, resulting in mismatches which cause problems for the collection and use of injury-related data. Such problems will arise where researchers are interested in investigating the role played by injuries in the emergence, for example, of some chronic pain disorder. To address problems of this sort, we are constructing the Injury Ontology (INJO), which is designed to provide a unified definition of 'injury' which will at the same time accommodate the specific injury subtypes identified in different domains. The ontology will include not only types of injury but also types of injury-causing process and of injury sequelae. It also identifies the relation between injury

on the one hand and tissue damage on the other. The ontology then distinguishes two sorts of injury. On the one hand are those caused by acute exposure to different forms of energy, for example in a car crash. On the other hand, are injuries caused by chronic exposure, for example in the case of repeated noise in a factory or of repeated exertions of pressure in typing on a keyboard.

Keywords: Injury Ontology, Injury classification, Tissue damage, Injury causation

Title: The 2026 Cell Line Ontology Updating and CellCards Web Display

Author list: Jie Zheng[1]

Detailed Affiliations:

[1]University of Michigan Medical School.

Abstract: The well-used Cell Line Ontology (CLO) was initiated in 2011 to standardize the representation of cell lines and integrate cell line information from multiple sources. Currently CLO includes over 41,000 terms including nearly 40,000 cell line cell types. Over the years, we identified many issues in CLO including outdated ontology term import, quality-control errors, and modeling inconsistency. To address these issues, we updated imported terms and revised associated CLO terms to ensure that deprecated external terms are no longer reused. We also resolved 609 ROBOT-identified errors, including duplicate labels, duplicate definitions, and terms with multiple definitions. In addition, we improved the modeling of cell line cell terms. We updated the primary CLO design pattern and ensured its consistent implementation by optimizing the object properties. Based on the optimized data pattern, we developed standardized tab-delimited template files and used ROBOT to convert the contents to OWL format, supporting consistent logical representation. CLO is also being integrated into CellCards, an integrated knowledge base of cells and associated biological information, including biomarkers, tissues, developmental stages, and diseases. CellCards provides a platform for exploring cell line cells based on their biological features, including additional details not represented in CLO. Future work will focus on incorporating additional cell lines from resources such as ATCC and Cellosaurus to better support cell line–based biomedical data annotation and analysis, particularly in vaccine research, including studies of cellular immune responses following vaccination and vaccine/adjuvant development.

Keywords: Cell Line Ontology, Ontology quality control, CellCards, Biomedical data integration

Title: A decade of progress in automated quality assurance of biomedical ontologies

Author list: Licong Cui[1]

Detailed Affiliations:

[1]University of Texas Health Science Center at Houston.

Abstract: Biomedical ontologies such as SNOMED CT, the NCI Thesaurus, and the Gene Ontology serve as foundational knowledge resources for a wide range of biomedical applications, including clinical decision support, EHR data analytics, and large-scale knowledge integration. Ensuring their quality is therefore critical yet increasingly challenging as these resources continue to grow in size and complexity. This talk presents a research program spanning nearly a decade that develops systematic, automated approaches to quality assurance of biomedical ontologies, with a focus on identifying missing and erroneous hierarchical IS-A relations. We begin with a principled structural-lexical method grounded in non-lattice subgraph theory, which identifies error-prone subgraph fragments and maps them to specific quality defects and remediations. Building on this foundation, we introduce an evidence-based lexical pattern framework that leverages existing relational evidence in layers to surface missing and erroneous relations. We then advance to learning-based methods, including a deep learning-based approach for SNOMED CT and a hybrid approach combining non-lattice subgraph filtering with fine-tuned pre-trained language models. Finally, we demonstrate that these ontology quality defects have statistically significant, measurable impacts on real-world clinical applications such as patient cohort identification over EHR data - motivating the practical importance of this line of work. Together, these methods form a complementary, evolving framework for scalable biomedical ontology quality assurance.

Title: Automating Intervention Discovery from Scientific Literature: A Progressive Ontology Prompting and Dual-LLM Framework

Author list: Jinjun Xiong[1]

Detailed Affiliations:
[1]University at Buffalo.

Abstract: Speech-Language Pathologists (SLPs) support individuals with communication challenges but identifying evidence-based practices (EBPs) from the scientific literature is challenging due to the high volume of publications, specialized terminology, and inconsistent reporting formats, making manual curation of EBPs laborious and prone to oversight. To address this challenge, we propose a novel framework leveraging large language models (LLMs), which integrates a progressive ontology prompting (POP) algorithm with a dual-agent system, named LLM-Duo. On the one hand, the POP algorithm conducts a prioritized breadth-first search (BFS) across a predefined ontology, generating structured prompt templates and action sequences to guide the automatic annotation process. On the other hand, the LLM-Duo system features two specialized LLM agents, an explorer and an evaluator, working collaboratively and adversarially to continuously refine annotation quality. We showcase the real-world applicability of our framework through a case study focused on speech-language intervention discovery. Experimental results show that our approach surpasses advanced baselines, achieving more accurate and comprehensive annotations through a fully automated process. Our approach successfully identified 2,421 interventions from a corpus of 64,177 research articles in the speech-language pathology domain, culminating in the creation of a publicly accessible intervention knowledge base with great potential to benefit the speech-language pathology community.

Keywords: Large language models, Ontology-guided annotation, Evidence-based practice, Speech-language pathology

Title: From Biomedical Ontologies to Knowledge Discovery: Ontology-Driven Applications in Ignet 2.0 and Vignet

Author list: Junguk Hur

Detailed Affiliations:
University of North Dakota

Abstract: Biomedical ontologies provide structured representations of biological entities and relationships, but their broader value depends on their integration into practical knowledge-discovery systems. Ignet 2.0 and Vignet are complementary web platforms that apply the Interaction Network Ontology, Vaccine Ontology, and Human Disease Ontology to literature-based gene interaction and vaccine research. Ignet 2.0 combines PubMed mining with BioBERT-based scoring of gene-gene interaction evidence and supports ontology-guided interaction classification, disease-centered exploration, and network visualization. Vignet extends this framework to vaccine research through ontology-guided vaccine exploration and heterogeneous networks connecting vaccines, genes, diseases, and drugs. The platforms also provide AI-assisted evidence summarization, a REST API, and a Model Context Protocol endpoint. Together, Ignet 2.0 and Vignet demonstrate how biomedical ontologies can support semantic integration, evidence retrieval, network construction, and AI-assisted knowledge discovery.

Workshop – Cancer Genomics
August 3rd
1:00 – 3:40 PM
Room: 2213B

Chairs: Siyuan Zheng, Xiaojing Wang

Title: An embedding-based framework for statistical testing of LLM-inferred gene-set functions

Author list: Yanhao Tan[1,2], Li-Ju Wang[1], Tianyuzhou Liang[1,2], Ying-Ju Lai[1,3], Chien-Hung Shih[1], Yibing Guo[1], Tyler M. Yasaka[1], George C. Tseng[3], Yu-Chiao Chiu[1,2,3]

Detailed Affiliations:
[1]UPMC Hillman Cancer Center, University of Pittsburgh School of Medicine, Pittsburgh, PA 15232, USA, [2]Department of Medicine, University of Pittsburgh School of Medicine, Pittsburgh, PA 15261, USA, [3]Department of Biostatistics and Health Data Science, University of Pittsburgh, Pittsburgh, PA 15261, USA.
Abstract: Emerging large language models (LLMs) can infer gene functions directly from gene lists, enabling hypothesis generation without predefined gene sets. However, these LLM-derived functional predictions are typically qualitative, and a principled statistical framework for testing them is lacking. Here, we develop an embedding-based statistical framework that transforms gene and function descriptions into vector representations, enabling quantitative evaluation of gene-gene and gene-function relationships. We benchmark seven state-of-the-art embedding models using both curated gene summaries and retrieval-augmented, literature-derived descriptions across diverse biological contexts. Overall, OpenAI's text-embedding-3-large and Google's gemini-embedding-001 achieved the highest performance, capturing gene-gene functional relationships in 88.7-92.5% of Gene Ontology biological processes and approximately 98.7% of canonical pathways. In gene-function association analyses, these models demonstrated high sensitivity (95.2-98.4%) and specificity (72.7-84.4%). Through systematic contamination analysis and application to experimentally derived protein assembly gene sets, our framework reliably distinguished biologically meaningful LLM-inferred gene function hypotheses from noise, outperforming confidence-based inference and conventional gene set enrichment analysis. Together, these results establish LLM-derived embeddings as a quantitative foundation for functional genomics and a generalizable statistical validation layer for LLM-based gene function inference.
Keywords: Large language model; text embedding; gene-set function; statistical hypothesis testing

Title: Creative applications of AI in biomedical research

Author list: Yan Guo[1]

Detailed Affiliations:
[1]University of Miami.
Abstract: Artificial intelligence (AI) has become a transformative force in biomedical research, enabling the integration, interpretation, and translation of complex, high dimensional biological data at unprecedented scale. This abstract highlight a series of AI driven methodological innovations developed across my research program, illustrating how creative combinations of machine learning architectures can address long standing challenges in genomics, immunology, spatial biology, and population health. Collectively, these studies demonstrate how AI can move beyond incremental performance gains to enable new scientific insights and analytic paradigms. One major theme of this work is the use of AI to enhance immune repertoire analysis. We developed ImmuSeeker, an AI enabled framework designed to sensitively detect and characterize immune receptor signatures from high throughput sequencing data. By integrating probabilistic modeling and machine learning, ImmuSeeker improves the identification of immune diversity and clonal expansion patterns, supporting applications in cancer immunology, infectious disease surveillance, and environmental health studies. This work illustrates how AI can expand the analytical resolution of immune profiling while maintaining robustness across heterogeneous datasets. Another innovation addresses incomplete and sparse epigenomic data. In DiffuCpG, we introduced a diffusion-based machine learning approach to impute DNA methylation at CpG sites, leveraging correlation structures across the genome. This method enables accurate reconstruction of methylation profiles from partial measurements, facilitating large scale epigenetic studies that would otherwise be limited by cost or data missingness. By framing imputation as a diffusion process, this work demonstrates a creative adaptation of machine learning principles to a core problem in epigenomics. More recently, we developed TransGCN, a hybrid framework that integrates transformer models, graph convolutional networks, and Monte Carlo sampling to address spatial transcriptomics imputation. TransGCN models both long-range gene–gene relationships and spatial cell–cell interactions, enabling accurate prediction of missing gene expression across multiple spatial platforms, including CosMx, Xenium, and Visium HD. Importantly, the framework is trained on complementary scRNA seq and bulk RNA seq data, demonstrating strong cross platform generalizability and providing uncertainty estimates that enhance interpretability for downstream biological inference. These studies exemplify the creative application of AI to biomedical research by combining novel model architectures with deep domain knowledge. Rather than treating AI as a black box predictor, this body of work emphasizes interpretability, generalizability, and biological relevance.

These innovations not only advance computational methodology but also create practical tools that accelerate discovery and support data driven precision medicine across diverse biomedical domains.

Keywords: AI, Spatial Transcriptomics, Methylation

Title: Proteogenomics guides tumor systems biology: opportunities and challenges

Author list: Chen Huang[1]

Detailed Affiliations:

[1]University Department of Genetics, Department of Biomedical Informatics and Data Science, and O'Neal Comprehensive Cancer Center, University of Alabama at Birmingham, Birmingham, AL 35232, USA

Abstract: The past two decades have witnessed significant progress in personalized cancer therapy, driven largely by advanced genomic studies such as The Cancer Genome Atlas (TCGA). However, our understanding of tumor systems biology at the protein level remains incomplete. Through large-scale, mass spectrometry (MS)-based proteomics, genome-wide protein expression and post-translational modifications have recently been profiled to deepen our understanding of tumor biology and identify novel therapeutic targets.

In our previous studies, we have generated unique insights into tumor biology that are inaccessible through the analysis of genomic or transcriptomic data alone. These insights include the functional interpretation of cancer genetic aberrations, the elucidation of signaling cascades, the identification of protein biomarkers and druggable targets, and improved patient stratification for precision oncology. Conversely, while large-cohort MS proteomics data combined with matched next-generation sequencing (NGS) data—collectively termed 'proteogenomics'—provide unprecedented resources for secondary analysis, it is critical to address the inherent limitations and pitfalls of these data. Challenges such as limited sensitivity and increased noise, compared to sequencing data, remain significant hurdles. Consequently, our recent research also focuses on delineating the constraints of MS proteomics data and improving data processing workflows to more effectively translate these findings into biological and clinical knowledge.

Keywords: Cancer proteogenomics, Mass spectrometry proteomics, Precision oncology, Protein biomarkers

Title: Phenylalanine reprograms TLS architecture and plasma cell fate in MASH-HCC

Author list: Lisa A. Huang[1,#], Yamei Chen[2,#], Jian Gao[1,#], Laura Beretta[1,*], Leng Han[2,*], Liuqing Yang[1,3,4,*] and Chunru Lin[1,4,*]

Detailed Affiliations:

[1]Department of Molecular and Cellular Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA, [2]Biostatistics & Health Data Science, Indiana University, [3]Center for RNA Interference and Non-Coding RNAs, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA, [4]The Graduate School of Biomedical Sciences, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA. [#]Co-first author, [*]Corresponding author

Abstract: Metabolic dysfunction–associated steatohepatitis (MASH) is a rapidly growing cause of hepatocellular carcinoma (HCC) and is associated with poor response to immune checkpoint inhibitors. Here, we identify a metabolic–immune mechanism linking altered phenylalanine (Phe) metabolism to defective humoral immunity in MASH-HCC. Human tumors showed selective loss of mature follicular tertiary lymphoid structures (TLS), and a diet-induced MASH mouse model revealed impaired plasma cell differentiation in tumor-associated TLS and secondary lymphoid organs. This defect was associated with excessive Phe accumulation and a more immunosuppressive tumor microenvironment. Spatial transcriptomic and single-cell analyses indicated that excess Phe reshapes plasma cell states. Mechanistically, Phe activated GCN2-dependent amino acid stress signaling, while chronic type I interferon signaling impaired PERK–eIF2α responses, creating an integrated stress response–resistant state that blocked B-cell differentiation. Restoring Phe metabolism by dietary, pharmacologic, or liver-targeted RNA approaches rescued TLS maturation, improved B-cell differentiation, and enhanced immunotherapy response, identifying Phe metabolism as a targetable vulnerability in MASH-HCC.

Keywords: Metabolic dysfunction-associated steatohepatitis (MASH), Hepatocellular Carcinoma (HCC), Tertiary Lymphoid Structure (TLS), Phenylalanine, Plasma cell, Immunotherapy

Title: HOXA9 maintains cancer-specific 3D genome organization in AML

Author list: Qili Shi[1*], Sajesan Aryal[2*], Chenhui He[1], Qiushi Jin[1], Xinyue Zhou[2], Paul Brent Ferrell Jr[3], Stanley C Lee[4,] Robert S Welner[2], Yang Zhou[5], Yu Luan[1,6,7#], Rui Lu[2#]

Detailed Affiliations:

[1]UT Health San Antonio, Department of Cell Systems and Anatomy, San Antonio, TX, United States; [2]Division of Hematology/Oncology and O'Neal Comprehensive Cancer Center, University of Alabama at Birmingham Heersink School of Medicine, Birmingham, AL, United States; [3]Department of Medicine, Division of Hematology/Oncology & Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center; Nashville Veterans Affairs Hospital, Tennessee Valley Health Care, Department of Veterans Affairs, Nashville, TN, United States; [4]Translational Science and Therapeutics Division, Fred Hutchinson Cancer Center, Seattle, WA, United States; [5]Department of Biomedical Engineering, University of Alabama at Birmingham Heersink School of Medicine and School of Engineering, Birmingham, AL, United States; [6]Greehey Children's Cancer Research Institute, University of Texas Health Science Center at San Antonio, San Antonio, TX, United States; [7]Mays Cancer Center at UT Health San Antonio, San Antonio, TX, United States. [*]Co-first author, [#]Corresponding author

Abstract: The three-dimensional (3D) genome organization is essential for cell identity, yet how lineage-restricted transcription factors actively maintain higher-order chromatin architecture remains unclear. In acute myeloid leukemia (AML), the homeobox transcription factor HOXA9 is a central driver of leukemic stemness, but its direct contribution to cancerous genome topology and regulatory circuitry has been poorly defined. Here, we developed a unique AML model for acute HOXA9 degradation that enables kinetic dissection of leukemia-specific genome organization. By integrating time-resolved nascent transcription, chromatin accessibility, histone modification, architectural protein occupancy, and deep Hi-C profiling, we define the hierarchical framework of HOXA9-dependent genome regulation. We show that HOXA9 directly maintains a core leukemic transcriptional regulatory network by sustaining chromatin accessibility at distal enhancers, including AML-specific super-enhancers, thereby preserving expression of stemness-associated transcriptional circuits. Acute HOXA9 loss triggers rapid enhancer closure and immediate silencing of functionally essential leukemia genes, followed by coordinated disruption of the broader regulatory network, progressive reorganization of higher-order genome architecture, and activation of differentiation programs. At the architectural level, HOXA9 depletion induces selective collapse of AML-specific active compartments, progressive weakening of a subset of topologically associating domain (TAD) boundaries, and widespread loss of enhancer–promoter loops. Mechanistically, HOXA9 interacts with cohesin and promotes its recruitment to HOXA9-sensitive cohesin-associated regulatory elements (HS-CoREs), which largely lack CTCF binding and function as internal loop anchors nested within larger CTCF-defined structures. Loss of HOXA9 at HS-CoREs initiates a hierarchical de-anchoring cascade, leading to secondary destabilization of surrounding CTCF-anchored structural loops and domain boundaries. Importantly, these HOXA9-dependent 3D genome features in our model mirror the AML-specific patterns observed in primary human samples. Together, these findings establish HOXA9 as a structural oncogene that integrates regulatory network control with enhancer-directed cohesin loading to enforce leukemia-specific genome topology.

Keywords: HOXA9, epigenetics, AML, 3D genome organization, cohesin, enhancer

Title: From unresolved variants to actionable biomarkers: defining when long-read sequencing is required

Author list: Xiaotu Ma, Pandurang Kolekar, Yanling Liu, Bensheng Ju, Heather Mulder, Natarajan, Sivaraman, John Easton

Detailed Affiliations:

Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, TN, 38105, USA.

Abstract: Long-read sequencing is often viewed as a higher-resolution alternative to short-read sequencing, rather than as a necessary tool for uncovering variants that are fundamentally unresolvable due to read-length limitations. Here, we present a series of representative cases that delineate conditions under which critical genomic alterations cannot be resolved by short-read approaches. These include complex intragenic rearrangements such as PAX5 intragenic tandem multiplication, structural variants in low-complexity or repetitive regions that evade detection due to mapping ambiguity, cryptic second-hit events in TP53 in osteosarcoma, and DNA-level resolution of canonical oncogenic fusions such as TCF3::PBX1. In each case, short-read data contain partial or ambiguous signals (e.g., soft clipping) but fail to resolve the underlying structural variant, whereas long-read sequencing provides direct, contiguous, and unambiguous characterization.

Importantly, we demonstrate that once these variants are resolved, they can be translated into clinically actionable biomarkers. Structural variants identified by long-read sequencing can be sensitively detected in downstream assays, including short-read sequencing of circulating tumor DNA, establishing a practical framework in which long reads enable discovery and short reads enable scalable monitoring. In addition, allele-resolved long-read data enable reconstruction of the formation of complex structural alterations and precise inference of the functional status of affected transcripts.

Together, these results demonstrate that the primary value of long-read sequencing lies not simply in higher resolution, but in expanding the observable genome and enabling the transition from previously invisible genomic alterations to actionable biological and clinical insights.

Keywords: Long-read sequencing; Structural variation; Cancer genomics; ctDNA

Title: AI-empowered virtual immunopeptidomics of neoantigen immunogenicity

Author list: Yuhao Tan[1,2], Ziqi Yang[1,2], Tong Wang[1,2], Hailong Hu[1,2], Julia Fleming[1,2], Bo Li[1,2]

Detailed Affiliations:

[1]The Children's Hospital of Philadelphia, [2]University of Pennsylvania.

Abstract: T Effective anti-tumor T cell responses depend on the abundance of pMHCs presented on cancer cells, yet neoantigen prioritization has largely focused on predicted pMHC affinity because abundance has been difficult to measure and model. Here we develop epiVIP, a deep neural network that predicts the within-sample rank of individual HLA-I peptide abundance from peptide–HLA sequence and transcriptomic regulators of antigen processing and presentation. Using 1.7 million HLA-I peptides uniformly quantified across 365 samples with paired transcriptomes, epiVIP generalized across cohorts and achieved AUC > 0.8 for ranking unseen peptides in independent test sets. When applied to 33,782 neoantigens from four immunogenicity cohorts, higher predicted abundance was associated with increased immunogenicity, and this association depended on self-discrimination, revealing compensatory effects between antigen abundance and T cell recognition. In three immune-checkpoint blockade cohorts, the abundance of the most highly presented clonal neoantigens was associated with survival. epiVIP also captures the regulatory effects of a proteasome regulator PSME4 on MAGEA3 epitope T cell response and links proteasome-associated shifts in C-terminal residue preferences to differences in the T cell repertoire. These results position pMHC abundance as a key quantitative determinant of neoantigen immunogenicity and motivate incorporating abundance into immunotherapy target selection.

Keywords: Antigen presentation, Deep learning, Immune response, Cancer immunotherapy, Immunopeptidomics

Title: Genomic alterations across racial and ethnic groups reveals their contributions to cancer outcome disparities

Author list: Mehwish Noureen[1], Siyuan Zheng[1,2,3], Yidong Chen[1,2,3], Amelie G. Ramirez[2,3], Jonathan Gelfond[2,3], Xiaojing Wang[1,2,3,4#]

Detailed Affiliations:

[1]Greehey Children's Cancer Research Institute, University of Texas Health San Antonio, San Antonio, TX, USA, [2]Department of Population Health Sciences, University of Texas Health San Antonio, San Antonio, TX, USA, [3]Mays Cancer Center, University of Texas Health at San Antonio, San Antonio, TX, [4]Barshop Institute for Longevity and Aging Studies, University of Texas Health San Antonio, San Antonio, TX, USA. [#]Corresponding author

Abstract: A systematic analysis of genomic alterations across racial and ethnic groups provides critical insights into cancer disparities. Using clinical sequencing data, we compared genomic profiles between minority groups (non-Hispanic Black, non-Hispanic Asian, Hispanic) and non-Hispanic White patients, identifying 87 gene-cancer pairs with significantly different alteration frequencies. Nearly one-third of these genomic disparities were associated with patient survival outcomes, approximately 55% of which linked to unfavorable survival in minority groups. We also observed disparities in tumor mutation burden (TMB) across select cancer types. In a cohort of patients treated with immune checkpoint inhibitors, the FDA recognized TMB-high threshold (TMB >=10 mutations/Mb) did not consistently predict improved survival in non-Hispanic Asian and Hispanic patients. Furthermore, significant disparities in genomic instability (GI) were observed across cancer types, persisting after adjustment for age, sex, metastatic status, and TP53 mutation status, particularly in non-small cell lung cancer and bladder cancer. Elevated GI was associated with worse prognosis in Hispanic bladder cancer patients. Collectively,

these results highlight the contribution of genomic variation to cancer disparities and underscore the need for personalized therapeutic strategies tailored to diverse populations.

Keywords: Cancer disparity, clinical sequencing, race and ethnicity

Title: Genome-scale discovery of RNA-centered regulation in cancer drug response and immunotherapy

Author list: Da Yang[1]

Detailed Affiliations:

[1]Department of Pharmaceutical Sciences, Department of Computational and Systems biology, University of Pittsburgh, Pittsburgh, PA 15261, USA

Abstract: The remarkable success of cancer genomics has transformed our understanding of oncogenic drivers and therapeutic vulnerabilities. However, much of cancer biology remains interpreted through a protein-centric and DNA-centric framework, leaving many RNA-mediated regulatory mechanisms unexplored. Emerging evidence suggests that noncoding RNAs, RNA-binding proteins, and RNA-centered regulatory networks constitute a major layer of gene regulation that shapes tumor evolution, therapeutic response, and immune surveillance.

Our research combines genome-scale functional genomics, CRISPR-based screening, transcriptomics, and RNA interactome technologies to systematically uncover previously unrecognized RNA regulators in cancer. Through large-scale CRISPR activation and interference screens, we have identified functional long noncoding RNAs (lncRNAs) that govern tumor growth, immune evasion, and response to therapy. These studies have revealed novel lncRNA-mediated mechanisms controlling innate immune signaling, tumor-immune interactions, and sensitivity to immune checkpoint blockade. In parallel, we have developed integrative computational and experimental approaches to define RNA-centered regulatory networks underlying cancer drug response and immunotherapy.

More recently, our work has uncovered unexpected RNA-binding activities in canonical transcription factors. We demonstrated that MYC functions as a nuclear RNA-binding protein that interacts with coding and noncoding RNAs to modulate chromatin engagement and transcriptional programs. In contrast, the xenobiotic receptor PXR acts as a cytoplasmic RNA-binding protein that directly regulates mRNA stability and metabolic adaptation. These findings expand the functional landscape of transcription factors beyond their traditional DNA-binding roles and reveal previously unappreciated mechanisms through which RNA–protein interactions regulate cancer phenotypes.

Together, these studies support a model in which RNA-centered regulation represents a fundamental and therapeutically actionable layer of cancer biology. By integrating large-scale discovery platforms with mechanistic investigation, we aim to define how regulatory RNAs and RNA-binding proteins shape cancer drug response and immunotherapy, and how these pathways may be leveraged for precision oncology.

Keywords: Regulatory RNAs; Long noncoding RNAs; CRISPR screening; RNA-binding proteins; Cancer immunotherapy; Drug response

Workshop – Deep Learning and CRISPR Screening August 3rd
1:00 – 3:40 PM
Room: 1220

Chair: Xueqiu Lin

Title: Decoding Non-Coding Regulatory Mutations in Cancer Using FFPE-CUTAC and Interpretable Deep Learning

Author list: Xueqiu (Chu) Lin

Detailed Affiliations:

Fred Hutchinson Cancer Center

Abstract: Non-coding somatic mutations in cis-regulatory elements (CREs) are pervasive across cancer genomes, yet their functional impact remains largely uncharacterized due to low recurrence and subtle individual effects. We present an integrated computational and experimental framework to detect and interpret functional regulatory mutations from formalin-fixed, paraffin-embedded (FFPE) tumor samples. Leveraging FFPE-CUTAC, a low-cost epigenomic profiling assay that simultaneously maps active CREs and detects somatic mutations within them, we

demonstrate sensitive and robust identification of regulatory mutations directly from archival clinical specimens. To interpret how these mutations collectively impact gene regulation, we develop interpretable deep learning approaches that extract functional sequence features associated with regulatory mutations in oncogenic networks. Our framework addresses a critical gap in cancer genomics, connecting the vast landscape of non-coding somatic variation to its functional consequences in gene regulation, and highlights the potential of integrating epigenomic profiling with deep learning for precision oncology.

Title: Harnessing tandem repeats to understand the genetic basis of human diseases

Author list: Ya (Allen) Cui[1]

Detailed Affiliations:
University of California

Abstract: Tandem repeats are a major yet historically underexplored source of genetic variation in the human genome. Recent advances in sequencing technologies and computational methods have enabled their systematic characterization at population scale, revealing important roles in gene regulation and disease susceptibility. In this talk, I will present our recent work on large-scale analyses of tandem repeats across human populations and multi-omic datasets. I will discuss how integrating tandem repeat variation with functional genomics helps identify causal variants, uncover molecular mechanisms, and provide new insights into the genetic basis of human complex diseases.

Title: Structured Modeling of Allelic-specific Gene Expression

Author list: William H. Majoros[1]

Detailed Affiliations:
Duke University, Durham, NC

Abstract: Abnormalities in gene expression can serve as important diagnostic signals in cases of undiagnosed genetic disease. Of particular relevance to the search for gene regulatory defects is allelic imbalance in expression. We describe a family of probabilistic graphical models for dissecting allelic signals in both endogenous genes and massively parallel reporter assays. For endogenous genes, we show that detailed modeling of haplotype structure can improve estimation accuracy when short read data are used, and in the context of pedigrees we show that joint modeling of the sharing of haplotypes and expression imbalance between individuals leads to improved identification of inheritance patterns of genetic causes of imbalance, even when causal factors are unobserved. In the case of pooled, multi-sample reporter assays, we show that imposing structure on the sample pools leads to higher estimation accuracy, due to heterogeneity in allele frequencies and an induced Poisson-binomial structure in the data-generating process. Finally, we show that deep learning models trained on these data can be used to dissect gene regulatory syntax and to detect patterns of selection against both gain- and loss-of-function mutations in noncoding DNA.

Title: Characterizing the genetic architecture and molecular mechanism of circulating fatty acids

Author list: Kaixiong Ye

Detailed Affiliations:
University of Georgia

Abstract:

Background: Fatty acids (FA) play crucial roles in human health, influencing the risk of developing various conditions, such as cardiovascular disease and dementia. While previous genome-wide association studies (GWAS) have identified hundreds of genetic loci associated with the circulating FA levels, these loci account for only a small fraction of trait variance, and the underlying biological mechanisms linking these identified loci to FA metabolism remain largely unclear.

Methods: We performed a multi-trait analysis of GWAS (MTAG) on 19 FA traits (N = 239,268) from UK Biobank to boost discovery power. SNP-based heritability was estimated using GREML-LDMS and partitioned across functional categories using LDSC. We integrated GWAS with a single-cell CRISPRi screen to functionally characterize the target genes and the molecular mechanisms underlying FA-associated loci.

Results: Our MTAG identified 224 genomic loci that are significantly associated with 19 FAs, including 150 novel discoveries not detected in single-trait GWAS. We estimated that genome-wide imputed SNPs explained 11-31% of phenotypic variance. FA-related signals are largely enriched in high LD regions with common variants (MAF > 5%), evolutionarily conserved regions, and regions with active enhancers. To explore the regulatory function of FA-associated loci, we conducted a single-cell CRISPRi screen in > 200,000 HepG2 liver cells, targeting 360 candidate cis-regulatory elements (CREs) from fine-mapped FA trait variants. We identified 501 CRE-gene pairs in cis, identifying 239 genes, such as *APOB* and *FGG*. We validated the top 16 candidate genes independently using targeted CRISPRi in bulk cell cultures. The identified cis-regulated genes are highly interconnected and enriched in lipid metabolism pathways (e.g., high-density lipoprotein particle remodeling). Notably, 20 of these genes were identified as transcription factors, with 12 (e.g., *TCF7L2*) having established roles in lipid metabolism. Inhibition of the CRE containing the putative causal variant rs12259064 significantly altered the expression of *CDK6*, *LGR5*, and *NKD1*. As these genes are responsive to *TCF7L2* depletion, our findings suggest that the CRE (rs12259064) locus exerts trans-effects through modulating *TCF7L2*.
Conclusion: Our integrative multi-trait and functional genomic analyses reveal novel genetic loci and the molecular mechanisms regulating circulating FA levels, providing mechanistic insights into the genetic architecture of fatty acid metabolism.
Keywords: Fatty acids, Genome-wide association studies (GWAS), Single-cell CRISPR screen, Integrative Omics

Title: How to Align Unpaired Single-Cell Omics: Benchmarks, Practical Guidelines, and Impacts on Gene Regulatory Modeling

Author list: Zhana Duren[1]

Detailed Affiliations:
[1]Indiana University School of Medicine, Indianapolis, IN

Abstract: Dissecting the cell-type-specific linkages between gene expression and epigenome is fundamental to understanding gene regulation. While joint multi-omic profiling technologies are advancing rapidly, vast repositories of single-cell data exist as unpaired single-cel sequencing datasets. Integrating these separate, unpaired modalities across distinct molecular spaces, often termed "diagonal integration", remains a formidable computational challenge.
Computational strategies for cross-modal alignment have rapidly expanded. However, because diagonal integration consists of a complex pipeline with distinct modular choices at every stage, from preprocessing and feature selection to latent embedding and batch alignment, users face an overwhelming array of potential workflow combinations without clear consensus on what works best in practice.
In this talk, I will address the central question: What is the most practical and reliable strategy for integrating unpaired single-cell transcriptomic and epigenomic data? I will present a comprehensive benchmarking study that systematically evaluates modular integration workflows using paired multi-omic datasets as ground-truth baselines. Furthermore, I will highlight how sub-optimal integration choices propagate errors downstream, significantly compromising the fidelity of inferred gene regulatory networks. Finally, I will provide actionable best-practice recommendations to guide researchers in executing robust, high-confidence single-cell data integration for regulatory biology.

Title: Perturbation of allele specific chromatin variants in human colon organoids links gene regulation to non-coding variants associated with digestive disease

Author list: Siming Zhao[1]

Detailed Affiliations:
[1]Dartmouth College, Hanover, NH

Abstract: Genetics influence digestive traits and diseases, including colorectal cancer (CRC) and inflammatory bowel disease (IBD). Although genome-wide association studies have identified thousands of associated variants, most causal variants remain unresolved. We used allele-specific open chromatin (ASoC) to identify causal variants in differentiated human colon organoids from 24 patients across all large-intestinal regions. To identify condition specific effect, we tested cytokines important in disease related environments, including TNFa, IL17 and IFNg. Bulk ATAC-seq identified over 20,000 ASoC variants (FDR < 0.1). ASoCs were approximately 70-fold enriched

for CRC and IBD GWAS signals. Annotation-informed fine-mapping identified ASoCs as likely causal variants at about 30% of GWAS loci, including noncoding regions near CTNNB1, SMAD3, and C1orf106. To elucidate their regulatory effects, we employed a high-throughput functional genomics approach, combining CRISPRi with single-cell RNA-seq (Perturb-seq). A MECP2-dCAS9-ZNF10KRAB stable hCO line was transduced with a lentiviral guide RNA library targeting within 100 bp of ASoC loci. This revealed cell-type-specific effects for IBD and CRC variants, linking ASoC loci to differential expression of nearby genes. Integrative analysis uncovered regulatory networks and molecular pathways altered by these variants, including Wnt and TGF-β signaling. Our findings demonstrate that ASoC is a powerful tool for resolving causal variants and provide molecular insights into the regulatory mechanisms underlying digestive disease, advancing precision medicine for CRC and IBD.

Title: Improving Disease Risk Prediction with Functional Genomics Annotations Using LLMs

Author list: Wanwen Zeng[1]

Detailed Affiliations:

[1]Stanford University, Stanford, CA

Abstract: Functional genomics annotations provide critical information for interpreting noncoding genetic variants and improving disease risk prediction by capturing regulatory effects across diverse cellular contexts. However, generating these context-specific annotations requires large-scale multi-omics data and computationally intensive analyses, limiting their application to biobank-scale whole-genome sequencing studies.

Here, we present an LLM-based framework that efficiently incorporates functional genomics annotations into disease risk prediction while substantially reducing the computational burden of annotation generation. By learning regulatory patterns across diverse cell types and tissues, our approach captures context-dependent variant effects without requiring exhaustive de novo annotation for every dataset.

We demonstrate that integrating these functional annotations consistently improves disease risk prediction while enabling scalable analysis of large whole-genome sequencing cohorts. Our framework provides an efficient and biologically interpretable strategy for functionally informed genetic risk prediction.

Title: Interpret human genetic variants via single-cell multimodal omics and AI/ML models

Author list: Yang Li[1]

Detailed Affiliations:

[1]Washington University in St. Louis Louis, MO

Abstract: It's challenging to interpret human genetic variants due to a lack of functional annotations in the disease-relevant cell types and model organisms. Recent advances in single-cell technologies are widely employed to resolve the spatiotemporal dynamics of the epigenome and 3D genome, either alone or in combination with the transcriptome in human and mammalian cells. We constructed the largest single-cell epigenetic atlas to date, comprising ~11 million cells curated from 64 high-quality human and mouse studies across 29 organs. We harmonized 346 cell types and identified over 2.7 million candidate cis-regulatory elements (cCREs) spanning various developmental stages, ages, and experimental conditions. Compared to bulk-level databases, we identified ~20% and 40% novel cCREs in human and mouse, respectively. Through integrative analysis, we have linked many putative transcriptional regulatory elements to potential target genes and characterized cell-type-specific gene regulatory programs. By comparing orthologous cell types, we evaluated regulatory conservation between humans and mice at both the sequence and functional levels. We also uncovered significant regulatory divergence between species, particularly in immune cell populations. We integrated this atlas with 101 genome-wide association studies (GWAS) of human disorders and traits to identify disease-relevant cell populations. Finally, we further applied multimodal deep learning models to prioritize genetic risk variants and their impacts on transcriptome, epigenome and 3D chromatin architecture in specific cellular contexts. This multimodal cell atlas is freely available to the public through an interactive web portal and AI agent, CATLAS (www.catlas.org).

Workshop – Clinical Imaging, Multimodal AI, and Healthcare Deployment
August 3rd
4:00 – 6:00 PM
Room: 2220A&B

Chair: Shaolei Teng

Title: Matching Speech Models to ADRD Subtypes: A Mechanistic Benchmark

Author list: Xiaoyu Zhang[1], Wei Bo[1], Anarghya Das[1], Longxiang Pan[1] and Wenyao Xu[1]

Detailed Affiliations:
[1]SUNY at Buffalo.

Abstract: Pretrained speech models have become the dominant approach for voice-based detection of Alzheimer's Disease and Related Dementias (ADRD), yet these models are trained with fundamentally different objectives—contrastive learning, cluster prediction, denoising, and automatic speech recognition (ASR)—that induce distinct representational biases. Meanwhile, ADRD subtypes disrupt different dimensions of speech: Alzheimer's disease (AD) and mild cognitive impairment (MCI) primarily affect temporal organization such as pause patterns and hesitations, whereas primary progressive aphasia (PPA) degrades linguistic content. Despite the theoretical expectation that training-induced feature biases should create differential subtype sensitivities, no study has systematically established this correspondence. We address this gap by benchmarking seven pretrained speech models—spanning all major pretraining paradigms— on five-class ADRD subtype classification using GloVoAD, the largest multilingual voice-ADRD dataset (2,058 participants, 9 corpora, 3 languages). Our benchmark reveals that no single model dominates: Whisper (ASR-supervised) achieves the highest PPA detection (F1 = 0.774), while HuBERT (cluster prediction) leads on AD (F1 = 0.453) and WavLM is competitive on MCI (F1 = 0.500). We then conduct two mechanistic verification experiments. First, region masking demonstrates that WavLM, Whisper-medium, and UniSpeech-SAT depend on pause regions for AD/MCI classification, whereas contrastive and cluster-prediction models rely on speech regions. Second, linguistic perturbation via audio reversal reveals that Whisper-medium's PPA detection collapses by 0.782 F1 points—a near-total failure—when linguistic content is destroyed, far exceeding the drops observed in other models. Together, these findings establish mechanistic evidence linking training objective through feature bias to subtype sensitivity, mirroring known clinical pathology. We propose a principled multi-model clinical decision framework in which each model is deployed where its training-induced bias aligns with the target subtype's speech disruption profile.

Keywords: ADRD subtype classification, pretrained speech models, speech biomarker, mech- anistic analysis

Title: A framework for dynamically training and adapting deep reinforcement learning models to different, low-compute, and continuously changing radiology deployment environments

Author list: Guangyao Zheng[2], Shuhao Lai[1], Vladimir Braverman[2], Michael A. Jacobs[3,4], and Vishwa S. Parekh[5]

Detailed Affiliations:
[1]Department of Computer Science, The Johns Hopkins University, Baltimore, MD 21208, USA, [2]Department of Computer Science, Rice University, Houston, TX, USA, [3]Department of Diagnostic And Interventional Imaging, McGovern Medical School, UTHealth Houston, Houston, TX, USA, [4]Department of Radiology, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA, [5]University of Maryland Medical Intelligent Imaging (UM2ii) and Department of Diagnostic Radiology and Nuclear Medicine, University of Maryland School of Medicine, Baltimore, MD 21201, USA.

Abstract: Combining AI (Artificial Intelligence) with edge devices is a powerful tool that can revolutionize the healthcare industry by enabling real-time diagnostics, reduced latency, and greater privacy and accessibility for people around the world. However, edge devices, by their nature, are generally not equipped with powerful GPUs; thus, models are trained offline and deployed on edge devices, with the hope that inference will not take long. In addition, deployment environments evolve rapidly (for example, changes in imaging protocols, new diseases, or different user behaviors), and portable, low-compute edge devices such as endoscopes and point-of-care ultrasound would be required to continually learn and adapt to these changes without powerful GPUs. To this end, we developed three image coreset algorithms to compress and denoise medical images for lifelong reinforcement

learning on edge devices. We implemented neighborhood averaging coreset, neighborhood sensitivity-based sampling coreset, and maximum entropy coreset on full-body DIXON water and DIXON fat MRI images to compress these images during training, lifelong learning, and testing on the localization of five different landmarks in wholebody MRI (left knee, right trochanter, left kidney, spleen, and lung). All three coresets produced 8x, 27x, 64x, and 125x compression with excellent performance in localizing five anatomical landmarks across both imaging environments. Maximum entropy coreset with 125x obtained the best performance of 14.61 ± 11.97 average distance error, compared to the conventional lifelong learning framework's 19.24 ± 50.77.

Keywords: TBD

Title: Multimodal Deep Learning for Alzheimer's Disease Prediction with Improved Training Stability and Flexibility

Author list: Jianan Liu[1,2], Jincheng Shi[3], Long Zhao[1,2], Yuanxin Yao[3], Shihua Li[1,2], Rongjie Liu[4]

Detailed Affiliations:

[1]School of Automation, Southeast University, Nanjing, 210096, P. R. China, [2]Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Nanjing, 210096, P. R. China, [3]Department of Statistics, Florida State University, Tallahassee, FL, 32304, USA, [4]Department of Statistics, University of Georgia, Athens, GA, 30602, USA.

Abstract: Alzheimer's disease (AD) is a progressive neurodegenerative disorder that requires early and accurate diagnosis, particularly at the mild cognitive impairment (MCI) stage. The increasing availability of multimodal datasets has enabled the development of deep learning models that integrate heterogeneous data sources to improve diagnostic performance. However, existing approaches often suffer from unstable training dynamics, including gradient vanishing and explosion, which may reduce model robustness and generalization in clinical applications. This study proposes a multimodal deep learning framework for AD prediction using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI). The framework integrates structural MRI, PET imaging, and non-imaging clinical variables to capture complementary disease-related information. A tunable nonlinear transformation mechanism is introduced to enhance the adaptability of model representations while maintaining stable gradient propagation during training. Theoretical analysis is provided to characterize its key properties, including nonlinearity, continuity, differentiability, and boundedness, and to demonstrate its capability to approximate commonly used activation functions. Experimental results show that the proposed framework achieves improved diagnostic performance compared with baseline models, especially in the clinically important task of MCI identification. Ablation analysis further reveals that different modalities contribute unequally to AD classification. Clinical features provide the dominant predictive signal, while neuroimaging features offer complementary disease-related information. In contrast, genetic or static features contribute only modestly, suggesting that they are more suitable as contextual risk indicators than as strong standalone diagnostic markers. These findings indicate that multimodal integration should be interpreted as a mechanism for supplementing dominant clinical information rather than as evidence that all modalities contribute equally. The proposed nonlinear transformation mechanism also improves training stability and generalization. It achieves better validation performance, suggesting an implicit regularization effect and reduced overfitting. This is particularly meaningful for MCI classification, where diagnostic boundaries are subtle and model robustness is critical. Overall, this work highlights the importance of stable nonlinear transformation design in multimodal AD prediction and provides a flexible framework for early diagnosis. Future work will extend the current cross-sectional classification setting to longitudinal prediction tasks, such as estimating MCI-to-AD conversion risk over time.

Keywords: Alzheimer's disease, multimodal transformer, activation function, disease prediction, training stability, nonlinear transformation

Title: RA-CMF: Region-Adaptive Conditional MeanFlow for CT Image Reconstruction

Author list: Md Shifatul Ahsan Apurba[1], Md Selim[1] and Jin Chen[1]

Detailed Affiliations:

[1]University of Alabama at Birmingham.

Abstract: The use of CT imaging is important for screening, diagnosis, therapy planning, and prognosis of lung cancers. Unfortunately, due to differences in imaging protocols and scanner models, CT images acquired by different means may show large differences in noise statistics, contrast, and texture. In this study, we develop a

novel conditional MeanFlow pipeline for CT image reconstruction. We introduce a conditional MeanFlow network that models the enhancement trajectory by predicting image-conditioned flow fields given intermediate image states. The image enhancement network is trained with a MeanFlow consistency loss along with the image reconstruction loss. In order to provide an adaptive refinement process in terms of spatial location of enhancements, we integrate a regional reinforcement learning-driven policy network into our approach. The policy network receives information about the MeanFlow rollouts and provides predictions in terms of tile-wise refinement budgets, stopping criteria, and total budget allocation of enhancement processes. Our policy network is trained through reinforcement learning in a policy gradient framework, where the goal of the training reward is to maximize improvement of enhancements while minimizing unnecessary computations and avoiding instabilities. In this way, our approach combines conditional flowbased enhancement with reinforcement learning–based spatial enhancement control. This allows our approach to focus more attention on enhancing difficult areas while stabilizing areas already showing sufficient quality. Experimentally, we validate the effectiveness of our approach for CT image reconstruction. Our results show high accuracy in the tumor ROI, with the average radiomic feature CCC being 0.96, an average PSNR of 31.30 ± 4.16, and average SSIM of 0.94 ± 0.07. Moreover, there is an improvement in the overall quality of images, with an average PSNR of 34.23 ± 1.71 and average SSIM of 0.95 ± 0.01.

Keywords: Computed tomography, CT image reconstruc- tion, image enhancement, conditional MeanFlow, reinforcement learning, region-aware refinement, medical image analysis

Title: WoundWise: A Mobile Deep Learning Solution for Accurate Pressure Ulcer Staging

Author list: Daniel Change[1#], Runxi Liu[1#], Xuanzhe Wang[1#], Zhongming Zhao[2], Qian Liu[1,3*]

Detailed Affiliations:

[1]Nevada Institute of Personalized Medicine, University of Nevada, Las Vegas, 4505 S Maryland Pkwy, Las Vegas, NV 89154, USA, [2]Center for Precision Health, McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA, [3]School of Life Sciences, College of Sciences, University of Nevada, Las Vegas, 4505 S Maryland Pkwy, Las Vegas, NV 89154, USA. [#]Co-first author, [*]Corresponding author

Abstract: Pressure ulcers represent a serious yet largely preventable healthcare burden in the United States, affecting millions of adults and contributing to billions of dollars in annual healthcare expenditures. Delayed or inaccurate staging can lead to severe complications, and pressure ulcers are associated with approximately 60,000 hospitalrelated deaths each year. Therefore, accurate and early staging is critical for effective treatment, timely intervention, and prevention of disease progression. Recent advances in deep learning have enabled image-based approaches for automated pressure ulcer staging. However, existing models mainly rely on large, computationally intensive architectures with limited training samples, making it challenging to balance predictive performance with efficient deployment on resource-constrained mobile devices. This limitation hinders their applicability in real-world settings, particularly for point-of-care and home-based monitoring. To address this challenge, we developed WoundWise, a lightweight deep learning solution designed for deployment on widely available smartphones. WoundWise achieves high accuracy in pressure ulcer staging while maintaining computational efficiency suitable for on-device inference. Independent evaluations demonstrate that WoundWise performs robustly across stages, with particularly strong performance in distinguishing early-stage and advanced ulcers. WoundWise is publicly available and has the potential to facilitate early detection, reduce disease progression, and ultimately lower the healthcare burden associated with pressure ulcers.

Keywords: Pressure ulcer staging; deep learning; image-based diagnosis; on-device artificial intelligence

Title: ICA-Based Analysis of Progressive Multi-Scale Structural–Functional Fusion for Alzheimer's Disease

Author list: Yanteng Zhang[1,*], Anees Abrol[1], Yuxiang Wei[1], Selim Suleymanoglu[1], Vince Calhoun[1,*]

Detailed Affiliations:

[1]Center for Translational Research in Neuroimaging and Data Science (GSU, GaTech, Emory), USA. [*]Corresponding author

Abstract: Detecting subtle brain alterations in Alzheimer's disease (AD) remains a major challenge for clinical diagnosis. Structural and functional MRI provide complementary information, yet their inherent differences make effective integration nontrivial. To address this, we propose a multiscale multimodal graph convolutional

framework that jointly models structural and functional connectivity across different spatial resolutions. In addition, we investigate the role of progressive fusion by examining how multimodal and multiscale integration influences learned representations. Specifically, based on Independent Component Analysis (ICA), we analyze the evolution of feature relationships across modalities and scales under different fusion settings, focusing on how the representation structure is progressively reorganized through the fusion process. Classification results indicate that multimodal fusion leads to more structured and consistent representation patterns, while also improving the interpretability of brain network organization. The proposed framework achieves improved performance compared to single-modality and singlescale approaches, demonstrating the benefit of progressive fusion for representation learning and analysis.

Keywords: Alzheimer's disease, magnetic resonance imaging, multimodalfusion, multiscaleanalysis, graphneural networks

Workshop – Special panel: AI in Research and Scientific Writing
August 3rd
4:00 – 6:00 PM
Room: 2213B

Chair: Wei Zhang

Title: TBD

Author list: Zhandong Liu
Detailed Affiliations: Baylor College of Medicine

Abstract: TBD

Keywords: TBD

Title: From Prompt to Publication: Applying AI Agents to Bioinformatics Research

Author list: Wei Zhang
Detailed Affiliations:
[1]University of Central Florida

Abstract: Bioinformatics research increasingly requires investigators to navigate large multimodal datasets, rapidly expanding literature, unfamiliar software repositories, and complex computational pipelines. This presentation introduces a practical framework for incorporating AI agents throughout the research lifecycle, including project scoping, literature review, dataset discovery, algorithm development, implementation, debugging, evaluation, and manuscript preparation. Using bioinformatics examples, the presentation will demonstrate how AI agents can assist with exploring repositories, explaining computational methods, debugging analysis pipelines, summarizing results, and supporting reproducible research workflows. It will also emphasize the importance of human oversight in validating biological interpretations, statistical analyses, data provenance, and reproducibility. Attendees will gain practical guidance for integrating AI agents into bioinformatics research while maintaining scientific rigor and responsible AI use.

Title: Agentic Loop Engineering for Bioinformatics Research: Data Gathering, Algorithm Development, and Paper Writing

Author list: Tejas Sandip Mahajan[1], Wei Zhang[1]
Detailed Affiliations:
[1] University of Central Florida

Abstract: TBD

Keywords: TBD

Title: TBD
Author list: TBD
Detailed Affiliations:
[1]TBD
Abstract: TBD
Keywords: TBD

Title: TBD
Author list: TBD
Detailed Affiliations:
[1]TBD
Abstract: TBD
Keywords: TBD

Workshop – AI-Driven Multi-Omics for Disease Mechanisms and Therapeutic Discovery August 4th
9:00 AM – 12:00 PM
Room: 2220A&B

Chairs: Hongying Sun, Kaifu Chen

Title: Longitudinal epigenome-wide associations and pre-chemotherapy prediction of post-chemotherapy inflammation in patients with breast cancer undergoing chemotherapy

Author list: Hongying Sun[1]
Detailed Affiliations:
[1]University of Rochester.

Abstract: Background: Inflammation is linked with cancer outcomes; systemic cytokine levels are associated with cognitive and functional impairments during chemotherapy. Epigenetic changes, such as DNA methylation, are also associated with cytokine levels, yet the relationship between epigenetic changes and inflammation during breast cancer treatment remains unclear. Methods: We performed epigenome-wide methylation association analyses to identify CpG sites whose DNA methylation levels were associated with circulating inflammatory cytokine concentrations in a nationwide cohort of 241 participants (192 patients with breast cancer and 49 age/sex-matched healthy controls). DNAm was measured using Illumina 450K and EPIC arrays at pre-chemotherapy (Timepoint [T] 1) and post-chemotherapy (T2). DNA methylation was analyzed on the M-value scale (log2 ratio of methylated to unmethylated probe intensities), allowing more stable variance for linear modeling. Cytokines measured from serum included IL-4, IL-6, IL-8, IL-10, and TNF-α, as well as soluble TNF receptors I and II (sTNFRI/II). Independent CpG-wise linear regression models were fit using limma, adjusting for age, group (chemotherapy vs control), and array type (450K vs EPIC). Analyses examined for cross-sectional associations between DNAm and inflammation at T1 and T2, as well as prediction of T2 inflammation from T1 DNAm. False discovery rate (FDR) < 0.05 defined epigenome-wide significance. Results: Across all analyses (N = 241; mean age = 60 ± 8 years; 78% college degree or above; 71% married/long-term relationship), we tested 449,521 CpG sites and identified five CpG–biomarker associations meeting FDR < 0.05. T1 DNAm at cg08458121 (chr17:1,028,676; CpG island within the gene body/5′UTR region of ABR) was positively associated with T1 IL-10 (log2) (T1→T1; β = 0.224, FDR = 0.032, N = 228). T2 DNAm at cg08458121 (chr17:1,028,676; CpG island within the gene body/5′UTR region of ABR) was positively associated with T2 IL-10 (log2) (T2→T2; β = 0.233, FDR = 0.026, N = 222), demonstrating consistent direction of association across timepoints. T2 IL-4 (log2) was positively associated with T2 DNAm at cg01963906 (chr1:27,677,240; intragenic CpG island in synaptotagmin-like protein 1 [SYTL1]; β = 0.126, FDR = 0.019, N = 222). In longitudinal prediction models (T1→T2), T1 DNAm was inversely associated T2 sTNFRI at cg01479664 (chr16:58,016,655; TEPP; N_Shore; β = −0.188, p = 1.6×10^{-7}, FDR = 0.041, N = 222) and cg03198678

chr3:194,381,165; LSG1; OpenSea; $\beta = -0.091$, $p = 2.1\times10^{-7}$, FDR = 0.041, N = 222). This epigenome-wide analysis identified DNAm sites associated with circulating inflammatory biomarker levels at pre-chemotherapy and post-chemotherapy in breast cancer.

Keywords: DNA methylation, breast cancer, chemotherapy, EWAS, biomarker

Title: Quantum Computing for Single-Cell Biology: Networks, Dynamics, and Generative Models

Author list: James J. Cai[1,2]

Detailed Affiliations:

[1]Department of Veterinary Integrative Biosciences, [2]Department of Electrical & Computer Engineering, Texas A&M University, College Station, TX, USA.

Abstract: Single-cell technologies are generating unprecedentedly complex datasets, but current computational methods struggle to capture their full richness. Our work explores how quantum computing can transform this landscape. First, we developed a gate-based quantum circuit model where qubits represent genes and entangling gates encode gene regulatory interactions, enabling us to infer network structure directly from single-cell transcriptomes (npj Quantum Inf., 2023). Next, we advanced a quantum annealing framework that formulates gene selection and cell fate dynamics as quadratic optimization problems, successfully revealing hidden drivers of differentiation and drug resistance (Quantum Mach Intell., 2025). Building on this foundation, our new research pioneers three directions: (1) quantum differential network algorithms to map gene regulatory network rewiring between cell states, (2) phase-controlled quantum circuits to jointly model single-cell gene expression and chromatin accessibility, and (3) quantum generative models that simulate single-cell data by leveraging entanglement to capture gene-gene and cell-cell interactions (arXiv:2510.12776, 2025). Together, these approaches show how quantum resources can unlock new biological insights, opening the path toward quantum-enabled single-cell biology.

Keywords: Single-cell biology, quantum computing, network science, generative models, systems biology

Title: Cross-Trait PRS and data-driven subtyping to uncover obesity heterogeneity and enable precision risk prediction

Author list: Huihuang Yan[1], Andres J. Acosta[2], Tahmina Sultana Priya[1], Tara L. Haynes[3], Konstantinos N. Lazaridis[2], Eric W. Klee[1], Daniel Schaid[1], Shulan Tian[1]

Detailed Affiliations:

[1]Division of Computational Biology, Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, USA, [2]Division of Gastroenterology and Hepatology, Department of Internal Medicine, Mayo Clinic, Rochester, MN, USA, [3]Center for Digital Health, Mayo Clinic, Rochester, MN, USA

Abstract: Obesity is a chronic, relapsing, systemic disease characterized by excessive adiposity that disrupts metabolic homeostasis across multiple organs. Although body mass index (BMI) remains the most widely used clinical surrogate, it fails to capture variations in fat distribution, muscle mass, and genetic predisposition. To address this limitation, we applied a compose polygenic risk score (cPRS) approach leveraging cross-trait genetic architecture in the Mayo Clinic Biobank (n = 51,186). We retrieved 174 single-trait PRSs for seven adiposity-related traits from the PGS Catalog and derived a cPRS using elastic net regression implemented in the R package PRSmix+, adjusting for age, sex, and ancestry PCs. We then performed latent class analysis (LCA) on individuals in the top 20% of the cPRS distribution (n = 4,650) using 10 key clinical variables to identify distinct obesity subtypes.

The cPRS markedly outperformed conventional sing-trait BMI PRS, explaining 39.5% of the phenotypic variance in BMI (95% CI: 37.1%–42.0%). In a large independent validation cohort (n=96,470), the cPRS demonstrated strong discrimination between individuals with obesity (BMI ≥30 kg/m²) and normal-weight controls (BMI ≤25) (AUC=0.78). Notably, 65.1% of participants in the top decile of cPRS had obesity (mean BMI 33.7 kg/m²), and these individuals had 5.1-fold increased odds of obesity (OR 5.1, 95% CI: 4.6–5.6) compared with those in the 40th–50th percentile.

LCA identified six biologically distinct obesity subtypes. The three largest groups, Metabolically Healthy Obese I (MHO-I, 27.7%), MHO-II (20.1%), and Dyslipidemia High-BMI Obese (DHBO, 19.2%), together accounted for about two-thirds of all cases. These were followed by Cardiovascular-Kidney-Metabolic Obese I (CKMO-I, 16.1%), CKMO-II (12.2%), and Genetic Hepatic Steatosis Obese (GHSO, 4.7%). Each subtype exhibited distinct clinical,

genetic, and comorbid profiles. In 10-year cumulative incidence analyses, both CKMO and GHSO subtypes showed significantly increased risks of cardiometabolic, renal, and hepatic complications. For example, compared with the MHO-I subtype, CKMO-II had a 1.89-fold higher risk of coronary heart disease (HR 1.89, 95% CI: 1.75–2.09), while GHSO demonstrated an even greater risk (HR 2.48, 95% CI: 2.09–3.02).

Our cPRS approach successfully constructed a genetically informed obese cohort. The resulting subtypes exhibited distinct clinical, genetic, and comorbid profiles, providing a strong rationale for subtype-specific risk prediction on obesity-related complications

Keywords: Polygenic risk score; Cross-trait; Data-driven clustering; Latent class analysis

Title: Agentic AI for atlas analysis of cell-cell communications

Author list: Rongbin Zheng[1,2], Jasmine Liu[1], Kulandaisamy Arulsamy[1,2], Zimo Zhu[1], Yang Zhang[3,4], Tadataka Tsuji[3,4], Hong Chen[5], Lili Zhang[1,2], Yu-Hua Tseng[3,4], Kaifu Chen[1,2,*]

Detailed Affiliations:

[1]Department of Cardiology, Boston Children's Hospital, Boston, 02115, MA, USA, [2]Department of Pediatrics, Harvard Medical School, Boston, 02115, MA, USA, [3]Section on Integrative Physiology and Metabolism, Joslin Diabetes Center, Harvard Medical School, Boston, 02115, MA, USA, [4]Department of Medicine, Harvard Medical School, Boston, 02115, MA, USA, [5]Vascular Biology Program, Boston Children's Hospital and Harvard Medical School, Boston, 02115, MA, USA

Abstract: Cell–cell communication is central to tissue function and disease, yet its interpretation across large-scale single-cell atlases remains challenging. Methods such as MEBOCOST enable inference of metabolite-mediated cell–cell communication (mCCC) but translating these complex landscapes into biological insight remains a bottleneck. Here, we present an agentic AI framework implemented within the Metabolite-Mediated Cell–Cell Communication Portal (MCCP), where large language model–driven agents coordinate communication prioritization, pathway interpretation, cross-dataset comparison, and hypothesis generation with transparent and reproducible reasoning. By unifying atlas-scale data with agent-based analysis, MCCP enables dynamic, user-guided exploration of intercellular communication and provides a generalizable paradigm for AI-driven biological discovery.

Keywords: Metabolite, cell-cell communication, cell signaling, single cell, scRNA-seq, AI agent, Agentic AI, artificial intelligence

Title: Cellular and molecular dynamics of smooth muscle cells during aortic disease development

Author list: Yanming Li[1,*], Chen Zhang[1], Hernan G. Vasquez[1], Deborah Vela[2], Abhijit Chakraborty[1], Yang Li[1], Xinghao Xu[3], Guizhen Zhao[3], Bowen Li[1], Kimberly Rebello[1], Samantha Xu[1], Alon R. Azares[2], Lin Zhang[1], Zhen Zhou[4], Yvan Saeys[5,6], George Tellides[7], Jay D Humphrey[8], Steven B. Eisenberg[9], Hong S. Lu[10, 11], Joseph S. Coselli[1, 2, 12], Kenneth K. Liao[1], Alexis E. Shafii[1], Gabriel Loor[1], Alan Daugherty[10, 11], Dianna M. Milewicz[4], Scott A. LeMaire[13, *], Ying H. Shen[1, 2, 12, *]

Detailed Affiliations:

[1]Division of Cardiothoracic Surgery, Michael E. DeBakey Department of Surgery, Baylor College of Medicine, Houston, Texas, USA, [2]The Texas Heart Institute, Houston, Texas, USA, [3]Department of Pharmacological and Pharmaceutical Sciences, University of Houston College of Pharmacy, Houston, Texas, USA, [4]Division of Medical Genetics, Department of Internal Medicine, McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, Texas, USA, [5]Data Mining and Modelling for Biomedicine, VIB Center for Inflammation Research, Ghent, Belgium, [6]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium, [7]Department of Surgery (Cardiac), Yale School of Medicine, New Haven, Connecticut, USA, [8]Department of Biomedical Engineering, Yale University, New Haven, Connecticut, USA, [9]Department of Cardiothoracic and Vascular Surgery, The University of Texas Health Science Center at Houston, Houston, Texas, USA, [10]Saha Cardiovascular Research Center, University of Kentucky, Lexington, Kentucky, USA, [11]Department of Physiology, University of Kentucky, Lexington, Kentucky, USA, [12]Cardiovascular Research Institute, Baylor College of Medicine, Houston, Texas, USA, [13]Department of Cardiothoracic Surgery, Geisinger, Danville, Pennsylvania, USA.

Abstract: The factors driving progression of ascending thoracic aortic aneurysm (ATAA) to dissection (ATAD) remain poorly understood. In this study, scRNA-seq analysis of patient ascending aortas revealed profound smooth

muscle cells (SMCs) phenotypic transition into extracellular matrix (ECM)-producing SMCs in ATAA. However, this transition was significantly inhibited in ATAD, instead, SMCs shifted toward pro-inflammatory and pro-death phenotypes. This differential SMC phenotypic transitions in ATAA and ATAD was also observed in angiotensin II (AngII)-infusion mouse model. Spatial transcriptomics detected ECM-producing SMCs in aortic wall and was associated with healing and wall remodeling. Furthermore, scATAC-seq analysis suggested this phenotypic transition was induced at epigenetic level. TGF-β/SMAD and YAP/TEAD were identified as the primary pathways epigenetically driving SMC transition to ECM producing SMCs, likely through direct chromatin remodeling at ECM gene loci. Finally, SMC-specific deletion of Yap1 or Tgfbr2 in mice prevented SMC transition to ECM-producing SMCs, that was associated with increased aortic dissection and rupture. Our findings suggest that SMC transition to an ECM-producing phenotype ATAA may represent a protective response essential for aortic wall stabilization, whereas compromise in this response may contribute to disease progression to ATAD. YAP/TEAD and TGF-β/SMAD pathways are potential therapeutic targets for preventing ATAD.

Keywords: Aortic Aneurysm and Dissection; Smooth Muscle Cell Phenotype Transition; Adaptive and Reparative Response; Single-cell Omics; Spatial Transcriptomics

Title: Computational biology in early drug discovery: turning complex data into biological insights

Author list: Yu Sun[1]

Detailed Affiliations:

[1]Biogen, Inc. Research Department, 225 Binney St., Cambridge, MA, 02142, USA.

Abstract: Drug discovery is a complex procedure that requires significant resources. Early research has been critical in selecting targets, identifying molecules, performing validation, and nominating the most promising candidate for advancement towards the clinic. Among these processes, computational biology plays a transformative role in modern drug discovery, powering target discovery and validation. Our team has been dedicated to developing novel tools that facilitate robust data processing and interpretation. We have developed a series of pipelines for omics data processing, including RNASequest (Journal of Molecular Biology, 2023) for bulk RNA-seq, scRNASequest (BMC genomics, 2023) and ScaleSC (Bioinformatics Advances, 2025) for single-cell RNA-seq, and the recently published SpaceSequest (JARPBS, 2026) for spatial transcriptomics data analysis. These workflows integrate state-of-the-art tools and deliver end-to-end data analysis solutions. Moreover, we have focused on improving data visualizations for end users, with two Shiny applications developed for single- and multi-omics data exploration respectively: Quickomics (Bioinformatics, 2021) and xOmicsShiny (Bioinformatics Advances, 2025). To enhance interactive single-cell RNA-seq data analysis, we developed CellxGene Visualization in Plugin, or CellxGene VIP (bioRxiv, 2020), which supports flexible data query and publication-ready figure generation without needing to write code. More recently, we created an LLM-powered artificial intelligence single-cell RNA-seq data exploration tool, CompbioAgent (bioRxiv, 2025). Together, these tools enable the efficient analysis of complex genomic, transcriptomic, and multi-omic datasets, facilitating critical steps in drug discovery, including target identification, validation, and prioritization.

Keywords: Bioinformatics, Multi-omics, Data visualization, RNA-seq, Single-cell RNA-seq, Spatial transcriptomics, AI/ML

Title: Machine learning-based prediction of gene expression using RNA motifs and structural properties

Author list: Mingyi Zhu[1], Wendy Gilbert[1]

Detailed Affiliations:

[1]Department of Molecular Biophysics & Biochemistry, Yale University, New Haven, CT, USA

Abstract: The 5′ untranslated region (5′ UTR) is a major regulatory element of mRNA that plays a central role in controlling translation, yet the determinants of 5′ UTR activity remain incompletely understood. Sequence variation within 5′ UTRs can generate more than 100-fold differences in protein output, but the functional consequences of many natural and disease-associated variants remain unknown. Deep learning offers a promising approach for predicting 5′ UTR activity directly from sequence; however, current models face three major limitations: they are commonly trained on translation datasets containing a mixture of upstream AUG-containing and non-uAUG sequences, which reduces predictive performance for sequences lacking uAUGs; they rely on fixed-length inputs

that do not capture the diversity of human 5′ UTRs; and they often oversimplify RNA secondary structure, failing to represent its dynamic regulatory roles.

Here, we present a developing model trained on direct analysis of ribosome targeting (DART) data, a sensitive high-throughput assay for quantifying 5′ UTR activity. Our model accommodates variable-length sequences and improves prediction accuracy for non-uAUG 5′ UTRs across multiple cell lines. Together, these advances are intended to improve prediction of human 5′ UTR activity and define the sequence and structural features that govern translational efficiency, providing a foundation for future functional studies of regulatory 5′ UTR variants.

Keywords: DART ,5′ UTR, translational regulation, deep learning

Workshop – From Metagenomes to Mechanisms: Functional Microbiome Discovery

August 4th
9:00 AM – 12:00 PM
Room: 2213A

Chairs: Qiyun Zhu, Xiaofang Jiang

Title: Spatiotemporal microbiome dynamics: from early disease prediction to single-cell functional analysis and sorting

Author list: Shi Huang[1]

Detailed Affiliations:
[1]University of Hong Kong, Hong Kong SAR, China

Abstract: The human microbiome exhibits complex spatial and temporal variations that hold immense potential for early disease detection and personalized intervention. However, translating these insights into clinical practice is hindered by challenges in analyzing low-biomass samples, resolving microbial strains, and linking taxonomy to metabolic function. This presentation will showcase our pioneering work in developing integrated methodological frameworks to address these bottlenecks, with a focus on oral diseases.

First, I will introduce our spatiotemporal approach to predicting early childhood caries (ECC). By longitudinally tracking microbiota at single-tooth resolution in over 100 children, we demonstrate that ECC disrupts normal interior-to-posterior ecological gradients, with microbial shifts preceding clinical symptoms. Leveraging these spatiotemporal variations, our Spatial Microbial Index of Caries (sMiC) achieves an AUROC of 0.93 for early ECC prediction. Similarly, we developed the Microbial Index of Gingivitis (MiG) and showed that gingivitis accelerates "oral microbiota age," linking periodontal health to host aging.

Second, to overcome limitations in low-biomass environments (e.g., blood, tissues, FFPE samples) where host DNA dominates, we developed 2bRAD-M, a reduced-genome sequencing method based on type IIB restriction sites. This approach enables cost-effective, quantitative, and highly reproducible microbiome profiling even from degraded DNA (<150 bp). We further extended this to Strain2bFunc for accurate strain-level resolution, demonstrating that oral microbial strains exhibit remarkable host specificity (e.g., Rothia mucilaginosa with Adonis $R^2 = 0.87$), enabling high-resolution host discrimination.

Finally, to move beyond taxonomy to function, we established a single-cell Ramanome platform (scRamanomics). Raman microspectroscopy provides a rapid, label-free "metabolic phenotype" of individual cells. I will present applications including: 1) fast species identification and viability testing of probiotics; 2) functional sorting of ECC-associated bacteria; and 3) characterization of fluoride-resistant cariogenic strains. Using FlowRACS and droplet microfluidics (DCP), we isolated over 100 fluoride-tolerant isolates within one week, revealing heterogeneous metabolic recovery under fluoride and acid stress. This functional enrichment pipeline bridges phenotypic screening with genomic sequencing.

Collectively, our integrated pipeline—from spatiotemporal metagenomics to single-cell functional Ramanomics—provides a powerful framework for precision microbiome diagnostics and therapeutics, significantly advancing early detection and prevention of microbiome-associated diseases.

Keywords: Precision microbiome diagnostics, Spatiotemporal metagenomics, Single-cell Ramanomics, Oral microbiome

Title: Decoding the gut microbiome's functional dark matter via comparative genomics

Author list: Xiaofang Jiang[1]

Detailed Affiliations:

[1]National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Abstract: The human gut microbiome encodes roughly 150 times more genes than the human host, yet most microbial proteins remain uncharacterized. This functional "dark matter," especially among microbial oxidoreductases, limits our understanding of how the microbiota shapes host metabolism, drug responses, and disease risk. Progress has been slowed by the fact that many key gut microbes are difficult to culture, making conventional biochemical screening hard to scale. To overcome this, we developed a comparative genomics framework that links gut microbial genes to enzyme function without requiring organism isolation. By comparing genomes from phenotypically distinct strains, we identify candidate genes and then confirm function through ancestral sequence reconstruction, structural modeling, and targeted experiments. Using this approach, we identified bilR, the bacterial reductase that converts bilirubin to urobilinogen, solving a biochemical mystery that persisted for more than 125 years. BilR, an Old Yellow Enzyme family member, is found across more than 650 gut Firmicutes. It is present in 99.9% of healthy adults but only 22% of neonates under 30 days old, matching the clinical window for neonatal jaundice. We have also used this platform to define six additional enzyme families involved in the metabolism of cholesterol, steroid hormones, and oxidized sugars. Together, these results show that comparative genomics can systematically reveal the gut microbiome's enzymatic repertoire and its clinical relevance.

Keywords: Gut microbiome enzymes, Comparative genomics, Microbial oxidoreductases, Bilirubin metabolism

Title: Microbiome-Informed Therapeutic Peptide Discovery

Author list: Shanlin Ke[1,2], Franz Georg Zingl[3], Matthew K Waldor[3], Yang-Yu Liu[4], Can Chen[5], Quan Zhao[5]

Detailed Affiliations:

[1]Division of Gastroenterology, Hepatology and Nutrition, Department of Internal Medicine, College of Medicine, The Ohio State University Wexner Medical Center, Columbus, OH, USA, [2]The Comprehensive Cancer Center, The Ohio State University, Columbus, OH, USA, [3]Department of Microbiology, Harvard Medical School, Boston, MA, USA, [4]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA, [5]School of Data Science and Society, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Abstract: The human microbiome harbors vast and largely untapped biosynthetic potential, encoding a diverse repertoire of peptides with therapeutic promise. Among these, antimicrobial peptides (AMPs) represent a compelling alternative to traditional antibiotics for combating drug-resistant infections. However, systematically identifying functional peptides for infectious diseases from complex microbial communities remains challenging. Urinary tract infections (UTIs) are common infections that pose a critical burden on healthcare and society. Although the human urinary tract harbors its own microbiome, current knowledge of its composition, functional potential, and alterations in UTI remains limited. Here, we used whole-metagenome shotgun sequencing data together with assembly and binning strategies to construct a microbial catalog from 450 urinary microbiome samples collected in four independent cohorts. An extensive human urinary microbiome catalog consisting of 1,294,428 non-redundant microbial genes and 705 non-redundant metagenome-assembled genomes was constructed. Microbiomes from UTI patients carry significantly more genes linked to antibiotic resistance and virulence vs controls. There was an enrichment of multiple Escherichia strains in UTI patients from two independent case-control cohorts. UTIs are becoming multidrug-resistant, and we used machine learning models to identify potential antimicrobial peptides (AMPs) in these genomes. We further conducted in vitro experiments, demonstrating that two of these AMPs exhibited strong inhibitory activity against uropathogenic Escherichia coli strains. Motivated by limitations encountered in this discovery process, we subsequently developed a systematic benchmarking framework to rigorously evaluate computational methods for peptide discovery. By assessing model robustness across diverse datasets and validation schemes, we establish best practices for improving reproducibility and prioritization of candidate AMPs and anticancer peptides. Together, these works provide a valuable urinary

microbiome resource and establishes both validated therapeutic candidates and a robust computational foundation for microbiome-derived peptide discovery.

Keywords: Urinary microbiome, Antimicrobial peptide discovery, Metagenomic catalog, Machine learning

Title: Microbiome-based early screening of intestinal tumors driven by intelligent sensing and privacy protection

Author list: Yangyang Sun[1], Shunyao Wu[1], Hao Gao[1], Jieqi Xing[1], Jin Zhao[1], Xiaoquan Su[1]

Detailed Affiliations:

[1]College of Computer Science and Technology, Qingdao University, Qingdao, China

Abstract: Non-invasive detection of intestinal tumors represents a critical strategy for reducing the global cancer burden. Although the gut microbiome has demonstrated substantial potential for disease prediction and early screening, current studies remain constrained by inconsistent analytical workflows, pronounced cross-cohort heterogeneity, and fragmented data resources, resulting in unstable model performance and limited clinical translation.

In this study, we systematically evaluated more than 2,000 metagenomic samples from 15 public and in-house cohorts worldwide. Through comprehensive benchmarking of 1,080 analytical workflow combinations, we established MiDx, an optimized analytical framework for microbiome-based intestinal tumor detection. Despite substantial geographic and cohort variability, MiDx successfully captured conserved microbial dynamic patterns associated with colorectal cancer (CRC), achieving an AUROC of 0.83.

However, early detection of precancerous adenomas (ADAs) remains challenging due to the instability of microbial signatures across cohorts. To address this limitation, we developed Meta-iTL, an instance-based transfer learning strategy that leverages prior cohort knowledge to enhance model generalization, improving ADA detection performance in independent cohorts from 0.58 to 0.77.

Recognizing that multi-cohort data integration is further restricted by privacy and data-sharing constraints, we additionally propose FedBio, a microbiome federated learning framework enabling cross-institutional collaborative modeling without sharing raw data. FedBio introduces a global prototype mechanism serving as a semantic anchor to align feature representations across participating centers, effectively reducing technical bias and distributional discrepancies. This approach achieves performance approaching centralized training while substantially outperforming isolated single-cohort models, with improved sensitivity and robustness for precancerous-stage detection.

Collectively, these works advance microbiome-driven intestinal tumor detection across three complementary levels—standardized analytical workflows, transfer learning–based knowledge reuse, and privacy-preserving federated learning. By enabling scalable collaboration while safeguarding data privacy, our framework establishes a practical paradigm for early screening and risk stratification of colorectal cancer and its precancerous stages.

Keywords: Microbiome-based cancer detection, Transfer learning, Federated learning, Colorectal cancer screening

Title: Multi-dimensional virome profiling from bulk metagenomes reveals non-redundant signatures of gut dysbiosis

Author list: Yiyan Yang[1], Dan Huang[1], Scott Weiss[1], Joshua Korzenik[2], Yang-Yu Liu[1], Zheng Sun[1]

Detailed Affiliations:

[1]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA,[2]Resnek Family Center for PSC Research, Crohn's and Colitis Center, Division of Gastroenterology, Hepatology and Endoscopy, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

Abstract: The human gut virome is a critical driver of microbiome ecology, yet its characterization from whole-metagenome sequencing (WMS) data remains severely bottlenecked by overwhelming bacterial DNA backgrounds and the high false-positive rates of standard reference-based profilers. Here, we present VIP2B, an assembly-free, multi-dimensional virome profiling pipeline that overcomes these limitations using an in silico multi-enzyme DNA-to-2Btag strategy paired with a virome-tailored false-positive recognition model. Across rigorous benchmarking, VIP2B achieved state-of-the-art precision and abundance fidelity, uniquely resisting spurious viral calls in host- or bacteria-dominated WMS data. Beyond taxonomic abundance, VIP2B directly extracts genome coverage, functional repertoires, host predictions, and lifestyle phenotypes. By applying VIP2B to 3,685 WMS samples across

20 disease datasets, we demonstrate that these multi-dimensional viral features substantially increase sensitivity to gut dysbiosis and routinely outperform bacteriome-centered predictive models. Furthermore, profiling 6,090 adult metagenomes redefines the baseline concordance between bulk and virus-like-particle sequencing and uncovers two reproducible, health-stratified virome community states. VIP2B provides a highly scalable, high-fidelity framework that establishes the virome as a non-redundant, clinically actionable dimension of microbiome research.

Keywords: Gut virome profiling, Whole-metagenome sequencing, Viral dysbiosis, VIP2B

Title: Predicting metabolite response to dietary intervention using deep learning

Author list: Tong Wang[1], Hannah D. Holscher[2], Sergei Maslov[3], Frank B. Hu[4], Scott T. Weiss[5], Yang-Yu Liu[5]
Detailed Affiliations:
[1]Department of Biological Sciences, Purdue University, [2]Department of Food Science and Human Nutrition, University of Illinois Urbana-Champaign, [3]Department of Bioengineering, University of Illinois Urbana-Champaign, [4]Department of Nutrition, Harvard T.H. Chan School of Public Health, [5]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School

Abstract: Due to highly personalized biological and lifestyle characteristics, different individuals may have different metabolite responses to specific foods and nutrients. In particular, the gut microbiota, a collection of trillions of microorganisms living in the gastrointestinal tract, is highly personalized and plays a key role in the metabolite responses to foods and nutrients. Accurately predicting metabolite responses to dietary interventions based on individuals' gut microbial compositions holds great promise for precision nutrition. Existing prediction methods are typically limited to traditional machine learning models. Deep learning methods dedicated to such tasks are still lacking. Here, we develop a method, McMLP (Metabolite response predictor using coupled Multilayer Perceptrons), to fill in this gap. We provide clear evidence that McMLP outperforms existing methods on both synthetic data generated by the microbial consumer-resource model and real data obtained from six dietary intervention studies. Furthermore, we perform sensitivity analysis of McMLP to infer the tripartite food-microbe-metabolite interactions, which are then validated using the ground-truth (or literature evidence) for synthetic (or real) data, respectively. The presented tool has the potential to inform the design of microbiota-based personalized dietary strategies to achieve precision nutrition.

Keywords: Precision nutrition, Gut microbiome, Metabolite response prediction, Deep learning

Title: Calculating differential genome coverages among metagenomic sources

Author list: Yuhan Weng[1,2], Caitlin Guccione[1,2,3], Daniel McDonald[2], Renee Oles[2,4], Suzanne Devkota[5], Evguenia Kopylova[6], Gregory D. Sepich-Poore[7,8], Rodolfo A. Salido[2], M. Omar Din[2], Se Jin Song[9], Kit Curtius[3,10,11], Hiutung Chu[12,13], Andrew Bartko[2,7,9], Jeff Hasty[7,14], Rob Knight[2,7,9,15,16]
Detailed Affiliations:
[1]Bioinformatics and Systems Biology Program, University of California San Diego, La Jolla, CA, USA, [2]Department of Pediatrics, University of California San Diego, La Jolla, CA, USA, [3]Division of Biomedical Informatics, Department of Medicine, University of California San Diego, La Jolla, CA, USA, [4]Department of Pathology, School of Medicine, University of California San Diego, San Diego, CA, USA, [5]Human Microbiome Research Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA, [6]Clarity Genomics, San Diego, CA, USA, [7]Shu Chien-Gene Lay Department of Bioengineering, University of California San Diego, La Jolla, CA, USA, [8]Feinberg School of Medicine, Northwestern University, Chicago, IL, USA, [9]Center for Microbiome Innovation, University of California San Diego, La Jolla, CA, USA, [10]VA San Diego Healthcare System, San Diego, CA, USA, [11]Moores Cancer Center, University of California San Diego, La Jolla, CA, USA, [12]Department of Pathology, University of California, San Diego, La Jolla, CA, USA, [13]Chiba University-UC San Diego Center for Mucosal Immunology, Allergy and Vaccines (cMAV), University of California, San Diego, La Jolla, CA, USA, [14]Department of Molecular Biology, Division of Biological Science, University of California San Diego, La Jolla, CA, USA, [15]Halıcıoğlu Data Science Institute, University of California San Diego, La Jolla, CA, USA, [16]Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA

Abstract: Interpreting metagenomic data often depends on summary statistics that do not fully capture how sequencing signals are distributed across reference genomes. Coverage breadth provides an additional perspective by describing the fraction of a reference genome covered by at least one read, making it useful for comparing microbial signals across samples, groups, and genomic regions. We present micov, a method for scalable analysis

of genome coverage breadth in metagenomic datasets. micov supports per-sample coverage quantification, comparison across metadata-defined groups, and visualization of genome-wide coverage patterns. In this presentation, I will introduce the general framework implemented in micov and discuss how these analyses can be used to support interpretation, comparison, and hypothesis generation in metagenomic studies.

Keywords: Metagenomic coverage breadth, Genome-wide coverage analysis, Microbial community profiling

Title: Enabling AI-driven discovery in complex biological systems with scikit-bio

Author list: Matthew Aton[1], Daniel McDonald[2], Rob Knight[2], James Morton[3], Qiyun Zhu[1]

Detailed Affiliations:

[1]Arizona State University, Tempe, AZ, USA,[2]University of California San Diego, La Jolla, CA, USA,[3]Gutz Analytics, New York, NY, USA

Abstract: Biological "omics" data are high-dimensional, sparse, and compositional, and are structured by underlying biological relationships such as phylogeny, functional hierarchies, and ecological interactions. Without computational infrastructure that explicitly encodes these properties, AI systems risk learning patterns and signatures that are biologically inconsistent or difficult to interpret and reproduce. We present scikit-bio, an open-source Python library developed over the past decade as a biologically informed computational foundation that bridges raw omics data and AI-driven workflows. Rather than providing end-to-end supervised learning models, scikit-bio focuses on supplying the data structures, transformations, and statistical tests that AI systems rely on when operating on complex biological data. Recent development has significantly expanded scikit-bio's capabilities to support AI-enabled discovery. We have implemented mmvec, a neural network framework for microbe-metabolite association modeling, with a redesigned kernel that runs efficiently on CPU. We have implemented high-performance statistical methods for differential abundance testing, such as ANCOM-BC, to enable feature selection and complement interpretable models within cross-validation workflows. We have enriched and upgraded compositional data transformation methods to support GPU and arbitrarily ranked tensors, allowing them to be embedded directly into deep learning pipelines. We have enhanced the classic PCoA and PERMANOVA algorithms to enable much accelerated GPU execution. At the same time, scikit-bio's core data structures have been modernized to adopt emerging Python array standards, enabling interoperability with PyTorch, JAX, CuPy, and other AI-oriented libraries. By providing trusted biological abstractions, such as compositional representations, phylogenetically informed distances, and multivariate embeddings, scikit-bio enables AI systems to engage more effectively with the complexity of living systems and supporting the next generation of predictive, automated biological discovery.

Keywords: scikit-bio, Compositional omics analysis, Biologically informed AI, High-performance bioinformatics

Workshop – AI and Data Science in Health Informatics
August 4th
9:00 AM – 12:00 PM
Room: 2213B

Chairs: Yu Huang, Yan Zhuang

Title: Agentic AI for complex clinical informatics workflows

Author list: Xing He[1,2]

Detailed Affiliations:

[1]Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN, USA, [2]Regenstrief Institute, Indianapolis, IN, USA

Abstract: Many clinical informatics tasks are challenging not because any single step is difficult, but because they require coordinating many interdependent steps, each involving domain judgment that depends on the outcome of the last. The knowledge that makes experts effective is largely procedural: knowing how to decompose a problem, which resources to consult, how to handle ambiguity or gaps along the way, and when the result is good enough to

finalize. This kind of process knowledge is central to reliable performance, yet it is seldom formalized—it lives in practitioners' heads, is learned through apprenticeship, and must be re-created by every new team. Current automation tends to target isolated subtasks rather than the end-to-end workflows that practitioners actually carry out.

Recent advances in agentic AI offer a way to close this gap. Unlike single-pass generation or fixed pipelines, agentic systems—built on large language models with tool use—can decompose complex tasks, invoke external resources at each decision point, maintain working memory of what has been tried and what remains unresolved, and revise their approach based on intermediate results. Critically, these agents can be guided: domain-specific tools ground their reasoning in authoritative sources, and structured procedural knowledge—what we call skills—channels their behavior along expert-informed strategies. The combination of tools, skills, and structured intermediate state turns an LLM from a general reasoner into a system that can execute expert-level workflows with transparency and auditability.

We explore how this paradigm applies to clinical informatics through three tasks. In chart review, agents navigate clinical records to extract structured findings with provenance, uncertainty handling, and explicit stopping criteria. In ontology development, agents align heterogeneous data sources to community standards by generating executable mapping artifacts and identifying coverage gaps. In clinical code set construction, agents translate natural-language cohort definitions into standardized concept sets using vocabulary tools and authoring skills. Across these cases, evaluation isolates the contributions of tool access, skill guidance, and agentic orchestration. Our preliminary results show that performance is driven primarily by domain-specific tools and explicit workflow knowledge rather than by parametric reasoning alone, suggesting that the tools-plus-skills paradigm can make complex clinical informatics workflows more reproducible, scalable, and accessible.

Keywords: Large Language Models, Agentic AI

Title: Multi-modal AI using neuroimaging genomics to study alzheimer's and dementia

Author list: Da Ma[1]

Detailed Affiliations:

[1]Department of Internal Medicine Section of Gerontology and Geriatric Medicine, Wake Forest University School of Medicine, Winston-Salem, NC, USA

Abstract: TThis talk will focus on multi-modal artificial intelligence (AI) methods for neuroimaging and genomics to improve the study of brain aging, Alzheimer's disease, and related dementias (ADRD). Alzheimer's disease develops over decades before clinical symptoms emerge, creating a critical need for tools that can detect risk earlier, characterize disease heterogeneity, and support individualized prediction. The research presented highlights efforts to improve the clinical translatability of AI for ADRD by emphasizing generalizability, explainability, longitudinal disease modeling, and biologically informed multimodal integration. This talk will feature several representative projects to address several challenges in the field of AI applications in ADRC, including: neuroimaging-based brain aging, explainable cortical graph neural networks to enhance dementia risk cortical atrophy pattern; spatial-temporal deep learning that utilize longitudinal neuroimaging data and deep survival analysis to improve the risk prediction of dementia onset time; generative AI approaches for counterfactual brain MRI synthesis to address data scarcity, demographic imbalance, and limited longitudinal coverage in neuroimaging datasets; deep-learning-based polygenic dementia risk modeling utilizing multiomic-based explainability to uncover disease etiology and to guide personalized disease prevention strategy. The talk will conclude with multimodal fusion strategies that integrate neuroimaging and genomic data to improve risk stratification and prediction, with an emphasis on explainability and robustness to missing modality.

Keywords: AI for ADRD, Multi-modal Modeling, Neuroimaging, Genomics

Title: Immune digital twin and vascular disease

Author list: Juilee Thakar[1,2]

Detailed Affiliations:

[1]Department of Microbiology and Immunology, Department of Biomedical Genetics, [2]Translational Biomedical Sciences graduate program, University of Rochester, Rochester, NY, USA

Abstract: The development of immune digital twins (IDTs) represents a critical step toward personalized medicine and *in silico* clinical trials. However, most current efforts remain at the level of conceptual frameworks or small-scale demonstrations. In this talk I will discuss the development of Digital Twin-ready datasets and a dynamic modeling framework. The dynamic modeling framework developed using single-cell RNA-seq data can reveal perturbation-informed analyses of disease phenotypes. We employ Boolean discrete-state modeling to define network-based State Transition Graphs (STGs) that capture intracellular signaling dynamics and identify dominant attractors corresponding to stable cellular states. By integrating these dynamic models with pseudotime inference, we reconstruct disease-specific trajectories and estimate transition probabilities between cellular states. To enable patient-specific characterization, we introduce clustering metrics that combine attractor similarity, proximity in reduced-dimensional space, and alignment with clinical phenotypes (e.g., AS+ vs AS−). Overall, this work advances a scalable and interpretable approach for modeling immune system dynamics.
Keywords: Digital twins, Dynamic Modeling

Title: Unlocking clinical development innovation with artificial intelligence and real-world evidence

Author list: Jinhai Stephen Huo[1], Chuan Gao[1]

Detailed Affiliations:

[1] Johnson & Johnson Innovative Medicine, Research & Development, Data Science and Digital Health, Real-World Evidence, New Brunswick, NJ, USA

Abstract: Artificial intelligence (AI) and Real-World Evidence (RWE) are fundamentally reshaping the clinical development paradigm in pharmaceutical research and development, enabling a transition from conventional, protocol-driven approaches toward more adaptive, data-informed, agile clinical decision-making processes. Advances in digital infrastructure, analytical methodologies, and evolving regulatory frameworks further support these enhancements, creating opportunities to improve trial design, accelerate development cycles, and ultimately lead to better patient outcomes.

The influences of AI and RWE on clinical development are throughout its entire lifecycle, from early drug discovery to late-stage post-marketing surveillance. Leveraging real-world multimodal datasets - including clinical, multi-omics, imaging, and digital health data - not only facilitates a more granular understanding of disease progression and treatment response but also allows for the identification of novel biomarkers and precise patient subgroups. Emerging methodologies such as digital twins and predictive simulations are revolutionizing the way clinical trials are designed and executed, offering the ability to augment traditional studies with in-silico models that simulate not only patient outcomes but trial operation scenarios. Advances in causal inference methodologies and theoretical frameworks further strengthen the ability to derive robust, decision-grade evidence from non-randomized real-world data sources, providing confidence in the validity of findings generated outside of conventional randomized controlled trials. In parallel, the development of AI-empowered novel endpoints derived from real-world data, include digital biomarkers, patient-reported outcomes captured through wearable devices, or composite measures that reflect the complexity of patient health, is expanding the definition of clinically meaningful outcomes and enhancing the relevance of clinical research to real-world patient experiences. Additionally, through innate capability of orchestrating multi-step analytic workflows, agentic AI offers the potential to further streamline analysis execution and reporting for all these processes.

Despite these advances, significant challenges remain. Regulatory acceptance continues to evolve, with persistent concerns around data quality, bias, and AI model transparency. Ensuring methodological rigor, explainability, and traceability is critical to building trust among regulators and clinicians. A practical approach that balances innovation with scientific rigor and delineates a roadmap for the scalable and sustainable integration of AI and real-world evidence is essential to enhance clinical development and making it more adaptive, patient-centric, and grounded in robust evidence.

Keywords: Artificial intelligence, Real-world evidence, Clinical development, Pharmaceutical, Trial design, Patient outcomes

Title: Agentic AI for automated RNA structural motif analysis and biomedical discovery

Author list: Jie Hou[1]

Detailed Affiliations:

[1] Department of Computational Mathematics, Science and Engineering, Michigan State University, East Lansing, MI, USA

Abstract: RNA structural motifs, such as internal loops, bulges, and hairpins, encode fundamental principles of RNA folding and molecular function directly relevant to drug target discovery and understanding disease-associated non-coding RNAs. However, existing motif analysis pipelines demand substantial computational expertise, creating barriers for experimental researchers and limiting the scale at which structural data can be systematically explored. We present ARSMA (Agentic AI Framework for RNA Structural Motif Analysis), a framework that leverages large language model (LLM)-powered autonomous agents to automate complex RNA structural analysis workflows while preserving transparency and reproducibility.

ARSMA implements end-to-end pipelines for large-scale motif extraction, redundancy removal, and structural characterization from crystallographic, NMR, and cryo-EM structures. The framework generates multi-level, machine-learning-ready feature representations including torsion angles, sugar puckers, stacking geometries, hydrogen-bonding patterns, and pairwise distance matrices. Applied to over 175,000 motifs from the CoSSMos database, ARSMA enables systematic clustering and classification at a scale previously impractical through manual analysis. A complementary sequence-centric pipeline leverages pre-trained RNA language models to derive biologically interpretable features, enabling motif identification without requiring experimentally resolved structures.

Central to ARSMA is a multi-agent orchestration layer in which specialized AI agents collaborate to automate tool selection, execution, and result interpretation. This agentic architecture reduces technical barriers while maintaining the rigor required for reproducible research. This talk will present the ARSMA framework, demonstrate its application to RNA motif characterization, and discuss how agentic AI architectures can serve as scalable models for AI-assisted scientific workflows in biomedical research, lowering barriers to computational discovery while addressing considerations of interpretability, validation, and responsible deployment.

Keywords: Agentic AI, RNA Structural Motifs, Structural Bioinformatics, Large Language Models, Automated Workflows, Biomedical Discover

Title: LLM-Agent-Based clinical trial cohort building: an experimental evaluation using MIMIC-IV

Author list: Hao Liu[1]

Detailed Affiliations:

[1] School of Computing, Montclair State University, Montclair, NJ, USA

Abstract: Clinical trial recruitment remains a critical bottleneck in translational medicine, with manual eligibility screening of electronic health records (EHR) being labor-intensive, error-prone, and poorly scalable. Large language models (LLMs) offer a promising path toward automated eligibility matching, yet fundamental questions about their reliability in this high-stakes clinical setting remain unanswered.

We present CTCB (Clinical Trial Cohort Building), an agent-driven system for automated clinical trial eligibility assessment. CTCB implements a 7-phase orchestrated pipeline that combines LLM reasoning with deterministic logic. Phase 0 parses trial criteria and classifies each criterion by determinability: structured-data-evaluable (S1), notes-dependent (S2), hybrid (S3), or retrospectively undeterminable (S0). Phases 1-2 use LLM agents to discover candidate patients and evaluate structured criteria (labs, vitals, diagnoses, medications) via dynamic SQL generation against MIMIC-IV. Phase 3 scores candidates by data richness without any LLM involvement. Phase 4 deploys parallel LLM agents to evaluate notes-dependent criteria using ChromaDB-indexed clinical notes with multi-strategy retrieval. Phase 5 applies deterministic boolean logic combining all prior evaluations to produce final eligibility determinations (ELIGIBLE, EXCLUDED, or UNDETERMINED), and Phase 6 generates a human-readable cohort report.

To systematically evaluate the impact of model choice and inference temperature on cohort building reliability, we designed a multi-factor experiment across two disease domains: sepsis/septic shock (10 completed Phase 2/3 clinical trials) and acute kidney injury (10 trials), totaling 20 trials sourced from ClinicalTrials.gov with 7 to 27

eligibility criteria each. We evaluate multiple LLMs served locally via vLLM at three temperature settings (0.1, 0.3, 0.7), with three repetitions per condition for reproducibility measurement. This design enables analysis of model-temperature interactions and domain-dependent performance patterns.
Our preliminary findings reveal that temperature sensitivity varies substantially across pipeline phases, with criteria parsing (Phase 0) and structured evaluation (Phase 2) showing high consistency at low temperatures, while notes-based evaluation (Phase 4) exhibits greater model-dependent variability. This work provides the first systematic, multi-model benchmark for LLM-based clinical trial eligibility matching.
Keywords: Large Language Models, Clinical Trial Cohort Building

Title: A scalable multi-omics representation framework for disease modeling across proteomic, single-cell, and spatial transcriptomic data
Author list: Daisy Shen[1], Tingyi Li[1], Kenneth Tsai[1], Roger Li[1], Xuefeng Wang[1], Xiaoqing Yu[1]
Detailed Affiliations:
[1] Department of Biostatistics and Bioinformatics, Moffitt Cancer Center, Tampa, FL, USA
Abstract: The rapid accumulation of large-scale multi-omics datasets across diseases necessitates computational frameworks capable of integrating molecular dynamics, cellular heterogeneity, and spatial organization. Here, we present a scalable and generalizable framework for multi-modal data integration and disease modeling across proteomics, single-cell transcriptomics, spatial transcriptomics, and bulk sequencing data. We first introduce a trajectory-based representation learning approach to model dynamic molecular changes from cross-sectional proteomics data. Applied to cerebrospinal fluid proteomics in autosomal dominant Alzheimer's disease (ADAD), our framework infers pseudo-temporal protein trajectories relative to disease onset, identifying stage-specific molecular programs spanning early metabolic dysregulation, neuronal injury, and immune activation. This representation further supports machine learning–based biomarker discovery, enabling robust prediction of disease status beyond conventional markers. To generalize across tissues and diseases, we constructed a large-scale single-cell reference atlas by integrating multi-cohort scRNA-seq datasets in solid tumors, including head and neck cancer. This atlas defines high-resolution cellular representations of malignant and microenvironmental states, which are projected onto bulk transcriptomic cohorts via Bayesian modeling to enable population-scale inference of tumor composition and clinically relevant programs. We further extend this framework to spatial transcriptomics by generating, to our knowledge, one of the largest single-cell–resolution spatial transcriptomic datasets using Xenium profiling in Merkel cell carcinoma. By integrating graph-based spatial representation learning, we model cellular neighborhoods, interaction networks, and spatial centrality, uncovering structured tumor–immune niches that are not detectable in dissociated data. By unifying trajectory inference, cellular state representation, and spatial graph modeling into a shared computational framework, our approach enables cross-modal and cross-disease integration of multi-dimensional omics data. This work establishes a scalable foundation for learning biologically meaningful representations from complex biomedical data and provides a powerful paradigm for disease modeling, biomarker discovery, and precision medicine.
Keywords: Multi-omics integration, Disease trajectory modeling, Single-cell and spatial transcriptomics, Precision medicine

Title: Cross-Hospital privacy-preserving distributed AI
Author list: Tsung-Ting Kuo[1]
Detailed Affiliations:
[1]Biomedical Informatics and Data Science and of Surgery, Yale University, New Haven, CT, USA
Abstract: To improve the predictive capability of clinical AI approaches and to identify medication-outcome associations for diseases and conditions, it is important to integrate data "horizontally" (i.e., when patients have the same types of data from different hospitals) or "vertically" (i.e., when patients has some types of clinical data in one hospital and some types elsewhere), from multiple healthcare systems. In this presentation, we will introduce a decentralized cross-institutional AI method with the following desirable technical features: (1) including both horizontally- and vertically-partitioned data without patient-level record dissemination; (2) providing a distributed and computationally-fair AI modeling across hospitals; (3) utilizing centralization-equivalent algorithms and community-supported distributed platforms with blockchain; and (4) building trustworthy, transparent and scalable predictive models. Our approach can help the use of data from multiple institutions and potentially address ethical

issues associated with AI models. As a decentralized cross-institutional algorithm, our method can support collaborative privacy-preserving AI modeling across multiple healthcare systems with horizontally- or vertically partitioned data.

Keywords: Federated learning, Privacy-preserving clinical AI, Cross-institutional data integration, Decentralized healthcare modeling

Workshop – Integrative omics to understand and predict disease progression

August 4th
9:00 AM – 12:00 PM
Room: 1220

Chairs: Jingwen Yan, Sha Cao

Title: Continuous Modeling of Disease Progression in Alzheimer's Disease and Related Dementias

Author list: Lu Zhang[1]

Detailed Affiliations:
[1]Department of Computer Science, Indiana University Indianapolis, Indianapolis, IN, USA

Abstract: Alzheimer's disease and related dementias (ADRD) are characterized by substantial heterogeneity and continuously evolving disease trajectories, posing significant challenges for accurately characterizing disease progression. Conventional approaches typically assign patients to discrete disease stages, which often fail to capture the gradual, dynamic, and individualized nature of disease evolution. In this talk, I will present our recent work on continuous modeling of disease progression in ADRD through disease representation learning. Instead of treating disease stages as isolated categories, our framework learns a continuous disease embedding space that encodes the intrinsic relationships among clinical stages and organizes them into an interpretable disease progression trajectory. This representation enables individualized characterization of disease status across the Alzheimer's disease continuum while preserving patient heterogeneity beyond conventional stage labels, providing a more faithful representation of disease progression and offering an interpretable view of disease evolution. Finally, I will highlight how continuous disease progression modeling can serve as a general computational framework for studying progressive neurological disorders and discuss its potential applications in disease monitoring, prognosis, and precision medicine.

Keywords: Alzheimer Disease, Neuroimaging, Disease progression

Title: Stability-aware integrative clustering for characterizing disease progression from multi-modal data

Author list: Rachael Hageman Blair[1]

Detailed Affiliations:
[1]Department of Biostatistics, New York University at Buffalo, Buffalo, NY, USA

Abstract: Understanding disease progression from integrative omics data is challenging due to heterogeneous signal, lack of ground truth, and sensitivity of clustering methods to analytic choices. We present Stability-based Partitions for Ensemble Clustering (SPEC), a graph-based framework that integrates clustering solutions across algorithms, parameter settings, and data modalities to identify reproducible patient subgroups. SPEC estimates stability via subsampling at both the observation and partition levels, encoding these measures within a weighted bipartite graph linking samples to cluster assignments. Community detection on this graph yields consensus clusters that emphasize stable structure while down-weighting spurious patterns . We apply SPEC to the METABRIC breast cancer cohort, integrating clinical variables, gene expression, and mutation profiles within a unified framework. The resulting clusters exhibit significantly different survival trajectories, capturing clinically meaningful patterns of disease progression and highlighting subgroups with distinct risk profiles. By prioritizing reproducible structure across modalities, SPEC provides a robust and flexible approach for integrative omics analysis, with potential applications in biomarker discovery, patient stratification, and understanding the dynamics of disease progression.

Keywords: Multi-omics, Ensemble Clustering, Patient Stratification

Title: Continuous staging of amyloid progression in Alzheimer's disease with enhanced sensitivity to early changes

Author list: Jingwen Yan[1]

Detailed Affiliations:

[1]Department of Biomedical Engineering and Informatics, Indiana University Indianapolis, Indianapolis, IN, USA

Abstract: Understanding of early Alzheimer progression is critical for timely diagnosis and treatment evaluation, but traditional diagnostic groups often lack sensitivity to subtle early-stage changes. We developed an unsupervised dimensionality reduction method that models the amyloid progression in AD on a continuous scale which preserves the temporal sequence of follow-up visits. Applied to longitudinal amyloid PET data, it derived staging scores with better preserved temporal consistency across diagnostic groups and longitudinal follow-up visits and additionally demonstrated robust generalization to held-out subjects. The progression trajectory revealed biologically consistent amyloid spreading patterns and greater sensitivity to early progression than global amyloid SUVR. Taken together, the new staging score provides a continuous quantification of amyloid pathology that complements global amyloid measures by capturing early localized progression.

Keywords: Alzheimer's disease, Amyloid progression modeling, Early Detection

Title: AD-LLaVA-3D: Forecasting Alzheimer's Disease Progression with a Unified Multimodal Video-Language Framework

Author list: Sha Cao[1]

Detailed Affiliations:

[1]Department of Biomedical Engineering, Oregon Health & Science University, Portland, OR, USA

Abstract: Accurate forecasting of Alzheimer's disease (AD) progression is vital for personalized clinical management, yet integrating high-dimensional 3D MRI with sparse, irregular clinical records remains a significant challenge. We developed AD-LLaVA-3D, a unified multimodal framework that conceptualizes 3D MRI volumes as temporal video sequences and integrates them with longitudinal clinical history via semantic prompting. The model was trained on the ADNI cohort and evaluated against multimodal, imaging-only, and traditional machine learning baselines. AD-LLaVA-3D significantly outperformed all benchmarks, achieving a peak R^2 of 0.68 for CDR-SB forecasting on the ADNI test set. Notably, the framework demonstrated exceptional zero-shot generalization on the external OASIS-3 cohort (R^2=0.82) and maintained robust predictive stability over extended 48-month horizons, where traditional models suffered significant performance degradation. By synthesizing structural neuroimaging with semantic clinical context, AD-LLaVA-3D provides a scalable, high-fidelity tool for longitudinal risk stratification and real-world clinical decision support.

Keywords: Alzheimer's disease, longitudinal prediction, multi-modal integration

Title: Staging and Pseudotime Inference of Alzheimer's Disease Progression Using Multimodal Imaging Data

Author list: Zixuan Wen[1]

Detailed Affiliations:

[1]Department of Applied Mathematics and Computational Science, University of Pennsylvania, Philadelphia, PA, USA

Abstract: Alzheimer's disease (AD) is a neurodegenerative disorder marked by progressive cognitive decline and the accumulation of amyloid-beta, tau, and neurodegeneration. We analyze amyloid PET, tau PET, and structural MRI from the ADNI dataset to characterize continuous and discrete patterns of AD progression. We use PHATE to identify low-dimensional disease trajectories and derive pseudotime with Slingshot. In parallel, we apply SuStaIn to infer biomarker event sequences and assign individuals to discrete disease stages. PHATE-derived pseudotime strongly corresponds with SuStaIn stages, supporting temporal consistency across the two frameworks. SuStaIn also identifies distinct, modality-specific biomarker sequences consistent with established AD models. We validate these patterns using pseudotime-aligned density models and clinical measures. Together, these results support an integrative framework for studying AD progression. Future work will anchor pseudotime to clinical timelines and validate predicted biomarker cascades longitudinally.

Keywords: Alzheimer's disease, disease progression, multi-modal imaging

Title: From 2D Snapshots to 3D Molecular Landscapes: Reconstructing Volumetric Spatial Transcriptomics with AI

Author list: Wei Zhang[1]

Detailed Affiliations:

[1]Department of Computer Science, University of Central Florida, Orlando, FL, USA

Abstract: Spatial transcriptomics has transformed our ability to characterize gene expression within intact tissues, yet current technologies remain largely limited to 2D tissue sections. Reconstructing volumetric spatial transcriptomics from these 2D measurements presents a fundamental computational challenge that requires advances in computer vision, machine learning, and spatial modeling. In this talk, I will present our ongoing research on developing AI methods for volumetric spatial transcriptomics. The presentation will cover computational frameworks for multimodal image registration, three-dimensional gene expression reconstruction, and spatial representation learning. I will discuss how deep learning can leverage complementary histological and molecular information to infer continuous 3D molecular landscapes from serial tissue sections, as well as future directions toward foundation models for volumetric spatial omics. This work aims to provide a general computational framework for enabling next-generation 3D molecular tissue analysis.

Keywords: Spatial Transcriptomics;3D Reconstruction; Artificial Intelligence; Volumetric Imaging.

Title: How Much Biological Knowledge Is Encoded in LLM-Derived Gene Embeddings?

Author list: Jun Li[1]

Detailed Affiliations:

[1]Department of Applied and Computational Mathematics and Statistics, Notre Dame University, Notre Dame, Indiana, USA

Abstract: LLM-derived gene embeddings constructed from brief NCBI gene descriptions have achieved strong performance in a range of biological applications, yet the nature of the information they encode remains poorly understood. Using a GSEA-based framework, we show that OpenAI's embeddings recover the vast majority of Hallmark and C2 pathways. Remarkably, embeddings generated from gene symbols alone retain substantial pathway information, suggesting that pretrained language models encode prior biological knowledge beyond the information provided in the input text. A comparison of eleven smaller models further demonstrates that even limited textual context is sufficient for all models to achieve similar pathway coverage. These results show that LLM-derived gene embeddings capture considerably richer biological information than their sparse textual inputs would imply.

Keywords: Gene embeddings, Language models, Biological pathways

Title: Toward integrative modeling of drug response in acute myeloid leukemia

Author list: Olga Nikolova[1]

Detailed Affiliations:

[1]Division of Oncological Sciences, Oregon Health & Science University, Portland, Oregon, USA

Abstract: Acute myeloid leukemia (AML) is an aggressive cancer, with roughly 21,000 new cases and over 10,000 related deaths reported annually in the United States. Although AML often initially responds to therapy, relapse is frequent, resulting in a very high mortality rate. AML is genetically heterogeneous, and although some targeted drugs have been developed for recurrent mutations, resistance is still prevalent. Resistance arises through an interplay between tumor intrinsic responses, microenvironment signals, and cell state plasticity. To address these challenges, our lab develops integrative probabilistic models that bridge mechanistic understanding of cell-state plasticity with genome-driven cellular processes by integrating information across molecular modalities, patient cohorts, and scale – from bulk to single-cell resolution. This talk will highlight our recent methodologies that integrate multi-omic and functional data across the largest AML patient cohorts to predict drug response and identify biologically grounded latent space representations that drive differential therapeutic response. It will also outline exciting new directions involving immune programs and the bone marrow microenvironment newly associated with cell state driven drug response.

Keywords: Acute myeloid leukemia; integrative omics; therapeutic response

Workshop – Multi-Omics, Microbiome, and Metabolomics
August 4th
1:40 – 5:20 PM
Room: 2220A&B

Chair: Hao Liu

Title: Multi-modality foundation model to link pathology with omics

Author list: Guangyu Wang[1]

Detailed Affiliations:
[1]Houston Methodist, Weill Cornell Medicine School.

Abstract: TBD

Keywords: TBD

Title: Analyzing Soil Microbial Diversity in Organically Managed Fields Amended with Dredged Material in Northwest Ohio

Author list: Shikshya Gautam[1], Mira Luna Ebersole[1], Jyotshana Gautam[1], Carlos D Soto[2], Pei Wang[3], Angélica Vázquez-Ortega[2], Zhaohui Xu[1]

Detailed Affiliations:
[1]Department of Biological Sciences, School of Earth Environment and Society, Department of Business Analytics, Economics, and Information Systems, Bowling Green State University, Bowling Green, OH 43403, USA.

Abstract: Each year, around 1.5 million tons of sediments are excavated from the federal navigation channels at Toledo Harbor, which used to be disposed in an open lake placement area in Lake Erie. After the ban on open water disposal, utilizing this dredged material (DM) as a soil amendment became an attractive approach to managing DM disposal issues. Previous research has shown the DM can improve soil physiochemical properties in greenhouse settings (Brigham et al., 2021; Rúa et al., 2023; Gautam et al., 2024; Gautam et al., 2025). This study aimed to characterize the taxonomic diversity of microbial communities in the fields amended with DM, using 16S rRNA analysis. Among the three organically managed fields, plot 1 and plot 3 received 40 tons per acre and 80 tons per acre of DM, respectively, and plot 2 received no DM (control treatment). Alpha diversity metrics (Shannon index) did not show significant differences among the plots. ($p > 0.05$). This suggests that the microbial richness and evenness remained relatively stable regardless of the addition of DM. Principal component analysis (PCoA) based on Bray-Curtis dissimilarity revealed clear separation of microbial communities among the three treatments. Each treatment formed a distinct cluster in the emperor plot, indicating compositional differences in microbial communities. PICRUSt2 based functional predictions further revealed variation in genes associated with phosphorus, nitrogen, and carbon cycling. Notably, phosphorus-cycling genes (phoD and phytase) were enriched in DM amended soils and showed positive associations with soil calcium, suggesting enhanced microbial capacity for organic phosphorus mineralization and solubilization under Ca-enriched conditions. This suggests that while DM influenced microbial structure, functional potential was regulated over time, reflecting microbial adaptation to evolving soil conditions, highlighting the importance of microbial community management for sustainable agricultural productivity and environmental health.

Keywords: Dredged Material/Sediment, Lake Erie, Soil Health, Soil Microbiome, Microbial diversity

Title: DUET-Knockoffs: FDR-Controlled Cross-Platform Feature Selection from Paired 16S and Shotgun Microbiome Count Data

Author list: Yiqiao Zhu,[1], Shiyuan Wang,[1] Qiwei Zhang,[1] Yong Ge[3] and Yuxuan Du[1,*]

Detailed Affiliations:
[1]Department of Electrical Engineering, The University of Texas at San Antonio, One UTSA Circle, 78249, TX, USA, [2]Department of Biostatistics, University of Michigan, Ann Arbor, 500 S State St, 48109, MI, USA,

[3]Department of Microbiology, Immunology & Molecular Genetics, University of Texas Health San Antonio, 7703 Floyd Curl Drive, 78229, TX, USA. *Corresponding author
Abstract: Motivation: Paired 16S rRNA gene amplicon and whole-genome shotgun metagenomic sequencing are increasingly collected on the same samples, providing complementary but imperfectly aligned views of microbial communities. However, most feature-selection methods are designed for a single sequencing platform. In practice, 16S and shotgun data are often analyzed separately and compared post hoc, which provides no formal error control for identifying taxa with reproducible evidence across platforms. There remains a need for statistical methods that jointly analyze paired microbiome count profiles while accounting for sequencing depth, sparsity, platform-specific bias, and cross-taxon dependence. Results: We propose DUET-Knockoffs, a knockoff-based framework for cross-platform feature selection from paired 16S and shotgun microbiome count data. For each taxon, DUET-Knockoffs targets the union-null hypothesis that the taxon is conditionally independent of the outcome in at least one sequencing platform. Rejection of this null prioritizes taxa with evidence of conditional association in both platforms. Within each platform, taxon counts are modeled using zero-inflated negative binomial marginal distributions with library-size adjustment and observed covariates, coupled through a Gaussian copula to approximate cross-taxon dependence. The fitted feature model is used to generate outcome free synthetic null count profiles, which serve as reference profiles for original-versus-synthetic contrast statistics. Platform-specific contrasts are then combined using a simultaneous-knockoff aggregation statistic, yielding a single selection statistic per taxon. Under the fitted synthetic-null feature model and the symmetry assumptions of the simultaneous knockoff framework, the resulting procedure controls the false discovery rate (FDR) for the cross-platform union-null target. In simulations calibrated to real microbiome count profiles, DUET-Knockoffs maintains empirical FDR control and achieves higher power than representative single-platform and integrative competitors across settings with varying community heterogeneity, overdispersion, structural zeros, platform bias, and covariate confounding. Applications to an oral microbiome study of prediabetes and a non-human primate gut microbiome study of dietary niche prioritize biologically plausible, count-scale cross-platform candidate genera that are not consistently recovered by single-platform analyses. Availability and Implementation: The DUET-Knockoffs software is publicly available at https://github.com/dyxstat/DUET-Knockoffs.
Keywords: Microbiome; 16S rRNA; Shotgun metagenomic sequencing; Feature selection; FDR control

Title: Multiomic Integration Reveals Regulatory Function of GWAS Variants within Transposable Elements in Brain Aging

Author list: Sreejato Chatterjee[1], Shuangshuang Feng[2], Dean Zhang[2] and Hongbo Liu[2*]

Detailed Affiliations:
[1]Department of Biology, University of Rochester, RC Box 270211, 14627, NY, USA,[2]Department of Biomedical Genetics & Aging Institute, University of Rochester School of Medicine & Dentistry, 601 Elmwood Avenue, 14642, NY, USA.
Abstract: Genome-wide association studies (GWAS) have identified thousands of variants associated with brain aging, yet most lie in noncoding regions, complicating functional interpretation because their effects must be inferred through regulatory mechanisms. In the brain, where long-lived cells rely on stable epigenetic regulation to maintain transcriptional homeostasis, transposable elements (TEs) represent a major component of the regulatory DNA landscape and a potential substrate through which noncoding variation may be interpreted. However, it remains unclear whether and how brain-aging GWAS variants within specific TE sequences are linked to biological pathways relevant to brain aging. To fill this gap, we applied an integrative genomic framework combining brain aging GWAS, TE annotations, single-nucleus ATAC-seq chromatin accessibility maps across major brain cell types, ENCODE candidate cis-regulatory elements (cCREs), HOMER motif enrichment analysis, and partitioned LD Score Regression. We found that nearly half of the brain aging-related GWAS variants were located within transposable element sequences. Among these TE families, the SINE elements, especially Alu, show the strongest genomic enrichment for these brain aging-related variants. Brain-aging GWAS variants overlapping TE sequences are preferentially localized within transcription factor-associated cCREs and depleted from promoter-like regions. Motif enrichment within SINE-associated cCREs revealed modest but reproducible signatures involving nuclear receptor (AR-halfsite), immune-associated (PU.1:IRF8), and chromatin-related (RFX1) motifs, suggesting heterogeneous regulatory programs rather than a unified TE-specific enhancer grammar. Together, these results distinguish between genomic localization, regulatory activity, sequence-level regulatory signatures, and heritability

contribution, and support a model in which brain-aging GWAS variants localized within transposable element sequences reflect underlying epigenetic regulatory architecture rather than independent TE-specific genetic effects.

Keywords: genome-wide association study, brain aging, transposable elements, epigenomics, chromatin accessibility, LD score regression, ENCODE cCREs, single-nucleus ATAC-seq, Alu elements, heritability partitioning

Title: metaSRNA: A Unified Framework for Reprocessing Bacterial Data from microRNA Sequencing Protocols and Drafting Bacterial Small RNA Catalogs

Author list: Mengyuan Zhou[1], Peerzada Tajamul Mumtaz[2], Janos Zempleni[2], Qiuming Yao[1,3]

Detailed Affiliations:
[1]School of Computing, University of Nebraska Lincoln, 256 Avery Hall, Lincoln, NE, 68588,USA, [2]Department of Nutrition and Health Sciences, 1700 N 35th St, Lincoln, NE 68583, USA, [3]Nebraska Center for Virology, University of Nebraska, 4240 Fair St., Lincoln, NE, 68583.

Abstract: Given the limited understanding of biogenesis and the lack of comprehensive annotations for short bacterial small RNAs, large-scale computational discovery and cataloging are required to define the search space, enable reproducibility, and guide downstream experimental investigation. We develop metaSRNA, a novel computational framework specifically designed for the prokaryotic small RNAs sequencing data from microRNA protocols and show that bacteria small RNAs with microRNA lengths constitute a complex landscape that is underrepresented in current resources. By systematically reanalyzing 111 publicly available datasets spanning single bacterial genomes, microbiome communities, and extracellular vesicles, we demonstrate the broad applicability of metaSRNA and uncover extensive sequence diversity, heterogeneous genomic mapping topologies, and variable reproducibility across studies. Genomic mapping further shows diverse and complex patterns beyond sequence level analysis, motivating a block-based representation of small RNA loci. Using this framework, we assemble an unprecedented draft catalog of candidate prokaryotic small RNAs across approximately 30 bacterial species and subspecies. Application to extracellular vesicle datasets reveals condition-dependent and species-specific small RNA signatures that may inform future studies of microbiome-host interactions. By providing both a discovery-oriented computational framework and a draft catalog of candidate short bacterial small RNAs, this work establishes a shared reference and resource for comparative analysis, sequence discovery, and rational design of small RNAs and microRNA-like molecules for future.

Title: SPLISUM: Spectral Prediction and Library Searching for Untargeted Metabolomics

Author list: Chhavi Thakur[1], Yuhui Hong[1], Sujun Li[1], and Haixu Tang[1,*]

Detailed Affiliations:
[1]Luddy School of Informatics, Computing, and Engineering, Indiana University Bloomington, IN, USA. [*]Corresponding author

Abstract: In metabolomics and natural product discovery, compound identification using untargeted tandem mass spectrometry (MS/MS) remains a primary bottleneck due to the limited coverage of experimental spectral libraries. While machine learning models now enable the in silico prediction of MS/MS spectra from molecular structures, the accuracy and sensitivity of compound identification against predicted spectral libraries have not been systematically studied.Moreover, statistical confidence in matches from such searching approaches remains a significant challenge.In this paper, we present a scalable metabolite identification pipeline SPLISUM that integrates large-scale predicted spectral libraries with a robust target–decoy framework for false discovery rate (FDR) estimation. To address this,leveraging the 3D conformation-based deep learning model 3DMolMS, we generated predicted spectra of the positive and negative ions for 65,671 metabolites from the Human Metabolome Database (HMDB), forming the target spectra library subject to be searched against for metabolite identification in untargeted metabolomics. To achieve high-throughput processing, SPLISUM utilizes locality-sensitive hashing (msCRUSH) for spectral clustering and an accelerated cosine similarity-based search (msSLASH). We also integrated a novel decoy strategy into SPLISUM that reshuffles precursor masses across spectral clusters to maintain realistic mass distributions for statistical validation.Evaluating SPLISUM using three benchmark datasets, we found that incremental False Discovery Rate (FDR) estimation stabilized around a cosine similarity threshold of 0.65. While structural similarity analysis suggests that mass-based decoys may slightly underestimate empirical error (by 0.02–0.03) due to isomeric overlap, SPLISUM outperforms the existing metabolite identification tool MCID with both

higher sensitivity and lower FDR.Therefore, SPLISUM provides a reliable and scalable framework for metabolite identification in large-scale untargeted MS/MS studies.

Keywords: TBD

Title: Celltaminate: A Bioinformatic Pipeline for Rapid Detection of Pathogenic Microorganisms

Author list: Ahmad Salman Sirajee[1], Bassel Ghaddar[1] and Subhajyoti De[1]

Detailed Affiliations:

[1]Rutgers University, New Brunswick.

Abstract: Rapid microbial detection is essential for effective clinical management and patient outcomes. While current diagnostic efforts are largely protein-based, emerging genomic techniques are rapidly evolving to bridge the gap in speed and sensitivity. Accurate interpretation of microbial reads in metagenomic sequence remains difficult because lowbiomass samples often contain spurious taxa arising from laboratory reagents, taxonomic misclassification, and contamination from sample handling. Here, we introduce Celltaminate (Contamination Evaluation for Low-Level Taxon-Abundance in Microbial Inference with Noise-Aware Triage Estimate), an interactive computational application for high-confidence, near real-time detection of microorganisms from both short and longread sequences. Celltaminate implements a composite scoring framework that integrates multiple key filters and incorporates a curated list of kitome and clinically important pathogens to recalibrate organism-level confidence. Across validation datasets, Celltaminate accurately prioritized known true organisms while downranking recurrent laboratory and environmental contaminants, significantly outperforming other alternative tools. We showcase its utility for rapid and sensitive detection and annotation of pathogenic microbial and viral signatures in diverse clinical samples from Nanopore sequencing. By combining taxonomic classification with contamination-aware scoring and an accessible user interface, Celltaminate provides a useful framework for rapid, interpretable detection of microorganisms in experimental and clinical settings.

Keywords: TBD

Workshop – Single-Cell, Spatial, and Regulatory Networks
August 4th
1:40 – 5:20 PM
Room: 2213A

Chair: Lu Li

Title: Single-cell Analysis of Intracellular Transport and Expression of Cell Surface Proteins

Author list: Rekha Mudappathi[1,2], Vaishali Bharadwaj[3], Li Liu[1,2,*]

Detailed Affiliations:

[1]College of Health Solutions, Arizona State University, Phoenix, AZ 85004, USA,[2]Biodesign Institute, Arizona State University, Tempe, AZ 85281, USA, [3]Division of Hematology and Internal Medicine Mayo Clinic, Rochester, MN, USA. [*]Corresponding author

Abstract: Motivation: Intracellular protein transport (ICT) plays a critical role in regulating surface protein localization and expression, with dysfunction contributing to disease progression and therapeutic resistance. However, the genome-wide regulatory role of ICT in modulating surface proteins remains poorly understood. Results: We present MPET (Modeling Protein Expression and Transport), a novel computational framework that integrates single-cell transcriptomic and proteomic data from CITE-seq to dissect post-transcriptional regulation of surface protein expression. MPET consists of three modules that identify ICT–surface protein regulatory circuits and evaluate their influence on cellular phenotypes using mixed-effects modeling and mediation analysis. Applying MPET to single-cell data from COVID-19 patients of varying disease severity, we uncovered contextspecific recruitment of ICT genes and pervasive rewiring of intracellular transport pathways. Notably, we found that surface protein levels were significantly altered in severe cases despite stable mRNA expression of their coding genes, highlighting the role of ICT in modulating protein expression independently of transcription. We further identified

ICT genes as potential therapeutic targets to rectify the abnormal immune responses in severe COVID-19 cases. Availability and Implementation: MPET is implemented as an R package and available at: https://github.com/liliulab/MPET. The processed CITE-seq Seurat object used for MPET benchmarking has been deposited on Figshare (DOI: 10.6084/m9.figshare.30434290). Contact: Li Liu (liliu@asu.edu)

Title: Knowledge Grounded Multi-Agent Orchestration for Interpretable Functional Summarization of Single-Cell Transcriptomics

Author list: Jinlian Wang, PhD[1], Hui Li, PhD[1], Cynthia Ju, PhD[2], Hongfang Liu, PhD[1*]

Detailed Affiliations:

[1]The McWilliams School of Biomedical Informatics UTHealth, Houston, TX, USA,[2]Department of Anesthesiology, Critical Care and Pain Medicine, McGovern Medical School, UTHealth Houston, TX, USA. [*]Corresponding author

Abstract: Interpreting single-cell RNA sequencing (scRNA-seq) data requires translating lists of differentially expressed genes (DEGs) into biologically meaningful functional summaries that integrate gene identity, pathway activity, and disease context. Current large language model-based approaches often generate fluent narratives from gene lists alone, relying solely on parametric knowledge. This leads to frequent hallucination, unverifiable claims, and a lack of traceability, limiting their utility for reproducible biological interpretation. We present a knowledge-grounded multi-agent orchestration framework that decomposes the functional interpretation pipeline into specialized agents, each grounding its outputs in real-time curated biological databases before the next step proceeds. A DEG validator queries the UniProt REST API to confirm gene annotations, a pathway agent verifies enrichments against the Reactome REST API, and a disease agent maps relevance using MyGene.info and DisGeNET. An orchestrator integrates the agent outputs, detects potential biological contradictions, computes a weighted confidence score, and assembles a structured evidence packet. This packet is then passed to a grounded LLM narrator (GPT-5.4) that generates natural-language summaries explicitly marking uncertain claims and acknowledging evidence limitations. We evaluated the framework on the primary tumor subset of GSE131907 as a representative benchmark. The framework achieved a GO-term F1 score of 0.663, outperforming naive GPT-5.4 (0.572, +16%) and GSEA-only baselines (0.267). This improvement was driven by substantially higher precision (0.577 vs. 0.470 for naive GPT-5.4). It produced confident functional summaries for 14 of 20 clusters while abstaining on the remaining 6 clusters where grounding evidence was insufficient. These abstention decisions were well-calibrated, as demonstrated by a positive calibration gap of +0.134 in GO-term recall between confident and uncertain clusters. On average, the framework generated 2.7 explicit uncertainty flags per cluster narrative, compared to zero for the naive LLM baseline. To our knowledge, this represents one of the scRNA-seq functional interpretation frameworks to integrate real-time knowledge grounding across the full interpretation chain (DEG validation, pathway confirmation, and disease mapping) with orchestrator-mediated uncertainty quantification and inter-agent consistency checking. The modular, open-source design provides an interpretable and reproducible alternative to black-box LLM annotation and is readily extensible to diverse scRNA-seq datasets, spatial transcriptomics, and multi-omics integration.

Keywords: single-cell RNA-seq, multi-agent orchestration, knowledge grounding, uncertainty quantification, functional interpretation, large language models, lung adenocarcinoma

Title: miR-CellMap: A Graph-Transformer Framework for Single-Cell microRNA-gene Regulatory Analysis

Author list: Jane Ohia[1], Juan Cui[1,*]

Detailed Affiliations:

[1]Department of Computer Science, System Biology and Biomedical Informatics Laboratory, School of Computing, University of Nebraska-Lincoln, United States. [*]Corresponding author

Abstract: Motivation: MicroRNAs (miRNAs) are key post-transcriptional regulators of gene expression, but their regulatory roles remain difficult to resolve at single-cell resolution. Emerging co-profiling technologies now allow miRNA and mRNA measurement in the same cell, creating new opportunities to study miRNA-mediated regulation in a cellular context. However, current single-cell computational frameworks are primarily designed for transcriptional or epigenomics analysis and lack dedicated strategies for integrating miRNA expression, regulatory priors, and cell-level heterogeneity into a unified regulatory model. Results: We present miR-CellMap, a graph-transformer-based computational framework for singlecell miRNA-gene regulatory analysis. The framework constructs a heterogeneous tripartite graph linking cells, genes, and miRNAs by integrating co-profiled expression

data with curated miRNA-gene interaction priors. Building on heterogeneous graph transformer architectures, miR-CellMap embeds cells, genes, and miRNAs in a shared latent space and prioritizes candidate miRNA-gene interactions in a context-aware manner. To improve regulatory interpretability, the framework further incorporates Bayesian topology optimization and competition-aware scoring to refine candidate regulatory structures and support gene-centered analysis of regulatory hierarchy, miRNA competition, and cooperative targeting. Applied to a representative co-profiled single-cell dataset, miR-CellMap demonstrates stable representation learning, biologically coherent regulatory organization, and improved recovery of experimentally supported miRNA-gene interactions after topology refinement. These results highlight the importance of combining regulatory priors with graph-based representation learning for post-transcriptional regulatory inference. Conclusion: miR-CellMap provides a scalable and extensible framework for incorporating miRNAmediated regulation into single-cell multi-omic analysis. By integrating prior biological knowledge, heterogeneous graph learning, and interpretable regulatory refinement, the framework offers a general computational foundation for studying post-transcriptional regulation at cellular resolution. Availability and implementation: miR-CellMap is available at https://github.com/sbbi-unl/miR-CellMap

Keywords: TBD

Title: Cell Segmentation Using Spatial Transcriptomic Data From The Multi-objective Optimization Perspective

Author list: Tszyi Kwok[1,2], Tara Garber[1,2], Ziying Liu[3], Hsu Kiang Ooi[3], Melanie Spears[4], Aleksandar Necakov[1,2], Youlian Pan[3] and Yifeng Li[1,2, 5,*]

Detailed Affiliations:

[1]Department of Biological Sciences, Brock University, 1812 Sir Isaac Brock Way, St. Catharines, L2S 3A1, Ontario, Canada, [2]Centre for Biotechnology, Brock University, 1812 Sir Isaac Brock Way, St. Catharines, L2S 3A1, Ontario, Canada, [3]Digital Technologies Research Centre, National Research Council Canada, 1200 Montreal Rd, Ottawa, K1A 0R6, Ontario, Canada, [4]Ontario Institute for Cancer Research, MaRS Centre 661 University Avenue, Suite 510 Toronto, M5G 0A3, Ontario, Canada, [5]Department of Computer Science, Brock University, 1812 Sir Isaac Brock Way, St. Catharines, L2S 3A1, Ontario, Canada. [*]Corresponding author

Abstract: Motivation: Cell segmentation is a central problem in spatial transcriptomics (ST): the accuracy of cell boundaries constrains every downstream analysis. State-of-the-art self-supervised frameworks such as BIDCell combine several biologically informed loss terms — nuclei encapsulation, cell calling, over-segmentation, overlap, and positive/negative marker priors — into a single weighted sum and optimize it with standard gradient descent. This scalarization implicitly assumes the losses are non-conflicting and of comparable scale; when either assumption fails, training is dominated by the largest-gradient objective. Result: A multi-objective optimization (MOO) framework is proposed for ST cell segmentation. Four recent MOO methods—Aligned Multi-Task Learning (AMTL), Smooth Tchebycheff (STCH) scalarization, Unconflicting Projection of Gradients (UPGrad), and Multiple-gradient descent algorithm (MGDA)—were applied to the training process of the BIDCell framework. A new multiple assignment loss function was introduced to replace the overlap loss which exhibited weak and irregular convergence and was unsuitable for gradient alignment. Across three platforms (Xenium, MERFISH, and an in-house NanoString CosMx cohort), lmu consistently improves morphological regularity and transcript density over lov. While MOO methods trace distinct, dataset-dependent trade-offs between shape quality and area coverage. Loss ablations show that nuclei-encapsulation and over-segmentation priors are universally critical, whereas sensitivity to the remaining objectives is strongly optimizer-specific. The main observation in this study is; MOO is a viable alternative to SOO but should be treated as a dataset-informed design choice rather than a universal upgrade. Availability: The implementation of this work is available at https://github.com/bkty1122/MOSeg. Contact: yli2@brocku.ca Supplementary information: Supplementary data are submitted along with the main text.

Keywords: Spatial Transcriptomics, Cell Segmentation, Multi-objective Optimization

Title: MolGene-E: Inverse Molecular Design to Modulate Single Cell Transcriptomics

Author list: Rahul Ohlan[1], Raswanth Murugan[3], Li Xie[3], Vikas Nallabolu[5], Mohammedsadeq Mottaqi[2], Shuo Zhang[3,4*], Lei Xie[1,2,3,4,5*]

Detailed Affiliations:

[1]Ph.D. program in Computer Science, The Graduate Center, The City University of New York, New York, NY, 10016, USA, [2]Ph.D. program in Biochemistry, The Graduate Center, The City University of New York, New York,

NY, 10016, USA, [3]Department of Computer Science, Hunter College, The City University of New York, New York, NY, 10065, USA,[4]Helen and Robert Appel Alzheimer's Disease Research Institute, Feil Family Brain & Mind Research Institute, Weill Cornell Medicine, Cornell University, New York, NY, 10065, U.S.A,[5]School of Pharmacy and Pharmaceutical Sciences & Center for Drug Discovery, Northeastern University, Boston, MA, 02115, USA. [*]Corresponding author

Abstract: Designing drugs that can restore a diseased cell to its healthy state is an emerging approach in systems pharmacology to address medical needs that conventional target-based drug discovery paradigms have failed to meet. Single cell transcriptomics can comprehensively map the differences between diseased and healthy cellular states, making it a valuable technique for systems pharmacology. However, single-cell omics data is noisy, heterogeneous, scarce, and high-dimensional. As a result, no machine learning methods currently exist to use single-cell omics data to design new drug molecules. We have developed a new deep generative framework named MolGene-E to tackle this challenge. MolGene-E combines two novel models: 1) a cross-modal model that can harmonize and denoise chemical-perturbed bulk and single-cell transcriptomics data, and 2) a contrastive learning-based generative model that can generate new molecules based on the transcriptomics data. MolGene-E consistently outperforms baseline methods in generating high-quality, hit-like molecules on gene expression profiles from two evaluation settings: CRISPR knock-out perturbation profiles from L1000-to-RNA-seq dataset, and single-cell gene expression profiles from Sciplex-3 dataset, both in zero-shot molecule generation setting. This superior performance is demonstrated across diverse de novo molecule generation metrics. Extensive evaluations demonstrate that MolGene-E achieves state-ofthe-art performance for zero-shot molecular generations. This makes MolGene-E a potentially powerful new tool for drug discovery.

Keywords: Systems Pharmacology, Drug Discovery, Phenotypic Screening, Generative AI

Title: Dysregulated intercellular communication and signaling pathways across brain regions in Alzheimer's Disease

Author list: Hani Alostaz[1], Songjian Lu[1], Citu Citu[1], Zhongming Zhao[1,2,3*]

Detailed Affiliations:

[1]Center for Precision Health, McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA, [2]MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences, Houston, TX 77030, USA, [3]Faillace Department of Psychiatry and Behavioral Sciences, McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, TX, USA. [*]Corresponding author

Abstract: Alzheimer's Disease (AD) is a progressive neurodegenerative disease characterized by profound neuronal loss and cognitive decline. Intercellular cell-cell communication (CCC) plays a critical role in maintaining brain homeostasis, yet its dysregulation in AD across distinct brain regions remains poorly understood. Here, we present a comprehensive analysis of 12 single nucleus RNA-seq (snRNA-seq) AD-specific datasets across four critical brain regions: the prefrontal cortex (PFC), entorhinal cortex (EC), hippocampus (HIP), and dorsolateral prefrontal cortex (DLPFC). In total, we identified 3,217 differentially regulated ligand-receptor (L-R) pairs, of which 990 were upregulated and 2,227 were downregulated across all four regions in six major brain cell types. Among the upregulated L-R pairs, 388 were shared across all brain regions, whereas 602 were region-specific. In contrast, 1,119 downregulated L-R pairs were common across regions, and 1,108 were region-specific. The EC region showed the highest number of region-specific downregulated L-R pairs, indicating that it is the most transcriptionally disrupted region with respect to reduced signaling activity in AD. The target gene analysis of dysregulated L-R pairs identified APP, CCL2, ICAM1, SPP1, and UQCRC1 as major targets across brain regions. Pathway analysis showed upregulation of glutamate transport in the PFC and activation of ERBB/MAPK cascade signaling in EC, DLPFC, and HIP. Overall, these findings highlight region-specific and shared disruptions in CCC as key drivers of AD progression and potential therapeutic targets.

Title: EVTS: A Topology-Aware Metric for Evaluating RNA Velocity Embeddings in Single-Cell Transcriptomics

Author list: Xiuquan Wang[1], Nan Nan[2], Kai Wang[3,4]

Detailed Affiliations:

[1]Department of Mathematics and Computer Science, Tougaloo College, Tougaloo, MS, USA, [2]School of Agricultural Sciences and Forestry, Louisiana Tech University, Ruston, LA 71272, USA, [3]Raymond G. Perelman

Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA, [4]Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.
Abstract: RNA velocity is widely used to infer the likely future states of single cells from the relative abundance of spliced and unspliced transcripts. In most analyses, the resulting velocity field is displayed on a two-dimensional embedding, such as PCA, UMAP, t-SNE, or diffusion maps. These plots are easy to interpret visually, but the embedding step can change the geometry of the velocity transition graph and may cause projected arrows to point in directions that do not match the transition probabilities estimated by the RNA velocity model. To address this issue, we propose the Embedding Velocity Topology Score (EVTS), a quantitative score for measuring how well projected velocity vectors agree with transition-predicted future cell directions in an embedding space. EVTS asks a simple question: for a given cell, does the displayed velocity arrow point toward the cells that the RNA velocity model identifies as likely future states? We also describe multi-scale and uncertainty-aware versions of EVTS, together with a KLdivergence formulation that compares the RNA velocity transition distribution with an embedding-induced attention distribution. In a synthetic branching trajectory with known groundtruth flow, EVTS separated embeddings that preserved the simulated direction of progression from embeddings that distorted it, and EVTS-guided refinement modestly improved flow agreement across the tested methods. In the scVelo pancreas dataset, UMAP gave the highest EVTS and MU-EVTS among the baseline embeddings, whereas t-SNE gave the strongest neighborhood preservation. These results suggest that EVTS measures something that is missed by standard embedding scores: whether the arrows shown in an RNA velocity plot actually follow the transition graph estimated by the velocity model.
Keywords: RNA velocity; single-cell RNA sequencing; embedding evaluation; trajectory inference; graph diffusion; topology preservation; dimensionality reduction; single-cell omics

Workshop – Computational Drug Discovery and Therapeutic Modeling August 5th 9:20 AM – 12:00 PM Room: 2220A&B

Chair: Zackary Falls

Title: KGPredict-Com: A Knowledge Graph Model for Drug Combination Prediction

Author list: Zhenxiang Gao[1], Rong Xu[1*]

Detailed Affiliations:
[1]Center for Artificial Intelligence in Drug Discovery, School of Medicine, Case Western Reserve University, Cleveland, Ohio, USA. [*]Corresponding author
Abstract: Identifying effective drug combinations is an important strategy for improving therapeutic outcomes, particularly for complex diseases. However, the large combinatorial space of possible drug pairs makes exhaustive experimental evaluation infeasible. Computational methods offer a practical way to prioritize candidates for further validation, but many existing approaches, developed in restricted settings such as specific cell lines or relying on limited data modalities with insufficient disease-specific information, have limited generalizability to broader disease contexts. In this study, we propose KGPredict-Com, a knowledge graph-based framework for diseasespecific drug combination prediction, designed to address the limitations of existing approaches in disease specificity and generalizability. The model consists of two modules. The foundation module constructs a foundational biological knowledge graph that integrates heterogeneous biomedical data and rich disease-level information, including disease-gene associations, comorbidities, and treatment information, enabling more comprehensive and disease-aware representations of drugs, diseases, and genes. The task-specific module performs diseaseconditioned combination scoring over the full combinatorial space, allowing the model to directly predict drug combinations for a given disease and capture context-dependent interactions through a drug-disease relevance scorer and a three-way interaction module. In evaluation, KGPredict- Com achieved excellent performance, with an average Hits@100 of 0.5, MRR of 0.1, AUROC of 0.99, and AUPRC of 0.29.
Keywords: TBD

Title: Sparsity, Structure, and Interpretability in Biologically Informed Neural Networks for Drug Response Prediction

Author list: Jason Chia[1,2], Yitan Zhu[2,*], Oleksandr Narykov[2], Thomas S. Brettin[2] and Rick L. Stevens[2,3]

Detailed Affiliations:

[1]Department of Computer Science, Purdue University, 610 Purdue Mall, West Lafayette, IN 47907, United States, [2]Computing, Environment and Life Sciences, Argonne National Laboratory, 9700 S. Cass Avenue, Lemont, IL 60439, United States, [3]Department of Computer Science, The University of Chicago, 5730 S. Ellis Ave, Chicago, 60637, IL, United States. [*]Corresponding author

Abstract: Biologically informed neural networks (BiNNs) incorporate prior knowledge of biological pathways and genomic interactions into their architectures, offering the potential for both predictive accuracy and mechanistic interpretability in tasks such as anticancer drug response prediction. However, whether these architectures genuinely leverage their embedded biological structure remains underexplored. We investigate this question using the Nested Systems in Tumors–Visual Neural Network (NeST-VNN), a model designed to predict cancer cell responses to replication stress–inducing drugs. The architecture of NeST-VNN is derived from NeST, a curated hierarchy of cancer protein systems. Through input gene-order permutation experiments, we show that disrupting gene-to-assembly mappings does not degrade predictive performance, indicating that the biological structure encoded in the architecture is not a necessary driver of accuracy. To better understand the architectural factors that do drive performance and to address the computational limitations of NeST-VNN, we develop the fast NeST-Neural Network (fNeST- NN), a computationally optimized variant that parallelizes computation across protein assemblies and aggregates intermediate-layer predictions to generate the final output. This approach achieves up to an approximately 20-fold training speedup while improving predictive accuracy. Through systematic comparisons with dense network baselines, statically initialized sparse architectures, and sparse architectures dynamically learned during training, we attribute the strong performance of fNeST-NN to three compounding factors: statically initialized sparsity; direct gene inputs to intermediary layers combined with aggregated intermediate layer predictions; and the specific NeST-derived sparsity architecture. These findings clarify the sources of predictive strength in NeST-based models and establish fNeST-NN as a faster, higher-performing alternative for practical drug response modeling workflows.

Keywords: TBD

Title: A DALEX-Based Explainable AI Framework for Lymphography Prediction

Author list: Tapas Si[1], Akash Deep Kumar[1], Saurav Mallik[2], and Zhongming Zhao[3,4,5]

Detailed Affiliations:

[1]UEMJ Advanced Machine Learning Research Group, AI Innovation Lab, Department of Computer Science and Engineering, University of Engineering & Management, Jaipur, Rajasthan 303807, India, [2]Department of Pharmacology and Toxicology, University of Arizona, Tucson, AZ 85721, United States, [3]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, United States, [4]Human Genetics Center, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, United States, [5]Department of Pathology and Laboratory Medicine, McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, TX, United States.

Abstract: Lymphatic disorders pose diagnostic challenges due to heterogeneous clinical patterns and the need for reliable decision support. This article proposes an explainable machine learning framework for lymphography disease classification using K-Nearest Neighbors (KNN), Decision Tree (DT), and Logistic Regression (LR). Hyperparameter optimization was performed using Random Search technique, and a 10-fold stratified crossvalidation framework was to ensure unbiased performance evaluation and prevent data leakage. Models were assessed using accuracy, precision, recall, F1-score, and Geometric Mean, followed by Technique for Order Preference by Similarity to Ideal Solution (TOPSIS)-based multi-criteria ranking. LR achieved the best overall performance with a mean accuracy of 83.19%, and the highest TOPSIS rank. To enhance interpretability, the Descriptive mAchine Learning EXplanations (DALEX) framework was employed to provide global and instance-level explanations, identifying key predictors such as lymph node involvement and structural abnormalities. The proposed framework integrates robust evaluation and model-agnostic explainability to support transparent and reliable AI-assisted lymphography diagnosis.

Keywords: lymphography, explainable AI, DALEX, ma- chine learning, healthcare

Title: TransGen: De Novo Enzyme Design via Catalytic-Conditioned Sequence Generation

Author list: Hongmin Cai[1], Zhihai Zhang[2], Xiaoqi Sheng[1], Fangwei Song[3], Sankar Mondal[1], Ting-He Zhang[1,*] and Ying Lin[3,*]

Detailed Affiliations:

[1]School of Future Technology, South China University of Technology, 777 Xingye Avenue East, 511442, Guangdong, China, [2]School of Computer Science and Engineering, South China University of Technology, 382 Zhonghuan Road East, 510006, Guangdong, China, [3]School of Biology and Biological Engineering, South China University of Technology, 382 Zhonghuan Road East, 510006, Guangdong, China. [*]Corresponding author

Abstract: De novo enzyme design remains challenging because abstract catalytic logic is difficult to encode directly into amino acid sequences. Existing approaches often rely on homology search, random mutation, or coarse functional conditioning, resulting in low efficiency and substantial manual intervention. To address these limitations, we introduce TransGen, a generative model that uses dilated convolutions and directly integrates reaction-level catalytic changes into protein sequence generation. The model is trained and evaluated on UDP-glycosyltransferases (UGTs), enzymes crucial in plant metabolism. Each reaction is encoded as a single Reaction Vector, defined as the difference between aggregated product and reactant embeddings. Incorporating the Reaction Vector through Adaptive Layer Normalization (AdaLN) conditions residue-level sequence patterns, while alignment of latent representations with the ESM-2 knowledge space promotes the biological plausibility of generated sequences. Docking analysis suggests that the generated sequences preserve pocket-level structural compatibility with the target reactions. By generating protein sequences in response to reaction-specific chemical signals, TransGen provides a promising framework for controllable enzyme sequence design. The code of TransGen is available at: https://github.com/falcooon/TransGen.

Keywords: Enzyme design; Catalytic transformations; Protein engineering; Artificial intelligence

Title: DrugPTM-Bench: A Large-Scale Dataset for Predictive Modeling of Drug-Induced Cell Type-Specific Protein Post-Translational Modifications

Author list: Amitesh Badkul,[1*] Mohammadsadeq Mottaqi[2], Li Xie[3] and Lei Xie[3,4]

Detailed Affiliations:

[1]Computer Science, The Graduate Center, CUNY, 365 Fifth Avenue, New York City, New York, USA, [2]Biochemistry, The Graduate Center, CUNY, 365 Fifth Avenue, New York City, New York, USA, [3]Computer Science, Hunter College, 695 Park Avenue, New York City, New York, USA, [4]Pharmaceutical Sciences, Northeastern, 360 Huntington Avenue, Boston, Massachusetts, USA. [*]Corresponding author

Abstract: Protein post-translational modifications (PTMs), particularly phosphorylation, serve as the primary ”molecular switches” that orchestrate cellular signaling and drug response. While PTM dysregulation is a hallmark of cancer and neurodegeneration, the lack of standardized, drug-perturbed datasets has hindered the development of predictive models capable of capturing context-dependent PTM responses. Effective predictive modeling must therefore integrate multidimensional data, including the specific drug, dosage, treatment duration, cellular background, and the modified site. However, existing PTM resources remain largely static and fail to capture drug-induced regulation across these critical dimensions. To address this gap, we present DrugPTM-Bench, a curated, large-scale benchmark derived from decryptM-derived dose-dependent PTM measurements, standardizing site-level drug response across 7 cancer cell lines, 27 drugs, and 11,167 proteins. Comprising 99.5% phosphorylation events, the dataset includes six time points, 16 dosage levels, and pEC50 potency values (half-maximal effective concentration). We formulate a classification task to identify upregulated, downregulated, or unchanged PTM sites (following a drug treatment), a critical step in deciphering drug Mechanism of Action (MoA) and target engagement. Our evaluation reveals that in protein-disjoint out-of-distribution (OOD) setting, baseline machine learning and deep learning models struggle to recover minority regulation classes, while standard rebalancing strategies improve recall only at the cost of precision and overall F1-score. These results indicate that current methods do not learn robust decision boundaries between regulated and unchanged PTM events. DrugPTM-Bench provides a phosphoproteomics benchmark for modeling drug-induced PTM regulation in imbalanced biological settings. Beyond classification, DrugPTM-Bench’s retention of pEC50 values, drug perturbation profiles, and site-level sequence context enables additional predictive tasks including drug potency regression, mechanism-of-action prediction from PTM fingerprints, and drug-specific PTM site sensitivity ranking, establishing a multi-

task benchmark for PTM-centric drug discovery. Ultimately, DrugPTM-Bench establishes a rigorous framework for developing robust, context-aware models to elucidate drug MoA and signaling dynamics.

Title: From Generalist to Specialist: Evolution of PS2 α-integrins and Implications for Drug Targeting

Author list: Shixi Sun[1], Yunfeng Chen[2], Rong-Guang Xu[2], Heng Zhang[3], Fahad Mostafa[4], Li Liu[1,5,*]

Detailed Affiliations:

[1]College of Health Solutions, Arizona State University, Phoenix, AZ 85004, USA, [2]Department of Biochemistry and Molecular Biology and Department of Pathology, The University of Texas Medical Branch, Galveston, TX 77555, USA, [3]Versiti Blood Research Institute, Milwaukee, WI 53226, USA, [4]School of Mathematical and Natural Sciences, Arizona State University, Glendale, AZ 85306, USA, [5]Biodesign Institute, Arizona State University, Tempe, AZ 85281, USA. [*]Corresponding author

Abstract: Integrins are heterodimeric transmembrane receptors that mediate cell–cell and cell– extracellular matrix interactions and play essential roles in development and disease. Within the PS2 α-integrin subfamily, four paralogs (αIIb, α5, α8, and αV) share a conserved RGD-binding motif yet exhibit diverse functional specializations. Integrins have been widely targeted therapeutically for various clinical conditions, though achieving subtype specificity remains a major challenge. Here, we performed an integrative evolutionary analysis of 114 PS2 α-integrin sequences across 28 vertebrate species, combining phylogenetic reconstruction, time calibration, ancestral sequence inference, and structural mapping. Our time-calibrated phylogeny indicates that the PS2 lineage originated ~862 Mya, with diversification of the four paralogs occurring prior to vertebrate radiation. Ancestral state reconstruction reveals that fibronectin and vitronectin binding are ancestral traits, whereas fibrinogen binding and β3 pairing arose independently in the αIIb and αV lineage. Evolutionary rate analysis shows domain specific divergence, with the β-propeller acting as a hotspot of evolutionary change, likely driven by combined pressures from ligand binding and β-subunit interaction. These pressures vary across paralogs: αIIb exhibits accelerated evolution in ligand-binding regions, while αV displays elevated rates in β-subunit interaction domains. Mapping sequence variation onto structural interfaces identifies lineage-specific substitutions underlying functional divergence, including distinct molecular solutions for fibrinogen binding in αIIb and αV. These findings collectively demonstrate that PS2 α-integrins evolved from a generalist ancestor through neofunctionalization and lineage-specific specialization. This work provides an evolutionary framework for identifying subtype-specific functional sites and highlights the potential of evolution-informed strategies to guide the development of more selective integrin-targeting therapeutics.

Title: A Reproducible Hybrid QSPR-Machine Learning Framework Integrating Graph-Theoretic and Cheminformatics Descriptors for Anticancer Drug Property Prediction

Author list: H Muhammad Fraz[1], Muhammad Faisal Nadeem[1], Muhammad Kamran Jamil[2], Saurav Mallik[3,4], Zhongming Zhao[5,6,7,*]

Detailed Affiliations:

[1]Department of Mathematics, COMSATS University Islamabad, Lahore Campus, Lahore, Pakistan, [2]Department of Mathematics, Riphah International University, Lahore, Pakistan, [3]Department of Environmental Health, Harvard T.H. Chan School of Public Health, Boston, MA, USA, [4]Department of Pharmacology & Toxicology, University of Arizona, Tucson, AZ, USA, [5]Center for Precision Health, McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA, [6]MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences, Houston, TX 77030, USA, [7]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37203, USA. [*]Corresponding author

Abstract: Accurate prediction of physicochemical properties plays a fundamental role in rational drug design, lead optimization, and virtual screening by enabling early assessment of molecular behavior prior to experimental synthesis. Quantitative structure– property relationship (QSPR) modeling provides an efficient mathematical framework for establishing relationships between molecular structure and macroscopic physicochemical responses. In this study, we develop a reproducible hybrid QSPR—machine learning framework that integrates mathematical graph-theoretic descriptors with cheminformatics molecular descriptors and structural fingerprints for predictive modeling of physicochemical properties of anticancer drug molecules. A curated dataset of 100 experimentally validated compounds was obtained from PubChem and ChEMBL databases and standardized through duplicate removal, salt elimination, structural normalization, and validation of molecular connectivity. 1 2 The descriptor space combines RDKit physicochemical descriptors with graphtheoretic descriptors derived from molecular graph

representations. Morgan fingerprints were incorporated to encode substructural patterns and pharmacophoric fragments, enabling representation of both global molecular topology and local chemical environments. Machine learning models including Ridge regression, Random Forest, and XGBoost were optimized using nested cross-validation in accordance with OECD principles for QSPR validation. Model reliability and predictive stability were evaluated using R2, Q2, RMSE, MAE, Y-randomization testing, and applicability domain analysis using Williams plots. The developed models demonstrate reliable predictive capability for multiple physicochemical properties relevant to drug-likeness and pharmacokinetic behavior, achieving Q2 values of 0.659 for rotatable bonds, 0.650 for topological polar surface area, 0.546 for heavy atom count, 0.539 for molecular complexity, and 0.728 for nPotency, while lipophilicity (logP) exhibited moderate predictability (Q2 = 0.271) due to complex intermolecular interaction effects. The results indicate that integration of graph-theoretic invariants with cheminformatics descriptors improves structural representation and enhances predictive robustness of machine learning models. The novelty of this work lies in the systematic integration of mathematically interpretable graph invariants with cheminformatics descriptors and molecular fingerprints within a unified and reproducible QSPR–machine learning framework. The proposed hybrid descriptor representation provides improved linkage between molecular topology and physicochemical behavior, supporting interpretable artificial intelligence approaches for data-driven drug discovery. The developed framework offers a transparent computational pipeline for virtual screening and molecular property prediction, and can be extended to graph neural networks, transformer-based molecular embedding, and multi-objective optimization strategies for next-generation AI-assisted drug design.

Keywords: TBD

Workshop – Multi-omics, AI, and the Future of Precision Health
August 5th
9:20 AM – 12:00 PM
Room: 2213A

Chair: Yuanyuan Fu

Title: From Genetic Susceptibility to Modifiable Biology: Multi-Omic and Dietary Integration Toward Precision Alzheimer's Disease Prevention

Author list: Yuxi Liu*[1,2,3], Dong D. Wang[1,3]

Detailed Affiliations:
[1]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA, USA, [2]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA, [3]Broad Institute of MIT and Harvard, Cambridge, MA, USA. *Corresponding author

Abstract: Blood-based biomarkers representing amyloid (A), tau (T), and neurodegeneration-related (X/N) pathophysiology are rapidly transforming Alzheimer's disease (AD) research by enabling scalable, minimally invasive characterization of disease biology and progression across the AD continuum. While plasma markers such as phosphorylated tau (e.g., pTau217 and pTau181) have shown promise for early risk stratification, the genetic architecture underlying variation across a broader panel of ATX(N) biomarkers remains largely unknown. Characterizing this genetic basis is essential for distinguishing proteins under predominantly local genetic control (cis) from those influenced by distal loci (trans), and for elucidating whether observed biomarker fluctuations reflect protein-specific signatures or convergent genetic pathways underlying neurodegeneration. Here, we present a genome-wide meta-analysis of plasma ATX(N) proteomics across two independent cohorts: the Nurses' Health Study and the DIRECT-PLUS trial. Utilizing the NULISAseq platform, we quantified a 131-protein panel optimized for the detection of low-abundance plasma neurodegenerative signals with high sensitivity, including core AD markers such as overall and brain-derived pTau species (pTau217, pTau181, pTau231), amyloid dynamics (Aβ42/40), and markers of axonal injury (NfL) and astrogliosis (GFAP), as well as proteins involved in immune, vascular, and metabolic pathways. Our analysis identified a robust landscape of cis- and trans-protein quantitative trait loci (pQTL), with several loci colocalizing with established AD risk regions. Notably, we identified a cis-pQTL for brain-derived pTau181 that maps to established AD risk loci, suggesting a mechanistic link between genetic susceptibility of AD and the regulation of tau proteostasis. Furthermore, we identified a novel trans-link between

AD risk variants and CRH, implicating a genetic connection between systemic stress-response signaling and AD susceptibility. Ongoing work utilizes Mendelian Randomization to assess the causality of these proteins in AD pathogenesis and network-based integration to identify protein modules under shared genetic control. By establishing this ATX(N) biomarker genetic atlas, we provide a foundation for multi-omic integration to dissect AD biology, prioritize therapeutic targets, and refine the interpretation of plasma biomarkers across clinical and research settings.

Keywords: Alzheimer's disease, ATX(N) biomarkers, protein quantitative trait loci (pQTL), multi-omics, genetic architecture

Title: Evaluation of statistical differential analysis methods for identification of senescent cells using single-cell transcriptomics

Author list: Dongmei Li, Pinxin Liu, Irfan Rahman, Martin Zand, Gloria Pryhuber, Timothy Dye, Maciej Goniewicz, Aditi Uday Gurkar, Melanie Königshoff, Oliver Eickelberg, Ana Mora, Mauricio Rojas, Qin Ma, Jose Lugo-Martinez, Ziv Bar-Joseph, Serafina Lanna, Toren Finkel, Zidian Xie

Detailed Affiliations:

Clinical and Translational Science Institute, School of Medicine and Dentistry, University of Rochester Medical Center, Rochester, NY, USA

Abstract: Differential gene expression (DGE) analysis is a crucial step in identifying senescent cells using single-cell RNA sequencing (scRNA-seq) data. However, few studies have evaluated the performance of DGE methods—particularly those implemented in the widely used Seurat package. In this study, we systematically assessed 10 DGE methods available in Seurat—Wilcox, Wilcox-limma, bimod, roc, t, negbinom, Poisson, LR, MAST, and DESeq2—using simulated and real scRNA-seq datasets. We evaluated each method's performance across varying sample sizes, levels of sparsity, and proportions of truly differentially expressed genes. Metrics assessed included false discovery rate (FDR), sensitivity, specificity, accuracy, area under the receiver operating characteristic curve (AUC), and area under the precision-recall curve (AUPRC). Among all methods, DESeq2 consistently demonstrated the best overall performance, showing the highest AUC and AUPRC across all tested conditions. Based on our findings, we recommend DESeq2 as the preferred method for DGE analysis in scRNA-seq data.

Keywords: single-cell RNA-seq, differential expression, senescent cells, DESeq2, method evaluation

Title: TRACED: A Graph-Based Algorithm to Infer Single-cell Time-lagged Gene Regulation

Author list: Yang Yu[1,2], Lili Zhang[1,2], Kaifu Chen[1,2,^]

Detailed Affiliations:

[1]Department of Cardiology, Boston Children's Hospital, [2]Department of Pediatrics, Harvard Medical School.
[^]Corresponding author

Abstract: Regulatory interactions in biological systems are inherently time dependent. However, current scRNA-seq analyses that adopt static correlation approaches to infer causal relationships cannot distinguish regulators from targets without additional data modalities such as ATAC-seq or ChIP-seq, which are limited to specific interaction types. We developed TRACED, a graph-based algorithm that leverages pseudo- or real-time information to infer causal relationships from expression data alone. Our approach models temporal progression by identifying concordant expression variation patterns along developmental trajectories of gene pairs. We validated our method in simulations and real biological contexts. For instance, in mouse liver circadian time series data, TRACED successfully identified the intrinsic time lag between unspliced and spliced mRNA dynamics. We further showed that TRACED-identified regulator-target pairs offer rich biological insights: in an iPSC-derived arterial endothelial cell (EC) differentiation model, TRACED identified 32 genes that are repressed by the EC identity transcription factor, MECOM. These genes are enriched for collagen and ECM functions, which led us to hypothesize that MECOM knockout impairs mesenchymal-to-endothelial transition (EndoMT). Such a hypothesis was experimentally validated. Finally, we showed that TRACED identified a subset of core genes that are sufficient to infer cell trajectory in a mouse endocrine cell development model. Together, TRACED provides a robust framework for inferring time-lagged regulatory relationships from expression data alone, enabling identification of driver genes and causal interactions across diverse biological systems.

Keywords: Single-cell, Time lag, Regulatory relationships, Cell trajectory

Title: Leveraging gene-environment interactions to understand colorectal cancer risk

Author list: Junyi Xin, *Bingxin Liu*, Zhutao Ding, Yu Shao, Yuming Chen, Qian Gong, Yutao Zhou, Yichu Chen, Jiajin Wu, Silu Chen, Meilin Wang, Mulong Du*, David C Christiani

Detailed Affiliations:

Department of Biostatistics, Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, China, *Presenter

Abstract: Metabolites and genetic factors have been widely reported to be associated with the risk of cancer; however, their potential interactions involved in tumorigenesis remain unclear.

In this study, leveraging plasma NMR metabolomics and genotyping data of 200,604 European individuals from UK Biobank cohort, we comprehensively evaluated the interactions between 251 metabolites and genome-wide SNPs on pan-cancer risk. Subsequently, a Chinese NJCRC cohort including 956 colorectal cancer cases and 330 controls was used to validate the gene-by-environment (G×E) interaction occurred in colorectal cancer susceptibility, followed by the biological mechanism experiments.

Using a multi-stage G×E interaction analysis framework, we identified a total of 5,803,196 nominally significant SNP-metabolite interaction pairs (P~interaction~ ≤ 0.001), involving 211 cancer-related metabolites and corresponding 1,931,919 interacted SNPs (iSNPs) across 15 cancer types. Furthermore, we developed a user-friendly online database, CGMI (https://njmu-edu.cn:40072/CGMI/), to help users access the above interaction results, along with functional annotations and clinical applications (including 6,546 polygenic risk scores interacted with metabolites) of these iSNPs. Interestingly, we further validated the interaction between rs930557 (8p23.1, located in MCPH1) and glucose underlying colorectal cancer susceptibility in the Chinese NJCRC cohort; and observed the genetic function that the rs930557 C allele may promote colorectal tumorigenesis under elevated blood glucose conditions through suppressing MCPH1 expression.

CGMI can serve as an important resource to systematically assess the genetic regulation of metabolite-cancer associations, providing novel insights into the genetic mechanisms underlying tumorigenesis.

Keywords: cancer, metabolite, genetic variants, interaction, risk stratification

Title: Machine learning-derived subtypes of autism reveal distinct genetic contributions from common and rare variants

Author list: Qing Tan, Jingyu Xie, LeeAnne Green Snyder, Bridget Fernandez, Nicholas Mancuso, Charleston Chiang, Chang Shu*

Detailed Affiliations:

Department of Population and Public Health Sciences and Pediatrics, Keck School of Medicine, University of Southern California, *Presenter

Abstract: Autism spectrum disorder (ASD) is characterized by substantial heterogeneity in core symptom severity and co-occurring conditions, complicating efforts to understand its genetic basis. Most genetic studies rely on binary case-control ASD definitions, which may obscure the underlying biological differences and the diverse genetic contributions across ASD subtypes.

We constructed ASD subtypes and corresponding continuous subtype scores by using Discriminative Dimensionality Reduction via learning a Tree (DDRTree), an unsupervised machine learning clustering approach that projects data into a reduced-dimensional space while constructing a principal tree that captures the intrinsic structure of the data. We applied this method to 14 key ASD clinical traits including assessing social, cognitive, behavioral scores and comorbidity among 6,355 ASD children in the Simons Foundation Powering Autism Research for Knowledge (SPARK). We also projected individuals from the Simons Simplex Collection (SSC; N = 1,403) onto the SPARK-derived subtype structure using a k-nearest neighbors (k-NN) approach. We compared the burden of de novo loss-of-function (LoF) variants and SNP-based heritability across ASD subtypes in SPARK and SSC.

We identified three distinct ASD subtypes: Subtype 1 represents the most severely affected group, with highest therapy needs especially in language skills; Subtype 2 shows high comorbidity with mood disorders, with specific needs in medication and mood regulation; Subtype 3 is the least affected group with lowest support needs. Projected ASD subtypes in SSC showed similar clinical profiles as SPARK, supporting the generalizability of our model. The

de novo LoF burden is consistently higher in Subtype 1 and 2 as compared with Subtype 3 in SPARK and SSC. The SNP heritability of the continuous Subtype 1 and 2 scores was 0.163 (SE = 0.053) and 0.129 (SE = 0.049) respectively, which are higher than SNP heritability of 0.118 observed for the binary ASD diagnosis.

We demonstrate that DDRTree can construct clinically meaningful ASD subtypes and continuous ASD subtype scores that capture both core and co-occurring symptoms. These subtypes are genetically distinct, supported by de novo LoF burden and SNP heritability analysis. Our findings highlight the potential of machine learning approaches in dissecting ASD heterogeneity and linking clinical phenotypes to common and rare variants.

Keywords: autism spectrum disorder, phenotypic heterogeneity, machine learning, rare variants, common variants

Title: Risk-stratified hepato-pancreato-biliary cancer screening with self-evolving trustworthy AI

Author list: Haojun Yu, Jiachen Zheng, Yiang Gao, Xiangran Jia, Xiaowen Tang, Caichao Lv, Xinhe Zhang, Hui Ding, Ze Zhang, Wenqing Yang, Runxin Shi, Yinbo Liu, Binbin Su, Guangyi Fan, James Zou, Haiyan Xin, Ashok Saluja, Ning Mao, Liwei Wang, Jie Nie, He Ren*

Detailed Affiliations:

Medical Research Center, The Affiliated Hospital of Qingdao University, Qingdao, China. *Presenter

Abstract: Hepato-pancreato-biliary (HPB) cancers are highly lethal malignancies, and early detection can significantly improve patient survival. However, effective population-wide screening approaches for HPB cancers are limited in clinical practice. In this presentation, we present risk-stratified HepatO-Pancreato-biliary cancer screening with self-Evolving trustworthy AI (HOPE), a population-wide HPB cancer screening approach with high accuracy and cost efficiency. To accomplish this, HOPE progressively concentrates screening efforts on higher risk candidates, named at-risk, high-risk and suspected tiers, through more precise data modality and AI tools. Moreover, to foster physician trust, we trained large language models (LLM) with a self-evolving algorithm to provide interpretable and evidence-based reasoning. We developed and tested HOPE on an internal cohort (n=10,269,153). Then, we conducted extensive evaluation on external cohorts (UK Biobank n=501,946; MIMIC-IV n=364,627; Multi-center CT n=1,127) and conducted a real-world study (n=102,324) to assess the ability of HOPE. We demonstrate HOPE is capable of accurately predicting high risk individuals with a median lead time of 1.66-2.03 years compared to clinical diagnosis, providing potential for treatment at an earlier stage.

Keywords: HPB cancer, risk stratification, AI screening, trustworthy AI, early detection

Title: Beyond exposure: computational calibration and ancestry-aware genetics of skin carotenoid spectroscopy for precision nutrition

Author list: Yixing Han

Detailed Affiliations:
National Institutes of Health (NIH)

Abstract: Noninvasive skin carotenoid spectroscopy provides a scalable biomarker of fruit and vegetable exposure, but skin carotenoid score is a composite phenotype rather than a direct substitute for plasma carotenoid concentration. We developed a layered, ancestry-aware framework to quantify how circulating exposure, carotenoid species composition, host biology, and genetic architecture shape skin–plasma carotenoid discordance in two deeply phenotyped multi-ancestry cohorts with plasma carotenoids, skin spectroscopy, dietary and clinical covariates, intervention data, genome-wide genotypes, and genetic ancestry information.

The exposure/composition layer showed that total plasma carotenoids strongly predicted skin carotenoid score, but plasma species mixture explained substantial additional variation. Adding plasma lycopene fraction to a research model containing total plasma carotenoids, age, sex, race/ethnicity, BMI, cholesterol, hemoglobin index, and melanin index increased explained variance from 0.591 to 0.703 and reduced AIC from 2428.9 to 2363.7. Higher lycopene fraction was associated with lower-than-expected skin scores, indicating a predictable source of skin-plasma mismatch. The host biology layer showed that composition-adjusted discordance was driven primarily by optical/readout-related factors, especially melanin and hemoglobin indices, with smaller contributions from adiposity and lipid transport. These associations replicated in an independent intervention cohort baseline subset, and discordance showed substantial longitudinal stability over six weeks.

The genetic layer distinguished calibration mismatch from the measured skin phenotype. SNP heritability was higher for measured skin carotenoid score than for total plasma carotenoids (h2=0.30 vs. 0.08), whereas discordance was not treated as a primary GWAS trait. An ancestry-aware linear mixed-model GWAS of measured skin score identified suggestive lead loci, with the top association at chr8:80665712:A:G (P=7.15e-8). Standard regression produced additional ancestry-structured signals with large allele-frequency differences across groups, underscoring the need for mixed-model adjustment. Conditional lead-locus analyses showed that plasma carotenoids and melanin/hemoglobin covariates attenuated genetic effects, suggesting contributions from both systemic carotenoid burden and skin-side measurement biology. This framework turns a convenient dietary readout into an interpretable genetic and biomarker phenotype and provides a general strategy for validating noninvasive health measures in diverse populations.

This work converts a convenient noninvasive dietary readout into an interpretable computational and genetic phenotype. The framework provides a general strategy for validating scalable biomarkers in diverse populations, reducing ancestry- and measurement-related bias, and improving the reliability of precision nutrition and public-health tools.

Workshop – Emerging BioHealth and Environmental Frontiers August 5th
9:20 AM – 12:00 PM
Room: 2213B

Chair: Jane Ohia

Title: Weakly-supervised Topology-preserving 3D Mitochondria Segmentation

Author list: Elliott Seo[1], Saumya Gupta[1], Susanne Rafelski[1], Matheus Viana[1] and Chao Chen[1]

Detailed Affiliations:
[1]Stony Brook University.

Abstract: Accurate segmentation of mitochondria from microscopy images is crucial for understanding cellular function and pathology. However, it is a challenging task due to the complex network-like morphology of this organelle. Conventional deep learning approaches frequently predict segmentation with topological errors such as broken connections or false merges, undermining biological interpretations at the network topologylevel. Furthermore, the prohibitive cost of creating large-scale manual annotations has hindered progress in this domain. In this paper, we propose a novel two-stage framework that addresses both challenges: first by leveraging weak

labels for initial model pre-training, then by fine-tuning with a topologybased objective function on a small subset of manually corrected ground truth volumes. This approach significantly reduces the annotation burden while preserving topological integrity. We validate our method on a Fluorescence Microscopy mitochondria dataset, demonstrating superior performance in both conventional segmentation metrics and topological accuracy compared to traditional deep learning approaches. Our experiments confirm that even with a small subset of accurately annotated data, the proposed framework achieves biologically relevant segmentations by effectively combining weak supervision with topological constraints. This work offers a practical solution for accurate mitochondria segmentation that balances efficiency with biological fidelity, with potential applications in studying mitochondrial dynamics in various pathological conditions. Index Terms—mitochondria segmentation, topology, deep learning, weak supervision, fluorescence microscopy

Keywords: mitochondria segmentation, topology, deep learning, weak supervision, fluorescence microscopy

Title: Learning-Augmented Resource Allocation for Intelligent Healthcare Scheduling

Author list: Fei Li [1]

Detailed Affiliations:

[1]Department of Computer Science, George Mason University, 4400 University Drive, 22030, Fairfax, VA, USA.

Abstract: Motivation: We study scheduling under uncertainty with predictions, motivated by healthcare applications such as physician assignment and ICU capacity planning. We model patients as jobs and treat both ICU capacity and physicians as constrained resources. Results: Our objective is to minimize total cost, defined as the amount of critical resources required to feasibly serve all patients. We develop a stochastic programming framework for structured scheduling and extend it to incorporate predictions within the learning-augmented algorithms paradigm. Our approach leverages predicted treatment durations to guide scheduling decisions, enabling the system to exploit accurate predictions for improved efficiency while maintaining robustness to prediction errors. We evaluate our methods through simulations calibrated to MIMIC-style intensive care data. The results show that our prediction-augmented scheduling algorithms consistently outperform baseline methods in reducing resource usage, leading to reductions in the number of ICUs/physicians required under realistic workloads. Furthermore, performance degrades gracefully as prediction noise increases, demonstrating strong robustness. Availability: Code supporting the experimental results will be made available upon request. Contact: fli4@gmu.edu

Keywords: Healthcare scheduling, learning-augmented algorithms, resource allocation, stochastic programming

Title: Hyperspectral Deep Learning for Compositional Classification of Municipal Solid Waste: An Environmental Health Informatics Application

Author list: Jason Li[1], Enyonam Ahadzi[2], Jiaqiong Li[3], Lingling Liu[4], Beiwen Li[4], Jian Shi[2] and Jin Chen[5]

Detailed Affiliations:

[1]Westwood High School, Austin, TX, USA, [2]College of Agriculture, Food and Environment, University of Kentucky, Lexington, KY, USA, [3]College of Agriculture, Food and Environment, University of Wisconsin-Oshkosh, Oshkosh, WI, USA, [4]College of Engineering, University of Georgia, Athens, GA, USA, [5]School of Medicine, University of Alabama at Birmingham, Birmingham, AL, USA.

Abstract: Accurate compositional characterization of municipal solid waste (MSW) is a critical environmental health informatics challenge, with direct implications for reducing toxic emissions, preventing contamination, and advancing sustainable waste management. MSW is a highly heterogeneous mixture whose material complexity renders manual sorting methods inadequate at operational scale. In this work, we developed an image-based computational framework combining hyperspectral imaging (HSI) and machine learning (ML) to classify common waste materials from their spectral signatures. HSI captured discriminative spectral features supporting robust material discrimination, and deep learning further improved classification performance by extracting subtle spectral patterns beyond the reach of traditional methods. These findings establish HSI-driven AI classification as a viable and scalable approach to waste stream characterization, with broader applicability to environmental sensing and public health monitoring applications involving complex, high-dimensional data.

Keywords: Hyperspectral imaging, Artificial intelligence, Material classification

Title: XLCenter: crosslink-induced truncation site identification reveals sequence and structural specificities of RNA binding protein motifs

Author list: Anthony Cheng[1,2], Yuping Zhang[3] and Zhengqing Ouyang[1,*]

Detailed Affiliations:
[1]Department of Biostatistics and Epidemiology, School of Public Health and Health Sciences, University of Massachusetts, Amherst, 01003, MA, USA, [2]Department of Genetics and Genome Sciences, UConn Health, Farmington, 06269, CT, USA and [3]Department of Statistics, University of Connecticut, Storrs, 06269, CT, USA.

Abstract: Motivation: RNA-binding proteins (RBPs) are key factors that control the metabolism of transcripts. Elucidating the binding of an RBP has significant implications for its function. While computational methods exist to analyze RBP binding assays such as CLIP-seq, revealing the binding specificities and properties of RBP motifs is still a daunting problem. Results: We present XLCenter, an RBP crosslink-induced truncation sites (CITS) caller that combines a site-specific local Poisson model and a beta-uniform mixture model for controlling the false discovery rate. XLCenter resolves binding sites from single-nucleotide resolution assays. The CITS size is 7 nucleotides (nt) and recapitulates known canonical motifs of sequence-specific RBPs in a 25-nt neighborhood. By analyzing eCLIP datasets in the ENCODE data portal, we show how the CITS center is more effective than peaks for interpreting an RBP's function. Lastly, we show how the CITS center is vital for effective integration with single-nucleotide structural probing assays to understand the relationship between RBPs and in vivo RNA structure. Availability and Implementation: Source code implemented in Python is freely available at https://people.umass.edu/ouyanglab/xlcenter/ Contact: 715 N Pleasant St, Amherst, MA 01003, USA. Email: ouyang@schoolph.umass.edu Supplementary material: Supplementary material is attached.

Keywords: CLIP, RBP motif, RNA structure, CITS

Title: Genetic heterogeneity underlying breast cancer subtypes: genetic epidemiologic investigations to disentangle disease complexity

Author list: Xiang Shu[1]

Detailed Affiliations:
[1]Memorial Sloan Kettering Cancer Center.

Abstract: TBD

Keywords: TBD

Title: TBD

Author list: TBD

Detailed Affiliations:
[1]TBD

Abstract: TBD

Keywords: TBD

Title: TBD

Author list: TBD

Detailed Affiliations:
[1]TBD

Abstract: TBD

Keywords: TBD

TECHNOLOGY SESSION

August 3rd
4:00 – 6:00 PM
Room: 2213A

Chair: TBD

Title: Empowering genomic discovery: Accelerating regulatory biology through multi-omic solutions

Author list: Yun Zhao[1] and Zonghui Peng[1]

Detailed Affiliations:
[1]Admera Health, South Plainfield, NJ, USA

Abstract: Advances in next-generation sequencing (NGS) and multi-omic technologies have transformed regulatory biology by enabling comprehensive analyses of genomic, transcriptomic, epigenomic, and spatial molecular landscapes. As biological systems become increasingly recognized as highly interconnected, integrated multi-omic approaches have become essential for uncovering complex regulatory mechanisms, identifying novel biomarkers, and accelerating translational research.

Admera Health provides a comprehensive portfolio of genomics and bioinformatics services designed to support researchers throughout the entire project lifecycle, from study design and sample processing to sequencing, data analysis, and biological interpretation. Our capabilities include whole-genome sequencing (WGS), whole-exome sequencing (WES), targeted DNA sequencing, bulk and single-cell RNA sequencing, long-read sequencing using the PacBio Revio platform, spatial transcriptomics, methylation and epigenetic profiling, microbiome sequencing, and customized multi-omic workflows. These laboratory services are supported by validated analytical pipelines for genome alignment, variant discovery, gene expression quantification, differential expression analysis, pathway enrichment, immune repertoire profiling, and integrated multi-omic data interpretation.

By combining complementary sequencing technologies with robust computational analyses, researchers can characterize gene regulation across multiple biological layers, providing deeper insights into transcriptional regulation, chromatin accessibility, genomic variation, cellular heterogeneity, and disease-associated molecular pathways. Our flexible workflows support a broad range of applications in oncology, immunology, neuroscience, rare diseases, infectious diseases, and precision medicine.

This presentation will highlight how Admera Health's integrated multi-omic solutions enable high-quality, reproducible data generation and accelerate discoveries in regulatory biology. Through representative case examples, we will demonstrate how combining multiple sequencing modalities with advanced bioinformatics improves biological interpretation, facilitates biomarker discovery, and supports translational and clinical research. Attendees will gain practical insights into selecting appropriate multi-omic strategies to address complex biological questions and advance precision medicine.

Keywords: Multi-omics, Regulatory Biology, Genomics, Next-Generation Sequencing, Bioinformatics, Precision Medicine

Title: Unlocking Archived Pathology Specimens: Robust Single-Cell Transcriptomics from FFPE Tissue

Author list: Jin Zhou[1]

Detailed Affiliations:
[1]Singleron Biotechnologies Inc., Woodbridge, CT, USA

Abstract: Formalin-fixed, paraffin-embedded (FFPE) tissue archives represent one of the largest and most clinically annotated biological resources available. However, severe RNA degradation and chemical cross-linking have historically restricted FFPE transcriptomic analysis to bulk or spatially averaged approaches, limiting the application of single-cell RNA sequencing. Here, we present a novel single-cell RNA analysis technology designed for robust, whole-transcriptome profiling of FFPE tissue-derived single nuclei. By combining optimized nuclei extraction with random reverse transcription (RT) priming and subsequent cDNA tagging, this automated approach overcomes key technical barriers of archived sample analysis. Crucially, the random-priming strategy ensures coverage across the entire gene body, which enables mutation detection and permits the sequencing of both mRNA and non-coding RNAs at single-cell resolution. We demonstrate the high reproducibility, sensitivity, and biological

interpretability of this technology across multiple FFPE tissue types. Beyond technical benchmarks, we showcase its capacity to extract actionable biological insight including accurate cell-type annotation, disease-relevant expression profiles, and treatment-induced cellular composition changes from samples previously deemed incompatible with single-cell genomics. Suitable for routine testing, this method expands the scope of single-cell transcriptomics to archived clinical specimens, bridging the gap between historical pathology archives and single-cell resolution to power retrospective studies and translational discovery.

Keywords: Single cell RNAseq, FFPE tissue, clinical biological samples, mutation detection, non-coding RNA

Title: Protein interactomics by proximity networks of single cells (Zoom)

Author list: Michael Förster

Detailed Affiliations:
[1]Pixelgen.

Abstract: TBD

Keywords: TBD

Title: Use of Stereo-seq to dissect cell type spatial differences in human brain organoids and mouse eyeball (Zoom)

Author list: Elaine Lim

Detailed Affiliations:
[1]MGI.

Abstract: We will introduce the Stereo-seq technology, tools and packages developed by STOmics for Stereo-seq data analysis in this session, including the Standard Analysis Workflow (SAW), Stereo Data Analysis Solution (SDAS) and additional advanced analysis tools. We demonstrate 2 real world examples of spatial discoveries made from Stereo-seq data on human brain organoids (PFA-fixed sample) and mouse eyeball (fresh frozen sample).

Title: Built for What's Next: How 10x's Flex Apex, Xenium, and Atera Platforms Are Powering the AI Era of Biology

Author list: Aleksandar Janjic[1]

Detailed Affiliations:
[1]10x Genomics.

Abstract: 10x Genomics, known for pioneering single-cell biology through Chromium, continues to deliver the fixed-cell, high-throughput data needed to train and validate predictive models in genomics through its Flex Apex assay. With Xenium, 10x's in situ spatial platform, that reach extends into tissue context: Xenium and Flex already underpin large-scale industry and academic efforts, including CareDx's ImmuneScape program, generating high-resolution maps of the immune mechanisms underlying organ rejection, and Bioptimus's STELA initiative, the world's largest clinically linked spatial biology atlas for its multimodal AI model, M-Optimus. Atera, 10x's newly launched whole-transcriptome, single-cell-resolution spatial platform, adds substantially higher throughput and broader gene coverage to that toolkit. This talk covers what Flex, Xenium, and Atera each contribute today, data structure, scale, and current limitations, and where the near-term opportunities lie for teams building predictive models on single-cell and spatial data.

FUTURE SCIENTIST IN AI SESSION

August 4th
1:40 – 5:20 PM
Room: 1220

Chairs: Chi Zhang, Jingwen Yan, Arhan Patel

Title: Building a Cell Type-Specific Genetically Predicted Gene Expression Resource in the Indiana Biobank: Application to Rheumatoid Arthritis

Author list: Derek Ai[1]

Detailed Affiliations:
[1]Park Tudor High School.

Abstract: TBD

Keywords: TBD

Title: Multimodal protein function inference via vision-language models and structural similarity representations

Author list: Christopher Barker[1]

Detailed Affiliations:
[1]Department of Computer Science, Pacific Lutheran University, Tacoma 98447, United States.

Abstract: TBD

Keywords: TBD

Title: Machine Learning Analysis of Facial Phenotype Features for Rare Disease Classification

Author list: Joshua Bie[1]

Detailed Affiliations:
[1]Harvard-Westlake School.

Abstract: TBD

Keywords: TBD

Title: Integrative Machine Learning and Transcriptomic Analysis of Severe Cardiovascular Adverse Drug Events Associated with Sotalol and Drug–Drug Interactions

Author list: Alyssa Chen[1]

Detailed Affiliations:
[1]Carrollwood Day High School.

Abstract: TBD

Keywords: TBD

Title: Pilot Evaluation of Seventeen Large Language Models on Supplement-Associated Disease Recommendations Using Literature-Based Evidence

Author list: Tim Chu[1]

Detailed Affiliations:
[1]Forest Hills Central High School.

Abstract: TBD

Keywords: TBD

Title: Prediction of cellular senescent programs in affecting retinoblastoma development
Author list: Bianca Ehling[1]
Detailed Affiliations:
[1]The Ohio State University.
Abstract: TBD
Keywords: TBD

Title: AD-Stage-Net: Cross-Dataset AI Classification for Alzheimer's Disease Using Brain MRI
Author list: Katelyn Hur[1]
Detailed Affiliations:
[1]Red River High School.
Abstract: TBD
Keywords: TBD

Title: Hyperspectral Deep Learning for Compositional Classification of Municipal Solid Waste: An Environmental Health Informatics Application
Author list: Jason Li[1]
Detailed Affiliations:
[1]Westwood High School, Austin, TX, USA
Abstract: TBD
Keywords: Hyperspectral imaging, Artificial intelligence, Material classification

Title: Three-Dimensional Genome Analysis Reveals Enrichment of Cancer Cell-Free DNA Methylation Markers in Regulatory Genome Architecture
Author list: Isabella Jiayi Li[1]
Detailed Affiliations:
[1]Westlake High School.
Abstract: TBD
Keywords: TBD

Title: Network Rewiring Reveals Coordinated Metabolic, Structural, and Immune-Proteostasis Programs in Heart Failure
Author list: Addison Liu[1], Sandali Dewni Lokuge[2], Sheng Liu[3], Jun Wan[2,3,4]
Detailed Affiliations:
[1]Park Tudor School, Indianapolis, IN, USA.
Abstract: TBD
Keywords: TBD

Title: mRehab 2.0: A Voice-Guided, Intelligent System for Home Stroke Rehabilitation
Author list: Emily Liu[1]
Detailed Affiliations:
[1]The Bishop Strachan School.
Abstract: TBD
Keywords: TBD

Title: Molecular profiling of the mystic human unique heterochromatin region of the Y chromosome
Author list: Ping Liang[1,2]
Detailed Affiliations:
[1]Department of Biological Sciences, Brock University, St. Catharines, Ontario, Canada.
[2]Centre for Biotechnology, Brock University, St. Catharines, Ontario, Canada.
Abstract: TBD
Keywords: TBD

Title: Computational hardness and flow-based optimization for neural connectivity inference
Author list: Amy Li[1]
Detailed Affiliations:
[1]Canyon Crest Academy, San Diego, CA 92130, USA.
Abstract: TBD
Keywords: TBD

Title: Association of Donor-Derived Cell-Free DNA with Biopsy-Proven Rejection in Kidney Transplant Recipients
Author list: Emma K. Molnar[1]
Detailed Affiliations:
[1]Waterford School.
Abstract: TBD
Keywords: TBD

Title: Comparative Structural and Mutational Analysis of CIRBP Reveals Evolutionary Conservation of Protein Stability Determinants
Author list: Christina Teng[1]
Detailed Affiliations:
[1]Thomas S. Wootton High School.
Abstract: TBD
Keywords: TBD

Title: Effects of Preprocessing Strategies and Covariate-Adjusted PCA on Population Stratification Detection for GWAS
Author list: Alexander Wu[1]
Detailed Affiliations:
[1]Arizona State University.
Abstract: TBD
Keywords: TBD

Title: Multimodal Speech Sound Disorder Assessment Using Ultrasound Tongue Imaging and Acoustic Signals
Author list: Austin Xu[1]
Detailed Affiliations:
[1]Williamsville East High School.
Abstract: TBD
Keywords: TBD

Title: Learning Brain Dynamics from Data: From Small Networks to a Digital Twin Zebrafish
Author list: Shaoyu Xu[1]
Detailed Affiliations:
[1]Shanghai World Foreign Language Academy.
Abstract: TBD
Keywords: TBD

Title: Dissecting Bladder Cancer Recurrence Through Spatial Molecular Profiling
Author list: Adrian Z. Yang[1]
Detailed Affiliations:
[1]Lawrence E. Elkins High School.
Abstract: TBD
Keywords: TBD

Title: Correlation between Motor Coordination and Behavior in Children with Autism: A Cross-Sectional Study
Author list: Henry Yang[1]
Detailed Affiliations:
[1]Sage Hill School, Newport Coast, CA, USA.
Abstract: TBD
Keywords: TBD

Title: Decoding Morphology–Gene Expression Relationships in Colon Cancer Using Spatial Transcriptomics
Author list: Alina Yu[1]
Detailed Affiliations:
[1]Germantown Friends School.
Abstract: TBD
Keywords: TBD

Title: NOTCH Pathway Downregulation and Transcription Factor Reprogramming in Anti-PD-1 Resistance in Clear Cell Renal Cell Carcinoma: A Transcriptomic Analysis
Author list: Daniel Zhang[1]
Detailed Affiliations:
[1]Canyon Crest Academy.
Abstract: TBD
Keywords: TBD

Title: Single-Cell Transcriptomics and Spatial Validation Identify SLIT2 Signaling as a Glomerular-Associated Pathway in Diabetic Kidney Disease
Author list: Simon Zhang[1]
Detailed Affiliations:
[1]Lexington High School.
Abstract: TBD
Keywords: TBD

Title: Computational analysis of metabolic changes and precision nutrition for atherosclerosis
Author list: Eva Zhao[1]
Detailed Affiliations:
[1]Middlesex School.
Abstract: TBD
Keywords: TBD

Title: Investigating the Causal Impact of COVID-19 on Acute Myocardial Infarction: A Multi-ancestry Mendelian Randomization Study
Author list: Katherine Zhou[1]
Detailed Affiliations:
[1]Choate Rosemary Hall (High School), Wallingford, CT, USA.
Abstract: TBD
Keywords: TBD

Title: Ensemble Machine Learning Models For Early Sepsis Detection
Author list: Shashvat Hariharan[1]
Detailed Affiliations:
[1]Linganore High School, Frederick, MD, USA
Abstract: TBD
Keywords: TBD

FLASH TALKS

Flash talks
August 3rd
4:00 – 6:00 PM
Room: 1220

Chair: Tong Wang

Title: CO-PICO/PECO: Condition-based Occupational PICO/PECO and its Applications for LLM extraction and Ontological Modeling of Structured Occupational Safety and Health Evidence from Biomedical Literature
Author list: Hasin Rehana[1,2], Jie Zheng[3], Yongqun He[3,4,§], and Junguk Hur[2,§]
Detailed Affiliations:
[1]School of Electrical Engineering & Computer Science, University of North Dakota, Grand Forks, North Dakota, 58202, USA, [2]Department of Biomedical Sciences, University of North Dakota School of Medicine and Health Sciences, Grand Forks, North Dakota, 58202, USA, [3]Unit for Laboratory Animal Medicine, Department of Learning Health Sciences, University of Michigan, Ann Arbor, Michigan, 48109, USA, [4]Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan, 48109, USA. [§]Corresponding author
Abstract: Motivation: Occupation is a critical social determinant of well-being, influencing exposure to risks and health outcomes across diverse work environments. However, most occupational safety and health (OSH) evidence is buried in unstructured texts, which limits the scalability of systematic synthesis and ontology development. While PICO (Population, Intervention, Comparison, and Outcome) and PECO (Population, Exposure, Comparison, and Outcome) frameworks have been applied in clinical domains, occupation-centered structured modeling remains underexplored. Moreover, traditional PICO and PECO frameworks do not explicitly capture occupational conditions or the specific working environment/context in which exposure occurs (e.g., frontline work, pandemic

settings), limiting their use in occupational evidence modeling. Method: We developed a condition-focused occupational PICO/PECO (CO-PICO/PECO) framework that explicitly incorporates occupational conditions into population definition. We implemented a Large Language Model (LLM) pipeline using GPT-4o with a schema-constrained prompting to extract standardized records capturing Population (occupation and contextual condition), Intervention/Exposure, Comparison group (occupation and contextual condition), and Outcome from OSH-related literature. Extracted entities were normalized and mapped to the Occupation Ontology (OccO) to ensure consistency and enable downstream analysis. Various occupational conditions were also ontologized. This structured outcome provides a consistent representation of occupation–intervention/exposure–outcome relations, forming a foundation for downstream ontology development and knowledge graph integration. As a proof-of-concept, we applied the framework to occupationally relevant COVID-19 abstracts and produced 10,105 CO-PICO/PECO records. Results: The dominant population groups were healthcare workers (n=3,286), workers/other (n=1,931), nurses (n=1,207), and physicians (n=526). The ontological classification of various occupations and conditions made the results easier to explain. The most frequent exposures/interventions were COVID-19 pandemic context (n=1,010), SARS-CoV-2 exposure (n=856), SARS-CoV-2 infection (n=285), and COVID-19 vaccination (n=281). The most frequent population conditions were COVID-19 context (n=2,192), working during COVID-19 (n=234), working frontline (n=193), and treating COVID-19 patients (n=190). Outcomes were predominantly mental health and infection-related, including mental health burden (n=1,693), risk of infection (n=826), burnout (n=366), and seroprevalence/seropositivity (n=351). These patterns highlight how occupational conditions shape exposure and health outcomes. Conclusion: The CO-PICO/PECO framework with LLM enabled an automated, scalable conversion of unstructured OSH-related literature into structured, ontology-ready data. Explicitly modeling occupational conditions improves condition-aware risk analysis and supports category-level synthesis and systematic OSH literature mining beyond manual curation. Availability: Source code is available at https://github.com/hurlab/CO-PICO-PECO-LLM.

Keywords: CO-PICO/PECO, Large Language Model, Occupational Safety and Health, COVID-19, Ontology, OccO

Title: CIRCA identifies cell-and niche-level interactions with amyloid-β plaques using causal representation learning

Author list: Tyler C Lovelace[1], Chia-Hao Tung[1], Zhongming Zhao[1]

Detailed Affiliations:

[1]Center for Precision Health, McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA.

Abstract: Alzheimer's disease (AD) affects one in nine Americans over age 65, and this population is expected to double by 2050. AD arises from complex interactions among cells, local tissue environments, and pathology such as amyloid-β (Aβ) plaques. Spatial transcriptomics enables these interactions to be studied in situ, but sparse gene expression measurements and the non-identifiable embeddings learned by existing deep learning-based methods limit the recovery of biologically meaningful latent cell states, spatial niches, and cell-niche interactions. To address these limitations, we developed CIRCA (Causal Interaction Representations for Cellular Architecture), a causal representation learning framework that learns identifiable regulatory programs capturing interactions among cells, local neighborhoods, and disease-associated pathology when available. We evaluated CIRCA using two cross-slide integration tasks constructed from sequential MERFISH sections from the Allen Brain Cell Atlas and compared it with NicheCompass, Novae, and BANKSY using cell type and anatomical annotations, spatial continuity, and batch integration metrics. We then applied CIRCA to a CosMx dataset from an AD mouse model with paired Aβ immunofluorescence and linked reproducible latent factors to gene expression programs to identify genes and pathways associated with cell-, niche-, and pathology-related factors. Across benchmark tasks, clustering of the CIRCA niche latent space better captured fine-grained anatomical regions and improved slide integration relative to the best baseline method for each metric, increasing AMI from 0.561±0.003 to 0.587±0.003 in coronal sections and from 0.578±0.003 to 0.590±0.008 in sagittal sections, while also improving kBET slide integration from 0.300±0.003 to 0.376±0.003 and from 0.338±0.005 to 0.457±0.010, respectively. CIRCA cell-state embeddings recovered both coarse- and fine-grained cell types, while the spatial cell latent space integrated cell identity with spatial context. In the AD mouse model, CIRCA identified major brain cell types, anatomically coherent healthy tissue niches, and three distinct plaque-associated niches: a cortical plaque niche, a hippocampal stratum plaque

niche, and a CA1 plaque niche. The model also consistently recovered two Aβ-associated latent factors: one reflecting high Aβ signal near the central cell and another reflecting high Aβ signal across the broader local neighborhood. Across random subsamples, 34 of 50 cell factors and 36 of 50 niche factors were reproducibly recovered in at least two of three runs. Annotation of high-confidence factors linked Aβ-associated spatial programs to microglial phagocytosis, inflammatory cytokine signaling, and antigen processing and presentation.

Keywords: Alzheimer's disease, spatial transcriptomics, amyloid-β pathology, cell-niche interactions, causal representation learning, deep learning

Title: Machine Learning Augmented Mechanistic Modeling for Identifying Predictive Tumor Biomarkers

Author list: Alexander R. J. Silalahi[1] and Joseph D. Butner[1]

Detailed Affiliations:

[1]MD Anderson Cancer Center.

Abstract: We present a mechanistic–radiomics fusion framework to identify predictive tumor biomarkers and improve treatment response modeling, which we have evaluated by investigating two complementary strategies. First, we use a serial approach that comprises a mechanistic model followed by a radiomics-driven machine learning model (elastic net regression). The mechanistic model is used to generate baseline predictions, and the elastic net regression is trained on the residual errors of the mechanistic model to capture unexplained variability. Feature selection is performed using penalized regression to identify a parsimonious set of informative biomarkers. Second, as guided by the feature selection analysis, we incorporate selected radiomic features (surface area, volume) directly into the mechanistic model. Although these features are not always the highest-ranked in terms of predictive importance, they are interpretable and can be readily integrated into the mechanistic model. Our analysis found that this improves the predictive capability of the mechanistic model. By taking advantage of the interpretability and robustness of mechanistic modeling and the high-dimensional descriptive power of radiomics, we aim to improve predictive performance and to enable more reliable identification of imagingbased tumor biomarkers for radiotherapy response.

Keywords: mechanistic modeling, radiomics, radiotherapy, machine learning

Title: Multimodal Protein Function Inference via Vision-Language Models and Structural Similarity Representations

Author list: Christopher Barker[1], Matthew Hou[1], Boen Liu[1], Brandon Ma[1], Xiao Chen[2], Jie Hou[3], Dong Si[4], Renzhi Cao[1*]

Detailed Affiliations:

[1]Department of Computer Science, Pacific Lutheran University, Tacoma 98447, United States, [2]Computer Science Department , Hamilton College, United States, [3]Department of Computational Mathematics, Science and Engineering, Michigan State University, United States, [4]School of Science, Technology, Engineering & Mathematics, University of Washington, USA, United States. [*]Corresponding author

Abstract: Predicting protein function from sequence remains a major challenge in computational biology. In this work, we propose a multimodal pipeline that embeds protein sequences using ESM-3, converts them into pairwise similarity matrices, and visualizes these representations as 2D heatmaps. These visual representations are processed by a BLIP-2 vision-language model to generate natural language descriptions, which are then translated into functional annotations using a rule-based mapping aligned with Gene Ontology terminology and evaluated via semantic similarity scoring. Through iterative prompt engineering across six experimental batches, two critical failure modes were identified: mode collapse and visual hallucination. To address these issues, a structurally constrained dynamic prompting architecture was developed, enforcing biologically grounded interpretation while preserving generative flexibility. The final pipeline achieves statistically significant improvement over baseline performance while maintaining functional diversity across predictions. These results demonstrate a scalable and interpretable framework for multimodal protein function inference and establish prompt design as a key control mechanism in scientific multimodal reasoning.

Keywords: protein function inference, multimodal learning, vision-language models, ESM-3, BLIP-2, Gene Ontology, prompt engineering

Title: scDCBC: Model-based scRNA-seq Deep Clustering with Batch Correction

Author list: Xiaohe Chen,[1] Zeliang Sun,[2] Yangshuang Xu,[1] Yue Julia Wang[3] and Chao Huang[2,*]

Detailed Affiliations:

[1]Department of Statistics, Florida State University, 117 N Woodward Ave, Tallahassee, 32304, FL, USA, [2]Department of Epidemiology & Biostatistics, University of Georgia, 101 Buck Rd, Athens, 30606, GA, USA, [3]College of Medicine, Florida State University, 1115 W Call St, Tallahassee, 32304, 32304, USA. [*]Corresponding author

Abstract: Unsupervised clustering is widely used in single-cell RNA sequencing (scRNA-seq) data analysis to identify distinct cell populations. Still, dropout events, high dimensionality, and batch effects hinder its effectiveness in multi-batch data. Although existing deep clustering methods improve representation learning for sparse count data, they often do not explicitly model batch heterogeneity within the clustering process, limiting their performance in integrated analyses. Here, we present scDCBC, a deep clustering framework for multi-batch scRNA-seq data that unifies batch correction and cell type clustering. scDCBC uses a zero-inflated negative binomial autoencoder to learn latent representations from sparse count data and introduces a cluster-matching strategy to identify similar cell populations across batches. Matched clusters are aligned in the latent space through maximum mean discrepancy regularization, while a self-training objective further improves clustering assignments. We benchmarked scDCBC against widely used scRNA-seq integration and clustering methods on multiple public datasets. scDCBC effectively corrected batch effects while preserving cell-type separation and achieved competitive overall performance across diverse benchmarking scenarios. Together, these results support scDCBC as a unified and adaptable framework for multi-batch scRNA-seq clustering analysis.

Title: A systematic investigation of molecular encoding methods for drug property predictions across neural network and Transformer encoder-based model

Author list: Sheng-Ya Chen[3,4] and Shan-Ju Yeh[1,2,3*]

Detailed Affiliations:

[1]School of Medicine, National Tsing Hua University, Hsinchu, Taiwan, [2]Institute of Bioinformatics and Structural Biology, National Tsing Hua University, Hsinchu, Taiwan, [3]Department of Life Science, National Tsing Hua University, Hsinchu, Taiwan. [4]Interdisciplinary Program of Life Sciences and Medicine, National Tsing Hua University, Hsinchu, Taiwan. [*]Corresponding author

Abstract: Fundamental investigations into how different molecular encoding methods affect molecular property prediction remain relatively limited. In this study, we extensively examined the optimal molecular encoding methods for molecular properties prediction using two prevalent structure designs: a classical neural network model (MLP) and a Transformer encoder-based model (MLP+TL). For molecular encoding methods, we investigated several types of fingerprints, including traditional topological fingerprints, substructure-based fingerprints, and string-based representations. These two models were trained on seven well-known molecular datasets to evaluate different input molecular encoding methods based on evaluation metrics. On several biologically relevant classification tasks, including toxicity, mutagenicity, and side-effect prediction, our models consistently achieved average AUC values above 0.9. Beyond model performance, we further explored whether the Transformer encoder-based framework could provide interpretable insight. Rather than relying on external post-hoc explanation methods such as the local interpretable model-agnostic explanation (LIME) or the Deep SHapley Additive exPlanations (SHAP), we leveraged the model's intrinsic attention weights as an internal interpretability signal for identifying potentially important feature. To further support this analysis, we conducted case studies on the BBBP and MUTAG datasets. The MLP+TL model using MACCS and PubChem as input can capture chemically interpretable groups that determined the major blood-brain barrier (BBB) permeability and mutagenicity in Salmonella typhimurium. In particular, a comparison between Morphine and Heroin highlighted the role of hydroxyl-related substructures in BBB permeability prediction, which was consistently reflected in the attention weights. Overall, our findings provide practical guidance for selecting effective molecular encoding methods and contribute to the development of interpretable molecular informatics approaches for drug discovery.

Title: Beyond PCA: Multiple Correspondence Analysis for Robust Cell-Type Annotation and Trajectory Inference

Author list: Tsai-Jung Lu[#,1,2,3], Tzu-Hung Hsiao[#,2,4,5], Hui-Mei Tsai[1,2,3], Eric Y. Chuang[1,6,7,8, §], Yidong Chen[3,9, §]

Detailed Affiliations:

[1]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei 106319, [2]Taiwan Department of Medical Research, Taichung Veterans General Hospital, Taichung 407219, Taiwan, [3]Greehey Children's Cancer Research Institute, University of Texas Health San Antonio, San Antonio, TX 78229, USA, [4]Department of Public Health, Fu Jen Catholic University, New Taipei 24205, Taiwan, [5]Institute of Genomics and Bioinformatics, National Chung Hsing University, Taichung 402202, Taiwan, [6]Biomedical Technology and Device Research Laboratories, Industrial Technology Research Institute, Hsinchu 310401, Taiwan, [7]Research and Development Center for Medical Devices, National Taiwan University, Taipei 106319, Taiwan, [8]Department of Electrical Engineering, College of Electrical Engineering and Computer Science, National Taiwan University, Taipei 106319, Taiwan, [9]Department of Population Health Sciences, University of Texas Health San Antonio, San Antonio, TX 78229, USA. [#]Contributed equally, [§]Corresponding author

Abstract: Single-cell RNA sequencing (scRNA-seq) enables analysis of cellular heterogeneity and dynamic biological processes, yet standard workflows separate cell representation from genelevel interpretation. Cells are embedded using PCA followed by nonlinear projections such as UMAP or t-SNE, which can distort global relationships and fragment cell-state connections, while gene functions are inferred post hoc. Moreover, trajectory inference reconstructs cellular transitions but often depends on predefined clusters and is not intrinsically linked to biological functions. Here, we present a Multiple Correspondence Analysis (MCA)-based dimension reduction approach that jointly represents cells and genes, enabling direct evaluation of cell-gene relationships. Proximity marker genes (PMGs) for each cell are identified using a distancebased framework and a permutation-based statistic without predefined group comparisons. These genes are used for nonlinear projection, joint cell-gene clustering, cell-type annotation, and trajectory inference within the same space. For trajectory analysis, the root state can be easily defined by its co-localization with marker genes. To link trajectory structure with function, segment-level enrichment analysis is performed, enabling gene programs to be evaluated continuously along differentiation paths. Applied to human hematopoiesis, we demonstrated that the proposed approach recovers known lineage relationships and functional transitions, all achieved without performing cluster- 3 based differential expression analysis. Therefore, the MCA-based dimension reduction method enables direct and continuous interpretation of gene programs along trajectories.

Keywords: TBD

Title: TBD

Author list: TBD

Detailed Affiliations:

[1]TBD

Abstract: TBD

Keywords: TBD

Title: TBD

Author list: TBD

Detailed Affiliations:

[1]TBD

Abstract: TBD

Keywords: TBD

Flash Talks
August 4th
1:40 – 5:20 PM
Room: 2220A&B

Chair: Jun Wan

Title: Cross-Species Gene Set and Cell Type Annotation Through Large Language Models and Agentic AI with Literature-Grounded Validation

Author list: Wenbo Chen[1], Rongbin Li[1], Neha Maurya[1], Arnav Solanki[1], Meaghan Ramlakhan[1], Zhuhao Wu[2], W. Jim Zheng[1,#]

Detailed Affiliations:
[1]McWilliams School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, USA, [2]Appel Alzheimer's Disease Research Institute, Feil Family Brain and Mind Research Institute, Weill Cornell Medicine, New York, NY, USA. [#]Corresponding author

Abstract: Background Single-cell RNA sequencing (scRNA-seq) has transformed our ability to identify diverse cell types and their transcriptomic signatures. Despite rapid progress, annotating newly discovered or poorly characterized cell clusters remains a major challenge in biomedical research. Traditional approaches such as Gene Set Enrichment Analysis (GSEA) rely on predefined gene sets that are often incomplete for novel biological contexts. Although large language models (LLMs) leverage vast biomedical literature, their application to cell type annotation is limited by insufficient grounding in structured biological ontologies. Therefore, there is a critical need for an advanced computational framework that can generate reliable and biologically meaningful cell type annotations from scRNA-seq data across species and contexts. Methods We leveraged fine-tuned LLMs to develop an agentic AI system to perform cross-species annotation for cell types. We first extracted gene modules from scRNA-seq datasets, which group genes into co-expression modules based on cell-cell similarity. We then developed a multi-agent AI framework by fine-tuning LLMs including Gemma and Llama architectures on curated human and mouse gene sets from MSigDB, separately. Retrieval-augmented generation (RAG) was incorporated to ground predictions in peer-reviewed literature from PubMed, reducing hallucinations and improving interpretability. The framework takes gene lists derived from cell clusters as input and generates literature-supported functional annotations. Performance was evaluated using accuracy and semantic similarity metrics, and domain experts conducted human evaluation to assess biological relevance. Results Task-specific fine-tuned models consistently outperformed models without fine-tuning, demonstrating that targeted training on structured biological knowledge is essential for accurate gene set annotation. Across benchmark gene sets, 77% of mouse and 74% of human gene sets achieved correct annotations among the top predictions, outperforming existing LLM-based approaches. Retrieval grounding further improved prediction reliability in both species. As a key application, we applied the framework to large-scale brain atlases with 5,322 mouse and 3,313 human whole-brain scRNA-seq clusters and generated biologically meaningful cell type descriptions that captured both conserved and species-specific transcriptional programs. Together, this end-to-end pipeline bridges the gap from raw scRNA-seq data to meaningful biological annotations across species. Conclusion By combining species-specific fine-tuning with retrieval-based validation, our framework enables accurate and literature-grounded gene set and cell type annotation at scale. This approach enables us to comprehensively evaluate marker gene sets and cell types for a broad biomedical domain where accurate functional annotation is critically important for the mechanistic understanding of disease and therapeutic development.

Title: Drug-drug interaction prediction using proteomic scale signatures within the CANDO platform

Author list: Lucille Tomin[1], Sarah Mullin, PhD[2], Ram Samudrala, PhD[1], Zackary Falls, PhD[1,2]

Detailed Affiliations:

[1]Department of Biomedical Informatics, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, Buffalo, NY, [2]Roswell Park Comprehensive Cancer Center, Buffalo, NY.

Abstract: Prescription medication uses in the United States increased by nearly 85% over the past twenty years. The number of older adults routinely taking five or more medications has also more than tripled from 12.8 to 39%. The use of concomitant medications, often from multiple prescribers, place patients at an increased risk for experiencing drug-drug interaction-induced side effects. It is estimated that over 22% of adverse drug reaction (ADR) related hospital admissions can be contributed to drug-drug interactions (DDIs). Accurate prediction of these events is therefore critical for improving patient safety and public health. Current DDI prediction tools lack explainability of identified interactions and contain minimal overlap of ADRs between different software and databases. We aim to overcome these shortcomings via the Computational Analysis of Novel Drug Opportunities (CANDO) platform. We have developed the platform's DDI assessment module based on compound-pair proteomic signature similarity enabling the accurate prediction of ADRs induced by any compound-pair. The drug-proteome interaction methodology scores interactions for 8,385 human proteins for every compound in our drug library producing their proteomic interaction signatures which are then used to predict DDI-induced ADRs based on similarity to known pairwise interactions. The prediction model was benchmarked using our standard leave-one-out method previously developed for the CANDO paradigm adapted for ADR prediction based on compound-pairs. DDI-ADR mappings were curated from the comprehensive structured ADR and drug labeling datasets from DrugBank. Our gold-standard DDI dataset includes 10,126 unique associations, 376 unique drugs, 377 unique drug-pairs, and a total of 1,597 unique ADRs. Given that at least two compound-pair associations are needed for leave-one-out benchmarking, we were able to use 940 ADRs for validation. The top 10 ADR prediction accuracy improved 9.4x compared to random control (2.65% to 24.92%) and top 100 accuracy improved 2.3-fold (26.53% to 61.45%). Prediction on an external validation set curated from the open source NSIDES database produced an average of 15.7% hit rate for the top 25 cutoff mark. The CANDO-DDI module benchmarking results demonstrate effective computational evaluation of compound-pairs for potential ADRs. Moreover, the system-level proteomic signature approach within CANDO enables researchers to investigate which proteins are associated with the prediction of a given DDI-ADR. Future work will focus on improving mapping methods, categorizing ADRs based on severity, expanding to higher order drug interactions, and validating with external databases or clinical datasets.

Keywords: TBD

Title: L2b-Probiotics: A Fusion Genetic Language Model for Probiotics Identification

Author list: Caihao Sun[1], Shuwen Hou[1], Jiachao Zhang[1], Pan Huang[1], Qixiao Zhai[1] and Shi Huang[1]

Detailed Affiliations:

[1]Faculty of Dentistry, The University of Hong Kong, Hong Kong SAR, China.

Abstract: Probiotic discovery increasingly relies on computational screening of microbial genomes, yet existing methods either depend on handcrafted sequence features or lack biologically interpretable functional evidence. We present L2b-Probiotics (Language 2bRAD-M-Probiotics), a hybrid framework that integrates genomic language modeling with gene-level functional annotation for probiotic identification. The method first applies 2bRAD-M genome sketching to generate compact, biologically informative genome representations. A pretrained Skip-Gram genomic language model learns contextual k-mer embeddings, which are processed by a four-block 1D-CNN to extract sequence-level features. In parallel, a gene-level branch uses Prokka annotation and log-odds ratio scoring to quantify genes enriched in probiotic versus non-probiotic genomes. The two feature streams are fused for final classification. On the iProbiotics benchmark, L2b-Probiotics achieved 98.46% accuracy, outperforming iProbiotics (91.98%) and metaProbiotics (95.38%). The model remained robust under simulated genome incompleteness, maintaining 89.23% accuracy even at 5% genome completeness. Biological analyses showed that the learned embedding space captures functionally meaningful signals, while external validation demonstrated applicability across independent metagenomic datasets. Screening 5,785 gut microbial genomes required approximately 10 minutes and identified 90 high-potential probiotic candidates. In a clinical cohort, predicted probiotics were 2.2-fold more abundant in healthy than diseased microbiomes. These results demonstrate that combining biologically

grounded genome sketching, language-model representations, and interpretable gene-level evidence provides an accurate and scalable strategy for large-scale probiotic discovery.

Keywords: Probiotics; Genomic language model; Microbiome; Deep learning; Genome annotation; Probiotic Identification

Title: Single-Cell Viral Detection from Unmapped scRNA-seq Reads

Author list: Mohammadamin Mahmanzar[1,2] , Karim Rahimian[3] , Kaveh Kavousi[2,3*], Youping Deng[1*]

Detailed Affiliations:
[1]Department of Quantitative Health Sciences, John A. Burns School of Medicine, University of Hawaii at Manoa, Honolulu, HI 96813, USA, [2]Department of Bioinformatics, Kish International Campus University of Tehran, Kish, Iran, [3]Laboratory of Complex Biological Systems & Bioinformatics (CBB), Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran. [*]Corresponding author

Abstract: Introduction: Unmapped reads in single-cell RNA sequencing (scRNA-seq) are routinely discarded despite their potential to encode non-host biological signals. However, interpretation of these reads remains controversial due to contamination risks, low-abundance noise, and lack of cellular resolution. Here, we present a bioinformatics-driven framework to systematically recover and validate viral signals from unmapped scRNA-seq reads, enabling reconstruction of the colorectal cancer (CRC) virome at single-cell resolution. Methods: We analyzed 570 scRNA-seq samples from 179 patients across healthy, tumor-adjacent, and CRC tissues. A barcode-aware computational pipeline was developed to recover unmapped reads, align them against a curated human-associated viral reference, and restore single-cell identity. High-confidence filtering (≥90% identity, ≥40 bp) minimized spurious alignments. Viral signals were integrated with cell-type annotation (SingleR/Seurat) and analyzed across tissue condition, anatomical location, and cellular compartments. A multi-level validation framework was implemented, including (i) cross-dataset reproducibility, (ii) paired within-patient consistency, and (iii) synthetic spike-in benchmarking (3D validation: dataset-level, patient-level, and signal-level). Machine learning (CatBoost) was applied to evaluate predictive power of viral features. Results: Global viral burden remained stable across conditions, while virome composition underwent significant restructuring. Healthy tissues exhibited higher diversity and compositional stability, whereas tumor-adjacent and CRC tissues showed reduced diversity and increased heterogeneity. Viral signals were highly structured across cell types, with maximal burden in epithelial cells and strongest condition-dependent remodeling in immune and stromal compartments. Spatial analysis revealed pronounced virome restructuring in distal colorectal regions, particularly the rectosigmoid colon. Paired analyses demonstrated coordinated within-host redistribution of viral signals, characterized by depletion in immune/stromal cells and retention in epithelial compartments during tumorigenesis. Machine learning achieved robust classification performance (macro AUC ≈ 0.77), indicating that viral features encode reproducible biological signatures rather than stochastic noise. Discussion: This study establishes unmapped scRNA-seq reads as a biologically informative resource when coupled with rigorous computational and multi-level validation. The observed virome remodeling reflects structured, cell-type– and location-specific dynamics rather than random contamination, supporting a model in which viral signals are integrated within tumor microenvironment organization. By combining single-cell resolution with 3D validation across datasets, patients, and synthetic controls, this work provides a robust bioinformatics framework for viral discovery and highlights the role of latent viral signatures in cancer biology.

Title: STEP: Spatial Transcriptomics Embedding Procedure for Multi-scale Biological Heterogeneities Revelation in Multiple Samples

Author list: Lounan Li[1,2], Zhong Li[1,2], Xiao-ming Yin[3], Xiaojiang Xu[3]

Detailed Affiliations: [1]School of Information Engineering, Huzhou University, Huzhou, Zhejiang, 313000, China, [2]College of Science, Zhejiang Sci-Tech University, Hangzhou, Zhejiang, 310018, China, [3]Pathology & Laboratory Medicine, School of Medicine, Tulane University, New Orleans, LA, 70112, USA

Abstract: In the realm of spatially resolved transcriptomics (SRT) and single-cell RNA sequencing (scRNA-seq), addressing the intricacies of complex tissues, integration across non-contiguous sections, and scalability to diverse data resolutions remain paramount challenges. We introduce STEP (Spatial Transcriptomics Embedding Procedure), a novel foundation AI architecture for SRT data, elucidating the nuanced correspondence between biological heterogeneity and data characteristics. STEP's innovation lies in its modular architecture, combining a

Transformer and β-VAE based backbone model for capturing transcriptional variations, a novel batch-effect model for correcting inter-sample variations, and a graph convolutional network (GCN)-based spatial model for incorporating spatial context—all tailored to reveal biological heterogeneities with un-precedented fidelity. Notably, STEP effectively scales the newly proposed 10x Visium HD technology for both cell type and spatial domain identifications. STEP also significantly improves the demarcation of liver zones, outstripping existing methodologies in accuracy and biological relevance. Validated against leading benchmark datasets, STEP redefines computational strategies in SRT and scRNA-seq analysis, presenting a scalable and versatile framework to the dissection of complex biological systems.

Flash Talks
August 4th
4:40 – 5:20 PM
Room: 2213A

Chair: Zhaohui Xu, Xiaojiang Xu

Title: SomAtt: A Foundation Model for the Tumor Genome

Author list: David Arredondo[1], Ala Jararweh[2], Luis Tafoya[1], Avinash Sahu[1]

Detailed Affiliations:
[1]Division of Molecular Medicine, University of New Mexico, Albuquerque, NM, USA, [2]Jordan University of Science and Technology, Irbid, Jordan.

Abstract: Somatic mutations drive diverse phenotypes that vary with mutation type and location within a gene, while interactions between mutated genes further shape tumor heterogeneity, pathology, and clinical outcome. Current computational methods often treat all mutations within a gene as functionally equivalent, ignore combinatorial effects between genes, and struggle to integrate the many sequencing platforms and gene panels used across datasets. We introduce SomAtt (Somatic Attention), a foundation model for tumor genomes that uses self-attention to model interactions among mutations within a tumor. Pre-trained on 215,410 tumors from 191,001 unique cancer patients spanning 73 sequencing panels, SomAtt generates tumor embeddings for diverse downstream tasks and addresses three challenges: (a) distinguishing different mutations within the same gene, (b) capturing interactions among multiple mutations, and (c) integrating diverse targeted panels without requiring uniformity. Fine-tuned on a cohort of 49,592 MSK-IMPACT patients (58,202 tumors), SomAtt achieves state-of-the-art cancer-type classification across 34 cancer types on the MSK-IMPACT panel, with 83.0% accuracy and 70.6% macro-precision, rising to 94.0% accuracy on the 76.5% of tumors it flags as high-confidence. These predictions extend to settings that lack a tissue diagnosis: SomAtt assigns a likely primary site to cancers of unknown primary (CUP; 75% accuracy, 94% on high-confidence calls). Because it integrates arbitrary gene panels without retraining, SomAtt also recovers the disease site directly from blood-based liquid-biopsy panels absent from training (e.g., 62% top 1 and 79% on high-confidence calls on the MSK-ACCESS panel). SomAtt further predicts patient survival, with per-cancer-type concordance indices (C-index) of 0.78 in prostate, 0.72 in endometrial, 0.70 in gastrointestinal stromal tumor, and 0.69 in breast cancer. It also enables patient stratification beyond traditional subtyping: within endometrial cancers of no specific molecular profile (NSMP), a clinically heterogeneous group, SomAtt identifies a previously unrecognized, molecularly defined aggressive subgroup with significantly worse survival — reproducible in an independent cohort (TCGA, log-rank $p < 10^{-6}$) — and nominates candidate biomarkers (broad copy-number burden and serous-like chromosomal instability) for this subgroup. Finally, a downstream survival model built on SomAtt embeddings and trained on MSK-CHORD treatment histories (18,460 patients, ~150 anticancer agents) forecasts treatment-specific progression-free survival across major solid tumors — for example, a C-index of ≈0.80 for abiraterone in prostate cancer and ≈0.78 for oxaliplatin in colorectal cancer. By modeling somatic mutations and their interactions in a panel-agnostic manner, SomAtt provides a robust, scalable framework for tumor-genome analysis, supporting molecular assessment of pre-cancerous lesions, early detection, and precision oncology.

Keywords: Somatic mutations, precision oncology, survival prediction, cancer of unknown primary, foundation model

Title: AI-based Characterization of Alzheimer's Disease Phenotypes from Population Scale Single Cell Data

Author list: Chenfeng He[1,2†], Athan Z. Li[2,3†], Kalpana Hanthanan Arachchilage[1,2†], Chirag Gupta[1,2†], Xiang Huang[2], Xinyu Zhao[2,4], Carissa L. Sirois[2,4], PsychAD Consortium[5], Kiran Girdhar[6,7,8,9,10,11], Georgios Voloudakis[6,7,8,9,10,11], Gabriel E. Hoffman[6,7,8,9,10,11], Jaroslav Bendl[6,7,8,9], John F. Fullard[6,7,8,9], Donghoon Lee[6,7,8,9], Panos Roussos[6,7,8,9,10*], Daifeng Wang[1,2,3*].

Detailed Affiliations:

[1]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, 53706, USA,[2]Waisman Center, University of Wisconsin-Madison, Madison, WI, 53705, USA,[3]Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI, 53706, USA,[4]Department of Neuroscience, University of Wisconsin-Madison, Madison, WI, USA, [5]PsychAD Consortium, [6]Center for Disease Neurogenomics, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA,[7]Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA,[8]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA,[9]Department of Genetics and Genomic Science, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA,[10] Center for Precision Medicine and Translational Therapeutics, James J. Peters VA Medical Center, Bronx, NY, 10468, USA. 11 Mental Illness Research Education and Clinical Center, James J. Peters VA Medical Center, Bronx, NY, 10468, USA. [†]Co-first author, [*]Corresponding author

Abstract: The complexity of Alzheimer's disease (AD) manifests in diverse clinical phenotypes, including cognitive impairment and neuropsychiatric symptoms. However, the etiology of these phenotypes remains elusive. To address this, the PsychAD project generated a population-level single-nucleus RNA-seq dataset comprising over 6 million nuclei from the prefrontal cortex of >1,000 individual brains, covering a variety of disease phenotypes. Leveraging this dataset, we developed a computational framework, called Phenotype Associated Single Cell encoder (PASCode), to score single-cell phenotype associations, and identified ~1.5 million phenotype associated cells (PACs) from 584 donors with AD-related phenotypes. PASCode ensembles multiple statistical methods into a graph neural model for robust scoring. Comparing PACs within 27 brain cell subclasses, we prioritized cell subpopulations and their expressed genes for various AD phenotypes. For instance, we identified microglia subpopulations implicated in AD pathology; reactive astrocyte subtypes with altered neuroprotective and neurotoxic gene expression that likely confer cognitive resilience; and enhanced excitatory/inhibitory imbalance and mitochondrial dysfunction in cognitively impaired AD donors. We also identified many PACs for multiple phenotypes, including the astrocytes between AD and depression showing specific gene expression patterns such as inflammation and Endoplasmic Reticulum stress pathways. These prioritized subpopulations, genes and pathways potentially offer valuable insights for precision diagnostic and therapeutic development. We also validated our findings in external population-scale datasets including AD and major depression disorder, compiled an AD-phenotypic single-cell atlas, and delivered the framework as an open-source tool with pre-trained models and a web application for community use.

Title: FACET: criterion-anchored semantic phenotype matching with a guarded LLM for DSM-5-TR face validity of psychiatric mouse models

Author list: An Le[1], Chunyu Liu[1]

Detailed Affiliations:

[1]Department of Psychiatry, SUNY Upstate Medical University, Syracuse, NY, USA.

Abstract: Mouse genetic models anchor psychiatric neuroscience, but their face validity (whether they reproduce the disorder's clinical features) is still judged by hand. Phenotype-matching methods such as Phenodigm rely on overlap between mouse and human phenotype ontologies, absent for 97% of candidate human-to-mouse pairs examined; the right correspondences are cross-level, linking DSM-5-TR clinical criteria to mouse behavioral-assay phenotypes, which ontology overlap misses but criterion-anchored matching recovers. We developed FACET (Face-validity Assessment by Criterion-anchored Expert-LLM Translation), scoring DSM-5-TR face validity across 76,060 Mouse Genome Informatics (MGI) genotypes. FACET sorts DSM-5-TR criteria into per-disorder symptom domains, shortlists candidate correspondences by embedding (text-embedding-3-large) and lexical similarity, then a large language model (gpt-5.4, fixed seed) adjudicates each under fixed schemas, disorder-neutral prompts, and logged rationales, with verdicts checked against hand-curated anchor sets. The language model only

proposes correspondences; deterministic code aggregates accepted ones into a face-validity score, with significance from a depth-stratified permutation null. We then benchmarked FACET against Phenodigm. We assessed face validity three ways. First, discrimination: FACET ranked a disorder's MGI-curated disease models above other genotypes, beating Phenodigm on behavioral-criterion disorders, autism AUROC 0.94, schizophrenia 0.84, intellectual disability 0.69, versus Phenodigm 0.69, 0.72, 0.59 (paired DeLong; nominal p = 2e-10, 0.002, 0.016; autism and schizophrenia survive Bonferroni, intellectual disability only FDR, q = 0.031). Phenodigm led only for major depression (0.77 vs 0.65); since it scores the whole profile, not behavior alone, restricting it to behavioral phenotypes lowered it to 0.53, 0.66, 0.52, 0.48 (autism/schizophrenia/ID/depression), leaving FACET ahead on three powered disorders (depression underpowered at n = 7). Second, a per-symptom-domain test (Cliff's delta) was significant for autism only, carried by its social domain (delta 0.35). Third, a depth-controlled logistic, separating a disorder's models from other psychiatric disorders', held only for autism and schizophrenia (odds ratios 3.5, 2.6 per SD). This ceiling tracks shared symptoms, not merely method failure: autism models outranked intellectual-disability models on intellectual-disability criteria (reversed ranking, AUROC 0.23). Likewise, FACET found no specific mouse analog for guilt or suicidality, criteria mice cannot model. A genetics-driven expansion (three evidence routes: de novo, GWAS, combined; four disorders) reached curated-model face validity in none; most carried no DSM-5-TR-relevant phenotype (91% in the autism de novo tier). FACET's top autism models concentrated on strong-evidence genes, agreeing with autism's known synaptic-adhesion, mTOR, and chromatin convergence. FACET shows that guarded LLM adjudication adds value where cross-species ontologies are sparse: a scalable, criterion-anchored, auditable face-validity measure.

Keywords: TBD

Flash Talks
August 4th
1:20 – 4:40 PM
Room: 2213B

Chair: Yufang Jin

Title: Nanoparticle-Induced Thermogenic Adipocytes Improve Endothelial Tube Formation and Network

Author list: Parker Olkowski[1], Mariana Chica Uribe[1], Christian Mendiondo[1], María A. Gonzalez Porras[1]

Detailed Affiliations:
[1]Department of Biomedical Engineering and Chemical Engineering, University of Texas at San Antonio, San Antonio, TX.

Abstract: During the proliferation phase of wound healing, the formation of new capillaries to replace damaged blood vessels and restore circulation occurs through angiogenesis. As such, angiogenesis plays a crucial role in supplying the injury site with nutrients, leukocytes, and growth factors necessary for granulation tissue formation and epithelialization. Disruption of angiogenesis during wound healing, hence, is characterized by delayed wound healing and chronic wound formation.

Adipocytes are increasingly recognized as regulators of wound healing through paracrine signaling. Notably, subcutaneous white adipose tissue (sWAT) undergoes browning, the conversion of white adipocytes into beige thermogenic adipocytes, following chronic skin injury. Thermogenic adipocytes are mitochondria-rich, energy-dissipating cells with emerging roles as secretory and metabolic regulators. Yet, how thermogenic adipocyte-derived secreted factors regulate endothelial angiogenic activity remains poorly understood.

The objective of this work is to determine whether thermogenic adipocyte-derived secretomes improve endothelial angiogenic activity. Our group developed a lipid-coated mesoporous silica nanoparticle (LCMSN) platform loaded with forskolin (FSK) to activate thermogenic pathways in sWAT. To investigate how thermogenic adipocyte-derived secretomes affect angiogenesis, we differentiated adipose stromal cells into beige or white adipocytes using established methods and generated nanoparticle-induced beige adipocytes by treating white adipocytes with LCMSN-FSK. Conditioned media from beige adipocytes, white adipocytes, and LCMSN-FSK-induced beige adipocytes were collected. Human umbilical vein endothelial cells (HUVECs) were then treated with beige adipocyte-conditioned media (BeAT-CM), white adipocyte-conditioned media (WAT-CM), or LCMSN-FSK-

induced adipocyte-conditioned media (LCMSN-FSK-CM) in an in vitro tube formation assay. Endothelial network formation was quantified using the Angiogenesis Analyzer plugin for ImageJ by measuring the number of nodes, master segments, meshes, branches, tube length, and tube thickness.

Our results demonstrate that HUVECs treated with LCMSN-FSK-CM significantly increased the number of nodes, master segments, branches, and tube thickness after eight hours, indicating improved endothelial network formation. These findings suggest that thermogenic adipocyte-derived factors promote endothelial angiogenic activity, supporting their potential to improve wound healing. This work supports nanoparticle-induced browning of adipocytes as a promising nanotechnology-enabled strategy to promote pro-regenerative wound healing.

Acknowledgement: This study is partially supported by NSF #2447769.

Keywords: Angiogenesis; Beige Adipocytes, Nanoparticle Drug Delivery

Title: Electroencephalography (EEG) for Remote Vehicle Control Applications

Author list: Seth Johnston[1], Roberto Navarijo[1], Sergio Eduardo[1], Chen Pan[1], and Yu-Fang Jin[1]

Detailed Affiliations:

[1]RISE Labs, Department of Electrical Engineering, The University of Texas at San Antonio, San Antonio, TX, USA

Abstract: Brain–computer interfaces (BCIs) based on electroencephalography (EEG) provide a non-invasive approach for recording and interpreting neural activity, enabling more natural human–machine interaction. However, translating EEG signals into reliable robotic control remains challenging because EEG data are noisy, user-dependent, and difficult to label accurately. Developing a flexible EEG-based control framework can support assistive technologies, medical rehabilitation, remote operation, robotics, and industrial human–machine collaboration by allowing users, including individuals with physical disabilities or limited mobility, to interact with robotic platforms through intuitive movement intentions.

This work presents the development of an EEG-based control platform for classifying user arm-movement intentions and translating the predicted movements into control commands for unmanned aerial vehicles (UAVs), remote-controlled ground vehicles, and other robotic systems. The proposed framework consists of four main components: EEG sensing, labeled data collection, machine-learning-based movement classification, and robotic controller interfacing. EEG signals are acquired using the OpenBCI Cyton board with dry electrodes mounted on a custom 3D-printed headset. The OpenBCI GUI is used for signal visualization, while a Python-based pipeline records EEG signals into CSV files for training and analysis.

To generate labeled training data, a MediaPipe-based tracking system is used to detect user arm movements and synchronize movement labels with the recorded EEG signals. The collected EEG data are then preprocessed and used to train a PyTorch-based machine learning model for classifying user movement intentions. Initial EEG datasets covering multiple movement classes have been collected and organized for model development and evaluation.

After classification, the predicted movement class is converted into a robotic control command. Since the machine-learning output is digital, while many drone and robotic controllers rely on analog input channels such as potentiometer-based control signals, the controller interface was modified to support digital-to-analog conversion. A DAC-based interface translates the classified digital command into compatible analog control signals, allowing the drone or robotic controller to interpret EEG-based decisions as physical control inputs.

As an initial testbed, a Bluetooth-controlled remote vehicle was developed to evaluate the feasibility of EEG-based robotic control. The broader impact of this work is the development of a low-cost and adaptable BCI framework that may benefit assistive robotics, rehabilitation, mobility support for disabled individuals, remote vehicle operation, human–robot interaction, and future autonomous systems by improving accessibility and creating more intuitive interfaces between humans and machines.

This study is partially supported by NSF #2447769 and NIFA 2026-68018-46052.

Keywords: Electroencephalography, Brain-Computer Interface, Machine Learning, Human-Machine Interaction, Robotics, Unmanned Systems

Title: Hardware-Oriented Programming for an Autonomous Robot

Author list: Angela Zhang[1], Dylan Xie[1], Jerry He[1], Max Zhao[1], Charlie Ai[1], Sophia Lu[1], Helen Liu[1], Jason Olkowsky[1]

Detailed Affiliations:

[1]ACE Labs, Department of Electrical Engineering, The University of Texas at San Antonio, San Antonio, TX, USA

Abstract: Robotics competitions provide valuable platforms for K–12 students to apply STEM knowledge and develop teamwork and problem-solving skills through hands-on robot design. Although millions of students participate in robotics competitions, many teams lack structured knowledge of the hardware-oriented programming required for complex robotic systems. This work presents a hardware-oriented programming pipeline developed for a robot designed for the FIRST Tech Challenge (FTC) Decode season. In this game, alliance teams collect colored balls from the field and shoot them into goals, earning higher scores when the shooting sequence matches a specified color pattern.

To accomplish these tasks, the robot integrates four subsystems: drivetrain, intake, transfer, and delivery. The drivetrain uses four goBILDA 435-rpm motors; the intake uses one 1,150-rpm motor; and the delivery subsystem includes two 6,000-rpm motors for shooting and one 435-rpm motor for turret aiming. The transfer mechanism employs two Axon MAX servos. Closed-loop control is implemented throughout the system using 19 sensors. Robot localization is achieved using a Pinpoint Inertial Measurement Unit (IMU) and two odometry sensors, while multiple distance and color sensors detect ball presence and color. Additional sensors support turret initialization, automatic aiming, and jam detection.

Based on this hardware design, a Java software package was developed using finite state machines. The hardware-oriented program coordinates sequential, sensor-based control within subsystems and parallel execution across subsystems. More than 20 finite states were defined for the individual sub-systems, with consideration given to the synchronization among all 4 sub-systems. To improve localization accuracy, a Kalman filter was applied to estimate and correct the difference between the robot's actual position and the position measured by the odometry and IMU sensors.

The proposed hardware-oriented programming approach achieves a control-loop time of 10ms during autonomous operations. The resulting system enables the robot to shoot three balls within 600 milliseconds and complete four scoring cycles during a 30-second autonomous period. The complete autonomous system ranked in top 6% of teams in the 2026 international FTC robotic competition.

This work is presented by FTC team 16458 (Technowizards), in support of UTSA, SWRI, Federation of Korean Associations, CATS, the Computer Shop, and NSF #2447769.

Keywords: Robotics, Unmanned Systems

Title: Bridging Computer Vision and Food Policy: A YOLOv11 Produce Freshness Detector Aligned with California AB 660 Date Label Vocabulary

Author list: Aston Liu[1], Hari Krishna Yalamanchili[2]

Detailed Affiliations:

[1]The Kinkaid School, Houston. [2]Baylor College of Medicine, Houston, TX, USA

Abstract: Date-label confusion is a measurable public-health and food-waste problem. A 2025 national survey by the Harvard Law School Food Law and Policy Clinic, ReFED, and the Johns Hopkins Bloomberg School of Public Health found that 43% of U.S. adults always or usually discard food near its printed date. ReFED estimates such confusion wastes three billion pounds of food, worth $7 billion, annually. Misread labels cut both ways: edible food discarded, unsafe food retained past safety dates. California AB 660 (signed September 28, 2024, effective July 1, 2026) mandates "Best if Used By" for quality and "Use By" for safety. A proposed federal equivalent (H.R.4987/S.2541) is now in subcommittee. To our knowledge, no consumer-facing visual AI tool operationalizes this quality/safety distinction at the consumer's point of decision. Existing YOLO and CNN-based freshness models stop short of regulatory mapping. Mukhiddinov et al. (2022, Sensors) reported 50.4% average precision across 10 produce types labeled fresh/rotten (20 classes) via improved YOLOv4. Fahad et al. (2022, CMC) extended to three ordinal stages (pure-fresh, medium-fresh, rotten), reaching 82-84% accuracy across 11 produce types. Neither map output to regulatory label vocabulary, and public datasets undersample the intermediate senescence zone.

We trained YOLOv11 on a primarily self-collected five-produce dataset (apples, strawberries, blackberries, raspberries, cucumbers) captured via time-lapse over five days under elevated humidity, supplemented by 100

curated endpoint images (Sriram, Kaggle). This design records gradual senescence (enzymatic browning, ethylene-driven softening, surface degradation) including the intermediate gray zone absent from endpoint-sampled benchmarks. Frames were sparsely sampled by day to prevent temporal leakage. The model was trained to 80 epochs. Two labels (fresh, rotten) map onto AB 660's regulatory vocabulary, with fresh aligning to the Best-if-Used-By window and rotten approaching the Use-By threshold. The model achieved mAP50 76.8%, precision 81.0%, and recall 67.6%. Mukhiddinov et al. reported 50.4% average precision on binary detection across 10 produce types, while Fahad et al. reached 82- 84% accuracy on three-stage classification, under different conditions than ours. The 67.6% recall reflects visual ambiguity across intermediate senescence stages, appropriate to an advisory rather than diagnostic instrument. The contribution is application framing: a consumer-education tool that maps detection output onto AB 660's vocabulary (fresh → Best if Used By; rotten → Use By as a precautionary visual signal, not a microbial safety determination), bridging food policy and computer vision toward household-level label literacy where prior CV models have served only industrial quality control.

Keywords: TBD

Title: DUET-Knockoffs: FDR-Controlled Cross-Platform Feature Selection from Paired 16S and Shotgun Microbiome Count Data

Author list: Yiqiao Zhu[1,2], Shiyuan Wang,[1] Qiwei Zhang,[1] Yong Ge[3] and Yuxuan Du[1,*]

Detailed Affiliations:

[1]Department of Electrical Engineering, The University of Texas at San Antonio, One UTSA Circle, 78249, TX, USA, [2]Department of Biostatistics, University of Michigan, Ann Arbor, 500 S State St, 48109, MI, USA, [3]Department of Microbiology, Immunology & Molecular Genetics, University of Texas Health San Antonio, 7703 Floyd Curl Drive, 78229, TX, USA. [*]Corresponding author

Abstract: Motivation: Paired 16S rRNA gene amplicon and whole-genome shotgun metagenomic sequencing is increasingly collected on the same samples, providing complementary but imperfectly aligned views of microbial communities. However, most feature-selection methods are designed for a single sequencing platform. In practice, 16S and shotgun data are often analyzed separately and compared post hoc, which provides no formal error control for identifying taxa with reproducible evidence across platforms. There remains a need for statistical methods that jointly analyze paired microbiome count profiles while accounting for sequencing depth, sparsity, platform-specific bias, and cross-taxon dependence. Results: We propose DUET-Knockoffs, a knockoff-based framework for cross-platform feature selection from paired 16S and shotgun microbiome count data. For each taxon, DUET-Knockoffs targets the union-null hypothesis that the taxon is conditionally independent of the outcome in at least one sequencing platform. Rejection of this null prioritizes taxa with evidence of conditional association in both platforms. Within each platform, taxon counts are modeled using zero-inflated negative binomial marginal distributions with library-size adjustment and observed covariates, coupled through a Gaussian copula to approximate cross-taxon dependence. The fitted feature model is used to generate outcome free synthetic null count profiles, which serve as reference profiles for original-versus-synthetic contrast statistics. Platform-specific contrasts are then combined using a simultaneous-knockoff aggregation statistic, yielding a single selection statistic per taxon. Under the fitted synthetic-null feature model and the symmetry assumptions of the simultaneous knockoff framework, the resulting procedure controls the false discovery rate (FDR) for the cross-platform union-null target. In simulations calibrated to real microbiome count profiles, DUET-Knockoffs maintains empirical FDR control and achieves higher power than representative single-platform and integrative competitors across settings with varying community heterogeneity, overdispersion, structural zeros, platform bias, and covariate confounding. Applications to an oral microbiome study of prediabetes and a non-human primate gut microbiome study of dietary niche prioritize biologically plausible, count-scale cross-platform candidate genera that are not consistently recovered by single-platform analyses. Availability and Implementation: The DUET-Knockoffs software is publicly available at https://github.com/dyxstat/DUET-Knockoffs.

Keywords: Microbiome; 16S rRNA; Shotgun metagenomic sequencing; Feature selection; FDR control

Title: An AI-powered single-cell and spatial omics ecosystem for neurodegenerative disease

Author list:

Weidong Wu[1,2], Tae Yeon Kim[3,16], Jielin Xu[4,16], Hao Cheng[1,16], Cankun Wang[1,2,16], Weiqing Chen[5,6], Qi Guo[1], Xiaojie Jin[1], Jeremy A. Miller[7], Kaixuan Xiao[1], Shaohong Feng[1], Jonathan G. Miller[1], Shaopeng Gu[1,2], Julie Zhu[1], Xiaoying Wang[1,2], Elizabeth Wang[1], Yunxiao Ren[4], Marie-Amandine Bonte[3], Jing Zhao[1], Dana M. McTigue[3,11], Lang Li[1], Benjamin M. Segal[12,13], Guangyu Wang[5,6,8,9,10], Phillip G. Popovich[3,11,14], Feixiong Cheng[4], Hongjun Fu[3,15,17], Anjun Ma[1,2,17], Qin Ma[1,2,17,18]

Detailed Affiliations:

[1]Department of Biomedical Informatics, The Ohio State University, Columbus, OH, USA, [2]Pelotonia Institute for Immuno-Oncology, The James Comprehensive Cancer Center, The Ohio State University, Columbus, OH, USA, [3]Department of Neuroscience, The Ohio State University, Columbus, OH, USA, [4]Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, OH, USA, [5]Center for Bioinformatics and Computational Biology, Houston Methodist Research Institute, Houston, TX, USA, [6]Department of Physiology, Biophysics & Systems Biology, Weill Cornell Graduate School of Medical Science, Cornell University, New York, NY, USA, [7]Allen Institute, Brain Health, Seattle, WA, USA, [8]Center for Cardiovascular Regeneration, Houston Methodist Research Institute, Houston, TX, USA, [9]Center for RNA Therapeutics, Houston Methodist Research Institute, Houston, TX, USA, [10]Department of Cardiothoracic Surgery, Weill Cornell Medicine, Cornell University, New York, NY, USA, [11]Belford Center for Spinal Cord Injury, The Ohio State University, Columbus, OH, USA, [12]Department of Neurology, College of Medicine and Wexner Medical Center, The Ohio State University, Columbus, OH, USA, [13]Neuroscience Research Institute, The Ohio State University, Columbus, OH, USA, [14]Center for Brain and Spinal Cord Repair, The Ohio State University, Columbus, OH, USA, [15]Center for Brain Injury Recovery & Discovery, The Ohio State University, Columbus, OH, USA, [16]These authors contributed equally, [17]Senior authors, [18]Lead contact

Abstract: Neurodegenerative diseases (NDs) share pathological features but differ in cellular vulnerability, anatomical involvement, and progression. Disentangling these shared and divergent programs demands an integrated view at the single-cell level yet remains missing in the ND domain. Here, we present ssKIND, an AI-powered, comprehensive resource integrating 24.12 million annotated brain cells or nuclei from 4,532 single-cell/single-nucleus RNA sequencing samples and 1,380 spatial transcriptomic samples across nine NDs. From that, ssKIND harmonizes human and mouse ND brain atlases for cross-region and -disease analyses. We further introduce NeuroCLIP, a brain-specialized vision-omics foundation model fine-tuned on 0.96 million ST H&E histology images, connecting molecular profiles with brain tissue morphology in NDs. Using ssKIND, we uncover shared and disease-specific transcriptional programs across NDs, specifically, the HSP90-centered proteostasis response that tracks Alzheimer's disease progression. Furthermore, we prioritize microglial disease signatures to identify small-molecule drug candidates capable of reversing microglial stress states across NDs. ssKIND delivers an AI-powered, large-scale single-cell and ST ecosystem for understanding molecular mechanisms in NDs. ssKIND are publicly available at https://bmblx.bmi.osumc.edu/sskind/.

Keywords: Neurodegenerative diseases, Single-cell transcriptomics, Spatial transcriptomics, Brain cell atlas, Multimodal foundation model, Cross-disease analysis, Drug repurposing, AI platform

Title: BindFuse: A Multimodal Graph Learning Framework for Protein–Ligand Binding Affinity Prediction

Author list: Bilal Ahmad[1], Ram Samudrala[1] and Zackary Falls[1,2]

Detailed Affiliations:

[1]Department of Biomedical informatics, Jacobs School of Medicine and Biomedical Sciences, State University of New York at Buffalo, NY, USA, [2]Department of Biostatistics and Bioinformatics, Roswell Park Comprehensive Cancer Center, Buffalo, NY, USA.

Abstract: Accurate prediction of protein–ligand binding affinity is a central problem in computational drug discovery due to the complex and multiscale nature of molecular recognition. Despite recent progress in protein language models and molecular representation learning, most existing approaches rely on a single modality, such as sequence, structure, or ligand topology, limiting their ability to fully capture interaction complexity and generalize across diverse protein–ligand systems. To address this limitation, we propose BindFuse, a multimodal graph learning framework that integrates evolutionary sequence information, structural representations, and ligand chemical features into a unified protein–ligand interaction graph. BindFuse leverages six complementary pretrained encoders and feature generators, including ESM-3, ProtBERT-BFD, AlphaFold3-inspired structural descriptors, structural diffusion representations, ChemBERTa, and RDKit-derived ligand features, producing 2,769-

dimensional node embeddings. Protein–ligand interaction graphs are constructed using a 5.0 Å spatial cutoff, encoding both covalent and non-covalent interactions. We evaluate four graph neural network architectures, including AffinityFormer, our proposed Graphormer3D-based architecture incorporating edge-augmented attention bias, dynamic structural feature augmentation, stochastic depth, and masked mean pooling, alongside SchNet, GemNet-T, and GATv2Conv baselines. Experiments on the PDBbind v2020 benchmark using random, scaffold-based, and cold-target splits show that BindFuse consistently improves predictive performance relative to single-modality representations, achieving Pearson correlation coefficients of 0.724, 0.691, and 0.695, with corresponding RMSE values of 1.031, 1.037, and 1.060 pK units, respectively. In addition, large-scale experiments on the BindingDB dataset demonstrate scalability and robustness, achieving a Pearson correlation coefficient of 0.807. These results suggest that integrating complementary sequence, structural, and chemical modalities provides a more effective representation for protein–ligand affinity prediction and highlights the potential of multimodal graph learning for structure-based drug discovery.

Keywords: Protein–Ligand Binding Affinity, Multimodal Learning, Graph Neural Networks, Structure-Based Drug Discovery, Protein Language Models, Geometric Deep Learning

Title: Integrative Mendelian Randomization and Protein–Protein Interaction Network Analysis Links Multi-Organ Aging to Complex Diseases

Author list: Chunxuan Ma[1,2], Yuan Hou[1], Feixiong Cheng[1,3,4,5]

Detailed Affiliations:
[1]Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, Ohio, USA, [2]Department of Population and Quantitative Health Sciences, School of Medicine, Case Western Reserve University, Cleveland, Ohio, USA, [3]Department of Molecular Medicine, Cleveland Clinic Lerner College of Medicine, Case Western Reserve University, Cleveland, Ohio, USA, [4]Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine, Cleveland, Ohio, USA, [5]Cleveland Clinic Genome Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, Ohio, USA

Abstract:

Background

Biological aging is heterogeneous across organ systems, with distinct but functionally interconnected genetic architectures. Genome-wide association studies (GWAS) have independently characterized organ-specific aging loci and disease risk variants. However, the molecular interactomes through which aging signals across multiple organs converge to drive complex disease, including Alzheimer's disease (AD), remain poorly understood. Here, we applied Mendelian randomization (MR) integrating organ-specific biological age gap (BAG) GWAS with organ-specific expression quantitative trait loci (eQTL) data to identify genes linking aging organ systems to AD. We conducted a multi-organ protein-protein interaction (PPI) network to identify convergent genes and pathways implicated in aging-linked AD mechanisms.

Methods

We applied two-sample MR using organ-specific eQTL as instrumental variables to analyze organ-specific BAG GWAS across nine systems and case-control AD GWAS. eQTL data were obtained primarily from the Genotype-Tissue Expression (GTEx v10) database, with immune and metabolic eQTL derived from the Database of Immune Cell Expression, eQTLs and Epigenomics (DICE) and published datasets. MR analyses were performed separately for each organ aging system and AD, identifying candidate genes. These organ aging and AD-associated genes were integrated into a PPI network to identify convergent genes between aging functional modules and the AD disease modules. Candidate genes were validated using single-cell RNA sequencing (scRNA-seq) and pathway enrichment analyses.

Results

MR analyses identified between 9 and 246 aging-associated genes per organ system and 113 genes associated to AD risk. MR analysis identified PLEKHA1 as an immune aging- and AD-associated gene. Network analysis further identified PLEKHA1 as a convergent hub connecting brain and lung aging (NINL), metabolic aging (WDR45B), brain aging (PTPN13), and AD-associated (APP) molecular nodes. The scRNA-seq analysis showed significant (Benjamini-Hochberg FDR < 0.05) upregulation of PLEKHA1 in neurons and astrocytes in AD compared with control counterpart. Network-connected genes NINL, WDR45B, and PTPN13 also showed significant differential expression between AD and control populations in neuronal and glial cell types. Pathway enrichment identified

phosphatidylinositol phosphate (PIP) synthesis, phosphatidylinositol (PI), phospholipid metabolism, and RND GTPase cycling as top processes. These pathways suggest that convergent multi-organ aging networks centered on PLEKHA1 may contribute to lipid membrane dysregulation and inflammatory processes relevant to AD.

Conclusion

This study identifies PLEKHA1 as a convergent hub centered multiple organ aging signals with AD risk. These findings provide evidence for the contribution of biological aging across multiple organ systems to AD risk and show the value of multi-organ network integration for investigating how aging drives the onset of complex diseases.

Keywords: Protein-Protein Interaction Network, Aging, Multi-organ System, Alzheimer's Disease

Title: STITCH: A Platform-Agnostic Pipeline for Barcoding-based Spatial Transcriptomics

Author list: Yuanhang Liu[1], Brenna Novotny[1], Humza Ashraf[1], Mariam Stein[1], Jeong-Heon Lee[2], Ting Zhang[2], Alejandro Stark[2], Sathiya Pandi Narayanan[2], Corey Seavey[2], Arthur Beyder[2], Tamas Ordog[2], Chen Wang[1]

Detailed Affiliations:

[1]Department of Quantitative Health Sciences, Mayo Clinic, MN, USA, [2]Department of Physiology and Biomedical Engineering, Mayo Clinic, MN, USA

Abstract: Barcoding-based spatial transcriptomics (ST) technologies have gained rapid adoption in recent years to investigate biological questions in spatial context. Major platforms include Visium and Visium HD (10x Genomics), Seeker and Trekker (Takara Bio), and Stereo-seq (Complete Genomics), providing options at varying spatial resolutions and sensitivities. Those platforms share a common output structure consisting of spatially indexed gene expression measurements associated with spatial units, e.g., spots, bins, beads, or cells. Despite this shared structure, no unified analytical workflow currently supports processing all those platforms.

To address this gap, we developed Spatial and single-cell Transcriptomics Integration Tool for CHaracterization (STITCH), a Nextflow-based pipeline for platform-agnostic analysis of barcoding-based ST datasets. STITCH provides an end-to-end workflow including quality control, normalization, spatially variable gene (SVG) identification, cell type annotation via mapping or deconvolution, spatial embedding, data integration, clustering, differential expression, pathway analysis, cell–cell interaction and colocalization analysis. To enable efficient analysis of large-scale datasets, STITCH incorporates on-disk computation through BPCells. Final outputs are generated as both Seurat and AnnData objects to support downstream analysis in R and Python ecosystems.

We additionally conducted benchmarking studies to evaluate combinations of normalization, SVG selection, and clustering methods for spatial domain identification using pathology-annotated datasets, including a public human dorsolateral prefrontal cortex (DLPFC) Visium dataset and in-house human colon Visium data. We demonstrated that incorporating spatial embedding, particularly Banksy, substantially improves spatial domain identification, especially in complex tissues such as brain where pathology-defined regions contain overlapping and heterogeneous cell populations. We further performed sparse data simulations using in-house generated mouse brain Trekker datasets with cell type annotations derived from the Allen Mouse Brain Atlas. Under high sparsity, clustering performance decreased across all method combinations, with TF-IDF showing marginally better robustness. For highly sparse data, cell type deconvolution using refined, cell type–specific marker genes provided improved cell typing performance. Overall, STITCH provides a unified and scalable framework for streamlined analysis of barcoding-based spatial transcriptomics data.

Keywords: spatial transcriptomics, data integration, spatial domain identification, data sparsity, single cell analysis

Title: Strategies for robust, accurate, and generalizable benchmarking of drug discovery platforms

Author list: Melissa Van Norden[1], William Mangione[1], Zackary Falls[1], and Ram Samudrala[1]

Detailed Affiliations:

[1]Department of Biomedical Informatics, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, Buffalo, NY, USA

Abstract: Consistent and accurate benchmarking is essential to the creation, improvement, and comparison of computational drug discovery technologies. We therefore revised the assessment protocols of our multiscale therapeutic discovery platform, the Computational Analysis of Novel Drug Opportunities (CANDO) platform, in order to bring them into alignment with best practices. These new assessment tools allowed us to optimize multiple parameters used in drug candidate prediction. We found that CANDO could predict 7.4% of drugs in the top 10 compounds for their respective diseases/indications with these optimized parameters when using the Comparative Toxicogenomics Database (CTD) as a gold standard drug-indication mapping; this rose to 12.1% when using only

approved drug-indication associations extracted from the Therapeutic Targets Database (TTD). This improvement in the performance of CANDO when using the TTD was maintained when examining only matched drug-indication associations between the two mappings. Finally, we found that performance on an indication correlated with features of that indication: there was a weak correlation between performance and the number of drugs associated with the indication (Spearman > 0.3), and a moderate correlation with intra-indication chemical similarity (Spearman > 0.5). Performance as assessed by our original and new benchmarking protocols was also moderately correlated (Spearman > 0.5). This study shows the utility of robust benchmarking protocols not only for accurate performance assessment, but for the improvement of drug discovery pipelines and comprehension of the determinants of their performance.

Keywords: Drug discovery, repurposing, assessment, benchmarking, similarity

Title: CLARA: A Containerized Lightweight Automated Reproducible Analysis for 16S rRNA Amplicon Sequencing

Author list: Andres Benavides Arevalo[1]

Detailed Affiliations:

[1]Escuela Ingeniería de Sistemas e Informática, Universidad Industrial de Santander, Bucaramanga, Colombia

Abstract: Amplicon sequencing of the 16S rRNA gene has become a widely adopted approach for characterizing bacterial communities due to its cost-effectiveness, scalability, and reliability. 16S amplicon sequencing analysis involves the selection, installation, configuration, and execution of multiple bioinformatics tools. In addition, there are different sequencing technologies capable of generating either short or long reads. These factors create significant challenges for researchers without advanced computational expertise in preparing suitable analysis tools for their amplicon experiments.

This study presents CLARA: Containerized Lightweight Automated Reproducible Analysis for 16S rRNA Amplicon Sequencing — a framework designed to automate amplicon data processing and simplify analysis. CLARA utilizes container abstraction to package popular amplicon analysis tools as images, which are then automatically executed by a workflow management system. Users interact with the system via configuration files, specifying the sequencing technology, the target dataset, the reference database, and the amplicon pipeline. The results are displayed as interactive tables and charts within a web interface.

CLARA's robustness and versatility were demonstrated by processing datasets from Illumina, PacBio, and Oxford Nanopore platforms using popular tools such as QIIME 2, EMU, and LotuS2, among others. Across all tested datasets, CLARA demonstrated the ability to automatically deploy and manage multiple tools while significantly reducing hands-on time for installation and execution procedures. Overall, CLARA lowers the computational barrier for amplicon sequencing analysis, making reproducible microbiome workflows accessible to researchers across technical backgrounds. Nevertheless, containerization may introduce performance overhead and reduce observability, making it difficult to diagnose pipeline-specific failures at runtime. Furthermore, analytical capabilities and taxonomic resolution are inherently constrained by the completeness and scope of the reference databases employed.

Keywords: 16S rRNA gene, Amplicon sequencing, Automated analysis, Containerization, Workflow management

Title: MetaAge: a Highly Accurate Ensemble Machine Learning Model for Biological Age Estimation

Author list: Liguo Wang[1], and Anping Chen[2]

Detailed Affiliations:

[1]Division of Computational Biology, Mayo Clinic College of Medicine, 200 1st Street SW Rochester, MN 55905，

[2]An-Ping Chen, Boston Harbor Biotechnologies, 17 Briden Street, Worcester MA, 01605

Abstract:

Biological age is a measure of an individual's physiological state that reflects the cumulative effects of genetics, lifestyle, environmental exposures, and diseases, which cannot be fully captured by chronological age. DNA methylation (DNAm)-based epigenetic clocks have emerged as the most widely used and accurate biomarkers of biological aging, often outperforming transcriptomic, proteomic, metabolomic, and clinical biomarker-based approaches.

Over the past decade, numerous DNAm clocks have been developed, including first-generation chronological age clocks (e.g., Horvath and ZhangEN), second-generation healthspan and mortality clocks (e.g., GrimAge and DunedinPACE), and more recent mechanistic aging clocks (e.g., YingCausAge). While these clocks have proven valuable for aging research, each was developed using specific training populations and optimization objectives, limiting their generalizability across diverse cohorts and applications. For example, ZhangEN achieves excellent accuracy in adults over 20 years old but performs poorly in adolescents; Horvath and Hannum primarily capture chronological age and are less predictive of morbidity and mortality; GrimAge is strongly associated with lifespan and disease risk but tends to exhibit systematic calibration biases across populations; DunedinPACE measures the pace of aging rather than biological age itself, making direct comparisons with age-prediction clocks challenging; and mechanistic clocks such as YingAdaptAge and YingDamAge capture specific biological processes but often show reduced accuracy for chronological age estimation.

These limitations suggest that no single clock fully captures the multidimensional nature of human aging, motivating the development of ensemble approaches that integrate complementary aging signals. Here, we developed MetaAge, an ensemble biological age estimator that integrates 17 DNAm clocks using gradient-boosted decision trees (XGBoost). By combining complementary aspects of the aging process captured by existing clocks, MetaAge achieved significantly higher accuracy and robustness than any individual clock across cross-validation analyses and multiple independent external validation cohorts. These findings demonstrate the power of ensemble learning frameworks to overcome the limitation of individual models and advance the precise quantification of human aging.

Keywords: Biological age, DNA methylation, Epigenetic clock, Aging, Ensemble learning

Title: Building a Psychoplastogenic Signature: A Comparative Evaluation of Compound-Based and Target-Based Approaches in the CANDO Platform

Author list: Justin Shinkle[1], William Mangione[1], Zackary Falls[1], and Ram Samudrala[1]

Detailed Affiliations:

[1]Department of Biomedical Informatics, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, Buffalo, NY, USA

Abstract:

Psychoplastogens are a pharmacologically diverse class of compounds that rapidly loosen static-state constraints on cortical neurons, inducing lasting structural and functional plasticity with therapeutic potential in depression, PTSD, and substance use disorders. Their mechanistic diversity suggests that psychoplastogenicity reflects a distributed, systems-level pharmacologic property rather than acute activity at any single acute protein target, motivating the construction of a proteome-wide psychoplastogenic signature to guide compound repurposing and de novo design.

Here we evaluate two complementary strategies for constructing such a signature within the CANDO platform: a compound-based approach and a target-based approach, comparing target-based methods for their ability to recapture known psychoplastogens relative to a CNS-active non-plastogenic control set. The compound-based approach, in which candidates are evaluated by similarity of their predicted proteome-wide interaction profiles to those of established psychoplastogens, recovered psychoplastogens at above-baseline rates across full proteome and synaptogenesis-relevant protein subset analyses, with psychoplastogens recovering each other at average recall rates ($k = 10, 100, 500, 1000, 5000$) of 0.0117, 0.045, 0.125, 0.241, and 0.69, compared to CNS control rates of 0.004, 0.021, 0.073, 0.126, and 0.498.

We then evaluated target-based methods across 95 curated protein libraries derived from SynGO and Reactome pathway annotations. Our de novo compound predictor, which counts instances and sums interaction scores above a defined threshold for a group of proteins, failed to detect psychoplastogen enrichment when filtering to the 90th percentile or greater of interaction scores, with the best Benjamini-Hochberg FDR-adjusted p-value at 0.627. A second function ranking compounds by interaction score across proteins and averaging placements into a master rank list similarly failed to place psychoplastogens above CNS controls for any protein library. However, BANDOCK+, a newer scoring function with improved interaction prediction accuracy, detected statistically significant differences for certain protein libraries. Notably, Reactome protein library R-HSA-375280, containing amine-binding GPCRs broadly targeted by psychoplastogens, placed psychoplastogens at an average 9.7th percentile versus 29.4th percentile for CNS controls (MWU, $p<0.00001$).

These findings indicate that target-based signature construction is sensitive to both method selection and protein library composition. The variable performance of BANDOCK+ warrants further investigation into which protein configurations best concentrate the psychoplastogenic pharmacologic signal. Identifying the optimal protein configuration would further enable evaluation of hybrid approaches that integrate compound-based and target-based signatures, which may yield greater predictive power for psychoplastogenicity than either strategy alone.

Keywords: Psychoplastogens, drug discovery, polypharmacy, repurposing

Title: Benchmarking variation detection methods with linear reference genomes and pangenomes in grapevines

Author list: Robert Robinson[1,*], Brooke Atamanyk[1,*], Ping Liang[1,2]

Detailed Affiliations:

[1]Department of Biological Sciences, [2] Centre for Biotechnology Brock University, St. Catharines, Ontario, Canada

Abstract:

Analysis of genomic variants through whole-genome sequencing (WGS) has become a standard approach for the genetic study of organisms, owing to the rapid development of next-generation DNA sequencing (NGS) technologies over the past few decades.

The rapid accumulation of WGS data has enabled the survey of genetic diversity across and within species at an unprecedented pace. In the meantime, technical challenges in accurately documenting genomic variants arise due to short read lengths, high sequencing error rates, and the extremely complex nature of genomes. Addressing these challenges necessitates continued development and optimization of data, analytical algorithms and tools, as well as the generation of high-quality datasets for training and assessing algorithm efficiency, often in an organism-specific fashion.

In this study, we focus on the performance comparison of variant identification using linear reference-genome and pangenome approaches and their optimization for genomic diversity analysis of grapevines, an important crop worldwide, for which there is a lack of training datasets and standards. We determined the minimum sequencing coverage to achieve false-negative rates below 5% to be 25, 10, and 15 times the genome size, respectively, for Illumina paired-end reads (150 bp x 2), PacBio HiFi reads, and Oxford Nanopore reads, indicating read length and sequencing accuracy as the determining factors. We also determined the optimal filtering parameters for the linear reference genome using BCFtools and for the pangenome approach using vg giraffe to achieve low false-positive and false-negative rates. Additionally, two unexpected observations were made: the superior performance of PacBio data aligned to a linear reference genome compared to a pangenome, and a deterioration in ONT data's ability to detect INDELs at coverage greater than 15X.

In conclusion, our results indicate that while PacBio HiFi, as a long-read platform with high accuracy, is the ideal WGS platform for variant analysis, the Illumina platform remains a competitive option due to its best cost-efficiency and sequencing accuracy.

Flash Talks
August 4th
4:40 – 5:20 PM
Room: 2213B

Chair: Shengyu Li

Title: Improved drug discovery through accurate prediction of adverse drug reactions within the CANDO platform

Author list: Vida Bodaghi-Namileh[1], Lucille Tomin[1],William Mangione, PhD[1], Ram Samudrala, PhD[1], Zackary Falls, PhD[1,2]

Detailed Affiliations:

[1]Department of Biomedical Informatics, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, Buffalo, NY, [2]Department of Biostatistics and Bioinformatics, Roswell Park Comprehensive Cancer Center, Buffalo, NY.

Abstract: Adverse drug reactions (ADRs) remain difficult to predict because drug effects are rarely driven by a single target. Compounds can perturb proteins beyond intended mechanisms, making it unclear which protein-level interactions, interaction combinations, or shared pathways contribute to specific side effects. These distributed patterns may explain why different compounds produce overlapping ADRs, but they are not fully captured by approaches based primarily on known targets, chemical similarity, or adverse event report frequency. We evaluated whether the Computational Analysis of Novel Drug Opportunities (CANDO) platform can use predicted proteome-wide interaction patterns to prioritize known compound–ADR associations from curated and postmarketing sources. CANDO represents each compound using a compound–proteome interaction signature generated through structure-based interaction scoring across a protein structure library. These signatures enable compounds to be compared by predicted interaction behavior across the proteome. DrugBank and OFFSIDES served as complementary reference sources, with DrugBank supporting internal benchmarking of curated compound–ADR associations and OFFSIDES providing postmarketing associations derived from adverse event reporting data. We constructed an OFFSIDES-unique external truth set by removing compound–ADR pairs already present in DrugBank. Using similarity relationships between compound–proteome interaction signatures, we evaluated two complementary association-recovery tasks in internal and external settings. In compound-to-ADR evaluation, ADRs were ranked for each queried compound to assess whether known side effects were prioritized. In ADR-to-compound evaluation, compounds were ranked for each queried ADR to assess whether known associated compounds were prioritized. Performance was assessed across multiple output rank cutoffs and compared with matched random controls. After preprocessing and identifier mapping, the cleaned DrugBank reference contained 99,242 deduplicated compound-ADR associations. OFFSIDES filtering produced 103,332 usable unique compound–ADR pairs; removing 3,103 DrugBank-overlapping pairs yielded 100,229 OFFSIDES-unique associations spanning 1,685 compounds and 10,306 ADRs. Across benchmarks, CANDO recovered associations above matched random controls. At top-100, external compound-to-ADR recovery retrieved 4.69% of true associations versus 2.12% for controls, while external ADR-to-compound recovery retrieved 6.16% versus 4.08%. Internal DrugBank benchmarks also exceeded controls, with reverse recovery reaching 23.25% versus 6.54% and forward recovery reaching 15.28% versus 2.12%. These corresponded to 2.2-, 1.5-, 3.6-, and 7.2-fold enrichments over control, respectively. These results suggest that proteome-wide CANDO interaction signatures capture useful signal for ADR recovery. By representing compounds through predicted interaction behavior across the proteome, CANDO may complement existing pharmacovigilance approaches and prioritize biologically plausible compound–ADR associations for further review.

Title: Prevalence and Genetic Determinants of Chronic Diseases in a University Student Population: A Cross-Sectional Analysis with International Comparisons

Author list: Afagh Bapirzadeh[1,2], Kiana Mosayebi[3],Arezoo Nikpay[3], Nasim Eghbali[3], Mohammad Mahdi Pashaei[3], Ali Heidari[3], Dr. Mitra Salehi, Dr. Zarrin Minoucher[2], Dr. Bijan Bambai[2]

Detailed Affiliations:

[1]Genetic group, Islamic Azad University, North Tehran Branch, P.O. Box: 1651153311, [2]National Institute of Genetic Engineering and Biotechnology, Tehran–Karaj Highway, Science and Technology Park, Research Boulevard, P.O. Box: 1497716316, [3]IslamicAzad University, Shahre Qods Branch, Kalaher Boulevard, Fath Expressway, Shahre Qods, Tehran, Iran, P.O. Box: 37515374.

Abstract: Chronic diseases such as diabetes, hypertension, thyroid disorders, and cardiovascular diseases represent a major global health burden, increasingly affecting younger populations and shaping long-term morbidity and mortality patterns. University students, as a key segment of young adults, are at a critical stage where lifestyle habits are formed and early manifestations of chronic conditions may emerge. Therefore, this cross-sectional study investigated the prevalence of these conditions among 215 university students (both male and female). The participants had diverse familial and ethnic backgrounds but were largely within a comparable age range. The study also examined potential gender differences. Data were collected using a structured questionnaire that documented

the presence or absence of relevant health conditions in the students themselves and among their close relatives. In addition, the study considered genetic factors reported in the literature that may contribute to susceptibility to these diseases. Descriptive statistical methods were applied to calculate frequencies, percentages, and ranges for each condition at both the individual and familial levels. The resulting findings provide an initial health profile that can help identify groups at elevated risk and guide preventive actions. This may include targeted education on healthy lifestyle, early screening initiatives, and interventions addressing modifiable factors such as diet, physical activity, stress, and smoking. Moreover, incorporating family history strengthens interpretation of both genetic predisposition and shared environmental influences that can shape disease risk in young adults. Beyond describing the distribution of conditions in the studied population, this research places local estimates within a broader epidemiological context. By comparing observed frequencies with internationally reported statistics, the study evaluates whether patterns among these students align with, exceed, or fall below global norms. Furthermore, based on our view that the "normal range" of these diseases should be reassessed periodically at regional and national levels—especially in a young country like Iran—such comparisons can support evidence-informed public health planning. Ultimately, the findings underline the importance of tailoring prevention and health-promotion policies to the needs of university-aged populations and can inform future analytical and longitudinal studies.

Keywords: chronic disease, cross-sectional study, university students, family history, genetic factors, health prevention

Title: AI in Fast Detection of Diseases

Author list: Afagh Bapirzadeh[1,2], Yasaman Rostami Safarloo[3], Samaneh Rostami Safarloo[3], Shadi Shahsavari[3], Mohaddeseh Sabzikar Ghane[3], Hasti Mostafavi Ghamsari[3], Mitra Salehi[2], Bijan Bambai[1] and Zarrin Minuchehr[1]

Detailed Affiliations:

[1]National Institute of Genetic Engineering and Biotechnology, Pajohesh Boulevard, Science and Technology City, Tehran–Karaj Highway, Tehran, Iran. Postal Code 1497716316, [2]Department of Genetics, Islamic Azad University, North Tehran Branch, Tehran, Iran. Postal Code 1651153511, [3]Islamic Azad University, Shahre-Qods Branch, Kalhor Boulevard, Fatah Highway, Shahr Qods, Tehran, Iran. Postal Code 37515374.Iran.

Abstract: The ability in early diagnosis of diseases can lead to preventing world epidemiological prevalence of an specific type of illness.it also helps to investigate repetitive patterns which can accelerate faster recognition and more effectiveness toward understanding the right moment and first place of occurrence of a disease.AI also works on pre-existing comparative analysis methods not only to recognize a situation but also to diminish the possibility of another world disaster such as COVID-19.Increase in urban life style along with its effects on environment and deforestation, biological hazards, unprecedented global instability can become relying factors of emerging new and unknown diseases. Health and medical centers have not been immune to human error, both at the level of identifying cases and at the level of assessing the causes and factors of the disease, clinical symptoms, and effective treatment methods. In addition, lack of data, lack of human resources, uneven resources and facilities in different countries, lack of appropriate infrastructure, and political and economic crises in some countries have caused mass medical measures not to be taken in a timely manner, leading to the spread of the disease. In this article, we examine the shortcomings and shortcomings of existing systems and then examine effective methods for collecting information, sharing information among specialized medical centers, and finding useful treatment solutions to prevent global epidemics.

Keywords: Artificial Intelligence (AI), Diagnosis, Treatment Centers, Information Sharing Systems

POSTER SESSION

Posters
August 3rd
6:00 – 7:30 PM
Room: Atrium

Poster P01

Title: Network Rewiring Reveals Coordinated Metabolic, Structural, and Immune-Proteostasis Programs in Heart Failure

Author list: Addison Liu[1], Sandali Dewni Lokuge[2], Sheng Liu[3,4], Jun Wan[2,3,4]

Detailed Affiliations:

[1]Park Tudor School, Indianapolis, IN, USA, [2]Department of Biomedical Engineering and Informatics, Luddy School of Informatics, Computing and Engineering, Indiana University at Indianapolis, IN 46202, USA, [3]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN USA, [4]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN USA.

Abstract: Heart failure with reduced ejection fraction (HFrEF) is a complex disease involving coordinated dysregulation across multiple biological systems. However, conventional differential expression analyses primarily identify static marker genes and fail to capture dynamic network reorganization underlying disease progression. Here, we applied DiCE-Ex, a recently developed network-based framework integrating gene expression with protein interaction topology, to uncover condition-specific rewiring events in HFrEF. Using a published HFrEF dataset, we identified 800 DiCE-prioritized genes by combining statistical significance (FDR < 0.05), network-based weighting (mean WIG), and condition- specific protein-protein interaction (PPI) networks constructed via Pearson correlation. Compared to 1,630 differentially expressed genes (DEGs) with traditional cutoffs, DiCE-specific genes revealed distinct biological programs not captured by expression-based approaches. Upregulated DiCE genes were enriched in immune and inflammatory pathways, including NF-κB, IL-17, and T cell receptor signaling, whereas downregulated genes were enriched in cardiac contraction, oxidative phosphorylation, fatty acid metabolism, and longevity pathways. Network partitioning using the Leiden algorithm identified six modules in both normal and HFrEF states, with substantial rewiring observed in disease. Notably, a normal module (NM5) split into two HFrEF-specific modules: DM3 which is enriched for immune-proteostasis processes, including T cell activation, cytokine signaling, ubiquitin-proteasome system, autophagy, etc., and DM6 enriched for cardiac structural remodeling, e.g., sarcomere organization, cytoskeleton, cell-cell junctions, extracellular matrix. Additionally, two metabolic modules (NM3 and NM4) in normal control, representing oxidative phosphorylation and fatty acid oxidation (FAO), respectively, merged into a unified DM4 module in HFrEF. This integration reflects increased coupling between lipid utilization and mitochondrial ATP production, suggesting reduced metabolic flexibility under bioenergetic stress. Collectively, these results support a three-layer systems model of heart failure consisting of metabolic dysfunction (DM4), cardiomyocyte structural remodeling (DM6), and immune-proteostasis responses (DM3). By identifying transition drivers rather than static markers, DiCE reveals coordinated network-level changes underlying disease progression, providing new insights into the systems biology of HFrEF.

Keywords: DiCE-Ex, HFrEF, network rewiring, module

Poster P02

Title: Isoform-resolved predicted dsRNA burden is concentrated in enterocyte-like malignant colorectal cancer cells identified by long-read single-cell RNA sequencing

Author list: Alex Fang[1], Yunyun Zhou[2]

Detailed Affiliations:

[1]Independent Researcher, Irvine, USA, [2]Fox Chase Cancer Center, Temple University, Philadelphia, USA

Abstract: Long-read single-cell RNA sequencing captures full-length isoform architecture, unlike short-read RNA-seq, which relies on fragmented transcript segments and can obscure which exons, UTRs, retained sequences, and repetitive elements are linked within the same RNA molecule. Accurate endogenous double-stranded RNA (dsRNA) prediction requires transcript resolution because duplex formation depends on intramolecular complementarity, inverted repeats, and Alu organization. Thus, long-read technology can nominate dsRNA-prone isoforms and map them to malignant epithelial subpopulations.

Here, we developed an isoform-resolved, three-component scoring framework integrating intra-isoform complementarity (S1), Alu inverted-repeat architecture (S2), and an isoform-level reliability filter (S3) accounting for long-read support, expression robustness, and patient recurrence. Component weights were calibrated against J2 anti-dsRNA immunoprecipitation sequencing from a colorectal cancer (CRC)-derived cell-line context. Applied to 26,091 curated isoforms from a 12-patient CRC long-read single-cell cohort, a pre-specified priority threshold identified 76 candidate isoforms with elevated predicted dsRNA-forming potential.

Predicted dsRNA burden was concentrated in CEACAM5$^+$/PHGDH$^+$ enterocyte-like iCMS3-like malignant CRC cells, consistent with an MSS-associated, CMS3-enriched program, compared with ASCL2$^+$/AXIN2$^+$ stem-like iCMS2-like cells, consistent with an MSS-associated, CMS2-enriched program. iCMS3-like cells showed higher median burden than iCMS2-like cells across 12 patients (14.5 vs 1.0; 12/12 concordant; one-sided binomial $p = 0.0002$). This pattern was not recovered by gene-level expression ranking, indicating that isoform architecture is required. Gene-level projection into an independent 62-patient 10x Chromium CRC atlas supported directional concordance across patients (41/62; $p = 0.0076$), providing expression-level support but not isoform-level validation. Three orthogonal biochemical analyses provided mechanistic support without establishing a causal pathway. J2 anti-dsRNA pulldown enriched 57 of 64 candidate loci above background ($p = 5.1$ x 10-3), with exon-restricted candidate Alu signal 194-fold above genomic background. MDA5-protected A-to-I editing overlapped candidate loci above chance (permutation $p < 0.001$), and ADAR1 knockdown induced duplex formation at TYRO3 ($p < 0.001$), supporting ADAR1-dependent suppression. Together, these results support a model in which isoform-specific exonic architecture contributes to endogenous dsRNA potential, with candidate loci showing evidence consistent with dsRNA formation, MDA5-associated recognition, and ADAR1-mediated constraint.

This study establishes a transferable framework for identifying isoform-resolved predicted endogenous dsRNA programs from long-read single-cell tumor data. By substituting tumor-specific cell state annotations, the workflow can be applied to other long-read single-cell tumor datasets. In CRC, CEACAM5$^+$/PHGDH$^+$ iCMS3-like malignant epithelial cells showed the strongest enrichment for elevated predicted dsRNA burden, nominating this cell state for future validation of dsRNA-linked innate immune biology.

Keywords: colorectal cancer, long-read scRNA-seq, transcript isoforms, endogenous dsRNA, innate immune surveillance, iCMS

Poster P03

Title: PGCC Explorer: An Interactive Web Platform for Integrative Analysis of Drug Responses in Polyploid Giant Cancer Cells

Author list: Li-Ju Wang[1], Hsiao-Chun Chen[1], Chien-Hung Shih[1], Yuan Zhang[1], Huikang Ye[1], Ying-Ju Lai[1], Yushu Ma[1], Tiffany Habib[1], Hsi-Chun Wang[1], Yu-Chih Chen[1], Yu-Chiao Chiu[1]

Detailed Affiliations:

[1]UPMC Hillman Cancer Center, University of Pittsburgh, Pittsburgh, PA 15232.

Abstract: Polyploid giant cancer cells (PGCCs), typically arising from whole-genome duplication, are major drivers of therapeutic resistance and tumor recurrence. Building on our recently published high-throughput single-cell drug screening platform and ongoing efforts to profile PGCC responses across multiple cell lines representing diverse cancer lineages, we developed an interactive web-based platform that enables integrative analysis and visualization of these data. The platform integrates high-throughput PGCC drug screening results with multi-omic features curated from public cancer data resources, including somatic mutations, copy number alterations, gene expression profiles, and pathway activity scores. It supports two primary analysis modules: (1) a geneor pathway-centric module that identifies compounds whose anti-PGCC efficacy is influenced by specific molecular features, and (2) a compound-centric module that identifies genes or pathways associated with the efficacy of a selected compound. Statistical comparisons, interactive visualizations, and annotations from external knowledge bases enable users to explore drug-feature relationships and generate new hypotheses. The platform facilitates systematic investigation of PGCC vulnerabilities through data-driven exploration. Example analyses demonstrate that compounds targeting oxidative stress response and cytoskeletal remodeling pathways preferentially suppress PGCC-enriched cell populations, consistent with their structural plasticity and adaptive signaling. Integration with baseline pharmacogenomic datasets further distinguishes PGCC-selective inhibitors from broadly cytotoxic agents, supporting the prioritization of therapeutic candidates. In summary, this interactive web resource provides an accessible and scalable framework for analyzing the molecular and pharmacologic landscape of PGCCs. By integrating our prior and ongoing high-throughput datasets with multi-omic data, it accelerates the discovery of biomarkers and therapeutic strategies to overcome treatment resistance driven by genome-doubled tumor populations.

Poster P04

Title: Associated Cell Types and Influential Genes Identified through Integration of Genome-Wide Association Studies and Single-cell Transcriptomics for Circulating Fatty Acid Traits

Author list: Elijah Sterling[1,2], Saurav Choudhary[3], Kaixiong Ye[1,3]

Detailed Affiliations:

[1]Department of Genetics, University of Georgia, Athens, GA, USA, [2]Regenerative Bioscience Center, University of Georgia, Athens, GA, USA, [3]Institute of Bioinformatics, University of Georgia, Athens, GA, USA.

Abstract: Background: Genome-Wide Association Studies (GWAS) identify single-nucleotide polymorphisms (SNPs) associated with traits. Fatty acid traits are of particular interest because they contribute to diverse biological processes, including membrane composition, cytokine signaling, and liver disease. A challenge in GWAS is that most association signals reside in non-coding regions of the human genome. Recent advances in multi-omic technologies enable the functional annotation of GWAS signals and the discovery of underlying biological mechanisms. Methods: We used a multi-method approach to identify cells, tissues, and genes associated with fatty acid traits. Our lab previously published a GWAS of 19 circulating fatty acid (cFA) traits in more than 200,000 European-ancestry individuals from the UK Biobank. In this study, we integrate the GWAS summary statistics with single-cell transcriptomic data from a human liver cell atlas spanning postnatal to older adult stages ($n > 100,000$ cells) and spatial transcriptomics data from a human embryo dataset representing Carnegie stage 8 ($n > 35,000$ spots). We applied several bioinformatic tools to identify associated cell types (scDRS v1.0.3, seismicGWAS v1.0.0), enriched tissues (gsMap v1.73.5), and influential genes (seismicGWAS v1.0.0). Results: We found that centrilobular and periportal hepatocytes were significantly associated with 19 cFA traits across multiple cell-type association methods (FDR < 0.05). We also identified significant enrichment in the endoderm for cFA traits ($p < 0.05$). In addition, we identified influential genes contributing to the signal in centrilobular hepatocytes (FDR < 0.05) with three genes being identified as influential across all 19 cFA traits (MLXIPL, APOB, and SLC22A1). KEGG enrichment analysis further highlighted significant pathways including complement and coagulation cascades, bile secretion, and lipid and atherosclerosis ($p < 0.05$). Conclusions: In summary, our findings suggest that two specific hepatocyte subpopulations are associated with cFA traits. We further identified the endoderm as an enriched developmental tissue, supporting a potential role for early lineage specification in fatty acid biology. In addition, we elicited three influential genes relevant for liver function that are present across all CFA traits. These results also implicate specific genes and pathways in mediating the effect of observed associated variants on the fatty acid traits.

Keywords: Fatty acids, Hepatocytes, Genome-Wide Association Studies (GWAS)

Poster P05

Title: From Screening to Differential Diagnosis: A 3D Point Cloud Benchmark for Craniofacial Conditions

Author list: Tingting Chen[1], Mian Umair Ahsan[1], Justin Cotney[2,3], Eric Liao[2], Kai Wang[1,4]

Detailed Affiliations:

[1]Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA, USA, [2]Center for Craniofacial Innovation, Department of Surgery, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA, [3]Department of Genetics, University of Pennsylvania, Philadelphia, PA, USA, [4]Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA.

Abstract: Background. Three-dimensional (3D) facial morphology encodes clinically meaningful information about craniofacial conditions, and most automated diagnostic approaches require mesh registration, in which a facial scan is aligned to a standard face model. However, registering scans with craniofacial dysmorphology to a normative template can introduce biases: the registration process may force atypical anatomy toward a “normal” shape, distort or attenuate clinically meaningful features. It remains unclear whether 3D point-cloud encoders that directly analyze raw unregistered mesh can capture diagnostically relevant facial shape across a broad spectrum of conditions, ranging from orofacial clefts to rare genetic syndromes. Method. We benchmarked four architectures (PointNet, PointNet++ with multi-scale grouping, DGCNN, and a handcrafted Geometric MLP) using 3D facial scan data from the FaceBase consortium (27,000+ scans on 18,076 subjects across 8 datasets from multiple collection sites). We evaluated five classification tasks related to orofacial cleft screening and rare syndrome prediction. All experiments used subject-level stratified splits (70/15/15) to prevent within-patient scan leakage.

For subjects with multiple scans, per-scan embeddings were averaged into a single subject-level representation at inference. For syndrome classification tasks, we paired a cross-entropy-trained backbone with prototype retrieval. Results. PointNet++ achieved the strongest performance in all tasks. Binary screening of orofacial cleft was robust: cleft detection (n=6,938 scans) reached accuracy of 84.1% and AUC of 0.856, and a pan-condition affected-vs-control task (n=24,973) achieved accuracy of 90.2% and AUC of 0.960. Fine-grained multi-class syndrome classification was more challenging: subject-aggregated prototype retrieval reached AUC of 0.812 for 19-category syndrome-pathway (n=12,990) and AUC of 0.898 for 33-class clinical-diagnosis classification (n=3,623), the latter exceeding previously reported AUC of 0.863 by Boehringer (Am J Med Genet A. 2011). In classification tasks, AUC far surpassed top-1 accuracy, suggesting that discriminative embeddings were limited by conventional softmax heads under severe class imbalance. Disease recognizability was driven by the presence of characteristic craniofacial abnormality than by class size (e.g. Williams syndrome is more readily recognized than CHARGE and Rett syndromes). Several of the best-recognized classes were single-gene syndromes with strong craniofacial phenotypes (e.g. Achondroplasia caused by FGFR3 mutations, Crouzon and Apert caused by FGFR2 mutations). For these monogenic syndromes, gradient-based feature attribution highlighted regions broadly consistent with the known phenotype (midface and forehead features), though saliency remained spatially diffuse. However, models trained on data collected from one site and evaluated on an independent site showed near-chance performance (AUC $\approx$ 0.50), revealing batch effects as architecture-independent barriers to generalization. Conclusion. Our results provided three lessons. First, subject-level embedding aggregation provides training-independent performance gain and should be a default strategy for subjects with multiple scans. Second, in multi-class diagnostic tasks, the primary bottleneck is the classification head rather than point-cloud encoder; pairing discriminative embeddings with prototype retrieval recovers substantial performance. Third, task granularity determines difficulty: binary screening is highly reliable, but fine-grained disease diagnosis remains limited by class imbalance, registration-independent representation learning, and batch effects. Together these findings motivate a two-stage framework for future studies: initial binary screening followed by retrieval-based differential diagnosis using subject-aggregated 3D facial embeddings.

Poster P06

Title: Characterizing the genetic architecture and molecular mechanism of circulating fatty acids

Author list: Kaixiong Ye[1,2], Huifang Xu[1], Saurav Kumar Choudhary[2], Haifeng Zhang[1], Jiayi Xue[1], Ge Yu[1], Ke Yi[1], Pengpeng Bi[1,3]

Detailed Affiliations:
[1]Department of Genetics, University of Georgia, Athens, Georgia, US, [2]Institute of Bioinformatics, University of Georgia, Athens, Georgia, US, [3]Center for Molecular Medicine, University of Georgia, Athens, Georgia, US.

Abstract: Background: Fatty acids (FA) play crucial roles in human health, influencing the risk of developing various conditions, such as cardiovascular disease and dementia. While previous genome-wide association studies (GWAS) have identified hundreds of genetic loci associated with the circulating FA levels, these loci account for only a small fraction of trait variance, and the underlying biological mechanisms linking these identified loci to FA metabolism remain largely unclear. Methods: We performed a multi-trait analysis of GWAS (MTAG) on 19 FA traits (N = 239,268) from UK Biobank to boost discovery power. SNP-based heritability was estimated using GREML-LDMS and partitioned across functional categories using LDSC. We integrated GWAS with a single-cell CRISPRi screen to functionally characterize the target genes and the molecular mechanisms underlying FA-associated loci. Results: Our MTAG identified 224 genomic loci that are significantly associated with 19 FAs, including 150 novel discoveries not detected in single-trait GWAS. We estimated that genome-wide imputed SNPs explained 11-31% of phenotypic variance. FA-related signals are largely enriched in high LD regions with common variants (MAF > 5%), evolutionarily conserved regions, and regions with active enhancers. To explore the regulatory function of FA-associated loci, we conducted a single-cell CRISPRi screen in > 200,000 HepG2 liver cells, targeting 360 candidate cis-regulatory elements (CREs) from fine-mapped FA trait variants. We identified 501 CRE-gene pairs in cis, identifying 239 genes, such as APOB and FGG. We validated the top 16 candidate genes independently using targeted CRISPRi in bulk cell cultures. The identified cis-regulated genes are highly interconnected and enriched in lipid metabolism pathways (e.g., high-density lipoprotein particle remodeling). Notably, 20 of these genes were identified as transcription factors, with 12 (e.g., TCF7L2) having established roles in lipid metabolism. Inhibition of the CRE containing the putative causal variant rs12259064 significantly altered

the expression of CDK6, LGR5, and NKD1. As these genes are responsive to TCF7L2 depletion, our findings suggest that the CRE (rs12259064) locus exerts trans-effects through modulating TCF7L2. Conclusion: Our integrative multi-trait and functional genomic analyses reveal novel genetic loci and the molecular mechanisms regulating circulating FA levels, providing mechanistic insights into the genetic architecture of fatty acid metabolism.

Keywords: Fatty acids; Genome-wide association studies (GWAS); Single-cell CRISPR screen; Integrative Omics;

Poster P07

Title: STITCH: A Platform-Agnostic Pipeline for Barcoding-based Spatial Transcriptomics

Author list: Yuanhang Liu[1], Brenna Novotny[1], Humza Ashraf[1], Mariam Stein[1], Jeong-Heon Lee[2], Ting Zhang[2], Alejandro Stark[2], Sathiya Pandi Narayanan[2], Corey Seavey[2], Arthur Beyder[2], Tamas Ordog[2], Chen Wang[1]

Detailed Affiliations:

[1]Department of Quantitative Health Sciences, Mayo Clinic, MN, USA, [2]Department of Physiology and Biomedical Engineering, Mayo Clinic, MN, USA.

Abstract: Barcoding-based spatial transcriptomics (ST) technologies have gained rapid adoption in recent years to investigate biological questions in spatial context. Major platforms include Visium and Visium HD (10x Genomics), Seeker and Trekker (Takara Bio), and Stereo-seq (Complete Genomics), providing options at varying spatial resolutions and sensitivities. Those platforms share a common output structure consisting of spatially indexed gene expression measurements associated with spatial units, e.g., spots, bins, beads, or cells. Despite this shared structure, no unified analytical workflow currently supports processing all those platforms. To address this gap, we developed Spatial and single-cell Transcriptomics Integration Tool for CHaracterization (STITCH), a Nextflow-based pipeline for platform-agnostic analysis of barcoding-based ST datasets. STITCH provides an end-to-end workflow including quality control, normalization, spatially variable gene (SVG) identification, cell type annotation via mapping or deconvolution, spatial embedding, data integration, clustering, differential expression, pathway analysis, cell–cell interaction and colocalization analysis. To enable efficient analysis of large-scale datasets, STITCH incorporates on-disk computation through BPCells. Final outputs are generated as both Seurat and AnnData objects to support downstream analysis in R and Python ecosystems. We additionally conducted benchmarking studies to evaluate combinations of normalization, SVG selection, and clustering methods for spatial domain identification using pathology-annotated datasets, including a public human dorsolateral prefrontal cortex (DLPFC) Visium dataset and in-house human colon Visium data. We demonstrated that incorporating spatial embedding, particularly Banksy, substantially improves spatial domain identification, especially in complex tissues such as brain where pathology-defined regions contain overlapping and heterogeneous cell populations. We further performed sparse data simulations using in-house generated mouse brain Trekker datasets with cell type annotations derived from the Allen Mouse Brain Atlas. Under high sparsity, clustering performance decreased across all method combinations, with TF-IDF showing marginally better robustness. For highly sparse data, cell type deconvolution using refined, cell type–specific marker genes provided improved cell typing performance. Overall, STITCH provides a unified and scalable framework for streamlined analysis of barcoding-based spatial transcriptomics data.

Poster P08

Title: Uncovering Sex-Biased COVID-19 Vaccine Adverse Events Using a Large Language Model

Author list: Anna He[1], Katelyn Hur[2], Xingxian Li[3], Jie Zheng[4], Yongqun He[4,5], Junguk Hur[6]

Detailed Affiliations:

[1]Huron High School, Ann Arbor, MI, USA, [2]Red River High School, Grand Forks, ND, USA, [3]Harvard Medical School, Boston, MA, USA, [4]Unit for Laboratory Animal Medicine, University of Michigan Medical School, Ann Arbor, MI, USA, [5]Department of Learning Health Sciences, University of Michigan Medical School, Ann Arbor, MI, USA, [6]Department of Biomedical Sciences, University of North Dakota School of Medicine & Health Sciences, Grand Forks, ND, USA.

Abstract: Introduction.

Licensed human vaccines are generally safe, but in rare cases, adverse events (AEs) occur. Vaccine AEs are unwanted and sometimes serious health outcomes following vaccination. Our recent study analyzing Vaccine

Adverse Event Reporting System (VAERS) data (December 2020 to June 2024) for five COVID-19 vaccines showed that females experienced more general and common AEs, while males were more associated with severe and cardiovascular AEs (ref. doi: 10.3389/fphar.2026.1741967). However, a comprehensive and systematic literature-based study is still lacking. To address this gap, we employed a large language model (LLM)-assisted literature assessment to evaluate literature support for these findings.

Methods.

We developed a Python-based pipeline using ChatGPT-4o prompt engineering to analyze 1,586 PubMed article abstracts, classifying evidence as supporting, partially supporting, contradicting, insufficient, or irrelevant. Comparative analyses were conducted between female- and male-associated results. The LLM-generated annotations were further manually reviewed by two independent human reviewers to assess accuracy and identify common sources of model error.

Results.

The assessment criteria were based on two sex-biased AE patterns observed in our VAERS analysis: females experience higher rates of common adverse events than males, whereas males experience higher rates of severe or cardiovascular adverse events than females. Our LLM analysis found that 98 articles supported, 46 partially supported, and 11 contradicted the female-pattern statement, while 130 articles supported, 28 partially supported, and 55 contradicted the male-pattern statement. Manual validation by two independent reviewers showed that 153 of 155 LLM-identified female-pattern annotations and 182 of 213 LLM-identified male-pattern annotations were correct. However, LLM identified male-pattern contradiction annotations were less reliable, with only 28 of 55 verified as correct. Manual review also showed that sex-based AE patterns varied by vaccine type and age. For example, myocarditis AE was more frequently reported after mRNA vaccine administration in young male adults, whereas headache and fever AEs were more frequently reported in females across all vaccine types.

Discussion.

This study presents the first comprehensive analysis of the effect of sex on COVID- 19 vaccine AEs using LLM-assisted literature assessment. Overall, the LLM evaluation supported the sex-biased patterns identified from VAERS analysis, and some condition- dependent contradictions were also found. Future directions include using electronic health record (EHR) resources (such as the National Clinical Cohort Collaborative N3C) and extending LLM analysis to full-text articles to enable deeper mechanistic investigation and guide sex-specific vaccine safety strategies.

Keywords: Large Language Model, COVID-19 vaccine, Adverse event, Sex difference, Vaccine safety, PubMed

---

Poster P09

Title: MetaAge: a Highly Accurate Ensemble Machine Learning Model for Biological Age Estimation

Author list: Liguo Wang[1], and Anping Chen[2]

Detailed Affiliations:

[1]Division of Computational Biology, Mayo Clinic College of Medicine, 200 1st Street SW Rochester, MN 55905

2An-Ping Chen, Boston Harbor Biotechnologies, 17 Briden Street, Worcester MA, 01605.

Abstract: Biological age is a measure of an individual's physiological state that reflects the cumulative effects of genetics, lifestyle, environmental exposures, and diseases, which cannot be fully captured by chronological age. DNA methylation (DNAm)-based epigenetic clocks have emerged as the most widely used and accurate biomarkers of biological aging, often outperforming transcriptomic, proteomic, metabolomic, and clinical biomarker-based approaches. Over the past decade, numerous DNAm clocks have been developed, including first-generation chronological age clocks (e.g., Horvath and ZhangEN), second-generation healthspan and mortality clocks (e.g., GrimAge and DunedinPACE), and more recent mechanistic aging clocks (e.g., YingCausAge). While these clocks have proven valuable for aging research, each was developed using specific training populations and optimization objectives, limiting their generalizability across diverse cohorts and applications. For example, ZhangEN achieves excellent accuracy in adults over 20 years old but performs poorly in adolescents; Horvath and Hannum primarily capture chronological age and are less predictive of morbidity and mortality; GrimAge is strongly associated with lifespan and disease risk but tends to exhibit systematic calibration biases across populations; DunedinPACE measures the pace of aging rather than biological age itself, making direct comparisons with age-prediction clocks challenging; and mechanistic clocks such as YingAdaptAge and YingDamAge capture specific biological processes

but often show reduced accuracy for chronological age estimation. These limitations suggest that no single clock fully captures the multidimensional nature of human aging, motivating the development of ensemble approaches that integrate complementary aging signals. Here, we developed MetaAge, an ensemble biological age estimator that integrates 17 DNAm clocks using gradient-boosted decision trees (XGBoost). By combining complementary aspects of the aging process captured by existing clocks, MetaAge achieved significantly higher accuracy and robustness than any individual clock across cross-validation analyses and multiple independent external validation cohorts. These findings demonstrate the power of ensemble learning frameworks to overcome the limitation of individual models and advance the precise quantification of human aging.

Poster P10

Title: An AI-powered single-cell and spatial omics ecosystem for neurodegenerative disease

Author list: Weidong Wu[1,2], Tae Yeon Kim[3,16], Jielin Xu[4,16], Hao Cheng[1,16], Cankun Wang[1,2,16], Weiqing Chen[5,6], Qi Guo[1], Xiaojie Jin[1], Jeremy A. Miller[7], Kaixuan Xiao[1], Shaohong Feng[1], Jonathan G. Miller[1], Shaopeng Gu[1,2], Julie Zhu[1], Xiaoying Wang[1,2], Elizabeth Wang[1], Yunxiao Ren[4], Marie-Amandine Bonte[3], Jing Zhao[1], Dana M. McTigue[3,11], Lang Li[1], Benjamin M. Segal[12,13], Guangyu Wang[5,6,8,9,10], Phillip G. Popovich[3,11,14], Feixiong Cheng[4], Hongjun Fu[3,15,17], Anjun Ma[1,2,17], Qin Ma[1,2,17,18]

Detailed Affiliations:

[1]Department of Biomedical Informatics, The Ohio State University, Columbus, OH, USA, [2]Pelotonia Institute for Immuno-Oncology, The James Comprehensive Cancer Center, The Ohio State University, Columbus, OH, USA, [3]Department of Neuroscience, The Ohio State University, Columbus, OH, USA, [4]Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, OH, USA, [5]Center for Bioinformatics and Computational Biology, Houston Methodist Research Institute, Houston, TX, USA, [6]Department of Physiology, Biophysics & Systems Biology, Weill Cornell Graduate School of Medical Science, Cornell University, New York, NY, USA, [7]Allen Institute, Brain Health, Seattle, WA, USA, [8]Center for Cardiovascular Regeneration, Houston Methodist Research Institute, Houston, TX, USA, [9]Center for RNA Therapeutics, Houston Methodist Research Institute, Houston, TX, USA, [10]Department of Cardiothoracic Surgery, Weill Cornell Medicine, Cornell University, New York, NY, USA, [11]Belford Center for Spinal Cord Injury, The Ohio State University, Columbus, OH, USA, [12]Department of Neurology, College of Medicine and Wexner Medical Center, The Ohio State University, Columbus, OH, USA, [13]Neuroscience Research Institute, The Ohio State University, Columbus, OH, USA, [14]Center for Brain and Spinal Cord Repair, The Ohio State University, Columbus, OH, USA, [15]Center for Brain Injury Recovery & Discovery, The Ohio State University, Columbus, OH, USA, [16]Equal-contributing authors: Tae Yeon Kim, Jielin Xu, Hao Cheng, and Cankun Wang, [17]Senior authors: Hongjun Fu, Anjun Ma, and Qin Ma, [18]Lead contact: Qin Ma.

Abstract: Neurodegenerative diseases (NDs) share pathological features but differ in cellular vulnerability, anatomical involvement, and progression. Disentangling these shared and divergent programs demands an integrated view at the single-cell level yet remains missing in the ND domain. Here, we present ssKIND, an AI-powered, comprehensive resource integrating 24.12 million annotated brain cells or nuclei from 4,532 single-cell/single-nucleus RNA sequencing samples and 1,380 spatial transcriptomic samples across nine NDs. From that, ssKIND harmonizes human and mouse ND brain atlases for cross-region and -disease analyses. We further introduce NeuroCLIP, a brain-specialized vision-omics foundation model fine-tuned on 0.96 million ST H&E histology images, connecting molecular profiles with brain tissue morphology in NDs. Using ssKIND, we uncover shared and disease-specific transcriptional programs across NDs, specifically, the HSP90-centered proteostasis response that tracks Alzheimer's disease progression. Furthermore, we prioritize microglial disease signatures to identify small-molecule drug candidates capable of reversing microglial stress states across NDs. ssKIND delivers an AI-powered, large-scale single-cell and ST ecosystem for understanding molecular mechanisms in NDs. ssKIND are publicly available at https://bmblx.bmi.osumc.edu/sskind/.

Poster P11

Title: Building a Psychoplastogenic Signature: A Comparative Evaluation of Compound-Based and Target-Based Approaches in the CANDO Platform

Author list: Justin Shinkle[1], William Mangione[1], Zackary Falls[1], and Ram Samudrala[1]

Detailed Affiliations:

[1]Department of Biomedical Informatics, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, Buffalo, NY, USA.

Abstract: Psychoplastogens are a pharmacologically diverse class of compounds that rapidly loosen static-state constraints on cortical neurons, inducing lasting structural and functional plasticity with therapeutic potential in depression, PTSD, and substance use disorders. Their mechanistic diversity suggests that psychoplastogenicity reflects a distributed, systems-level pharmacologic property rather than acute activity at any single acute protein target, motivating the construction of a proteome-wide psychoplastogenic signature to guide compound repurposing and de novo design. Here we evaluate two complementary strategies for constructing such a signature within the CANDO platform: a compound-based approach and a target-based approach, comparing target-based methods for their ability to recapture known psychoplastogens relative to a CNS-active non-plastogenic control set. The compound-based approach, in which candidates are evaluated by similarity of their predicted proteome-wide interaction profiles to those of established psychoplastogens, recovered psychoplastogens at above-baseline rates across full proteome and synaptogenesis-relevant protein subset analyses, with psychoplastogens recovering each other at average recall rates (k = 10, 100, 500, 1000, 5000) of 0.0117, 0.045, 0.125, 0.241, and 0.69, compared to CNS control rates of 0.004, 0.021, 0.073, 0.126, and 0.498. We then evaluated target-based methods across 95 curated protein libraries derived from SynGO and Reactome pathway annotations. Our de novo compound predictor, which counts instances and sums interaction scores above a defined threshold for a group of proteins, failed to detect psychoplastogen enrichment when filtering to the 90th percentile or greater of interaction scores, with the best Benjamini-Hochberg FDR-adjusted p-value at 0.627. A second function ranking compounds by interaction score across proteins and averaging placements into a master rank list similarly failed to place psychoplastogens above CNS controls for any protein library. However, BANDOCK+, a newer scoring function with improved interaction prediction accuracy, detected statistically significant differences for certain protein libraries. Notably, Reactome protein library R-HSA-375280, containing amine-binding GPCRs broadly targeted by psychoplastogens, placed psychoplastogens at an average 9.7th percentile versus 29.4th percentile for CNS controls (MWU, $p<0.00001$). These findings indicate that target-based signature construction is sensitive to both method selection and protein library composition. The variable performance of BANDOCK+ warrants further investigation into which protein configurations best concentrate the psychoplastogenic pharmacologic signal. Identifying the optimal protein configuration would further enable evaluation of hybrid approaches that integrate compound-based and target-based signatures, which may yield greater predictive power for psychoplastogenicity than either strategy alone.

Poster P12

Title: Small Molecule and Peptide Based Inhibition Mechanisms for Mycobacterium Tuberculosis

Author list: Arhan Patel[1]

Detailed Affiliations:

[1]Department of Pharmacy and Pharmaceutical Sciences, University at Buffalo, Buffalo, NY, USA

Abstract: Mycobacterium tuberculosis uses the PPE18 protein to suppress immune responses by binding human TLR2 on macrophages. Since existing TB drugs don't target this mechanism, blocking PPE18-TLR2 interaction offers a novel therapeutic angle. This project develops computational tools to design inhibitors that prevent this binding. We first used AlphaFold2 and I-TASSER to generate the PPE18 structure. Using that structure, HotPoint and HADDOCK identify the TLR2-binding interface and hotspot residues to target for an interaction. For peptide generation, RFdiffusion generates binder scaffolds, which are then optimized with ProteinMPNN. Top candidates are validated using Rosetta scoring and AlphaFold2-Multimer to predict binding interfaces. For small molecule generation, inhibitors were designed by generating drug-like compounds via REINVENT and screening against the PPE18 TLR2 domain with AutoDockVina. Additionally, FDA-approved compounds from DrugBank are screened against the same site to identify drugs for repurposing. Designs are ranked by binding affinity and drug-likeness

and top candidates validated through binding/docking simulations to assess binding stability. This combined approach returns multiple lead candidates. Targeting an immune-evasion mechanism distinct from conventional TB drugs, this work opens a new avenue for treating drug-resistant TB without the risk of resistance.

Keywords: Mycobacterium tuberculosis, PPE18, TLR2, peptide inhibitor design, small-molecule docking, drug repurposing

---

Poster P13

Title: BindFuse: A Multimodal Graph Learning Framework for Protein-Ligand Binding Affinity Prediction

Author list: Bilal Ahmad[1], Ram Samudrala[1] and Zackary Falls[1,2]

Detailed Affiliations:

[1]Department of Biomedical informatics, Jacobs School of Medicine and Biomedical Sciences, State University of New York at Buffalo, NY, USA, [2]Department of Biostatistics and Bioinformatics, Roswell Park Comprehensive Cancer Center, Buffalo, NY, USA.

Abstract: Accurate prediction of protein–ligand binding affinity is a central problem in computational drug discovery due to the complex and multiscale nature of molecular recognition. Despite recent progress in protein language models and molecular representation learning, most existing approaches rely on a single modality, such as sequence, structure, or ligand topology, limiting their ability to fully capture interaction complexity and generalize across diverse protein–ligand systems. To address this limitation, we propose BindFuse, a multimodal graph learning framework that integrates evolutionary sequence information, structural representations, and ligand chemical features into a unified protein–ligand interaction graph. BindFuse leverages six complementary pretrained encoders and feature generators, including ESM-3, ProtBERT-BFD, AlphaFold3-inspired structural descriptors, structural diffusion representations, ChemBERTa, and RDKit-derived ligand features, producing 2,769-dimensional node embeddings. Protein–ligand interaction graphs are constructed using a 5.0 Å spatial cutoff, encoding both covalent and non-covalent interactions. We evaluate four graph neural network architectures, including AffinityFormer, our proposed Graphormer3D-based architecture incorporating edge-augmented attention bias, dynamic structural feature augmentation, stochastic depth, and masked mean pooling, alongside SchNet, GemNet-T, and GATv2Conv baselines. Experiments on the PDBbind v2020 benchmark using random, scaffold-based, and cold-target splits show that BindFuse consistently improves predictive performance relative to single-modality representations, achieving Pearson correlation coefficients of 0.724, 0.691, and 0.695, with corresponding RMSE values of 1.031, 1.037, and 1.060 pK units, respectively. In addition, large-scale experiments on the BindingDB dataset demonstrate scalability and robustness, achieving a Pearson correlation coefficient of 0.807. These results suggest that integrating complementary sequence, structural, and chemical modalities provides a more effective representation for protein–ligand affinity prediction and highlights the potential of multimodal graph learning for structure-based drug discovery.

Keywords: Protein–Ligand Binding Affinity, Multimodal Learning, Graph Neural Networks, Structure-Based Drug Discovery, Protein Language Models, Geometric Deep Learning,

---

Poster P14

Title: High CRISPR-Cas System Prevalence and Discovery of the Hyper-CRISPR Phenomenon in Antibiotic Resistant Escherichia coli: Comprehensive Genomic Analysis of 972 Clinical Isolates Reveals Phylogenetic Clustering and Temporal Evolution

Author list: Zhabiz Golkar[1], Dawn Freeman[1] and Anil K. Challa[1]

Detailed Affiliations:

[1]The Center of Excellence for Research and Program Innovation, Voorhees University, Denmark, SC, USA.

Abstract: Background: CRISPR-Cas systems function as bacterial adaptive immune systems against mobile genetic elements, yet their role in shaping the accessory genomes of antibiotic-resistant pathogens remains poorly characterized. Escherichia coli, a critical opportunistic pathogen with rising multidrug resistance, represents an ideal model for investigating CRISPR-mediated genomic dynamics. Objective: We conducted a comprehensive genomic survey to characterize CRISPR-Cas system distribution, evolutionary dynamics, and resistance correlations across a large clinical E. coli population. Methods: A total of 972 complete Escherichia genomes (952 E. coli and 20 closely related Escherichia species) spanning 2002–2019 were analyzed using CRISPRCasFinder, with systematic

characterization of CRISPR arrays, cas gene operons, and anti-CRISPR (acr) proteins. Results: We report near-universal CRISPR prevalence (99.6%; 968/972 strains) with 81.0% harboring complete CRISPR-Cas systems. Notably, 37.6% of strains exhibited a "hyper- CRISPR" phenotype (≥7 arrays), with 9.2% displaying "ultra-hyper" characteristics (≥10 arrays). Temporal analysis revealed a modest but statistically significant increase in mean array counts from 5.31 (2008) to 6.36 (2019) (individual genome regression: $R^2 = 0.012$, $p < 0.001$; n=964). Phylogenetic analysis demonstrated non-random clustering of hyper-CRISPR strains within specific sequence types, with the ST11 lineage showing 45.1% hyper-CRISPR prevalence (permutation test $p = 0.003$). Conclusions: This study reveals high CRISPR-Cas prevalence in clinical E. coli and documents the hyper-CRISPR phenomenon, contributing to understanding of bacterial adaptive immunity and its relationship to antibiotic resistance evolution. These findings establish the descriptive foundation for understanding CRISPR dynamics in clinical populations; the mechanistic relationship between CRISPR architecture and antibiotic resistance coexistence remains a hypothesis requiring experimental validation.

Keywords: CRISPR-Cas systems; Escherichia coli; antibiotic resistance; bacterial adaptive immunity; comparative genomics; phylogenetic analysis; temporal evolution; hyper-CRISPR; mobile genetic elements

---

Poster P15

Title: MetaboAgent: A Reliable LLM-Agent Architecture for Reproducible, Code-Free Metabolomics Analysis

Author list: Kasim Mawari[1], Fadhl Alakwaa[2]

Detailed Affiliations:

[1]Nova Southeastern University, Fort Lauderdale, FL, USA; [2]Division of Nephrology, Department of Internal Medicine, University of Michigan, Ann Arbor, MI, USA.

Abstract:

Background: As LLM-based agents enter scientific workflows, a core tension emerges between usability and rigor. Natural-language interfaces could open complex omics analysis to researchers without coding expertise, yet the stochastic behavior of LLMs makes them unreliable for analyses that must be reproducible: the same biological question, paraphrased, can yield different computational outcomes. Metabolomics, where insight depends on a brittle sequence of preprocessing and statistical steps, exemplifies this challenge.

Methods: MetaboAgent resolves this tension through an architecture that decouples language understanding from statistical computation. Rather than letting the model perform analysis, it uses Claude Sonnet 4.6 solely to route a user's plain-English request to one of eleven deterministic Python functions implementing the MetabolonR pipeline (quality control, KNN imputation, log transformation, scaling, PCA, sources-of-variation, pathway variation, differential abundance, and session export). Because every numerical operation executes in fixed, version-controlled code, identical intent yields identical output regardless of wording, and a structured-JSON session log permits exact, model-free re-execution. A deployed Streamlit interface accepts user CSV uploads and renders interactive PCA projections annotated by clinical covariates such as disease state or batch.

Results: To quantify reliability, we ran a controlled prompt-perturbation study: 20 paraphrases of equivalent analytical requests, each issued three times (60 total runs), against the Progredir chronic kidney disease cohort (450 samples; 151 metabolites after QC). Across unambiguous requests, configuration parameters matched in 100% of runs and the recovered set of significant metabolites was perfectly stable (Jaccard = 1.000). When a request was genuinely ambiguous, the agent solicited clarification every time instead of inferring intent. Observed variability in tool ordering (42% agreement) traced entirely to optional diagnostic steps and never altered the substantive findings.

Conclusions: These results show that an LLM agent can deliver conversational ease without compromising reproducibility, provided that interpretation and computation are kept strictly separate. Beyond metabolomics, this intent-routing-over-deterministic-tools pattern offers a transferable blueprint for building auditable, trustworthy AI assistants across bioinformatics and biomedical informatics. MetaboAgent is openly accessible at https://metaboagent.streamlit.app/

Keywords: metabolomics, large language models, conversational agents, reproducibility, chronic kidney disease, bioinformatics

---

Poster P16

Title: Benchmarking variation detection methods with linear reference genomes and pangenomes in grapevines

Author list: Robert Robinson[1], Brooke Atamanyk[1] and Ping Liang[1]

Detailed Affiliations:

[1]Department of Biological Sciences.

Abstract: Analysis of genomic variants through whole-genome sequencing (WGS) has become a standard approach for the genetic study of organisms, owing to the rapid development of next-generation DNA sequencing (NGS) technologies over the past few decades. The rapid accumulation of WGS data has enabled the survey of genetic diversity across and within species at an unprecedented pace. In the meantime, technical challenges in accurately documenting genomic variants arise due to short read lengths, high sequencing error rates, and the extremely complex nature of genomes. Addressing these challenges necessitates continued development and optimization of data, analytical algorithms and tools, as well as the generation of high-quality datasets for training and assessing algorithm efficiency, often in an organism-specific fashion. In this study, we focus on the performance comparison of variant identification using linear reference-genome and pangenome approaches and their optimization for genomic diversity analysis of grapevines, an important crop worldwide, for which there is a lack of training datasets and standards. We determined the minimum sequencing coverage to achieve false-negative rates below 5% to be 25, 10, and 15 times the genome size, respectively, for Illumina paired-end reads (150 bp x 2), PacBio HiFi reads, and Oxford Nanopore reads, indicating read length and sequencing accuracy as the determining factors. We also determined the optimal filtering parameters for the linear reference genome using BCFtools and for the pangenome approach using vg giraffe to achieve low false-positive and false-negative rates. Additionally, two unexpected observations were made: the superior performance of PacBio data aligned to a linear reference genome compared to a pangenome, and a deterioration in ONT data's ability to detect INDELs at coverage greater than 15X. In conclusion, our results indicate that while PacBio HiFi, as a long-read platform with high accuracy, is the ideal WGS platform for variant analysis, the Illumina platform remains a competitive option due to its best cost-efficiency and sequencing accuracy. Similarly, while the pangenome approach offers better overall performance than the linear reference-genome approach, the latter remains an attractive option due to its ease of use and lower computational requirements. These observations, along with the optimal filter parameters we identified for each approach, should be valuable for genomic research of grapevine and other crops.

Keywords: grapevine, variation, linear reference genome, pangenome, benchmarking

---

Poster P17

Title: Towards a novel paper-based detection of the oomycete Saprolegnia

Author list: Hsin-Ho Wei[1], Bartolomeo Gorgoglione[2], Vipa Phuntumart[1]

Detailed Affiliations:

[1]Biological Department, Bowling Green State University, Bowling Green, OH 43402 USA, [2]Fish Pathobiology and Immunology Laboratory, College of Veterinary Medicine, Michigan State University, East Lansing, MI 48824 USA.

Abstract: Oomycetes of the genus *Saprolegnia* are important pathogens affecting natural ecosystems and aquaculture worldwide. Saprolegniasis (cotton-wool disease) is commonly diagnosed based on the general morphology of clinical lesions rather than species level molecular identification, particularly *S. parasitica* being a primary pathogen. Developing reliable, rapid, and cost-effective molecular diagnostics remains challenging because these pathogens exhibit high genetic diversity and evolutionary incongruence, including discordance between nuclear internal transcribed spacer (ITS) and mitochondrial cytochrome c oxidase I (COX1) markers, which can lead to species misidentification. In our study, we observed this discordance, which may be attributed to potential hybridization, database misidentifications or other factors. Given the distinct evolutionary histories represented by nuclear and mitochondrial markers, we evaluated ITS and COX1 for diagnostic suitability. Because ITS showed greater consistency in species-level identification and phylogenetic resolution, it was selected as the target region for developing a paper-based loop-mediated isothermal amplification (LAMP)-CRISPR-Cas12a diagnostic assay targeting the nuclear ITS region of *Saprolegnia*. Two guide RNAs (gRNAs) were designed to enable either broad detection of *Saprolegnia* species or specific detection of *S. parasitica* and *S. ferax*. Under laboratory conditions, the assay demonstrated high specificity for *Saprolegnia*, with no cross-reactivity with non-target aquatic fungi, including *Fusarium* and *Aspergillus*. The assay was validated using 70 *Saprolegnia* isolates collected between 2023 and 2024 from natural ecosystems, hatcheries, and aquaculture facilities across Ohio, Michigan and Indiana. Our

developed assay achieved a detection sensitivity of 20 fg of DNA. Overall, this study highlights a value of translational bioinformatics for diagnostic target selection and establishes a promising, instrument-free molecular platform for field-ready diagnostics and precision aquaculture management.
Keywords: Oomycete, cotton-wool disease, LAMP-CRISPR, Phylogenetic analysis, Aquaculture diagnostics

---

Poster P18

Title: Inferring and Predicting Clade-Level Relative Transmission Fitness in Seasonal Influenza A Using Differential Population Growth Rate and Deep Learning

Author list: Ursula Senam Nkonu[1], Richard Annan[2], Letu Qingge[2], Hong Qin[3]

Detailed Affiliations:
[1]Biomedical Sciences PhD Program, Old Dominion University, Norfolk, VA, USA, [2]Department of Computer Science, North Carolina A&T State University, Greensboro, NC, USA, [3]School of Data Science, Department of Computer Science, Old Dominion University, Norfolk, VA, USA.

Abstract: Seasonal influenza A evolves rapidly, allowing newly emerged clades to replace previously dominant lineages and complicate surveillance and vaccine evaluation. Here, we applied the Differential Population Growth Rate (DPGR) framework to GISAID-derived H3N2 and H1N1 surveillance data collected from 1 January 2014 to 12 February 2026, including the 2025–2026 influenza season, to estimate clade-level relative transmission fitness across continents and within the United States. We identified windows of cocirculation with sliding-window regression, reconstructed relative-fitness relationships among clades, and compared inferred growth advantages with independent WHO and CDC surveillance patterns. We further trained subtype-specific convolutional neural networks on complete viral genomes to predict DPGR from sequence, quantified predictive uncertainty with conformal prediction, and used SHAP to localize genomic contributors to fitness. DPGR recovered recurrent lineage turnover in both subtypes and consistently identified the emerging H3N2 subclade K as fitter than the 2025-2026 vaccine-lineage background across multiple regions. Genome-based models predicted DPGR accurately for H3N2 ($R2 = 0.9577$) and H1N1 ($R2 = 0.9871$), while interpretation highlighted known haemagglutinin antigenic sites together with contributions from internal genes. These results support DPGR as an interpretable surveillance signal and show that influenza fitness can be linked to genomic prediction and biological interpretation in a unified framework.

Poster P19

Title: Integrative Mendelian Randomization and Protein-Protein Interaction Network Analysis Links Multi-Organ Aging to Complex Diseases

Author list: Chunxuan Ma[1,2], Yuan Hou[1], Feixiong Cheng[1,3,4,5]

Detailed Affiliations:
[1]Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, Ohio, USA, [2]Department of Population and Quantitative Health Sciences, School of Medicine, Case Western Reserve University, Cleveland, Ohio, USA, [3]Department of Molecular Medicine, Cleveland Clinic Lerner College of Medicine, Case Western Reserve University, Cleveland, Ohio, USA, [4]Case Comprehensive Cancer Center, Case Western Reserve University School of Medicine, Cleveland, Ohio, USA, [5]Cleveland Clinic Genome Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, Ohio, USA.

Abstract: Background: Biological aging is heterogeneous across organ systems, with distinct but functionally interconnected genetic architectures. Genome-wide association studies (GWAS) have independently characterized organ-specific aging loci and disease risk variants. However, the molecular interactomes through which aging signals across multiple organs converge to drive complex disease, including Alzheimer's disease (AD), remain poorly understood. Here, we applied Mendelian randomization (MR) integrating organ-specific biological age gap (BAG) GWAS with organ-specific expression quantitative trait loci (eQTL) data to identify genes linking aging organ systems to AD. We conducted a multi-organ protein-protein interaction (PPI) network to identify convergent genes and pathways implicated in aging-linked AD mechanisms. Methods: We applied two-sample MR using organ-specific eQTL as instrumental variables to analyze organ-specific BAG GWAS across nine systems and case-control AD GWAS. eQTL data were obtained primarily from the Genotype-Tissue Expression (GTEx v10) database, with immune and metabolic eQTL derived from the Database of Immune Cell Expression, eQTLs and Epigenomics

(DICE) and published datasets. MR analyses were performed separately for each organ aging system and AD, identifying candidate genes. These organ aging and AD-associated genes were integrated into a PPI network to identify convergent genes between aging functional modules and the AD disease modules. Candidate genes were validated using single-cell RNA sequencing (scRNA-seq) and pathway enrichment analyses. Results: MR analyses identified between 9 and 246 aging-associated genes per organ system and 113 genes associated to AD risk. MR analysis identified PLEKHA1 as an immune agingand AD-associated gene. Network analysis further identified PLEKHA1 as a convergent hub connecting brain and lung aging (NINL), metabolic aging (WDR45B), brain aging (PTPN13), and AD-associated (APP) molecular nodes. The scRNA-seq analysis showed significant (Benjamini-Hochberg FDR < 0.05) upregulation of PLEKHA1 in neurons and astrocytes in AD compared with control counterpart. Network-connected genes NINL, WDR45B, and PTPN13 also showed significant differential expression between AD and control populations in neuronal and glial cell types. Pathway enrichment identified phosphatidylinositol phosphate (PIP) synthesis, phosphatidylinositol (PI), phospholipid metabolism, and RND GTPase cycling as top processes. These pathways suggest that convergent multi-organ aging networks centered on PLEKHA1 may contribute to lipid membrane dysregulation and inflammatory processes relevant to AD. Conclusion: This study identifies PLEKHA1 as a convergent hub centered multiple organ aging signals with AD risk. These findings provide evidence for the contribution of biological aging across multiple organ systems to AD risk and show the value of multi-organ network integration for investigating how aging drives the onset of complex diseases.

Poster P20

Title: Bridging Computer Vision and Food Policy: A YOLOv11 Produce Freshness Detector Aligned with California AB 660 Date Label Vocabulary

Author list: Aston Liu[1], Hari Krishna Yalamanchili[2]

Detailed Affiliations:
[1]The Kinkaid School, Houston, TX, USA, [2]Baylor College of Medicine, Houston, TX, USA

Abstract: Date-label confusion is a measurable public-health and food-waste problem. A 2025 national survey by the Harvard Law School Food Law and Policy Clinic, ReFED, and the Johns Hopkins Bloomberg School of Public Health found that 43% of U.S. adults always or usually discard food near its printed date. ReFED estimates such confusion wastes three billion pounds of food, worth $7 billion, annually. Misread labels cut both ways: edible food discarded, unsafe food retained past safety dates. California AB 660 (signed September 28, 2024, effective July 1, 2026) mandates “Best if Used By” for quality and “Use By” for safety. A proposed federal equivalent (H.R.4987/S.2541) is now in subcommittee. To our knowledge, no consumer-facing visual AI tool operationalizes this quality/safety distinction at the consumer’s point of decision. Existing YOLO and CNN-based freshness models stop short of regulatory mapping. Mukhiddinov et al. (2022, Sensors) reported 50.4% average precision across 10 produce types labeled fresh/rotten (20 classes) via improved YOLOv4. Fahad et al. (2022, CMC) extended to three ordinal stages (pure-fresh, medium-fresh, rotten), reaching 82-84% accuracy across 11 produce types. Neither map output to regulatory label vocabulary, and public datasets under sample the intermediate senescence zone.

We trained YOLOv11 on a primarily self-collected five-produce dataset (apples, strawberries, blackberries, raspberries, cucumbers) captured via time-lapse over five days under elevated humidity, supplemented by 100 curated endpoint images (Sriram, Kaggle). This design records gradual senescence (enzymatic browning, ethylene-driven softening, surface degradation) including the intermediate gray zone absent from endpoint-sampled benchmarks. Frames were sparsely sampled by day to prevent temporal leakage. The model was trained to 80 epochs. Two labels (fresh, rotten) map onto AB 660’s regulatory vocabulary, with fresh aligning to the Best-if-Used-By window and rotten approaching the Use-By threshold. The model achieved mAP50 76.8%, precision 81.0%, and recall 67.6%. Mukhiddinov et al. reported 50.4% average precision on binary detection across 10 produce types, while Fahad et al. reached 82-84% accuracy on three-stage classification, under different conditions than ours. The 67.6% recall reflects visual ambiguity across intermediate senescence stages, appropriate to an advisory rather than diagnostic instrument. The contribution is application framing: a consumer-education tool that maps detection output onto AB 660’s vocabulary (fresh → Best if Used By; rotten → Use By as a precautionary visual signal, not a microbial safety determination), bridging food policy and computer vision toward household-level label literacy where prior CV models have served only industrial quality control.

Keywords: food freshness detection, YOLOv11, object detection, food date labeling, California AB 660, food waste reduction

---

Poster P21

Title: Exposure, host biology, and ancestry-aware genetics shape a noninvasive skin carotenoid biomarker

Author list: Yixing Han[1*], Savannah Mwesigwa[2,4], Nancy E. Moran[3*], Neil A. Hanchard[1,3,4*]

Detailed Affiliations:

[1]Center for Precision Health Research, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, [2]Department of Medical Microbiology, College of Health Sciences, Makerere University, Kampala, Uganda, [3]USDA/ARS Children's Nutrition Research Center, Department of Pediatrics, Baylor College of Medicine, Houston, TX, [4]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX.

Abstract: Noninvasive skin carotenoid spectroscopy provides a scalable biomarker of fruit and vegetable exposure, but skin carotenoid score is a composite phenotype rather than a direct substitute for plasma carotenoid concentration. We developed a layered, ancestry-aware framework to quantify how circulating exposure, carotenoid species composition, host biology, and genetic architecture shape skin–plasma carotenoid discordance in two deeply phenotyped multi-ancestry cohorts with plasma carotenoids, skin spectroscopy, dietary and clinical covariates, intervention data, genome-wide genotypes, and genetic ancestry information. The exposure/composition layer showed that total plasma carotenoids strongly predicted skin carotenoid score, but plasma species mixture explained substantial additional variation. Adding plasma lycopene fraction to a research model containing total plasma carotenoids, age, sex, race/ethnicity, BMI, cholesterol, hemoglobin index, and melanin index increased explained variance from 0.591 to 0.703 and reduced AIC from 2428.9 to 2363.7. Higher lycopene fraction was associated with lower-than-expected skin scores, indicating a predictable source of skin-plasma mismatch. The host biology layer showed that composition-adjusted discordance was driven primarily by optical/readout-related factors, especially melanin and hemoglobin indices, with smaller contributions from adiposity and lipid transport. These associations replicated in an independent intervention cohort baseline subset, and discordance showed substantial longitudinal stability over six weeks. The genetic layer distinguished calibration mismatch from the measured skin phenotype. SNP heritability was higher for measured skin carotenoid score than for total plasma carotenoids (h2=0.30 vs. 0.08), whereas discordance was not treated as a primary GWAS trait. An ancestry-aware linear mixed-model GWAS of measured skin score identified suggestive lead loci, with the top association at chr8:80665712:A:G (P=7.15e-8). Standard regression produced additional ancestry-structured signals with large allele-frequency differences across groups, underscoring the need for mixed-model adjustment. Conditional lead-locus analyses showed that plasma carotenoids and melanin/hemoglobin covariates attenuated genetic effects, suggesting contributions from both systemic carotenoid burden and skin-side measurement biology. This framework turns a convenient dietary readout into an interpretable genetic and biomarker phenotype and provides a general strategy for validating noninvasive health measures in diverse populations. This work converts a convenient noninvasive dietary readout into an interpretable computational and genetic phenotype. The framework provides a general strategy for validating scalable biomarkers in diverse populations, reducing ancestry and measurement-related bias, and improving the reliability of precision nutrition and public-health tools.

Poster P22

Title: Sex differences in AD at single-cell subtype level

Author list: Ayauzhan Arystambekova[1], Yuan Hou[1], Lijun Dou[1] and Feixiong Cheng[1]

Detailed Affiliations:

[1]Genome Center, Cleveland Clinic Research, Cleveland Clinic, Cleveland, OH, USA Department of Data Science, Cleveland State University, Cleveland, OH, USA.

Abstract: Sex differences in AD can be explained by factors other than the longevity gap between men and women. Among those include genetics (APOE4 gene), hormonal fluctuations (menopause), and immune/metabolic responses. However, the cellular and molecular mechanisms of such sex differences remain largely unexplored. We used a large multi-cohort single-cell dataset integrating 21 independent sources, 2,200 individuals, ~12.3 million cells, 7 brain regions, and 3 pathologies: Alzheimer's disease (AD), primary age-related tauopathy (PART) and cognitive resilience (CR) – at the clinical, transcriptional, pathway, and cellular levels. At the clinical level, sex was

a significant predictor only in AD (p=0.007). At the cell-subtype level, sex differences were subtype-specific and disease-context dependent. Microglia showed sex bias across all pathologies: DAM1 (higher in females, p=0.016) and Tau microglia (higher in males, p=0.005) in AD; Homeostasis Microglia1 in both PART (p=0.043) and CR (p=0.018). In AD, inhibitory PVALB neurons were male-enriched (p=0.043), while excitatory L4 IT (p=0.022) and L5 ET (p=0.043) showed female-enrichment. In CR, a distinct set of subtypes was found:–LAMP5 LHX6, Chandelier, SST neurons, Synapse-associated astrocytes, and Homeostasis Microglia1. It suggests a sex-specific neuroprotective cellular program. Differential expression analysis identified two independent layers of sex-biased gene regulation. Sex-biased autosomal gene expression programs were found in glial cells, specifically microglia (DAM1) were female-biased for inflammatory mediators (PRORP, ATF7, and CYBB) and astrocytes–related to oxidative phosphorylation and mitochondrial ATP production (ZC3H12B and PUDP). Across all cell types, sex-chromosome-linked genes (KDM6A, UTY) formed a recurrent axis of sex bias, implicating dosage as a shared epigenetic mechanism. Cell-cell communication revealed two separate sex-biased axes: male-related to a DAM1/MHC1 program, supports crosstalk between microglia and excitatory neurons via the common pathways (APOE/TREM2) and (SPP1/integrin); whereas female-related was complementary-based and pro-inflammatory microglia. Together, these findings reveal context-dependent sex difference in AD, highlighting subtype-specific and pathology-specific molecular targets for sex-informed therapeutic strategies.

Poster P23

Title: 15 years of the WashU Epigenome Browser

Author list: Chanrung (Chad) Seng[1], Wenjing Zhang[1], Tianjie Liu[1], Daofeng Li[1], Ting Wang[1]

Detailed Affiliations:

[1]Department of Genetics, The Edison Family Center for Genome Sciences & Systems Biology, Washington University School of Medicine, St. Louis, MO, USA.

Abstract: The WashU Epigenome Browser (https://epigenomegateway.wustl.edu/) is a web-based tool for exploring genomic data and providing visualization, investigation, and analysis of epigenomic datasets. Since its 2018 update, the redesigned user interface and newly developed features have enhanced how investigators interact with both the Browser and the extensive genomic data it hosts. The rapid evolution of the JavaScript ecosystem has presented new challenges and opportunities in maintaining and developing the WashU Epigenome Browser. In this update, we present a completely rewritten codebase. This new codebase minimizes the use of external libraries whenever possible, resulting in a significantly smaller code bundle size after production compilation. The reduced code size improves loading efficiency and boosts the Browser's performance, with improved scripting, graphics rendering, and painting performance. Lowering external dependencies also allows for faster and more straightforward installation. Additionally, the update includes a redesign of the user interface to further enhance user experience and features a new modular design in the codebase that enables the Browser to be exported as stand-alone modules for use in other web applications. Several novel track types for long-read methylation data and single-cell methylation data visualization have been added, and we continue to update and expand the data hubs we host for major consortia. We constructed the first data hub to systematically compare genomic data mapped to different genome assemblies, focusing on comparisons between hg38 and the first human T2T genome, chm13, using our new comparative genomics track function. The WashU Epigenome Browser also serves as a foundation for other genomics platforms, such as the WashU Virus Genome Browser, developed for SARS-COV-2 research, the WashU Comparative Epigenome Browser, and the WashU Repeat Browser.

Poster P24

Title: An Integrated Single-Cell and Epigenomic Resource for Comparative Analysis of the Basal Ganglia from BICAN

Author list: Wenjin Zhang[1], Wubin Ding[2], Kai Li[3,4], Lei Chang[3,4], Amit Klein[2], Cindy Tatiana Báez-Becerra[5], Jonathan A. Rink[5], Anna Bartlett[2], Huaming Chen[2], Natalie Schenker[2], Nelson Johansen[6], Tyler Mollenkopf[6], Yuanyuan Fu[6], Yang Xie[3,4], Shane Liu[1], Chanrung Seng[1], Benpeng Miao[1], Tianjie Liu[1], Quan Zhu[7], Rebecca D.

Hodge[6], Trygve E. Bakken[6], Ed S. Lein[6], Michael Hawrylycz[6], Xiangmin Xu[8,9,10,11,12], M. Margarita Behrens[5], Bing Ren[3,4,7,13], Joseph R. Ecker[2,14], Ting Wang[1,15], Daofeng Li[1]

Detailed Affiliations:

[1]Department of Genetics, The Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, MO 63110, USA, [2]Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA, [3]Department of Cellular and Molecular Medicine, University of California San Diego, San Diego, CA, USA, [4]New York Genome Center, New York, NY, USA,[5] Computational Neurobiology Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA, [6]Allen Institute for Brain Science, Seattle, WA 98109, USA, [7]Center for Epigenomics, Department of Cellular and Molecular Medicine, University of California San Diego, La Jolla, CA, USA, [8]Department of Anatomy and Neurobiology, School of Medicine, University of California, Irvine, Irvine, CA 92697, USA, [9]Department of Microbiology and Molecular Genetics, School of Medicine, University of California, Irvine, Irvine, CA 92697, USA, [10]Center for Neural Circuit Mapping, University of California, Irvine, Irvine, CA 92697, USA, [11]Department of Biomedical Engineering, University of California, Irvine, Irvine, CA 92697, USA, [12]Department of Computer Science, University of California, Irvine, Irvine, CA 92697, USA, [13]Department of Genetics and Development, Systems Biology, Biochemistry and Molecular Biophysics, Columbia University Irving Medical Center, New York, NY, USA, [14]Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA, [15]McDonnell Genome Institute, Washington University School of Medicine, St. Louis, MO 63108, USA.

Abstract: The basal ganglia regulate motor, cognitive, and affective behaviors, and their dysfunction underlies diverse neurological and psychiatric disorders. Comprehensive, accessible multi-omics resources are needed to understand the regulatory mechanisms governing basal ganglia cell types. Here we present an open, interactive web-based platform for exploring single-cell multi-omics datasets from basal ganglia, generated using 10XMultiome, snm3C-seq, and Paired-Tag technologies by the BICAN (NIH BRAIN Initiative Cell Atlas Network) consortium. The platform is available athttps://basalganglia.epigenomes.net/ and enables integrated visualization of gene expression, chromatin accessibility, DNA methylation, histone modifications, and chromatin conformation across cell types and human, macaque, marmoset, and mouse species, with direct genome browser support and comparative epigenomic functionality. Representative analyses demonstrate cell-type-specific regulatory landscapes, conserved and species-specific regulatory elements, and links between epigenomic regulation and transcription. This resource provides a scalable, community-oriented foundation for advancing basal ganglia biology and interpreting regulatory mechanisms relevant to brain function and disease.

Poster P25

Title: Multi-sample iStar enables joint enhancement of spatial transcriptomics data across tissue samples

Author list: Xiaokang Yu[1†], Anthony Q.[2†], Mingyao Li[3*]

Detailed Affiliations:

[1]Statistical Center for Single-Cell and Spatial Genomics, Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA, [2]Graduate Group in Genomics and Computational Biology, University of Pennsylvania, Philadelphia, PA, USA, [3]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [†]These authors contributed equally to this work. [*]Corresponding author

Abstract: Spatial transcriptomics technologies provide genome-wide gene expression measurements while preserving tissue context, but their spatial resolution remains limited. iStar was recently developed to enhance spatial transcriptomics data by integrating histology images and transcriptomic measurements to reconstruct super-resolution gene expression patterns. However, the current framework analyzes each sample independently and cannot leverage information shared across multiple samples. We developed multi-sample iStar, an extension of iStar that jointly models multiple spatial transcriptomics samples within a unified framework. To enable information sharing across samples, multi-sample iStar first removes batch effects at two levels—across histology (H&E) images and across gene expression measurements—and then trains a prediction model on the harmonized data. This design allows the framework to borrow strength across samples while accounting for inter-sample variation, improving gene expression enhancement beyond what single-sample analysis can achieve. We evaluated multi-sample iStar on three independent spatial transcriptomics cohorts, comprising 36 HER2-positive breast cancer samples, 15 gastric cancer samples, and 56 lung cancer samples. Across all datasets, multi-sample iStar consistently

outperformed the original single-sample iStar framework and generated higher-quality super-resolution reconstructions. The resulting spatial expression patterns showed improved consistency with underlying tissue morphology, demonstrating the benefits of leveraging multiple samples simultaneously. In summary, multi-sample iStar extends image-guided spatial transcriptomics enhancement from single-sample to multi-sample analysis. By harmonizing both histology and expression data before joint modeling, the framework provides a practical and scalable strategy for improving gene expression resolution across tissue samples and broadens the applicability of super-resolution spatial transcriptomics analysis.

Poster P26

Title: Domain-Aware Evaluation Evo2 Across Evolutionary Molecular, Clinical, and Generative Contexts

Author list: Sama Salarian[1,†], Anurag Shandilya[2,†], Mridul Sharma[2], Nishtha Vaghela[2], Swati Padhee[1], Ranjith Padinahateeri[2], Srinivasan Parthasarathy[1], Ganesh Ramakrishnan[2] and Raghu Machiraju[1,*]

Detailed Affiliations:

[1]Department of Computer Science and Engineering, The Ohio State University, Columbus, 43210, Ohio, USA and [2]Koita Centre for Digital Health, Indian Institute of Technology, Bombay, 400076, Maharashtra, India. [†]Co-first author, [*]Corresponding author

Abstract: Motivation: Genome-scale foundation models aim to learn generalizable representations of biological sequences, supporting tasks such as variant effect scoring and sequence generation. However, model evaluation is frequently reported using aggregate performance metrics that may obscure variation across genomic contexts and biological domains. Here, we present a domain-aware evaluation of the genome foundation model Evo2 [8], examining how its predictions and generated sequences relate to evolutionary constraint, molecular context, and biological organization. We perform a four-stage analysis encompassing evolutionary alignment, molecular context dependence, trait-stratified variant-effect prediction, and genome-scale sequence generation. Results: First, we evaluate evolutionary alignment by quantifying the association between Evo2-derived variant loglikelihood ratios and PhastCons conservation scores across genomic contexts and trait domains. We observe heterogeneous correspondence, with consistent associations in selected coding regions and weaker or variable alignment in regulatory and other noncoding elements. Second, analysis of long noncoding RNA essentiality indicates dependence on cellular context, suggesting limited sensitivity to some context-specific regulatory dynamics. Third, trait-stratified evaluation on TCGA-IDH glioma, TCGA-BRCA, and ClinVar datasets shows that Evo2 achieves relatively strong performance on IDH glioma and ClinVar, but performance on TCGA-BRCA is sensitive to model scale, with Evo2-40B showing large and consistent gains over Evo2-7B across all trait categories. Finally, genome-scale analysis of Evo2-generated sequences shows that synthetic genomes contain predicted open reading frames (ORFs), but most predicted proteins lack detectable Pfam domains, indicating reduced inferred functional content relative to the natural reference genome. Rather than proposing a new model, this work provides a biologically grounded evaluation framework for identifying context-specific strengths and limitations of genome foundation models. Taken together, these results indicate that Evo2's performance varies across genomic and biological contexts, highlighting the value of disaggregated, biologically grounded evaluation when applying genome foundation models to functional genomics and translational use cases. Availability and Implementation: Our code and data paths are available at: https://github.com/samasalarian/ domain-aware-evaluation Contact: All correspondence should be addressed to: machiraju.1@osu.edu

Keywords: genome foundation models; Evo2; variant effect prediction; evolutionary conservation; trait-stratified evaluation; sequence generation; functional annotation

---

Poster P27

Title: Computational Identification of Natural Inhibitors Targeting Fiber Proteins of FAdV-1 and FAdV-4 Through Integrated Virtual Screening and Molecular Dynamics Simulations.

Author list: kardoudi Aminaa[1,2]*, Salaheddine Redouane[2], Benani Abdelouahebc[2,3], Kichou Faouzia[1], Charifa Drissi Touzania,[1] and Fellahi Sihama[1]

Detailed Affiliations:

[1]Department of Veterinary Pathology and Public Health, Hassan II Agronomic and Veterinary Institute , B.P. 6202, Rabat, 10 000, Morocco, [2]Laboratory of Genomics and Human Genetics, Institut Pasteur du Maroc, Casablanca,

Morocco. [3]Medical Biology Department, Molecular Biology Laboratory, Pasteur Institute of Morocco, 11 20360, Casablanca, Morocco. [*]Corresponding author
Abstract: Haut du formulaire Fowl adenoviruses (FAdVs) represent a major threat to poultry health, with serotypes FAdV-1 and FAdV-4 causing adenoviral gizzard erosion (AGE) and hepatitis-hydropericardium syndrome (HHS), respectively. A wide variety of afflicted birds, including chicken, pigeon, and psittacine species, have been reported to carry aviadenoviruses. The disease is highly contagious and spreads rapidly between flocks and farms through vertical and horizontal transmission. In this study, we implemented a multi-stage computational drug-discovery pipeline to identify natural inhibitors of the viral fiber proteins for both FAdV-1 and FAdV-4. A curated library of 7523 natural compounds from the African Natural Products Database (ANPDB) and the South African Natural Compounds Database (SANCDB) was subjected to pharmacophore-based filtering, molecular docking, ADMET prediction, and 500-ns molecular dynamics simulations against four structural targets: Fiber-1 and Fiber-2 of FAdV-4, and the Short and Long Fibers of FAdV-1. Three ligands ANPDB_6449 (-10.3 kcal/mol), ANPDB_2908 (-10.2 and -10.0 kcal/mol), and SANCDB_254 (-9.2 kcal/mol) consistently emerged as strong candidates across the entire computational workflow. While ANPDB_2908 demonstrated notable multi-target capability by binding to fiber proteins from both FAdV-1 and FAdV-4, ANPDB_6449 and SANCDB_254 exhibited strong serotype-specific potential, supported by stable interaction profiles and favorable drug-likeness characteristics. Together, these compounds highlight promising natural scaffolds for the development of targeted antiviral interventions against pathogenic FAdV serotypes. Runing Title: Natural Inhibitors of FAdV Fiber Proteins

---

Poster P28

Title: Strategies for robust, accurate, and generalizable benchmarking of drug discovery platforms

Author list: Melissa Van Norden[1], William Mangione[1], Zackary Falls[1]

Detailed Affiliations:
[1]Department of Biomedical Informatics, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, Buffalo, NY, USA.
Abstract: Consistent and accurate benchmarking is essential to the creation, improvement, and comparison of computational drug discovery technologies. We therefore revised the assessment protocols of our multiscale therapeutic discovery platform, the Computational Analysis of Novel Drug Opportunities (CANDO) platform, in order to bring them into alignment with best practices. These new assessment tools allowed us to optimize multiple parameters used in drug candidate prediction. We found that CANDO could predict 7.4% of drugs in the top 10 compounds for their respective diseases/indications with these optimized parameters when using the Comparative Toxicogenomics Database (CTD) as a gold standard drug-indication mapping; this rose to 12.1% when using only approved drug-indication associations extracted from the Therapeutic Targets Database (TTD). This improvement in the performance of CANDO when using the TTD was maintained when examining only matched drug-indication associations between the two mappings. Finally, we found that performance on an indication correlated with features of that indication: there was a weak correlation between performance and the number of drugs associated with the indication (Spearman > 0.3), and a moderate correlation with intra-indication chemical similarity (Spearman > 0.5). Performance as assessed by our original and new benchmarking protocols was also moderately correlated (Spearman > 0.5). This study shows the utility of robust benchmarking protocols not only for accurate performance assessment, but for the improvement of drug discovery pipelines and comprehension of the determinants of their performance.

Poster P29

Title: Quantum Computing for Single-Cell Biology: Networks, Dynamics, and Generative Models

Author list: James J. Cai[1,2]

Detailed Affiliations:
[1]Department of Veterinary Integrative Biosciences, [2]Department of Electrical & Computer Engineering, Texas A&M University, College Station, TX, USA.
Abstract: Single-cell technologies are generating unprecedentedly complex datasets, but current computational methods struggle to capture their full richness. Our work explores how quantum computing can transform this landscape. First, we developed a gate-based quantum circuit model where qubits represent genes and entangling

gates encode gene regulatory interactions, enabling us to infer network structure directly from single-cell transcriptomes (npj Quantum Inf., 2023). Next, we advanced a quantum annealing framework that formulates gene selection and cell fate dynamics as quadratic optimization problems, successfully revealing hidden drivers of differentiation and drug resistance (Quantum Mach Intell., 2025). Building on this foundation, our new research pioneers three directions: (1) quantum differential network algorithms to map gene regulatory network rewiring between cell states, (2) phase-controlled quantum circuits to jointly model single-cell gene expression and chromatin accessibility, and (3) quantum generative models that simulate single-cell data by leveraging entanglement to capture gene-gene and cell-cell interactions (arXiv:2510.12776, 2025). Together, these approaches show how quantum resources can unlock new biological insights, opening the path toward quantum-enabled single-cell biology.

Poster P30

Title: Toward Computationally Complete Spatial Omics

Author list: Wei Li[1], Liran Mao[1,2,3], Yunhe Liu[4], Fuduan Peng[4], Nadja Sachs[5,6], Wenrui Wu[7], Stephanie Pei Tung Yiu[7], Hanying Yan[1], Amelia Schroeder[1], Xiaokang Yu[1], Kaitian Jin[1], Shunzhou Jiang[1], Zihao Chen[1], Melanie L. Loth[1], Lorena Gomez[8], Idania Lubo[8], Niklas Blank[9], Laith Z. Samarah[10,11,12], Ankit Basak[13], Ye Won Cho[14], Chia-Yu Chen[14], David M. Kim[14], Alex K. Shalek[13], Luisa Maren Solis Soto[8], Joshua D. Rabinowitz[10,11,12], Muredach P. Reilly[15,16], Xuyu Qian[17], Christoph A. Thaiss[9], Lars Maegdefessel[5,6,18], Linghua Wang[4], Humam Kadara[8], Sizun Jiang*[7], Yanxiang Deng*[2,19], Mingyao Li*[1,2]

Detailed Affiliations:

[1]Statistical Center for Single-Cell and Spatial Genomics, Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA,[2]Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA,[3]Graduate Group in Genomics and Computational Biology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA,[4]Department of Genomic Medicine, University of Texas MD Anderson Cancer Center, Houston, TX, USA,[5]Department for Vascular and Endovascular Surgery, TUM Klinikum Rechts der Isar, Technical University of Munich, Munich, Germany,[6]German Center for Cardiovascular Research (DZHK), Berlin, Germany; Partner Site Munich Heart Alliance,[7]Center for Virology and Vaccine Research, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA,[8]Department of Translational Molecular Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA,[9]Arc Institute, Palo Alto, CA, USA; Department of Pathology, Stanford University, Stanford, CA, USA,[10]Department of Chemistry, Princeton University, Princeton, NJ, USA,[11]Lewis-Sigler Institute of Integrative Genomics, Princeton University, Princeton, NJ, USA,[12]Ludwig Institute for Cancer Research, Princeton University, Princeton, NJ, USA,[13]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA, USA,[14]Department of Oral Medicine, Infection and Immunity, Harvard School of Dental Medicine, Boston, MA, USA,[15]Division of Cardiology, Department of Medicine, Vagelos College of Physicians and Surgeons, Columbia University, New York, NY, USA,[16]Irving Institute for Clinical and Translational Research, Columbia University Irving Medical Center, New York, NY, USA,[17]Department of Pediatrics, Children's Hospital of Philadelphia, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA,[18]Department of Medicine, Karolinska Institute, Stockholm, Sweden,[19]Epigenetics Institute, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.

*Corresponding author

Abstract: Multimodal spatial omics has transformed biology by mapping molecular complexity within intact tissues, yet current technologies remain limited in the number of modalities measured simultaneously and often produce lower-quality data than single-modality assays. We present COSIE, a computational framework that generates high-resolution, multilayered molecular landscapes across tissue sections, individuals, and platforms. COSIE integrates histology, epigenome, transcriptome, proteome, and metabolome into a unified representation. Applied to 12 datasets spanning 10 spatial technologies, eight modalities, and nine tissue types, ranging from thousands of spots to millions of cells, COSIE outperforms existing methods. It resolves tissue structures, enhances noisy measurements, predicts unmeasured modalities, and captures dynamic processes. In human tumors, COSIE identifies invasive subregions linked to clinical outcomes and predicts spatial gene expression in TCGA samples using only histology images. By transforming fragmented data into comprehensive spatial maps, COSIE advances computationally complete spatial omics and the creation of digital tissue twins for biomedicine.

Poster P31

Title: Interpreting Omics Data Analysis with Large Language Models for Disease Target and Drug Discovery

Author list: Zixi Xu[1*], Weihang Chen[1,2*], Wuyu Ren[1*], Tianqi Xu[1], Somadina Amaechina[1], Raad Khan[1,4], Yixin Chen[2], Michael Province[4], Philip Payne[1], Fuhai Li[1,2,5#]

Detailed Affiliations:

[1]Institute for Informatics, Data Science and Biostatistics (I2DB), [2]Computer Science & Engineering, [3]Department of Biology, [4]Department of Genetics, [5]Department of Pediatrics, Washington University in St Louis, MO, USA. [*]Equal contributions, [#]Correspondence.

Abstract: In biomedical scientific discovery, synthesizing prior knowledge from the literature is an essential component of interpreting numerical omics data analyses for disease target identification and drug discovery. Large language models (LLMs) alone can rapidly retrieve disease mechanisms from biomedical text, but text-only outputs are general and unreliable for target and drug prioritization without cohort-specific quantitative evidence. Herein, we propose a provenance-aware Text-to-Target framework that couples schema-constrained multi-model LLM retrieval with numeric omics data analysis. The key design is a modality-aware fusion step: candidates are partitioned into overlap-supported anchors, retrieval-only hidden hubs, and network-emergent novelty nodes, then propagated into staged hypothesis and strategy generation under topology constraints. We evaluate the model in Alzheimer's disease (AD) and pancreatic ductal adenocarcinoma (PDAC). In PDAC, the workflow produced a balanced 75-gene candidate universe and a 23-strategy portfolio, with significant DepMap support at both target level and strategy level. In AD, stricter candidate controls yielded a compact 34-gene universe and 14 strategies; under an expanded CRISPRbrain registry, both target-level axes were significant, with strong strategy-level enrichment. Across both diseases, final strategies preserved full provenance closure to the candidate pool, enabling end-to-end auditability from retrieval 1 2 Omics and LLM for Disease Target and Drug Discovery artifacts to validation outputs. These results support a transferable discovery architecture in which omics evidence constrains biological activity, LLM retrieval expands mechanistic search space, and network-aware fusion preserves interpretability. The framework provides a reproducible basis for dual-disease target prioritization and motivates continuous literature-mechanism concordance with agentic evidence-refresh loops.

Keywords: Large Language Models (LLMs), Single Cell, Omics, Precision Medicine, Network Biology, Target Discovery, Drug Discovery. Omics and LLM for Disease Target and Drug Discovery

---

Poster P32

Title: Radiomic Analysis of Organ-specific Response in Phase Ib/II Trial of p38 Inhibitor Pexmetinib Plus Nivolumab Following PD-1 Failure in Advanced Solid Tumors

Author list: Caroline Tyndall[1,2,3], Andrew S. Poklepovic[4,5], Sinef Huvaj-Aksoy[1,6], Mohammadreza Amjadzadeh[1,6], Hyun Lee[4,5], Lorenzo Sellitto[1], Riley Santiago[1,2,3], Nils Wildberg[1,2,3], Mark Jelinek[1,7], Carl Kim[1,2,3], Sarah Newman[1], Christopher Deitrick[1,2,8], Brian Isett[1,2,8], Qiangqiang Gu[9], Julie Urban[1], Amy Rose[1], Diwakar Davar[1,2], Yana Najjar[1,2], Liza C. Villaruz[1,2], Lazar Vujanović[1,10], Rivka R. Colen[1,6], Robert L. Ferris[11], Dan P. Zandberg[1,2,3], Riyue Bao[1,2,3,*]

Detailed Affiliations:

[1]Hillman Cancer Center, UPMC, Pittsburgh, PA, USA, [2]Department of Medicine, University of Pittsburgh, Pittsburgh, PA, USA, [3]Division of Malignant Hematology and Medical Oncology, University of Pittsburgh, Pittsburgh, PA, USA, [4]Departments of Massey Cancer Center, Virginia Commonwealth University, Richmond, VA, USA, [5]Departments of Internal Medicine, Virginia Commonwealth University, Richmond, VA, USA, [6]Department of Radiology, University of Pittsburgh, Pittsburgh, PA, USA, [7]Biostatistics Core, UPMC Hillman Cancer Center, Pittsburgh, PA, USA, [8]Cancer Bioinformatics Core, UPMC Hillman Cancer Center, Pittsburgh, PA, USA, [9]Department of Pathology, University of Pittsburgh, Pittsburgh, PA, USA, [10]Department of Otolaryngology-Head & Neck Surgery, University of Pittsburgh, Pittsburgh, PA, USA, [11]UNC Lineberger Comprehensive Cancer Center, Chapel Hill, NC, USA. *Corresponding author

Abstract: Background: Predicting a patient's response to immunotherapy remains a challenge in cancer care. P38 MAPK activation has been associated with an immunologically cold tumor microenvironment. A p38 inhibitor, pexmetinib, has therefore been evaluated in combination with PD-1/PD-L1 checkpoint blockade therapy in a phase Ib/II clinical trial for patients with advanced solid tumors who progressed on PD-1/PD-L1 monotherapy. Radiomic

analysis of these patients' tumors provides a noninvasive method of characterizing lesions at baseline or early in treatment, with the aim of predicting response on this combination of therapy. Methods: Baseline and on-treatment CT scans from the 29 radiographically evaluable patients enrolled in the phase II trial were analyzed. 93 lesions, primarily from the lung (n=31) and lymph nodes (n=23), were identified and manually delineated using Eclipse. 46 radiomic features were extracted from segmented lesions and the annular region surrounding the lesion (lesion ring; 4mm in width) using the LIFEx software. Delta values for each feature were calculated as the difference between the feature value at final follow-up and baseline. Baseline features were scaled to a range of 0 to 1 and visualized along with delta values using organ specific 2D heatmaps organized by disease control status, defined by partial response or stable disease versus progressive disease classified by RECIST1.1. Statistical comparisons on features of interest were completed with a Wilcoxon rank sum test. Results: For lung and lymph node lesions and lesion rings, there was no statistical difference in baseline features between lesions from patients with controlled disease and progressive disease. However, lymph node lesions exhibited treatment-induced changes that were not present in lung lesions. Disease controlled lesions at lymph node sites (n=8) exhibited greater absolute delta values for gray-level run length matrix (GLRLM; p=0.025-0.037) and gray level zone length matrix (GLZLM; p=0.005-0.020) features compared to progressive disease lesions (n=13). Lymph node lesion rings exhibited a similar pattern with greater absolute delta values in controlled disease lesions for one GLRLM feature (p=0.006) and multiple GLZLM features (p=0.01-0.045). Conclusions: Lymph node lesions and lesion rings exhibited greater treatment related changes in patients with controlled disease than those with progressive disease, suggesting that radiomic texture patterns could be used as markers for treatment response in patients on combined p38 inhibition and PD-1/PD-L1 checkpoint blockade therapy. Future directions include the development of organ-specific machine-learning models to identify early responders and expand to larger phase III/IV trials.

Keywords: Radiomics, immunotherapy, PD-1/PD-L1, p38 MAPK, Organ-specific

---

Poster P33

Title: A Spatial Framework for Defining the Tumor-Lung Interface Reveals Distinct Gene Signatures in Metastatic Breast Cancer

Author list: Hsuan-Yu Hung[1], Yu-Ting Kang[1], Tzu-Hung Hsiao[1], Eric Y Chuang[1] and Yidong Chen[1]

Detailed Affiliations:
[1]Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei 106319, Taiwan.

Abstract: Background: Tumor invasion is a spatially heterogeneous process that occurs at specific regions along the tumor boundary. Identifying these invasion-associated regions is important for understanding metastatic progression and discovering molecular features linked to tumor dissemination. However, spatial regions of interest (ROIs) are often defined by manual pathological annotation, which may be subjective and difficult to reproduce. Methods: We developed a spatial framework to define the tumor–lung interface and identify invasion-associated tumor regions in a lung metastatic triple-negative breast cancer specimen. After preprocessing and cell-type annotation, tumor cells and lung epithelial cells were identified using canonical marker genes. Spatial proximity between tumor cells and lung epithelial cells was used to define the invasive front. A tumor core region was subsequently identified as the compact internal tumor compartment located far from the invasive front. Within the invasive front, tumor-core distance was used to identify the outermost tumor-cell population, defined as the leading edge (LE), while the remaining invasive-front cells were classified as the invasion interface (II). Tumor cells located between the invasive front and tumor core were classified as intermediate regions. Results: The proposed framework partitioned tumor cells into spatially distinct regions, including tumor core, intermediate region, invasion interface, and leading edge. Comparative transcriptomic analysis revealed substantial spatial variation in gene expression across these regions. LE cells displayed distinct molecular characteristics compared with other tumor compartments and showed enrichment of invasionassociated gene sets, including epithelial–mesenchymal transition and TNFA signaling via NF-κB. Differential expression analysis identified 50 LE-specific genes that were consistently enriched at the tumor boundary, including FOSB, JUNB, JUN, EGR1, and DUSP1. Among these, 26 genes were further supported by gene set enrichment analysis, suggesting their potential roles in tumor invasion and metastatic progression. Spatial mapping confirmed that these gene signatures were preferentially localized at the tumor–lung interface. Conclusions: This study presents a reproducible spatial framework for defining the tumor–lung interface and characterizing invasion-associated tumor regions. By integrating spatial organization with gene-expression

profiling, the framework identifies distinct gene signatures associated with the tumor boundary and provides insights into the molecular features underlying metastatic breast cancer invasion.

Keywords: Proximity-Anchored ROI , Spatial transcriptomics, Triple-negative breast cancer (TNBC),Tumor heterogeneity, Spatial trajectory, Leading edge

---

Poster P34

Title: Cell-Type Specific APA Changes are Associated with Gene Dysregulation and Predict Survival Outcome in Head and Neck Cancer

Author list: Amra M. Vranjkovina[1], Michael J. Buck[1,2], and Jonathan E. Bard[1]

Detailed Affiliations:
[1]Department of Biochemistry, Jacobs School of Medicine and Biomedical Sciences, State University of New York at Buffalo, Buffalo, NY, 14203, United States, [2]Department of Biomedical Informatics, Jacobs School of Medicine and Biomedical Sciences, State University of New York at Buffalo, Buffalo, NY, 14203, United States.

Abstract: Alternative polyadenylation (APA) occurs in over 70% of human genes and results in the formation of multiple mRNA isoforms with varying 3' untranslated region (UTR) length. While single-cell transcriptomic techniques have improved over the years, cell-type specific alternative polyadenylation in head and neck squamous cell carcinoma (HNSCC) remains poorly characterized. We utilize a computational workflow to investigate alterations in 3'UTR length in single-cell RNA sequencing (scRNA-seq) HNSCC data, aiming to identify APA-mediated gene dysregulation and potential oncogenic advantages associated with the inclusion or exclusion of specific miRNA-binding sites. Our analysis has identified distinct populations of genes that undergo 3'UTR shortening and lengthening in cancer epithelial cells compared with normal epithelial cells. To assess the clinical relevance of these findings, we extended this to TCGA HNSC bulk RNA-seq data and computed per-gene 3'UTR length changes relative to GTEx minor salivary gland as a normal reference. Lasso-penalized Cox modeling identified APA events associated with survival, with significant interaction by molecular subtype. We hypothesize that cell-type specific 3'UTR APA changes lead to an oncogenic advantage by modulating miRNA binding sites and result in gene dysregulation in HNSCC. These findings establish a clinically relevant APA regulatory landscape in HNSCC linking tumor specific 3'UTR changes to dysregulated gene expression, and poor patient survival.

Poster P35

Title: CLARA: A Containerized Lightweight Automated Reproducible Analysis for 16S rRNA Amplicon Sequencing

Author list: Andres Benavides Arevalo[1]

Detailed Affiliations:
[1]Escuela Ingeniería de Sistemas e Informática, Universidad Industrial de Santander, Bucaramanga, Colombia.

Abstract: Amplicon sequencing of the 16S rRNA gene has become a widely adopted approach for characterizing bacterial communities due to its cost-effectiveness, scalability, and reliability. 16S amplicon sequencing analysis involves the selection, installation, configuration, and execution of multiple bioinformatics tools. In addition, there are different sequencing technologies capable of generating either short or long reads. These factors create significant challenges for researchers without advanced computational expertise in preparing suitable analysis tools for their amplicon experiments. This study presents CLARA: Containerized Lightweight Automated Reproducible Analysis for 16S rRNA Amplicon Sequencing — a framework designed to automate amplicon data processing and simplify analysis. CLARA utilizes container abstraction to package popular amplicon analysis tools as images, which are then automatically executed by a workflow management system. Users interact with the system via configuration files, specifying the sequencing technology, the target dataset, the reference database, and the amplicon pipeline. The results are displayed as interactive tables and charts within a web interface. CLARA's robustness and versatility were demonstrated by processing datasets from Illumina, PacBio, and Oxford Nanopore platforms using popular tools such as QIIME 2, EMU, and LotuS2, among others. Across all tested datasets, CLARA demonstrated the ability to automatically deploy and manage multiple tools while significantly reducing hands-on time for installation and execution procedures. Overall, CLARA lowers the computational barrier for amplicon sequencing analysis, making reproducible microbiome workflows accessible to researchers across technical backgrounds. Nevertheless, containerization may introduce performance overhead and reduce

observability, making it difficult to diagnose pipeline-specific failures at runtime. Furthermore, analytical capabilities and taxonomic resolution are inherently constrained by the completeness and scope of the reference databases employed.

Poster P36

Title: mRehab 2.0: A Voice-Guided, Intelligent Assistant for Home Stroke Rehabilitation

Author list: Emily Liu[1] and Henry Yang[1]

Detailed Affiliations:

[1]The Bishop Strachan School, Toronto, ON, Canada.

Abstract: Background: Over 7 million stroke survivors live in the United States, and insurance often ends supervised therapy long before recovery is complete. mRehab, a portable system embedding a smartphone in 3D-printed everyday objects to guide and measure daily-living exercises, has an established evidence base: consistent measurement, feasible six-week unsupervised home use with improved clinical outcomes, and acceptable usability. However, extended home use also documented concrete barriers — a crash-prone history screen (9/11 users), opaque numeric feedback (5/11), lost paper manuals requiring phone support, and fragile printed parts driving home visits. Methods: Guided by this published user feedback, we upgraded the system toward a more user-friendly, humancentric, and intelligent design: a cross-platform re-implementation; in-app demonstration videos with voice-over guidance; redesigned activity-detection algorithms (filtered sensing, gyroscope-based twist detection, dynamic-target pouring); normalized 0–100 performance scores; a rebuilt history/calendar view; a workout scheduler with reminders; and refined object fit. To verify the upgrade, we conducted a structured task-scenario usability evaluation covering ten end-to-end scenarios — setup, device insertion, activity initiation and execution, session controls, restart, error recovery, history review, data persistence, and scheduling — recording completion, errors, efficiency, and observation notes, which were coded into findings. Results: All ten scenarios were completable; four passed with zero issues, including the redesigned history, data persistence, and scheduling. Coding yielded 17 findings (8 positive, 9 mostly minor issues). All four legacy pain points assessable in this evaluation were resolved (4/4): the rebuilt history was navigated and interpreted without error, demonstration videos were rated easy to follow, weekly performance was correctly interpreted, and reminders were set in ~45 seconds. Remaining issues concentrated on screen occlusion when the phone sits inside objects (3 findings) and edge-case activity termination. Conclusions: The feedback-driven upgrade was validated at the task level, closing the loop from documented user barriers to verified design responses. The residual occlusion issues motivate the next upgrade layer: voice-first conversational-AI coaching that reduces reliance on an occluded touchscreen. Limitations: a small usability walkthrough, task-level rather than clinical validation, and three legacy issues not assessed.

Keywords: stroke rehabilitation; mHealth; usability evaluation; human-centered design; intelligent health systems; conversational AI

Poster P37

Title: Multimodal Protein Function Inference via Vision-Language Models and Structural Similarity Representations

Author list: Christopher Barker[1], Matthew Hou[1], Boen Liu[1], Brandon Ma[1], Xiao Chen[2], Jie Hou[3], Dong Si[4], Renzhi Cao[1*]

Detailed Affiliations:

[1]Department of Computer Science, Pacific Lutheran University, Tacoma 98447, United States, [2]Computer Science Department , Hamilton College, United States, [3]Department of Computational Mathematics, Science and Engineering, Michigan State University, United States, [4]School of Science, Technology, Engineering & Mathematics, University of Washington, USA, United States.

Abstract: Predicting protein function from sequence remains a major challenge in computational biology. In this work, we propose a multimodal pipeline that embeds protein sequences using ESM-3, converts them into pairwise similarity matrices, and visualizes these representations as 2D heatmaps. These visual representations are processed by a BLIP-2 vision-language model to generate natural language descriptions, which are then translated into functional annotations using a rule-based mapping aligned with Gene Ontology terminology and evaluated via semantic similarity scoring. Through iterative prompt engineering across six experimental batches, two critical

failure modes were identified: mode collapse and visual hallucination. To address these issues, a structurally constrained dynamic prompting architecture was developed, enforcing biologically grounded interpretation while preserving generative flexibility. The final pipeline achieves statistically significant improvement over baseline performance while maintaining functional diversity across predictions. These results demonstrate a scalable and interpretable framework for multimodal protein function inference and establish prompt design as a key control mechanism in scientific multimodal reasoning.

Keywords: protein function inference, multimodal learning, vision-language models, ESM-3, BLIP-2, Gene Ontology, prompt engineering

---

Poster P38

Title: Unveiling Signals Lost in the Noise: Single-Cell Viral Discovery from Unmapped Single-Cell RNA-seq Data in Colorectal Cancer

Author list: Mohammadamin Mahmanzar[1,2], Karim Rahimian[3], Kaveh Kavousi[2,3*], Youping Deng[1*]

Detailed Affiliations:

[1]Department of Quantitative Health Sciences, JohnA. Burns School of Medicine, University of Hawaii at Manoa, Honolulu, HI 96813, USA, [2]Department of Bioinformatics, Kish International Campus University of Tehran, Kish, Iran, [3]Laboratory of Complex Biological Systems & Bioinformatics (CBB), Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran. [*]Corresponding author

Abstract: Unmapped reads in RNA sequencing data have long been considered technical byproducts, yet growing evidence suggests they may encode biologically relevant non-host signals. However, their interpretation remains controversial due to concerns regarding contamination, low-abundance noise, and lack of cellular resolution. Here, unmapped reads from single-cell RNA sequencing (scRNA-seq) data were systematically interrogated to reconstruct the colorectal cancer (CRC) virome at cellular resolution across 570 sequencing samples from 179 patients. By bypassing conventional reference-dependent microbiome pipelines, reproducible viral signals were identified across epithelial, immune, and stromal compartments and mapped across distinct anatomical regions of the colorectal tract. Global viral burden remained comparable across healthy, tumor-adjacent, and CRC tissues, indicating that disease-associated differences arise from compositional restructuring rather than overall viral load. Viral communities in healthy tissues exhibited higher diversity and stability, whereas tumor-adjacent and CRC tissues showed reduced diversity and increased heterogeneity. Cell-type–resolved analyses demonstrated that viral signals are highly structured, with the highest burden observed in epithelial cells, while the most pronounced condition-associated remodeling occurred within immune and stromal populations. Anatomical stratification further revealed that distal regions, particularly the rectosigmoid colon, exhibit the strongest virome restructuring. Paired within-patient analyses uncovered coordinated redistribution of viral signals during tumorigenesis, characterized by reduced viral abundance in immune and stromal compartments alongside retention within epithelial cells. Machine learning models further demonstrated that viral features encode reproducible signatures capable of distinguishing tissue states, supporting their biological relevance. This study serves as a pioneering effort in defining the colorectal cancer virome at single-cell resolution using unmapped transcriptomic data, enabling characterization of cell-type–specific viral landscapes within the tumor microenvironment and across anatomical regions of the colon. Together, these findings establish that colorectal tumorigenesis is accompanied by structured, celltype– and location-specific remodeling of the local virome, providing a foundation for future mechanistic and translational investigations.

Keywords: Single-cell RNA sequencing (scRNA-seq), Colorectal cancer (CRC), Virome, Unmapped reads, Tumor microenvironment

---

Poster P39

Title: Gene networks uncover extensive trans-genetic regulation from transcriptional to proteomic levels in the human brain

Author list: Qiuman Liang[1,2] Yu Wei[2], Yizhi Wang[3], Yue Wang[3], Chao Chen[1,4*], Chunyu Liu[2*]

Detailed Affiliations:

[1]MOE Key Laboratory of Rare Pediatric Diseases & Hunan Key Laboratory of Medical Genetics, School of Life Sciences, and Department of Psychiatry, The Second Xiangya Hospital, Central South University, Changsha, Hunan,

China, [2]Department of Psychiatry, SUNY Upstate Medical University, Syracuse, NY, USA, [3]Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA, USA, [4]Furong Laboratory, Changsha, Hunan, China. *Corresponding author
Abstract: Most studies of genetic regulation have focused on cis effects, whereas the contribution of trans genetic regulation remains poorly understood. Here, we used matched transcriptomic, translatomic, and proteomic data from the human prefrontal cortex to characterize gene co-expression modules and infer direct association networks (DANs) using bootstrap graphical lasso at the transcriptional, translational, and proteomic levels, separately. Although co-expression modules were largely preserved, DANs underwent extensive rewiring across molecular levels. Leveraging DAN-connected genes as mediators in expression prediction models revealed widespread trans-genetic regulation, with trans effects detected in nearly half of genes at each level. Notably, DANs-derived trans signals explained more expression variance than cis SNPs for 17.6%, 42.4%, and 35.5% of genes at the transcriptional, translational, and proteomic levels, respectively. Applying these models to schizophrenia GWAS identified 28, 18, 9 risk genes, including 25.0%, 77.8%, and 55.6% that were missed by conventional cis-TWAS analyses. Together, these findings reveal pervasive rewiring of gene DANs across molecular levels and demonstrate that network-mediated trans-genetic regulation represents a major and previously underappreciated contributor to schizophrenia risk, particularly at the translational and proteomic levels.

Poster P40

Title: A Comparison between CARLIN and DNA Typewriter in CRISPR-mediated Lineage Tracing

Author list: Fengshuo Liu[1], Xiang Zhang[2], Yipeng Yang[3]

Detailed Affiliations:
[1]Lester and Sue Smith Breast Center, Baylor College of Medicine, One Baylor Plaza, Houston, 77030, TX, USA, [2]Lester and Sue Smith Breast Center, Baylor College of Medicine, One Baylor Plaza, Houston, 77030, TX, USA, [3]Department of Mathematics and Statistics, University of Houston - Clear Lake, 2700 Bay Area Blvd, 77058, TX, USA.

Abstract: Background CARLIN and DNA Typewriter are two major breakthroughs in CRISPR-based lineage tracing technology. It is essential to understand the potential and performance of these methods in lineage tracing, which provides important guidance on experimental design. Results In this study, we systematically compare these two strategies using a unified stochastic simulation framework with known ground-truth lineages. By explicitly modeling CRISPR editing dynamics, barcode evolution, and cell division processes, the framework enables quantitative benchmarking of lineage reconstruction accuracy across diverse experimental parameter regimes. Both methods are evaluated using multiple accuracy metrics, including Robinson–Foulds accuracy and triplet accuracy, allowing a comprehensive assessment of lineage reconstruction performance under various editing probabilities, sampling depths, and lineage lengths. Conclusions DNA Typewriter consistently outperforms CARLIN in lineage reconstruction accuracy when sufficient numbers of recording targets are used, particularly in more cell divisions. Sequential and ordered recording in DNA Typewriter substantially reduces ambiguity in lineage inference compared to unordered CRISPR barcode editing. CARLIN's lineage-recording potential exhausts rapidly under continuous induction, limiting its effectiveness in long-term lineage tracing. Triplet accuracy provides a more permissive and informative metric than Robinson–Foulds accuracy, especially under partial sampling scenarios.

Keywords: Lineage Tracing, CARLIN Model, DNA Typewriter

Poster P41

Title: Cross-Species Gene Set and Cell Type Annotation Through Large Language Models and Agentic AI with Literature-Grounded Validation

Author list: Wenbo Chen[1], Rongbin Li[1], Neha Maurya[1], Arnav Solanki[1], Meaghan Ramlakhan[1], Zhuhao Wu[2], W. Jim Zheng[1,#]

Detailed Affiliations: [1]McWilliams School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, USA, [2]Appel Alzheimer's Disease Research Institute, Feil Family Brain and Mind Research Institute, Weill Cornell Medicine, New York, NY, USA. [#]Corresponding author

Abstract:

Background

Single-cell RNA sequencing (scRNA-seq) has transformed our ability to identify diverse cell types and their transcriptomic signatures. Despite rapid progress, annotating newly discovered or poorly characterized cell clusters remains a major challenge in biomedical research. Traditional approaches such as Gene Set Enrichment Analysis (GSEA) rely on predefined gene sets that are often incomplete for novel biological contexts. Although large language models (LLMs) leverage vast biomedical literature, their application to cell type annotation is limited by insufficient grounding in structured biological ontologies. Therefore, there is a critical need for an advanced computational framework that can generate reliable and biologically meaningful cell type annotations from scRNA-seq data across species and contexts.

Methods

We leveraged fine-tuned LLMs to develop an agentic AI system to perform cross-species annotation for cell types. We first extracted gene modules from scRNA-seq datasets, which group genes into co-expression modules based on cell-cell similarity. We then developed a multi-agent AI framework by fine-tuning LLMs including Gemma and Llama architectures on curated human and mouse gene sets from MSigDB, separately. Retrieval-augmented generation (RAG) was incorporated to ground predictions in peer-reviewed literature from PubMed, reducing hallucinations and improving interpretability. The framework takes gene lists derived from cell clusters as input and generates literature-supported functional annotations. Performance was evaluated using accuracy and semantic similarity metrics, and domain experts conducted human evaluation to assess biological relevance.

Results

Task-specific fine-tuned models consistently outperformed models without fine-tuning, demonstrating that targeted training on structured biological knowledge is essential for accurate gene set annotation. Across benchmark gene sets, 77% of mouse and 74% of human gene sets achieved correct annotations among the top predictions, outperforming existing LLM-based approaches. Retrieval grounding further improved prediction reliability in both species. As a key application, we applied the framework to large-scale brain atlases with 5,322 mouse and 3,313 human whole-brain scRNA-seq clusters and generated biologically meaningful cell type descriptions that captured both conserved and species-specific transcriptional programs. Together, this end-to-end pipeline bridges the gap from raw scRNA-seq data to meaningful biological annotations across species.

Conclusion

By combining species-specific fine-tuning with retrieval-based validation, our framework enables accurate and literature-grounded gene set and cell type annotation at scale. This approach enables us to comprehensively evaluate marker gene sets and cell types for a broad biomedical domain where accurate functional annotation is critically important for the mechanistic understanding of disease and therapeutic development.

Keywords: Cell type annotation, Large Language Models, Single-cell RNA sequencing, Retrieval-augmented generation, Multi-agent AI system

---

Poster P42

Title: Multimodal Deep Learning for Alzheimer's Disease Prediction with Improved Training Stability and Flexibility

Author list: Jianan Liu[1,2], Jincheng Shi[3], Long Zhao[1,2], Yuanxin Yao[3], Shihua Li[1,2], Rongjie Liu[4]

Detailed Affiliations:

[1]School of Automation, Southeast University, Nanjing, 210096, P. R. China, [2]Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Nanjing, 210096, P. R. China, [3]Department of Statistics, Florida State University, Tallahassee, FL, 32304, USA, [4]Department of Statistics, University of Georgia, Athens, GA, 30602, USA.

Abstract: Alzheimer's disease (AD) is a progressive neurodegenerative disorder that requires early and accurate diagnosis, particularly at the mild cognitive impairment (MCI) stage. The increasing availability of multimodal datasets has enabled the development of deep learning models that integrate heterogeneous data sources to improve diagnostic performance. However, existing approaches often suffer from unstable training dynamics, including gradient vanishing and explosion, which may reduce model robustness and generalization in clinical applications. This study proposes a multimodal deep learning framework for AD prediction using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI). The framework integrates structural MRI, PET imaging, and non-imaging clinical variables to capture complementary disease-related information. A tunable nonlinear transformation

mechanism is introduced to enhance the adaptability of model representations while maintaining stable gradient propagation during training. Theoretical analysis is provided to characterize its key properties, including nonlinearity, continuity, differentiability, and boundedness, and to demonstrate its capability to approximate commonly used activation functions. Experimental results show that the proposed framework achieves improved diagnostic performance compared with baseline models, especially in the clinically important task of MCI identification. Ablation analysis further reveals that different modalities contribute unequally to AD classification. Clinical features provide the dominant predictive signal, while neuroimaging features offer complementary disease-related information. In contrast, genetic or static features contribute only modestly, suggesting that they are more suitable as contextual risk indicators than as strong standalone diagnostic markers. These findings indicate that multimodal integration should be interpreted as a mechanism for supplementing dominant clinical information rather than as evidence that all modalities contribute equally. The proposed nonlinear transformation mechanism also improves training stability and generalization. It achieves better validation performance, suggesting an implicit regularization effect and reduced overfitting. This is particularly meaningful for MCI classification, where diagnostic boundaries are subtle and model robustness is critical. Overall, this work highlights the importance of stable nonlinear transformation design in multimodal AD prediction and provides a flexible framework for early diagnosis. Future work will extend the current cross-sectional classification setting to longitudinal prediction tasks, such as estimating MCI-to-AD conversion risk over time.

Keywords: Alzheimer's disease, multimodal transformer, activation function, disease prediction, training stability, nonlinear transformation

---

Poster P43

Title: Long-read methylome profiling in the human pangenome reveals ancestry-associated methylation states and genetic-variant-coupled regulatory effects

Author list: Tianjie Liu[1], Juan F. Macias-Velasco[1], Xiaoyu Zhuo[1], Wenjin Zhang[1], Daofeng Li[1], Juan Jiang[1], Zheng Dong[1], Chad Tomlinson[1], Eddie Belter[1] and Ting Wang[1]

Detailed Affiliations:

[1]Department of Genetics, Washington University School of Medicine, Saint Louis, United States.

Abstract: DNA methylation is fundamental to chromatin architecture, imprinting, transcriptional regulation, and disease, yet short-read approaches incompletely resolve complex loci and non-reference sequences. Here, we combine long-read methylation profiling with the human pangenome graph to define population-scale methylation diversity, CpG island polymorphism, and the local and cis-regulatory coupling between methylation and genetic variation. Using a graph-projected harmonization framework anchored to the telomere-to-telomere CHM13 backbone, we enable direct comparison of methylation across 462 haplotypes, 5 populations, and structurally variable sequences, including previously inaccessible non-reference insertions. Graph-based CpG genotyping captures more than 40 million nonredundant CpGs and supports cross-haplotype analyses within variant-bearing regions. Across the genome, we identify ~5 Mb of highly variable methylated regions (HVMRs), enriched in distal regulatory elements and CpG island shores. Methylation variability shows a weak overall negative correlation with nucleotide and synteny diversity, with a particularly strong negative relationship in SVA elements. At a small subset of loci with extreme synteny diversity, however, methylation no longer conforms to canonical sequence-embedded patterns, instead, methylation states extend from variant sequence into adjacent flanks, reshaping local regulatory landscapes. We further construct a nonredundant pangenome-wide CpG island set and define its population polymorphism and saturation dynamics, providing a systematic view from CpG island sequence variation to methylation divergence. PCoA of HVMR methylation resolves ancestry-associated methylome states, with African-ancestry haplotypes showing the clearest separation from non-African haplotypes. Graph-based differential methylation analyses further reveal lower methylation in African-ancestry haplotypes at LTR-rich repeats and identify ancestry-specific differentially methylated regions. Finally, by integrating long-read Iso-Seq-defined novel transcripts, we perform haplotype-resolved, allele-specific xQTL analyses that link genetic variation to methylation, isoform expression, and alternative splicing. Together, these results establish a pangenome-enabled framework for dissecting how structural variation shapes human methylation and transcription, and reveal ancestry-associated epigenomic modules with broad implications for human biology and evolution.

Keywords: Human pangenome, DNA methylation, CpG island, mQTL, isoQTL, sQTL

Poster P44

Title: Investigating the Causal Impact of COVID-19 on Acute Myocardial Infarction: A Multi-ancestry Mendelian Randomization Study

Author list: Katherine Zhou[1] and Ying Ma[2]

Detailed Affiliations:
[1]Choate Rosemary Hall (High School), Wallingford, CT, USA, [2]Department of Biostatistics, School of Public Health, Brown University, Providence, RI, USA.

Abstract: Observational studies suggest an association between COVID-19 and acute myocardial infarction (AMI). However, whether this relationship is truly causal across different ancestries remains unclear. Here, we performed multi-ancestry Mendelian randomization (MR) analysis across European, South Asian, East Asian, and African populations using three COVID-19 phenotypes: very severe respiratory confirmed COVID-19, hospitalized COVID-19, and general infection. GWAS summary statistics were obtained from the COVID-19 Host Genetics Initiative Round 7 (16,171 very severe, 39,612 hospitalized, 144,834 infected cases; 3,314,635 controls) and UK Biobank/Biobank Japan for AMI outcomes (39,051 cases, 599,423 controls). In European ancestry, MR analysis revealed a significant inverse association between very severe respiratory confirmed COVID-19 and AMI risk, with no evidence of directional horizontal pleiotropy. This finding persisted despite significant heterogeneity. Sensitivity analyses using MR-Egger, weighted median, and weighted mode methods yielded effect estimates with the same negative sign but did not reach significance. A similar negative association was observed for hospitalized COVID-19 on AMI risk without detectable pleiotropy, while sensitivity analyses using the other methods yielded effect estimates with the same sign but did not reach significance. In South Asian ancestry, hospitalized COVID-19 was significantly associated with reduced AMI risk, with no observable heterogeneity or pleiotropy, and with no significance in sensitivity analyses. No significant causal associations were identified for any COVID-19 phenotype in East Asian or African ancestries, with no evidence of excess heterogeneity or horizontal pleiotropy. These findings highlight the complexity of COVID-19 and AMI association and shed light on potential strategies to promote post-COVID cardiovascular health.

Keywords: Mendelian randomization, COVID-19, Cardiovascular system, Multi-ancestry, Acute Myocardial Infarction

---

Poster P45

Title: Computational Pathway Analysis of Biologic Drug Innovation in Alzheimer's Disease: A Structural Equation Modeling Framework for Care Delivery Transformation

Author list: Sang-Yong Oh,[1] Janice Tran,[1] Jacqueline S Markle[1] and Satish Mahadevan Srinivasan[1,*]

Detailed Affiliations:
[1]School of Professional Graduate Studies at Great Valley, The Pennsylvania State University, 30 East Swedesford Rd, Malvern, 19355, PA, USA. [*]Corresponding author

Abstract: Biologic therapeutics targeting amyloid-beta and tau pathways represent a paradigm-level intervention in Alzheimer's disease, yet their systemic effects on care delivery infrastructure remain poorly characterized. We present a computational pathway analysis framework, implemented as a Python-based structural equation modeling (SEM) pipeline, to model the causal architecture through which biologic innovation reshapes healthcare delivery patterns. Treating the multiequation SEM system as a directed acyclic graph (DAG), we integrate prescription claims data, physician specialty records, and drug characteristic annotations across 475 medication-market observations spanning January 2023 through September 2024. Feature engineering procedures include log-transformation of utilization volumes and binary encoding of biologic status, producing a biologically interpretable variable set aligned with computational pharmacology conventions. The pipeline estimates four interconnected path equations representing sequential stages of care transformation. Our core findings demonstrate that biologic drug status reduces neurology care intensity by 13.9 percentage points ($\beta = -0.139$, $p < 0.001$), while simultaneously increasing total patient utilization by approximately 42-fold ($\beta = 3.747$, $p < 0.001$). Mediation analysis within the DAG framework reveals that 71% of the biologic effect on specialist penetration is transmitted through an indirect pathway via reduced neurology intensity (ab = −0.133; total c = −0.187), with the care-to-specialist path achieving excellent model fit (R2 = 0.815). These computational findings demonstrate a "democratization" effect in which biologic innovation restructures the care delivery network by enabling broader primary care engagement. The resulting pathway model constitutes a reusable health informatics asset applicable to clinical decision support

system design, predictive modeling of market transformation, and translational research on precision medicine adoption. All data and code are publicly available, supporting reproducible bioinformatics practice.

Keywords: Structural Equation Modeling, causal pathway modeling, directed acyclic graph, biologic therapeutics, Alzheimer's disease, clinical decision support, health informatics, computational pharmacology, care delivery transformation, machine learning for clinical applications, EHR-based phenotyping, precision medicine

---

Poster P46

Title: Dissecting genetic control of human neuronal DNA methylation states refines interpretation of brain disorder risk

Author list: Yu Wei[1], Qiuman Liang[2], Jo-Fan Chien[3], Heng Xu[4], Junhao Li[5], Alexey Kozlenkov[6,7], Ramu Vadukapuram[6,7], Eran A. Mukamel[5], Stella Dracheva[6,7], Chunyu Liu[1]

Detailed Affiliations:
[1]Department of Psychiatry, SUNY Upstate Medical University, Syracuse, NY, USA, [2]Center for Medical Genetics & Hunan Key Laboratory of Medical Genetics, School of Life Sciences, Central South University, Changsha, Hunan, China, [3]Department of Physics, University of California, San Diego, La Jolla, CA, USA, [4]Bioinformatics and Systems Biology, University of California, San Diego, La Jolla, CA, USA, [5]Department of Cognitive Science, University of California, San Diego, La Jolla, CA, USA, [6]Friedman Brain Institute and Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [7]Research & Development and VISN2 MIREC, James J, Peters VA Medical Center, Bronx, NY, USA.

Abstract: DNA methylation in the human brain primarily exists as 5-methylcytosine (5mC) and its oxidized form, 5-hydroxymethylcytosine (5hmC), both of which are stable epigenetic marks linked to gene regulation and brain disorder risk. However, current DNA methylation quantitative trait loci (mQTL) studies have largely relied on bulk brain tissue and aggregate methylation signals (the sum of 5mC and 5hmC), limiting resolution across neuronal subtypes and methylation states. To address this gap, we generated cell-type-resolved multi-omics data from glutamatergic (GLU) and GABAergic (GABA) neurons isolated from the human dorsolateral prefrontal cortex (DLPFC) using fluorescence-activated nuclei sorting (FANS). We mapped mQTLs for three distinct CpG methylation states (hmCG, True mCG, and Total mCG), corresponding to 5hmC, 5mC, and their combined signal, respectively, together with expression quantitative trait loci (eQTLs). We integrated these xQTL datasets with brain disorder GWAS data through colocalization analyses and used Mendelian randomization (MR) to investigate directional relationships between DNA methylation and gene expression at disease-associated loci. We observed broadly concordant xQTL signals between neuronal subtypes (π_1 = 0.84–0.98). Although Total mQTLs captured most significant genetic associations identified by hmQTLs and True mQTLs, they obscured methylation-state-specific regulatory patterns. Integrative colocalization and MR analyses demonstrated that increased neuronal subtype and DNA methylation-state resolution refined the functional interpretation of brain disorder GWAS loci across schizophrenia, bipolar disorder, major depressive disorder, and Alzheimer's disease. Directional analyses revealed diverse relationships between DNA methylation and gene expression, including bidirectional effects, at disease-associated loci. Together, these findings highlight the value of resolving both neuronal subtypes and DNA methylation states for understanding the regulatory mechanisms underlying brain disorder risk.

Poster P47

Title: From Pixels to Biological Reasoning: Semantic, Hierarchical, and Graph-Structured Learning for Interpretable Biomedical AI

Author list: Chengzhi Cao[1], Qirong Ho[1], Min Xu[1,2]

Detailed Affiliations:
[1]Mohamed bin Zayed University of Artificial Intelligence, UAE, [2]Carnegie Mellon University, PA, USA.

Abstract: Biomedical images contain rich information across multiple levels, from fine-grained visual patterns to organ-level and system-level relationships. However, conventional deep learning methods often process these data as isolated pixels, regions, or global embeddings, limiting their ability to preserve biological structure and provide clinically meaningful explanations. This poster presents a unified research direction for structure-aware biomedical intelligence, drawing on three complementary works. First, semantic-aware knowledge learning enhances cryo-EM micrograph denoising by using foundation-model-derived semantic priors to distinguish particle regions from noisy

non-semantic regions, combining cross-local relationship modeling with spatial-frequency complementary learning for improved structural preservation. Second, intrinsic hierarchical organization is explored for explainable medical diagnosis by modeling biological data across multiple scales, such as organs, systems, and the whole body, through hypergraph-based intra-level and inter-level reasoning. Third, MedGraphGen generates interpretable medical knowledge graphs directly from CT images, representing organs as nodes and anatomical relationships as edges, with relation-aware hypergraph reasoning to capture higher-order multi-organ interactions. Together, these works suggest a broader framework for biomedical AI that moves beyond visual recognition toward semantic perception, hierarchical organization, and relational reasoning. By embedding biological priors into neural architectures, the proposed direction improves both performance and interpretability, supporting downstream tasks such as denoising, diagnosis, medical report generation, visual question answering, and anatomical knowledge graph construction.

Poster P48

Title: Decoding Solid Tumor Prognosis and Immunotherapy Response Through NK Cell Gene Signatures

Author list: Emily Tang[1], Xiaoqi Li[2]

Detailed Affiliations:

[1]North Carolina School of Science and Mathematics, Durham, NC, USA, [2]Eli Lilly and Company, Indianapolis, IN, USA

Abstract: Introduction

Chimeric antigen receptor natural killer (CAR-NK) therapy is a promising immunotherapeutic approach, but its efficacy in solid tumors is limited by the metabolic and physical barriers of the tumor microenvironment (TME). Understanding NK cell-TME interactions and identifying biomarkers that predict prognosis and immunotherapy response are critical for developing precision CAR-NK therapies. This study evaluated three functional gene groups, CX3CR1 and CXCR3 (trafficking receptors), ZEB2 and SPON2 (maturation regulators), and PRF1 and GZMB (cytolytic effectors), in Lower-Grade Glioma (LGG), Skin Cutaneous Melanoma (SKCM), and Lung Adenocarcinoma (LUAD), using FCGR3A as a marker of mature cytotoxic NK cells.

Methods

Gene expression and clinical data were obtained from The Cancer Genome Atlas (TCGA) through TIMER3 platform. Correlations between functional genes and FCGR3A were analyzed using partial Spearman correlation and highly functional genes identified were verified in NK linage by deconvolution algorithms. Prognostic significance was assessed using Cox proportional hazards models adjusted for age, tumor purity, and/or stage to evaluate overall survival (OS) and progression-free interval (PFI). Predictive biomarkers for immunotherapy response were evaluated using receiver operating characteristic (ROC) analysis across selected immunotherapy cohorts.

Results

Computational analyses revealed distinct tumor-specific NK cell phenotypes. SKCM demonstrated high infiltration of mature, functional NK cells (CX3CR1 $\rho = 0.429$; ZEB2 $\rho = 0.421$), which was associated with significantly reduced risks of mortality and disease progression ($p=0.006$ and $p<0.001$ for OS and PFI, respectively). LUAD exhibited moderate infiltration of mature NK cells (CX3CR1 $\rho = 0.293$; ZEB2 $\rho = 0.621$), without significant protection for tumor progression ($p=0.879$) or long-term survival benefit ($p=0.562$). In contrast, LGG showed robust NK-cell infiltration (CX3CR1 $\rho = 0.618$) accompanied by absent maturation signaling (ZEB2 $\rho = 0$), and elevated NK homing was associated with increased risks of progression ($p=0.004$) and mortality ($p<0.001$). Cytolytic gene expression remained relatively consistent across tumor types (GZMB $\rho = 0.399–0.512$). Across all tumors, CX3CR1 emerged as a good predictor of immunotherapy response, with AUC values of 0.873, 0.643, and 0.613 for SKCM, LGG, and LUAD, respectively.

Conclusion

NK function is highly dependent on TME, supporting the need for tumor-specific immune engineering strategies. In LGG, impaired NK maturation may contribute to poor clinical outcomes despite substantial NK infiltration. Engineering CAR-NK cells to restore maturation and sustain cytotoxicity within immunosuppressive TMEs may improve therapeutic efficacy. These findings provide a framework for developing precision CAR-NK therapies tailored to immune landscapes of solid tumors.

Keywords: NK, glioma, skin cutaneous melanoma, lung adenocarcinoma, tumor microenvironment, computational biology

Poster P49

Title: Tetris: Transformative Embedding through Multilevel Contrastive learning for hypothesis discovery

Author list: Xiaoying Wang[1,2], Yuke Hou[1], Guangyu Wang[3], Ana Mora[4], Mauricio Rojas[4], Qin Ma[1,2,$]

Detailed Affiliations:

[1]Department of Biomedical Informatics, College of Medicine, Ohio State University, Columbus, OH, [2]Pelotonia Institute for Immuno-Oncology, The James Comprehensive Cancer Center, The Ohio State University, Columbus, OH, [3]Center for Bioinformatics and Computational Biology, Houston Methodist Research Institute, Houston, TX, USA, [4]Dorothy M. Davis Heart and Lung Research Institute, Division of Pulmonary, Critical Care, & Sleep Medicine, Department of Internal Medicine, School of Medicine, Ohio State University, Columbus, OH. [$]Corresponding author

Abstract: Although scRNA-seq enables large-scale characterization of cellular states, most analytical frameworks rely on a single representation of the data, limiting discovery of biological systems shaped by interacting phenotypes or lacking definitive markers. Here we present Tetris, a multilevel contrastive learning framework for biological system discovery that transforms cellular representations under phenotype constraints. We demonstrate that Tetris decodes biological systems jointly shaped by multiple phenotypes, identifies system-relevant cell populations and system-defining molecular programs across diverse biological contexts, and uncovers hidden biological systems lacking definitive markers, including a senescence-associated system inferred from aging and fibrosis phenotypes. We further show that biological systems discovered by Tetris can be transferred into pathology foundation models to generate system-aware representations for morphology-to-system inference. Finally, Tetris incorporates a multi-level evaluation framework for prioritizing experimentally actionable genes and hypotheses through convergent computational, multimodal, and knowledge-based evidence. Together, Tetris provides a framework for decoding, transferring, and validating biological systems from single-cell data.

Keywords: single-cell RNA sequencing, biological system discovery, multilevel contrastive learning, phenotype-guided representation learning, foundation models, hypothesis prioritization

---

Poster P50

Title: Ensemble Machine Learning Models For Early Sepsis Detection

Author list: Shashvat Hariharan[1], Sanjay Gobu[1], Gautham Venkatramanan[2]

Detailed Affiliations:

[1]Linganore High School, Frederick, MD, USA; [2]Oakdale High School, Frederick, MD, USA

Abstract: Sepsis is a life-threatening condition that requires early detection to improve patient safety, yet its initial symptoms are often subtle and difficult to identify. This project investigates whether an ensemble machine learning (ML) approach can improve the accuracy and timeliness of sepsis prediction. An ensemble model using python programming was developed that combines the outputs of three distinct ML algorithms - Logistic Regression, Random Forest, and XGBoost. A geometric mean of the risk probabilities from each algorithm's output was used to generate a combined risk probability. The system was trained on structured electronic medical record (EMR) datasets with anonymized personally identifiable information (PII) that complies with the Health Insurance Portability and Accountability Act (HIPAA) privacy standards. These EMR datasets were obtained from online databases with real static data from Johns Hopkins University and real time-series data from the PhysioNet/Computing in Cardiology Challenge 2019. This real data needed pre-processing through iterative imputation to handle missing data and file concatenation to combine data spread over multiple files. For training purposes, synthetic datasets were also developed using computer generated values with specified ranges for each feature. In total, the project trained the ensemble model on four different datasets.

Results show that the ensemble model achieved strong predictive performance in detecting sepsis, with high true negatives (85.6%) and true positives (13.9%), alongside very low false positives (0.2%) and false negatives (0.3%) utilizing static datasets. The model was also able to predict sepsis with a detection probability of 98%, approximately six hours before clinical onset using time-series datasets. To enhance model interpretability for physicians, SHapley Additive exPlanations (SHAP) analysis was used to identify key clinical predictors. Some of these include temperature, white blood cell count (WBC), lactate, and PaO_2/FiO_2 ratio, matching with commonly identified predictors by physicians.

A user-friendly interface was developed to display sepsis risk probability along with contributing features, supporting clinical decision-making. This interface was developed in HTML to call the sepsis prediction model that has been packaged as an app. In conclusion, this project demonstrates that the developed ensemble model can provide early, accurate and interpretable sepsis predictions. Future work will focus on using larger datasets, incorporating unstructured clinical data and improving accuracy in real-world healthcare settings.

Keywords: Sepsis prediction, machine learning, time-series data, iterative imputation, SHapley Additive exPlanations (SHAP) analysis

---

Poster P51

Title: Comparative Metagenomic Profiling of Soil Microbiota in San Antonio Urban Ecosystems: Medina River Natural Area vs. Loop 1604 Trailhead

Author list: Raymond Cai[1], Claire Wang[2], Jianhua Ruan[3]

Detailed Affiliations:
[1]Keystone School, San Antonio, TX, USA, [2]Westfield High School, Westfield, NJ, USA, [3]Department of Computer Science, The University of Texas at San Antonio, San Antonio, TX, USA.

Abstract: Soils experience different ecological stresses depending on land use, vegetation, and surrounding infrastructure. However, little is known about the composition and functional roles of soil microbial communities across San Antonio's expanding urban landscape. To address this knowledge gap, we examined how environmental conditions and human disturbance affect urban soil ecosystems. We performed a comparative metagenomic analysis between two ecological areas in South Texas: the largely undisturbed, moisture-rich Medina River Natural Area and the heavily trafficked Loop 1604 Trailhead. DNA extracted from surface soils was used for 16S rRNA gene amplicon sequencing targeting the V3-V4 region. The raw reads were processed through the DADA2 pipeline to determine Amplicon Sequence Variants (ASVs). Downstream ecological analyses, including alpha diversity, Bray-Curtis distance matrices, and PERMANOVA tests, were performed using the phyloseq and vegan R packages. We found that soil communities at the Medina River and the Loop 1604 Trailhead differed in composition, indicating differences in tree canopy cover, soil moisture, and proximity to paved infrastructure. Differential abundance analysis revealed shifts in indicator taxa. The Medina River zone was enriched by deep-soil nutrient-cycling specialists, whereas several groups associated with disturbed environments were more abundant at the Loop 1604 location. The observed differences in bacterial community composition between the two sites indicate that local environmental conditions are associated with variation in urban soil microbiomes. These assessments of soil bacterial communities from contrasting urban habitats in San Antonio have established a baseline to understand how urbanization may influence soil biodiversity and ecosystems in South Texas.

Keywords: soil, South Texas, 16s, microbiota, metagenomics

Poster P52

Title: AI-Driven Innovations in Chinese Eight-Ball Pool: Motion Analysis and Intelligent Coaching Systems

Author list: Yanguang Wei[1], Henry Yang[2]

Detailed Affiliations:
[1]Mater Dei Catholic High School1, CA, USA, [1]Sage Hill School, CA, USA

Abstract: Background
With the rapid advancement of AI, sports training has increasingly incorporated AI technologies to enhance performance, and motion analysis has become widely used in sports. In recent years, Chinese Eight-Ball Pool has gained global popularity; however, scientific research in the field of billiards remains limited, with key aspects like posture evaluation relying heavily on subjective assessments by coaches rather than objective, data-driven methods.
Objective
This study aims to bridge this gap by leveraging AI to analyze the shooting motions of Chinese Eight-Ball Pool players and provide actionable feedback to improve shot accuracy, motion consistency, and overall training efficiency.
Methods

This designed coaching system integrates two core AI-powered functionalities: Motion Optimization and Personalized Shot Route Guidance. Firstly, the project combines the VICON motion capture system with computer vision techniques to create a robust motion analysis framework. VICON system, acting as the ground-truth measurement tool, utilizes reflective markers to capture three-dimensional body joint coordinates and angles with millimeter-level precision. Simultaneously, RGB cameras record the player's shooting actions, with computer vision algorithms detecting and extracting key body points. The system analyzes these biomechanical data with metrics such as head stability, elbow angle, cue stroke trajectory, and body balance—elements known to influence player performance. Second, the AI system analyzes the table layout by detecting cue balls, object balls, and pockets. It uses algorithms to calculate optimal shot routes, accounting for factors such as shot angle, collision path prediction, success probability, and positional advantage for subsequent shots. For beginner players, the system recommends simpler, safer routes with higher success rates, avoiding complex wide-angle or spin-based shots.

Results

The effectiveness of the AI-powered coaching system is evaluated through experiments comparing traditional training methods and AI-assisted training. Experimental data indicate that beginners using the AI-assisted system demonstrate significantly enhanced potting success rates and better posture after adopting the system's feedback. In terms of shot route guidance, beginners achieved higher shot success rates as the AI prioritized safer, simpler routes. Players reported increased confidence and strategic understanding, transitioning from novice-level decisions to intermediate levels of play with improved accuracy.

Conclusion

These results demonstrate that the system not only enhances immediate performance but also accelerates the training process through real-time, personalized coaching. By integrating motion analysis with trajectory insights, this innovative approach elevates player development, enabling skill refinement through intelligent, data-driven feedback. This validates its potential as a transformative tool for advancing Chinese Eight-Ball Pool training.

Keywords: Biomechanics Analysis, Pose Estimation, Personalized Coaching, Computer Vision, Artificial Intelligence, Chinese Eight-Ball Pool

---

Poster P53

Title: Assessing the Homogeneity of Urban Park Soil Microbiomes: A Comparative Metagenomic Analysis of McAllister and Eisenhower Parks in San Antonio

Author list: Chloe Peiyi Ruan[1], Yufeng Wang[2]

Detailed Affiliations:

[1]Ronald Reagan High School, San Antonio, TX, USA, [2]Department of Molecular Microbiology and Immunology, South Texas Center for Emerging Infectious Diseases, University of Texas at San Antonio, San Antonio, TX, USA.

Abstract: Urban parks in growing cities serve as important ecological refuges, yet the influence of park design and management, and topographic variation on soil microbial communities remains poorly understood. In San Antonio, Texas, municipal parks constitute actively managed urban landscapes that vary in recreational intensity and encompass a range of environmental conditions. This study focuses on the degree of microbial divergence between two major municipal green spaces: McAllister Park, an expansive, low-elevation, multi-use park, and Eisenhower Park, characterized by rugged limestone hills and canyon microclimates. Surface soil samples were collected from both sites. Total community DNA was extracted, and the V3-V4 hypervariable region of the bacterial 16S rRNA gene was amplified. Raw sequences were processed utilizing the DADA2 pipeline to resolve Amplicon Sequence Variants (ASVs). Alpha and beta diversity indices were computed and statistically evaluated using the phyloseq and vegan software. Alpha diversity analysis showed that McAllister Park possessed a more diverse and even bacterial community than Eisenhower Park, indicating a stable ecosystem. Beta diversity analysis indicated no compartmentalization between the two parks. Our testing for differential abundance pinpointed distinct indicator groups for each site. McAllister Park favored organic-matter decomposers like Actinomycetaceae, while Eisenhower's soil selected for anaerobic lineages like Succinivibrionaceae and Fusobacteriaceae. Even though these parks share the same city, these clear differences show that unique terrain, plant cover, and land management heavily influence local soil biology.

Keywords: metagenomics, San Antonio, soil microbiota

Poster P54

Title: PIWI-RNA-Based Signatures for Colorectal Cancer: Diagnostic and Prognostic Insights from Machine Learning and Survival Modeling

Author list: Gino Quintal, MS[1],Youping Deng, PhD[2],Yuanyuan Fu, PhD[2]

Detailed Affiliations:

[1]Western University of Health Sciences College of Osteopathic Medicine, Oregon, USA, [2]University of Hawaii John A. Burns School of Medicine, Hawaii, USA Co-responding author: Yuanyuan Fu, PhD, University of Hawaii John A. Burns School of Medicine, Hawaii, USA.

Abstract: Background Despite decreasing trends in incidence and mortality, colorectal cancer remains a leading cause of cancer-related deaths world-wide. Identifying novel diagnostic and prognostic biomarkers, as well as potential therapeutic targets is crucial for improving patient outcomes. This study investigates PIWI-interacting RNAs (piRNAs) as potential biomarkers for CRC. Materials and Methods Micro-RNA (miRNA) sequencing data from the Gene Expression Omnibus (GEO) were analyzed to identify differentially expressed piRNAs. These piRNAs were used to develop a diagnostic model based on the best performing machine learning algorithm. The diagnostic model was then validated with The Cancer Genome Atlas Colon Adenocarcinoma (TCGA-COAD) data set. A prognostic signature was constructed using LASSO Cox Regression with the TCGA-COAD data set and compared with 109 published prognostic signatures. Potential gene targets of the candidate piRNAs were predicted, and their associated pathways were analyzed. Results Among 4 machine learning classifiers, support vector machine (SVM), demonstrated the highest performance in the diagnostic model (AUC = 0.932). In the multivariable prognostic model, a 10-piRNA signature was significantly associated with worse survival outcomes. [$p<0.001$, HR = 2.76 (1.68-4.5)]. Advanced disease stages further increased risk {Stage III [($p<0.05$, HR = 3.22 (1.12-9.2)] and Stage IV [($p<0.05$, HR = 8.28 (2.88-23.8)]}. The 10-piRNA based signature gains significance when used individually in younger (40-60) and early-stage cancer patients (Stage I-II). When applied to the TCGA-COAD dataset, this piRNA-based signature outperformed 109 previously published models. Finally, piR-36011 was identified in both the prognostic and diagnostic model, with predicted targets involved in key CRC oncogenic pathways, including Netrin-1 signaling, MAPK, BRAF and RAF1. Conclusion Differentially expressed piRNAs demonstrate robust diagnostic potential. The superior performance of the piRNA-based prognostic model compared to existing signatures suggests that piRNAs may serve as more reliable biomarkers for CRC. Notably, piR-36011 emerges as a promising dual-purpose biomarker, given its strong association with colorectal cancer oncogenesis.

Keywords: piRNA, biomarker, colorectal cancer, machine learning, oncology

---

Poster P55

Title: Molecular profiling of the mystic human unique heterochromatin region of the Y chromosome

Author list: Ping Liang[1,2] and Sunny Lowell[1]

Detailed Affiliations:

[1]Department of Biological Sciences, [2]Centre for Biotechnology Brock University.

Abstract: The human Y chromosome has a distinctive molecular architecture shaped by suppressed recombination, lineage-specific rearrangements, and extensive accumulation of repetitive DNA. With >30 Mb of new sequences (and correction of errors) over the current standard human reference genome, GRCh38, the recent telomere-to-telomere Y chromosome assembly enabled examination of regions previously unresolved, including ampliconic gene families, palindromic sequences, satellite arrays, and other repeat-rich domains. Importantly, it added 41 protein-coding genes and many more non-coding genes. Among these regions, the heterochromatic Yq12 segment is especially notable. Once considered largely inaccessible because of its dense, repetitive structure, Yq12 is now recognized as a complex domain enriched for large satellite DNA blocks, including the human satellite 1 and satellite 3 arrays, as well as additional repeat-associated sequence features. This composition gives Yq12 a molecular profile that is distinct from both the euchromatic portions of the Y chromosome and comparable regions in other primates. Here, we focus on characterizing the sequence organization and epigenetic profiles of the human Yq12 heterochromatic region and its contribution to the unique identity of the human Y chromosome. Our analysis revealed several highly unusual features of this region, including AluY as the exclusive type of transposable elements, forming a high level of tandem repeat clusters with HSAT1B and contributing to lncRNAs, and a high content of long non-coding genes. For the first time, we can provide the specific sequence identities of the previously reported mystic, a large population of RNA species that exhibit developmental-stageand testis-specific expression

from this region, along with the associated epigenetic signatures. We hope that our results will provide new insights into Y chromosome biology and the function of its heterochromatin region in gene regulation and male-associated diseases.

Keywords: human Y chromosome, heterochromatin, epigenetic profiling, lncRNAs

---

Poster P56

Title: Prediction of Rapid Fibrosis Progression in Metabolic Dysfunction–Associated Steatotic Liver Disease with Interpretable Machine Learning Models

Author list: Tahmina Sultana Priya[1,2,3], Huihuang Yan[3], Konstantinos N. Lazaridis[4], Eric W. Klee[3], Danfeng (Daphne) Yao[1,2], Shulan Tian[3]

Detailed Affiliations:

[1]Department of Computer Science, Virginia Tech, Blacksburg, VA, USA. [2]Sanghani Center for Artificial Intelligence and Data Analytics, Virginia Tech, Blacksburg, VA, USA. [3]Division of Computational Biology, Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, USA. [4]Division of Gastroenterology and Hepatology, Department of Internal Medicine, Mayo Clinic, Rochester, MN, USA.

Abstract:

Introduction

Metabolic dysfunction-associated steatotic liver disease (MASLD) is characterized by substantial clinical and biological heterogeneity, leading to highly variable rate of fibrosis progression. Accurately identifying patients at risk of rapid progression remains a critical unmet clinical need. Although existing non-invasive tests can effectively stage current fibrosis, they have limited ability to predict future disease risk and often fail to identify patients who have minimal baseline fibrosis but later experience rapid progression.

Methods

We developed and validated a partitioned machine learning (ML) framework using two independent Mayo Clinic cohorts, the Mayo Clinic Biobank and the Tapestry Study, comprising a total of 13,227 patients with MASLD. The framework integrates longitudinal laboratory measurements, diagnosis code trajectories, and genetic information from electronic health records to predict the risk of rapid fibrosis progression. We first applied latent class analysis with clinical variables to identify clinically meaningful subgroups. Within each subgroup, we performed multiple feature selection and trained supervised ML models, including a stacking ensemble. To enhance clinical interpretability, we derived the partitioned Rapid Fibrosis Progression Score (pRFPS), an interpretable point-based risk score, using SHAP (SHapley Additive exPlanations)-based feature attribution via the AutoScore framework.

Results

Unsupervised phenotypic partitioning identified two reproducible subgroups with distinct metabolic and hepatic characteristics. The cardiometabolic-dominant subgroup was characterized by higher BMI, poorer glycemic control, and elevated triglycerides, whereas the liver-centric subgroup exhibited relatively milder metabolic dysfunction but more pronounced hepatic injury, a stronger genetic predisposition, and higher rates of hepatocellular carcinoma and liver transplantation. Subgroup-specific ML models consistently outperformed non-partitioned population-level models, achieving up to 4% increase in accuracy and a maximum accuracy of 85%. The pRFPS effectively stratified patients into clinically meaningful risk groups. The high-risk individuals had significantly higher rates of hepatocellular carcinoma, liver transplantation, and other adverse liver-related outcomes. Notably, the framework identified 23%–25% of rapid progressors who would have been classified as low risk by conventional clinical tools.

Conclusion

This study demonstrates that, with routinely collected clinical and laboratory data, our partitioned, subtype-aware ML framework could reliably predict the risk of rapid fibrosis progression in MASLD. By explicitly accounting for disease heterogeneity, our approach improves predictive performance and uncovers subtype-specific drivers of progression that are not captured by conventional fibrosis staging systems. The pRFPS provides a clinically actionable tool for risk stratification and early identification of high-risk patients with MASLD. This work offers a generalized framework for precision medicine in MASLD and other heterogeneous chronic diseases.

Poster P57

Title: Genome-scale profiling of RNA capping levels uncovers sequence determinants and roles in gene regulation

Author list: Zhenguo Lin[1]

Detailed Affiliations:
[1]Saint Louis University.

Abstract: TBD

Keywords: TBD

Poster P58

Title: Investigating the Effects of Microplastics on miRNA Using Novel Bioinformatics Pipeline on Mus musculus Prefrontal Cortex Transcriptomic Data

Author list: Thomas Yu[1]

Detailed Affiliations:
[1]Allendale Columbia School.

Abstract: TBD

Keywords: TBD

Poster P59

Title: Single-molecule RNA modification mapping by AI analysis of nanopore sequencing

Author list: Yunxia Wang[1]

Detailed Affiliations:
[1]Boston Children's Hospital, Harvard Medical School.

Abstract: TBD

Keywords: TBD

Poster P60

Title: Classification of Cancer Types Using Cross-Cancer SNP Signatures from a Large-Scale Genomic Profile

Author list: Sunwoo Han[1], Limin Jiang[2] and Yan Guo[2,*]

Detailed Affiliations:
[1]Department of Biostatistics, Florida International University, Miami FL 33199, USA, [2]Department of Public Health Sciences, University of Miami, Miami FL 33316, USA. [*]Corresponding author

Abstract: Single-nucleotide polymorphisms (SNPs) represent the most prevalent form of genetic variation and play a central role in cancer research through genome-wide association studies (GWAS). While traditional GWAS have identified risk-associated SNPs by comparing cases with healthy controls, these variants often lack specificity for distinguishing among different cancer types. To address this limitation, we conducted a cross-cancer GWAS using twelve cancer types from a large-scale genomic profile. By contrasting each cancer type against the remaining types, this design aims to identify cancer-type-specific SNPs that capture inter-tumor heterogeneity. We identified significant SNPs for each cancer type and evaluated their discriminative performance using random forest models in both binary and multiclass classification settings. We further compared these SNPs with variants reported in the GWAS Catalog derived from conventional case-control studies. Notably, the identified SNPs showed minimal overlap with those in the GWAS Catalog, suggesting that cross-cancer GWAS captures distinct genetic signals. Overall, this approach improves the identification of cancertype-specific genetic signatures and enhances discrimination among different cancers, providing a complementary strategy for refining genetic biomarkers.

Keywords: cross-cancer GWAS, cancer specificity, random forests, classification

---

Poster P61

Title: SCALE-ST: Multi-Scale Learning for Histology-Based Prediction of Spatial Gene Expression

Author list: Sayed Asaduzzaman[1,2], Hasin Rehana[1,2], Brett McGregor[2], and Junguk Hur[2,*]

Detailed Affiliations:

[1]School of Electrical Engineering & Computer Science, University of North Dakota, Grand Forks, North Dakota, 58202, USA, [2]Department of Biomedical Sciences, School of Medicine and Health Sciences, University of North Dakota, Grand Forks, North Dakota, 58202, USA. [*]Corresponding author

Abstract: Spatial transcriptomics allow mapping gene expression within intact tissue, but current platforms require a trade-off between resolution and spatial coverage. Visium v2 offers whole-transcriptome profiling at a coarse resolution of about 55 μm, while the newer Visium HD reaches near-cellular resolution (2–16 μm), though it requires significant resources. Overcoming this computational gap is crucial for advancing high-resolution spatial biology using standard histological techniques. Here, we present SCALE-ST (Spatial Cross-scale Learning of Expression for Spatial Transcriptomics), a multiscale deep learning framework designed to infer high-resolution gene expression directly from hematoxylin and eosin (H&E)-stained image patches. Trained on co-registered hierarchical multi-scale patches (2 μm, 8 μm, and 16 μm) from Visium HD data, SCALE-ST captures cross-scale morphological features because gene expression also depends on surrounding tissue structure. Its architecture combines a Vision Transformer backbone with a center-focused convolutional branch and employs a Zero-Inflated Negative Binomial (ZINB) output head to accurately model sparse, overdispersed data. Evaluation on Visium v2 data shows that SCALE-ST delivers strong predictive performance. Gene-level correlation between true and predicted gene expression values was 0.25 at 16 μm, 0.21 at 8 μm, and 0.17 at 2 μm. Spot-level analysis yields high Spearman (0.91–0.93) and lognormalized Pearson correlations (0.73–0.75) across scales, indicating good preservation of relative expression structure. Although scatter plots show strong spatial and structural agreement, a systematic overestimation bias increases absolute errors at higher spatial resolutions. Implementing a strict blankpatch suppression strategy effectively removes spurious background predictions. SCALE-ST thus provides a histology-based framework for sub-spot gene expression inference, enabling exploratory analysis of morphology-driven spatial gene expression patterns beyond conventional spot-level resolution.

Keywords: Multi-Scale Learning, Gene Expression Prediction, Spatial Transcriptomics, Visium, Histology, Transfer Learning

---

Poster P62

Title: Proteinext 2.0: Protein Function Prediction with Sequence Embeddings, Structure Embeddings, and Natural Language Processing

Author list: Naufal Alavi[1], Sanjina Kumari[1], Boen Liu[1], Christopher Barker[1], Ashnaa George[1], Xiao Chen[2], Jie Hou[3], Dong Si[4], Renzhi Cao[1]

Detailed Affiliations:

[1]Department of Computer Science, Pacific Lutheran University, Tacoma 98447, United States, [2]Computer Science Department , Hamilton College, United States, [3]Department of Computational Mathematics, Science and Engineering, Michigan State University, United States, [4]School of Science, Technology, Engineering & Mathematics, University of Washington, USA, United States.

Abstract: Proteins are fundamental to life, but understanding their functions remains challenging due to their complex structures and diverse roles. Existing computational methods often suffer from slow performance, high computational demands, and difficulty handling highly specific or low-homology proteins. Proteinext 2.0 is a next-generation model that builds upon the foundation of Proteinext by incorporating both sequence and structural information for improved protein function prediction. The model predicts Gene Ontology (GO) terms across Molecular Function (MF), Cellular Component (CC), and Biological Process (BP) categories. Proteinext 2.0 leverages Meta's Evolutionary Scale Modeling (ESM-3) model to generate high-quality protein sequence embeddings and integrate structural features derived from AlphaFold2-predicted 3D protein structures. Additional features such as secondary structure (SS) and relative solvent accessibility (RSA) are extracted using DSSP and fused with sequence embeddings to form a multimodal representation. These combined features are processed using a fine-tuned BigBird transformer model, enabling efficient handling of long input sequences without truncation. Our model demonstrates strong predictive capability, achieving a precision of 0.64, a recall of 0.15, and an Fmax score of 0.18. Benchmark comparisons indicate that Proteinext 2.0 attains precision on par with state-of-the-art

approaches such as GAT-GO and TransFun, underscoring the effectiveness of integrating both sequence and structural information. Overall, these results highlight the promise of multimodal deep learning methods for scalable and accurate protein function prediction. Proteinext 2.0 is available at https://github.com/Cao-Labs/lutesAI2025/tree/main/naufal

HOTEL INFO & PARKING

Conference location, lodging, and transportation

Conference Venue

Jacobs School of Medicine and Biomedical Sciences

University at Buffalo | 955 Main Street | Buffalo, NY 14203-1121

CAMPUS	METRO	PARKING
Buffalo Niagara Medical Campus	**Allen/Medical Campus Station**	**854 Ellicott Street garage**

Hotel Information

Hilton Garden Inn Buffalo Downtown

The ICIBM 2026 discounted hotel room block has been fully booked. A limited number of rooms may still be available at the current regular rate, subject to availability.

Nearby alternatives

- Hyatt Regency Buffalo
- Wyndham Garden Buffalo Downtown
- Residence Inn by Marriott Buffalo Downtown
- Aloft by Marriott Buffalo Downtown
- Holiday Inn Express & Suites Buffalo Downtown – Medical Ctr by IHG
- Embassy Suites by Hilton Buffalo

Getting to Buffalo and the Conference Venue

By Air: Buffalo Niagara International Airport (BUF) is approximately 15 miles from the downtown campus. Taxi, rideshare, and NFTA-Metro Route 24 service are available.

By Train: Amtrak provides service to Buffalo-Depew Station and Buffalo-Exchange Street Station.

By Car: Buffalo is accessible via I-90 and I-190. Use the paid garage at 854 Ellicott Street for conference parking.

Hotel to Venue by Metro Rail: Walk to Lafayette Square Station, take Metro Rail toward University Station, and exit at Allen/Medical Campus Station.

Hotel to Venue on Foot: The venue is approximately a 20-minute walk from the Hilton Garden Inn Buffalo Downtown.

Current travel links

NFTA-Metro Route 24 schedule

NFTA-Metro fare information

Buffalo airport transportation and tourism

SPECIAL ACKNOWLEDGEMENTS

We extend our sincere appreciation to all members of the ICIBM 2026 Program Committee and the many reviewers who generously contributed their time and expertise to the peer review of submitted manuscripts. We are deeply grateful to the General Chairs and to the Steering, Publication, Workshop, Tutorial, Award, and Future Scientist in AI Committees for their leadership and service.

We also thank the staff and technical support teams at the Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, for their assistance, along with the many volunteers who helped coordinate the keynote and invited speakers and support early-career researchers and trainees.

Finally, we gratefully acknowledge the conference sponsors and everyone whose dedication and contributions helped make ICIBM 2026 possible.

SPONSORS

ICIBM 2026 gratefully recognizes its current sponsors

PLATINUM SPONSOR

admera health

GOLD SPONSORS

MGI

PIXELGEN TECHNOLOGIES

Singleron

SILVER SPONSORS

Computational and Structural Biotechnology Journal
A SCIENCE PARTNER JOURNAL

BMEF
A SCIENCE PARTNER JOURNAL

10X GENOMICS

TRAVEL AWARD SUPPORT

The Computational and Structural Biotechnology Journal supports four ICIBM 2026 travel awards.

Sponsorship contact: icibm.common@gmail.com

PLATINUM SPONSOR

Admera Health

Advancing discovery through true collaboration.

HISTORY **2014** launched as a Genewiz spinout | **2024** added BioEcho and EchoLUTION purification | **2026** named the first global Shasta service provider

AT A GLANCE

2014	19,000+	30+	99.9%
Established	Publications supported	Omic platforms	Data pass-through*

ABOUT

Established in 2014 as a spinout from Genewiz, Admera Health is a South Plainfield, New Jersey genomics and bioinformatics services partner. Its CLIA/CLEP-certified and CAP-accredited environment supports research from experimental design and sample processing through multi-omic data generation and analysis.

FLAGSHIP OFFERINGS

RNA & Multi-Omics

Bulk and total RNA-seq, Iso-Seq, epigenomics, metabolomics, proteomics, and integrated bioinformatics.

Single-Cell & Spatial

Certified workflows spanning 10x, Parse, Singleron, Shasta, Visium HD, and Stereo-seq technologies.

Genomics & Clinical-Grade

WGS/WES, clinical-grade WGS, long-read sequencing, sequencing-only, and project-specific analysis.

Standout feature: a scientist-to-scientist partner for complex translational, single-cell, spatial, and integrated multi-omics study designs.

EXPLORE admerahealth.com | 2026 services catalog

GOLD SPONSOR

Singleron Biotechnologies

End-to-end single-cell solutions—from sample to interpretable data.

HISTORY **2018** founded around clinical single-cell translation | **2025** launched MobiuSCOPE full-length scRNA-seq | **2026** expanded with RetroSCOPE, Tensor, and new regional access

AT A GLANCE

2018	1B+	9,000+	1,000+
Founded	Cells sequenced*	Projects supported*	Publications supported*

ABOUT

Founded in January 2018, Singleron develops single-cell multi-omics solutions for research, drug development, and precision medicine. Its global portfolio combines tissue preparation, SCOPE-chip workflows, kits, automation, services, and accessible bioinformatics across Europe, Asia, China, and the United States.

FLAGSHIP OFFERINGS

RetroSCOPE FFPE

Whole-transcriptome single-nucleus profiling of coding and non-coding RNA from archived FFPE tissue.

MobiuSCOPE

Full-length single-cell transcript insight—including splicing, SNPs, and fusions—using short-read sequencing.

Matrix NEO & Tensor

Automation from chip processing to scalable library generation; Tensor supports up to 32 samples in 24 hours.

Standout feature: an integrated path from difficult or archived samples to single-cell libraries, publication-ready analysis, and translational interpretation.

EXPLORE singleron.bio | Explore all products

GOLD SPONSOR

Pixelgen Technologies

Mapping cell-surface architecture through protein interactomics.

HISTORY **2020** founded in Stockholm | **2025** introduced the Proximity Network Assay product line | **2026** launched Proxiome Kit v2 for translational-scale studies

AT A GLANCE

2020	155	~50 nm	64
Founded	Surface proteins	Average resolution	Samples/experiment

ABOUT

Pixelgen was founded in Stockholm in 2020 to develop DNA-based visualization of cell-surface proteomics. Its Proximity Network Assay converts nanoscale protein neighborhoods into sequencing-readable single-cell networks, enabling analysis of abundance, clustering, and colocalization without specialized imaging equipment.

FLAGSHIP OFFERINGS

Proxiome Kit v2

A 155-plex human immune-cell panel producing abundance, clustering, and colocalization scores at nanoscale resolution.

Custom & CAR T Add-ons

Custom antibody conjugation plus CAR-specific reagents extends the core panel for targeted translational studies.

Pixelator Software

Open-source processing and analysis tools connect sequencing reads to proximity networks and single-cell workflows.

Standout feature: a new computational data type for studying immune-cell states, protein neighborhoods, biomarkers, and cell-therapy mechanisms.

EXPLORE pixelgen.com | Proxiome Kit v2

SILVER SPONSOR

Computational and Structural Biotechnology Journal

Computational and Structural Biotechnology Journal

A SCIENCE PARTNER JOURNAL

Open access research across computational and structural biotechnology.

HISTORY **2012** launched at the biology-computation interface | **Today** four editorial sections span established and emerging fields | **2026** ICIBM event partner and travel-award supporter

AT A GLANCE

2012	4.8	8.2
First issue	2025 Impact Factor	2025 CiteScore

ABOUT

Launched with its first issue in April 2012, Computational and Structural Biotechnology Journal (CSBJ) advances functional and mechanistic understanding of biological systems through computational, structural, and data-driven research. Its current scope reaches from molecular and systems biology to digital health, advanced materials, quantum biology, and biophotonics.

CORE ARTICLE TYPES

Research & Methods

Original studies, methods, databases, software/web servers, and innovation reports across computational biotechnology.

Reviews & Perspectives

Reviews, mini-reviews, perspectives, and editorials connect methods with biological mechanism and translation.

Current Frontiers

Smart hospitals, nanoscience, advanced materials, quantum biology, biophotonics, and cross-disciplinary special issues.

TRAVEL AWARD PARTNER: CSBJ supports four ICIBM 2026 travel awards, helping early-career researchers participate in the scientific program.

EXPLORE CSBJ journal home | Information for authors

SILVER SPONSOR

BME Frontiers

BMEF

A SCIENCE PARTNER JOURNAL

Engineering the future of biomedicine through open, multidisciplinary research.

HISTORY **2020** launched after a year of preparation | **2025** reported a 7.7 JIF and Q1 biomedical-engineering rank | **2026** continues a broad basic-to-clinical engineering remit

AT A GLANCE

2020	7.7	9.8
Journal launch	2024 Impact Factor*	CiteScore

ABOUT

BME Frontiers (BMEF) launched in April 2020 as the third Science Partner Journal. Published in association with the Suzhou Institute of Biomedical Engineering and Technology, Chinese Academy of Sciences, it connects foundational engineering, preclinical investigation, clinical translation, and technologies that improve human health.

CORE ARTICLE TYPES

Research Articles

Original biomedical-engineering studies spanning mechanisms, devices, materials, systems, imaging, sensing, and translation.

Rapid Reports

Concise, validated, ground-breaking advances designed for timely communication across the multidisciplinary community.

Reviews & Perspectives

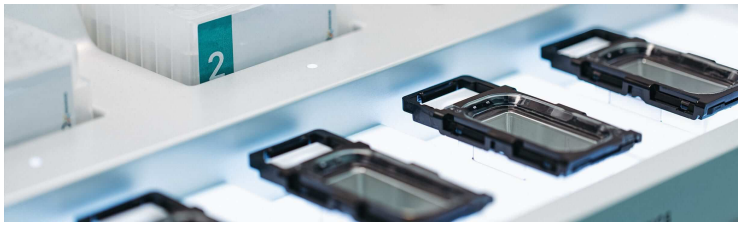
Synthesis and forward-looking viewpoints across biomaterials, neuroengineering, digital medicine, and emerging BME frontiers.

Standout feature: an open-access venue for researchers bridging AI, computation, imaging, sensing, biomaterials, devices, and translational medicine.

EXPLORE BMEF journal home | Information for authors

SILVER SPONSOR

10x Genomics



Decoding biology. Transforming health.

HISTORY **2012** incorporated in California | **2026** introduced the Atera In Situ platform | **2026** expanded capabilities through Proteintech Genomics

AT A GLANCE

2012	10,000+	2,200+	>$1.5B
Incorporated	Publications*	Patents*	R&D investment*

ABOUT

Incorporated in 2012, 10x Genomics develops single-cell and spatial platforms that help researchers characterize cell states, regulatory programs, immune diversity, and tissue context. Its integrated portfolio spans assays, instruments, analysis pipelines, cloud tools, datasets, and visualization software.

FLAGSHIP OFFERINGS

Atera In Situ

Whole-transcriptome spatial analysis with single-cell sensitivity, 18,000+ genes, and large-scale tissue throughput.

Chromium Single Cell

Flex Apex and Universal chemistries support scalable gene expression, immune profiling, CRISPR, protein, and epigenomics.

Xenium & Visium

Targeted in situ exploration and unbiased whole-transcriptome spatial discovery provide complementary views of tissue biology.

Standout feature: linked single-cell, spatial, and analysis ecosystems for building cell atlases, biomarkers, disease models, and AI-ready biological datasets.

EXPLORE 10xgenomics.com | Explore Atera