The 13th International Conference on Intelligent Biology and Medicine (ICIBM 2025)

August 3rd - 5th, 2025
Columbus, Ohio, USA

Hosted by:
The International Association for Intelligent Biology and Medicine (IAIBM),
Department of Biomedical Informatics, The Ohio State University, and
Translational Data Analytics Institute, The Ohio State University

Table of Contents

Welcome to ICIBM 2025!

On behalf of all our conference committees and organizers, we are thrilled to welcome you to the 2025 International Conference on Intelligent Biology and Medicine (ICIBM 2025). 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 2025, 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 2025 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. Veera Baladandayuthapani, Jiang Bian, Qing Nie, and Julie A. Johnson. Additionally, we are honored to feature four eminent scholar speakers: Drs. Yu-Ping Wang, Kaifu Chen, Yufeng Shen, and Xiang Zhou. These distinguished researchers are global leaders in their respective fields, and we are privileged to have them share their insights at ICIBM 2025. The conference will also include twelve workshops, along with presentations from faculty members, postdoctoral fellows, PhD 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 exciting opportunities ICIBM 2025 will offer. We would like to express our deepest gratitude to our sponsors, whose generous support has made this event possible. Our sponsors include Admera Health, Arcegen, Complete Genomics, Computational Biology and Chemistry Journal, Computational and Structural Biotechnology Journal, Olink, Singleron, and 10x Genomics.

Finally, our heartfelt thanks go to all members of the ICIBM 2025 committees and our volunteers for their dedication and hard work. Their commitment to making ICIBM 2025 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 2025. We look forward to your active participation in all that the conference has to offer!

Sincerely,

Lianbo Yu, PhD
Program Co-Chair
Associate Professor
The Ohio State University

Leng Han, PhD
Program Co-Chair
Professor
Indiana University

Maciej Pietrzak, PhD
Publication Chair
Associate Professor
The Ohio State University

Lang Li, PhD
General Co-Chair
Professor
The Ohio State University

Zhongming Zhao, PhD
General Co-Chair
Professor
University of Texas Health Science Center at Houston

ACKNOWLEDGEMENTS

General Chairs
Lang Li, The Ohio State University
Zhongming Zhao, University of Texas Health Science Center at Houston
Qin Ma, The Ohio State University

Program Committee
Lianbo Yu, The Ohio State University
Leng Han, Indiana University
Oindrila Bhattacharyya, The Ohio State University
Weidan Cao, The Ohio State University
Sapuni Chandrasena, The Ohio State University
Xiao Chang, Children's Hospital of Philadelphia
Lijun Cheng, The Ohio State University
Mohamed Elsaid, The Ohio State University
Rejuan Haque, The Ohio State University
Matthew Hayes, Xavier University of Louisiana
Tao Huang, Shanghai Institute of Nutrition and Health
Weichun Huang, U.S. Environmental Protection Agency
Peilin Jia, University of Texas Health Science Center at Houston
Garrett Kinnebrew, The Ohio State University
Jiaying Lai, Johns Hopkins University
Aimin Li, Xi'an University of Technology
Fuhai Li, Washington University at St. Louis
Yuan Lin, Kristiania University of Applied Science, Norway
Xiaoming Liu, University of South Florida
Tianle Ma, Oakland University
Joseph McElroy, The Ohio State University
Xiaokui Mo, The Ohio State University
Jiang Qian, Johns Hopkins University
Michal Seweryn, University of Lodz
Rama Shankar, Michigan State University
Li Shen, University of Pennsylvania
Yang Shen, Texas A&M University
Jiao Sun, University of Central Florida
Shulan Tian, May Clinic
Manabu Torii, Kaiser Permanente
Alper Uzun, Brown University
Ece Uzun, Brown University
Jiayin Wang, Xi'an Jiaotong University
Junbai Wang, Radium Hospital
Qing Wang, Univerity of Florida
Junfeng Xia, Anhui University
Min Xu, Carnegie Mellon University
Jianhua Xuan, Virginia Tech
Yu Xue, Huazhong University of Science and Technology

Huihuang Yan, Mayo Clinic
Rui Yin, University of Florida
Shaojie Zhang, University of Central Florida
Wei Zhang, University of Central Florida
Jim Zheng, University of Texas Health Science Center at Houston

Publication Committee
Maciej Pietrzak, The Ohio State University
Xinghua (Mindy) Shi, Temple University

Workshop/Tutorial Committee
Hongbo Liu, University of Rochester
Yusi Fu, Texas A&M University
Qianqian Song, University of Florida
Pengyue Zhang, Indiana University
Travis Johnson, Indiana University
Chi Zhang, Oregon Health & Science University
Xiang Gao, Loyola University Chicago
Ece Uzun, Brown University

Award Committee
Fuhai Li, Washington University in St. Louis

Trainee Committee
Chi Zhang, Oregon Health & Science University
Jingwen Yan, Indiana University

Publicity Committee
Shibiao Wan, University of Nebraska
Alper Uzun, Brown University

Local Committee
Joseph McElroy, The Ohio State University
Gang Peng, Indiana University
Zhiyong Ding, MD Anderson Cancer Center

Schedule

The International Conference on Intelligent Biology and Medicine (ICIBM 2025) Program, August 3-5, 2025
Pomerene Hall, 1760 Neil Ave, Columbus, OH

Sunday, August 3rd, 2025

7:30 AM - 5:30 PM	**Registration**				
CONCURRENT SESSIONS/WORKSHOPS					
Room: **320**		Room: **301**		Room: **350**	
Genomics and Translational Bioinformatics Working Group Chairs: Ece Uzun, Wenyu Song		**Advanced Computational Statistics and Artificial Intelligence to Address Public Health Epidemics** Chairs: Naleef Fareed, Soledad Fernandez		**Microbiome Data Analysis: Advanced Methods and Practical Applications** Chairs: Qunfeng Dong, Xiang Gao	
8:30 AM - 8:50 AM	**Calibration of Computational Prediction Tools for Improved Clinical Variant Classification and Interpretation** Vikas Pejaver, Mount Sinai	8:30 AM - 8:50 AM	**Leveraging Urinary Drug Test (UDT) Results as a Novel Data Source and Proxy for Drug Use** Naleef Fareed/Soledad Fernandez, The Ohio State University	8:30 AM - 8:50 AM	**A Deep Learning Feature Importance Test Framework for Integrating Informative High-dimensional Biomarkers to Improve Disease Outcome Prediction** Baiming Zou, University of North Carolina at Chapel Hill
8:50 AM - 9:10 AM	**Opioid Prescriptions and Associated Patient Response: An Integrated Genetic Analysis Using Clinical Biobank**	8:50 AM - 9:10 AM	**Predicting Opioid Overdose Mortality Using UDT Data with a Bayesian Approach** John Myers, The Ohio State University	8:50 AM - 9:10 AM	**Enhancing Microbiome-Trait Prediction through Phylogeny-Aware Modeling and Data Augmentation**

	Wenyu Song, Brigham and Women's Hospital				Yang Lu, University of Waterloo
9:10 AM - 9:30 AM	**Leveraging Deep Learning to Infer Cellular Dynamics** Shengyu Li, Houston Methodist Research Institute	9:10 AM - 9:30 AM	**Implementing a Spatial-Temporal Graph Neural Network (ST-GNN) Framework, a Novel, Multi-Modal Data Approach for Predicting Opioid Overdose Death Rates** Xianhui Chen, The Ohio State University	9:10 AM - 9:30 AM	**VirusPredictor: Software to Predict Virus-related Sequences in Human Data** Dawei Li, Texas Tech University Health Sciences Center
9:30 AM - 9:50 AM	**Clinical and Genomic Investigation of Immune-Related Adverse Events** Qianqian Song, University of Florida	9:30 AM - 9:50 AM	**Flexible Copula-Based Capture–Recapture Modeling of Opioid Misuse Using Urine Drug Testing Data: Evidence from Franklin County, Ohio (2016–2023)** Fode Tounkara, The Ohio State University	9:30 AM - 9:50 AM	**Bayesian Spatial Statistical Models for Quantifying Relationships Among Cell Types in Image Data** Jacqueline R. Starr, Brigham and Women's Hospital, Harvard Medical School
9:50 AM - 10:10 AM	**Machine Learning-Based Integration of Transcriptomic and Epigenetic Data for Cancer Biomarker Discovery** Alper Uzun, Brown University	9:50 AM - 10:10 AM	**Evaluating the Public Health Decision Support Landscape for Opioid Outcomes** Brandon Slover/Neena Thomas, The Ohio State University	9:50 AM - 10:10 AM	**Multimedia: An R Package for Multimodal Mediation Analysis of Microbiome Data** Kris Sankaran, University of Wisconsin–Madison
10:10 AM -10:30 AM		***Coffee/Tea Break***			
10:30 AM - 10:50 AM	**Subtyping Metabolic Dysfunction-**	10:30 AM - 10:50 AM	**Paper 46: PCORsearch: A Scalable, User-**	10:30 AM - 10:50 AM	**Leveraging New Genomic LLMs for Studying Under-**

	Associated Steatotic Liver Disease using Electronic Health Record-Linked Genomic Cohorts Reveals Diverse Etiologies and Progression Shulan Tian, Mayo Clinic		**Centric Platform for Self-Service Cohort Discovery and Feasibility Analysis of PCORnet Data** Jacob Herman, The Ohio State University		**Annotated Microbial Genes** Siyuan Ma, Vanderbilt University
10:50 AM - 11:10 AM	**Predicting Cancer Recurrence Using Deep Learning Based Models** Ece Uzun, Brown University	10:50 AM – 11:10 AM	**Paper 52: Towards AI Co-Scientists for Scientific Discovery in Precision Medicine** Hao Li, Washington University in St. Louis	10:50 AM - 11:10 AM	**Integrated Transcriptomics Analysis on Human Respiratory Viral Inoculation and Vaccine Challenge Studies** Fei Zou, University of North Carolina at Chapel Hill
11:10 AM - 11:30 AM	**Genetic Impact of Alternative Transcription Initiation Reveals a Novel Molecular Phenotype for Human Diseases** Lei Li, Shenzhen Bay Laboratory	11:10 AM - 11:30 AM	**Paper 3: Tokenvizz: GraphRAG-Inspired Tokenization Tool for Genomic Data Discovery and Visualization** Zhenxiang Gao, Case Western Reserve University	11:10 AM - 11:30 AM	**AI-Powered Discovery of Novel Antimicrobial Peptides in *Trichomonas vaginalis*** Xiang Gao, Loyola University Chicago
11:30 AM - 1:30 PM		**Lunch Break / Poster Session I (Atrium)**			
1:30 PM - 1:40 PM		**Opening Remarks (Room 320)**			
1:40 PM - 2:20 PM		**Keynote Lecture** (Room 320) **Veera Baladandayuthapani, PhD** **University of Michigan**			
CONCURRENT SESSIONS/WORKSHOPS					

Room: **320**		Room: **301**		Room: **350**	
Advancements in AI and Large Language Models for Biomedical Research Chairs: Jing Su, Gangqing Hu		**Big Data for Better Studying Disease Systems** Chair: Xiuwei Zhang		**Advanced Omics Platforms and Tools** Chairs: Kaixong Ye, Hongbo Liu	
2:30 PM - 2:50 PM	**Preliminary Evaluation of ChatGPT Model Iterations in Emergency Department Diagnostics** Gangqing Hu, West Virginia University	2:30 PM - 2:50 PM	**Eminent Scholar Presentation** Yu-Ping Wang, Tulane University	2:30 PM - 2:50 PM	**CCLLM: Cellular Community Large Language Model to identify motifs of cell organization in spatial transcriptomics** Juexin Wang, Indiana University
2:50 PM - 3:10 PM	**Thinking, Fast and Slow: DualReasoning Enhances Clinical Knowledge Extraction from Large Language Models** Haining Wang, Indiana University	2:50 PM - 3:10 PM	**ShinyEvents: Harmonizing Longitudinal Data for Real World Survival Estimation.** Timothy Shaw, Moffitt Cancer Center	2:50 PM - 3:10 PM	**A Universal Gene Representation of Atlas Single Cell Data** Hao Chen, University of Illinois Chicago
3:10 PM - 3:30 PM	**mcDETECT: Decoding the Dark Transcriptomes in 3D with Subcellular-Resolution Spatial Transcriptomics** Jian Hu, Emory University	3:10 PM - 3:30 PM	**Harnessing Big Data to Advance Understanding of Novel Therapeutic Strategies** Yuan Liu, Indiana University	3:10 PM - 3:30 PM	**Decoding Kidney Disease at Single-Cell Resolution: A Cross-Platform Spatial Transcriptomics Study** Haojia Wu, Washington University in St. Louis
3:30 PM - 3:50 PM		***Coffee/Tea Break***			
3:50 PM - 4:10 PM	**A Visual-Omics Foundation Model**	3:50 PM - 4:10 PM	**Spatially Resolved Transcriptomics and**	3:50 PM - 4:10 PM	**DNA Methylation Predictors of**

	for Integrating Histopathology Images and Transcriptomics Weiqing Chen, Houston Methodist Research Institute		**Proteomics to Interrogate Biological Mechanisms Underlying Cancer Disparities** Nina Steele, University of Cincinnati		**Inflammatory Cytokine Changes in Breast Cancer Survivors Undergoing Chemotherapy** Hongying Sun, University of Rochester
4:10 PM - 4:30 PM	**Large Language Models in Cancer Pharmacogenomics: from Drug-Gene Association to Response Prediction** Yu-chiao Chiu, University of Pittsburgh	4:10 PM - 4:30 PM	**Studying Single Cells Through Multi-Omics and Multi-Condition scRNA-seq** Xiuwei Zhang, Georgia Institute of Technology	4:10 PM - 4:30 PM	**Age-Related Patterns of DNA Methylation Changes** Gang Peng, Indiana University
4:30 PM - 4:50 PM	**STHD: Probabilistic Cell Typing of Single Spots in Whole Transcriptome Spatial Data with High Definition** Yi Zhang, Duke University	4:30 PM - 4:50 PM	**High-Resolution Reconstruction of Single-Cell Specific Spatial Genome Architectures in 3D Space Reveals Context-Specific Mechanisms of Long-Range Gene Regulation** Jianrong Wang, Michigan State University	4:30 PM - 4:50 PM	**Uncovering Hidden Biological and Technical Links from Large-scale DNA Methylome Data** Wanding Zhou, Children's Hospital of Philadelphia
4:50 PM - 5:10 PM	**Predicting Protein-Protein Interactions with Structure-Based ML/DL Modeling** Haiqing Zhao, University of Texas Medical Branch	4:50 PM - 5:10 PM	**Integrating Amyloid Imaging and Genetics for Early Risk Stratification of Alzheimer's Disease** Jingwen Yan, Indiana University	4:50 PM - 5:10 PM	**A BLAST from the Past: Revisiting BLAST's E-value** Yang Lu, University of Waterloo

5:10 PM - 5:30 PM	**A Benchmarking Framework for Foundation Models in Drug Response Prediction** Qianqian Song, University of Florida	5:10 PM - 5:30 PM	**Integrated Multi-Omics Study in Early Onset of Type 1 Diabetes.** Wenting Wu, Indiana University	5:10 PM - 5:20 PM	**Resting with Rhythm: Brain Functional Network Connectivity and Music Habits in Adolescents** Anaiah Calhoun TReNDS

Monday, August 4th, 2025

8:00 AM - 6:00 PM	**Registration**				
8:30 AM - 9:10AM	**Keynote Speaker** (Room 320) **Jiang Bian, PhD** **Indiana University**				
CONCURRENT SESSIONS/WORKSHOPS					
Room: **320**		Room: **301**		Room: **350**	
Advances in Target Discovery and Computational Drug Design Chairs: Pengyue Zhang, Yijie Wang		**Future Scientists in AI** Chairs: Chi Zhang, Jingwen Yan		**Biosignals and Omics in Neurological and Cancer Diseases: Opportunities and Challenges** Chairs: Haoqi Sun, Chen Huang	
9:20 AM - 9:40 AM	**Eminent Scholar Presentation** Kaifu Chen Harvard Medical School	9:20 AM - 9:30 AM	**Unlocking Fine-Grained Features: Vision Foundation Models for Improved Skin Cancer Classification** Alex Fu Union County Magnet High School	9:20 AM - 9:40 AM	**Bioinformatics Meets Biosignals: Opportunities and Challenges** Haoqi Sun, Harvard Medical School
		9:30 AM - 9:40 AM	**Automated Clinical Diagnosis using ML and Electronic Health Records: A Prototype**		

			in IBD Anshu Mukherjee University of California, San Francisco		
9:40 AM - 10:00 AM	**Drug Repurposing for Substance Use Disorders by Genome-Wide Association Studies and Real-World Data Analyses** Dongbing Lai Indiana University	9:40 AM - 9:50 AM	**Applications of Neural Networks in Chaperonin Generation for Complement Factor Renaturalization** Arhan Patel Matthew Fang Massachusetts Institute of Technology	9:40 AM - 10:00 AM	**Leveraging Clinical Biobanks and Genetics to Understand Sleep Apnea and Related Comorbidities** Brian Cade, Harvard Medical School
		9:50 AM - 10:00 AM	**MODE: High-Resolution Digital Dissociation with Deep Multimodal Autoencoder** Ayesha A. Malik University of Central Florida		
10:00 AM - 10:20 AM	**An Informatics Bridge to Improve the Design and Efficiency of Phase I Clinical Trials for Anticancer Drug Combinations** Lei Wang The Ohio State University	10:00 AM - 10:10 AM	**Empowering Confident Communication: Artificial Intelligence Based Detection of** David Li Thousand Oaks High School	10:00 AM - 10:20 AM	**Sleep Architecture Biomarkers of Psychiatric Disease** Shaun Purcell, Harvard Medical School
		10:10 AM - 10:20 AM	**Identifying Morin Hydrate as an Anti-Aging Drug with Machine Learning** Joanna Hou Princeton High School		

10:20 AM - 10:40 AM	***Coffee/Tea Break***	10:20 AM - 10:30 AM	***Coffee/Tea Break***		***Coffee/Tea Break***
		10:30 AM - 10:40 AM	**Deep Learning-Based Fall Detection for Autonomous Real-Time Emergency Notification: Integrating YOLO and Twilio** Daniel Zhou William Lyon Mackenzie Collegiate Institute		
10:40 AM - 11:00 AM	**Building an Explainable Graph Neural Network by sparse Learning for the Drug-Protein Binding Prediction** Yijie Wang Indiana University	10:40 AM - 10:50 AM	**Bulk and Spatial Single Cell Transcriptomic Analysis and Machine Learning based Disease Classification of ALS** Aayush Veerabhadran, Pioneer High School	10:40 AM - 11:00 AM	**Reimagining Sleep Medicine using AI-based Physiology-guided Digital Twins** Ankit Parekh, Icahn School of Medicine at Mount Sinai
		10:50 AM - 11:00 AM	**Evaluating the Prognostic Value of Mutational Signatures in Small-Cell Lung Cancer Through Data-Driven Threshold Optimization and Signature Assignment** Rishabh Garg Yale University		
11:00 AM- 11:20 AM	**Combining Genetics and Real-World Patient Data Fuel Ancestry-Specific Target and Drug**	11:00 AM- 11:10 AM	**LLM-Powered Web Agents and Their Impact on Automation** Jasmine Zhang	11:00 AM- 11:20 AM	**Y-Chromosome Loss in Cancer: Single-Cell Insights into Origins and Consequences**

	Discovery in Alzheimer's Disease Yuan Hou Cleveland Clinic		Carmel High School		Jun Xia, Texas A&M University
		11:10 AM-11:20 AM	**AI-Enhanced Aptamer Design: Addressing the Blood-Brain Barrier Challenge of Alzheimer's Disease Therapeutics** Christina He Carnegie Vanguard High School		
11:20 AM - 11:40 AM	**Identifying Repurposable Treatments in Patient Subpopulations** Pengyue Zhang, Indiana University	11:20 AM - 11:30 AM	**Application of a Quantitative Systems Pharmacology Model to Predict Amyloid Plaque Reduction in Alzheimer's Disease Therapies** Alex Mi Carmel High School	11:20 AM - 11:40 AM	**Computational Techniques for Deciphering Cancer Genomics and the Tumor Microenvironment at Single-Cell Resolution** Jinzhuang Dou, The University of Alabama at Birmingham
		11:30 AM - 11:40 AM	**Illuminating Dark Proteins with AI** Andy Dong Illinois Junior STEM Society		
11:40 AM - 11:50 AM	**Pan-Cancer Analysis of the Immune Microenvironment's Role in Tumor Genomic Evolution** Elaine Li, Arizona State University	11:40 AM - 11:50 AM	**Optimizing Torch-MISA for Efficient Signal Separation in IVA of fMRI via Definitive** Orit Yohannes Georgia State University	11:40 AM - 12:00 PM	**Distinct Signatures of Tumor-Associated Macrophages in Shaping Immune Microenvironment and Patient Prognosis**

		11:50 AM - 12:00 PM	**Distinct DNA Methylation Patterns in Alzheimer's Disease Brain Tissue** Raymond Cheng Arizona State University		Chongming Jiang, Terasaki Institute for Biomedical Innovation
12:00 PM - 1:30 PM		**Lunch Break**			
1:30 PM - 2:10 PM		**Keynote Speaker** (Room 320) **Qing Nie, PhD** **University of California, Irvine**			
CONCURRENT SESSIONS/WORKSHOPS					
Room: **320**		Room: **301**		Room: **350**	
AI and Applications for Better Understanding Disease Mechanisms Chair: Xubo Song		**Advances in Bioinformatics** Chair: Juexin Wang		**Integrative Genomics and Epigenomics to Link GWAS Variants to Function** Chairs: Hongbo Liu, Kaixiong Ye	
2:20 PM - 2:40 PM	**Reprogramming Protein Language Models for Protein Function Annotation and Engineering** Yunan Luo, Georgia Institute of Technology	2:20 PM - 2:40 PM	**Eminent Scholar Presentation** Yufeng Shen, Columbia University	2:20 PM - 2:40 PM	**Precision Nephrology: The Role of Genetics in Kidney Health** Atlas Khan, Columbia University
2:40 PM - 3:00 PM	**MARVEL: Microenvironment Annotation by Supervised Graph Contrastive Learning** Yuying Xie, Michigan State University	2:40 PM - 2:50 PM	**Benchmarking Cellular Deconvolution Algorithms to Predict Cell Proportions: A Literature Review** Ayesha Malik, University of Central Florida	2:40 PM - 3:00 PM	**Unraveling the Molecular Heterogeneity of Severe Acute Malnutrition: Multi-omic Insights** Yixing Han, National Institutes of Health

		2:50 PM - 3:00 PM	**Landscape of gene essentiality in cancer cell death pathways** Shangjia Li, The Ohio State University		
3:00 PM - 3:20 PM	**Leveraging AI for Characterizing Pediatric Cancer** Shibiao Wan, University of Nebraska Medical Center	3:00 PM - 3:20 PM	**Technology Session** Chair: Yu-Chiao Chiu **Boosting Analysis Pipeline Efficiency in Bioinformatics Through Snakemake** Shunian Xiang, Admera Health	3:00 PM - 3:20 PM	**Leveraging chromatin accessibility data to understand complex traits** Siming Zhao, Dartmouth College
3:20 PM -3:40 PM		***Coffee/Tea Break***			
3:40 PM - 4:00 PM	**Deep Learning Models for Image Enhancement, Translation, and Harmonization** Xubo Song, Oregon Health & Science University	3:40 PM - 4:00 PM	**Spatial Transcriptomics at Scale with Stereo-seq: Big Data for Impactful Science** Yongfu Wang, Complete Genomics	3:40 PM - 4:00 PM	**Integrative Genomics and Epigenomics Reveal Functions of Non-Coding Variants** Hongbo Liu, University of Rochester
4:00 PM - 4:20 PM	**Advancing AI for Individualized Diagnosis and Prognosis: From Prenatal Heart Defects to Prostate Cancer Survival** Jieqiong Wang, University of Nebraska	4:00 PM - 4:20 PM	**Access the Full Richness of Biological Complexity with Single Cell and Spatial Multiomics from 10x Genomics** Nicole Jaymalin, 10x Genomics	4:00 PM - 4:20 PM	**Mechanistic Annotation of GWAS Loci for Circulating Fatty Acids by Single-Cell Omics and CRISPR Screens** Huifang Xu, University of Georgia
4:20 PM - 4:40 PM	**Large Scale, AI-Enabled, Spatial Signal Processing of Breast Cancer**	4:20 PM - 4:40 PM	**Directed Evolution of Molecular Enzymes Empowers NGS Library Preparation**	4:20 PM - 4:40 PM	**Linking Rare Non-Coding Variants Associated with Human Longevity to**

	Pathology Identifies Consensus Tissue Structures Related to Biology and Outcomes Jordan Krull, The Ohio State University		Robin Song, Yeasen Biotechnology Co., Ltd.		**Cellular Senescence via Integrative Functional Genomic Approaches** Jiping Yang, Columbia University
4:40 PM - 5:00 PM	**Evaluate, Standardize, and Optimization Bioinformatics Software Documentation Using AI-Agents** Shaopeng Gu, The Ohio State University	4:40 PM - 5:00 PM	**Uncover Cellular Heterogeneity with Advanced Single Cell Multi-Omics Approaches** Julie Laliberte, Singleron Biotechnologies Inc., USA	4:40 PM - 5:00 PM	**Identification of Replicative Aging and Inflammatory Aging Signatures via whole-Genome CRISPRi Screens and GWAS Meta-analysis** Xueqiu Lin, Fred Hutchinson Cancer Center
5:00 PM - 6:00 PM	**Poster Session II (Atrium)**				

Tuesday, August 5th, 2025

8:00 AM - 5:00 PM	**Registration**				
8:30 AM - 9:10 AM	**Keynote Speaker** (Room 320) **Julie A. Johnson, PharmD** **The Ohio State University**				
CONCURRENT SESSIONS/WORKSHOPS					
Room: **320**		Room: **301**		Room: **350**	
Data Science Solutions for Spatial Transcriptomics Chair: Travis Johnson		**Computational Omics for Precision Medicine and Drug Discovery** Chairs: Bin Chen, Qian Li		**Integrative Bioinformatics for Translational and Precision Medicine** Chairs: Yuan Liu, Shilin Zhao	
9:20 AM - 9:40 AM	**Eminent Scholar Presentation**	9:20 AM - 9:40 AM	**Protein Language Model ESM3**	9:20 AM - 9:40 AM	**A Novel Immune-Related Risk**

	Xiang Zhou, Yale University		**Enables Superior Prediction of Complex Variant Effects Across ClinVar and DMS Benchmarks** Xiaoming Liu, University of South Florida		**Stratification Model to Predict Prognosis, Immunotherapy and Chemotherapy Response for Neuroblastoma** Xiaohui Zhan, Chongqing Medical University
9:40 AM - 10:00 AM	**SpatialGE: An Interactive Web Platform for Accessible and Reproducible Spatial Transcriptomics Analysis** Xiaoqing Yu, Moffit Cancer Center	9:40 AM - 10:00 AM	**Massive Labeled Transcriptomics as a Resource of Transcriptome Representation Learning and Drug Discovery** Bin Chen, Michigan State University	9:40 AM - 10:00 AM	**The Impact of HLA Diversity on Immune Cell Composition, Tumor Mutation Burden, and Cancer Survival** Shilin Zhao, Vanderbilt University
10:00 AM - 10:20 AM	**Spatial Resolved Gene Regulatory Networks Analysis** Zhana Duren, Indiana University School of Medicine	10:00 AM - 10:20 AM	**Generative AI for Human Genetics and Functional Genomics** Xinghua (Mindy) Shi, Temple University	10:00 AM - 10:20 AM	**Horizontal Gene Transfer Networks Reveal Resistance of Plasmid-Mediated Communication in Antibiotic Exposure** Lijia Che, City Univesity of Hong Kong
10:20 AM - 10:40 AM		***Coffee/Tea Break***			
10:40 AM - 11:00 AM	**Identifying Key Regulators of Amyloid Beta Clearance from Single Cell Spatial Transcriptomics using Generalized Linear Mixed Effect Models**	10:40 AM - 11:00 AM	**Distinct Mutational Profiles in Primary Sclerosing Cholangitis-Associated Cholangiocarcinoma Compared to *de novo* Cholangiocarcinoma**	10:40 AM - 11:00 AM	**Boolean Network Modeling-Guided Identification of FDA-Approved Drug Combinations for Targeted Treatment Strategies in Head and Neck Cancer**

	Debolina Chatterjee, Indiana University School of Medicine		Shulan Tian, Mayo Clinic		Pranabesh Bhattacharjee, Texas A&M University
11:00 AM - 11:20 AM	**Leveraging Spatial Transcriptomics of Brain Tissue in Neurological Diseases** Oscar Harari, The Ohio State University	11:00 AM - 11:20 AM	**High-Resolution Multi-Omic Dissociation of Brain Tumors with Multimodal Autoencoder** Qian Li, Ph.D., St. Jude Children's Hospital	11:00 AM - 11:20 AM	**Comparison of Nanopore Sequencing, MethylationEPIC Array, and EM-Seq for DNA Methylation Detection** Steven Brooks, Indiana University
11:20 AM - 11:40 AM	**A Statistical Framework to Improve the Design of Spatial Transcriptomics Experiments** Dongjun Chung, The Ohio State University	11:20 AM - 11:40 AM	**CoMPaSS: A Computational Pipeline for Cross-Platform Concordance Assessment and Navigating Study Design in Microbiome Research** Xi Qiao, University of Utah	11:20 AM - 11:40 AM	**A Hierarchical Adaptive Diffusion Model for Flexible Protein-Protein Docking** Rujie Yin, Texas A&M University
11:40 AM - 12:00 PM	**Integrative Modeling of Gene Expression and Histology via Cross-Modal Alignment and Multi-Scale Graph Inference** Chao Chen, Stony Brook University	11:40 AM - 12:00 PM	**SEHI-PPI: An End-to-End Sampling-Enhanced Human-Influenza Protein-Protein Interaction Prediction Framework with Double-View Learning** Rui Yin, University of Florida	11:40 AM - 12:00 PM	**"Frustratingly Easy" Domain Adaptation for Cross-Species Transcription Factor Binding Prediction** Mark Maher Ebeid, University of Pittsburgh
12:00 PM - 12:20 PM	**Utilizing Deep Transfer Learning to Identify High Risk**	12:00 PM - 12:20 PM	**Cyclin D1 Induces Epigenetic and Transcriptional**	12:00 PM - 12:10 PM	**Multi-Omic Analysis Integrating Global Transcriptional and**

	Subpopulations of Cells in Single Cell Spatial Omics Data Travis Johnson, Indiana University School of Medicine		**Alterations in Multiple Myeloma with t(11;14)(q13;q32)** Huihuang Yan, Mayo Clinic		**Post-Transcriptional Profiles Reveals Predominant Role of Post-Transcriptional Control in Three Human Cell Lines** Alexander Krohannon, Indiana University
12:20 PM - 1:30 PM		**Lunch Break**			
CONCURRENT SESSIONS					
Room: **320**		Room: **301**		Room: **350**	
AI and Machine Learning in Translational Genomics Chairs: Huihuang Yan, Yixing Han		**Data-Driven Insights into Disease Modeling** Chairs: Shulan Tian, Joseph McElroy		**Flash Talks** Chair: Zhifu Sun	
1:30 PM - 1:50 PM	**Adaptive Chebyshev Graph Neural Network for Cancer Gene Prediction with Multi-Omics Integration** Sa Li, Oakland University	1:30 PM - 1:50 PM	**Compositional Bayesian Co-Clustering of DTI biomarkers with Clinical Measures for Enhanced Prediction of Parkinson Disease Severity** Chandrajit Bajaj, University of Texas at Austin	1:30 PM - 1:40 PM	**A Multimodal Vision Transformer using Fundus and OCT Images for Interpretable Classifications of Diabetic Retinopathy** Shivum Telang, North Allegheny High School
				1:40 PM - 1:50 PM	**Abnormal ERV Expression and its Clinical Relevance in Colon Cancer** Aditya Bhagwate, Mayo Clinic
1:50 PM - 2:10 PM	**A Generative Imputation Method for Multimodal Alzheimer's**	1:50 PM - 2:10 PM	**Latent Factor Modeling Reveals Unexpected Spatial Heterogeneity in**	1:50 PM - 2:00 PM	**From Bench to Insight: Rapid Pathogen Genomic Surveillance**

	Disease Diagnosis Reihaneh Hassanzadeh, Georgia Institute of Technology		**Human Alzheimer's Disease Brain Transcriptomes** Hu Chen, Baylor College of Medicine		**Workflow for SARS-CoV-2 and Emerging Pathogens** Venkat Sundar Gadepalli, The Ohio State University
				2:00 PM - 2:10 PM	**LoRA-BERT: a Natural Language Processing Model for Robust and Accurate Prediction of long non-coding RNAs** Nicholas Jeon, Texas A&M University
2:10 PM - 2:30 PM	**A User-Friendly R Shiny App for Predicting Surface Protein Abundance from scRNA-seq Expression Using Deep Learning in blood cells** Yidong Chen, University of Texas Health San Antonio	2:10 PM- 2:30 PM	**DuAL-Net: A Hybrid Framework for Alzheimer's Disease Prediction from Whole-Genome Sequencing via Local SNP Windows and Global Annotations** Eun Hye Lee, Indiana University School of Medicine	2:10 PM - 2:20 PM	**ICM-MD: Integrating TM-Specific Features and MD-Derived Structures for Accurate Prediction of Inter-Chain Contacts in Alpha-Helical Transmembrane Homodimers** Bander Almalki, University of Delaware
				2:20 PM - 2:30 PM	**OmicsSankey: Crossing Reduction of Sankey Diagram on Omics Data** Bowen Tan, City University of Hong Kong
2:30 PM - 2:50 PM	**HELP-TCR Harmonized Explainable Language Processing toolkit**	2:30 PM - 2:50 PM	**Resolving Gene Heterogeneity in DEG Analysis: A Novel Pipeline for**	2:30 PM - 2:40 PM	**TCR Convergence as a Proxy for Tumor-Specific Immunity in HSV1-Positive rGBM**

	for T-Cell antigen Receptor repertoires. Yulyana Kalesnik, University of Lodz		**Precision Genomics** Jiasheng Wang, Baylor College of Medicine		**Patients Treated with CAN-3110** Ayse Selen Yilmaz, The Ohio State University
				2:40 PM - 2:50 PM	**Vritra: A Streamlined Pipeline for Species-resolved Functional Profiling of Target Genes in Microbiome Data** Boyan Zhou, New York University
2:50 PM - 3:10 PM	**Efficient and Valid Large Molecule Generation via Self-supervised Generative Models** Doyoung Kwak, Texas A&M University	2:50 PM - 3:10 PM	**Multimodal Imaging and Cell-Free DNA Methylation Analysis for Noninvasive Lung Cancer Diagnosis** Ran Hu, University of California – Los Angeles	2:50 PM - 3:00 PM	**Transcriptomic Signatures in Nucleus Accumbens, Midbrain, Pre-Frontal Cortex, and Amygdala Regions Identifies Shared and Unique Gene Signatures for Substance Use** Avinash Veerappa, University of Nebraska
				3:00 PM - 3:10 PM	**Endophenotype-based in silico network medicine prediction and real-world patient data validation identify potential drug combinations for Alzheimer's disease** Zhendong Sha, Cleveland Clinic Genome Center
3:10 PM - 3:30 PM		***Coffee/Tea Break***			

3:30 PM - 3:50 PM	**DG-scRNA: Deep Learning with Graphic Cluster Visualization to Predict Cell Types of Single Cell RNAseq Data** Yimin Liu, The Ohio State University	3:30 PM - 3:50 PM	**Multidimensional Impact of Microbiota Absence on Thymic T Cell Development in Mice: A Study Based on Single-Cell and Spatial Transcriptomics** Yifei Sheng, University of Chinese Academy of Sciences	3:30 PM - 3:40 PM	**VaxLLM: An End-to-End Framework Leveraging a Fine-Tuned Large Language Model for Automated Vaccine Annotation and Database Integration** Xingxian Li, University of Michigan
				3:40 PM - 3:50 PM	**Supervised and Unsupervised Classification with Feature Selection for Single-Cell RNAseq Based on an Artificial Immune System** Dawid Krawczyk, University of Lodz
3:50 PM - 4:10 PM	**A Machine Learning–Enhanced Pipeline for Detecting Disruption of Transcription Termination (DoTT) in RNA-Seq Data** Michael Levin, Temple University	3:50 PM - 4:10 PM	**The Drug Overdose Surveillance in Ohio: What We Can See with the Geospatial Shared Component Analysis of the Urine Drug Test Results** Joanne Kim The Ohio State University	3:50 PM - 4:00 PM	**Multi-Modal Domain-Specific Foundation Model for Prostate Cancer Explanation: Utilizing H&E Image and Spatial Proteomics** Kyeong Joo Jung, The Ohio State University
				4:00 PM - 4:10 PM	**Static and Dynamic Cross-Network Functional Connectivity Shows Elevated Entropy in Schizophrenia Patients** Natalia Maksymchuk, TReNDS

4:10 PM - 4:30 PM	**THANOS: An AI Pipeline for Engineering Antibodies** Arnav Solanki, University of Texas Health Science Center at Houston	4:10 PM - 4:30 PM	**Telehealth Utilization and Patient Experiences: The Role of Social Determinants of Health Among Individuals with Hypertension and Diabetes** Jiancheng Ye	4:10 PM - 4:20 PM	**A Network-Based Systems Genetics Framework Identifies Pathobiology and Drug Repurposing in Parkinson's Disease** Lijun Dou, Cleveland Clinic Genome Center
				4:20 PM - 4:30 PM	**MetaphorPrompt2-A Structure and Function Focused Approach for Extracting Causal Events from Biological Text** Parth Patel, University of Texas at San Antonio
4:30 PM - 4:40 PM	**GRN-Integrated Heterogeneous Attentive Graph Autoencoder for Cell-Cell Interaction Reconstruction from Spatial Transcriptomics** Aiwei Yang, Beijing Normal-Hong Kong Baptist University	4:30 PM - 4:40 PM	**AutoRADP: An Interpretable Deep Learning Framework to Predict Rapid Progression for Alzheimer's Disease and Related Dementias Using Electronic Health Records** Qiang Yang, University of Florida	4:30 PM - 4:40 PM	**DisSubFormer: A Subgraph Transformer Model for Disease Subgraph Representation and Comorbidity Prediction** Ashwag Altayyar, University of Delaware
5:00 PM - 5:45 PM		Award Presentation (Room 320)			
5:45 PM - 6:00 PM		Closing Remarks (Room 320)			

Keynote Speakers

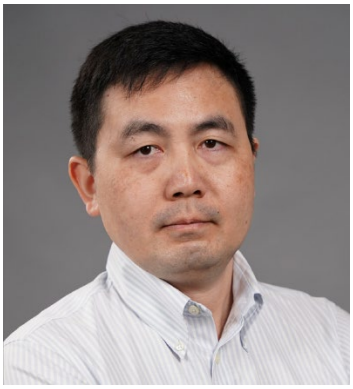
Keynote Speaker
Veera Baladandayuthapani, Ph.D.
August 3rd
1:40 PM - 2:20 PM
Room: 320

Dr. Veera Baladandayuthapani is currently Jeremy M.G. Taylor Collegiate Professor and Chair in the Department of Biostatistics at University of Michigan (UM), where he also serves as the Associate Director of Quantitative Data Sciences and Director of the Cancer Data Science Shared Resource at UM Rogel Cancer Center. He obtained his Ph.D. in Statistics from Texas A&M University (in 2005), M. A. in Statistics from University of Rochester (in 2000) and BSc in Mathematics from Indian Institute of Technology, Kharagpur in 1998. He joined UM in Fall 2018 after spending 13 years in the Department of Biostatistics at University of Texas MD Anderson Cancer Center, Houston, Texas, where was a Professor and Institute Faculty Scholar and held adjunct appointments at Rice University, Texas A&M University and UT School of Public Health. His research explores the potential of Bayesian probabilistic models and machine learning methods to assist in medical and health sciences. These methods are motivated by large and complex datasets such as high-throughput genomics, epigenomics, transcriptomics and proteomics as well as high-resolution neuro- and cancer- imaging. A special focus is on developing integrative and spatial models combining different sources of data for biomarker discovery and clinical prediction to aid precision/translational medicine. His work has resulted in 160+ papers published in top statistical, biostatistical, bioinformatics, biomedical & oncology journals. He has also co-authored a book on Bayesian analysis of gene expression data. He has received several prestigious awards that include being selected as Myrto Lefkopoulou Distinguished Lectureship from Harvard School of Public Health; H. O. Hartley award from the Department of Statistics at Texas A&M University; Theodore. G. Ostrom from Washington State University; MD Anderson Faculty Scholar Award; Young Investigator Award from the International Indian Statistical Association (IISA) and Editor's Invited Paper for Biometrics, a top biostatistics journal and the flagship journal of the International Biometrics Society. He is a fellow of the American Association for Advancement in Science and the American Statistical Association and an elected member of the International Statistical Institute. He serves or has served on the Editorial board for major bio/-statistical journals such as Journal of American Statistical Association, Annals of Applied Statistics and Biometrics.

Title: Artificially Intelligent BioSpatial Modeling: Decoding Tumor Geography

Abstract: The tumor microenvironment (TME) is increasingly recognized as a critical frontier in cancer research, revealing how the spatial organization and dynamic interactions among diverse cell populations

govern immune responses, tumor progression, and therapeutic outcomes. Recent advances in spatial profiling technologies—including spatial multiplex imaging, spatial transcriptomics, and digital pathology—have enabled unprecedented high-resolution characterization of these complex ecosystems. Yet these data introduce significant computational and statistical challenges: intricate spatial dependencies, substantial heterogeneity within and across samples, and non-conformable spaces that complicate integrative, population-level analyses. I will discuss my perspective on how the conflation of AI techniques and biologically-informed rigorous statistical modeling can address these challenges and unlock actionable biological insights. Specifically, I will discuss frameworks for modeling spatially varying genomic networks and transcriptional programs, approaches for quantifying intercellular interactions within the TME, and strategies for linking spatial features to patient-specific clinical outcomes. The utility and translational potential of these methods will be illustrated through multiple case studies spanning diverse cancer types.



Keynote Speaker
Jiang Bian, Ph.D.
August 4th
8:30 AM - 9:10 AM
Room: 320

Dr. Bian specializes in biomedical informatics and health data science, interdisciplinary fields focused on leveraging data, information, and knowledge to drive scientific discovery, problem-solving, and decision-making, all aimed at improving human health. Dr. Bian brings extensive experience in developing real-world data infrastructure, informatics tools, and systems, as well as applying advanced AI and data science methods to analyze and interpret multimodal clinical and biomedical data. Dr. Bian serves as Chief Data Scientist at Regenstrief, Chief Data Scientist at IU Health, and Associate Dean of Data Science among other leadership roles.

Title: Real-World Data to Real-World Evidence: Successes, Challenges, and Opportunities

Abstract: This presentation delves into the methods and tools enable the transformation of real-world data (RWD) into actionable real-world evidence (RWE). It emphasizes the central role of data science in addressing the inherent challenges of working with large-scale, messy, and heterogeneous data sources such as EHRs and claims data. Specific case studies—including target trial emulation for evaluating GLP-1 receptor agonists (GLP-1RAs) and outcomes in cancer risk and survivorship—demonstrate how advanced analytical frameworks and causal inference techniques can generate RWD complement randomized controlled trials. Also study design issues and target trial emulation.

Keynote Speaker
Qing Nie, Ph.D.
August 4th
1:30 PM - 2:10PM
Room: 320

Dr. Qing Nie is a University of California Presidential Chair and a Distinguished Professor of Mathematics and Developmental & Cell Biology at University of California, Irvine. Dr. Nie is the director of the *NSF-Simons Center for Multiscale Cell Fate Research* jointly funded by NSF and the Simons Foundation – one of the four national centers on mathematics of complex biological systems. In research, Dr. Nie uses systems biology and data-driven methods to study complex biological systems with focuses on single-cell analysis, multiscale modeling, cellular plasticity, stem cells, embryonic development, and their applications to diseases. Dr. Nie has published more than 250 research articles, including more than 50 papers in journals such as Nature, Science, Nature Methods, PNAS, Nature Machine Intelligence, Cancer Cells, Nature Communications. In training, Dr. Nie has supervised more than 60 postdoctoral fellows and PhD students, with many of them working in academic institutions. In 2025, Dr. Nie was ranked #1 by ScholarGPS based on citation metrics as *Highly Ranked Scholar* in two areas: a) Single-cell transcriptomics & b) Transcriptomics technologies for *Prior Five Years.* Dr. Nie has been recognized by various professional societies for his interdisciplinary research achievements. Dr. Nie is a fellow of the *American Association for the Advancement of Science (AAAS)*, *American Physical Society (APS)*, *Society for Industrial and Applied Mathematics (SIAM),* and *American Mathematical Society (AMS)*.

Title: Systems Learning of Single Cells

Abstract: Cells make fate decisions in response to dynamic environments, and multicellular structures emerge from multiscale interplays among cells and genes in space and time. While single-cell omics data provides an unprecedented opportunity to profile cellular heterogeneity, the technology requires fixing the cells, often leading to a loss of spatiotemporal and cell interaction information. How to reconstruct temporal dynamics from single or multiple snapshots of single-cell omics data? How to recover interactions among cells, for example, cell-cell communication from single-cell gene expression data? I will present a suite of our recently developed computational methods that learn the single-cell omics data as a spatiotemporal and interactive system. Those methods are built on a strong interplay among systems biology modeling, dynamical systems approaches, machine-learning methods, and optimal transport techniques. The tools are applied to various complex biological systems in development, regeneration, and diseases to show their discovery power. Finally, I will discuss the methodology challenges in systems learning of single-cell data.

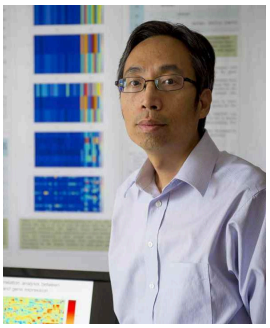
Keynote Speaker
Julie A. Johnson, Pharm.D.
August 5th
8:30 AM - 9:10 AM
Room: 320

Julie A. Johnson, Pharm.D. is the Dr. Samuel T and Lois Felts Mercer Professor of Medicine and Pharmacology at The Ohio State University's Colleges of Medicine and Pharmacy. She is the Director of OSU's Clinical and Translational Science Institute, Associate Dean for Research (Medicine) and Associate Vice President for Research at OSU. Dr. Johnson's research focuses on pharmacogenomics discovery and implementation and documenting outcomes of precision medicine approaches in clinical practice. She is an internationally recognized leader in clinical pharmacology, pharmacogenomics and genomic medicine, with over 340 peer reviewed original publications and over $55M in research funding as principal investigator, excluding the CTSA award. From 2015-2018 she was named a Clarivate Analytics Highly Cited Researcher, an accomplishment of about 1 in 1000 scientists globally. Dr. Johnson has received numerous awards and honors, including election to the National Academy of Medicine and election as fellow of the American Association for the Advancement of Science, and three other societies, along with top research awards from several multiple organizations. She was recently appointed to the National Academie's Forum on Drug Discovery, Development and Translational. She has received teaching awards from the University of Tennessee and the University of Florida and mentoring awards from the American Society for Clinical Pharmacology and Therapeutics and the American Heart Association.

Title: Using real world evidence to advance pharmacogenomics

Abstract: This presentation will cover the opportunities to advance discoveries and validation of genetic associations with efficacious or adverse reponses to drug therapy, including discussion of datasets in which this is possible. There will then be specific data presented from studies evaluating associations between CYP2D6 genotype and drug interactions in patients treated with opioid therapy and emergency department visits based on CYP2D6 phenotype status.

Eminent Scholar Talks



Eminent Scholar Talk

Yu-Ping Wang, Ph.D.
August 3rd
2:30 PM - 2:50 PM
Room: 301

Dr. Yu-Ping Wang received the BS degree in applied mathematics from Tianjin University, China, in 1990, and the MS degree in computational mathematics and the PhD degree in communications and electronic systems from Xi'an Jiaotong University, China, in 1993 and 1996, respectively. After his graduation, he had visiting positions at the Center for Wavelets, Approximation and Information Processing of the National University of Singapore and Washington University Medical School in St. Louis. From 2000 to 2003, he worked as a senior research engineer at Perceptive Scientific Instruments, Inc., and then Advanced Digital Imaging Research, LLC, Houston, Texas. In the fall of 2003, he returned to academia as an assistant professor of computer science and electrical engineering at the University of Missouri-Kansas City. He is currently a Professor of Biomedical Engineering, Computer Sciences, Neurosciences, and Biostatistics & Data Sciences at Tulane University. Dr. Wang's recent effort has been bridging the gap between biomedical imaging and genomics, where has over 250 journal publications. Dr. Wang is a fellow of AIMBE and won the 2022 Tulane Convergence Award for his effort in bridging gaps between science, engineering and biomedicine. He has served for numerous program committees and NSF and NIH review panels and is currently an associate editor for J. Neuroscience Methods, IEEE/ACM Trans. Computational Biology and Bioinformatics (TCBB) and IEEE Trans. Medical Imaging (TMI). More about his research can be found at his lab website: http://www.tulane.edu/~wyp/

Title: Integration of brain imaging and genomics with interpretable multimodal collaborative learning

Abstract: Recent years have witnessed the convergence of multiscale and multimodal brain imaging and omics techniques, showing great promise for systematic and precision medicine. In the meantime, they bring significant data analysis challenges when integrating and mining these large volumes of heterogeneous datasets. In this work, we first introduce a linear collaborative learning model to combine both regression and correlation analysis such as CCA. To further capture complex interactions both within and across modalities, we develop an interpretable multimodal deep learning-based integration model to perform heterogeneous data integration and result interpretation simultaneously. The proposed model can generate interpretable activation maps to quantify the contribution of imaging or omics features. Moreover, the estimated activation maps are class-specific, which can therefore facilitate the identification of biomarkers. Finally, we apply and validate the model in the study of brain development with integrative

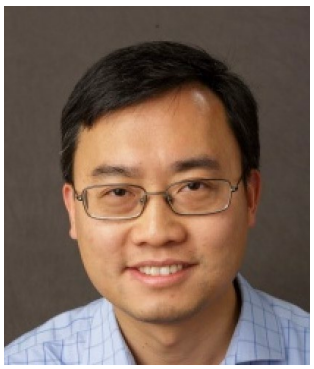
analysis of multi-modal brain imaging and genomics data. We demonstrate its successful application to both the classification of cognitive function groups and the discovery of underlying genetic mechanisms.

Eminent Scholar Talk
Kaifu Chen, Ph.D.
August 4th
9:20 AM -9:40 AM
Room: 320

Kaifu Chen, PhD, is currently an Associate Professor in the Pediatrics Department of Harvard Medical School and the Director of the Computational Biology Program in the Cardiology Department of Boston Children's Hospital. His research focused on understanding the expression regulation of cell identity genes by the binding of transcription factors to enhancers, histone modifications on nucleosomes, 3D genome folding, and cell-cell signaling in a tissue environment. His lab conducts bioinformatics analysis of multiomics data to understand these molecular mechanisms and develop AI models to uncover cell identity regulators based on these mechanisms, with a particular interest in applications to cardiovascular diseases and cancers.

Title: AI modeling of Cell Identity regulation in biological development and diseases

Abstract: Precise regulation of cell identity is fundamental in biological development and diseases, yet the underlying mechanisms remain incompletely understood. We present an integrated AI-driven framework that models cell identity regulation by combining large-scale single-cell transcriptomics, epigenomics, and regulatory network inference. Our analysis of millions of single cells across various human tissue types revealed that cell identity heterogeneity is shaped by both chromatin epigenetics and microenvironmental context. We developed a computational pipeline, MEBOCOST, to systematically characterize cell–cell communication and identified tissue-specific signaling patterns that modulate cell identity. To resolve the regulatory logic of cell identity, we introduced SCIG and CEFCIG—machine learning frameworks that uncover cell identity genes (CIGs) and their master regulators using genetic, epigenetic, and expression signatures. Applying this framework, we identified MECOM as a key regulator of endothelial lineage specification and demonstrated its role in enhancer looping, VEGF signaling, and angiogenesis. Together, our approach offers a unified model of cell identity regulation, advancing our understanding of tissue development, diseases, and regenerative interventions.



Eminent Scholar Talk
Yufeng Shen, Ph.D.
August 4th
2:20 PM - 2:40 PM
Room: 301

Yufeng Shen is an Associate Professor of Systems Biology and Biomedical Informatics at Columbia University. He received his B.Sc. in biochemistry from Peking University and his Ph.D. in computational biology from Baylor College of Medicine. At Baylor, he led the analysis of the first human genome sequenced by next-generation technologies. His research group is currently working on predicting effect of genetic variation using statistical and machine learning methods and to apply genomics and computational biology in genetic studies of human diseases. His group developed CANOES (Backenroth et al 2014) for calling copy number variants from exome sequencing data, gMVP (Zhang et al 2022), SHINE (Fan et al 2023), and MisFit (Zhao et al 2025) for predicting pathogenicity and fitness effect of protein variants. They discovered that epigenomic patterns in tissues under normal conditions are associated with risk genes of developmental disorders (Han et al 2018). In addition, his research led to the discovery of novel risk genes of congenital heart disease (Homsy et al 2015), congenital diaphragmatic hernia (Qi et al 2018, 2024), and autism (Zhou et al 2022).

Title: Representation and prediction of the impact of protein mutations

Abstract: Accurate prediction of genetic effect of missense variants is fundamentally important for disease gene discovery, clinical genetic diagnosis, personalized treatment, and protein engineering. Commonly used computational methods predict pathogenicity, which does not capture the quantitative impact on fitness in human. We developed a method, MisFit, to estimate selection coefficient of missense variants. MisFit jointly models the effect at a molecular level (D) and a population level (selection coefficient, S), assuming that in the same gene, missense variants with similar D would have similar S. We trained it by maximizing the probability of observed germline variant allele counts in 234,992 individuals of European ancestry. We show that S is informative in predicting allele frequency across ancestries and consistent with the fraction of de novo mutations observed in sites under strong selection. Further, S outperforms previous methods in prioritizing de novo missense variants in individuals with neurodevelopmental disorders. Finally, we show that predicted D and S are consistent with functional readout of deep mutational scan experiments of clinically important genes.

Eminent Scholar Talk
Xiang Zhou, Ph.D.
August 5th
9:20 AM -9:40 PM
Room: 320

Xiang Zhou is a Professor in the Department of Statistics and Data Science at Yale University. He earned a BS in Biology from Peking University in 2004, followed by an MS in Statistics (2009) and a PhD in Neurobiology (2010) from Duke University. He completed postdoctoral training in the Department of Human Genetics at the University of Chicago (2010–2013), where he later served as the Williams H. Kruskal Instructor in the Department of Statistics (2013–2014). Dr. Zhou joined the Department of Biostatistics at the University of Michigan as an Assistant Professor in 2014. He held the John G. Searle Assistant Professorship from 2018 to 2019 and was promoted to Associate Professor in 2019 and to full Professor in 2023. He served as Assistant Director of Precision Health (2022–2025) and, in 2025, became Assistant Director of Artificial Intelligence and Digital Health Innovation (AI&DHI). He joined Yale University in 2025. Dr. Zhou is a Fellow of the American Statistical Association and the recipient of the 2024 Mid-career Biosciences Faculty Achievement Recognition (MBioFAR) Award and the 2025 ICIBM Eminent Scholar Award. He is a standing member of the NIH MRAA Study Section and serves as an Associate Editor for PLOS Genetics, Journal of the American Statistical Association, and Annals of Applied Statistics. In 2024, he was Program Chair for the Section on Statistics in Genomics and Genetics of the American Statistical Association. His research centers on genomic data science, with a focus on developing advanced statistical and machine learning methods, including deep learning and AI tools, for the analysis of large-scale, high-dimensional genetic and genomic data. His work spans a range of application areas, including genome-wide association studies, single-cell sequencing, and spatial multi-omics.

Title: Statistical Methods for Single Cell Spatial Transcriptomics

Abstract: Spatial transcriptomics comprises a transformative set of genomic technologies that enable the measurement of gene expression with spatial localization information in tissue sections or cell cultures. In this talk, I will present several statistical methods recently developed by our group for analyzing spatial transcriptomics data. These include SPARK, a method for rigorous statistical detection of spatially variable genes (SVGs); SPARK-X, a fast and scalable approach for detecting SVGs in large-scale spatial transcriptomic studies; SpatialPCA, which enables spatially informed dimension reduction; CARD, a method for spatially guided cell type deconvolution; IRIS, a framework that integrates external single-cell data to support scalable spatial domain detection; and BASS, a hierarchical Bayesian model designed for multi-scale and multi-sample spatial transcriptomic analysis. Together, these methods provide a comprehensive toolkit for advancing the analysis and interpretation of complex spatial transcriptomic datasets.

Workshop – Genomics and Translational Bioinformatics Working Group

August 3rd
8:30 AM – 11:30 AM
Room: 320

Chairs: Ece Uzun, Wenyu Song

Title: Calibration of computational prediction tools for improved clinical variant classification and interpretation

Author list: Vikas Pejaver[1,2]
Detailed Affiliations:
[1]Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA;
[2]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

Abstract: The classification of genetic variants as being pathogenic or not is essential to the proper and timely diagnosis of genetic disorders. In 2015, the American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP) provided formal guidance on how to weight and integrate different lines of evidence (e.g., population, functional, computational, among others) about a variant's pathogenicity or benignity towards its classification into one of five clinically relevant categories: *pathogenic*, *likely pathogenic*, *likely benign*, *benign* or of *uncertain significance*. As per these guidelines, evidence from computational and machine learning-based tools that use molecular and/or evolutionary information to predict the functional or phenotypic effects of variants, such as REVEL and AlphaMissense, was restricted to the weakest level of evidential strength (*Supporting*). However, the 2015 ACMG/AMP standards for the use of such predictors in variant classification and interpretation were based on developer-defined score thresholds, which are not always appropriate for the clinical context. Furthermore, these guidelines generally lacked quantitative support, predisposing them to being applied in non-standard ways that could lead to the misestimation of the evidential strength of *variant effect* predictors and inappropriate and/or inconsistent variant classification. To this end, we recently introduced a new calibration approach that standardizes scores from any variant effect predictor to 2015 ACMG/AMP evidence strength levels, from *Supporting* to *Very Strong*. Our approach estimates the local posterior probability of pathogenicity/benignity at a given prediction score to relate evidence strength as quantified by an existing Bayesian framework, with typical predictor performance measures such as precision and recall. Using carefully assembled independent data sets, we estimated score intervals corresponding to each level of evidential strength for pathogenicity and benignity for several different missense variant effect predictors and demonstrated that some predictors can reach up to *Moderate* and *Strong* evidence levels for a subset of variants. We then validated our proposed score intervals and estimated their impact on clinical variant classification using real-world data sets. Based on our findings, we recommended revisions of the ACMG/AMP criteria with respect to the use of variant effect predictors in the clinical context. Our work makes the use of such predictors in clinical variant classification and interpretation more rigorous and suggests a more prominent role for them in clinical genetic testing in the future.

Keywords: Variant effect predictor, calibration, variant classification, variant interpretation, ACMG/AMP guidelines, clinical genetic testing

Title: Opioid Prescriptions and Associated Patient Response: An Integrated Genetic Analysis Using Clinical Biobank

Author list: Wenyu Song[1, 10], Max Lam[2, 4], Ruize Liu[2, 3], Aurélien Simona[9], Scott G. Weiner[5, 10], Richard D. Urman[8], Kenneth J. Mukamal[6, 10], Adam Wright[7], David W. Bates[1, 10]

Detailed Affiliations:

1. Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA. 2. Stanley Center for Psychiatric Research, The Broad Institute of MIT and Harvard, Cambridge, MA, USA. 3. Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston MA, USA. 4. Neurogenomics Laboratory @ IMH Research Division, Institute of Mental Health, Singapore. 5. Department of Emergency Medicine, Brigham and Women's Hospital, Boston, MA, USA. 6. Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA. 7. Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA. 8. Department of Anesthesiology, The Ohio State University Wexner Medical Center, Columbus, OH, USA. 9. Division of Clinical Pharmacology and Toxicology, Geneva University Hospitals and Faculty of Medicine, Geneva, Switzerland. 10. Harvard Medical School, Boston, MA, USA

Abstract: Opioids are among the most powerful pain relievers available. Opioid drugs have been successfully used to treat both acute and chronic pain. While they can be effective for pain control especially acutely, opioids also have serious side effects and are prone to misuse and possible overdose. In 2023, there were 81,083 opioid related overdose deaths occurred in the United States. Genome-wide association studies (GWAS) have suggested that opioid related adverse events, including opioid use disorder (OUD), have strong genetic underpinnings. These genetic factors are located within genes that can affect efficacy, metabolism, and adverse effects of opioid drugs, which can in turn cause heterogeneous individual responses to drugs, including both pain levels and addiction. Electronic health records (EHRs) offer a largely untapped source of information to conduct genetic studies, which could facilitate the investigation of the genetic background of complex diseases and their comorbidities. EHR can be a particularly valuable data source for disorders like OUD that tend to be underrepresented in the cohort studies that comprise many genetic consortia.

We utilized patient-level clinical data from a large clinical biobank to develop opioid related phenotypes for genetic research. We first examined the genetic architecture of EHR-derived phenotypes of opioid use disorder (OUD) using GWAS and identified one novel significant OUD-associated locus on chromosome 4. Furthermore, we screened ~16 million rows of prescription records to develop codeine, one commonly prescribed opioid medicine, prescription-frequency phenotypes based on the number of recorded prescriptions for a given patient. Both low- and high-prescription counts were captured by developing 8 types of phenotypes with selected ranges of prescription numbers to reflect potentially different levels of opioid risk severity. We identified one significant locus associated with low-count codeine prescriptions (1, 2 or 3 prescriptions), while up to 7 loci were identified for higher counts, with a strong overlap across different thresholds. We identified 9 significant genomic loci with all-count phenotype. Further, using the polygenic risk approach, we identified a significant correlation between an externally derived polygenic risk score for opioid use disorder and numbers of codeine prescriptions. Our research provides a generalizable and clinical meaningful phenotyping pipeline for the genetic study of opioid-related risk traits.

Keywords: Electronic health record, Genome-wide association study, Opioid use disorder, Polygenic risk score, Opioid prescription phenotype

Title: Leveraging Deep Learning to Infer Cellular Dynamics

Author list: Shengyu Li[1,2,3,4], Pengzhi Zhang[1,2,3,4], Weiqing Chen[1,5], Lingqun Ye[1,2,3], Kristopher W. Brannan[2,3,4], Nhat-Tu Le[2,4], Jun-ichi Abe[6], John P. Cooke[2], Guangyu Wang[1,2,3,4,*]
Detailed Affiliations:
1. Center for Bioinformatics and Computational Biology, Houston Methodist Research Institute, Houston, TX, USA. 2. Center for Cardiovascular Regeneration, Houston Methodist Research Institute, Houston, TX, USA. 3. Center for RNA Therapeutics, Houston Methodist Research Institute, Houston, TX, USA. 4. Department of Cardiothoracic Surgery, Weill Cornell Medicine, Cornell University, New York, NY, USA. 5. Department of Physiology, Biophysics & Systems Biology, Weill Cornell Graduate School of Medical Science, Weill Cornell Medicine, Cornell University, NY, USA. 6. The University of Texas MD Anderson Cancer Center, Department of Cardiology, Houston, TX, USA
Abstract: RNA velocity provides an approach for inferring cellular state transitions from single-cell RNA sequencing (scRNA-seq) data. Conventional RNA velocity models infer universal kinetics from all cells in an scRNA-seq experiment, resulting in unpredictable performance in experiments with multi-stage and/or multi-lineage transition of cell states where the assumption of the same kinetic rates for all cells no longer holds. Here we present cellDancer, a scalable deep neural network that locally infers velocity for each cell from its neighbors and then relays a series of local velocities to provide single-cell resolution inference of velocity kinetics. In the simulation benchmark, cellDancer shows robust performance in multiple kinetic regimes, high dropout ratio datasets and sparse datasets. We show that cellDancer overcomes the limitations of existing RNA velocity models in modeling erythroid maturation and hippocampus development. Moreover, cellDancer provides cell-specific predictions of transcription, splicing and degradation rates, which we identify as potential indicators of cell fate in the mouse pancreas.
Keywords: cell fate, RNA velocity, scRNA-seq, deep learning, cellDancer, relay velocity model

Title: Clinical and Genomic Investigation of Immune-Related Adverse Events
Author list: Yanfei Wang, PhD[1], Tyler A. Shugg, PharmD[2], Michael T. Eadon, MD[3], Jing Su, PhD[4], Thomas J George, MD[5], Jiang Bian, PhD[1], Steven M Smith[6], Yan Gong[7], Qianqian Song, PhD[1]
Detailed Affiliations:
[1]Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL, USA. [2]Division of Clinical Pharmacology, Indiana University School of Medicine, Indianapolis, IN, USA. [3]Department of Medicine, Indiana University School of Medicine, Indianapolis, IN, USA. [4]Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN, USA. [5]Division of Hematology & Oncology, University of Florida & UF Health Cancer Center, Gainesville, FL, USA. [6]Department of Pharmaceutical Outcomes and Policy, College of Pharmacy, University of Florida, Gainesville, FL, USA. [7]Pharmacotherapy and Translational Research, College of Pharmacy, University of Florida, Gainesville, FL, USA
Abstract: Immune checkpoint inhibitors (ICIs) have revolutionized cancer therapy, significantly improving survival outcomes across a range of malignancies. However, their use is frequently complicated by immune-related adverse events (irAEs), including acute kidney injury (AKI) and cardiovascular adverse events (CVAE), which can lead to treatment discontinuation and increased morbidity. To comprehensively characterize these risks, we leveraged large-scale clinical and genomic data from the OneFlorida+ Clinical Research Network and the All of Us Research Program. In a cohort of ICI-treated patients from OneFlorida+, 56.2% developed irAEs within one year, with severe cases notably impacting overall survival. Younger patients, females, and those with specific comorbidities such as myocardial infarction and renal disease were at higher risk. Combination ICI regimens further increased irAE incidence, and cancer type

also influenced risk profiles. In parallel, those ICI-treated patients revealed that 19.5% developed CVAEs, most commonly arrhythmias, myocardial infarction, and heart failure. Patients with pre-existing cardiometabolic conditions—including hypertension, diabetes, and hyperlipidemia—showed significantly elevated CVAE risk. Combination regimens, especially those involving CTLA-4 and PD-(L)1 inhibitors, were strongly associated with higher CVAE rates. Furthermore, genomic analysis of the All of Us cohort identified the rs16957301 variant in the PCCA gene as a novel genetic risk factor for ICI-AKI among Caucasian patients, with risk genotypes associated with earlier and higher incidence of AKI. Collectively, these findings highlight the critical importance of integrating clinical and genetic risk assessments to personalize ICI therapy, improve patient monitoring, and mitigate severe treatment-related toxicities. Tailored strategies addressing both renal and cardiovascular risks will be essential to ensure the safe and effective use of ICIs across diverse patient populations.

Keywords: Immune checkpoint inhibitors, Immune-related adverse events, risk factors

Title: Machine Learning-Based Integration of Transcriptomic and Epigenetic Data for Cancer Biomarker Discovery

Author list: Alper Uzun[1,2,3]

Detailed Affiliations:

[1]Department of Pathology and Laboratory Medicine, Warren Alpert Medical School of Brown University, Providence, RI 02903, USA; [2]Legorreta Cancer Center, Brown University, Providence, RI 02912, USA; [3]Brown Center for Clinical Cancer Informatics and Data Science (CCIDS), Brown University, Providence, RI 02912, USA

Abstract: ASCEND is a novel computational framework developed to integrate transcriptomic and DNA methylation data for accurate cancer prediction and biomarker discovery. Gene expression is a critical indicator of cellular function, and its dysregulation, often driven by epigenetic modifications like DNA methylation, plays a central role in cancer development. ASCEND bridges these two molecular layers to identify genes predictive of cancer and pinpoint methylation markers that may regulate their expression. Developed in Python using libraries such as scikit-learn, pandas, and numpy, ASCEND processes raw data from The Cancer Genome Atlas (TCGA), filters outliers and incomplete entries, and standardizes expression values across patients. Its workflow involves two main stages: the first predicts cancer presence using a Multilayer Perceptron Classifier trained on gene expression data to identify high-impact biomarker genes; the second uses a linear regression model to associate CpG methylation sites with those selected genes, revealing potential regulatory mechanisms. The tool automatically selects the top five genes based on importance scores but allows user customization. ASCEND was applied to datasets from breast, lung, and prostate cancers, comprising 370 samples from both healthy individuals and patients. In the case of breast adenocarcinoma, ASCEND achieved an 87% classification accuracy and identified WEE2P1, SUPT20HL1, TBC1D4, DGCR11, and TEX26 as the top candidate genes. Methylation analysis identified key CpG sites such as cg00396667, cg00493804, and cg00554640 from a pool of 27,577, many of which were mapped to these genes using ASCEND's feature importance scoring system. Literature review confirmed the relevance of four of the five identified genes to breast cancer, supporting the model's reliability and biological relevance. ASCEND's results are visualized through a user-friendly interface, offering customizable parameters, graphical outputs, and modular adaptability to different cancer types. The tool is openly available on GitHub, supporting transparent, reproducible research. By linking gene expression and DNA methylation in a single platform, ASCEND offers researchers a scalable and insightful

method for understanding the molecular basis of cancer, prioritizing diagnostic and therapeutic targets, and accelerating discoveries in precision oncology. Its design accommodates future expansion, enabling integration with additional omics layers and broader clinical applications, making ASCEND a valuable addition to the cancer bioinformatics toolkit.

Keywords: Cancer Biomarkers, Transcriptomics, DNA Methylation, Machine Learning, Multilayer Perceptron, Epigenetics

Title: Subtyping Metabolic Dysfunction-Associated Steatotic Liver Disease using Electronic Health Record-Linked Genomic Cohorts Reveals Diverse Etiologies and Progression

Author list: Tahmina Sultana Priya[1,7], Huihuang Yan[2,3], Kirk J. Wangensteen[4], Stephen Wu[2], Anthony C. Luehrs[5], Filippo Pinto e Vairo[3], Fan Leng[6], Andres J. Acosta[4], Robert, A. Vierkant[5], Alina M. Allen[4], Konstantinos N. Lazaridis[4], Eric W. Klee[2,3*], Danfeng (Daphne) Yao[1,7*], Shulan Tian[2,3*]

Detailed Affiliations:

[1]Department of Computer Science, Virginia Tech, Blacksburg, VA, USA; [2]Division of Computational Biology, Department of Quantitative Health Sciences, Rochester, MN, USA; [3]Center for Individualized Medicine and Department of Clinical Genomics, Mayo Clinic, Rochester, MN, USA; [4]Division of Gastroenterology and Hepatology, Department of Internal Medicine, Mayo Clinic, Rochester, MN, USA; [5]Division of Clinical Trials and Biostatistics, Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, USA; [6]Data Analytics and Integration, Mayo Clinic, Rochester, MN, USA; [7]Sanghani Center for Artificial Intelligence and Data Analytics, Virginia Tech, Blacksburg, VA, USA;

Abstract: Metabolic dysfunction-associated steatotic liver disease (MASLD) is a heterogeneous condition with diverse etiologies and clinical presentations. Stratifying patients into homogenous subgroups, or subtypes, could potentially reveal molecular mechanisms driving disease risk and progression. Yet, a consensus of subtypes is lacking in MASLD, posing major challenges in developing tailored interventions. In this study, we developed a subtyping framework based on latent class analysis of significant MASLD-related clinical variables, followed by centroid-based assignment of new patients to established subtypes. We identified five subgroups with distinct genetic, clinical, and risk profiles, which were well recapitulated in an independent MASLD cohort. Polygenic risk score and genetic variant analysis revealed genetic contributions across all subgroups. In particular, two of the subgroups, male-predominant cardiorenal (C2) and female-predominant with obesity and mood disorders (C3), are associated with high prevalence of type 2 diabetes, obesity and sleep apnea. The latter also had relatively high usage of antidepressant medicine. On the other hand, the idiopathic subtype (C4) was characterized by the lowest incidence of metabolic comorbidities and ischemic heart disease. Nevertheless, this subgroup overall had the highest rate of liver transplant, which is likely driven, in part, by the combinatorial genetic effects of high-prevalent risk alleles in *TM6SF2* and *MBOAT7* together with low-prevalent protective allele in *HSD17B13*. Finally, the hepatic injury subtype C5 showed an increased risk of developing advanced fibrosis and acute renal failure. Together, our study provides key insights into MASLD heterogeneity, highlighting the need for personalized therapies.

Keywords: Metabolic dysfunction-associated steatotic liver disease; Subtyping; Latent class analysis; Polygenic risk score; Precision medicine

Title: Predicting Cancer Recurrence Using Deep Learning Based Models

Author list: Jessica A. Patricoski-Chavez[1,2,3], Seema Nagpal[4], Ritambhara Singh[1,5], Jeremy L. Warner[2,6,7], Ece D. Gamsiz Uzun[1,2,3,6,7,8]

Detailed Affiliations:

[11]Center for Computational Molecular Biology, Brown University, Providence, RI; [2]Brown Center for Clinical Cancer Informatics and Data Science (CCIDS), Legorreta Cancer Center, Brown University, Providence, RI; [3]Department of Pathology and Laboratory Medicine, Brown University Health, Providence, RI; [4]Department of Neurology, Division of Neuro-oncology, Stanford University, Palo Alto, CA; [5]Department of Computer Science, Brown University, Providence, RI; [6]Departments of Medicine and Biostatistics, Brown University, Providence, RI; [7]Brown University Health Cancer Institute, Rhode Island Hospital, Providence, RI; [8]Department of Pathology and Laboratory Medicine, Warren Alpert Medical School of Brown University, Providence, RI

Abstract: Cancer is one of the leading causes of morbidity and mortality worldwide, with millions of new cases diagnosed each year. According to World Health Organization (WHO), nearly 10 million people worldwide lost their lives to cancer in 2020. Cancer recurrence remains a significant challenge, as it can occur months or even years after the initial treatment, underscoring the need for effective monitoring and predictive strategies. Understanding a patient's likelihood of cancer recurrence is crucial for optimizing treatment selection and timing, which can improve overall outcomes, and enhance quality of life. Deep learning (DL) models have shown increasing promise in various medical applications, including predicting disease recurrence. Gliomas represent approximately 25.5% of all primary brain and central nervous system (CNS) tumors and 80.8% of malignant brain and CNS tumors. Approximately 62% of patients experience recurrence within five years, and 17%–32% progress from low to high-grade glioma. Patients with low-grade gliomas (LGGs) have 5-year survival rates of up to 80%, while patients with higher-grade gliomas (HGGs) often experience rates below 5%. To explore the capability of DL models for predicting recurrence, we developed gLioma recUrreNce Attention-based classifieR (LUNAR), to predict early vs. late glioma recurrence using clinical, mutation, and mRNA-expression data from patients with primary grade II-IV gliomas from The Cancer Genome Atlas (TCGA). As an external validation set, we used the Glioma Longitudinal Analysis Consortium (GLASS). LUNAR outperformed all traditional ML models achieving area under the receiver operating characteristic curve (AUROC) of 82.84% and 82.54% on the TCGA and GLASS datasets, respectively.

Keywords: Cancer, deep learning, genomics, glioma, recurrence

Title: Genetic Impact of Alternative Transcription Initiation Reveals a Novel Molecular Phenotype for Human Diseases

Author list: Hui Chen[1], Xudong Zou[1], Wei Wang[2], Shuxin Chen[1], Yu Chen[2], Lei Li[1]

Detailed Affiliations:

[1]Shenzhen Bay Laboratory, Institute of Systems and Physical Biology. [2] Shenzhen Bay Laboratory, Institute of Cancer Research.

Abstract: Alternative transcriptional initiation (ATI) is a fundamental layer of gene regulatory mechanisms, characterized by multiple transcription start sites (TSSs) for a single gene, generating functionally distinct

isoforms. ATI plays crucial regulatory roles in mediating tissue-specific gene expression and contributes significantly to the dysregulated transcriptomes observed in cancer. However, the genetic architecture underlying ATI variation and its mechanistic links to cancer susceptibility remain an important knowledge gap in the field. Here, we present the first comprehensive characterization of the genetic regulation of alternative transcription initiation (ATI) spanning 49 human normal tissues and 33 tumor tissues. We identified 9,075 5′UTR alternative transcription initiation trait loci (5′aQTLs), encompassing approximately 0.41 million common genetic variants associated with the usage of distal transcription start sites (TSSs) of 5,436 genes, 32.1% of which were overlooked by eQTLs. We found that 7.6% of disease variants are colocalized with 5′aQTL signals, and 74.0% of them were overlooked by eQTLs. By integrating our 5′aQTL with well-powered GWAS datasets through transcriptome-wide association studies (TWAS), we identified 156 cancer susceptibility genes, including established cancer markers such as *MAFF* and *MLLT10*, as well as novel candidates. Collectively, our study reveals ATI as a critical mechanism linking non-coding variants to cancer risk, providing new insight for cancer target discovery.

Keywords: alternative promoter; quantitative trait loci; transcriptome-wide association study; cancer susceptibility

Workshop – Advanced Computational Statistics and Artificial Intelligence to Address Public Health Epidemics

August 3rd

8:30 AM – 11:30 AM

Room: 301

Chairs: Naleef Fareed, Soledad Fernandez

Title: Leveraging urinary drug test (UDT) results as a novel data source and proxy for drug use

Author List: Naleef Fareed[1], Ping Zhang[1], Joanne Kim[1], Penn Whitley[2], Charles Mark[2], Brandon Slover[1], Steven Passik[2], Eric Dawson[2], John Myers[1], Xianhui Chen[1], Changchang Yin[1], Fode Tounkara[1], Neena Thomas[1], Bridget Freisthler[3], Tim Huerta[4], Soledad Fernandez[1], (Rebecca Jackson[5])

Detailed Affiliations:

[1]Department of Biomedical Informatics, College of Medicine, The Ohio State University; [2] Millenium Health, LLC; [3] Department of Social Work, The Ohio State University; [4] Department of Family and Community Medicine, College of Medicine, The Ohio State University;; [5] Department of Internal Medicine, College of Medicine, The Ohio State University

Abstract: UDT data is example of a novel data source, like wastewater treatment data, to proxy fluctuating patterns of a public health phenomenon and support public health officials with decision making using signals to plan and predict crises. These signals could allow for better problem diagnosis, identification of cold/hot clusters in a geographical area, and enable tailored responses to critically hit areas in a timely manner. Our first presentation provides context for the workshop by: 1) embedding the use of novel measures such as UDT to characterize differential human behavior and social processes; 2) discussing the

scientific rationale for using novel measures such as UDT data within the context of the opioid crisis; and 3) describing the various analytical challenges of acquiring and processing multimodal and novel data sources for public health forecasting; and 4) lessons learned from our National Institute on Drug Abuse funded project to use UDT data, along with other multi-modal data sources, to develop a model for predicting opioid-related mortality outcomes. We will provide the audience with the logistics for the subsequent presentations and the learning objectives from each presentation. Drs. Fareed and Fernandez will field questions throughout the workshop as part of a Q&A session that will be held shortly after the presentations. There will also be a discussion of current challenges and limitations, along with strategic recommendations for the effective use of routinely collected data and emerging computational tools in forecasting public health trends and informing timely interventions.

Keywords: Prediction models; public health; epidemiology; artificial intelligence; community health; opioid crisis

Title: Predicting opioid overdose mortality using UDT data with a Bayesian approach

Authors: John Myers[1], Joanne Kim[1], Charles Marks[2], Penn Whitley[2], Brandon Slover[1], Naleef Fareed[1], Soledad Fernandez[1], Neena Thomas[1]

Detailed Affiliations:

[1]Department of Biomedical Informatics, College of Medicine, The Ohio State University; [2] Millenium Health, LLC

Abstract: The opioid crisis represents an ongoing emergency in the United States, and Ohio has one of the highest overdose death rates in the country. Timely interventions are needed to combat this crisis; however, delays in overdose death reporting remain a significant hurdle in developing strategies to combat overdose deaths. Urine drug test (UDT) data provides near real-time weekly updates with valuable insights into drug use patterns in communities. We use UDT data as a proxy for drug use in communities in overdose death prediction and expect it to fill the gap in the lagged overdose death data.

We constructed a hierarchical Bayesian model implemented with Integrated Nested Laplace Approximation (INLA). The model allows spatiotemporal random effects to capture unobserved factors. Spatiotemporal effects were implemented with correlated random effects for space (Besag model) and time (first order random walk). UDT data, sociodemographic factors, and EMS events with naloxone distribution were used as parameters in the model.

Predictions were made at the quarterly level using the moving window approach to utilize the up-to-date drug overdose trend. We used a training-testing schema with a forgetting mechanism to train on eight quarters of data and predicted two quarters at a time, corresponding with the reporting lag of overdose death. We compared our model with baseline models to confirm that the predictive performance improved with the addition of UDT data and EMS naloxone events.

Keywords: Opioid overdose, Bayesian methods, Integrated Nested Laplace Approximation, Prediction, Urine Drug Test, Spatiotemporal

Title: Implementing a Spatial-Temporal Graph Neural Network (ST-GNN) framework, a novel, multi-modal data approach for predicting opioid overdose death rates

Author List: Zishan Gu[2], Xianhui Chen[2], John Myers[1], Joanne Kim[1], Changchang Yin[1], Naleef Fareed[1], Neena Thomas[1], Soledad Fernandez[1], Ping Zhang[2]

Detailed Affiliations:
[1]Department of Biomedical Informatics, College of Medicine, The Ohio State University; [2] Department of Computer Science and Engineering, College of Engineering, The Ohio State University
Abstract: The opioid crisis has severely impacted Ohio, with overdose death rates surpassing national averages and disproportionately affecting rural and Appalachian regions. Timely resource allocation and response are essential for preventing the further escalation of overdose deaths. Consequently, many studies focus on developing predictive models that enable timely interventions. Although these methods have demonstrated promising performance, there are three limitations: (1) they do not integrate the spatial relationships inherent in geographic information; (2) existing methods fail to integrate static information with dynamic data; and (3) current studies apply the same loss function to both large and small counties, leading to unfair modeling. To address these challenges, we propose the Spatial-Temporal Graph Neural Network (ST-GNN) framework, a novel approach for predicting opioid overdose death rates at county level. Our framework leverages the strengths of GNNs to model spatial relationships between counties, augmented by Long Short-Term Memory (LSTM) networks to capture temporal dynamics. In particular, this study uses Ohio's quarterly opioid overdose data from 2017 to 2023, enriched with dynamic features such as EMS naloxone administration events and static SDoH features, to train and evaluate the ST-GNN framework. By jointly training these components, the ST-GNN framework dynamically models the evolution of opioid overdose deaths across time and geography. Furthermore, to account for heterogeneity among counties with varying population sizes, we tailor prediction tasks accordingly with a joint training loss: for small counties where monthly or quarterly death counts are close to three, we formulate a binary classification task to predict whether the death count will exceed three. In contrast, for larger counties, we perform a standard regression task to predict the actual death counts. Compared with LSTM, DCRNN and GConvLSTM, our work not only advances the baselines in predictive modeling for opioid overdose deaths but also provides a scalable and flexible solution for addressing public health crises driven by complex spatial-temporal phenomena. Experiments conducted on data reported by the Ohio Department of Health demonstrate that our proposed method outperforms all baseline models for both large and small counties. For large counties, our method achieves an RMSE of 9.149 and an SMAPE of 0.242. For small counties, it achieves an ROC-AUC of 0.738 and an F1 score of 0.579.
Keywords: Opioid Overdose Prediction, Long Short-Term Memory (LSTM) networks, Graph Neural Network, Public Health

Title: Flexible Copula-Based Capture–Recapture Modeling of Opioid Misuse Using Urine Drug Testing Data: Evidence from Franklin County, Ohio (2016–2023)

Author List: Fode Tounkara[1], Naleef Fareed[1], Charles Marks[2], Penn Whitley[2], Neena Thomas[1], Soledad Fernandez[1]
Detailed Affiliations:
[1]Department of Biomedical Informatics, College of Medicine, The Ohio State University; [2] Millenium Health, LLC;
Abstract: Understanding the hidden burden of opioid misuse remains a critical public health priority. Capture–recapture methods using urine drug testing (UDT) data offer a powerful framework for estimating the prevalence of people who misuse opioids (PWMO), especially when direct enumeration is not feasible. However, traditional models often fail to capture the complex heterogeneity and dependency structures among individuals.

We analyzed quarterly UDT data from Franklin County, Ohio, covering the years 2016 to 2023. Individuals were considered "captured" if they tested positive for any illicit opioid (e.g., heroin, fentanyl, oxycodone) during a given capture occasion. For each year, we derived binary capture histories and applied (1) a traditional generalized linear model (GLM) and (2) a suite of flexible copula-based zero-truncated binomial mixture models incorporating individual covariates (e.g., age). Copula families included Clayton, Gumbel, Joe, and Frank. For each model, we estimated the total number of PWMOs ($\hat{N}$), standard error (SE), 95% CI, and model fit (AIC). Subgroup analyses were conducted by sex.

Throughout the study period, the Frank copula model consistently outperformed or matched other models, especially when addressing asymmetric dependence structures. Estimates of $\hat{N}$ sometimes varied by over 20% between the Generalized Linear Model (GLM) and the best copula model, emphasizing the need to model latent correlations accurately. Subgroup analyses showed different trends in opioid misuse by gender, with females experiencing sharper increases in estimated prevalence post-2020, while male estimates remained more stable. Copula-based capture–recapture models provide a robust alternative for estimating hidden populations from complex surveillance data. By incorporating varied dependencies and individual covariates, our approach enhances opioid misuse prevalence estimates at the county level, informing public health resource allocation and overdose prevention strategies. This framework is also applicable to other substance use and surveillance areas.

Keywords: Opioid misuse, capture–recapture, copula model, urine drug testing, Franklin County, public health informatics

Title: Evaluating the public health decision support landscape for opioid outcomes

Author List: Naleef Fareed[1], Joanne Kim[1], Brandon Slover[1], Neena Thomas[1], Fernandez, S.

Detailed Affiliations:

[1]Department of Biomedical Informatics, College of Medicine, The Ohio State University

Abstract: A data dashboard is an example of a digital tool that could effectively translate complex datasets into interactive visuals to support public health decision making. However, it can be very difficult to create a dashboard that provides both clear insight into a problem, and the ability to act on this insight. To do so, it is crucial to understand why the dashboard will be used, who will be using it, and what the desired impact of the dashboard will be. The overarching goal of our study is to develop a dashboard to communicate both statistical and machine learning prediction models to identify potential opioid overdose outbreaks across Ohio communities. By displaying these models on a dashboard along with simpler visualizations and filters, stakeholders in Ohio could gain some knowledge about the models and use the results to preemptively address opioid overdose outbreaks. As the dashboard was being built, questions emerged about what was necessary to create an effective tool for addressing real-world problems. What factors could be included that were the most impactful? What is the best way to arrange visualizations? How could the dashboard be created to hold the user's attention, and ensure it was not too difficult to use? It was decided that a more structured process was needed to develop an effective dashboard that would accomplish its desired goal. The research team systematically explored other existing state opioid dashboards to be leveraged as inspiration. First, a rubric was developed for grading eleven different components of each dashboard. Three reviewers were then tasked with reviewing 42 of the state opioid dashboards. Cluster analysis was then performed to group the results based on the overall score, providing insight on which dashboards performed the best. Then, researchers met with the stakeholders who will be leveraging the dashboard to address the opioid overdose epidemic to prevent future outbreaks. These interviews were used to gauge what these individuals thought would be the most and least impactful to be included and ensure that all avenues of

addressing the problem were considered. During our presentation, we will discuss our approach and how other researchers can adopt similar techniques to design and implement public health prediction tools using interactive tools such as dashboards.
Keywords: Dashboard, public health, statistical models, machine learning models, Ohio, opioids, overdose, prediction, review, cluster analysis, prevent, interviews, implement, interactive tools

Title: PCORsearch: A Scalable, User-Centric Platform for Self-Service Cohort Discovery and Feasibility Analysis of PCORnet Data
Author list: [1]Jacob Herman, [1]Maciej Pietrzak, [1]Neena Thomas, [1]Soledad Fernandez
Detailed Affiliations: [1]Department of Biomedical Informatics, College of Medicine, The Ohio State University
Abstract: Regulatory and institutional restrictions on electronic health record (EHR) access often cause delays, sometimes extending weeks, as researchers depend on independent data analysts for feasibility analysis. To address this bottleneck, we developed the Patient-Centered Outcomes Research Search Tool (PCORsearch), a web-based application enabling investigators to independently analyze deidentified EHR-derived data while maintaining regulatory compliance. PCORsearch allows interactive exploration of medical terminology codes and data definitions from the PCORnet Common Data Model, construction of custom queries, and retrieval of feasibility counts. Additionally, it supports cohort discovery, summary statistics, and visualization to characterize identified cohorts. By streamlining feasibility assessment and reducing reliance on data analysts, PCORsearch accelerates early-stage research planning and enables broader access to large-scale clinical data resources in a secure, compliant manner.
Keywords: Cohort Discovery; EHR Analytics; Research Feasibility; Feasibility Analysis; Research Automation; Deidentified Data; PCORnet; Patient-Centered Outcomes; EHR Tools; PCORI

Title: Towards AI Co-Scientists for Scientific Discovery in Precision Medicine
Author list: [1]Hao Li, [1]Di Huang, [1]Wenyu Li, [1]Heming Zhang, [1]Patricia Dickson, [1]J Philip Miller, [1]Carlos Cruchaga, [1]Michael Province, [1]Yixin Chen, [1]Philip Payne, [1]Fuhai Li
Detailed Affiliations:
[1]Washington University in St. Louis
Abstract: AI agents are emerging as transformative tools in precision medicine (AI4PM), tackling complex, poorly understood disease pathogenesis. We developed a multi-agent system coordinated by an Orchestrator agent that manages workflows, categorizes known facts, identifies gaps, and generates sequential task plans. This exploratory study demonstrates the system's potential to accelerate scientific discovery in AI4PM by structuring collaborative problem-solving in precision medicine research.
Keywords: Multi-agent; Medical agent; Ai-Medicine

Title: Tokenvizz: GraphRAG-Inspired Tokenization Tool for Genomic Data Discovery and Visualization
Author list: [1]Cerag Oguztuzun, [1]Zhenxiang Gao, [1]Jing Li, [1]Mehmet Koyuturk, [1]Rong Xu
Detailed Affiliations:
[1]Case Western Reserve University
Abstract: Interpreting complex genomic relationships and predicting functional interactions remain key challenges in biomedical research. Traditional sequence-based methods often lack interpretability which limits the exploration of genomic language model predictions. To address this gap, we introduce Tokenvizz, a GraphRAG-inspired tool that transforms genomic sequences into intuitive graph representations, where DNA tokens become nodes connected by edges weighted by attention scores derived from genomic

language models. This novel approach translates genomic sequences into structured graph visualizations that reveal latent token relationships that are difficult to interpret through purely sequential methods. Tokenvizz provides an integrated pipeline that includes data preprocessing, graph construction from tokenized sequences, and an interactive web-based visualization interface. Users can dynamically adjust edge weight thresholds, perform position-based searches, and examine contextual sequence information interactively. This facilitates intuitive, multi-resolution analysis of genomic sequences and enhances the interpretability and exploratory capabilities of genomic language models.To validate Tokenvizz, we applied its graph representations to promoter-enhancer interaction prediction using a Graph Convolutional Network (GCN) on six datasets from the GUE+ benchmark. Tokenvizz consistently outperformed existing sequential deep learning models such as DNABERT2 and Nucleotide Transformer, demonstrating the utility of attention-derived graph structures for genomic prediction tasks. By effectively bridging attention-based genomic language modeling and interactive graph visualization, Tokenvizz offers researchers a visualization tool for exploratory genomic analyses. Future work will explore integrating external genomic annotation databases to further strengthen its interpretability and utility for genomics research. Tokenvizz, along with its user guide, is freely accessible on GitHub at: https://github.com/ceragoguztuzun/tokenvizz
Keywords: genomic language models; graph visualization; DNA sequence analysis; attention mechanism; regulatory elements; genomic interpretation

Workshop – Microbiome Data Analysis: Advanced Methods and Practical Applications

August 3rd
8:30 AM – 11:30 AM
Room: 350

Chairs: Qunfeng Dong, Xiang Gao

Title: A Deep Learning Feature Importance Test Framework for Integrating Informative High-dimensional Biomarkers to Improve Disease Outcome Prediction
Author List: Baiming Zou[1,2], James G. Xenakis[3], Meisheng Xiao[1], Apoena Ribeiro[4], Kimon Divaris[4], Di Wu[1,4], Fei Zou[1,5]
Detailed Affiliations:
[1]Department of Biostatistics, Giling School of Global Public Health, University of North Carolina, Chapel Hill, NC, USA; [2]School of Nursing, University of North Carolina, Chapel Hill, NC, USA; [3]Department of Statistics, Harvard University, Cambridge, MA, USA; [4]School of Dentistry, University of North Carolina, Chapel Hill, NC, USA; [5]Department of Genetics, School of Medicine, University of North Carolina, Chapel Hill, NC, USA.
Abstract: Many human diseases result from a complex interplay of behavioral, clinical, and molecular factors. Integrating low-dimensional behavioral and clinical features with high-dimensional molecular profiles can significantly improve disease outcome prediction and diagnosis. However, while some

biomarkers are crucial, many lack informative value. To enhance prediction accuracy and understand disease mechanisms, it is essential to integrate relevant features and identify key biomarkers, separating meaningful data from noise and modeling complex associations. To address these challenges, we introduce the high-dimensional feature importance test (HdFIT) framework for machine learning models. HdFIT includes a feature screening step for dimension reduction and leverages machine learning to model complex associations between biomarkers and disease outcomes. It robustly evaluates each feature's impact. Extensive Monte Carlo experiments and a real microbiome study demonstrate HdFIT's efficacy, especially when integrated with advanced models like deep neural networks (DNN), termed HdFIT-DNN. Our framework shows significant improvements in identifying crucial features and enhancing prediction accuracy, even in high-dimensional settings.

Keywords: Complex association, Dimension reduction, Interpretable and scalable predictive modeling, Non-parametric feature selection, Stable deep neural network

Title: Enhancing Microbiome-Trait Prediction through Phylogeny-Aware Modeling and Data Augmentation

Author list: Yifan Jiang[1], Disen Liao[1],Matthew Aton[2], Qiyun Zhu[2], Yang Lu[1]

Detailed Affiliations:

[1]Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, Canada; [2]School of Life Sciences, Arizona State University, Tempe, AZ, USA.

Abstract: Understanding how microbial communities influence human traits is central to microbiome research and precision medicine. However, microbiome data present significant analytic challenges due to their high dimensionality, compositional constraints, and strong phylogenetic structure. In this talk, I will present two complementary methods that advance trait prediction from microbiome profiles by leveraging domain-specific priors: MIOSTONE, a taxonomy-aware neural network for interpretable prediction, and PhyloMix, a phylogeny-guided data augmentation technique. MIOSTONE improves interpretability and predictive accuracy by mimicking microbial taxonomy within the model architecture, enabling it to determine whether variations in microbial taxa at different taxonomic levels best explain the outcome. Complementarily, PhyloMix enhances learning by generating synthetic microbiome samples through subtree-level recombination guided by phylogenetic relationships. This strategy introduces meaningful diversity while respecting compositional constraints, boosting performance across multiple models and tasks, including supervised and contrastive learning. Together, these methods demonstrate the power of integrating biological structure into machine learning workflows for robust, interpretable, and effective microbiome-trait association studies.

Keywords: Microbiome-disease association; Biomarker detection; Taxonomy; Phylogeny; Data augmentation;

Title: Leveraging new genomic LLMs for studying under-annotated microbial genes

Author list: Siyuan Ma

Detailed Affiliations:

Department of Biostatistics, Vanderbilt University Medical Center

Abstract: Recent advancements in genomic large language models (LLMs) promise novel bioinformatics solutions for microbiome research. Microbial genomic sequences, like natural languages, form a *language of life*, enabling the adoption of LLMs to extract useful insights from complex microbial ecologies. In this

talk, we will first review recent genomic LLMs, emphasizing their application towards metagenomics data. We will then present a particular application, namely, the study of unannotated microbial genes. We will demonstrate, with both bioinformatics evaluations and epidemiological findings based on public data, that novel genomic LLMs can be utilized to congregate otherwise unannotated microbial genes for more powerful downstream analysis and biologically interpretable findings.

Title: Bayesian spatial statistical models for quantifying relationships among cell types in image data
Author list: Jacqueline R. Starr[1]
Detailed Affiliations:
[1]Brigham and Women's Hospital, Harvard Medical School

Abstract: Our laboratory develops Bayesian spatial models to quantify and test relationships in biofilm or other FISH-based image data. I will describe these methods and how they can be used to investigate the role of bacteria (or other cells) in human health.

Title: Multimedia: An R package for multimodal mediation analysis of microbiome data
Author list: Hanying Jiang[1], Xinran Miao[1], Margaret W. Thairu[2], Mara Beebe[2], Dan W. Grupe[3], Richie J. Davidson[3,4,5], Jo Handelsman[6], Kris Sankaran[1]
Detailed Affiliations:
[1]Statistics Department, University of Wisconsin—Madison, Madison, Wisconsin, USA; [2]Wisconsin Institute for Discovery, University of Wisconsin—Madison, Madison, Wisconsin, USA; [3]Center for Healthy Minds, University of Wisconsin—Madison, Madison, Wisconsin, USA; [4]Psychology Department, University of Wisconsin—Madison, Madison, Wisconsin, USA; [5]Psychiatry Department, University of Wisconsin—Madison, Madison, Wisconsin, USA; [6]Plant Pathology Department, University of Wisconsin—Madison, Madison, Wisconsin, USA
Abstract: Mediation analysis has emerged as a versatile tool for answering mechanistic questions in microbiome research because it provides a statistical framework for attributing treatment effects to alternative causal pathways. Using a series of linked regressions, this analysis quantifies how complementary data relate to one another and respond to treatments. Despite these advances, existing software's rigid assumptions often result in users viewing mediation analysis as a black box. We designed the multimedia R package to make advanced mediation analysis techniques accessible, ensuring that statistical components are interpretable and adaptable. The package provides a uniform interface to direct and indirect effect estimation, synthetic null hypothesis testing, bootstrap confidence interval construction, and sensitivity analysis, enabling experimentation with various mediator and outcome models while maintaining a simple overall workflow. The software includes modules for regularized linear, compositional, random forest, hierarchical, and hurdle modeling, making it well-suited to microbiome data. Our case study revisits a study of the microbiome and metabolome of Inflammatory Bowel Disease patients, uncovering potential mechanistic interactions between the microbiome and disease-associated metabolites, not found in the original study. In addition to summarizing the package, we will explain the software design patterns that we drew inspiration from and how they could inform reproducible multi-omics integration more generally. A gallery of examples and reference page can be found at https://go.wisc.edu/830110.
Keywords: Mediation analysis, data integration, multi-omics, microbiome, software design

Title: VirusPredictor: Software to Predict Virus-related Sequences in Human Data
Author list: Dawei Li[1]
Detailed Affiliations:
[1]Department of Immunology and Molecular Microbiology, Texas Tech University Health Sciences Center, Lubbock, TX 79430, USA
Abstract: Detecting disease-associated viruses without reference genomes, i.e., uncharacterized viruses, in human high-throughput sequencing data is challenging, as such sequences often evade alignment-based methods. Machine learning offers a promising alternative by classifying unmapped reads, potentially revealing novel viral elements. We developed VirusPredictor, a fast, open-source Python tool based on XGBoost and an in-house viral genome database. VirusPredictor uses a two-step classification approach: first, it categorizes sequences as infectious virus, endogenous retrovirus (ERV), or non-ERV human. Accuracy improves with sequence length, reaching 0.76 for short reads (150-350 bp, e.g., Illumina), 0.93 for mid-length reads (850-950 bp, e.g., Sanger), and 0.98 for long reads (2,000-5,000 bp). Sequences predicted as infectious viruses are then classified into one of six viral taxonomic subgroups, with accuracy increasing from 0.92 at 150-350 bp to >0.98 at >850 bp. These results suggest that assembling short reads into contigs (>1,000 bp) enhances prediction accuracy. VirusPredictor achieved high performance on real-world genomic and metagenomic datasets. To our knowledge, this is the first machine learning framework to incorporate both ERV classification and viral subgroup prediction, offering a practical solution for characterizing unmapped sequences potentially derived from uncharacterized viruses.
Keywords: XGBoost; Alignment-free prediction; Unmapped sequences; Uncharacterized virus; Endogenous retrovirus

Title: Integrated Transcriptomics Analysis on Human Respiratory Viral Inoculation and Vaccine Challenge Studies
Author list: Fei Zou
Detailed Affiliations:
Department of Genetics, School of Medicine, UNC
Abstract: Respiratory viral infections cause significant acute and chronic illness and substantial healthcare and economic burden. Human inoculation and vaccine challenge studies offer a unique opportunity to monitor immune cell responses with a controlled timeline. In this talk, I will first present HR-VILAGE-3K3M, the largest human respiratory viral immunization longitudinal gene expression repository database that we have built with publicly accessible transcriptomics data. I will then describe a set of integrated analyses that we perform on HR-VILAGE-3K3M to demonstrate its utility and to investigate cell mediated systemic and local immunity responses to viral inoculation and vaccine challenges.

Title: AI-Powered Discovery of Novel Antimicrobial Peptides in *Trichomonas vaginalis*
Author list : Qunfeng Dong[1], Xiang Gao[1]
Detailed Affiliations:
Department of Medicine, Stritch School of Medicine, Loyola University Chicago, 2160 S 1st Ave, Maywood, IL 60153
Abstract: Antimicrobial peptides (AMPs) are key components of innate immunity but are challenging to identify using traditional sequence-based methods due to their structural diversity. To address this, we fine-

tuned ESM-2, a BERT-style protein language model, on a balanced dataset of AMP and non-AMP sequences. The fine-tuned model demonstrated strong performance in distinguishing AMPs and was applied to a large set of uncharacterized protein sequences from the NCBI RefSeq database. Among the candidates identified, many were from *Trichomonas vaginalis*, a protozoan pathogen known to disrupt vaginal microbial balance by suppressing *Lactobacillus* species. Additional analyses, including transcriptomic data and genomic context, suggest that a substantial portion of these candidates are expressed and associated with mobile genetic elements, pointing to possible roles in host interaction and adaptation. These findings highlight the potential of AI-driven approaches to uncover novel antimicrobial peptides and offer new perspectives on parasite biology and pathogenesis.

Workshop – Advancements in AI and Large Language Models for Biomedical Research

August 3rd
2:30 PM – 5:30 PM
Room: 320

Chairs: Jing Su, Gangqing Hu

Title: Preliminary Evaluation of ChatGPT Model Iterations in Emergency Department Diagnostics
Author list: Jinge Wang[1], Kenneth Shue[1], Li Liu[2,3], Gangqing Hu[1]
Detailed Affiliations:
[1]Department of Microbiology, Immunology & Cell Biology, West Virginia University, Morgantown, WV 26506, USA; [2]College of Health Solutions, Arizona State University, Phoenix, AZ 85004, USA; [3]Biodesign Institute, Arizona State University, Tempe, AZ, 85281 USA.
Abstract: Large language model chatbots such as ChatGPT have shown the potential in assisting health professionals in emergency departments (EDs). However, the diagnostic accuracy of newer ChatGPT models remains unclear. This retrospective study evaluated the diagnostic performance of various ChatGPT models—including GPT-3.5, GPT-4, GPT-4o, and o1 series—in predicting diagnoses for ED patients (n=30) and examined the impact of explicitly invoking reasoning (thoughts). Earlier models, such as GPT-3.5, demonstrated high accuracy for top-three differential diagnoses (80.0% in accuracy) but underperformed in identifying leading diagnoses (47.8%) compared to newer models such as chatgpt-4o-latest (60%, $p < 0.01$) and o1-preview (60%, $p < 0.01$). Asking for thoughts to be provided significantly enhanced the performance on predicting leading diagnosis for 4o models such as 4o-2024-0513 (from 45.6% to 56.7%; $p = 0.03$) and 4o-mini-2024-07-18 (from 54.4% to 60.0%; $p = 0.04$) but had minimal impact on o1-mini and o1-preview. In challenging cases, such as pneumonia without fever, all models generally failed to predict the correct diagnosis, indicating atypical presentations as a major limitation for ED application of current ChatGPT models.
Keywords: ChatGPT; large language models; emergency medicine; diagnosis; model iterations

Title: Thinking, Fast and Slow: DualReasoning Enhances Clinical Knowledge Extraction from Large Language Models

Author list: Haining Wang[1], Chenxi Xiong[1,2], Suthat Liangpunsakul[3], Wanzhu Tu[1], Jing Su[1]

Detailed Affiliations:

[1]Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indiana, IN, USA; [2]Department of Computer and Information Technology, Purdue University, West Lafayette, IN 47907, USA; [3]Department of Medicine, Indiana University School of Medicine, Indiana, IN, USA

Abstract: Large language models (LLMs), trained on vast repositories of medical knowledge, present transformative opportunities for clinical machine learning. However, their potential for extracting actionable insights from unstructured real-world data remains underexplored. We introduce DualReasoning, a novel framework that leverages LLMs to extract meaningful clinical information from medication records, enhancing predictive modeling and disease phenotyping. DualReasoning integrates deliberative, slow (Chain-of-Thought, CoT) and instinctive, fast (non-CoT) reasoning, reflecting human cognitive processes. We applied this approach to the All of Us cohort (N=247,652), which included 26,987 Type 2 Diabetes (T2D) cases, to extract diabetes-related knowledge from medication records. The extracted knowledge was used to enhance downstream predictive models for T2D phenotyping, including logistic regression, random forest, XGBoost, and multi-layer perceptron (MLP). These models incorporated demographic characteristics and Charlson comorbidities as predictors. DualReasoning's outperforms conventional feature engineering approaches (i.e., Polydrug risk scores [PdRS]) and LLM-based medication embeddings. By synergizing slow and fast reasoning, DualReasoning enhances clinical knowledge extraction through better use of medication information. This hybrid framework not only improves disease phenotyping but also shows promise for analyzing complex behavioral patterns, social determinants of health, and mental health conditions—areas where traditional approaches often fall short. Future work will explore its adaptability to broader clinical applications.

Keywords: large language model, drug informatics, knowledge extraction, knowledge presentation, phenotyping

Title: mcDETECT: Decoding the Dark Transcriptomes in 3D with Subcellular-Resolution Spatial Transcriptomics

Author list: Chenyang Yuan[1,2], Krupa Patel[1], Hongshun Shi[1,3,4], Hsiao-Lin V. Wang[1,5], Feng Wang[1], Ronghua Li[1,5], Yangping Li[1†], Victor G. Corces[1,5], Hailing Shi[1,4,5], Sulagna Das[1,6], Jindan Yu[1,3,4], Peng Jin[1,5], Bing Yao[1 *] & Jian Hu[1,2 *]

Detailed Affiliations:

1. Department of Human Genetics, Emory University School of Medicine, Atlanta, GA 30322, USA. 2. Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA. 3. Department of Urology, Emory University School of Medicine, Atlanta, GA 30322, USA. 4. Winship Cancer Institute, Emory University School of Medicine, Atlanta, GA 30322, USA. 5. Emory Center for Neurodegenerative Diseases, Emory University School of Medicine, Atlanta, GA 30322, USA. 6. Department of Cell Biology, Emory University School of Medicine, Atlanta, GA 30322, USA

Abstract: Spatial transcriptomics (ST) has shown great potential for unraveling the molecular mechanisms of neurodegenerative diseases. However, most existing analyses of ST data focus on bulk or single-cell resolution, overlooking subcellular compartments such as synapses, which are fundamental structures of the brain's neural network. Here we present mcDETECT, a novel framework that integrates machine

learning algorithms and *in situ* ST (iST) with targeted gene panels to study synapses. mcDETECT identifies individual synapses based on the aggregation of synaptic mRNAs in three-dimensional (3D) space, allowing for the construction of single-synapse spatial transcriptome profiles. By benchmarking the synapse density measured by volume electron microscopy and genetic labeling, we demonstrate that mcDETECT can faithfully and accurately recover the spatial location of single synapses using iST data from multiple platforms, including Xenium, Xenium 5K, MERSCOPE, and CosMx. Based on the subsequent transcriptome profiling, we further stratify total synapses into various subtypes and explore their pathogenic dysregulation associated with Alzheimer's disease (AD) progression, which provides potential targets for synapse-specific therapies in AD progression.

Keywords：spatial transcriptomics, RNA granule, subcellular structure, Alzheimer's disease, machine learning

Title: A Visual-Omics Foundation Model for Integrating Histopathology Images and Transcriptomics

Author list: Weiqing Chen[1,5,#], Pengzhi Zhang[1,2,3,4,#], Guangyu Wang[1,2,3,4]

Detailed Affiliations:

1. Center for Bioinformatics and Computational Biology, Houston Methodist Research Institute, Houston, TX, 77030, USA. 2. Center for Cardiovascular Regeneration, Houston Methodist Research Institute, Houston, TX, 77030, USA. 3. Center for RNA Therapeutics, Houston Methodist Research Institute, Houston, TX, 77030, USA. 4. Department of Cardiothoracic Surgery, Weill Cornell Medicine, Cornell University, New York, NY, 10065, USA. 5. Department of Physiology, Biophysics & Systems Biology, Weill Cornell Graduate School of Medical Science, Cornell University, New York, NY, 10065, USA

Abstract: Artificial intelligence has revolutionized computational biology, particularly with the emergence of omics technologies such as single-cell RNA sequencing (scRNA-seq) and spatial transcriptomics (ST), which provide detailed genomic insights alongside tissue histology. However, existing computational models typically focus on either omics- or image-based analysis, lacking an integrated approach.

To bridge this gap, we developed OmiCLIP, a visual-omics foundation model that links hematoxylin and eosin (H&E) images with transcriptomic data using tissue patches from Visium datasets. For transcriptomic representation, we generated 'sentences' by concatenating the top-expressed gene symbols from each tissue patch. We compiled a dataset of 2.2 million paired tissue images and transcriptomic profiles across 32 organs to train OmiCLIP, enabling a robust integration of histology and transcriptomics.

Building upon OmiCLIP, we developed the Loki platform, which offers five core functionalities: (1) tissue alignment, (2) tissue annotation using bulk RNA-seq or marker genes, (3) cell type decomposition, (4) image–transcriptomics retrieval, and (5) ST gene expression prediction from H&E images. Compared with 22 state-of-the-art models across five simulated datasets, 19 public datasets, and four in-house experimental datasets, Loki consistently demonstrated superior accuracy and robustness across all tasks.

Keywords: large language model, contrastive learning, histology images, omics

Title: Large language models in cancer pharmacogenomics: from drug-gene association to response prediction

Author list: Yu-Chiao Chiu[1]

Detailed Affiliations:

[1]UPMC Hillman Cancer Center, University of Pittsburgh, Pittsburgh, PA, USA

Abstract: Large language models (LLMs) are emerging as powerful tools in pharmacogenomics, driving both qualitative and quantitative advances in cancer drug studies. This talk highlights two recent applications. First, we introduce a retrieval-augmented framework that qualitatively infers drug-gene-cancer associations by synthesizing evidence from PubMed literature. This efficient approach supports a pan-cancer interaction network and a web-based inference tool. Second, we present a quantitative strategy using an ensemble machine learning model that integrates molecular fingerprints with LLM-derived drug embeddings to predict responses in polyploid giant cancer cells (PGCCs), a chemoresistant breast cancer subpopulation. Together, these studies highlight the potential of LLMs to advance cancer pharmacogenomics through accelerated discovery and predictive modeling.
Keywords: Large language models; Pharmacogenomics; Drug-gene interactions; Cancer drug response

Title: STHD: probabilistic cell typing of single spots in whole transcriptome spatial data with high definition
Author list: Chuhanwen Sun[1]*, Yi Zhang[1,2,3,4,5]*#
Detailed Affiliations:
[1]Department of Neurosurgery, Duke University; [2]Department of Biostatistics and Bioinformatics, Duke University; [3]Department of Cell Biology, Duke University; [4]Brain Tumor Omics Program, The Preston Robert Tisch Brain Tumor Center, Duke University; [5]Duke Cancer Institute. *These authors contributed equally.
Abstract: Recent spatial transcriptomics (ST) technologies have enabled single- and sub-cellular resolution profiling of gene expression across the whole transcriptome. However, the transition to high-definition ST significantly increased data sparsity and dimensionality, posing computational challenges in identifying cell types, deciphering neighborhood structure, and detecting differential expression - all are crucial steps to study normal and disease ST samples. Here we present STHD, a novel machine learning method for probabilistic cell typing of single spots in whole-transcriptome, high-resolution ST data. Unlike the current binning-aggregation-deconvolution strategy, STHD directly models gene expression at single-spot level to infer cell type identities without cell segmentation or spot aggregation. STHD addresses sparsity by modeling count statistics, incorporating neighbor similarities, and leveraging reference single-cell RNA-seq data. We show in VisiumHD data that STHD accurately predicts cell type identities at single-spot level, which achieves precise segmentation of both global tissue architecture and local multicellular neighborhoods. The high-resolution labels facilitate various downstream analyses, including cell type-stratified bin aggregation, spatial compositional comparisons, and cell type-specific differential expression analyses. Moreover, STHD labels further reveal frontlines of inter-cell type interactions at immune hubs in cancer samples. STHD is scalable and generalizable across diverse samples, tissues, diseases, and different spatial technological platforms, facilitating genome-wide analyses in various spatial organization contexts. Overall, computational modeling of individual spots with STHD facilitates discoveries in cellular interactions and molecular mechanisms in whole-genome spatial technologies with high resolution. STHD is available at https://github.com/yi-zhang/STHD/.
Keywords: Machine learning, spatial transcriptomics, high definition

Title: Predicting Protein-Protein Interactions with Structure-based ML/DL Modeling
Author list: Haiqing Zhao[1,2], Zhiyuan Song[1,2], Diana Murray[3], Barry Honig[3]
Detailed Affiliations:

[1]Department of Biochemistry and Molecular Biology, University of Texas Medical Branch, Galveston, TX, USA; [2]Sealy Center for Structural Biology and Molecular Biophysics, University of Texas Medical Branch, Galveston, TX, USA; [3]Department of Systems Biology, Columbia University Medical School, New York, NY, USA
Abstract body: Predicting whether two proteins physically interact has become a central challenge in computational biology. While recent deep learning (DL) approaches have shown promise in predicting protein–protein interactions (PPIs), they often lack the computational efficiency required to interrogate the vast number of possible interactions across the proteome. To address this, we developed **PrePPI-AF**, an algorithm that integrates structural information and additional sources of biological evidence to predict PPIs across most of the human proteome. PrePPI-AF leverages AlphaFold-predicted structures, which are parsed into individual domains to construct potential interaction models. In parallel, we introduced **ZEPPI** (Z-score Evaluation of Protein-Protein Interfaces), a framework that evaluates structural models of protein complexes using residue-level sequence co-evolution within interface regions. ZEPPI demonstrated superior performance over other deep learning–based approaches, particularly in evaluating models from the CASP-CAPRI benchmark experiments. We integrated the PrePPI and ZEPPI pipelines with a protein language model-based method to predict the *E. coli* PPI interactome. Clustering the resulting high-confidence predictions revealed functionally coherent subnetworks—even though our methods incorporated no explicit functional annotations. Together, these findings suggest that our proteome-wide prediction framework can serve as an efficient large-scale screening tool, which can be followed by more computationally intensive structural modeling for specific PPIs of interest.
Keywords: Protein-Protein Interaction, Protein Structure Prediction

Title: A Benchmarking Framework for Foundation Models in Drug Response Prediction
Author list: Qing Wang[1,2], Yining Pan[1], Minghao Zhou[1], Qianqian Song[1]
Detailed Affiliations:
[1]Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, FL 32611, USA
Abstract: Understanding and overcoming drug resistance remains a critical challenge in improving cancer treatment outcomes. Advances in single-cell technologies have enabled high-resolution profiling of cellular heterogeneity, offering new insights into therapeutic response. At the same time, large foundation models are rapidly transforming computational biology, yet their applications in predicting drug response using single-cell data remain underexplored. Our work has introduced an integrated benchmarking framework designed to evaluate the performance of diverse foundation models for drug response prediction. This framework incorporates ten foundation models, including models trained on single-cell and general language data, across a curated data resource from a wide range of tissues, cancers, and treatment conditions. Our results demonstrate that model performance varies substantially depending on the evaluation scenario and adaptation strategy. Certain models achieve high accuracy when fine-tuned on labeled data, while others maintain strong generalization in zero-shot settings without retraining. These findings not only provide practical guidance for selecting appropriate models for specific research or clinical contexts, but also highlight the diverse strengths of different modeling paradigms. By offering a flexible, extensible, and user-accessible framework, our study lays a critical foundation for advancing AI-driven drug discovery, supporting the broader goal of enhancing therapeutic decision-making and improving patient outcomes.
Keywords: Single-cell Profiling, Foundation Models, Drug Response Prediction, Low-Rank Adaptation, Zero-shot Learning, Computational Drug Discovery

Workshop – Big data for Better Studying Disease Systems

August 3rd
2:30 PM – 5:30 PM
Room: 301

Chairs: Xiuwei Zhang

Title: ShinyEvents: harmonizing longitudinal data for real world survival estimation.
Author list: Timothy Shaw
Detailed Affiliations:
Moffitt Cancer Center
Abstract: Longitudinal data analysis of the patient's treatment course is critical to uncovering variables that influence outcomes. However, existing tools have significant limitations in integrating multilayered time-series data. Here, we developed ShinyEvents, a web-based framework for complex longitudinal data analysis. https://shawlab-moffitt.shinyapps.io/shinyevents/. ShinyEvents allows users to upload data and generate interactive timelines of the patient's clinical events. Our tool can perform cohort-level analysis, including the assignment of treatment clusters and clinical endpoints. Our tool also provides informative cohort visualizations, such as a Sankey diagram of the treatment line and Swimmer diagram of the clinical course. Finally, our tool can infer a real-world progression-free survival (rwPFS) based on user-defined endpoints to perform Kaplan-Meier and Cox proportional hazards regression analysis. With these features, the tool can then associate the lines of treatment with clinical outcomes. Altogether, ShinyEvents facilitates the integration of multilayered longitudinal data and enables survival analysis in real-time.

Title: Harnessing Big Data to Advance Understanding of Novel Therapeutic Strategies
Author list: Yuan Liu
Detailed Affiliations:
Indiana University
Abstract: A data driven computational framework was developed to systematically expand the landscape of targeted protein degradation by focusing on the characterization of E3 ligases, a central component in proteolysis targeting chimera strategies. By integrating a wide range of large scale datasets spanning transcriptomics, proteomics, protein interaction networks, structural biology, chemical ligandability, functional genomics, and subcellular localization, the study provides a comprehensive and multidimensional profile of the human E3 ligase repertoire. Through the use of machine learning based virtual screening and multi modal data integration, the analysis identifies a broad set of previously underutilized E3 ligases with strong potential to serve as recruiters for induced protein degradation. These candidates are prioritized based on ligand availability, tumor specific expression, low expression in normal tissues, protein interaction potential with disease relevant targets, and other biological features that inform their therapeutic viability. The results are compiled into a publicly accessible and user friendly web portal that allows researchers to explore

E3 ligases based on customizable criteria relevant to disease context and therapeutic design. This work demonstrates how integrative, large scale analytics can be used to uncover new molecular opportunities in drug discovery and guide the rational development of targeted therapies in cancer and other complex diseases.

Title: Spatially Resolved Transcriptomics and Proteomics to Interrogate Biological Mechanisms Underlying Cancer Disparities
Author list: Nina Steele
Detailed Affiliations:
University of Cincinnati **Abstract:** Background: Pancreatic ductal adenocarcinoma (PDAC) is a lethal malignancy and is 20% more likely to develop in Black African American (BAA) patients, compared to Non-African American (NAA) patients. Comprehensive analysis of the large-scale gene expression signatures of tumor subtypes and cellular microenvironments of BAA PDAC versus NAA PDAC has not been rigorously evaluated. Methods: Here we present, for the first time, a multi-omics interrogation of BAA versus NAA PDAC using laser capture microdissection, spatial transcriptomics (ST), and 100-plex spatial proteomics platforms, with key findings verified by immunohistochemistry. Results: Whole exome sequencing confirmed African genetic ancestry in self-reported race of BAA samples. We analyzed 7 BAA and 15 NAA PDAC samples, combining our cohort with Human Tumor Atlas Network (HTAN) ST data. We found that BAA PDAC tumors exhibit a significant decrease in classical subtype cells accompanied by no change in the aggressive basal-like subtype. Intriguingly, BAA patients show a significant enrichment of the intermediate/hybrid phenotype that co-expresses genes in both subtypes, as well as unique markers, compared to NAA PDAC. BAA PDAC tissues displayed increased B cell interactions and reduced myeloid cell presence, compared to NAA tumors. Conclusions: Our findings suggest BAA tumors show a shift toward intermediate subtypes and a distinct immune cell composition within the tumor microenvironment. This study highlights that health disparities in patients of different ancestry may have a molecular basis that influences tumor aggressiveness and the immune response that may inform therapeutic efficacy and options. Significance: Our study reveals BAA tumors exhibit reduced classical content, increased intermediate subtype and B cell interactions, compared to NAA tumors. These findings underscore the importance of including diverse populations in cancer research.

Title: Studying single cells through multi-omics and multi-condition scRNA-seq
Author list: Xiuwen Zhang
Detailed Affiliations:
Georgia Institute of Technology
Abstract: With the advances in single cell technologies, cells are profiled through multiple modalities, and data on samples from an increasing number of individuals are obtained. Data integration methods across batches, modalities and conditions are being developed to learn representations of cells in a unified space. I will discuss how we learn biological insights in addition to building integrated datasets through integration methods developed in our lab. In addition, I will discuss how we use scMultiSim to evaluate these methods and other methods for single cell omics.

Title: High-resolution reconstruction of single-cell specific spatial genome architectures in 3D space reveals context-specific mechanisms of long-range gene regulation

Author list: Jianrong Wan
Detailed Affiliations:
Michigan State University
Abstract: Spatial chromosomal conformations in 3D space play pivotal roles in modulating interactions of long-range transcription regulation and demonstrates 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 demonstrate 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, geometric properties of chromatin loops, 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.

Title: Integrating amyloid imaging and genetics for early risk stratification of Alzheimer's disease
Author list: Jingwen Yan
Detailed Affiliations:
Indiana University
Abstract: Alzheimer's disease (AD) initiates years prior to symptoms, underscoring the importance of early detection. While amyloid accumulation starts early, individuals with substantial amyloid burden may remain cognitively normal, implying that amyloid alone is not sufficient for early risk assessment. Given the genetic susceptibility of AD, a multi-factorial pseudotime approach was proposed to integrate amyloid imaging and genotype data for estimating a risk score. Validation involved association with cognitive decline and survival analysis across risk-stratified groups, focusing on patients with mild cognitive impairment (MCI). Our risk score outperformed amyloid composite standardized uptake value ratio in correlation with cognitive scores. MCI subjects with lower pseudotime risk score showed substantial delayed onset of AD and slower cognitive decline. Moreover, pseudotime risk score demonstrated strong capability in risk stratification within traditionally defined subgroups such as early MCI, apolipoprotein E (*APOE*) ε4+ MCI, *APOE* ε4– MCI, and amyloid+ MCI, highlighting its great potential to improve the precision of early risk assessment.

Title: Integrated Multi-Omics Study in Early Onset of Type 1 Diabetes
Author list: Wenting Wu
Detailed Affiliations:
Indiana University
Abstract: This work integrates multimodal single-cell RNA sequencing (scRNA-seq), CITE-seq and RNA splicing analysis to uncover immune and transcriptomic signatures associated with early-onset type 1 diabetes (T1D). scRNA-seq profiling of PBMCs from newly diagnosed youth revealed NK cells displaying the most pronounced transcriptional alterations. Cross-tissue comparisons with CITE-seq data from pancreatic lymph nodes identified conserved NK phenotypes, including two major subsets (CD56^brightCD16^lo and CD56^dimCD16^hi) with altered composition in T1D. Meanwhile, elevated expression of PDIA3 was observed, which is a gene induced by type I interferons and predicted to be regulated by IRF1 by SCENIC prediction. Functional validation using type I IFN–treated NK-92 cells confirmed upregulation of IRF1 and activation markers. To assess translational relevance, we used CIBERSORTx and our in-house deconvolution tool ICTD to analyze rituximab trial data, finding strong concordance in immune composition estimates and treatment response stratification was in analysis. A Shiny app was developed to facilitate gene-level exploration. Complementing these results, deep RNA sequencing and machine learning were applied to whole blood from individuals with new-onset T1D and matched controls. This analysis uncovered distinct RNA splicing patterns, particularly retained introns, that robustly distinguished T1D cases and were enriched for antiviral response pathways. Together, these studies demonstrate that immune cell transcriptional states and alternative RNA splicing in circulation reflect key features of T1D pathogenesis. The findings support the utility of integrated transcriptomic and computational approaches for biomarker discovery, mechanistic insight, and therapeutic stratification in autoimmune disease.

Workshop – Advanced Omics Platforms and Tools

August 3rd
2:30 PM – 5:30 PM
Room: 350

Chairs: Kaixiong Ye, Hongbo Liu

Title: CCLLM: Cellular Community Large Language Model to identify motifs of cell organization in spatial transcriptomics
Author list: Chunyang Chai[1], Yang Yu[2], Shuang Wang[3], Dong Xu[2], Huiyan Sun[1], Juexin Wang[4]
Detailed Affiliations:
[1]School of Artificial Intelligence, Jilin University, Changchun, 130012, China; [2]Department of Electrical Engineering and Computer Science, Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA; [3]Department of Computer Science, Luddy School of Informatics, Computing, and Engineering, Indiana University Bloomington, Bloomington, IN 47405, USA; [4]Department of

Biomedical Engineering and Informatics, Luddy School of Informatics, Computing, and Engineering, Indiana University Indianapolis, Indianapolis, IN 46202, USA
Abstract: The organization of diverse cellular types and states is recognized to be associated with tissue function. However, spatial principles and underlying mechanisms governing that organization remain largely unresolved across most physiological and pathological contexts. Detecting explicit conserved patterns of spatial cell organization as topological cell type combinations, known as Cellular Community motifs (CC motifs), suffer from high computational costs and limited detection accuracy. We introduce Cellular Community Large Language Model (CCLLM) to identify CC motifs leveraging fine-tuned large language models (LLMs) modeling graphs constructed from spatial transcriptomics data. By converting spatial cellular distributions into structured textual prompts, CCLLM accurately identifies and counts subgraph patterns within cellular communities. We apply CCLLM on synthetic and real-world datasets to show its effectiveness and robustness in identifying disease-specific CC motifs with varying spatial resolutions, pathological conditions, and treatment responses. CCLLM harnesses the reasoning capabilities of LLMs to generate biologically meaningful interpretations of CC motif functions. This framework underscores the potential of graph-based LLMs modeling biological systems, offering insights into cellular communication dynamics and therapeutic target discovery.
Keywords: Spatial transcriptomics, cellular community, large language model, graph modeling, motifs

Title: A universal gene representation of atlas single cell data
Author list: Hao Chen[1,2], Nam D. Nguyen[1], Matthew Ruffalo[1], Ziv Bar-Joseph[1]
Detailed Affiliations:
[1]Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA, USA; [2]Department of Computer Science, University of Illinois Chicago, IL, USA.
Abstract: Recent efforts to generate atlas-scale single cell data provide opportunities for joint analysis across tissues and across modalities. Most of the existing methods for single cell atlas analysis use cells as the reference unit to combine datasets. However, such methods suffer from the limitations of inability to effectively integrate cross-modality data, hindering downstream gene-based analysis, and loss of genuine biological variations. Here we present a new data integration method, GIANT, which is for the first time designed for the atlas-scale analysis from the gene perspective. GIANT first converts datasets from different modalities into gene graphs, and then recursively embeds genes in the graphs into a latent space without additional alignment. Applying GIANT to the HuBMAP datasets creates a unified gene embedding space across multiple human tissues and data modalities, where gene representations reflect the functions of genes in their cells. Further evaluations demonstrate the usefulness of GIANT in discovering diverse gene functions, and underlying gene regulations in cells of different tissues.
Keywords: Single cell, Spatial transcriptomics, Representation learning, Multi-modal

Title: Decoding Kidney Disease at Single-Cell Resolution: A Cross-Platform Spatial Transcriptomics Study
Author list: Haojia Wu[1], Benjamin D. Humphreys[1,2]
Detailed Affiliations:
[1]Division of Nephrology, Department of Medicine, Washington University in St. Louis School of Medicine, St. Louis, MO, USA; [2]Department of Developmental Biology, Washington University in St. Louis School of Medicine, St. Louis, MO, USA

Abstract: Recent advances in spatial transcriptomics have transformed our understanding of how cells are organized in space and how their interactions affect organ function and dysfunction. This understanding is especially important for studying complex organs such as the kidney, where disease progression depends not only on cell-intrinsic changes but also on spatial relationships within the renal microenvironment. While multiple platforms are available for generating high-resolution spatial transcriptomic profiles, it remains unclear which technologies are best suited for studying the kidney and kidney diseases. In this study, we focus on commercially available platforms that provide single-cell resolution spatial transcriptomics data and compare the performance of four technologies, including full transcriptome profiling platforms (Stereo-seq and VisiumHD) and targeted gene panel platforms (Xenium and MERFISH), in both healthy mouse kidneys and those affected by acute kidney injury. All platforms demonstrated strong ability to identify major kidney cell types at single-cell resolution. However, missed cell types were observed in targeted gene panel platforms (Xenium and MERFISH) when key marker genes were absent from the panel design. Stereo-seq and VisiumHD had comparable sensitivity in transcript and gene detection, although Stereo-seq uniquely detected long non-coding RNAs that were not captured in VisiumHD data. On the other hand, VisiumHD offers the advantage of integrating high-quality histological staining images aligned to the spatial transcriptomics data, which can provide enhanced context for downstream data interpretation. When comparing Xenium and MERFISH, Xenium consistently produced cleaner cell clustering and more robust results. While MERFISH detected a higher number of transcripts per cell, its signal distribution was noisier than Xenium. Accurate gene panel selection was found to be critical for both platforms in capturing diverse cell types and cellular states. We have determined the minimal number of genes required to reliably identify all kidney cell types using targeted platforms. Furthermore, we evaluated gene imputation and data integration tools to enhance analysis when only limited gene sets were available. Finally, by integrating data from Stereo-seq, Xenium, and MERFISH in the context of acute kidney injury, we identified disease-associated microenvironments and spatial gene expression shifts that were not captured in our previous single-cell RNA-seq analysis of the same disease model.
Keywords: Spatial transcriptomics, Kidney diseases, VisiumHD, StereoSeq, Xenium, MERFISH

Title: DNA Methylation Predictors of Inflammatory Cytokine Changes in Breast Cancer Survivors Undergoing Chemotherapy

Author list: Hongying Sun[1], Michelle C. Janelsins[1,2]

Detailed Affiliations:

[1]Department of Surgery, Division of Supportive Care in Cancer, University of Rochester Medical Center, Rochester, NY, USA; [2] Wilmot Cancer Institute, Rochester, NY, USA

Abstract: Background: Inflammation is a critical mechanism driving cancer-related cognitive dysfunction, fatigue, and other long-term adverse effects in breast cancer survivors. However, predictive biomarkers for post-treatment inflammatory burden remain limited. DNA methylation, an epigenetic marker that reflects environmental and biological exposures, may serve as an early predictor of systemic inflammation following chemotherapy. **Methods:** We analyzed data from 241 participants enrolled in a multi-center observational cohort: 192 breast cancer patients treated with chemotherapy and 49 age- and sex-matched controls. Whole blood DNA methylation was profiled using Illumina HumanMethylation450 and EPIC BeadChip arrays at three timepoints: baseline (T1), post-chemotherapy (T2), and a change score representing within-person differences (T2_1). A panel of inflammatory cytokines—including IL-6, IL-10, IL-4, IL-8, TNF-α, soluble TNF receptors I and II (sTNFRI, sTNFRII)—was measured from plasma and log-transformed. Epigenome-wide association studies (EWAS) were conducted using linear regression

models adjusted for treatment group and array type. Each CpG site was modeled as a predictor of cytokine concentration at follow-up timepoints. False discovery rate (FDR < 0.05) was used to identify statistically significant associations. **Results:** The most robust epigenome-wide associations were observed for soluble TNF receptor II (sTNFRII) at outcome timepoint T2_1 (post-treatment change). Baseline (T1) methylation predicted T2_1 sTNFRII levels with 2,475 significant CpGs (FDR < 0.05), while follow-up (T2) and change (T2_1) methylation yielded 36 and 843 significant CpGs, respectively. Notably, many of the top differentially methylated CpGs mapped to immune- and inflammation-related genes, including *TNFRSF1B*, *IL1RAP*, *SOCS3*, and *NFKBIA*. In contrast, other cytokines such as IL-6, IL-10, and IL-8 showed fewer significant CpG associations (generally fewer than 25 CpGs per model). No substantial associations were observed in the control group, suggesting chemotherapy-specific epigenetic responses. **Conclusions:** Our findings demonstrate that DNA methylation at baseline is strongly associated with downstream inflammatory activation, particularly for sTNFRII, a marker of TNF-α signaling and immune aging. These results suggest that methylation profiles prior to chemotherapy may serve as predictive biomarkers for inflammation-related late effects in cancer survivors. Future work will focus on validating these epigenetic signals in independent cohorts and integrating multi-omics data to understand causal pathways of immune dysregulation.

Keywords: DNA methylation, cancer-related cognitive impairment, breast cancer, chemotherapy, inflammation

Title: Age-Related Patterns of DNA Methylation Changes

Author list: Kevin Chen[1], Wenshu Wang[2], Hari Naga Sai Kiran Suryadevara[3], Gang Peng[4]

Detailed Affiliations:

[1]DeBakey High School, Houston, TX, USA; [2]Herricks High School, New Hyde Park, NY, USA; [3]Bioinformatics Core, Pennington Biomedical Research Center, Louisiana State University, Baton Rouge, LA, USA; [4]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, USA

Abstract: DNA methylation is a dynamic process that adds methyl groups to DNA molecules, which plays an important role in gene expression regulation, genome stability, and aging. Epigenetic clocks developed over the past decade have used DNA methylation values at specific CpG sites combined with factors such as environmental influences and lifestyle factors to predict biological age with remarkable accuracy. However, these epigenetic clocks, built using conventional penalized regression models, have several limitations: (1) they assume consistent linear changes in methylation with age, even though methylation patterns may vary at different ages; (2) the selected CpGs in these clocks often lack specific biological significance; and (3) there is minimal overlap among the selected CpGs across different clocks, suggesting that only a small subset of a larger group of age-correlated CpGs is used to build an effective clock. Using data from 4,899 samples across 23 GEO datasets, we first analyzed the methylation patterns of 1,868 CpGs from 9 widely used epigenetic clocks and observed the expected consistently linear patterns in a subcluster. Next, we applied our own CpG selection method, which does not assume a lifelong linear correlation between DNA methylation and age. we divided the lifespan into overlapping age windows— [0, 20], [5, 25], [10, 30] …, [55, 75], [60, 80]— and identified CpGs that showed strong correlations with age within each specific window. Most CpGs were selected in early and later life windows, while few were selected in the middle life windows, indicating fast DNA methylation changes during child development and aging. Among the 12,903 unique CpGs selected within any of the windows, we performed clustering and uncovered from main patterns: (1) increase before 20 years old, (2) decrease before 20 years old, (3)

increase before 20 and then decrease after 65, and (4) decrease before 20 and then increase after 65. Comparing this to the clock CpGs, both sets contained CpGs that decreased during adolescence. When examining gender differences, we noted that female samples had notably slowed methylation changes in old age windows compared to male samples, possibly due to hormonal changes during menopause. These findings suggest that age-related methylation changes are more complex than previously thought, with implications for refining epigenetic clocks and redefining the way researchers study aging. Future research should explore the biological mechanisms behind these non-linear and sex-specific patterns.
Keywords: DNA methylation, Biological Age, Sex-specific

Title: Uncovering Hidden Biological and Technical Links from Large-scale DNA Methylome Data
Author list: David Goldberg[1], Sol Moe Lee[1], Cameron Cloud[1], Wanding Zhou[1,2]
Detailed Affiliations:
[5]Children's Hospital of Philadelphia, Philadelphia, PA, USA,
[2]Department of Pathology and Lab Medicine, Perelman School of Medicine University of Pennsylvania, Philadelphia, PA, USA
Abstract: Epigenome-wide association studies (EWAS) are transforming our understanding of the interplay between epigenetics and complex human traits and phenotypes. We performed scalable and quantitative screening of trait-associated DNA cytosine modifications in larger, more inclusive, and stratified human populations. We profiled the ternary-code DNA methylations—dissecting 5-methylcytosine (5mC), 5-hydroxymethylcytosine (5hmC), and unmodified cytosine—yielding multiple biological insights. We revealed a previously unappreciated role of 5hmC in trait associations and epigenetic clocks. We demonstrated that 5hmCs complement 5-methylcytosines (5mCs) in defining tissues and cells' epigenetic identities. In-depth analyses highlighted the cell type context of EWAS and GWAS hits. Using multiple platforms, we conducted a comprehensive human 5hmC aging EWAS, discovering tissue-invariant and tissue-specific aging dynamics, including distinct tissue-specific rates of mitotic hyper- and hypomethylation. These findings chart a landscape of the complex interplay of the two forms of cytosine modifications in diverse human tissues and their roles in health and disease.
Keywords: Epigenetics; DNA methylation; EWAS; Hydroxymethylation

Title: A BLAST from the past: revisiting BLAST's E-value
Author list: Yang Lu[1], William Stafford Noble[2], Uri Keich[3]
Detailed Affiliations:
[1]Cheriton School of Computer Science, University of Waterloo, Waterloo, ON,; [2]Department of Genome Sciences and Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA; [3]School of Mathematics and Statistics, University of Sydney, Camperdown, NSW, Australia.
Abstract: The Basic Local Alignment Search Tool, BLAST, is an indispensable tool for proteomic and genomic research. BLAST's E-values have provided scientists with a meaningful statistical evaluation of reported sequence similarity searches over the past 30 years. Here we critically reevaluate these E-values, showing that they can be at times significantly conservative while at others too liberal. We offer an alternative approach based on generating a small sample from the null distribution of random optimal alignments and testing whether the observed alignment score is consistent with it. In contrast with BLAST, our significance analysis seems valid through extensive simulated and real data experiments. One advantage of our approach is that it works with any reasonable choice of substitution matrix and gap penalties, avoiding

BLAST's limited options of matrices and penalties. In addition, we can formulate the problem using a canonical family-wise error rate control setup, thereby dispensing with E-values, which can at times be difficult to interpret.

Keywords: BLAST; E-value; Sequence Comparison

Title: Resting with Rhythm: Brain Functional Network Connectivity and Music Habits in Adolescents

Author list: Anaiah Calhoun1, Armin Iraji1

Detailed Affiliations:

1Tri-institutional center for Translational Research in Neuroimaging and Data Science (TReNDS), Georgia State, Georgia Tech, Emory, Atlanta, GA, USA

Abstract: Music plays a key role in daily life, yet many of its effects remain unseen. There are the visible factors, such as seeing people dance or sing, but music has also been shown to slow heart rate and increase dopamine. Functional neuroimaging allows us to noninvasively quantify brain activity and examine how music influences or relates to brain function. In this study, we leveraged a large resting-state functional magnetic resonance imaging (rs-fMRI) data of over 7000 adolescents (ages 9-13) from the Adolescent Brain Cognitive Development (ABCD) study to analyze how differentiating patterns of functional network connectivity (FNC), the degree to which each network is co-fluctuating with other networks, are associated with documented music listening duration. We hypothesize that networks involved in authority processing and the salience system show significant associations with music listening1-4.

We used NeuroMark 2.2, an automated independent component analysis (ICA) pipeline, to isolate 105 reproducible multiscale intrinsic connectivity networks (ICNs) from preprocessed rs-fMRI data5. These ICNs cover key functional domains such as auditory, visual, sensorimotor, attention, default mode, salience, and cognitive control. For each participant, the FNC matrix was generated by computing pairwise temporal correlations between the ICN time courses.

We compared FNC associations with the number of hours per week participants self-reported listening to music (sai_p_lmusic_hours), using baseline (N=7797) data. While many of the ABCD music-related variables were limited to a binary response, the use of this continuous variable allowed for a deeper analysis. We consistently found that the more often one listens to music, the stronger connectivity in networks involving auditory processing and dorsal attention. We observed that the FNC between ICNs in the salience subdomain of the triple network domainand those in the insular-temporal and temporoparietal subdomains of the higher cognition domain was negatively associated with the duration of music listening. Notably, all of these regions are involved in salience and authority processing, supporting our hypothesis.

Our findings suggest that frequent engagement with music may facilitate more efficient integration of sensory and cognitive processes, thereby reducing the functional communication demands between these regions, even during resting states. This relationship suggests that music listening may influence the intrinsic functional organization of the brain. By leveraging a large-scale dataset and through multiscale NeuroMark, the reliability and generalizability of our findings are strengthened. Collectively, these results offer a footing for future investigations into how long-term music engagement may contribute to mental health outcomes.

Workshop – Advances in Target Discovery and Computational Drug Design

August 4th
9:20 AM – 12:20 PM
Room: 320

Chairs: Pengyue Zhang, Yijie Wang

Title: Drug repurposing for substance use disorders by genome-wide association studies and real-world data analyses

Author list: Dongbing Lai

Detailed Affiliations:
Indiana University

Abstract: Substance use disorders (SUDs, including alcohol, cannabis, opioids, nicotine, etc.) represent significant public health challenges and the prevalence is rising rapidly. Many individuals with SUDs use more than one substance and they often experience increased risks of overdose, mental health problems, and chronic diseases, and interactions of multiple substances complicate diagnosis and treatment. Drugs treating SUDs are available; however, their efficacy is limited, and relapse rates of SUDs remain high. Studies have shown that genetic factors contribute to ~50% variations of SUDs and there exist genes responsible for multiple SUDs (SUD-shared genes). Repurposing drugs targeting SUD-shared genes provides an efficient and effective way to develop novel drugs to treat SUDs, especially for those using multiple substances. In this study, we conducted the largest genome-wide association studies of SUDs to date to identify SUD-shared genes using samples from European, African, and Latinx ancestries (N=1,683,739). We innovatively considered variants having the same directions of effects across different SUDs as SUD-shared and developed a pipeline to prioritize SUD-shared genes. In total, 220 loci were identified with 40 novel loci not reported as SUD-associated. We prioritized 785 SUD-shared genes and identified 183 FDA approved drugs targeting these genes. By using a large real-world data from Optum® Clinformatics®, 7 drugs showed significantly reduced hazard ratio (HR) to develop SUDs: Topiramate (HR=0.44, 95% confidence interval (CI): 0.42 -0.47), Aripiprazole or Cariprazine (HR=0.88, 95% CI: 0.78-0.88), Desipramine, Imipramine, or Nortriptyline (HR=0.89, 95% CI: 0.84-0.94), Methylphenidate (HR=0.84, 95% CI: 0.78-0.91), demonstrating that they may be repurposed to treat SUDs.

Keywords: Drug repurposing, genome-wide association studies, substance use disorders, substance use disorders-shared genes, real-world data analysis.

Title: An Informatics Bridge to Improve the Design and Efficiency of Phase I Clinical Trials for Anticancer Drug Combinations

Author list: Lei Wang

Detailed Affiliations:

The Ohio State University

Abstract: Prior preclinical and clinical knowledge is critical for designing effective and efficient cancer drug combinatory trials. In this study, we summarized critical databases of drug combination toxicity and pharmacokinetics. We further conducted a feasibility and utility study that demonstrates how different data sources can contribute to and assist phase I trial designs. Single-drug and drug combination toxicity and pharmacokinetic data were primarily reviewed from several databases. We focused on the MTD, dose-limiting toxicity (DLT), toxicity, and pharmacokinetic profiles. To demonstrate the feasibility and utility of these data sources in improving trial designs, phase I studies reported in ClincalTrials.gov from January 1, 2018 to December 31, 2018 were used as examples. We evaluated whether and how these studies could have been designed differently given toxicity and pharmacokinetic data. None of the existing pharmacokinetic and toxicity databases contain either MTD or DLT. Among 268 candidate trials, four drug combinations were studied in other phase I trials before 2018; 185 combinations had complete or partial information on drug interactions or overlapping toxicity, and 79 combinations did not have available information. Two drug combination trials were selected as case studies. The nivolumab-axitinib trial could have been designed as a dose deescalating study, and the vinorelbine-trastuzumab emtansine trial could have been designed with a lower dose of either drug. Public data sources contain significant knowledge of the drug combination phase I trial design. Some important data (MTD and DLT) are not available in existing databases but in the literature. Some phase I studies could have been designed more efficiently with additional preliminary data.

Keywords: Phase I clinical trial; cancer; drug combination; knowledge base

Title: Building an explainable graph neural network by sparse learning for the drug-protein binding prediction

Author list: Yijie Wang

Detailed Affiliations:
Indiana University

Abstract: Explainable Graph Neural Networks have been developed and applied to drug-protein binding prediction to identify the key chemical structures in a drug that have active interactions with the target proteins. However, the key structures identified by the current explainable GNN models are typically chemically invalid. Furthermore, a threshold must be manually selected to pinpoint the key structures from the rest. To overcome the limitations of the current explainable GNN models, we propose SLGNN, which stands for using Sparse Learning to Graph Neural Networks. It relies on using a chemical-substructure-based graph to represent a drug molecule. Furthermore, SLGNN incorporates generalized fused lasso with message-passing algorithms to identify connected subgraphs that are critical for the drug-protein binding prediction. Due to the use of the chemical-substructure-based graph, it is guaranteed that any subgraphs in a drug identified by SLGNN are chemically valid structures. These structures can be further interpreted as the key chemical structures for the drug to bind to the target protein. We test SLGNN and the state-of-the-art competing methods on three real-world drug-protein binding datasets. We have demonstrated that the key structures identified by our SLGNN are chemically valid and have more predictive power.

Keywords: Graph Neural Network, Interpretable model, Sparse learning, Drug-protein binding prediction

Title: Combining genetics and real-world patient data fuel ancestry-specific target and drug discovery in Alzheimer's disease

Author list: Yuan Hou

Detailed Affiliations:
Cleveland Clinic

Abstract: Although high-throughput DNA/RNA sequencing technologies have generated massive genetic and genomic data in human disease, translation of these findings into new patient treatment has not materialized. Method: To address this problem, we have used Mendelian randomization (MR) and large patient's genetic and functional genomic data to evaluate druggable targets using Alzheimer's disease (AD) as a prototypical example. We utilized the genetic instruments from 6 celltype specific eQTLs and tested the outcome of MR independently across 7 genome-wide association studies (GWAS). Results: We identified 25 drug targets for AD. We pinpointed that the inflammatory target of epoxide hydrolase 2 (EPHX2) emerged as a potent AD target in EAs, and treatment of AD transgenic rats with an EPHX2 inhibitor was therapeutic. We also identified 23 candidate drugs associated with reduced risk of AD in mild cognitive impairment (MCI) patients after analysis of ~80 million electronic health records. Using a propensity score-matched design, we identified that usage of either apixaban (hazard ratio [HR] = 0.74, 95% confidence interval [CI] 0.69 – 0.80) and amlodipine (HR = 0.91, 95% CI 0.88 – 0.94) were both significantly associated with reduced progression to AD in people with MCI. Conclusion: In summary, combining genetics and real-world patient data identified ancestry-specific therapeutic targets and medicines for AD and other neurodegenerative diseases if broadly applied.

Keywords: Mendelian randomization, AD, GWAS, EPHX2, drug, target

Title: Identifying repurposable treatments in patient subpopulations.

Author list: Pengyue Zhang

Detailed Affiliations:
Indiana University

Abstract: Real-world data mining has the potential to identify precise relationships between drug responses and patient characteristics. We investigated drug responses in Alzheimer's disease (AD) with a special awareness on patient characteristics. In a multidisciplinary study, we observed both real-world evidence and genetic associations supporting telmisartan as a candidate repurposable drug for AD in African Americans. Additionally, we identified candidate repurposable drugs for AD in patients with neuroinflammation-related conditions. **Keywords:** Alzheimer's disease, drug repurposing, neuroinflammation, real-world data, subpopulation

Title: Pan-Cancer Analysis of the Immune Microenvironment's Role in Tumor Genomic Evolution

Author list: Elaine Li[1], Li Liu[2]

Detailed Affiliations

[1]College of Arts and Sciences, Emory University, Atlanta, GA, USA; [2]College of Health Solutions, Arizona State University, Phoenix, AZ, United States

Abstract

The tumor microenvironment (TME) plays a critical role in tumorigenesis, progression, and response to treatment. By imposing selective pressures to promote or suppress the growth of specific cell populations (subclones), TME may significantly modulate the evolutionary trajectory of a tumor. However, the relationship between TME and tumor evolutionary dynamics are poorly understood. In this study, we integrated whole exome sequencing and transcriptomic data from tumors in The Cancer Genome Atlas (TCGA) project to investigate how TME shapes subclonal evolution.

For each tumor, we used the TEATIME program to infer evolutionary parameters including subclone emergence time, selection strength, and growth rate from somatic exome variants. We retrieved pre-computed estimates of the abundances of various immune cell types generated by CIBERSORT and xCell based on transcriptomic data. We then conducted a pan-cancer analysis to test the relationship between tumor evolutionary dynamics and immune cell infiltration. Using Pearson correlation and hierarchical clustering, we identified a cluster of T follicular helper (Tfh) cells, $CD4^+$ T cells, and activated natural killer (NK) cells that correlated with delayed subclonal emergence (pickt1, $r = 0.31$, FDR < 0.01) and later evolutionary endpoints (picktend). Conversely, neutrophils, naïve B cells, and M2 macrophages consistently had a negative correlation with the pickt1 and picktend TeaTime parameters, linked to accelerated evolution. These patterns were further validated through LASSO-based multivariate regression (glmnet, $\alpha = 1$), which selected Tfh cells as top predictors of delayed evolution. We used tumor purity as a covariate in the multiple regression, and found it to be frequently retained across parameters, indicating that it played a modulatory role in immune–evolution dynamics.

To determine patterns within specific cancer types, we stratified the data by cancer types that had a sufficient sample size ($n > 20$), which included BRCA (breast cancer), KIRC (renal clear cell carcinoma), and LUAD (lung adenocarcinoma). With these subsets, we repeated the correlation and LASSO analyses. We found that relationships between immune cells and tumor evolution were different depending on the type of cancer: for example in LUAD, $CD8^+$ T and NK cells negatively correlated with selection strength. Our results suggest that immune infiltration can constrain or accelerate tumor evolution both in specific cancers and across all cancer types. These findings present new information about our understanding of the evolutionary influence that the immune microenvironment has on cancer.

Keywords

Tumor evolution, immune microenvironment, TEATIME, LASSO regression, correlation analysis, hierarchical clustering

Workshop – Biosignals and Omics in Neurological and Cancer Diseases: Opportunities and Challenges

August 4th

9:20 AM – 12:20 PM

Room: 350

Chairs: Haoqi Sun, Chen Huang

Title: Bioinformatics Meets Biosignals: Opportunities and Challenges

Author list: Haoqi Sun[1]
Detailed Affiliations:
[1]Department of Neurology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA
Abstract: Biosignals, such as brain wave (electroencephalography, EEG) during sleep, contain rich information about the function and health. However, the intersection between biosignals and bioinformatics is understudied due to the lack of interdisciplinary collaborations. Here, I describe opportunities and challenges in this field, focusing on the molecular basis of sleep electrophysiology as measured by EEG microstructures, as well as their implications on brain health.
Keywords: Biosignal, omics, sleep, electroencephalography

Title: Leveraging Clinical Biobanks and Genetics to Understand Sleep Apnea and Related Comorbidities
Author list: Brian Cade[1]
Detailed Affiliations:
[1]Department of Medicine, Brigham Women's Hospital, Harvard Medical School, Boston, MA, USA
Abstract: We have recently identified associations between sleep apnea (OSA) and hundreds of diseases in a large clinical biobank. Validation of these associations using sleep clinic data, identifying shared genetic architecture, and organizing prioritized comorbidities into multimorbidity clusters are important next steps to improve our understanding of sleep apnea and its consequences. In this proposed talk, I will describe our approach to phenotyping thousands of clinical sleep recordings using advanced polysomnographic traits (endotypes and burdens) that have increased associations with comorbidities compared to traditional measures of sleep apnea. Multivariate analyses indicate that no single polysomnographic trait best captures these case-control associations. We have started to measure the heritability and genetic correlations of OSA endotypes and burdens and anticipate performing multivariate genome-wide association studies to improve study power. Finally, I will describe recent analyses to group associated comorbidities into age-dependent topic models.
Keywords: Genomics, genome-wide association studies, obstructive sleep apnea, endotype

Title: Sleep Architecture Biomarkers of Psychiatric Disease
Author list: Shaun Purcell[1]
Detailed Affiliations:
[1]Department of Psychiatry, Brigham Women's Hospital, Harvard Medical School, Boston, MA, USA
Abstract: Sleep can now be measured using increasingly scalable and non-invasive sensor technologies, making it an attractive potential source of future novel objective biomarkers, with applications across a range of physical and mental conditions. Here I outline applications to psychiatric disease (primarily schizophrenia) and cognitive aging, and discuss some of the challenges faced when developing biomarkers based on sleep biosignals.

Keywords: sleep, psychiatric disease, biosignal

Title: Reimagining Sleep Medicine using AI-based Physiology-guided Digital Twins
Author list: Ankit Parekh[1]
Detailed Affiliations:
[1] Icahn School of Medicine at Mount Sinai, New York, NY, USA
Abstract: Despite affecting over a billion people worldwide, sleep disorders such as obstructive sleep apnea (OSA) remain poorly characterized in terms of symptom severity, treatment response, and long-term health consequences. Conventional metrics like the Apnea-Hypopnea Index (AHI) fall short in capturing the true physiological burden of these conditions and offer limited insight into patient heterogeneity. In this talk, I will present a transformative framework for sleep medicine—anchored in the development of physiology-guided, AI-based digital twins. These virtual representations of patients integrate time-series data from sleep studies (e.g., EEG, airflow, oxygen saturation) with multimodal clinical inputs to simulate disease trajectories, predict outcomes, and personalize therapy.
Keywords: sleep, digital twin, artificial intelligence

Title: Multi-omics in Neurodegenerative Diseases
Author list: Bruno Benitez[1]
Detailed Affiliations:
[1]Department of Neurology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA
Abstract: Integrating multiple omics by aligning genomic, proteomic, and metabolomic data provides a comprehensive view of biological and pathological processes. Genome sequencing enables a comprehensive analysis of the entire genome, while bulk and single-cell transcriptomics provide an extensive view of the transcriptome. Meanwhile, advancements in proteomics and metabolomics have not kept up with this technology. There is a need for a statistical framework or bioinformatics tools to enable the seamless integration of these large datasets. Artificial intelligence is beginning to bridge these gaps. The application of cross-omics to neurodegenerative diseases has allowed us to stratify patients based on their molecular landscape.
Keywords: Multi-omics, neurodegenerative diseases, artificial intelligence

Title: Y-chromosome loss in cancer: single-cell insights into origins and consequences
Author list: Shiwei Yin[1], Yusi Fu[1], Jun Xia[1]
Detailed Affiliations:
[1] Institute of Biosciences and Technology, Texas A&M University, Houston, TX, 77030, USA
Abstract: Mosaic loss of the Y chromosome (LOY) is observed in cancer, aging, and cardiomyopathy; however, its origins, mechanisms, and functional impact remain insufficiently understood. Bulk sequencing has limited sensitivity for capturing copy-number heterogeneity, leaving the landscape of LOY in clinical specimens unresolved. To address this gap, we applied an ultra-sensitive single-cell DNA copy-number method—capable of detecting copy-number alterations as low as 10 kb resolution—to thousands of cells from multiple gastrointestinal (GI) cancers. Our analysis quantified LOY frequency across major GI tumor

types and determined whether LOY could be detected in peripheral blood mononuclear cells (PBMCs) or arose solely as a somatic event within tumors. We further characterized the extent of genomic instability in LOY clones, revealing frequent co-occurrence of additional genomic events. Integrating single-cell RNA-seq data showed that LOY correlates with altered immune cell compositions and key immune-response genes, suggesting a potential role in modulating sensitivity to immunotherapy. In parallel, longitudinal culture of GI cancer and precancer cell lines demonstrated the progressive expansion of LOY subclones, shedding light on the dynamics of Y-chromosome loss over time. Collectively, these single-cell genomic analyses offer a high-resolution view of copy-number evolution in GI cancers, clarify the sources and consequences of LOY, and identify potential biomarkers and therapeutic targets associated with Y-chromosome loss.
Keywords: LOY, single-cell, copy-number, gastrointestinal, neoantigen, immunotherapy

Title: Computational Techniques for Deciphering Cancer Genomics and the Tumor Microenvironment at Single-Cell Resolution
Author list: Jinzhuang Dou[1]
Detailed Affiliations:
[1]Biomedical Informatics and Data Science, The University of Alabama at Birmingham, AL, USA
Abstract: Our understanding of cancer gene mutations and how cancer evades the immune system has led to the development of targeted therapies and immunotherapies. Single-cell sequencing further enhances these treatments by revealing tumor genetic heterogeneity and elucidating their interactions with immune cells. In this talk, I will present our computational efforts to advance cancer research at single-cell resolution. First, I will introduce Monopogen, a computational tool for single-nucleotide variant (SNV) calling in single-cell sequencing data. Leveraging Monopogen maximizes the genetic information from available single-cell sequencing data, leading to immediate benefits in genetic ancestry mapping and somatic clonal lineage delineation. Second, I will demonstrate a novel mathematical solution, bi-order canonical correlation analysis (bi-CCA), which iteratively aligns rows and columns between data matrices. Bi-CCA effectively integrates two distinct single-cell modalities derived from the same sample. Through bi-CCA, we deepen our understanding of immune cell therapy processes from a comprehensive multi-omic perspective. Looking forward, these computational tools hold great promise for uncovering new insights and improving personalized cancer treatments.
Keywords: Cancer evolution, Immunotherapy, Single cell, Ancestry, Multi-omics

Title: Distinct Signatures of Tumor-Associated Macrophages in Shaping Immune Microenvironment and Patient Prognosis

Author list: Chongming Jiang[1]

Detailed Affiliations:
[1] Terasaki Institute for Biomedical Innovation, Los Angeles, CA 90024, USA

Abstract: Background: Renal cell carcinoma (RCC) comprises 90% of adult kidney cancers, characterized by significant heterogeneity within its tumor microenvironment. This study tests the hypothesis that tumor-associated macrophages (TAMs) influence RCC progression and treatment responses. Using immunomics, we investigated the prognostic value of TAM signatures in the RCC tumor immune microenvironment (TIME). **Methods**: Single-cell RNA sequencing data from RCC patients were analyzed to develop eight

distinct TAM signatures. A machine learning model predicting patient survival was built using TCGA data and validated across multiple independent RCC cohorts. Model performance was evaluated using Kaplan-Meier survival analysis, receiver operating characteristic (ROC) curves, principal component analysis (PCA) visualization. **Results**: We identified diverse TAM subpopulations within the RCC TIME, highlighting significant prognostic implications. Specific TAM signatures correlated strongly with patient survival, macrophage infiltration, and known TAM markers. A TAM risk model effectively stratified patients into distinct risk categories, with the low-risk group showing significantly improved overall survival. **Conclusions**: Our findings clarify the complex roles of TAMs within RCC and their impact on patient prognosis. The established TAM risk model offers valuable prognostic markers and identifies potential therapeutic targets to enhance RCC treatment efficacy.

Keywords: tumor-associated macrophages, renal cell carcinoma, prognosis; tumor immune microenvironment, machine learning

Workshop – AI and Applications for Better Understanding Disease Mechanisms

August 4th
2:20 PM – 5:20 PM
Room: 320

Chairs: Xubo Song

Title: Reprogramming Protein Language Models for Protein Function Annotation and Engineering
Author list: Yunan Luo
Detailed Affiliations:
Georgia Institute of Technology
Abstract: In this talk, I will introduce our recent works on protein function annotation and protein engineering. I will describe a suite of machine learning methods that reprogram protein language models to address the challenges of data bias and data scarcity in these problems.

Title: MARVEL: Microenvironment Annotation by Supervised Graph Contrastive Learning
Author list: Yuying Xie
Detailed Affiliations:
Michigan State University
Abstract: Recent advancements in in situ molecular profiling technologies, including spatial proteomics and transcriptomics, have enabled detailed characterization of the microenvironment at cellular and subcellular levels. While these techniques provide rich information about individual cells' spatial coordinates and expression profiles, extracting biologically meaningful spatial structures from the data remains a significant challenge. Current methodologies often rely on unsupervised clustering followed by

cell type annotation based on differentially expressed genes within each cluster and most of the time will require other information as the reference (e.g., HE-stained images). This is labor-intensive and demands extensive domain knowledge. To address these challenges, we propose a supervised graph contrastive learning framework, MARVEL. MARVEL is a supervised graph contrastive learning method that can effectively embed local microenvironments represented by cell neighbor graphs into a continuous representation space, facilitating various downstream microenvironment annotation scenarios. By leveraging partially annotated examples as strong positives, our approach mitigates the common issues of false positives encountered in conventional graph contrastive learning. Using real-world annotated data, we demonstrate that MARVEL outperforms existing methods in three key microenvironment-related tasks: transductive microenvironment annotation, inductive microenvironment querying, and the identification of novel microenvironments across different slices.

Title: Leveraging AI for Characterizing Pediatric Cancer
Author list:: Shibiao Wan
Detailed Affiliations:
University of Nebraska
Abstract: Pediatric cancers, such as medulloblastoma or leukemia, are very heterogeneous. Characterizing pediatric cancers is an essential step for downstream risk stratification and tailored treatment design. Conventional wet-lab approaches for pediatric cancer characterization are time-consuming, costly and laborious. Artificial intelligence (AI) and machine learning (ML) would be an efficient alternative for assisting in characterizing different types of pediatric cancer. In this talk, I will discuss our recent progress by developing different types of AI/ML approaches to characterize two common pediatric cancer, including medulloblastoma and leukemia. We believe our proposed approaches will bring significantly positive impacts on downstream diagnosis, prognosis, and treatment of different pediatric cancers.

Title: Deep Learning models for image enhancement, translation, and harmonization
Author list:: Xubo Song
Detailed Affiliations:
Oregon Health & Science University
Abstract: While machine learning and deep learning are accelerating biological and medical discovery, their performances deteriorate rapid when image quality is poor, when undesired factors bring variabilities which obscure true physiological changes, and when data is too sparse to train robust models. We will discuss our recent work to address these challenges. Specifically, we will present physics-guided diffusion restoration of optical aberrations in whole-slide microscopy, zero-shot medical image translation via structural guidance, and harmonization of histopathology images.

Title: Advancing AI for Individualized Diagnosis and Prognosis: From Prenatal Heart Defects to Prostate Cancer Survival
Author list:: Jieqiong Wang
Detailed Affiliations:
University of Nebraska

Abstract: Artificial intelligence (AI) offers powerful tools for enhancing both diagnosis and prognosis in clinical care. In this talk, I will present two translational AI projects aimed at improving individualized decision-making. The first focuses on leveraging fetal ultrasound imaging and deep learning models to support early and accurate diagnosis of congenital heart disease (CHD), with the goal of reducing disparities in prenatal detection, particularly in underserved populations. The second project involves developing AI models based on tabular clinical data to predict personalized survival curves for prostate cancer patients, offering a data-driven approach to individualized risk stratification. Together, these projects demonstrate the promise of AI in bridging imaging and non-imaging modalities to improve health outcomes across diverse clinical domains.

Title: Large Scale, AI-Enabled, Spatial Signal Processing of Breast Cancer Pathology Identifies Consensus Tissue Structures Related to Biology and Outcomes

Author list: Jordan E. Krull[1,2], Mirage Modi[1], Karthik Chakravarthy[3], Daniel Spakowicz[2,3], Qin Ma[1,2]
Detailed Affiliations:
[1]Department of Biomedical Informatics, Columbus, OH, USA 43210; [2]Pelotonia Institute for Immuno-Oncology, The Ohio State University Comprehensive Cancer Center – The James, Columbus, OH, USA 43210; [3]Division of Medical Oncology, The Ohio State University Comprehensive Cancer Center – The James, Columbus, OH, USA 43210

Abstract: Breast cancer remains a leading cause of cancer-related mortality worldwide, in women, underscoring the critical need for innovative diagnostic and prognostic approaches. Spatial evaluation of tissue structure in breast cancer provides valuable insights into the tumor microenvironment, including cellular organization, stromal interactions, and molecular heterogeneity. However, large-scale, high-resolution profiling of tumor structure is not financially nor logistically feasible. In this study, we sought to interrogate spatial signal organization in a large set of breast pathology images to identify common structural features and associated biology, in an unsupervised manner, and connect the presence of these features to clinically relevant endpoints. To accomplish this objective, we utilized a large cohort (n=1988) of normal breast and breast cancer biopsy digital, whole-slide, pathology images (WSI), from TCGA (n=1058), OSU's TCC (n=401), and GTEx (n=529). WSI processing and QC included color normalization and artifact segmentation. Quality tissue areas were tiled into non-overlapping 224μm x 224μm boxes of 1μm/px and feature embeddings were subsequently extracted from tiles using a deep learning pathology foundation model (CTransPath). Notably, we tested Resnet-50, UNI, and CTransPath as embedding models, and CTransPath provided the most modularity and was not as sensitive to tissue size limitations and color variations as UNI and Resnet-50. Each embedding was subsequently converted to spatial Fourier coefficients (FCs), specific to each sample, using spatial graph Fourier transform, and converted to feature spatial-coordination maps generated from cosine similarity of embedding FCs. To identify conserved tissue structures, we developed a customized graph-neural-network (GNN), trained to identify a common latent feature space of all tile embeddings, based on spatial similarity among all samples, and performed Louvain clustering of the resultant feature latent space. We identified 21 conserved tissue structures (FTU) consisting of 5 to 78 spatially coordinated image embedding values. We quantified a probabilistic feature prominence in each sample through reconstruction and compared sample FTU quantitation to gene expression programs and clinical features in 1459 breast cancers from TCGA and TCC. Using a 36 marker CODEX panel on a breast cancer H&E+CODEX slide pair, we confirmed cell type architectures like tumor, stroma, and immune rich regions, overlapping with reconstructed FTUs, highlighting their biological significance. Across 1230 unique breast cancer samples, almost half of the FTU's correlated strongly with molecular

subtypes (ER, PR, HER2 status, $p<0.05$) and several exhibited significant hazards for overall survival, independent of subtype. Every identified FTU associated with bulk gene expression modules including those enriched in lymphocyte infiltration, which paired immune related gene expression modules with at least two FTUs. We additionally, computed associations between FTU's and mutations as well as inferred microbial counts, finding significant associations between FTUs and these tumor intrinsic and extrinsic biological factors, illuminating the influence genomic and host factors may have on local tissue structure, detectable in H&E images. This work highlights the growing appreciation for spatial ecology in tumors and particularly breast cancer tissue structure. We also demonstrate that computationally identifiable, conserved tissue structures in breast cancer, derived from digital pathology, can serve as a biomarker for diagnosis and prognosis, with direct relationship to biology and clinical course.

Title: Evaluate, standardize, and optimization bioinformatics software documentation using AI-agents

Author list: Shaopeng Gu[1], Cankun Wang[1,2], Shaohong Feng[1], Qin Ma[1,2], Anjun Ma[1,2,*]

Detailed Affiliations:

[1]Department of Biomedical Informatics, Columbus, OH, USA 43210; [2] Pelotonia Institute for Immuno-Oncology, The Ohio State University Comprehensive Cancer Center – The James, Columbus, OH, USA 43210

Abstract: The exponential growth of omics data and the proliferation of computational tools have transformed biomedical research, enabling the exploration of complex biological systems at unprecedented resolution. Thousands of tools now exist for single-cell RNA sequencing, spatial transcriptomics, and other omics technologies, many powered by advanced artificial intelligence (AI) and deep learning methods. However, significant challenges remain in usability, reproducibility, and scalability—often stemming from inconsistent and non-standardized software documentation. These issues limit access for non-specialist users, reduce community engagement, and hinder reproducibility. While large language models (LLMs) are beginning to show promise in multi-omics data analysis, their potential for improving software engineering and documentation remains largely underexplored. To address this gap, we developed BioGuider, an innovative LLM-powered platform designed to standardize and enhance documentation during omics tool development. BioGuider operates through multiple specialized AI agents that (1) parse and analyze source code and documentation to identify logical structure, (2) evaluate documentation quality and reproducibility using an in-house scoring framework, and (3) automatically revise and generate key documentation components—including in-line code comments, user tutorials, vignettes, README files, and user guides. BioGuider is the first chat-based assistant specifically built for bioinformatics tool development, establishing a new standard for documentation generation. These standards quantitatively evaluate the readability, documentation density, code-documentation ratio, structure quality, content quality, example coverage. Beyond assessing existing packages, we envision BioGuider as a proactive guide that helps developers meet publication and journal documentation standards prior to manuscript submission.

Keywords: Software development, quality control, documentation standardization, large language model

2:20 PM – 5:20 PM
Room: 301

Chair: Juexin Wang

Title: Benchmarking Cellular Deconvolution Algorithms to Predict Cell Proportions: A Literature Review
Author list: Ayla Bratton, Ayesha Malik, Jiao Sun, Qian Li and Wei Zhang
Abstract: Computational cellular deconvolution has recently emerged as a powerful alternative to traditional experimental approaches for analyzing RNA-sequencing (RNA-seq) data. Instead of relying solely on physical separation techniques or histological analysis, researchers can now use in silico methods, such as statistical modeling and machine learning, to estimate the cellular composition of complex tissue samples. A key application of these methods is the deconvolution of bulk RNA-seq data using a reference derived from single-cell RNA-seq data. This review focuses on six widely used reference-based deconvolution algorithms: BayesPrism, CIBERSORTx (CSx), DISSECT, Scaden, TAPE, and scpDeconv. Each method employs a different modeling strategy and offers unique strengths, depending on the context of use. To benchmark these algorithms, both pseudobulk RNA-seq data and single-cell reference data were generated from a publicly available human cerebral cortex scRNA-seq dataset. A ground truth cell fraction matrix and a cell type–specific gene expression signature matrix were also constructed from the same dataset to enable evaluation. The pseudobulk samples were input into each deconvolution algorithm, with the corresponding single-cell data used as the reference. After running the models, the estimated cell type proportions were compared against the ground truth to assess performance. The evaluation revealed that while some algorithms excel at accurately estimating cell type proportions, others also provide reliable predictions of gene expression for individual cell types. As computational deconvolution continues to evolve, selecting an appropriate method for a given dataset and biological question remains a critical step in transcriptomic analysis.
Keywords: cellular fraction prediction; cell-type-specific gene expression profile; scRNA-seq

Title: Landscape of gene essentiality in cancer cell death pathways
Author list: Shangjia Li, Zhimo Zhu, Chen Yang, Nuo Sun, Lijun Cheng and Lang Li
Abstract: Regulated cell death (RCD), a process that relies on a series of molecular mechanisms, can be targeted to eliminate superfluous, irreversibly damaged, and potentially harmful cells. To better understand how the cell death pathway contributes to cancer therapy, we studied 1150 cancer cells in Dependency Map (DepMap) database for 12 distinct cell death pathways and assessed their gene essentialities. Genes who are essential in 90% or more cancer cell lines are called always essential; or partial essential if falling into (10%, 90%); or rare essential if they are essential in less than 10% of cancer cell lines. Overall, among these 12 cell death pathways, 23, 47, 551 genes were classified as always essential, partial essential, and rare essential, respectively. In two cell death pathways, Parthanatos, and Pyroptosis, all genes were rare essential. Among the other nine cell death pathways, Apoptosis, Autosis, Necroptosis, Efferocytosis, Ferroptosis, Mitotic cell death, Autophagy, Lysosome cell death, MPT driven necrosis and Immunogenic, there are (10, 1, 13, 6, 3, 6, 11, 1,1,0) partial essential genes (PEG), and (2,1,3,1,1,11,4,0,0,1) always essential genes (AEG). As of the date we collected the data, eleven AEGs and eighteen PEGs did not have targted drugs that under-going clinical trials. These cell death pathways essential genes could be viable targets for therapeutic drug development for cancer therapies.
Keywords: cell death pathway; gene essentiality; pan-cancer

Workshop – Integrative Genomics and Epigenomics to Link GWAS Variants to Function

August 4th
2:20 PM – 5:20 PM
Room: 350

Chairs: Hongbo Liu, Kaixiong Ye

Title: Precision Nephrology: The Role of Genetics in Kidney Health
Author list: Atlas Khan
Detailed Affiliations:
Division of Nephrology, Department of Medicine, Vagelos College of Physicians & Surgeons, Columbia University, New York, NY, USA
Abstract: Chronic kidney disease (CKD) affects an estimated 10-13% of the global population and represents a major cause of morbidity, premature mortality, and healthcare burden worldwide. The pathogenesis of CKD is influenced by a complex interplay of genetic and environmental factors. Over the past decade, genome-wide association studies (GWAS) have identified hundreds of common variants associated with estimated glomerular filtration rate (eGFR), a central biomarker for kidney function. These discoveries have led to the development of genome-wide polygenic scores (GPS) that aggregate the effects of common variants to quantify an individual's genetic predisposition to CKD. In our recent work, we developed and validated a CKD GPS across diverse ancestral populations, demonstrating its utility for population-level risk stratification for CKD (**Khan et al. Nature Medicine 2022**).
While GPS provides a valuable tool for risk stratification in the general population, it fails to capture rare, high-penetrance protein-coding variants that underlie monogenic kidney diseases. Notably, disorders such as autosomal dominant polycystic kidney disease (ADPKD) and COL4A-related nephropathies, caused by pathogenic variants in *PKD1*, *PKD2*, and *COL4A3-5*, account for a significant portion of early-onset and progressive CKD. To address this gap, we leveraged large-scale exome sequencing data from population-based cohorts, including the UK Biobank and All of Us Research Program, to study the combined impact of polygenic and monogenic variation on kidney disease risk (**Khan et al., Nature Communications, 2022**).
In this talk, I will present a comprehensive risk prediction framework that integrates common and rare genetic variation to more accurately assess CKD risk. Our approach enhances the precision of risk prediction, particularly for individuals with an intermediate GPS who also carry a high-risk monogenic variant, and might improve our ability to detect early kidney dysfunction across diverse ancestral backgrounds.
Keywords: CKD, GWAS, PRS

Title: Unraveling the Molecular Heterogeneity of Severe Acute Malnutrition: Multi-omic Insights

Author list: Natasha C. Lie[1,2], Yixing Han[2], Qing Li[2], Aarti Jajoo[2], Aparna Haldipur[2], Emilyn Banfield[2], Shanker Swaminathan[3], Sharon Howell[4], Orgen Brown[4], Roa Sadat[3], Nancy J. Hall[3], Kwesi Marshall[4], Katharina V. Schulze[3], Thaddaeus May[5], Marvin E. Reid[4], Carolyn Taylor-Bryan[4], Mark J. Manary[6,7,8], Indi Trehan[7,9], Mamana Mbiyavanga[10], Wisdom A. Akurugu[10], Colin A. McKenzie[4^], Dhriti Sengupta[11], Elizabeth G. Atkinson[3,12], Ananyo Choudhury[11], Neil A. Hanchard[1,2,3,6,*]

Detailed Affiliations:

[1] Graduate Program in Integrative Molecular and Biomedical Sciences, Baylor College of Medicine, Houston, TX, USA. [2] Childhood Complex Disease Genomics Section, National Human Genome Research Institute, National Institutes of Health, Bethesda, USA. [3] Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA. [4] Tropical Metabolism Research Unit, Caribbean Institute for Health Research, University of the West Indies, Mona, Jamaica. [5] Department of Internal Medicine, Baylor College of Medicine, Houston, TX, USA. [6] USDA/ARS/Children's Nutrition Research Center, Baylor College of Medicine, Houston, TX, USA. [7] Departments of Paediatrics and Child Health and Community Health, Kamuzu University of Health Sciences, Blantyre, Malawi. [8] Department of Pediatrics, Washington University in St. Louis, St. Louis, MO, USA. [9] Departments of Pediatrics, Global Health, and Epidemiology, University of Washington, Seattle, WA, USA. [10] Computational Biology Group, Faculty of Health Sciences, University of Cape Town, Western Cape, South Africa. [11] Sydney Brenner Institute for Molecular Bioscience (SBIMB), University of the Witwatersrand, Johannesburg, South Africa. [12]Jan and Dan Duncan Neurological Research Institute, Texas Children's Hospital, Houston, TX, USA

Abstract: Severe acute malnutrition (SAM) remains a global health emergency, directly or indirectly responsible for over 400,000 childhood deaths each year. Among its clinical forms, edematous SAM (ESAM, including kwashiorkor and marasmic-kwashiorkor) is more lethal than non-edematous SAM (NESAM or marasmus), despite comparable nutritional deficiencies and environments. ESAM is also the predominant form in populations from East Africa and the Caribbean. However, the molecular mechanisms underlying why some children get ESAM and others NESAM remain poorly understood.

To address this gap, we have developed a multi-omic framework to interrogate clinically phenotyped cohorts from Jamaica and Malawi. Our initial analyses included genome-wide DNA methylation profiling of buccal cells from 309 children revealed widespread hypomethylation in ESAM cases, particularly at genes implicated in metabolic and liver-related pathways. These epigenetic alterations were absent in adults recovered from SAM, pointing to disease-specific, dynamic methylation changes potentially driven by disrupted OCM. In parallel, intracontinental admixture mapping and targeted genotyping of 103 genes involved in one-carbon metabolism (OCM) across 711 children. We identified seven loci—such as *MTHFR*, *PRICKLE2*, and *PLD2*— with evidence of significant association with ESAM. These loci were enriched on East African ancestral haplotypes and located within genomic regions under recent positive selection, suggesting a potential evolutionary influence on disease susceptibility.

To deepen our understanding of SAM heterogeneity, we are integrating whole-genome sequencing, transcriptomics, methylation, and metabolomics data from ESAM and NESAM patients on the NHGRI AnVIL cloud platform. Scalable, machine learning–enabled pipelines will allow for population-aware variant calling, methylation quantitative trait loci (meQTL) mapping, functional enrichment, and network-based analyses. This integrative approach aims to elucidate genotype–epigenotype–phenotype relationships

and identify molecular networks contributing to ESAM, with a strong emphasis on reproducibility, equity, and cross-cohort validation.
Together, our findings support a model in which inherited variation in OCM pathways interacts with environmental and epigenetic factors to drive ESAM pathogenesis. By revealing molecular signatures unique to high-risk populations, our study lays the foundation for precision nutrition strategies and demonstrates the potential of ancestry-aware, cloud-based multi-omic research in addressing hidden drivers of pediatric disease.
Keywords: kwashiorkor; marasmus; multi-omics; genetic variants, admixture; local ancestry

Title: Leveraging chromatin accessibility data to understand complex traits
Author list: Liyang Yu[1], Siming Zhao[1]
Detailed Affiliations:
[1]Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College
Dartmouth Cancer Center
Abstract: Thousands of significant signals have been identified from Genome-Wide Association Studies (GWAS). As many associated variants locate in regulatory regions, cell type or tissue specific regulatory elements offer valuable insights in delineating disease related cell types or tissues. The development of scATAC-seq offers unprecedented resolution of different cell states defined by their chromatin accessibility profiles. Open chromatin regions often indicate regulatory elements; the enrichment of GWAS variants in open chromatin regions of a particular cell state indicates that this cell state is associated with the GWAS trait. Most current efforts identify disease associated cell types using bulk ATAC-seq data or construct pseudo bulk data for cell types identified from scATAC-seq; a disease relevance score can be derived on a cell-type level. However, such analyses neglect heterogeneity within a cell type, and for some scATAC-seq data, the cell states can be continuous rather than discrete, leading to ambiguity in cell typing. In this talk, I will present a new method to assess the disease relevance at cell level leveraging single cell ATAC-seq data. Our method leverages the polygenic signals of disease variants in GWAS data to assess its enrichment over the background at cell level. We overcome the sparsity issue of single cell ATAC-seq data through co-regulatory patterns of open chromatin regions across cells. Through simulations we found our method outperformed the states of art methods, providing more accurate cell level disease relevance scores and more effectively leverage single cell ATAC data to identify causal variants. We demonstrate the usefulness our method on single cell ATAC atlas data for a variety of complex traits.
Keywords: Chromatin accessibility, complex traits, statistical genetics, single cell ATAC-seq.

Title: Integrative genomics and epigenomics reveal functions of non-coding variants
Author list: Hongbo Liu[1,2,3,4] and Katalin Susztak[2,3,4]
Detailed Affiliations:
[1]Department of Biomedical Genetics, University of Rochester Medical Center, Rochester, NY 14642, USA. [2]Department of Medicine, Renal Electrolyte and Hypertension Division, University of Pennsylvania, Philadelphia, PA 19104, USA. [3]Institute of Diabetes Obesity and Metabolism, University of Pennsylvania, Philadelphia, PA 19104, USA. [4]Department of Genetics, University of Pennsylvania, Philadelphia, PA 19104, USA.
Abstract: Background: Genome-wide association studies (GWAS) have identified numerous DNA sequence variants associated with complex human diseases. However, over 90% of disease-associated

variants reside in noncoding genome regions, and their functions in complex diseases remain largely unknown—a problem often referred to as the 'variant-to-function' problem. **Methods**: To link non-coding variants to functions in human diseases, we integrated various genomic and epigenomic datasets to identify the regulatory variants by developing several computational strategies, including methylation quantitative trait loci (meQTL), allele-specific expression (ASE), and allele-specific expression accessibility (ASA). In particular, we developed a statistical model, Open4Gene, to link non-coding variants to their target genes using single cell multiome data, which simultaneously profiles DNA accessibility and gene expression within the same cell. **Results**: We conducted a multi-ancestry GWAS mapping in 2.2 million individuals and identified over 1,000 independent loci associated with kidney function. Ancestry-specific analysis indicated an attenuation of newly identified signals on common variants in European ancestry populations and the power of population diversity for further discoveries. We defined genotype effects on allele-specific gene expression and regulatory circuitries in human kidneys and cells. We developed a statistical approach named Open4Gene, which identified 1,351 target genes of genetic variants located within open chromatin regions. By integrating these GWAS and multiome datasets (total 32 types), we found over 24,000 regulatory variants targeting more than 1,000 genes, with over 600 genes also targeted by coding variants. In particular, we discovered the convergence of coding and regulatory variants on 161 key disease genes, critical cell types (including proximal tubules), transcriptional regulators (including HNF4A), and potential drug targets for kidney disease, providing an integrative strategy for functional annotation of noncoding variants in complex human diseases.

Keywords: Human Genetics, epigenetics, kidney disease, non-coding variants, single-cell multiome

Title: Mechanistic annotation of GWAS loci for circulating fatty acids by single-cell omics and CRISPR screens

Author list: Huifang Xu[1], Haifeng Zhang[1], Ge Yu[1], Yitang Sun[1], Elijah Sterling[1,2], Saurav Choudhary[3], Pengpeng Bi[1], Kaixiong Ye[1,3]

Detailed Affiliations:

[1]Department of Genetics, University of Georgia, Athens, Georgia, USA; [2]Regenerative Bioscience Center, University of Georgia, [3]Institute of Bioinformatics, University of Georgia, Athens, Georgia, USA

Abstract: 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, the underlying biological mechanism linking these identified loci to FA metabolism remains largely unclear. Here, we integrate GWAS with single-cell multi-omics and single-cell CRISPR screens to systematically uncover the cellular and molecular mechanisms underlying FA-associated genetic loci. We colocalized GWAS signals for 19 FA traits with six types of multi-omics quantitative trait loci (QTL), including gene expression, protein abundance, DNA methylation, splicing, histone modification, and chromatin accessibility, to identify intermediate molecular phenotypes that mediate the associations between the genetic loci and 19 FA traits. We found that 35% of GWAS loci overlapped with QTL signals for at least one molecular phenotype. Notably, a locus (around genes *GSTT1/2/2B*) associated with total fatty acids, the percentage of omega-6 polyunsaturated fatty acids (PUFA) in total FAs, and total monounsaturated fatty acids overlapped with QTL signals across all six molecular phenotypes. We analyzed single-cell RNA-seq data of over 100,000 cells from liver tissue to explore the cellular context. We discovered that hepatocyte cell populations, particularly those located in the periportal region, are enriched for genes associated with FA traits. To explore the regulatory function of FA-associated loci, we conducted a single-cell CRISPR screen

in over 200,000 HepG2 liver cells, targeting 360 candidate regulatory elements (CREs) from fine-mapped FA trait variants. We identified target genes in cis for 298 CREs, providing a direct map of the regulatory relationship. Our integrative analysis reveals the molecular and cellular mechanisms regulating circulating fatty acid levels, providing mechanistic insights into the genetic architecture of fatty acid metabolism.
Keywords: Fatty acids, GWAS, single-cell multi-omics, single-cell CRISPR screen, Mechanistic annotation

Title: Linking Rare Non-Coding Regulatory Variants Associated with Human Longevity to Cellular Senescence via Integrative Functional Genomic Approaches
Author list: Jiping Yang[1]; HyeRim Han[1]; Jhih-Rong Lin[2]; Zhengdong Zhang[2]; Sofiya Milman[2]; Nir Barzilai[2]; Yousin Suh[1]
Detailed Affiliations:
[1] Department of Obstetrics and Gynecology, Columbia University Medical Center, New York, USA; [2] Department of Genetics, Albert Einstein College of Medicine, Bronx, New York, USA
Abstract: Centenarians, despite representing a tiny proportion of the global population, hold the key to access longevity. By decoding the genomes in a unique Ashkenazi Jewish (AJ) centenarian cohort, we have identified rare coding variants protective against age-related diseases, along with numerous non-coding variants with unknown functions. Non-coding variants, once considered "Junk DNA", are now known to enrich in cis-regulatory elements (CREs) that control transcriptional activity. However, the functional interpretation of non-coding variants remains challenging due to incomplete knowledge of regulatory elements, their mechanisms of action, and the cellular states and processes in which they function, let alone the identification of truly causal variants and their target genes. To partially address this challenge, we employed phenotypic CRISPR screens to discover longevity-associated variant-residing CREs capable of modulating cellular senescence. We prioritized rare regulatory variants identified in our AJ centenarian cohort by mapping non-coding variants in linkage disequilibrium (LD) to potential CREs annotated by Cis-element Atlas (CATlas). Pooled activation (CRISPRa) or inhibition (CRISPRi) using CRE-targeting sgRNAs alleviated cellular senescence in human mesenchymal stromal cells compared to non-targeting sgRNAs. Sequencing-based sgRNA enrichment analysis in endpoint cells identified putative senescence-modulating CREs. Surprisingly, almost all these CREs were located in intergenic or intronic non-promoter regions. To further elucidate the role of these senescence-modulating CREs in transcriptional regulation, we conducted transcriptome-based single-cell CRISPR interference screens to identify their cis-regulated causal genes and trans-effect networks, leading to the discovery of novel genes driving cellular senescence and potential targets for extending human healthspan and lifespan.
Keywords: Functional genomics, rare non-coding variant, longevity, CRISPR screen

Title: Identification of replicative aging and inflammatory aging signatures via whole-genome CRISPRi screens and GWAS meta-analysis
Author list: Lingling Wu[1,2,3†], Xiang Zhu[4,5,6,9†], Yanxia Liu[1,10], Dehua Zhao[1,11], Betty Chentzu Yu[2,3], Zheng Wei[2,3], Xueqiu Lin[1,2,3*], Lei S. Qi[1,7,8*]
Detailed Affiliations:
[1]Department of Bioengineering, Stanford University, Stanford, CA 94305, USA. [2]Computational Biology Program, Fred Hutchinson Cancer Center, Seattle, WA 19024, USA. [3]Translational Data Science IRC, Fred Hutchinson Cancer Center, Seattle WA 19024, USA. [4]Department of Statistics, The Pennsylvania State

University, University Park, PA 16802, USA. [5]Institutes of the Life Sciences, The Pennsylvania State University, University Park, PA 16802, USA. [6]Department of Statistics, Stanford University, Stanford, 94305, CA, USA. [7]Sarafan ChEM-H, Stanford University, Stanford, CA 94305, USA [8]Chan Zuckerberg Biohub, San Francisco, CA 94158, USA. [9]Currrent address: Calico Life Sciences, South San Francisco, CA 94080, USA. [10]Currrent address: Epicrispr Biotechnologies, South San Francisco, CA, USA. [11]Currrent address: Genomic Sciences, GSK, San Francisco, CA, USA
[†]These authors contributed equally to this work.
Abstract: Aging is a major risk factor for chronic diseases and cancer. Cellular aging, particularly in adult stem cells, offers a high-throughput framework for dissecting the molecular mechanisms of aging. We performed multiple genome-wide CRISPR interference (CRISPRi) screenings in human primary mesenchymal stem cells (MSC) derived from adipose tissue during either replicative senescence or inflammation-induced senescence. These screens revealed distinct sets of potential novel regulators specific to each senescence pathway. Combining our perturbation-based functional genomic data with 405 genome-wide association study (GWAS) datasets, including 50 aging-related studies, we found that the inflammatory aging signatures identified from CRISPRi screenings were significantly associated with diverse aging processes, suggesting novel molecular signatures for analyzing and predicting aging status and aging-related disease. The signatures verified through comprehensive functional genomics and genetic analyses may provide new targets for modulating the aging process and enhancing the quality of cell therapy products.
Keywords: Inflammatory aging, CRISPR-screening, GWAS

Workshop – Data Science Solutions for Spatial Transcriptomics

August 5th
9:20 AM – 12:20 PM
Room: 320

Chairs: Travis Johnson

Title: SpatialGE: An Interactive Web Platform for Accessible and Reproducible Spatial Transcriptomics Analysis

Author list: Xiaoqing Yu

Detailed Affiliations:
Moffit Cancer Center

Abstract: Spatial transcriptomics (ST) enables the study of gene expression in spatial context, offering insights into the tumor microenvironment. However, analyzing ST data remains a challenge for non-computational researchers. We developed *spatialGE*, a point-and-click web application built on the spatialGE R package to make ST analysis accessible, reproducible, and interactive. The platform supports

domain detection (SpaGCN, MILWRM), cell deconvolution (STdeconvolve), and phenotyping (InSituType), with workflows for both Visium and single-cell ST data such as CosMx and Xenium. Users can perform multi-step analyses, compare results across samples, and overlay results on tissue images. Reproducibility is ensured through parameter tracking and downloadable outputs. We demonstrated the applications of *spatialGE* to melanoma brain metastases and Merkel cell carcinoma, highlighting how the platform enables hypothesis generation and biological insight without requiring programming experience.

Keywords: Spatial Transcriptomics, Deconvolution, Web applications

Title: Spatial Resolved Gene Regulatory Networks Analysis

Author list: Zhana Duren

Detailed Affiliations:
Indiana University School of Medicine

Abstract: Integrating spatial transcriptomics – which maps gene expression location within tissues – with single-cell multi-omics data, profiling gene expression and chromatin accessibility (or other epigenomic data), offers powerful insights into gene regulation. However, commercially available kits for simultaneous spatial multi-omics profiling are currently unavailable, hindering widespread data generation. Here, we present ISON (Integrated Spatial Omics Network), a novel computational method to infer spatial-resolved gene regulatory networks by leveraging existing single-cell multiome data and spatial transcriptomics data. ISON accurately predicts omics profiles for spatial spots and reconstructs spatially resolved gene regulatory networks, demonstrating scalability in both time and memory. Importantly, ISON omics prediction preserves cis- and trans- regulatory information and enables estimation of transcription factor (TF) activity at the spot level, distinguishing between TFs even within the same family – a capability absents in approaches relying solely on ATAC-seq data. Application of ISON to Alzheimer's disease data reveals disease- and age-specific spatial gene regulatory modules, highlighting its potential for uncovering spatially organized mechanisms driving complex biological processes.

Keywords: Single cell, Multiomics, Gene regulatory networks, Spatial omics

Title: Identifying Key Regulators of Amyloid Beta Clearance from Single Cell Spatial Transcriptomics using Generalized Linear Mixed Effect Models

Author list: Debolina Chatterjee

Detailed Affiliations:
Indiana University School of Medicine

Abstract: Single-cell spatial transcriptomics (ST) enables the simultaneous profiling of gene expression and spatial organization within tissues, offering unprecedented insights into cellular microenvironments. To harness the full potential of such data, we present TAWGLE (Topology AWare Generalized Linear mixed Effect model), a novel statistical framework that integrates spatial topology, cell type identity, intercellular interactions, and disease context to dissect gene expression patterns. Focusing on Alzheimer's disease, where genome-wide association studies have highlighted *INPP5D* as a critical regulator of microglial activity and amyloid beta (Aβ) accumulation, we applied our approach to three mouse models: wild-type (B6), amyloid-bearing (5xFAD), and amyloid-haplodeficient (5xFAD:KO). Our analysis reveals

that genes associated with Aβ pathology show spatially-resolved interactions between microglia and astrocytes. Notably, *PSEN1, GNAQ,* and *COX8A* were upregulated in astrocytes near microglia in 5xFAD, while *CACNA1D, COX6C,* and *COX4I1* were downregulated in 5xFAD:KO. In microglia near astrocytes, *APH1B.C, PLCB1,* and *COX4I1* were upregulated in 5xFAD, while *SLC39A11, SLC11A2, PSEN1, APP,* and *TUBB5* were upregulated in 5xFAD:KO. Thus, we explore cellular neighborhoods within brain tissue, and identify genes associated with enhanced Aβ clearance.

Keywords: Spatial transcriptomics, Single cell, multiomics, Alzheimer's disease, Linear mixed effect models

Title: Leveraging Spatial Transcriptomics of Brain Tissue in Neurological Diseases

Author list: Oscar Harari

Detailed Affiliations:
The Ohio State University

Abstract: Single-cell omics approaches have revolutionized the profiling of brain cell molecular composition with remarkable detail. However, these methods often lose cellular context during tissue processing. Spatial transcriptomics offers the possibility of in situ profiling, providing crucial information about the cell neighborhood and, when combined with immunohistochemistry, relating cellular states to neuropathological lesions. Our research focuses on leveraging single-cell and spatial omics to profile both affected and healthy brain tissue, aiming to reveal cell-type specific changes associated with the etiology and progression of neurological diseases. In this presentation, we will explore how spatial transcriptomics can be employed to uncover novel insights into the pathogenesis of various neurological disorders.
Alzheimer's disease presents as a heterogeneous disorder marked by diverse molecular mechanisms. Given the intricate nature of AD pathology, the relationships between cellular components and their spatial context is critical to understanding disease mechanisms. We employed the Visium HD platform to investigate dorsolateral prefrontal cortex tissues from late-stage AD patients, seeking to reveal spatial gene expression patterns in affected and unaffected regions at single-cell resolution. Immunofluorescence staining detected amyloid-beta and phosphorylated tau accumulations. This approach enabled us to examine the correlation between AD hallmarks and cellular gene expression profiles. Our findings indicate significant differences among cells proximal to amyloid plaques compared to the surrounding unaffected tissues, even among cell types with low representation like microglial and vascular populations.
Following ischemic stroke, pan-necrosis occurs at the injury core. Selective neuronal death is frequently observed in surrounding regions, but the molecular determinants underlying differential neuronal vulnerability remain unclear. To address this, we investigated whether homeostatic molecular pathways could predict the susceptibility or resilience of select neuronal populations to ischemia. Single-nuclei and spatial RNA sequencing were performed on the peri-infarct region of mice 24 hours post-tMCAO, the corresponding contralateral region, and sham mice to identify selectively vulnerable or resilient neurons. We identified genes expressed under homeostatic conditions in sham that predicted selective neuronal resilience or vulnerability. Utilizing the Vizgen MERSCOPE assay, we generated spatial maps, enabling observation of unique glial cell distribution within the infarct as well as spatial distribution of resilient and vulnerable cells within the peri-infarct and contralateral hemisphere.

In conclusion, the integration of single cell, spatially resolved transcriptomics, and immunochemistry significantly enhances our understanding of neurological diseases by elucidating specific spatial cellular niches in which cells mediate disease risk and progression.

Keywords: Spatial transcriptomics, Single cell, Neurodegenerative disease, Alzheimer's disease,

Title: A Statistical Framework to Improve the Design of Spatial Transcriptomics Experiments

Author list: Dongjun Chung

Detailed Affiliations:
The Ohio State University

Abstract: High-throughput spatial transcriptomics has recently gained significant attention and it can capture high-dimensional gene expression profiles in tissue samples at or near single-cell level while retaining the spatial location of each sequencing unit. This new technology provides unprecedented opportunities for biomedical research and has recently gained significant attention from various fields such as cancer research, neuroscience, and developmental biology. To effectively analyze this new type of data, various statistical and computational methods for spatial transcriptomics data analysis have been developed in recent years. However, while some efforts have been made to improve the design of these studies, it is still significantly understudied how to optimize key experimental factors of these experiments. In this talk, I will discuss spaDesign, our novel statistical framework for the design of spatial transcriptomics experiments, which aims to address this critical need. spaDesign is a statistically rigorously designed framework that employs Poisson Gaussian process and Fisher-Gaussian kernel mixture. It can easily simulate a range of spatial transcriptomics data with various sequencing depths, effect sizes, and spatial patterns, which allows rigorous estimation of needed total sequencing depths to detect spatial domains based on spatial transcriptomics experiments. We will demonstrate the utility and power of spaDesign using 10X Visium data from the human brain and the chicken heart.

Keywords: Spatial transcriptomics, Statistical frameworks, Poisson gaussian processes, Fisher-gaussian kernal mixtures

Title: Integrative Modeling of Gene Expression and Histology via Cross-Modal Alignment and Multi-Scale Graph Inference

Author list: Chao Chen

Detailed Affiliations:
Stony Brook University

Abstract: Spatial transcriptomics (ST) offers a powerful way to map gene activity within tissues, providing crucial insights into cellular diversity and spatial organization. When combined with histology images, ST data opens new avenues for enhancing whole slide image (WSI) prediction tasks, such as disease diagnosis and outcome forecasting. However, existing approaches often struggle to align gene expression with tissue images due to spatial distortions and modality differences, and they typically overlook the complex relationships between distant tissue regions. In this talk, we present methods that address these challenges. First, we introduce a novel ranking-based alignment method that captures nuanced cross-modal relationships between gene and image features while maintaining robustness across scales. This is further

stabilized using a teacher-student self-supervised learning strategy to handle the noise and sparsity in gene expression data. Second, we propose **MERGE** (Multi-faceted hiErarchical gRaph for Gene Expressions), a graph-based approach that models interactions across tissue patches by clustering them based on both spatial proximity and morphology. Through a hierarchical graph neural network (GNN), MERGE enables both local and long-range tissue interactions to inform gene prediction. We also examine the impact of various data smoothing techniques in ST, advocating for biologically grounded, gene-aware smoothing methods to reduce technical artifacts. Across multiple public datasets, our methods significantly outperforms current methods in tasks including gene prediction, slide-level classification, and survival analysis, demonstrating the promise of advanced feature alignment and multi-scale graph modeling for spatially informed biomedical insights.

Keywords: Spatial transcriptomics, Histopathology, Whole slide images, Graph neural networks, Multimodal models

Title: Utilizing Deep Transfer Learning to Identify High Risk Subpopulations of Cells in Single Cell and Spatial Omics Data

Author list: Travis S. Johnson

Detailed Affiliations:
Indiana University School of Medicine

Abstract: Leveraging single-cell gene expression profiles can significantly improve our understanding of diseases by associating single cells with traits such as disease subtypes, prognosis, and drug response. Although previous efforts have linked single cell clusters and groups with these attributes, they have primarily focused on changes in cell proportions while overlooking transcriptional changes at the single cell level. To further unravel cell heterogeneity with clusters and reveal nuanced behaviors of cellular subtypes, it is essential to assess the disease associations of individual cells. Previous methods often fail to capture complex patterns that are only discernible through summarizing non-linear relationships across multiple genes. The Diagnostic Evidence GAuge of Single-cells/Spatial-transcriptomics (DEGAS) framework advances these efforts by aligning single cells and/or spatial transcriptomics regions with patients through a unified latent space using nonlinear transformations learnt from deep neural networks (DNNs). Here, we present DEGAS version 2 (DEGASv2), which has been updated with optimal transport based transfer learning and improved time-to-event functionality, more advanced model architecture, and improved model baseline evaluations. DEGASv2 achieves superior performance in analyzing single cell and spatial transcriptomics datasets, including Alzheimer's disease (AD), multiple myeloma (MM) and prostate cancer (PDAC). On the MM discovery dataset, DEGASv2 enabled us to discover cell types that exhibited different drug response patterns over various time frames and were validated with multi-omic data from a time series of single cells that we generated, demonstrating a dangerous subtype of cell and a new therapeutic target.

Keywords: Spatial transcriptomics, Single cell, Transfer learning, Multiple myeloma, Alzheimer's disease, Prostate cancer

Workshop – Computational Omics for Precision Medicine and Drug Discovery

August 5th
9:20 AM – 12:20 PM
Room: 301

Chairs: Bin Chen, Qian Li

Title: Protein Language Model ESM3 Enables Superior Prediction of Complex Variant Effects Across ClinVar and DMS Benchmarks
Author list: Chang Li[1] and Xiaoming Liu[1]
Detailed Affiliations:
[1]Department of Global, Environmental, and Genomic Sciences, College of Public Health, University of South Florida, Tampa, Florida, USA.
Abstract: Accurate prediction of variant pathogenicity, particularly for complex mutations beyond single nucleotide variants (SNVs), remains a major challenge in genomic medicine. Traditional protein language models (PLMs) like ESM2 and AlphaFold focus on sequence or structure alone, limiting their ability to fully assess functional disruptions. Here, we demonstrate that ESM3, a multimodal protein model integrating sequence, structure, and function via geometric attention mechanisms, substantially advances variant effect prediction. Without relying on multiple sequence alignments, ESM3 shows strong performance across a wide range of variant types, including non-frameshift insertions/deletions (InDels) and complex variants such as stop-gain/loss mutations.
Benchmarking on Deep Mutational Scanning (DMS) and novel ClinVar datasets, ESM3 achieves superior prediction accuracy compared to current state-of-the-art tools, particularly excelling at complex and non-canonical variants where other models falter. Through case studies on GABRB3 and IRF6, we demonstrate that ESM3's cross-modality divergence and entropy metrics provide unique mechanistic insights, distinguishing gain-of-function (GOF) and loss-of-function (LOF) variants and highlighting domain-specific functional vulnerabilities. Our findings establish ESM3's zero-shot potential for variant effect prediction, particularly for poorly characterized or structurally disruptive mutations, offering new avenues for variant interpretation and precision medicine.
Keywords: protein language model, pathogenicity prediction, SNV, InDel

Title: Massive labeled transcriptomics as a resource of transcriptome representation learning and drug discovery
Author list: Bin Chen[1,2,3]
Detailed Affiliations:
[1]Department of Pharmacology and Toxicology, Michigan State University, East Lansing, MI 48824, USA.
[2]Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824,

USA. [3]Department of Pediatrics and Human Development, Michigan State University, Grand Rapids, MI 49503, USA

Abstract: Gene Expression Omnibus (GEO), the largest repository of transcriptomics data, houses more than millions of gene expression profiles from 200,000 studies and has been extensively explored for research; however, its metadata is presented in unstructured text and often lacks consistency due to varying submission formats. Manual curation is time-intensive and error-prone, posing challenges for dataset integration and downstream analyses. We introduce a GPT-based AI model to automate and standardize GEO metadata annotation, significantly improving efficiency, accuracy, and consistency. We establish a structured annotation framework, integrating domain-specific mega prompts and standardization protocols to ensure uniformity across including strain, genotype, disease, and treatment details. We present a comprehensive annotation dataset that encompasses >100K mouse and human samples each, along with their transcriptome profiles. We further develop benchmarks for the prediction of labels from gene expression profiles using state-of-the-art transcriptome embedding methods. By combining the large-scale transcriptome data and our drug discovery platforms OCTAD and GPS, we can predict the therapeutic potential for any drugs or compounds against hundreds of diseases. We expect this dataset will become an essential resource of learning transcriptome and large-scale drug discovery.

Keywords: Large Language Models (LLMs); GPT-based Annotation; Drug Repositioning

Title: Generative AI for Human Genetics and Functional Genomics

Author list: Xinghua Shi[1]

Detailed Affiliations:

[1]Department of Computer & Information Sciences, Temple University, Philadelphia, PA 19122, USA

Abstract body: Generative artificial intelligence (GenAI) techniques have been addressing key challenges and introducing transformative opportunities in human genetics and functional genomics. In general, GenAI techniques that have been widely adopted in the field include variational autoencoders (VAEs), generative adversarial networks (GANs), large language models (LLMs), Transformers and diffusion models. Using our research products as examples, I will present the applications of these models to classical and emerging problems in genotype imputation, synthetic genotype generation, augmented chromatin characterization, and disease prediction.

Keywords: Generative Artificial Intelligence (GenAI), Functional Genomics, Large Language Models (LLMs), Computational Biology

Title: Distinct Mutational Profiles in Primary Sclerosing Cholangitis-Associated Cholangiocarcinoma Compared to *de novo* Cholangiocarcinoma

Author list: Shulan Tian[1], Filippo Pinto e Vairo[3], Ahmad H. Ali[2], Huihuang Yan[1], Bryan M. McCauley[4], Brian D. Juran[2], Tony C. Luehrs[4], Fan Leng[5], Cameron M. Callaghan[6], Jacob A. Frank[4], Sicotte Hugues[1], Sebastian M. Armasu[4], Robert A. Vierkant[4], Jan B. Egan[3], Zhifu Sun[1], Nicholas B. Larson[4], Eric W. Klee[1,3], Konstantinos N. Lazaridis[2,3]

Detailed Affiliations:

[1]Division of Computational Biology; [2]Division of Gastroenterology and Hepatology; [3]Center for Individualized Medicine and Department of Clinical Genomics; [4]Division of Clinical Trials and Biostatistics; [5]Data Analytics and Integration; [6]Department of Radiation Oncology, Mayo Clinic

Abstract: Introduction: Cholangiocarcinoma (CCA) is a rare and aggressive malignancy characterized by etiologic heterogeneity and poor survival. Primary sclerosing cholangitis (PSC) is the most recognized risk factor for CCA in Western countries. PSC-associated CCA (PSC-CCA) is a leading cause of morbidity and mortality in PSC patients and exhibits distinct clinical features compared to those in *de novo* CCA. However, the molecular mechanisms driving these two subtypes of CCA remain largely unexplored. This study aimed to characterize the spectrum and prevalence of germline genetic variants in pathologically confirmed PSC-CCA, and *de novo* CCA as well as PSC patients without CCA (PSC-w/o CCA). **Material and method:** This retrospective cross-sectional study included 301 patients with PSC-w/o CCA and 170 patients with CCA (PSC-CCA, n=88; *de novo* CCA, n=82) identified from two population genomics studies conducted at Mayo Clinic between 2016 and 2023. Their diagnoses, phenotypes, outcomes, as well as medical and family histories were obtained from electronic health records (EHRs) and self-reported questionnaires. Exome sequencing of these patients was conducted with genomic DNA, and genetic variants were identified using bioinformatics workflow following the Genome Analysis Toolkit (GATK) best practices. A comprehensive list of c*ancer susceptibility genes* was compiled from prior cancer studies. Functional annotation and pathogenicity assessment of cancer-associated genetic variants were performed according to current American College of Medical Genetics and Genomics and the Association for Molecular Pathology (ACMG/AMP) guidelines. **Result and discussion**: Analysis of exome sequencing data from 471 patients identified 53 pathogenic/likely pathogenic (P/LP) germline variants across 25 cancer susceptibility genes (CSGs) in 10.8% (51/471) of patients. The highest prevalence of P/LP germline variants was observed in PSC-CCA patients (13.6%, 12/88), followed by PSC-w/o CCA (10.0%, 31/301) and *de novo* CCA (9.76%, 8/82). Interestingly, PSC-CCA patients exhibited P/LP germline variants mainly in moderate-, low-penetrance, and/or autosomal recessive genes, with significant enrichment in the Fanconi anemia DNA repair pathway. In contrast, patients with *de novo* CCA predominantly carried P/LP germline variants in the tumor suppressor genes that are key players in homologous recombination repair pathway. Similarly, germline variants led to differentially altered metabolic and signal pathways observed between PSC-CCA and *de novo* CCA patients. **Conclusion:** These findings provide key insights into distinct CCA subtypes and call for an effort to systematically study germline testing of patients with PSC-CCA and *de novo* CCA as an approach to inform personalized approaches to screening, clinical management and targeted therapy of CCA in these patients.
Keywords: primary sclerosing cholangitis; cholangiocarcinoma; cancer susceptibility genes; exome sequencing; DNA repair

Title: High-resolution multi-omic dissociation of brain tumors with multimodal autoencoder
Author list: Jiao Sun[1], Ayesha Malik[2], Tong Lin[1], Ayla Bratton[2], Kyle Smith[3], Yue Pan[1], Arzu Onar-Thomas[1], Giles W. Robinson[4], Wei Zhang[2], Paul A. Northcott[3], Qian Li[1*]
Detailed Affiliations:
[1]Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN, 38105. [2]Department of Computer Science, University of Central Florida, Orlando, FL, 32816. [3]Department of Developmental Neurobiology, St. Jude Children's Research Hospital, Memphis, TN, 38105. [4]Department of Oncology, St. Jude Children's Research Hospital, Memphis, TN, 38105
Abstract: Single-cell technologies enable high-resolution profiling of molecular dynamics in developmental and cancer biology. But heterogeneity and complexity of tumors may hinder the lineage cell mapping in developmental origins or dissection of tumor microenvironment, requiring digital dissociation of bulk tissues. Many deconvolution methods focus on transcriptomic assay using scRNA-seq as reference,

not easily applicable to other omics due to ambiguous cell markers and unexpected biological difference between reference and target tissues. Here, we present MODE, a multimodal autoencoder pipeline linking multi-dimensional molecular features to jointly predict personalized multi-omic profiles and estimate modality-specific cellular compositions, using pseudo-bulk data constructed by internal non-transcriptomic signature matrix recovered from target tissues and external scRNA-seq reference. The accuracy of MODE was evaluated through extensive simulation experiments generating realistic multi-omic data from distinct tissue types. MODE outperformed seven deconvolution pipelines with superior generalizability and enhanced fidelity across five independent datasets, elucidating multi-omic signatures for developmental origins, evolution, subtyping of pediatric medulloblastoma and the prognosis of adult glioblastoma.
Keywords: multimodal, autoencoder, high-resolution purification, origin cell mapping, tumor microenvironment

Title: CoMPaSS: A Computational Pipeline for Cross-Platform Concordance Assessment and Navigating Study Design in Microbiome Research
Author list: Xi Qiao[1,2], Ruitao Liu[3], Daoyu Duan[3], Qian Li[4], Liangliang[3]
Detailed Affiliations:
[1]Department of Internal Medicine, Epidemiology, School of Medicine University of Utah, Salt Lake City, UT, USA; [2]Huntsman Cancer Institute, Salt Lake City, UT, USA; [3]Department of Population and Quantitative Health Sciences, School of Medicine Case Western Reserve University, Cleveland, OH, USA; [4]Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN, USA.
Abstract body: Microbiome analysis is essential for understanding microbial interactions and their impact on human health. Advances in next-generation sequencing (NGS) have led to two widely used methods: 16S rRNA gene sequencing and metagenomic shotgun sequencing. While 16S sequencing enables broad taxonomic classification and phylogenetic analysis, shotgun sequencing provides higher taxonomic resolution and functional insights but at a higher cost. However, their comparative efficacy remains uncertain, complicating study design. To address this, we introduce CoMPaSS (Concordance of Microbiome Sequencing Platforms and Study Initiation Strategy), a computational pipeline for navigating microbiome study design. CoMPaSS systematically evaluates sequencing concordance across multiple levels, from community diversity to taxonomic composition, and provides power analysis, sample size estimation, and cost assessment to support study planning. Through extensive simulations and real-world microbiome studies, we found moderate concordance at broader taxonomic levels but significant discrepancies at finer levels and for rare taxa, emphasizing the impact of sequencing method selection on study outcomes. By integrating statistical and computational insights, CoMPaSS helps researchers optimize study design based on scientific and budgetary constraint.
Keywords: 16S, metagenomic shotgun, concordance, power calculation

Title: SEHI-PPI: An End-to-End Sampling-Enhanced Human-Influenza Protein-Protein Interaction Prediction Framework with Double-View Learning
Author list: Qiang Yang[1], Xiao Fan[2], Haiqing Zhao[3], Zhe Ma[4], Megan Stanifer[4], Jiang Bian[5], Marco Salemi[6], and Rui Yin[1, *]
Detailed Affiliations:
[1] Department of Health Outcomes & Biomedical Informatics, University of Florida, Gainesville, FL, USA. [2] Department of Biomedical Engineering, University of Florida, Gainesville, FL, USA. [3]Department of Biochemistry & Molecular Biology, University of Texas Medical Branch, Galveston, TX, USA. [4]

Department of Molecular Genetics and Microbiology, University of Florida, Gainesville, FL, USA. [5] School of Medicine, Indiana University, Indianapolis, IN, USA. [6] Department of Pathology, Immunology and Laboratory Medicine, University of Florida, Gainesville, FL, USA
Abstract: Influenza continues to pose significant global health threats, hijacking host cellular machinery through protein-protein interactions (PPIs), which are fundamental to viral entry, replication, immune evasion, and transmission. Yet, our understanding of these host-virus PPIs remains incomplete due to the vast diversity of viral proteins, their rapid mutation rates, and the limited availability of experimentally validated interaction data. Additionally, existing computational methods typically struggle with limited high-quality samples and inability for the modeling on the intricate nature of host-virus interactions. To address these challenges, we present SEHI-PPI, an end-to-end framework for human-influenza PPI prediction. SEHI-PPI integrates a double-view deep learning architecture that captures both global and local sequence features, coupled with a novel adaptive negative sampling strategy to generate reliable and high-quality negative samples. Our method outperforms multiple benchmarks, including state-of-the-art large language models, achieving a superior performance in sensitivity (0.986) and AUROC (0.987). Notably, in a stringent test involving entirely unseen human and influenza protein families, SEHI-PPI maintains strong performance with an AUROC of 0.837. The model also demonstrates high generalizability across other human-virus PPI datasets, with an average sensitivity of 0.929 and AUROC of 0.928. Furthermore, AlphaFold3-guided case studies reveal that viral proteins predicted to target the same human protein cluster together structurally and functionally, underscoring the biological relevance of our predictions. These discoveries demonstrate the reliability of our SEHI-PPI framework in uncovering biologically meaningful host-virus interactions and potential therapeutic targets.
Keywords: Protein-Protein Interaction, Machine Learning, Host-Virus, Double-view Learning

Title: Cyclin D1 induces epigenetic and transcriptional alterations in Multiple Myeloma with t(11;14)(q13;q32)
Author list: Huihuang Yan, PhD[1], Suganti Shivaram, M.B.B.S[2], Hongwei Tang, PhD[2], Hans Anderson[2], Shulan Tian, PhD[1], Michael D Howe[3], Abiola Bolarinwa, MBBS,FMCPath[4], Cinthya Zepeda Mendoza, PhD[5], Stacey Lehman[2], Leif Bergsagel, MD[6], Esteban Braggio, PhD[7], Rafael Fonseca, MD[8], Shaji Kumar, MD[4], Francesco Maura, MD[9], Linda B. Baughn, PhD[2]
Detailed Affiliations
[1]Division of Computational Biology, Mayo Clinic, Rochester, MN 55905, USA; [2]Division of Laboratory Genetics, Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN 55905, USA; [3]Division of Hematology, Department of Internal Medicine, Mayo Clinic, Rochester, MN 55905, USA; [4]Division of Hematology, Mayo Clinic, Rochester, MN 55905, USA; [5]Division of Laboratory Genetics, Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN 55905, USA; [6]Division of Hematology/Oncology, Mayo Clinic, Phoenix, AZ 85054, USA; [7]Division of Hematology and Oncology, Mayo Clinic, Scottsdale, AZ 85259, USA; [8]Division of Hematology, Mayo Clinic, Phoenix, AZ 85054, USA; [9]Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA.
Abstract: Cyclin D1, encoded by the *CCND1* gene, is primarily known as a major regulator of cell cycle progression. Emerging studies have demonstrated a role for cyclin D1 in regulating gene transcription. In mantle cell lymphoma with t(11;14)(q13;q32) associated with *CCND1::IGH*, cyclin D1 binds strongly to the promoters of highly transcribed genes; its overexpression results in global transcriptional downregulation and activation of a specific gene set. However, the regulatory role for cyclin D1 in t(11;14) multiple myeloma (MM) remains undefined. We hypothesize that cyclin D1 modulates the expression of a

distinct gene set through its impact on the landscape of epigenetic modifications. To address this hypothesis, CRISPR/Cas9 was used to generate knockout (KO) line of *CCND1* in U266B1, an MM cell line with a IGH Ea1 super-enhancer inserted centromeric to *CCND1* resulting in *CCND1::IGH*. ChIP-seq (H3K4me3, H3K4me1, and H3K27ac), ATAC-seq, and RNA-seq were performed for KO and wild-type (WT) clones. Differential expression analysis identified 2-fold more genes that were down-regulated than up-regulated in KO compared to WT (467 vs. 214). Pathway analysis revealed an enrichment of TNF-alpha signaling via NF-kappaB, inflammatory response, and apoptosis in association with down-regulated genes, while up-regulated genes are enriched in myogenesis, KRAS signaling, and epithelial mesenchymal transition pathway. In parallel, we identified 2.5% (1,161) H3K27ac peaks and 0.5-1% (359-610) of the peaks from the other three marks that showed significant changes between KO and WT, predominantly in distal regulatory regions. For ATAC- seq and H3K4me1, about 70% of the differential peaks showed increased signal in KO, in contrast with H3K27ac for which 60% of differential peaks showed decreased signal in KO. The preferential loss of H3K27ac occupancy is consistent with the observation of a higher proportion of genes being down-regulated in KO. These results indicate that knockout of *CCND1* results in a global increase of chromatin accessibility, but a reduction of H3K27ac and gene expression. Further analysis of differential peaks in the promoter regions revealed that, for ATAC-seq and H3K4me3, 47% and 67% of the differential peaks were associated with differentially expressed genes, vs. ~30% for the two enhancer marks. We conclude that knockout of *CCND1* impacts local chromatin state, particularly at enhancer regions, and the transcriptional program. Further studies will identify the genes targeted by the differential enhancers and their possible roles in MM pathogenesis mediated by cyclin D1 overexpression following the t(11;14) event.

Keywords: ChIP-seq; Cyclin D1; Multiple myeloma; RNA-seq; Translocation

Integrative Bioinformatics for Translational and Precision Medicine

August 5th
9:20 AM – 12:20 PM
Room: 350

Chairs: Yuan Liu, Shilin Zhao

Title: A novel immune-related risk stratification model to predict prognosis, immunotherapy and chemotherapy response for Neuroblastoma

Author list: Jiamei Xu, Peng Zhao, Jing Qiao, Yan Chang, Xiaoling Mu, Jingxian Wu, Jin Zhu and Xiaohui Zhan

Abstract: Neuroblastoma (NBL) characterized by high morbidity and mortality is a prevalent pediatric cancer originating from neural crest cells. Unsatisfactory prognostic and treatment effects persist due to NBL patients' clinical diversity and individual variations. Despite immunotherapy as a promising therapy has been used in NBL, it still fails in many cases. Thus, there is a strong need to develop an innovative

model to optimize therapeutic outcomes and improve patient survival. In this study, the local maximal quasi-clique merger (lmQCM) algorithm was employed to identify gene co-expression network (GCN) modules, with module 1, linked to immune function, selected for further analysis. A novel immune-related risk stratification model (NIRSM) was developed based on module 1's key genes, demonstrating associations with poor prognosis, immunotherapy and chemotherapy responses, and superior predictive performance compared to age, INSS stage, and MYCN status. The low-risk patients showed enhanced immunotherapy response and higher immune cell infiltration, while module 1 genes exhibited elevated expression in this group. The clinical features like age, INSS stage, and MYCN status differed significantly between two risk groups ($p < 0.001$). Single-cell analysis confirmed the cell-type-specific expression patterns of NIRSM-related genes in immune cells, underscoring the model's biological and clinical relevance. In total, we established a robust model with important implications for predicting NBL prognosis, immunotherapy and chemotherapy response. Our findings not only provide crucial clinical implications for personalized treatment strategies but also offer potential therapeutic targets.
Keywords: NBL; GCN; Immune-related risk stratification model; Prognosis; Immunotherapy; Chemotherapy

Title: The Impact of HLA Diversity on Immune Cell Composition, Tumor Mutation Burden, and Cancer Survival
Author list: Judy Bai, Lilly Wei, Kyle Yang, Qing Luo, Yu Liu, Justin Guo, Fenyao Yan, Limin Jiang, Yan Guo and Shilin Zhao
Abstract: HLA molecules play a crucial role in immune responses by influencing antigen presentation and immune cell composition. While previous studies have focused on specific HLA genes or alleles, the impact of overall HLA diversity remains understudied. In this study, we analyzed HLA diversity in relation to immune cell composition, TMB, and mutational signatures across multiple cancers. Higher HLA diversity correlated positively with cytotoxic immune cells, such as CD8+ T cells and activated NK cells ($p = 7.14 \times 10^{-10}$), while correlating negatively with immunosuppressive cells, including monocytes ($p = 6.00 \times 10^{-8}$). HLA diversity generally showed a negative correlation with TMB, except in GBM ($R = 0.15$, $p = 0.02$), where immune suppression may allow highly mutated cells to persist. In LGG ($R = -0.16$, $p = 0.0002$), higher HLA diversity appeared to enhance immune selection against mutated clones. Additionally, HLA diversity was negatively associated with mutational signatures from tobacco (LUSC: $p = 3.07 \times 10^{-5}$, LUAD: $p = 1.18 \times 10^{-5}$), UV exposure (SKCM: $p = 0.04$), and aflatoxin (LIHC: $p = 0.03$), suggesting a role in limiting mutation accumulation. Survival analysis showed that higher HLA diversity improved survival in SKCM (HR: 0.61, $p = 0.0005$) and LUAD (HR: 0.69, $p = 0.02$) but was linked to poorer survival in LGG (HR: 2.09, $p = 0.0001$), likely due to chronic inflammation and immune evasion.
Keywords: HLA; Cancer; Survival; Mutational Signature; Aflatoxin; UV light; Tobacco

Title: Horizontal gene transfer networks reveal resistance of plasmid-mediated communication in antibiotic exposure
Author list: Shuai Cheng Li, Lijia Che, Shuai Wang, Yiqi Jiang, Jingwan Wang, Bowen Tan and Xinyao Li
Abstract: Plasmids play an important role in microbial evolution and adaptation, serving as mediators of horizontal gene transfer (HGT) that facilitates the exchange of genetic material across diverse species. We have deduced the first comprehensive plasmid-mediated HGT network using 214,950 plasmid taxonomic

units (PTUs) sourced from the IMG/PR database. In this network, taxa serve as vertices, with edges symbolizing potential gene exchanges facilitated by plasmids. This network demonstrates a hierarchical structure and high robustness. The network edges exhibit strong specificity to particular environments, while they exhibit similarity and generality across various categories of antibiotic-resistance genes (ARGs). Further, we observed a consistent preservation of plasmid-mediated communication ability in gut microbiome after antibiotic exposure in two independent experiments of antibiotic exposure.
Keywords: horizontal gene transfer; antibiotic resistance; metagenomics; network analysis

Title: Boolean Network Modeling-Guided Identification of FDA-Approved Drug Combinations for Targeted Treatment Strategies in Head and Neck Cancer
Author list: Pranabesh Bhattacharjee and Aniruddha Datta
Abstract: Head and neck cancer (HNC) presents significant therapeutic challenges due to pathway redundancies and resistance mechanisms. To address this, we developed a Boolean network model integrating key signaling pathways—EGFR, Wnt, Hippo-YAP, MAPK/ERK, and PI3K/mTOR—to systematically assess single and combination drug therapies. Using the Normalized Size Difference (NMSD) metric, we quantified the efficacy of FDA-approved drugs against tumors with multiple mutations. Our simulations identified VT3989 (YAP/TEAD inhibitor) as the most effective monotherapy. Among two-drug combinations, Ulixertinib (ERK inhibitor) and VT3989 exhibited the lowest NMSD, indicating strong synergistic inhibition of MAPK and Hippo pathways. Adding Vorinostat (FBXW7 modulator) further enhanced efficacy, achieving 80% efficacy. The most effective combination—Temsirolimus (mTOR inhibitor), Ulixertinib, VT3989, and Vorinostat—demonstrated an 88.3% improvement over untreated conditions. Our findings support a shift from sequential to concurrent multi-pathway targeting, mirroring clinical evidence that combination approaches delay resistance. The hierarchical NMSD reductions from 0.685 (single-agent) to 0.120 (four-drug therapy) highlight the advantage of combination depth in pathway control. This computational framework provides a rationale for prioritizing Temsirolimus-containing quadruple therapies, offering a novel precision oncology strategy for HNC with complex mutational landscapes.
Keywords: Boolean Network; Combination Therapy; Drug Repurposing; Head and Neck Cancer; Targeted Therapy

Title: Comparison of Nanopore Sequencing, MethylationEPIC Array, and EM-Seq for DNA Methylation Detection
Author list: Steven Brooks, Hongyu Gao, Xuhong Yu, Yunlong Liu and Gang Peng
Abstract: DNA Methylation is an important biological process in epigenetics, and many methods have been developed to profile DNA methylation. An increasing number of recent studies have employed Oxford Nanopore long-read sequencing technology for DNA methylation detection, presenting an alternative to the widely utilized Infinium arrays and short-read whole-genome sequencing methods. In this study, we evaluate the performance of Nanopore sequencing in DNA methylation detection by comparing it to the Illumina's MethylationEPIC microarray (EPIC) and Enzymatic Methyl-Sequencing (EM-Seq). The initial comparison was conducted between the Nanopore platform and the EPIC array. Among the ~850,000 CpG sites covered by both methods, we observed high concordance (Pearson correlation coefficient, $r \geq 0.94$ across all four samples). After downsampling Nanopore data from an average coverage of 25.9 reads per site to 10 reads per site, the correlation in CpG methylation remained high ($r \geq 0.93$). Lower correlation of

CpG methylation (r: 0.79 - 0.88) was detected between Nanopore and EM- Seq, which can be attributed to biased and reduced coverage of hypomethylated CpG sites by EM-Seq. We also investigated new features detected by Nanopore sequencing, such as native DNA sequencing that can differentiate 5mC and 5hmC, as well as haplotype-specific methylation. Overall, the Nanopore platform exhibited a high degree of concordance with the EPIC array and provided more uniform genomic coverage than EM-Seq. This study provides insights for researchers in selecting appropriate DNA methylation detection methods, considering factors such as cost, DNA input, and the complexity of downstream analysis.
Keywords: Nanopore; 5mc/5hmc; haplotype-specific methylation

Title: A Hierarchical Adaptive Diffusion Model for Flexible Protein-Protein Docking
Author list: Rujie Yin and Yang Shen
Abstract: Protein–protein interaction prediction is critical but challenged by significant conformational changes. We propose a hierarchical adaptive diffusion model separating global rigid-body motions and local flexibility, with noise schedules mimicking induced fit effects. Local flexibility is adaptively conditioned on predicted conformational change levels. Using the DIPS-AF dataset, the model outperformed AlphaFold2-like and DiffDock-PP models, especially in flexible cases, demonstrating improved docking accuracy.
Keywords: Protein docking; Conformational changes; Generative models; Diffusion models

Title: “Frustratingly easy" domain adaptation for cross-species transcription factor binding prediction
Author list: Mark Maher Ebeid, Ali Tugrul Balcı, Maria Chikina, Panayiotis V Benos and Dennis Kostka
Abstract: Motivation: Sequence-to-function models, designed to interpret genomic DNA and predict functional outputs, have demonstrated success in characterizing regulatory sequence activity. However, interpreting these models remains an open challenge, raising questions about whether they learn generalizable biochemical properties. Cross-species prediction of transcription factor (TF) binding offers a promising avenue to push models toward generalization, leveraging variation across species to potentially uncover a conserved regulatory code. Nonetheless, accounting for systematic differences between the genomes of different species presents a significant challenge.
Results: We introduce MORALE, a framework leveraging an established domain adaptation approach that is “frustratingly easy". MORALE trains on sequences from one or more source species and predicts TF binding on a single target species; in order to learn an invariant cross-species representation, MORALE simultaneously aligns the moments (i.e., 1st and 2nd) between all species. This direct approach integrates readily into models with an embedding layer. Unlike adversarial alternatives, it requires no additional parameters and does not alter the standard gradient computation. We apply MORALE to two ChIP-seq datasets of liver-essential TFs: one comprising human and mouse, and another comprising five mammalian species. Compared to both the baseline and gradient reversal (GRL), MORALE demonstrates improved performance across all TFs in the two-species case, avoiding the performance degradation observed with the GRL approach in this study. Furthermore, gradient inspection revealed that the de novo motifs discovered by MORALE adhered more strictly to CTCF compared to the GRL approach. For the five-species case, our method significantly improved TF binding site prediction for all TFs when predicting on human data, surpassing the performance of a human-only model — a result not observed in the two-species comparison. Overall, MORALE is a direct and competitive approach that leverages domain adaptation techniques to improve cross-species TF binding site prediction.

Keywords: unsupervised domain adaptation; regulatory genomics; transcription factor binding site prediction; moment alignment; invariant representation learning

Title: Multi-omic analysis integrating global transcriptional and post-transcriptional profiles reveals predominant role of post-transcriptional control in three human cell lines
Author list: Alexander Krohannon, Mansi Srivasta, Neel Sangani and Sarath Janga
Abstract: Gene expression is regulated through a complex interplay between transcriptional and post-transcriptional mechanisms, yet their relative contributions and relationships remain incompletely understood. In this study, we propose normalized metrics to quantify regulatory density at each of these two levels by integrating ATAC-seq, RNA-seq, and Protein Occupancy Profiling sequencing (POP-seq) data across three human cell lines - HEK293, HepG2, and K562, to assess gene-specific regulatory variations. This analysis revealed 3 distinct regulatory classes: predominantly post-transcriptionally regulated, predominantly transcriptionally regulated, and neutrally regulated genes. Using this metric, significant associations between regulatory strategies and gene properties were uncovered; with transcriptionally regulated genes exhibiting greater length, post-transcriptionally regulated genes displaying higher isoform diversity and expression levels, and specific transcript types showing consistent enrichment patterns across regulatory categories. Remarkably, 55.8% of genes maintained identical regulatory classification across the three cell lines examined, with functional pathway analysis demonstrating high conservation of regulatory-functional relationships despite different cellular origins. The strong correlation between transcriptional and post-transcriptional regulation suggests coordinated interaction rather than independent operation. This study provides a novel framework for understanding gene regulatory strategies and demonstrates that the relationship between gene properties and regulatory mechanisms represents a fundamental organizational principle that transcends cell-type specificity, with implications for understanding dysregulation in disease states.
Keywords: Systems Biology; Multi-omics; Gene Regulation; ATAC-Seq; POP-Seq

AI and Machine Learning in Translational Genomics

August 5th
1:30 AM – 4:50 PM
Room: 320

Chairs: Huihuang Yan, Yixing Han

Title: Adaptive Chebyshev Graph Neural Network for Cancer Gene Prediction with Multi-Omics Integration
Author list: Sa Li
Abstract: Identifying cancer driver genes is computationally challenging due to diverse genetic and non-genetic factors. We present ACGNN, integrating pan-cancer multi-omics data and PPI networks into graph convolutional networks refined with adaptive Chebyshev filters for flexible feature aggregation. ACGNN achieved a 25.9% AUPRC improvement over state-of-the-art methods, accurately identifying established and novel cancer driver genes, providing valuable insights for cancer research and precision medicine.
Keywords: Cancer driver genes; Graph neural network; Node embeddings; Chebyshev networks

Title: A Generative Imputation Method for Multimodal Alzheimer's Disease Diagnosis
Author list: Reihaneh Hassanzadeh, Anees Abrol, Hamid Reza Hassanzadeh and Vince D. Calhoun
Abstract: Multimodal data analysis can lead to more accurate diagnoses of brain disorders due to the complementary information that each modality adds. However, a major challenge of using multimodal datasets in the neuroimaging field is incomplete data, where some of the modalities are missing for certain subjects. Hence, effective strategies are needed for completing the data. Traditional methods, such as subsampling or zero-filling, may reduce the accuracy of predictions or introduce unintended biases. In contrast, advanced methods such as generative models have emerged as promising solutions without these limitations. In this study, we proposed a generative adversarial network method designed to reconstruct missing modalities from existing ones while preserving the disease patterns. We used T1-weighted structural magnetic resonance imaging and functional network connectivity as two modalities. Our findings showed a 9% improvement in the classification accuracy for Alzheimer’s disease versus cognitive normal groups when using our generative imputation method compared to the traditional approaches.
Keywords: Generative Adversarial Networks; Multi-Modal Classification; Alzheimer's Disease

Title: A user-friendly R Shiny app for Predicting Surface Protein Abundance from scRNA-seq Expression Using Deep Learning in blood cells
Author list: Hui-Mei Tsai, Tzu-Hung Hsaio, Yu-Chiao Chiao, Eric Y. Chuang and Yidong Chen
Abstract: Understanding accurate immune cell heterogeneity and function in single-cell datasets requires access to protein-level information, which is often missing due to experimental limitations. To address this gap, we present shinyDeepGxP, an interactive web application that implements our deep learning model, DeepGxP, for predicting surface protein abundance from single-cell RNA-sequencing (scRNA-seq) data. The platform makes DeepGxP accessible to researchers without programming expertise. Users can upload scRNA-seq count matrices and perform “Predict Protein”, which predicts the abundance of 224 biologically relevant surface proteins. shinyDeepGxP also provides visualization to aid in identifying distinct cell populations based on predicted protein profiles. Moreover, users can access to “Interpret Model”, which reveals key RNA predictors and their associated biological pathways for each protein. In summary, shinyDeepGxP is a user-friendly and freely available web tool that brings protein-level resolution to RNA-only single-cell datasets, supporting multimodal discovery without the need for additional experiments.
Keywords: Shiny web; Protein prediction; Single-cell; Deep learning

Title: HELP-TCR Harmonized Explainable Language Processing Toolkit for T-Cell Antigen Receptor Repertoires
Author list: Michal Seweryn, Yulyana Kalesnik, Dawid Krawczyk, and Maciej Pietrzak
Abstract: Functional characterization of T-cell antigen receptor (TCR) repertoires is critical for advancing our understanding of adaptive immune responses across diverse contexts, including infectious diseases, cancer, autoimmune conditions, and allergic disorders. Detailed analysis of TCR repertoires can reveal disease-specific signatures, support biomarker discovery, and facilitate the development of immunotherapies and vaccines. However, current computational approaches often prioritize either global repertoire metrics or employ deep learning models that, while powerful, offer limited interpretability. We introduce HELP-TCR, a novel machine learning framework based on natural language processing that combines low-dimensional, explainable feature extraction with robust classification performance. HELP-TCR represents TCR repertoires by modeling the position-specific distributions of single amino acids and amino acid pairs, transforming sequences into multidimensional tensor structures. To increase reproducibility, a consensus grouping method merges features with highly similar position-wise distributions. A modified ResNet-18 deep learning architecture, adapted to process these tensors, enables accurate classification, while post-hoc saliency map analysis highlights the most informative features contributing to model predictions. Using a dataset of bootstrapped TCR sequences, HELP-TCR achieved an AUC of 0.96, outperforming existing methods including DeepTCR (AUC 0.76) and TCR-BERT embeddings. Beyond performance, HELP-TCR enables identification of position-specific amino acid motifs associated with classification decisions, offering biologically interpretable insights into TCR repertoire differences. By emphasizing model interpretability alongside predictive accuracy, HELP-TCR provides a versatile platform for functional TCR repertoire analysis with potential applications in immunotherapy development, vaccine design, and immune monitoring.
Keywords: T-cell antigen receptors; explainable AI; deep learning; neural network-based classification; Wasserstein distance

Title: Efficient and Valid Large Molecule Generation via Self-supervised Generative Models
Author list: Doyoung Kwak, Raiyan Chowdhury, Byung-Jun Yoon and Xiaoning Qian
Abstract: The realm of molecular design, particularly for large molecules, presents unique challenges and opportunities in drug discovery and materials science. Large molecule design is inherently more complex and less explored compared to designing small molecules, adding significant difficulty in generative modeling. We aim to establish strong baselines for better scalability, efficiency, and generative performance in this domain. We evaluate the scalability and performance of generative AI models, initially effective for small molecule design, in generating large molecules for potential drugs in gene-based therapies, immunotherapies, hormonal regulators, and targeted cancer therapies. Our findings indicate that computational strategies and model architectures designed for small molecules may not readily extend to large molecular structures. To address these limitations, we explore masked language modeling strategies alongside advanced tokenization methods, including Atom-Pair Encoding (APE), to enhance generative AI models. We probe how incorporating such strategies, particularly the APE tokenization method that explicitly captures structural and chemical characteristics, can significantly improve design capabilities for complex molecular structures. Overall, our results demonstrate both the potential and challenges of deep generative modeling for large molecules and how the proposed enhancements may bridge the gap in generating large molecules when novel discovery is the ultimate goal.

Keywords: Molecule generation; Large molecule; Generative model; Molecule representation; Self-supervised Learning; String representation; SMILES; SELFIES; Tokenization; BPE; APE

Title: DG-scRNA: Deep Learning with Graphic Cluster Visualization to Predict Cell Types of Single-Cell RNAseq Data
Author list: Birkan Gokbag, Yimin Liu, Abhishek Majumdar, Lang Li, Chongwen Dong, Yanan Song, Wei Xia, and Lijun Cheng
Abstract: Single-cell RNA sequencing (scRNA-seq) has revolutionized understanding of cellular heterogeneity, but accurate cell type annotation remains challenging. Marker genes, essential for distinguishing cell types, vary by tissue origin, disease state, and experimental methods. Here, we present DG-scRNA, a deep learning framework with graphic cluster visualization that systematically optimizes marker selection based on sample context, improving cell type identification. Validated with T-cell annotation from thyroid cancer scRNA-seq data, DG-scRNA achieved superior performance (F1 score: 95.19%) compared to existing annotation methods. Its automated, context-specific marker selection matches optimal markers based on species, tissue, and disease state. Applying DG-scRNA to papillary thyroid carcinoma samples uncovered distinct T-cell subpopulations associated with metastasis patterns. DG-scRNA offers an accurate, context-aware solution for cell type annotation across diverse systems and disease settings.
Keywords: Thyroid cancer; Single-cell RNA sequencing (scRNA-seq); Cell type annotation

Title: A Machine Learning–Enhanced Pipeline for Detecting Disruption of Transcription Termination (DoTT) in RNA-Seq Data
Author list: Michael Levin, Igor Astsaturov, and Yunyun Zhou
Abstract: Disruption of transcription termination (DoTT) occurs when RNA polymerase II fails to stop at the 3′ end of a gene, producing readthrough transcripts often missed by standard RNA-seq pipelines. We developed a pipeline that extends gene annotations downstream, quantifies readthrough reads, and applies differential expression analysis. A Random Forest classifier further boosts sensitivity (25% → 65%). Applied to high-carbohydrate diet and HSV-1 infection datasets, the pipeline detected ~50 and 707 DoTT events, respectively, recovering ~78% of known HSV-1 events. Pathway analysis revealed immune-related pathways, highlighting the pipeline's broad applicability and improved DoTT detection.
Keywords: Transcriptional readthrough; Transcription termination failure; RNA-seq; High-carbohydrate diet; HSV-1; Machine learning; Random Forest; DoGFinder; Readon

Title: THANOS: An AI Pipeline for Engineering Antibodies
Author list: Arnav Solanki1, Neha S Maurya1, Wenjin Jim Zheng1
Detailed Affiliations:
1McWilliams School of Bioinformatics, The University of Texas Health Science Center at Houston, Houston, TX, USA
Abstract: The Brain Research Through Advancing Innovative Neurotechnologies (BRAIN) initiative seeks to unravel the complexities of brain cell types, connections, and functions. One powerful approach to studying neural circuits is imaging brain proteins using antibodies. However, generating high-quality antibodies is often slow and expensive. Recent advances in AI tools like AlphaFold3 and RFdiffusion offer

a fast, fully digital alternative to traditional experimental screening. This project introduces THANOS (Targeted High-throughput Antibody Notation, Optimization, & Screening), a novel pipeline for rapidly engineering. THANOS was used to design de novo antibodies targeting human proteins by redesigning antigen-binding sites of antibodies using a high-performance GPU server.
This case study focuses on Parvalbumin (PVALB), a calcium-binding protein abundant in neural cells. Alphafold3 was used to model the 3D complexes of human PVALB and 12 mouse antibody variable fragments. With these 12 models as initial states, 300 structural variants were generated by using RFdiffusion to diffuse the complementarity determining regions (CDRs) at the binding sites. ProteinMPNN was used to predict optimal sequences that fold these structures, yielding new antibody chain sequences containing new residues in the CDRs. These 300 variants were screened against PVALB using AlphaFold3 to predict their binding. 4 candidates were observed to bind strongly based on low predicted alignment error. These candidates were validated through:

1) Structural inspection using ChimeraX.
2) Molecular dynamics simulations with GROMACS. The best candidate demonstrated a MMGBSA energy of -35 kcal/mol over 100 ns.
3) Solubility checks using Aggrescan3D to confirm the absence of aggregation-prone residues.
4) Sequence alignment to ensure minimal mutations and preserve nativeness.

These antibodies are currently undergoing experimental validation. THANOS demonstrates how AI can accelerate antibody engineering from weeks to hours without a wet lab. This pipeline can be applied on any desired target protein (beyond the interest of the BRAIN initiative) for numerous applications such as viral or cancer therapy, and will be invaluable to the fields of immunology and pharmacology.
Keywords: Antibodies, Protein Engineering, AI, Machine Learning, AlphaFold

Title: GRN-Integrated Heterogeneous Attentive Graph Autoencoder for Cell-Cell Interaction Reconstruction from Spatial Transcriptomics
Author list: Aiwei Yang, Yujian Lee, Yue Guan and Jiaxing Chen
Abstract: The reconstruction of cell-cell interaction networks (CCIs) is pivotal to unraveling the regulatory mechanisms governing orchestrated multicellular systems. While deep learning models show advances in inferring CCIs from spatial transcriptomes; most current approaches neglect intracellular gene regulatory dynamics that mediate communication or only get to suboptimal integration of spatial cellular contexts with gene regulatory networks (GRNs), while neglecting the potential data noise amplification through model process. To address these challenges, we propose SFNET, a robust graph neural framework that synergizes noise-resilient graph construction with topology-aware deep learning. Our Shared Factor Neighborhood (SFN) algorithm constructs cell graphs through joint optimization of spatial coordinates and genetic features, better reducing noise sensitivity compared to conventional KNN approaches.The Heterogeneous Attention Embedding (HAE) module then explicitly models cell-GRN interactions via multi-head cross-domain attention to preserve biological specificity. Finally, our Triple-Enhancement Graph Neural Network (CECB/SEB/EEB) combats feature degradation through multi-scale enhancement blocks. Enabling precise modeling of both local and long-range interactions in heterogeneous networks. When benchmarked versus existing models, SFNET reduces training time by 50% without compromising accuracy. It improves the average precision (AP) by 2.02%, ROC by 2.56%, and AUROC by 1.72%, highlighting the gains in both accuracy and computational efficiency in CCI inference.
Keywords:

Cell-cell Interaction Network; Graph Neural Network; Shared Neighbourhood Algorithm; Single-cell Spatial Transcriptomics

Data-Driven Insights into Disease Modeling

August 5th
1:30 AM – 4:50 PM
Room: 301

Chairs: Shulan Tian, Joseph McElroy

Title: Compositional Bayesian Co-Clustering of DTI Biomarkers with Clinical Measures for Enhanced Prediction of Parkinson Disease Severity
Author list: Ashwin Vinod and Chandrajit Bajaj
Abstract: Parkinson's disease (PD) exhibits inter-patient heterogeneity complicating prognosis and precision therapy. We propose an end-to-end Compositional Bayesian Co-Clustering (SRVCC) framework that integrates tissue-specific diffusion-tensor imaging (DTI) biomarkers and clinical assessments to uncover multimodal patterns predictive of disease severity. Four-dimensional DTI scans from the Parkinson's Progression Markers Initiative were processed to generate extracellular-contamination-free fractional anisotropy and mean diffusivity. Combined with clinical scores (UPDRS and MoCA), SRVCC jointly clusters imaging and clinical data in a variational latent space, suppressing noise while preserving discriminative latent modes. Cross-validation identified three patient subtypes corresponding to mild, intermediate, and severe PD, with imaging metrics mirroring progression. SRVCC outperformed k-means, spectral bi-clustering, and deep-clustering baselines, offering biologically consistent clusters differentiating cognitive status, motor severity, and microstructural changes, bridging DTI alterations and clinical manifestations.
Keywords: DTI imaging; Parkinson's Disease; Compositional Bayesian Co-Clustering

Title: Latent factor modeling reveals unexpected spatial heterogeneity in human Alzheimer's disease brain transcriptomes
Author list: Rami Al-Ouran, Chaozhong Liu, Linhua Wang, Ying-Wooi Wan, Chaohao Gu, Xiqi Li, Gerarda Cappuccio, Mirjana Maletic-Savatic, Aleksandar Milosavljevic, Joshua Shulman, Hu Chen and Zhandong Liu
Abstract: Alzheimer's disease is characterized by complex molecular and cellular heterogeneity, which complicates efforts to identify consistent biomarkers and therapeutic targets. To gain a deeper understanding of the heterogeneity, we applied latent factor modeling to RNA-seq data from approximately 2,500 human Alzheimer's disease brain samples, uncovering underlying patterns in gene expression. These

transcriptional groups demonstrated unique gene expression profiles related to synaptic and neuronal pathways, vasculature development, and protein folding and antigen processing. We demonstrated that this latent factor emerges from variations in spatial sampling. Adjusting for the latent factor recovers nearly three times more differentially expressed genes than analyses not stratified by this factor. This finding suggests that spatial heterogeneity is a pervasive element across various cellular and molecular brain profiles and has far-reaching implications for future studies of Alzheimer's disease and related neurological disorders.
Keywords: Brain transcriptome; Latent factor; Sampling variations

Title: DuAL-Net: A Hybrid Framework for Alzheimer's Disease Prediction from Whole Genome Sequencing via Local SNP Windows and Global Annotations
Author list: Eun Hye Lee, Taeho Jo
Abstract: Alzheimer's disease (AD) dementia is the most common formof dementia. With the emergence of disease-modifyingtherapies for AD such as anti-amyloid monoclonalantibodies, the ability to predict disease risk beforesymptom onset has become increasingly important. Wholegenome sequencing (WGS) data is a promising data form forearly AD prediction, despite several analytical challenges.In this study, we introduce DuAL Net, a hybrid deeplearning framework designed to predict AD dementia usingWGS data.DuAL-Net integrates two components: local probabilitymodeling, which segments the genome into non-overlappingwindows, and global annotation-based modeling, whichannotates each SNP and reorganizes the WGS input to capturelong range functional relationships. Both components employout of fold stacking with TabNet and Random Forestclassifiers. The final prediction is generated by combininglocal and global probabilities using an optimized weightingparameter alpha.We applied DuAL-Net to WGS data from 1,050 individuals (443cognitively normal and 607 with AD dementia), usingfive-fold cross validation for training and evaluation. Onaverage across the 100, 500, and 1000 SNP subset sizesevaluated, DuAL-Net achieved an Area Under the Curve (AUC)of 0.671 using top-ranked SNPs prioritized by the model,representing 35.0% and 20.3% higher predictive performancecompared to the average AUCs of bottom-ranked and randomlyselected SNPs, respectively. Assessment of modeldiscriminative ability via ROC analysis across differentSNP subset sizes consistently demonstrated a strongpositive correlation between the SNPs' prioritization rankand their predictive power. The model identified SNPs withknown associations to AD as top contributors to prediction,alongside potentially novel variants also ranked highly bythe model.In conclusion, DuAL Net presented a promising framework forAD prediction that improved predictive accuracy andenhanced biological interpretability. The framework and itsweb-based implementation offer an accessible platform forbroader research applications.
Keywords: Alzheimer's disease; Disease prediction; Whole genome sequencing; Deep learning; Machine learning

Title: Resolving Gene Heterogeneity in DEG Analysis: A Novel Pipeline for Precision Genomics
Author list: Jiasheng Wang, Ilia Buralkin, Rami Ai-Ouran and Zhandong Liu
Abstract: Gene heterogeneity, driven by extensive genetic variation across samples, poses significant challenges in identifying disease-associated genes through Differentially Expressed Gene (DEG) analysis. Traditional global DEG methods often fail to capture subtle yet biologically meaningful signals, which are obscured by genetic variability. To address this, we developed a novel DEG analysis pipeline that integrates DNA and RNA data to amplify signals within genetically similar local regions. This approach combines

dimensionality reduction techniques, local contrast identification, and co-regulation modeling to uncover subgroup-specific DEG signals. Our results demonstrate that this local analysis method significantly outperforms traditional global DEG approaches, particularly in detecting weak signals or those localized to small subgroups of samples. Applying this pipeline to the ROSMAP dataset identified key Alzheimer's disease (AD)-related pathways, such as ATP biosynthesis, nervous system development, and synaptic signaling, which were missed by conventional methods. Cross-dataset validation further confirmed the robustness of our approach, showing improved consistency in DEG detection and capturing a broader spectrum of gene expression changes. This study highlights the importance of addressing genomic heterogeneity in DEG analysis and offers a powerful tool for uncovering biologically relevant pathways and disease mechanisms. The proposed method has broad implications for precision medicine, enabling the identification of subgroup-specific signals in complex, heterogeneous datasets.
Keywords: Gene Heterogeneity; Alzheimer's Disease; Differentially Expressed Genes

Title: Multimodal Imaging and Cell-Free DNA Methylation Analysis for Noninvasive Lung Cancer Diagnosis
Author list: Ran Hu, Stephen Park, Paul Li, Weihua Zeng, Yonggang Zhou, Chun-Chi Liu, Shuo Li, Xiaohui Ni, Kostyantyn Krysan, Steven Dubinett, Denise Aberle, Ashley Prosper, Wenyuan Li, William Hsu and Xianghong Zhou
Abstract: Background: Low-dose computed tomography (LDCT) is an effective noninvasive screening tool for lung cancer. However, imaging-detected lesions often require invasive follow-up procedures for definitive diagnosis, increasing healthcare costs and the risk of overdiagnosis. There is a pressing need for additional noninvasive methods to improve diagnostic accuracy in patients with CT-detected lung lesions. Objectives: This study aims to develop a multimodal approach that integrates CT imaging and cell-free DNA (cfDNA) methylome data for noninvasive lung cancer diagnosis.
Methods: We utilized large single-modality datasets to pretrain deep learning (DL) models that extract robust and informative features from high-dimensional imaging and methylome data. For imaging, a foundation model was fine-tuned on 677 lung CT lesions to improve its ability to capture lung-specific imaging patterns. Handcrafted radiomic features were also extracted from the CT scans. For cfDNA methylation, lung cancer-specific biomarkers were first identified, followed by training an autoencoder model on 513 normal and lung cancer plasma samples to generate meaningful methylation feature embeddings. After feature extraction, we integrated the multimodal features using early, intermediate, and late fusion strategies, and evaluated model performance on 77 individuals with paired imaging and methylome data using support vector machine (SVM) and neural network (NN) classifiers under 5-fold cross-validation (CV).
Results: On this multimodal dataset, intermediate fusion of imaging and methylation features achieved the highest area under the receiver operating characteristic curve (AUC = 0.870 ± 0.128). The model also demonstrated strong performance in early-stage lung cancer detection, achieving an AUC of 0.806 for Stage I cases versus controls with CT-detected benign lesions.
Conclusions: Integrating multimodal imaging and cfDNA methylation features enhances the accuracy of lung cancer diagnosis and holds promise as a noninvasive approach for distinguishing malignant from benign CT-detected lesions.
Keywords: lung cancer; foundation model; deep learning; machine learning; medical imaging; cell-free DNA; methylation; multimodal data

Title: Multidimensional Impact of Microbiota Absence on Thymic T Cell Development in Mice: A Study Based on Single-Cell and Spatial Transcriptomics

Author list: Yifei Sheng, Qian Zhang, Zhao Zhang and Juan Shen

Abstract: Background: Gut microbiota plays an important role in host immune development, but the mechanisms of its influence on primary lymphoid organs such as the thymus remain unclear. Germ-free mice provide an ideal model for studying the impact of microbiota absence on thymus development.

Methods: This study utilized single-cell transcriptomics and spatial transcriptomics techniques to systematically compare thymic cellular composition and gene expression characteristics between germ-free mice and specific pathogen-free mice at different developmental stages (0, 2, 4, and 10 weeks of age).

Results: Spatial transcriptome analysis revealed that GF mouse thymus had 5 basic transcriptome classifications and lacked plasma cells and neutrophils, while SPF mouse thymus had 7 basic classifications. Although early T cell development was not affected under germ-free conditions (no significant differences in DN T cell numbers and expression of activation marker Itgal), T cell proliferation in pre-pubescent GF mice was significantly lower than in SPF mice, a difference that largely disappeared after puberty (10 weeks). Longitudinal analysis found that with increasing age, double-positive T cells gradually decreased while immature T cells increased. GF mice exhibited more pronounced Th1/Th2 imbalance and greater cell number fluctuations compared to SPF mice. Additionally, PLZF+ innate lymphoid cells in GF mice showed impaired early development but significant increase at 10 weeks of age, possibly representing a compensatory mechanism. Aire expression in thymic mesenchymal cells was significantly lower in GF mice compared to SPF mice.

Conclusion: The impact of microbiota absence on thymic T cell development exhibits time- and cell type-specific patterns. While early T cell development remains unaffected, microbiota absence leads to restricted T cell proliferation, Th1/Th2 imbalance, abnormal innate lymphoid cell development, and decreased Aire expression. These findings provide new perspectives for understanding the role of microbiota in shaping the host immune system.

Keywords: Germ-free mice; thymus;T cell development; single-cell transcriptomics;spatial transcriptomics

Title: The Drug Overdose Surveillance in Ohio: What we can see with the geospatial shared component analysis of the Urine Drug Test Results

Author list: Joanne Kim[1], John Myers[1], Charles Marks[2], Penn Whitley[2], Brandon Slover[1], Xianhui Chen [1], Neena Thomas[1], Ping Zhang[1], Naleef Fareed [1], Soledad Fernandez[1].

Detailed Affiliations: [1]Department of Biomedical Informatics, College of Medicine, The Ohio State University; [2] Millenium Health, LLC;

Abstract: Geospatial analysis of the substance use disorder (SUD) population has provided various insights for the surveillance of the SUD population. Numerous data sources have been investigated but the chronic challenge regarding delayed reporting and the scarcity of the data still remains.

To overcome this challenge, we conducted the Bayesian multivariate spatiotemporal modeling analysis using the real-time Urine drug test results for diverse sets of drugs (e.g. Fentanyl, Cocaine, Heroine and Methamphetamine). We use the multivariate Bayesian spatiotemporal approach to investigate the shared geospatial pattern of the substance use population. By looking at their shared components, we can investigate the co-evolving pattern of the drug substance use population in each county from 2013 to 2023. With this effort, we can confirm the existing belief about polysubstance use, and identify new shared

patterns with newly emerged substances. We also expect information sharing of multiple drugs can help improve the estimation results of small areas. This talk will discuss the analysis results for various sets of drugs and how the map of substance use population changes in the 10-year period in Ohio.
Keywords: Opioid overdose, Bayesian methods, shared component model, Urine Drug Test, Spatiotemporal

Title: Telehealth Utilization and Patient Experiences: The Role of Social Determinants of Health Among Individuals with Hypertension and Diabetes
Author list: Haoxin Chen, Will Simmons, Malak Hashish and Jiancheng Ye
Abstract: Objective: To evaluate the utilization patterns, effectiveness, and patient satisfaction of telehealth services among individuals with hypertension and/or diabetes, and to investigate the influence of social determinants of health (SDOH) on telehealth access and utilization in this population.
Methods: We conducted a cross-sectional analysis using data from the 2022 Health Information National Trends Survey (HINTS 6) by the National Cancer Institute. The study sample included 3,009 respondents with self-reported diabetes, hypertension, or both conditions. Telehealth usage was assessed through 14 survey questions, and participant characteristics were analyzed using sociodemographic, baseline health, and SDOH data.
Results: Of the 6,252 HINTS 6 survey respondents, 3,009 met the inclusion criteria. Significant sociodemographic differences were observed across the diabetes and/or hypertension groups. No significant differences were found in telehealth usage among the groups, with 43.9% of respondents utilizing telehealth in the past year. Common reasons for telehealth use included provider recommendation, convenience, and infection avoidance. Social determinants of health, such as food insecurity and transportation issues, were more prevalent among individuals with both conditions, though no significant differences in telehealth experiences were noted across groups.
Conclusion: Telehealth shows potential for managing chronic conditions like hypertension and diabetes, demonstrating substantial adoption and universal accessibility. However, disparities influenced by SDOH highlight the need for targeted interventions to ensure equitable access. Addressing privacy concerns, leveraging healthcare providers' recommendations, and tackling SDOH barriers are crucial for fostering wider telehealth adoption and improving outcomes. Future research should focus on the long-term impacts of telehealth and further investigate SDOH factors to develop tailored interventions that enhance engagement and equitable access across diverse patient populations.
Keywords: Telehealth; Technology utilization; Social determinants of health; Hypertension; Diabetes; Multiple chronic conditions

Title: AutoRADP: An Interpretable Deep Learning Framework to Predict Rapid Progression for Alzheimer's Disease and Related Dementias Using Electronic Health Records
Author list: Qiang Yang, Weimin Meng, Pei Zhuang, Stephen Anton, Yonghui Wu and Rui Yin
Abstract: Alzheimer's disease (AD) and AD-related dementias (ADRD) exhibit heterogeneous progression rates, with rapid progression (RP) posing significant challenges for timely intervention and treatment. The increasingly available patient-centered electronic health records (EHRs) have made it possible to develop advanced machine learning models for risk prediction of disease progression by leveraging comprehensive clinical, demographic, and laboratory data. In this study, we propose AutoRADP, an interpretable autoencoder-based framework that predicts rapid AD/ADRD progression using both

structured and unstructured EHR data from UFHealth. AutoRADP incorporates a rule-based natural language processing method to extract critical cognitive assessments from clinical notes, combined with feature selection techniques to identify essential structured EHR features. To address the data imbalance issue, we implement a hybrid sampling strategy that combines similarity-based and clustering-based upsampling. Additionally, by utilizing SHapley Additive exPlanations (SHAP) values, we provide interpretable predictions, shedding light on the key factors driving the rapid progression of AD/ADRD. We demonstrate that AutoRADP outperforms existing methods, highlighting the potential of our framework to advance precision medicine by enabling accurate and interpretable predictions of rapid AD/ADRD progression, and thereby supporting improved clinical decision-making and personalized interventions.
Keywords: Alzheimer's Disease and Related Dementias; Deep learning; Interpretable learning; Electronic Health Records; Data Imbalance

Technology Session
October 10th
3:40 PM – 5:40 PM
Room: 320

Chair: Yu-Chiao Chiu

Title: Boosting Pipeline Efficiency in Bioinformatics Through Snakemake
Author list: Shunian Xiang[1], Hua ke[1], Jingling Hou[1], Nihir Patel[1], Yaoqi Li[1], Haixin Shu[1], Si Chen[1], Yaping Feng[1]
Detailed Affiliations:
[1]Department of Bioinformatics, Admera Health, New Jersey, NJ, USA
Abstract: Given the complexity and diversity of modern bioinformatics (BI) analyses, automation has become an essential priority—particularly for biotechnology companies that manage high volumes of multi-species projects with custom client requirements. Automating BI workflows through the integration of scripts, pipelines, and AI-assisted systems enables a more streamlined, assembly line-like approach, improving scalability, analysis speed, reproducibility, and reducing both manual effort and human error. Ultimately, this allows researchers to focus more on scientific discovery in biology and medicine.
We present a modular bioinformatics workflow framework built with Snakemake, a powerful and scalable workflow management system. Our system supports a broad spectrum of next-generation sequencing (NGS) data types, including bulk RNA-seq, small RNA-seq, microRNA-seq, single-cell RNA-seq, spatial transcriptomics, ChIP-seq, ATAC-seq and so on. Each pipeline is composed of independent, reusable modules—for example, the RNA-seq workflow includes quality control, adapter trimming, genome alignment, quantification, differential expression analysis, and pathway enrichment—which can be flexibly assembled and automatically executed through a simple command-line interface.
The system can automatically assemble workflows by selecting and combining appropriate modules based on user input, allowing flexible customization without sacrificing automation. This design significantly

reduces manual workload, shortens turnaround time, and adapts easily to diverse project requirements—enabling efficient, reproducible, and scalable bioinformatics analysis.
Keywords: snakemake, automation, bioinformatics, efficiency, bulk RNA-seq, small RNA-seq, microRNA-seq, single-cell RNA-seq, spatial transcriptomics, chip-seq, atac-seq

Title: Spatial Transcriptomics at Scale with Stereo-seq: Big Data for Impactful Science
Author list: Yongfu Wang
Detailed Affiliations:
Complete Genomics
Abstract: Stereo-seq, originated from DNBSEQTM technology, is the highest resolution spatial transcriptomics platform available today. With 0.5 μm resolution and chip sizes up to 13 cm × 13 cm, Stereo-seq enables precise molecular mapping across whole organs or large tissue sections—ideal for developmental biology, oncology, neuroscience, and cross-species studies. This open, species agnostic platform supports both Fresh Frozen and FFPE samples, integrates transcriptomics with proteomics or histology, and captures total RNA—including non-coding RNAs and microbiome content—from FFPE samples. Hundreds to thousands of terabytes of data have been generated, and the scientific community continues to develop new tools to mine these datasets in pursuit of answers to the secret of life. This seminar will highlight the exciting advancements Stereo-seq has brought to the forefront of scientific discovery.

Title Access the full richness of biological complexity with single cell and spatial multiomics from 10x Genomics
Author list: Nicole Jaymalin
Detailed Affiliations:
10x Genomics, Pleasanton, California, USA.]
Abstract: Developing treatments for complex diseases requires building a complete understanding of both disease and treatment-response mechanisms. As we navigate a century where transformative advances in biology will reshape the way we deliver human health, translational and clinical researchers need approaches that provide actionable insights that can, ultimately, be leveraged to improve how diseases are diagnosed and treated.
Join us to learn how single cell, spatial, and in situ innovations from 10x Genomics can help you push the boundaries of your translational and clinical research. Discover novel therapeutic targets, explore how therapeutics modulate disease-associated cell populations and states, gain insights into mechanisms governing therapeutic toxicity, and understand resistance mechanisms governed by transcriptomic and epigenetic remodeling. Enabling deeper insight into cancer, immunology, neuroscience, and immuno-oncology, 10x Genomics gives researchers the ability to see biology in new ways.
Keywords: Single cell, Spatial Transcriptomics, In Situ

Title: Directed Evolution of Molecular Enzymes Empowers NGS Library Preparation
Author list: Robin Song
Detailed Affiliations:
Yeasen Biotechnology Co., Ltd.

Abstract: Next-generation sequencing (NGS) has transformed genomics, transcriptomics, and precision medicine by enabling high-throughput, large-scale nucleic acid analysis. At the heart of this revolution lies molecular enzyme evolution, which continuously drives improvements in the sensitivity, specificity, and efficiency of sequencing workflows. Through advanced protein engineering and directed evolution, novel enzymes are developed to overcome technical challenges in library preparation, amplification, and data quality, paving the way for faster, more accurate, and cost-effective solutions.
This presentation will explore how cutting-edge enzyme innovation is reshaping the future of genomics by empowering genome and transcriptome sequencing , methylation analysis, and ultra-low input detection. We will highlight the role of enzyme-optimized reagent kits in enhancing experimental robustness and reliability, accelerating research breakthroughs, and expanding applications. By combining enzyme engineering with innovative sequencing approaches, we aim to help promote the future of genomics and enable new frontiers in life science and healthcare.
Keywords: Next-Generation Sequencing (NGS), Molecular Enzyme Engineering, Directed Evolution, Library Preparation, Genomic study, Transcriptome Analysis

Title: Uncover Cellular Heterogeneity with Advanced Single Cell Multi-Omics Approaches
Author list: Julie Laliberte[1], Jing Zhou[1]
Detailed Affiliations:
[1]Singleron Biotechnologies Inc., USA
Abstract: Singleron Biotechnologies is a pioneering molecular diagnostics company dedicated to advancing clinical diagnostics, drug development, and health management through cutting-edge single-cell analysis technologies. Singleron provides comprehensive solutions for single-cell sequencing, offering both instrument-based workflows for in-lab use and full-service options through our global network of service laboratories.
What sets Singleron apart is our proprietary single-cell partitioning system. Our SCOPE-chip platform utilizes a gravity-driven micro-well microfluidics system for gentle and efficient cell partitioning and RNA barcoding. This approach ensures robust performance—even with fragile, rare cell types or even nuclei.
In addition to high-throughput single-cell RNA sequencing, Singleron has developed a suite of innovative multi-omics technologies to extract deeper insights from each cell:

- **DynaSCOPE** detects nascent RNA, capturing transcriptional dynamics and providing valuable temporal information—particularly useful in applications like drug screening.
- **MobiuSCOPE** enables full-length transcript sequencing, making it ideal for detecting splice variants and SNPs across entire transcripts—overcoming the limitations of conventional 3′ or 5′ RNA-seq approaches.
- **FocuSCOPE** enriches specific transcripts alongside whole transcriptome profiling, increasing sensitivity for targeted gene expression analysis.
- **ProMoSCOPE** simultaneously profiles cell surface glycans and transcriptomes—offering critical insights into cell–cell interactions and immune responses.
- **Scircle** captures full-length T-cell and B-cell receptor sequences along with the whole transcriptome, allowing researchers to identify immune clonotypes of interest in cancer, autoimmune diseases, and vaccine development.

To support data interpretation, Singleron offers advanced analysis tools and resources. Our SynEcoSys platform hosts over 46 million single cells from 731+ datasets across multiple species, all uniformly processed and expertly annotated to enable reliable cross-study comparisons.

Keywords: Single cell sequencing, Multiomics, Tissue dissociation, Advanced Bioinformatics tools

Future Scientists in AI Session

August 4th
9:20 AM – 12:20 PM
Room: 301

Chair: Chi Zhang, Jingwen Yan

Title: Unlocking Fine-Grained Features: Vision Foundation Models for Improved Skin Cancer Classification

Author list: Alex Fu1, Jiachen Yao2, Chao Chen3
Detailed Affiliations:
1Union County Magnet High School, Scotch Plains, NJ, USA; 2Department of Computer Science, Stony Brook University, Stony Brook, NY, USA; 3Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY, USA.

Abstract: Skin cancer is one of the most common and deadly forms of cancer worldwide. Its diagnosis traditionally relies on expert visual assessment, which can be subjective and resource-intensive.
Automated, AI-based methods offer a promising path to more accessible, scalable, and reliable diagnostic workflows. However, multiclass skin cancer classification remains challenging due to subtle inter-class differences and intra-class variability in lesion appearance.
Recent advances in vision foundation models have demonstrated remarkable generalizability across visual tasks. These models are large neural networks pre-trained on diverse datasets, excel at learning transferable and generalized representations. Despite their success in other domains, their application in dermatological image analysis is still underexplored. In this study, we investigate the potential of finetuned foundation models to improve the accuracy of multiclass skin cancer classification.
We specifically utilized the DINOv2 (DIstillation with NO labels version 2) foundation model, developed by Meta. DINOv2 is a vision transformer (ViT) trained by a self-supervised learning approach. It was trained on an extensive and diverse unlabeled image dataset sourced from the public domain; these images have no explicit image labels. The training methodology involves a teacher-student network architecture that enforces consistent predictions through knowledge distillation, with an exponential moving average used for updating the teacher network's weights.
For our study, we selected this DINOv2 vision transformer as our core foundation model. We then finetuned it using two widely-used dermatoscopic image datasets: HAM10000 and HIBA. Our specific finetuning strategy involved freezing the backbone of the pre-trained DINOv2 model and subsequently training a custom prediction head positioned on top of this frozen backbone. We compared the performance of this finetuned DINOv2 approach against several established baseline architectures, including ResNet-50, DenseNet, EfficientNet, and FixCaps.

Our results show that the finetuned foundation model outperformed all baselines, achieving a 6.8% improvement in classification accuracy on HAM10000 and a 7.2% improvement on HIBA. These findings suggest that vision foundation models can capture fine-grained visual cues critical for distinguishing between similar skin lesion types, making them highly suitable for dermatological diagnostics.
This project demonstrates the value of leveraging advanced AI models in biomedical imaging and highlights a promising direction for future research in automated skin cancer detection.

Title: Automated Clinical Diagnosis using ML and Electronic Health Records: A Prototype in IBD
Author list: Anshu Mukherjee1, Vivek Rudrapatna1,2
Detailed Affiliations:
1Bakar Computational Health Sciences Institute, University of California, San Francisco, CA, USA;
2Division of Gastroenterology, Department of Medicine, University of California, San Francisco, CA, USA.
Abstract: Background: Inflammatory Bowel Disease (IBD) is a chronic autoimmune condition characterized by intense inflammatory flares. Diagnostic delay for IBD remains a challenge, spanning 5-16 months on average, endangering patients: delays of just 3 months lead to increased risk of adverse events by 18.6%. To address this, we created a novel decision support tool (DST) integrating electronic health records (EHR) with environmental and public health data to predict IBD onset up to six months before standard clinical diagnosis.
Methods: Clinical history from 9000 patients were extracted from the UCSF EHR system. Environmental and public health data, such as water quality, air quality, and food deserts, were collected and processed by ZIP code. After rigorous processing, imputation, and feature selection, approximately 250 clinical & environmental features were used to create a final feature matrix. Patients were split into three different cohorts by time before diagnosis: six months, three months, and at diagnosis. Data was further segregated by adding a generic GI patients dataset (Tier 1), and an IBD specific dataset (Tier 2). Three ML models were selected: random forest, logistic regression, and light gradient boosting machine (LGBM). This resulted in 18 distinct models across 3 time points, 2 tiers, and 3 model types.
Results: All models had accuracy of over 80% on testing sets using 5-fold cross-validation for tuning. LGBM performed best, achieving over 90% accuracy (>0.95 AuROC) across the board for 6 months, 3 months, and diagnosis date for tier 1. For tier 2, it achieved 88% accuracy (0.95 AuROC) for 6 months, 94% accuracy (0.98 AuROC) for 3 months, and 98% accuracy (0.99 AuROC) at diagnosis date. Feature importance analysis showed kidney function indicators like estimated glomerular filtration rate (eGFR) and anion gap drove model performance. This supports deeper connections between the kidney and gut, consistent with the observations of a recent long-term observational study. Other high-importance features, such as mean corpuscular hemoglobin concentration (MCHC), eosinophil levels, and reported stress, were consistent with medical consensus around IBD.
Discussion: Overall, the DST shows promise for use in real-life clinical application, and could provide countless IBD patients with quicker answers, better outcomes, and relief from painful symptoms. Furthermore, the pipeline developed here could be easily expanded to create early diagnosis tools and identify novel risk factors for a multitude of rare and chronic diseases.

Title: Applications of Neural Networks in Chaperonin Generation for Complement Factor Renaturalization
Author list: Arhan Patel, Matthew Fang

Detailed Affiliations:
1Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA; 1Department of Biomedical Informatics, University at Buffalo, Buffalo, NY, USA; 1Department of Biological Sciences, University at Buffalo, Buffalo, NY, USA; 1,2Department of Pharmacy and Pharmaceutical Sciences, University at Buffalo, Buffalo, NY, USA
Abstract: Recent advances in the field of artificial intelligence have enabled the development of predictive models for complex biological systems. Here, we present a novel strategy to address malfunctioning complement factors and proteins caused by genetic mutations. We hypothesized that delivering mRNA encoding for mutation-specific chaperonins could allow for the restoration of correct complement factor folding in the liver. To test this, we designed a dual-output transformer neural network model, OmegaFold, to predict the 3-D model of mutated complement factor while simultaneously generating an amino acid sequence for the optimal mutation-specific chaperonin. Our results demonstrate the model's viability in generating compatible chaperonins for complement factors, indicating a promising therapeutic strategy for addressing complement factor dysregulation.

Title: MODE: high-resolution digital dissociation with deep multimodal autoencoder
Author list: Ayesha A. Malik2, Jiao Sun1, Tong Lin1 , Ayla Bratton2 , Yue Pan1 3 , Kyle Smith3,5, Arzu Onar-Thomas1 , Giles W. Robinson4 , Wei Zhang2 4 , Paul A. Northcott3,5, and Qian Li1,*
Detailed Affiliations:
1 Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN, 38105.
2 Department of Computer Science, University of Central Florida, Orlando, FL, 32816.
3 Department of Developmental Neurobiology, St. Jude Children's Research Hospital, Memphis, TN, 38105.
4 Department of Oncology, St. Jude Children's Research Hospital, Memphis, TN, 38105.
5 Center of Excellence for Neuro-Oncology Sciences, St. Jude Children's Research Hospital, Memphis, TN, 38105.
Abstract: In single cell biology, the complexity of tissues may hinder lineage cell mapping or tumor microenvironment decomposition, requiring digital dissociation of bulk tissues. Many deconvolution methods focus on transcriptomic assay, not easily applicable to other omics due to ambiguous cell markers and reference-to-target difference. Here, we present MODE, a multimodal autoencoder pipeline linking multi-dimensional features to jointly predict personalized multi-omic profiles and cellular compositions, using pseudo-bulk data constructed by internal non-transcriptomic reference and external scRNA-seq data. MODE was evaluated through rigorous simulation experiments and real multi-omic data from multiple tissue types, outperforming nine deconvolution pipelines with superior generalizability and fidelity.

Title: Empowering Confident Communication: Artificial Intelligence Based Detection of Stuttering Patterns for Speech Therapy
Author list: David Yuanze Li1, Frank Alber2
Detailed Affiliations:
1 Thousand Oaks High School, 2323 N Moorpark Road, Thousand Oaks, CA 91360, USA; 2 University of California at Los Angeles, Los Angeles, CA, 90095, USA

Abstract: Stuttering affects over three million Americans, profoundly impacting communication, social interactions, and confidence. Witnessing my father's lifelong struggle with stuttering inspired me to explore how cutting-edge technology could help others facing similar challenges. In this study, I investigated four primary dysfluency patterns: prolongation (stretched sounds), block (involuntary pauses mid-speech), sound repetition, and word repetition. Accurate identification of these patterns is essential for developing personalized therapy and providing real-time feedback.

While recent research has explored various audio feature extraction and machine learning techniques for stuttering detection, most existing approaches yield suboptimal performance and fail to fully capture the temporal dynamics of speech dysfluencies during feature extraction and modeling. Therefore, I employed a range of features, from traditional human-engineered acoustic features such as mel-frequency cepstral coefficients (MFCC) to advanced self-supervised speech representations from models like wav2vec2 and HuBERT. I proposed an attention-augmented bidirectional long short-term memory network (BiLSTM), which captures past and future context critical for detecting dysfluent patterns.

I systematically compared these features and models. My results demonstrate that self-supervised speech features, particularly HuBERT embeddings, combined with BiLSTM, significantly outperform traditional features and models. Specifically, this approach improved stuttering detection performance by over 35% compared to models using handcrafted features and classical machine learning methods across all four stuttering event types.

Specifically, I conducted experiments using two datasets: SEP-28k (> 28,000 audio clips) and FluencyBank (> 4,000 clips). After filtering out low-quality or ambiguous clips, I assembled 5,762 fluent samples and 2,252 single-event stuttering clips (679 block, 716 prolongation, 254 sound repetition, 603 word repetition) for four binary classification tasks: fluent speech vs. each single stuttering event. These tasks were evaluated via 5-fold cross-validation. I systematically evaluated combinations of three acoustic feature types (MFCC, mel spectrogram, zero-crossing rate), four self-supervised feature types (two wav2vec2 models, two HuBERT models), three feature selection methods (mutual information, ANOVA, chi-square), and three classifiers (SVM, random forest, BiLSTM). The combination of HuBERT embeddings and BiLSTM achieved the highest performance, AUROC of 0.81 for blocks, 0.92 for prolongations, 0.91 for sound repetitions, and 0.90 for word repetitions, compared to traditional feature-based machine-learning models, which all AUROCs <0.68.

This work highlights the potential of AI to assist speech therapists and empower individuals who stutter. My hope is that this project not only contributes to technological advancement but also helps reduce stigma and provides practical support for individuals, like my father, to communicate with greater confidence.

Title: Identifying Morin Hydrate as an Anti-Aging Drug with Machine Learning

Author list: Joanna Hou1

Detailed Affiliations:

1Princeton High School, NJ, USA

Abstract: Senescent cells accumulate naturally throughout the aging process as products of a tumor-suppressive mechanism that permanently inhibits their division. Though these cells are unable to divide, they become resistant to apoptosis and release Senescence-Associated Secretory Phenotype (SASP) factors that inflame neighboring cells and encourage tumor formation. Senescent cells are also closely associated with age-related health issues such as Alzheimer's Disease, Parkinson's Disease, and osteoarthritis as cells throughout the body may senesce. Fortunately, senescent cells can be selectively

eradicated by senolytics, a novel approach to anti-aging. However, our collection of confirmed senolytics is currently limited as little is known about the molecular targets of senolytics.
In this study, we develop a random forest machine learning model to analyze chemical features of published drugs and to predict their senolytic behavior. After training on data compiled by Smer-Barreto et al., the model suggests with 82.5% likelihood that morin hydrate, confirmed to be anti-oxidative, anti-inflammatory, and anti-cancer, is a senolytic. As oxidative stress is known to induce senescence, the effects of morin were then observed on Drosophila melanogaster fed food incorporated with hydrogen peroxide and were compared to the effects of a known senolytic, quercetin. Preliminary trials indicate that the lifespans and health of oxidative stress-induced flies were significantly extended and improved with the treatment of morin, and similar results were observed with the flies that were fed quercetin. Future additional trials will be conducted as well as a thiol assay test to obtain more concrete data on the senolytic effects of morin. These initial results as well as the prediction result of the machine learning model suggests that morin may be a senolytic.

Title: Deep learning-based fall detection for autonomous real-time emergency notification: integrating YOLO and Twilio
Author list: Daniel Zhou[1], Angela Zhang[2], Jerry Guo[3], David Guo[4,5]
Detailed Affiliations:
[1]William Lyon Mackenzie Collegiate Institute, Toronto, ON Canada, [2]Basis San Antonio - Shavano Campus, San Antonio, TX, USA, [3]Texas A&M University, College of Engineering, College Station, TX USA, [4]Department of Information Systems and Cyber Security, University of Texas at San Antonio, San Antonio, TX, USA, [5]United Services Automobile Association, San Antonio, TX USA.
Abstract: Falls poses a significant risk to elderly patients in geriatric care, especially when immediate assistance is unavailable, contributing to an estimated $50 billion in annual healthcare costs in the United States. This highlights the urgent need for real-time fall detection and response systems. This project presents a Deep Learning (DL)-based autonomous system designed to detect falls in real-time and promptly alter caregivers. Build on the open source computer vision model You-Only-Look-Once (YOLO) v8x-cls, which includes 5.6 million trainable parameters, the system accurately identifies fall events and triggers emergency notifications via Twilio, delivering both voice and text alerts to predesignated contacts. The model was trained on 458 annotated public images, 258 depicting falls (positive samples) and 193 as non-falls (negative samples), and evaluated on a separate test set of 106 images, including 70 with falls and 36 without. All images were gathered via Google image search and preprocessed through clipping and rotation. The fine-tuned model achieved a top-1 accuracy of 95.12% and top-5 accuracy of 100% in fall detection. The system was implemented on a Raspberry Pi 4 connected to a Raspberry Pi Camera Module 3, with Twilio's Programmable Voice API managing real-time emergency calls. Unlike traditional surveillance systems that provide only post-incident footage review, this system enables proactive and real-time intervention, ensuring safety and reducing response times in geriatric care environments.

Title: Bulk and Spatial Single Cell Transcriptomic Analysis and Machine Learning based Disease Classification of ALS
Author list: Aayush Veerabhadran
Detailed Affiliations:

Ann Arbor Pioneer High School, 601 W Stadium Blvd, Ann Arbor, MI 48103
Abstract: Introduction: ALS is a neurodegenerative disease that is characterized by the rapid degeneration of neurons in the brain and spinal cord. Patients experience symptoms of progressive muscular atrophy, increased fatigue, and respiratory problems, with it most commonly resulting in death by respiratory failure. Accurate identification of biological mechanisms and disease classification can significantly aid translational research in ALS.
Methods: From NCBI's Gene Expression Omnibus, the dataset GSE112680 (Swindell et al. [3]) comprises transcriptomic data derived from the blood cells of a cohort of 376 patients consisting of 164 samples from ALS patients, 137 healthy control patients, and 75 ALS mimic disease
(MIM) patients. Using the GEO2R analysis tool, I identified the top 250 differentially expressed genes with adjusted p-values < 2.85 x 10^-6 (t-test; controlling for false discovery rate) and inputted those genes into STRING-db to find enriched pathways. Further evaluation of the spatiotemporal dynamics in molecular pathology in ALS [4] provided a spatially resolved view of gene expression changes in mouse models and postmortem human tissues to understand the spatial molecular mechanisms that drive motor neuron degeneration. The findings from spatial transcriptomics data matched my pathway findings of related mechanisms in the integrated stress response and microglial signaling via the TREM2-TYROBP axis. Applying a machine learning-based LASSO regression model to the GEO data, I developed a predictive model to classify samples as either ALS or control, which was subsequently used to predict the "ALS State" of MIM patients and assess the predictive ability of differentially expressed genes.
Results: The analysis in STRING-db identified the Response of EIF2AK4 (GCN2) to amino acid deficiency pathway (count in network 9 of 100, strength = 0.92, FDR = 0.0016), and was corroborated with the dysregulation of protein processing in the endoplasmic reticulum mechanism from the spatiotemporal study. The significantly expressed genes from both analyses included DDIT3, CX3CR1, and TREML2, with both CX3CR1 and TREML2 playing a part in microglial phagocytosis of apoptotic neurons, which was a recurring mechanism of ALS in the spatial gene expression analysis from the study. Results from my ML-based predictive model, which was trained on ALS vs Control samples and outputted a predicted probability of ALS ranging from 0-1 (thresholded at 0.5), showed that 73% of the MIM samples were classified as ALS and the 27% were classified as healthy, showing ALS has a heterogeneous spectrum of disease progression.

Title: Evaluating the prognostic value of mutational signatures in small-cell lung cancer through data-driven threshold optimization and signature assignment
Author list: Rishabh Garg1, Kira A. Glasmacher1,2, Arnaud Augert3,4, Jeffrey P. Townsend1,4
Detailed Affiliations:
1Department of Biostatistics, Yale School of Public Health, New Haven, CT
2Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT
3Department of Pathology, Yale School of Medicine, Yale University, New Haven, CT 06510, USA;
4Yale Cancer Center, New Haven, CT 06520, USA;
Abstract : Background: Small-cell lung cancer (SCLC) is an aggressive malignancy with poor survival outcomes. Identification of robust prognostic biomarkers is a critical unmet need. Using de novo non-negative matrix factorization signature extraction and fixed binary signature contribution thresholds, previous studies have reported associations between single-base substitution (SBS) signatures—particularly SBS13 and SBS4—and patient prognosis. In particular, they have linked the presence of SBS4 signatures to lower tumor mutational burden (TMB) and worse overall survival, and linked SBS13 signatures to higher

TMB and better overall survival. However, these approaches have relied on arbitrary cutoffs and have encountered challenges in terms of reproducibility. Meticulous data curation, a rigorous clinically operable reference-based computational signature analysis, and appropriate threshold optimization are necessary to evaluate the utility of these signature attributions for prognosis.
Methods: We analyzed 45,882 somatic variants from 152 SCLC samples, of which 101 with survival data. Our analysis pipeline integrated both de novo and reference-based mutational signature assignment, evaluated signature activities as both continuous and as binary variables, and employed computational optimization of signature thresholds. We then examined associations between SBS signatures, TMB, and overall survival.
Results: Our analysis revealed marked discrepancies with prior findings from less rigorous analyses. Specifically, SBS4 activity was positively associated with TMB, a finding that is consistent with the underlying mutagenic and carcinogenic effects of tobacco smoke, but contradicting prior analysis that had indicated a negative association. SBS13 activity showed no consistent relationship with TMB. Furthermore, neither SBS4 nor SBS13 predicted overall survival across diverse rigorously implemented analytical methodologies. Notably, optimized activity thresholds provided statistically robust stratifications, outperforming previously used fixed cutoffs.
Conclusions: Our results re-evaluate the prognostic value of mutational signatures in SCLC using reproducible, data-driven methods. First, we demonstrate that reference-based signature assignment methods offer computational efficiency, stability, and power, which are essential features for clinical translation. Second, we show that statistically optimized binary thresholds provide a more robust analytical framework than arbitrary fixed cutoffs, ensuring statistical power and reproducibility. These methodological advances support more reliable translation of mutational signature biomarkers into clinical tools for outcome prediction in SCLC.

Title: LLM-Powered Web Agents and their Impact on Automation
Author list: Jasmine Zhang1
Detailed Affiliations:
1Carmel High School, Carmel, IN, USA
Abstract: Introduction: As the digital sphere expands, corporations are increasingly relying on complex web-based platforms with continuously changing user interfaces. Traditional automation using web agents has become a foundational tool for working with modern digital systems. It allows repetitive and time-consuming tasks to be performed faster and more consistently and is integral to testing and ensuring code reliability, data processing, and infrastructure management.
Gap: However, the complexity of these new systems has proven difficult for traditional automation to interact with these platforms. Pages with pop-ups, authentication flows, or hidden elements may cause automation to break down by attempting actions before elements become interactable. The emergence of Large Language Models (LLMs), Artificial Intelligence (AI) models trained on large amounts of textual data, provides solvency. They utilize machine learning techniques to understand and generate human language, allowing them to perform tasks. While traditional automation may not keep up with the complexity of new digital systems, LLM-powered web agents change the automation landscape. Whereas traditional automation relies on hard-coded instructions, LLMs allow agents to behave more like human assistants, reducing the need for manual script maintenance and allowing for resilience to complex and multi-step workflows.

Analysis: BabyAGI is a web agent powered by one of OpenAI's LLMs: GPT-4. The success of this LLM-powered model was compared to that of traditional web agents in two ways: automating browser actions through a Selenium-based script and web scraping using BeautifulSoup. This comparative analysis found that BabyAGI had higher accuracy and consistency than the traditional web agents. BabyAGI could autonomously perform tasks using instructions that were purely natural language, requiring no code. It could also produce its own tasks and execute them based on what the web pages contained. In fields with hidden elements, authentication flows, and drop-down menus, tasks were still executed correctly, depicting resilience to complex environments.
Conclusion: LLM-powered web agents are more adaptable and efficient in modern web environments than their traditional counterparts. As these digital systems continue to evolve and become more complex, these intelligent web agents will play a crucial role in redefining automation.

Title: AI-Enhanced Aptamer Design: Addressing the Blood-Brain Barrier Challenge of Alzheimer's Disease Therapeutics
Author list: Christina He[1], Jake Y. Chen[2]
Detailed Affiliations:
[1]Alphamind Club, Carnegie Vanguard High School; [2]University of Alabama at Birmingham, Birmingham, AL, USA
Abstract: Alzheimer's disease (AD) is a neurodegenerative disorder that is the most common cause of dementia that is marked by amyloid beta aggregation and cognitive decline. It leads to gradual nerve cell degeneration and impaired brain communication. Existing antibody-based therapeutics face challenges including large molecular size, limited tissue penetration, and lack of natural transport mechanisms for crossing the blood-brain barrier, hindering their effectiveness in targeting neurological disorders and efficacy in central nervous system. Antibody designs achieve 0.6 nM brain concentrations compared to 20-40 nM with direct CNS injection. Aptamers, short, structured oligonucleotides, offer a promising alternative as BBB shuttles when targeted to transferrin receptor (TfR). We also want to demonstrate proof-of-concept for a LYTAC construct utilizing TfR-targeting aptamers as the BBB penetration module for future amyloid-β therapeutic delivery.
We implemented an AI-driven pipeline to generate and screen candidate TfR-binding aptamer sequences. An initial pool of 27 AI-designed and benchmark sequences (35–72 nt, 44.4–73.3% GC) underwent secondary-structure filtering (minimum free energy ≤ –18 kcal/mol) and ViennaRNA folding analysis. Selected candidates were 3D-modeled via RNACOMPOSER and docked against the human TfR structure (PDB: 3KAS) using AutoDock Vina. Binding energies and receptor-contact profiles were computed to evaluate BBB penetration potential.
Docking simulations identified 18 out of 27 candidates with binding energies stronger than –9.0 kcal/mol, exceeding the threshold for effective TfR engagement. Top performers included tJBA8.26_candidate_5 and XQ-2d_candidate_3 (both –12.0 kcal/mol). The overall mean binding energy across all candidates was –9.61 kcal/mol (SD = 0.88), with multiple AI-generated aptamers outperforming DNA and RNA benchmarks. Sequence ranking by binding affinity guided selection of five lead candidates for future in vitro BBB assays. Our AI-enhanced computational strategy successfully produced novel TfR-targeting aptamers with superior predicted BBB penetration compared to established benchmarks. These candidates provide a proof-of-concept for aptamer-LYTAC constructs in AD therapeutic delivery and warrant experimental validation.

Title: Application of a Quantitative Systems Pharmacology Model to Predict Amyloid Plaque Reduction in Alzheimer's Disease Therapies
Author list: Alex Mi1, Jingwen Yan2
Detailed Affiliations:
1 Carmel High School, Carmel, IN, USA
2 Department of Biohealth Informatics, Indiana University Indianapolis, Indianapolis, IN, USA
Abstract: Introduction: A quantitative systems pharmacology (QSP) model was previously developed to characterize β-amyloid (Aβ) biology and the mechanism of action of aducanumab, an Aβ-targeting antibody for the treatment of Alzheimer's disease (AD). Donanemab, a recently approved Aβ-targeting antibody, has a slightly different binding mechanism than aducanumab. The objective of the current work is to update the Aβ QSP model and apply it to explore key dosing considerations for donanemab, including treatment cessation upon plaque clearance and optimization of dosing frequency.
Methods: The previous model was first updated to incorporate the pharmacokinetics (PK) and binding affinity of donanemab. It was then calibrated against PK and Aβ plaque data from a Phase 1 donanemab study to account for differences in binding mechanisms between donanemab and aducanumab. The calibrated model was further validated using data from a Phase 2 donanemab study. Subsequently, the validated model was used to simulate plaque reduction following various dosing regimens and plaque re-accumulation after treatment cessation, to confirm the optimal dosing strategy for donanemab.
Results: In addition to PK and binding affinity, the parameter of antibody-dependent cellular phagocytosis was updated during model calibration. The predicted Aβ plaque reduction from the calibrated model was consistent with observed plaque changes in the Phase 2 clinical study, in which donanemab was administered intravenously at 1400 mg every 4 weeks for up to 76 weeks. Model simulations showed similar plaque reduction effects with dosing intervals of every 8, 12, and 24 weeks, provided the total dose was equivalent to 1400 mg every 4 weeks. Simulations also indicated slow plaque re-accumulation following treatment cessation upon plaque clearance, suggesting that continuous dosing may not be necessary after plaque removal.
Conclusion: The Aβ QSP model was successfully adapted to predict the plaque-reducing effects of donanemab, further supporting its predictive capabilities. This model holds promise for guiding the design of future clinical trials in AD by enabling a priori predictions of plaque reduction outcomes.

Title: Illuminating Dark Proteins with AI
Author list: Andy Dong
Detailed Affiliations:
Illinois Junior STEM Society, Illinois
Abstract: Background: "Dark proteins" are uncharacterized proteins in the human genome with unknown functions. Some are suspected to be involved in diseases like cancer, making it critical to study them for potential biomedical applications.
Research Question: Can AI-based models provide deeper functional insights into dark proteins than conventional bioinformatics tools like BLAST.
Methods: We collected 7,264 dark protein sequences from PeptideAtlas and 572,970 known protein sequences from SWISS-PROT. Using BLAST, we searched for functional similarity through sequence alignment. In parallel, we developed a Python-based pipeline to analyze the same dark proteins using ProtT5, an AI-driven protein language model. ProtT5 generates 1024-dimensional embeddings that capture

structural and functional features. We calculated Euclidean distances between dark and known protein embeddings to predict functional similarity and compared results from both methods.
Results: ProtT5 identified 336 potential functional matches, which is over 10 times more than BLAST. Of the BLAST matches, 90% overlapped with ProtT5's results, confirming its reliability. Notably, several dark proteins identified by ProtT5 were linked to cancer-related genes.
Conclusions: AI-based models like ProtT5 can provide more comprehensive and accurate functional predictions than traditional tools. This approach offers a promising new path for uncovering the roles of previously uncharacterized proteins inhuman health and disease.

Title: Optimizing Torch-MISA for Efficient Signal Separation in IVA of fMRI via Definitive Screening Design

Author list: Orit Yohannes1, Fareya Borhan1, Xinhui Li3, Poorav Rawat1, Aditya Sule1, Calvin McCurdy2, Noah Lewis4, Bradley T. Baker4, Vince D. Calhoun3,4, Rogers F. Silva4
Detailed Affiliations:
1Department of Computer Science, Georgia State University, Atlanta, GA, USA; 2Department of Mathematics and Statistics, Georgia State University, Atlanta, GA, USA;3Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA; 4Tri-Institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Georgia State University, Georgia Institute of Technology, Emory University, Atlanta, GA, USA

Abstract: The independent vector analysis (IVA) framework, implemented in Torch-MISA is key for blind source separation in multi-subject functional neuroimaging. Torch-MISA's performance depends on hyperparameters like learning rate and batch size. However, their interaction in large IVA problems, where computational costs scale quadratically with subject count, is poorly understood. Optimizing these settings could improve both accuracy and efficiency. This study identifies optimal hyperparameters for large IV A tasks, enhancing source separation accuracy and reducing convergence rates to lower computational costs. We use a Definitive Screening Design (DSD) to evaluate hyperparameter effects with minimal experiments. A quadratic model is then built to capture interactions and nonlinearities, predicting settings for optimal accuracy and convergence. Confirmatory experiments validate the model's predictions based on the multidataset intersymbol interference (MISI) measure, showing significant gains in computational efficiency while retaining signal separation accuracy.
We simulated brain imaging data for 100 subjects using different random seeds (7, 14, and 21) to evaluate Torch-MISA's ability to extract independent components while maintaining source alignment across subjects. We studied eight hyperparameters affecting convergence and accuracy, including the learning rate, batch size, and the beta parameters of the Adam algorithm, using the DSD to efficiently capture main effects and interactions with fewer experiments than traditional designs. Performance was measured using Epochs (convergence speed), MISI (source separation quality), and MxE (the product of Epochs and MISI, indicating efficiency). A linear regression model identified optimal hyperparameter combinations, revealing new candidate settings for improved accuracy and convergence efficiency.
Our initial screening analysis showed that only learning rate, batch size, beta 1, and beta 2 significantly affected the performance metrics. By focusing on these key hyperparameters, we uncovered critical relationships impacting model convergence, signal separation quality, and training efficiency. Leveraging these insights, we were able to optimize Torch-MISA's performance in large IVA problems, improving efficiency and accuracy in a simulated multi-subject dataset. Using Definitive Screening Design (DSD), we reduced experiments and computational costs while identifying key hyperparameter relationships. Ongoing

work will extend this approach to larger datasets, validating scalability and robustness with independent test sets.

Title: Distinct DNA Methylation Patterns in Alzheimer's Disease Brain Tissue

Author list: Raymond Cheng1,2,3 *, Jingmin Shu1,2 *, Hai Chen1,2, Runxin Su4, Li Liu1,2

Detailed Affiliations:

1. College of Health Solutions, Arizona State University, Phoenix, AZ 85004, USA.
2. Center for Personalized Diagnostics, Biodesign Institute, Arizona State University, Tempe, AZ 85281, USA.
3. Case Western Reserve University, Cleveland, OH 44106, USA
4. Department of Psychology, Seattle, WA 98105, USA

Abstract: Alzheimer's disease (AD) is a progressive, incurable brain disorder causing memory loss, cognitive decline, and dependence. It affects over 6.9 million Americans and 33–38 million people globally, with numbers expected to double by 2050. Current treatments offer limited relief, and early diagnosis remains a challenge. Abnormal DNA methylation in AD affects genes linked to neuronal survival, inflammation, synaptic function, amyloid, and tau—often emerging before symptoms, making it a promising target for early detection and therapy. However, inconsistent findings across studies may result from unrecognized molecular subgroups, which mask true disease-associated epigenetic changes.

We analyzed dataset GSE59685 (55 AD samples, 24 controls) using a linear regression model (Methylation ~ Diagnosis + Age + Sex). Traditional statistical methods were limited: among 100,000 variable probes, only 94 were significant (adjusted $p < 0.05$), with minimal biological relevance, reflecting the subtle nature of methylation changes in AD. To address this, we applied the CIMP (CpG island methylator phenotype) concept from cancer, assuming AD subgroups. Unsupervised clustering revealed three distinct methylation clusters (hypo-, hypermethylation, transitional) across 9,024 probes. Similar patterns in controls suggested influence from baseline variation. Since bulk brain tissue was used, we estimated that cell-type composition explained 6.8% of methylation variance, while cluster membership explained 12.6%. A regression model with main effects and interactions identified 703 differentially methylated probes (adj. $p < 0.1$). Clustering improved diagnostic accuracy from ~50% to ~70% across three external datasets. Although overlap reproducibility was unchanged, clustering identified 369 new biomarkers than without clustering. Pathway enrichment analysis revealed both shared and cluster-specific AD-related pathways. For example, cluster 1 was enriched in developmental programs, synaptic plasticity, and epigenetic regulation, while cluster 2 showed enrichment in metabolic, neurotransmission, histone modification pathways, etc. These findings suggest that methylation-based subgroups capture underlying biological heterogeneity relevant to AD mechanisms and biomarker discovery.

To evaluate potential genetic and clinical influences on clustering, we analyzed a fifth dataset (GSE212682). The protective APOE ε2 allele was enriched in Cluster 3, which also exhibited differences in AD pathology severity based on the Alzheimer's Disease Severity Score (adseverscore, per NIA-AA criteria). In conclusion, DNA methylation–based subgroups in AD reflect underlying heterogeneity shaped by cell composition, epigenetic regulation, genetic interactions, and disease stage. These subgroups provide a valuable framework for enhancing biomarker discovery and guiding targeted therapeutic strategies in AD.

Flash Talk Session

August 5th
1:30 PM – 4:50 PM
Room 350

Chairs: Zhifu Sun

Title: A Multimodal Vision Transformer Using Fundus and OCT Images for Interpretable Classifications of Diabetic Retinopathy
Author list: Shivum Telang and Wei Chen
Abstract: Diabetic Retinopathy (DR) is a leading cause of vision loss worldwide, requiring early detection to preserve sight. Limited access to physicians often leaves DR undiagnosed. To address this, AI models leverage lesion segmentation for interpretability, but manually annotating lesions is impractical for clinical use. Physicians also require models that explain why a classification was made, not just where lesions are located. Current models often rely on a single imaging modality and achieve limited effectiveness. This study introduces RetGEN, a self-supervised learning-based framework that enhances DR classification through a multimodal vision-transformer architecture with a multimodal contrastive loss function. By integrating OCT and fundus scans, RetGEN improves classification accuracy while providing explainable insights for ophthalmologists. For interpretability, the model generates paired Grad-CAM heatmaps showcasing neuron weights across OCT images, visually highlighting regions contributing to DR severity classification. Trained on 3,000 fundus images, 1,000 OCT images, and 125 paired images, RetGEN outperforms state-of-the-art models, delivering more accurate, interpretable, and clinically meaningful assessments of DR severity. This methodology addresses key limitations in current DR diagnostics, offering a practical and comprehensive tool for improving patient outcomes.
Keywords: Vision Transformer; Contrastive Loss Function; Convolutional Neural Networks; Global Average Pooling; Weight Matrices

Title: Abnormal ERV expression and its clinical relevance in colon cancer
Author list: Aditya Vijay Bhagwate, Jason Ding, William Taylor, John Kisiel and Zhifu Sun
Abstract: Background: Human endogenous retroviruses (HERVs or ERVs) are genomic sequences that have integrated into the human genome from ancestral exogenous retroviruses and account for nearly 9% of human DNA. Many ERVs are expressed during embryogenesis but are epigenetically silenced afterward. However, growing evidence suggests that the reactivation of certain ERVs may be associated with human disease development, progression, and patient outcomes such as cancers and autoimmune diseases. Most studies selected a subset of ERVs and comprehensive profiling of well-annotated ERVs in colon cancer is lacking. This study aims to perform comprehensive profiling of ERVs and their associations with clinical phenotypes of colon cancer.
Methods: Cell line RNA-sequencing data, seven from colorectal cancer (CRC) and one from monocytes with both total RNA and polyA library preparations, and one from normal colon epithelium from PolyA protocol, were downloaded from RNA Atlas (GSE138734). RNA sequencing data for colon adenocarcinomas (COAD) and adjacent normal tissues were downloaded from GDC TCGA (https://portal.gdc.cancer.gov/). After alignment, ERV expression was quantified against comprehensively compiled ERVs (3,320). ERV expression profiles were compared between sequencing protocols, cancer

and normal cells, and matched tumor and normal tissue pairs. ERV enrichment was performed with their overlapping or closest protein coding genes. Unsupervised clustering was used to identify ERV expression defined tumor subtypes and their associations with clinical and other molecular features. ERV association with disease specific survival (DSS) was performed using Cox regression model or Kaplan-Meier curves.
Results: ERV expressions between PolyA and total RNA protocols were comparable where both showed a higher number of expressed ERVs in cancer cells. The increased ERV expression was even more dramatic in primary COAD tumor samples. The "reactivated" or highly expressed ERVs in COAD were mainly located in intergenic region or intronic region of protein coding genes or lncRNAs. Host or nearby genes of these up-expressed ERVs were significantly enriched in viral protein interactions with cytokine and cytokine receptors while the down expressed genes were enriched in vitamin and ascorbate metabolism. ERV expression defined tumor classes were significantly associated with tumor mutation burden (TMB) and immuno-phenotypes such as antigen processing and presenting machinery (APM) and tumor immune infiltration score (TIS). Survival analysis identified 152 ERVs to be independently associated with DDS and 51 of them were also differentially expressed between tumors and normal samples.
Conclusions: ERV abnormal up expression is common in CRC. The ERV defined subtypes are associated with tumor immunity and some individual ERVs are independently associated with patient outcomes. These findings provide further evidence abnormal ERV expression has clinical and treatment implications.
Keywords: Human endogenous retrovirus; ERV; colon cancer; tumor immune response; patient survival

Title: From Bench to Insight: Rapid Pathogen Genomic Surveillance Workflow for SARS-CoV-2 and Emerging Pathogens
Author list: Chelsea Zimmer, Selena McVay, Keely Starke, Kimily Hughley, Sara N Koenig and Venkat Sundar Gadepalli
Abstract: Clinical surveillance of infectious diseases caused by viruses, such as SARS-CoV-2, is important for effective intervention and preventing potential epidemics or pandemics. The frequent mutations of the SARS-CoV-2 genome, caused by its RNA nature and lack of proofreading mechanisms, allow it to adapt to its host organisms. This adaptation can lead to new strains or variants of the virus. Historically, next-generation sequencing techniques required complex chemistry and specialized training of laboratory technicians and other specialized personnel. However, with improvements in automation and nanotechnology, these inherently specialized methodologies have been simplified and are easily adapted by the novice user in a clinical molecular lab. Parallel improvements in sequencing technologies and decreased costs associated with whole genome sequencing resulted in a worldwide effort of sequencing viral genomes from patients identified for SARS-CoV-2 infection. Large-scale analysis on these sequence data is now feasible with new bioinformatics pipelines and various reporting tools. Such pipelines allow a clinical laboratory to perform surveillance sequencing workflows without requiring advanced technical expertise, creating endless prospects. The array of sequence data generated across the globe offers diverse opportunities to study SARS-CoV-2 evolutionary dynamics and serves as a foundation for different research questions in the future. To enhance data accessibility for various research and global surveillance projects, public data repositories have been developed. These publicly accessible repositories host diverse data from different countries, thereby assisting in determining regional variants or identifying emerging variants. Even though bioinformatics tools are rapidly developed for identifying mutations and variant reporting, they require some computational expertise. We have developed a COVID-19 mutational analysis pipeline using Workflow Description Language (WDL), which is open-source and combines various steps in an analysis workflow with human-readable syntax. Thus, users with minimal informatics background can easily adapt

the workflow while creating a local data repository within their institution. The pipeline processes input FASTA files and quality control files from Ion Torrent S5, performs clade and variant assignment, integrates patient metadata, and stores the results into a REDCap database. Further, in the pursuit of tracking sample records and relevant metrics during a sequencing run, our team has innovated a REDCap-based data capture system. This user-friendly REDCap form records essential details of each sequencing run and stores it in a REDCap database along with patient demographics info. To further enhance the utility of our REDCap-based data capture system, we have developed an intuitive interactive dashboard. This interface seamlessly connects with the REDCap data sources, providing real-time monitoring, interactive visualization, and the ability to create a consolidated variant report. Our overall approach streamlines processes in managing complex genomic data and offers easy adaptation to empower other molecular labs.
Keywords: WDL Workflow; Clade; Lineage; S5 Ion torrent suite

Title: LoRA-BERT: a Natural Language Processing Model for Robust and Accurate Prediction of long non-coding RNAs
Author list: Nicholas Jeon, Paul de Figueiredo, Lamin Saidykhan, Xiaoning Qian and Byung-Jun Yoon
Abstract: Long non-coding RNAs (lncRNAs) serve as crucial regulators in numerous biological processes. Although they share some sequence features similar to messenger RNAs (mRNAs), lncRNAs perform entirely different roles, providing important new avenues for biological research. The emergence of next generation sequencing technologies has greatly advanced the detection and identification of lncRNA transcripts and deep learning-based approaches have been introduced to classify lncRNAs. Although these methods have significantly improved the efficiency of identifying lncRNAs, they often lack robustness, and the prediction accuracy tends to vary significantly depending on the quality of the transcript. To tackle this issue, we introduce LoRA-BERT, a novel lncRNA prediction algorithm built on the bidirectional encoder representation from transformers (BERT). Lora-BERT is designed to effectively capture information at the nucleotide level that is important for lncRNA classification, leading to more robust and accurate prediction outcomes that are not significantly affected by the quality of the input transcript. Through performance evaluations on comprehensive benchmarks, we demonstrate that LoRA-BERT outperforms existing schemes in terms of accuracy, efficiency, and robustness. Especially, unlike other methods, LoRA-BERT retains good predictive capability for partial transcripts, a critical feature that makes it applicable for reliable lncRNA prediction even when the sequencing depth is relatively low.
Keywords: Natural language processing; Bidirectional Encoder Representations from Transformers (BERT); long non-coding RNA (lncRNA); lncRNA prediction

Title: ICM-MD: Integrating TM-Specific Features and MD-Derived Structures for Accurate Prediction of Inter-Chain Contacts in Alpha-Helical Transmembrane Homodimers
Author list: Bander Almalki and Li Liao
Abstract: Characterizing the interactions of alpha-helical transmembrane homodimers at the residue level is crucial for understanding their structure and function. However, most computational tools designed for globular proteins fail to translate to transmembrane (TM) proteins, largely due to the unique environment of the membrane and the limited availability of high-resolution structural data. To address this challenge, we present a machine learning framework geared for TM homodimers. Our method integrates sequence-based and structure-based features to enhance inter-chain residue contact prediction in TM homodimers.We address the challenge of limited training data by utilizing structures derived from molecular dynamics (MD)

simulations as surrogate ground truth. Our model leverages a simple yet effective feed-forward neural network, designed to enhance model's interpretability and scalability. Comparativeevaluation against state-of-the-art models, including DeepHomo1, DeepHomo2, Glinter, and DeepTMP, demonstrates that our method achieves superior performance. On a test set of eight alpha-helical TM homodimers, our model outperforms DeepHomo1 and Deep-Homo2 by 155.5% and 261.0% respectively, surpasses Glinter by 92.0%, and achieves 11.6% higher precision compared to DeepTMP in the mean top L ranking metric.
Keywords: Transmembrane Proteins; Alpha-Helical homoidmers; Dimerization; Inter-Chain contact; Machine Learning

Title: OmicsSankey: Crossing Reduction of Sankey Diagram on Omics Data
Author list: Shiying Li, Bowen Tan, Si Ouyang, Zhao Ling, Miaozhe Huo, Tongfei Shen, Jingwan Wang and Xikang Feng
Abstract: In bioinformatics, Sankey diagrams have been widely used to elucidate complex biological insights by visualizing gene expression patterns, microbial community dynamics, and cellular interactions. However, computational scalability remains a challenge for large-scale biological networks. In this work, we present OmicsSankey, a novel formulation of the layout optimization problem for Sankey Diagrams that employs eigen decomposition as a closed-form solution, addressing graph disconnection through a teleportation mechanism that enhances connectivity and stabilizes eigenvector solutions. Experimental results on synthetic datasets with varying layers and nodes validate the efficacy of OmicsSankey compared to state-of-the-art layout-optimizers. Improving the Sankey layouts for Cell Layers, BioSankey, and Sequence Flow further validates OmicsSankey in enhancing the interpretability of biological insights.
Keywords: Sankey diagram; Layout optimization; Omics data visualization; Microbiome visualization

Title: TCR Convergence as a Proxy for Tumor-Specific Immunity in HSV1-Positive rGBM Patients Treated with CAN-3110
Author list: Ayse Selen Yilmaz[1,2], Alexander Ling[4], Hiroshi Nakashima[4], Xiaokui Mo[1,3], E. Antonio Chiocca[4]
Detailed Affiliations:
[1] Department of Biomedical Informatics, College of Medicine, The Ohio State University Wexner Medical Center, Columbus, OH, USA [2] Bioinformatics Shared Resources, James Comprehensive Cancer Center, The Ohio State University; Columbus, OH, USA [3] Center for Biostatistics, College of Medicine, The Ohio State University Wexner Medical Center, Columbus, OH, USA [4] Harvey Cushing Neuro-oncology Laboratories, Department of Neurosurgery, Brigham and Women's Hospital, Boston, MA, USA
Abstract: Despite limited understanding on the mechanisms and predictors of outcome for oncolytic viral therapies, our previous work demonstrates that HSV1-positive rGBM patients exhibit prolonged survival following CAN-3110 viral therapy [1]. We hypothesize that this survival benefit is T cell mediated. T cells play an important role in adaptive immunity by utilizing the T cell receptors (TCRs) on their surfaces to recognize a wide range of antigens. Antigen specific TCRs are essential for eliminating tumor cells, but they are difficult to identify. A possible proxy for measuring tumor-specific TCRs is "TCR convergence", the phenomenon where TCRs have identical CDR3 amino acid sequences but different DNA sequences. TCR convergence has been shown to be a novel prognostic marker for immunotherapy [2]. We investigated relationships between TCR convergence, HSV1 serology, and survival in 21 IDH wild-type rGBM patients who received CAN-3110 treatment. DNA from pre- and post-treatment PBMCs was sequenced to extract

TCRβ sequences. Patients were stratified into high and low TCR convergence groups based on the convergent TCR count in each pre-treatment sample, using the mean as a cutoff, and classified as positive or negative based on their HSV1 serology before treatment. We found that HSV1 seropositive patients are more likely to exhibit higher TCR convergence (Fisher's test, p-value=0.0071). Although a strong correlation between convergent TCR count and overall survival (OS) was not identified, high TCR convergence group shows higher OS compared to the low TCR convergence group (393.1 vs. 277.5 days, t-test, p=0.08). Given the observed association between TCR convergence and HSV1 serostatus, we next examined the structural organization and antigen specificity of the T cell repertoire. Using the TCRosetta platform [3], we identified TCR clusters at the CDR3 amino acid level, constructed TCR interaction networks, and predicted high-confidence peptide targets from VDJdb in pre- and post-CAN-3110 samples. Notably, top TCR clusters before and after treatment shared conserved CDR3 motifs, suggesting their sustained role in mediating anti-tumor immune responses. Our findings with this limited number of patients may guide future investigations into the role of TCR convergence in immunotherapy and its potential as a prognostic marker.

Title: Vritra: a streamlined pipeline for species-resolved functional profiling of target genes in microbiome data

Author list: Boyan Zhou1, Menghan Liu2, Lama Nazzal3, Huilin Li1#

Detailed Affiliations:

[1]Division of Biostatistics, Department of Population Health, New York University School of Medicine, New York, NY, USA; [2]Department of Biological Sciences, Columbia University in the City of New York, New York, NY, USA; [3] Department of Medicine, New York University School of Medicine, New York, NY, USA.

Abstract: Microbiome functional profiling tools have advanced our understanding of microbial pathways and gene families. Increasingly, researchers are examining specific gene sets—for example, the frc/oxc genes involved in oxalate degradation or the bai clusters responsible for bile acid metabolism. Such targeted studies demand not only precise abundance quantification but also accurate species attribution. Two challenges remain: 1) ambiguous protein annotations hinder construction of comprehensive yet specific target gene sets, and 2) mainstream workflows align reads to UniRef90 clusters and then infer species in a manner that lacks the standardized boundaries (e.g., the ≥95 % ANI cutoff in GTDB) used to delineate species.

To overcome these limitations, we present Vritra (Versatile Reads-identification with Impartial Taxonomic Refinement and Assignment), a flexible pipeline for targeted gene detection and species-level profiling in shotgun microbiome sequencing. Vritra comprises two modules:

1. Gene-specific database construction. Users supply a single seed sequence plus a curated set of related sequences (e.g., from UniProt or InterPro). Vritra then employs a label-propagation algorithm to expand this into a refined UniRef90 reference tailored to the target gene family.
2. Species-resolved read analysis. Each UniRef100 cluster in the custom UniRef90 set is first assigned to a species using GTDB-style boundaries. Raw reads are then mapped to the representative sequence of each UniRef100 cluster to generate species-level gene abundance profiles.

Because Vritra builds only the gene-centric UniRef90 subset, its reference database is orders of magnitude smaller than the full UniProt UniRef90, greatly improving computational efficiency. We applied Vritra to three gene families across two publicly available microbiome cohorts, demonstrating accurate abundance

estimates, robust species assignment, and broad applicability to both metagenomic and metatranscriptomic data.
Keywords: Microbiome; metagenomics; targeted gene profiling; species-level resolution; functional analysis

Title: Transcriptomic signatures in nucleus accumbens, midbrain, pre-frontal cortex, and amygdala regions identifies shared and unique gene signatures for substance use
Author list: Avinash Veerappa, Chittibabu Guda
Detailed Affiliations:
Department of Genetics, Cell Biology and Anatomy, University of Nebraska Medical Center, Omaha, NE 68198, USA
Abstract: Background: Chronic substance use is a neuropsychiatric disorder involving persistent craving, pleasure, and reward, ultimately progressing to addiction. The midbrain controls hunger, reward, and pleasure traits, whereas the dorsolateral prefrontal cortex (DLPFC) controls craving, decision making, and tolerance. In contrast, nucleus accumbens (NAc) controls feeding, sexual, reward, stress-related, and drug self-administration behaviors, while amygdala regulates emotion and memory. Design: To understand these complex and dynamic events in the context of substance use disorders, we performed transcriptome analysis on the four brain regions (midbrain, DLPFC, NAc, and amygdala) combined with a multi-pronged strategy involving transcriptome clustering, bi-clustering, weighted correlation network analysis (WGCNA), and pathway enrichments. Findings: Upregulation of gene expression was dominant in all the four brain regions of the cases compared to controls. Distinct differential transcriptomic signatures were observed that were both unique to one region and shared across multiple regions leading us to identify 186 genes exclusive to the midbrain, 29 genes to DLPFC, 160 genes to NAc, and 442 genes for amygdala regions. Network analysis revealed that the DEGs across all brain regions were flanked and interconnected with neuropeptide-neurotransmitter axis suggesting the interference of substances with maintaining the equilibrium of neurotransmitters and neuropeptides. Significant upregulation of genes CSF3, GADD45B, SOCS3, and NPAS4 across all four brain regions was observed resulting in the enrichment of CREB Signaling in Neurons pathway leading us to postulate their involvement in long-lasting maladaptations of neurocircuitry due to chronic substance use. Conclusions: By unraveling the unique and shared transcriptomic signatures, our study advances the current understanding of the crosstalk among the key players in each brain region in substance use, thus implying that induction and exclusion signals drive distinctive pathway signaling and sustain addiction behavior. This study, while recognizing known genes in the field of substance biology and addiction, also identified several novel biomarkers that could confer susceptibility for addiction risk.

Title: Endophenotype-based in silico network medicine prediction and real-world patient data validation identify potential drug combinations for Alzheimer's disease
Author list: Zhendong Sha[1,2], Seungyeon Lee [3,4], Yadi Zhou [1,2], Yuan Hou 1,2, Ping Zhang [3,4], Pengyue Zhang[5], Feixiong Cheng [1,2,6*]
Detailed Affiliations
[1]Cleveland Clinic Genome Center, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44195, USA; [2]Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44195, USA; [3]Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210,

USA; [4]Department of Biomedical Informatics, The Ohio State University, Columbus, OH 43210, USA; [5]Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN 46202, USA; [6]Department of Molecular Medicine, Cleveland Clinic Lerner College of Medicine, Case Western Reserve University, Cleveland, OH 44195, USA
Abstract: Background: The multifactorial disease mechanism of Alzheimer's disease (AD) renders the conventional "one drug, one disease" paradigm of drug discovery insufficient to address its complexity. Drug combination therapy allows for the simultaneous targeting of multiple disease mechanisms, offering the potential to slow, or even reverse, AD progression. This study leveraged endophenotypes—pathobiological phenotypes underlying AD—as multifactorial disease modules to design drug combinations. The endophenotype targeting of drugs was evaluated based on the network proximity of their therapeutic targets to endophenotype modules on the protein-protein interactome (PPI).
Method: Eight AD endophenotypes covering the disease progression of AD were considered in this research. Endophenotype genes were curated from various biological knowledge sources, including Gene Ontology (GO), the Kyoto Encyclopedia of Genes and Genomes (KEGG), and relevant literature. These genes were refined for AD relevance using genome-wide association study (GWAS) summary statistics and brain tissue quantitative trait locus (xQTL) data. Drug
combinations were prioritized using a scoring framework that evaluated coverage and complementary exposure to the eight endophenotypes. High-confidence candidate drug combinations were further validated by two large, independent health insurance claims databases.
Result: We prioritized seven drug combinations by considering their drug combination scores, effectiveness in a real-world patient cohort, and their proximity to three key AD endophenotypes (Amyloid, Neuroinflammation, and Tau). Among them, we highlighted one combination demonstrating synergistic effectiveness and complementary endophenotype coverage. Carvedilol + Diclofenac collectively targeted Amyloid, Neuroinflammation, and Tau endophenotypes, with Carvedilol significantly targeting Neuroinflammation (network proximity z-score: -3.22) and Tau (z-score: -2.17) and Diclofenac targeting Amyloid (z-score: -3.23) and Neuroinflammation (z-score: -5.39), featuring a complementary exposure to Neuroinflammation. This combination also achieved a synergistic combination hazard ratio (HR) of 0.30 (individual HRs: Carvedilol 0.89, Diclofenac 0.73). In addition to HR analysis, we further evaluated the treatment effects of this combination using an independent patient cohort and a deep learning-based causal inference model to estimate average treatment effect (ATE), where ATE < 0 indicates improved outcomes compared to control. The ATE for Carvedilol + Diclofenac was -0.030 (95% CI: -0.052, -0.007), supporting the observed synergistic effects observed from combination HR and their complementary network proximity coverage to the three key AD endophenotype modules.
Conclusion: This study employed an in silico network medicine-based framework to rationally design drug combinations that target the complex pathobiology of AD. By integrating real-world patient data, this work provides robust evidence supporting the clinical potential of the predicted combination.
Keywords: Alzheimer's disease, Drug combination, Drug discovery, Endophenotype, Network medicine

Title: VaxLLM: An end-to-end framework leveraging a fine-tuned Large Language Model for automated vaccine annotation and database integration
Author list: Xingxian Li[1,2], Matthew Asato[1], YuPing Zheng[3], Joy Hu[1], Feng-Yu (Leo) Yeh[2], Zhigang Wang[4], Jie Zheng[2], Yongqun He[2]
Detailed Affiliations
[1]College of Literature, Science, and the Arts, University of Michigan, Ann Arbor, MI, USA; [2]Unit for Laboratory Animal Medicine, Center for Computational Medicine and Bioinformatics, Department of

Learning Health Science, University of Michigan Medical School, Ann Arbor, MI, USA; [3]Chinese University of Hong Kong, Shenzhen, Guangdong, China; [4]Department of Biomedical Engineering, Institute of Basic Medical Sciences Chinese Academy of Medical Sciences, School of Basic Medicine Peking Union Medical College, Beijing, China.
Abstract: Vaccines play a vital role in enhancing immune defense and preventing hosts against a wide range of diseases. However, vaccine annotation remains a labor-intensive task due to the ever-increasing volume of scientific literature. This study introduces the Vaccine Large Language Model (VaxLLM) framework to explore the application of Large Language Model (LLM) in automating the annotation of scientific literature and database integration on vaccines.

To develop VaxLLM, we first fine-tuned the Llama 3 model using the training data from VIOLIN vaccine knowledgebase and PubMed articles. The VIOLIN knowledgebase has so far included 4,708 vaccines for 217 pathogens or non-infectious diseases (e.g., cancer), which provides comprehensive vaccine information. The paper processing started with the automatic fetching of articles by literature mining from PubMed. The fine-tuned model was first used to classify the articles to filter the relevant articles containing specific information about vaccine development. If the article was classified as "yes", the fine-tuned Llama 3 model then annotated the article, specifically capturing key vaccine properties such as vaccine platform, antigen, formulation, target host species, experimental methods, protocols, immune response, efficiency, and experiment results. The PubTator tool was also used as an integrative component to extract the biomedical entities such as genes, diseases, chemicals, and species. To increase the accuracy of the model output, we also developed an annotation tool using the GPT API to extract more details from the full-text manuscript, such as detailed immune response and vaccine development stage, , and other related properties. For greater accuracy, a data harmonizer website was developed to help experts validate the results of these outputs in a clear manner. The final validated output could be directly exported into a database format and incorporated into the VIOLIN database.

Using keyword search, 143 PubMed articles about Brucella vaccines from the year 2024 to 2025 were used as testing data. The results of the VaxLLM were reviewed manually. The VaxLLM system achieved a classification precision of 0.92, recall of 1.0, AUROC of 0.88, and F1-score of 0.95. The annotation accuracy is 97.9%, outperforming the baseline Llama 3 model by a significant margin. Through rapid retrieval, the VaxLLM system can help the database increase its capacity at an extremely fast pace, such as adding thousands of vaccine information from literature in a year. Such a method may also be utilized for other domains of literature annotation.
Keywords: Vaccine, Large Language Model (LLM), Fine-tuning, Llama 3, PubTator, PubMed

Title: Supervised and unsupervised classification with feature selection for single-cell RNAseq based on an artificial immune system.

Author list: Dawid Krawczyk[1,4], Maciej Pietrzak[2], Michał Seweryn[3,4]

Detailed Affiliations:

[1]University of Lodz Doctoral School of Exact and Natural Sciences, University of Lodz, Poland;[2]Department of Biomedical Informatics, The Ohio State University, Columbus, OH, 43210, USA;[3]Centre for Digital Biology and Biomedical Sciences, Faculty of Biology and Environmental Protection, University of Lodz, Poland; [4]Regional Digital Medicine Center, Copernicus Memorial Hospital and University of Lodz, Poland

Abstract: In this study, we present a tool scAIS which enables both supervised and unsupervised classification of cells based on the expression profiles from single-cell RNA-seq experiments. Our approach

is an extension of the method proposed by Dudek in doi: 10.1109/TEVC.2011.2173580. The main novelty of scAIS is related to the feature selection part which is performed simultaneously with the main task of learning – either supervised or unsupervised classification. In principle, scAIS is based on two main steps: (1) selection of epitopes (combinations of features/genes) which best separates the points of interest in high dimensional subspaces and (2) estimation of local neighborhood of data points (or clusters) which defines the local structure in lower-dimensional subspaces. The main advantage of scAIS is the ability to perform feature selection and clustering without dimension reduction performed on the initial step of preprocessing which is known to bias the final outcomes of the clustering as well as differential expression analysis – please refer to Rafael's Irizarry's blogpost https://simplystatistics.org/posts/2024-12-23-biologists-stop-including-umap-plots-in-yourpapers. We compare scAIS to the recent as well as classical methods of machine learning used for classification of single-sell RNAseq data. To this aim we use a set of benchmarking datasets available through the github repository as well as R package provided by prof Martin Hemberg's lab. In the side-by-side comparisons, based on both the real world data as well as simulation studies, to the novel scMINER and the classical SC3 algorithms our scAIS achieves comparable sensitivity of cluster detection and at the same time retains higher specificity.

Keywords
single-cell RNA sequencing, clustering, feature selection, artificial immune system, machine learning

Title: Multi-modal Domain-specific Foundation Model for Prostate Cancer Explanation: Utilizing H&E Image and Spatial Proteomics

Author list: Kyeong Joo Jung[1†], Sourav S Rout[6†], Saksham Gupta[5], Sarikaa Sridhar[1], Dongjun Chung[2,3], Parag Mallick[4], Kshitij Jadhav[6], Raghu Machiraju[1,*]

Detailed Affiliations:
[1]Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, USA; [2]Department of Biomedical Informatics, The Ohio State University, Columbus, Ohio, USA; [3]Pelotonia Institute for Immuno-Oncology, The James Comprehensive Cancer Center, The Ohio State University, Columbus, Ohio, USA; [4]Canary Center for Cancer Early Detection, Department of Radiology, Stanford University School of Medicine, Stanford, CA, USA; [5]Center for Machine Intelligence and Data Science, IIT Bombay, India; [6]Koita Center for Digital Health, IIT Bombay, India.
[†]Authors contributed equally

Abstract: Prostate cancer remains a significant global health challenge, ranking as the second most common cancer and fifth leading cause of death among men [1]. Its inherent heterogeneity presents challenges to accurate diagnosis. Current diagnostic methods involve manual review of Hematoxylin and Eosin (H&E) stained images by pathologists, a process that is labor-intensive, time-consuming, and highly dependent on specialized expertise. While recent advancements in spatial proteomics, such as Cell Dive [2], offer a more comprehensive view of prostate tissue by detecting various cell types, these experiments are often costly and technically demanding. To address these limitations, our study proposes the development of a novel domain-specific foundation model. This model is trained using multi-modal data derived from spatial proteomics, integrating features extracted from H&E images and single-cell data. Our foundation model will perform solely on H&E images once trained, enabling the prediction of patch level description which includes diverse cell type distributions, and relevant clinical outcomes, and the segmentation of tissue maps highlighting tumor glands and immune regions. For model training, we leveraged 720 H&E slides from 200 prostate-cancer patients, combining their image features with labels/images generated from single-cell data and corresponding clinical outcomes. Feature extraction from H&E images is specifically

performed using Gigapath [3], a powerful foundation model that provides robust visual representations learned from extensive pathology image datasets. To segment the tissue maps, a classification head is employed, utilizing labeled images from single-cell data. A regression head predicts continuous outputs such as clinical outcomes and cell type proportions. The losses generated from these three distinct tasks—cell type distribution prediction, clinical outcome prediction, and tissue segmentation—are individually calculated and subsequently summed to optimize the overall model training process. After training, the model converts each tile from the H&E image into a vector that estimates the local mix of tumor, immune, stromal cells, and infiltration zones. It then overlays a tissue map on the slide that labels every cell type—including cancerous and immune-rich regions—and displays concise text summaries of cell composition and predicted clinical outcomes, all without extra staining, helping pathologists better understand the tissue microenvironment. Future work will (i) aggregate neighboring tiles to capture whole-slide tissues, and (ii) share a simple web viewer so users can overlay our predictions on their own H&E images.
Keywords: Spatial proteomics, Domain-specific foundation model, multi-modal data

Title: Static and Dynamic Cross-Network Functional Connectivity Shows Elevated Entropy in Schizophrenia Patients
Author list: Natalia Maksymchuk[1], Robyn L. Miller[1], Juan R. Bustillo[2], Judith M. Ford[3,4], Daniel H. Mathalon[3,4], Adrian Preda[5], Godfrey D. Pearlson[6,7], and Vince D. Calhoun[1]
Detailed Affiliations:
[1]Tri-Institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS): Georgia State University, Georgia Institute of Technology and Emory University, Atlanta, GA, USA;[2] Department of Psychiatry and Behavioral Sciences, University of New Mexico, Albuquerque, NM, USA;[3]Department of Psychiatry and Behavioral Sciences, Weill Institute for Neurosciences, University of California, San Francisco, CA, USA;[4]Mental Health Service, San Francisco Veterans Affairs Healthcare System, San Francisco, CA, USA;[5]Department of Psychiatry and Human Behavior, University of California, Irvine, CA, USA;[6]Departments of Psychiatry and Neuroscience, Yale University School of Medicine, New Haven, CT, USA;[7]Institute of Living, Hartford Healthcare Corp, Hartford, CT, USA
Abstract: Schizophrenia (SZ) patients exhibit abnormal static and dynamic functional connectivity across various brain domains. We present a novel approach based on static and dynamic inter-network connectivity entropy (ICE), which represents the entropy of a given network's connectivity to all the other brain networks. This novel approach enables the investigation of how connectivity strength is heterogeneously distributed across available targets in both SZ patients and healthy controls. We analyzed fMRI data from 151 SZ patients and 160 demographically matched healthy controls (HC). Our assessment encompassed both static and dynamic ICE, revealing significant differences in the heterogeneity of connectivity levels across available functional brain networks between SZ patients and HC. These networks are associated with subcortical (SC), auditory (AUD), sensorimotor (SM), visual (VIS), cognitive control (CC), default mode network (DMN), and cerebellar (CB) functional brain domains. Elevated ICE observed in individuals with SZ suggests that patients exhibit significantly higher randomness in the distribution of time-varying connectivity strength across functional regions from each source network, compared to HC. C-means fuzzy clustering analysis of functional ICE correlation matrices revealed that SZ patients exhibit significantly higher occupancy weights in clusters with weak, low-scale functional entropy correlation, while the control group shows greater occupancy weights in clusters with strong, large-scale functional entropy correlation. K-means clustering analysis on time-indexed ICE vectors revealed that cluster with highest ICE have higher occupancy rates in SZ patients whereas clusters characterized by lowest ICE have larger occupancy rates

for control group. Furthermore, our dynamic ICE approach revealed that in HC, the brain primarily communicates through complex, less structured connectivity patterns, with occasional transitions into more focused patterns. Individuals with SZ are significantly less likely to attain these more focused and structured transient connectivity patterns. The proposed ICE measure presents a novel framework for gaining deeper insight into mechanisms of healthy and diseased brain states and represents a useful step forward in developing advanced methods to help diagnose mental health conditions.

Keywords: schizophrenia, entropy, brain states, static functional connectivity, dynamic functional connectivity, functional connectivity patterns, mental health, biomarkers, fMRI, image data analysis

Title: A network-based systems genetics framework identifies pathobiology and drug repurposing in Parkinson's disease

Author list: Lijun Dou[1,2], Zhenxing Xu[3], Jielin Xu[1,2], Chengxi Zang[3], Chang Su[3], Andrew A. Pieper[4,5,6,7,8,9], James B. Leverenz[10,11], Fei Wang [3], Xiongwei Zhu[9], Jeffrey Cummings[12] & Feixiong Cheng [1,2,9]

Detailed Affiliations:

[1]Cleveland Clinic Genome Center, Lerner Research Institute, Cleveland Clinic, Cleveland, OH, 44195, USA;[2]Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, OH, 44195, USA;[3]Department of Population Health Sciences, Weill Cornell Medical College, Cornell University, New York, NY, 10065, USA;[4]Department of Psychiatry, Case Western Reserve University, Cleveland, OH, USA;[5]Brain HealthMedicines Center, Harrington Discovery Institute, University Hospitals Cleveland Medical Center, Cleveland, OH, 44106, USA;[6]Geriatric Psychiatry, GRECC, Louis Stokes Cleveland VA Medical Center, Cleveland, OH, 44106, USA;[7]Institute for Transformative Molecular Medicine, School of Medicine, Case Western Reserve University, Cleveland, OH, 44106, USA;[8]Department of Neurosciences, Case Western Reserve University, School of Medicine, Cleveland, OH, 44106, USA;[9]Department of Pathology, Case Western Reserve University, School of Medicine, Cleveland, OH, 44106, USA;[10]Department of Molecular Medicine, Cleveland Clinic Lerner College of Medicine, Case Western Reserve University, Cleveland, OH, 44195, USA;[11]Lou Ruvo Center for Brain Health, Neurological Institute, Cleveland Clinic, Cleveland, OH, 44195, USA;[12]Chambers-Grundy Center for Transformative Neuroscience, Department of Brain Health, School of Integrated Health Sciences, UNLV, Las Vegas, NV, 89154, USA.

Abstract: Parkinson's disease (PD) is the second most prevalent neurodegenerative disorder. However, current treatments only manage symptoms and lack the ability to slow or prevent disease progression. We utilized a systems genetics approach to identify potential risk genes and repurposable drugs for PD. First, we leveraged non-coding genome-wide association studies (GWAS) loci effects on five types of brain-specific quantitative trait loci (xQTLs, including expression, protein, splicing, methylation and histone acetylation) under the protein–protein interactome (PPI) network. We then prioritized 175 PD likely risk genes (pdRGs), such as SNCA, CTSB, LRRK2, DGKQ, and CD44, which are enriched in druggable targets and differentially expressed genes across multiple cell types. Integrating network proximity-based drug repurposing and patient electronic health record (EHR) data observations, we identified Simvastatin as being significantly associated with reduced incidence of PD (hazard ratio (HR) = 0.91 for fall outcome,95%confidence interval (CI): 0.87–0.94; HR = 0.88 for dementia outcome, 95% CI: 0.86–0.89) after adjusting for 267 covariates. In summary, our network-based systems genetics framework identifies potential risk genes and repurposable drugs for PD and other neurodegenerative diseases if broadly applied.

Keywords: Parkinson's disease; protein-protein network, GWAS, xQTLs, deep learning

Title: MetaphorPrompt2 - A Structure and Function Focused Approach for Extracting Causal Events from Biological Text
Author list: Parth Patel, Yu-Chiao Chiu, Yufei Hunag, and Jianqiu Zhang
Abstract: Biomedical literature is crucial for building knowledge graphs that explain disease mechanisms and guide drug discovery. However, even advanced large language models (LLMs) using in-context learning often misinterpret complex domain-specific causal statements or omit intermediary steps, resulting in incomplete pathway representations. MetaphorPrompt2 is motivated by cognitive theories of causal event representation and analogical reasoning. It improves molecular regulation pathway (MRP) extraction by emphasizing the structural relations and functional roles of biological entities rather than relying on surface-level grammar. This enables more effective metaphor construction and structural alignment between expert and general domains. The system integrates five components that collectively reduce parsing complexity and mitigate error propagation. MetaphorPrompt2 outperformed previous approaches, achieving a 24% improvement in edge prediction F1 score over a previous method without analogical reasoning. Notably, it eliminated missed entity errors and reduced m6A-related initiator extraction failures by 72.2%. These advances support the construction of more comprehensive biomedical knowledge graphs and enhance causal reasoning in LLMs, potentially facilitating automated hypothesis generation and accelerating drug discovery. Our findings highlight the value of a structure and function-focused approach for extracting complex causal knowledge from scientific text.
Keywords: Analogical Reasoning; LLM; Molecular Regulation Pathways; Knowledge Graphs; MetaphorPrompt; MetaphorPrompt2; Causal Events; Prompt Engineering

Title: DisSubFormer: A Subgraph Transformer Model for Disease Subgraph Representation and Comorbidity Prediction
Author list: Ashwag Altayyar and Li Liao
Abstract: Considering the complexity of diseases, comorbidity arises from intricate molecular interactions, functional relations, and shared pathological mechanisms, making comorbidity prediction a challenging yet prominent research topic in bioinformatics. As disease etiologies are inherently multifaceted, integrating multi-source data, such as the protein-protein interaction (PPI) network and Gene Ontology (GO), is crucial for identifying potential comorbid diseases. In this work, we develop DisSubFormer, a novel framework that combines GO-derived semantic features with PPI-based molecular interactions to generate biologically enriched representations of disease-associated proteins. These representations are leveraged within a subgraph Transformer model to represent disease subgraphs by capturing both local structural patterns and global relational information within the PPI network. More specifically, to enhance the scalability of the subgraph Transformer model on a large-scale PPI network, we introduce a biologically informed anchor patch sampling strategy integrated with a head-specific relational attention mechanism to learn context-aware disease subgraph representations for comorbidity prediction while simultaneously reducing computational complexity. We evaluate our proposed method on a benchmark dataset, achieving superior performance compared to state-of-the-art methods in disease comorbidity prediction, with an AUROC of 0.97.
Keywords: diseases; comorbidity; subgraph Transformer; protein-protein interaction; gene ontology; subgraph embedding; Hawkes process

Poster Session I
August 3rd
11:30 AM – 1:30 PM
Room: First floor Atrium

Title: Atlas of cell type-specific genetic impacts of alternative polyadenylation in immune-related diseases
Author list: Ting Zhang[1], Hui Chen[1], Shuxin Chen[1], Stephen Montgomery[2], Qin Li[3], Lei Li[1]

Detailed Affiliations
[1]Institute of Systems and Physical Biology, Shenzhen Bay Laboratory; Shenzhen, 518055, China;[2]Departments of Pathology and Genetics, Stanford School of Medicine; Stanford, CA 94305, USA; [3]Department of Genetics, Perelman School of Medicine, University of Pennsylvania; Philadelphia, PA 19104, USA;

Abstract
Genetic variants linked to immune disease risk often exert effects through regulatory mechanisms that vary across cell types and contexts, yet the role of post-transcriptional processes like alternative polyadenylation (APA) remains poorly understood. To address this, we analyzed single-cell RNA-sequencing data from over 7.27 million peripheral blood mononuclear cells across 2,022 individuals. We further developed a new computational method to map cell-type- and context-specific APA variation and identified 10,211 single-cell APA quantitative trait loci (sc-aQTLs) connecting 4,484 independent variants to 5,448 nearby genes, including autoimmune disease genes such as STAT6, which acts through APA independently of gene expression changes. Through colocalization and TWAS, we identified 267 APA-linked putatively causal disease genes, with 80.52% acting through expression-independent regulatory pathways. This sc-aQTL atlas highlights APA as a critical and underrecognized layer of immune disease regulation and a potential source of novel therapeutic targets.

Keywords
Single-cell, GWAS, eQTL

Title: Improving Metabolite Prediction Performance with Probability Filtering and Model Intergration
Author list: Yingtong Zhou[1] , Gabriel R. C. Pereira[2], Anja Conev[1], Bárbara de A. Abrahim-Vieira[2], Lydia E. Kavraki[1]

Detailed Affiliations
[1]Computer Science Department, Rice University, 6100 Main Street, Houston 77005, TX, USA;[2]Federal University of Rio de Janeiro (UFRJ), 373 Carlos Chagas Filho Avenue, Rio de Janeiro 21941-971, RJ, Brazil.

Abstract
Predicting drug metabolites is critical for understanding pharmacokinetics and ensuring drug safety in humans. However, traditional experimental approaches are slow, costly, and difficult to scale. Computational models offer a faster, more scalable alternative with strong predictive capabilities. Weevaluated four leading metabolite prediction tools: rule-based model (SyGMa) and MLbased models (MetaTrans, MetaPredictor), using a benchmark of 754 drugs and 1,478 known human metabolites from ChEMBL. Although MetaPredictor achieved the highest standalone accuracy, all models showed low precision from 4.0% to 13.9% in top-k predictions due to the generation of large volumes of low-confidence candidates. To address this, we implemented a probabilistic filtering framework that leverages modelspecific probability scores. Rule-based model outputs were scored by rule frequency, while ML models assigned probabilities reflecting the likelihood of each prediction based on learned patterns. These confidence scores enabled our threshold-based filtering to eliminate lowprobability and likely incorrect outputs. Specifically, prediction probability thresholds from 0 to 1 were evaluated in 0.01 increments for each model, using drugs from the MetaTrans training set to optimize precision-recall trade-offs. Each model obtains its own optimal threshold, and predictions falling below the threshold or identical to the input drugs were excluded. This filtering significantly reduced false positives but at the cost of recall. To recover lost recall and enhance robustness, we introduced a model integration that concatenates and deduplicates filtered outputs from MetaTrans, MetaPredictor, and SyGMa. This integration not only restored correct predictions filtered out by individual models but also provided a more balanced output set. Applying this pipeline significantly improved metabolite precision. In MetaTrans, 30.2%, 54.7%, and 67.1% of incorrect predictions were removed from the top-5, top-10, and top-15 outputs. In MetaPredictor, we filtered 30.0%, 73.1%, and 73.6% of false positives, while in SyGMa, we removed 62.3%, 81.3%, and 90.8% in the same ranks. The final model improved recall by 22.5% for top-5 and 10.3% for top-10 predictions compared to the best model, MetaPredictor, alone.For top-15 predictions, the integrated output not only preserved a comparably high recall but also improved precision by 20.6%. These results demonstrate that combining probabilistic filtering with model integration enhances the accuracy of metabolite prediction models. By reducing result clutter and helping users focus on the most reliable predictions, our framework offers a practical path toward more actionable and transparent drug metabolism predictions. Future work will explore deeper integration of confidence estimation and attention mechanisms into generation processes to further improve performance.

Keywords
drug metabolite prediction, machine learning, data science, precision-recall optimization, postprocessing, model integration

---

Title: Inferring Drug–Gene Relationships in Cancer Using Literature-Augmented Large Language Models

Author list: Ying-Ju Lai[1], Li-Ju Wang[1], Yufei Huang[1], Yu-Chiao Chiu[1]

Detailed Affiliations
[1]UPMC Hillman Cancer Center, University of Pittsburgh, Pittsburgh, PA, USA

Abstract

Understanding drug–gene relationships is essential for advancing targeted cancer therapies and drug repurposing strategies. However, the vast volume of biomedical literature poses significant challenges in efficiently extracting relevant insights. In this study, we developed an automated pipeline that leverages retrieval-augmented large language models (LLM) to infer drug–gene interactions using the most up-to-date biomedical literature. By integrating PubMed and state-of-the-art LLMs, our pipeline generates accurate, evidence-based inferences while addressing the limitations of static LLMs, such as outdated knowledge and the risk of producing misleading results. We systematically validated the pipeline's performance using curated databases and demonstrated its ability to accurately identify both well-established and emerging drug targets. Using our pipeline, we constructed a pan-cancer drug–gene interaction network among hundreds of FDA-approved drugs and key oncogenes. In a case study on liver cancer, we identified and validated an association between CTNNB1 mutations and enhanced sensitivity to sorafenib, highlighting a potential therapeutic strategy for this challenging mutation. To facilitate broad accessibility, we developed GeneRxGPT, a user-friendly web application that enables cancer researchers to utilize the pipeline without programming expertise or extensive computational resources. It provides intuitive modules for drug–gene inference and network visualization, streamlining the exploration and interpretation of drug–gene relationships. We anticipate that GeneRxGPT will empower researchers to accelerate drug discovery and development, making it a valuable resource for the cancer research community. This study was recently published in Cancer Research Communications. 2025;5(4):706–718. doi:10.1158/2767-9764.CRC-25-0030. PMID: 40293950.

Keywords
Drug–Gene Interactions; Large Language Models; Retrieval-Augmented Generation; R Shiny Application

---

Title: Pan-Cancer Single-Cell Profiling Uncovers the Biological Characteristics of Cancer-Testis Genes

Author list: Chunyang Fu[1], Ke Liu[1]

Detailed Affiliations
[1]Department of Medical Dataology, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan, China

Abstract
Cancer-testis genes (CTGs), characterized by their restricted expression in testicular and neoplastic tissues, have emerged as pivotal modulators of tumor progression and immunogenicity. Here, leveraging single-cell and single-nucleus RNA sequencing (scRNA-seq and snRNA-seq) data from 828 tumor samples spanning 13 cancer types, we systematically identified 407 CTGs based on their expression level in malignant cells. Compared to previously reported CTG sets, our curated genes exhibited broader expression among tumor samples and were predominantly enriched on autosomes. These CTGs displayed pronounced co-expression patterns, with a subset consistently activated in proliferating malignant cells, whereas non–cell cycle CTGs were highly correlated with epigenetic regulatory pathways. Functional analyses pinpointed chromatin-modifying enzymes, including ARID4B and YEATS2, as putative upstream regulators, whose knockdown resulted in widespread CTG downregulation. Moreover, we demonstrated that high CTG-count constitutes a robust and scalable feature for malignant cell identification across diverse omics data, including scRNA-seq, single-cell ATAC sequencing, and spatial transcriptomics. By quantifying transcriptomic similarity to a high CTG-count reference population, we developed a

streamlined classification framework that enables rapid and accurate annotation of malignant cells in tumor sc/snRNA-seq datasets across diverse cancer types. Clinically, CTG expression levels were associated with patient prognosis and immune microenvironment composition. In hepatocellular carcinoma, elevated CTG expression defined a stem-like, immune-suppressive malignant cell state enriched for regulatory T cells, linked to advanced disease and poor survival. Collectively, our study refines the understanding of CTGs in cancer biology and provides a valuable resource to facilitate future research and therapeutic targeting.

Keywords
Cancer-testis genes, Single-cell RNA sequencing, Epigenetic regulation, Malignant cell identification, Tumor immunogenicity

---

Title: A Multi-Agent Large Language Model Framework for Assessment of Patient Education Materials
Author list: Billy Zeng[1], Jin Ge[2], Alice Tang[3], Sijie Zheng[4]

Detailed Affiliations
[1]Department of Medicine, Kaiser Permanente Oakland Internal Medicine Residency, Oakland, CA, USA; [2]Division of Gastroenterology & Hepatology, University of California at San Francisco, San Francisco, CA, USA; [3]School of Medicine, University of California at San Francisco, San Francisco, CA, USA; [4]Department of Nephrology, Kaiser Permanente East Bay, Oakland, CA, USA

Abstract
Background: Large language models (LLMs) are rapidly being adopted for patient education materials (PEMS), but inaccurate content and hallucinations remain ongoing problems. There is an unmet need for systems capable of continuous auditing and flagging LLM-generated medical content for review and revision. Reliance on human review of LLM-generated content to ensure consistency with clinical guidelines and readability is impractical, especially for PEMs generated on-demand by LLMs post-clinical encounters. We, therefore, created a scalable multi-agent LLM system for initial screening and assessment of PEMs against clinical guidelines and established patient communication standards, using Anemia of Chronic Kidney Disease (CKD) as the specific use case.
Methods: We developed a multi-agent system using Python and the Langchain framework, integrating it with a locally-served Mistral-7B LLM for this proof-of-concept. The agents are:
1. A Resource Augmented Generation (RAG) Accuracy Assessment Agent to identify and verify medical claims against the Kidney Disease Improving Global Outcomes 2025 Anemia of CKD guidelines, which serve as a dynamic knowledge base.
2. A Readability and Actionability Assessment Agent to evaluate the PEM on the Agency for Healthcare Research and Quality's Patient Educational Material Assessment Tool (PEMAT).
3. An Integrative Assessment Agent to consolidate findings from the accuracy and PEMAT agents, producing a summary score and qualitative analysis.
To demonstrate the framework's lifecycle, we drafted a PEM with adversarial content, containing intentional inaccuracies, which underwent assessment.
Results: The system utilizing local LLM successfully generated comprehensive assessment reports for PEMs concerning Anemia of CKD. These reports provided detailed, claim-by-claim accuracy evaluations against guidelines and assigned PEMAT-P ratings for understandability and actionability. Analysis of these

automated reports demonstrates the framework's capacity to systematically identify strengths, inaccuracies, and areas for improvement in PEMs. (Scroll down to view example results/figures/tables)
Discussion & Conclusion: This proof-of-concept underscores the significant potential of LLMs for machine generated PEMs. The strength of our multi-agent framework is its systematic methodology for quality assurance – via integrating updated medical content with RAG and applying established PEMAT criteria for patient-centric evaluation. The PEMAT agent assesses readability and actionability, while the accuracy agent flagged adversarial content as contradicted by the knowledge base. This workflow supports human-in-the-loop validation, offering a flexible and rigorous method for quality control. This assessment-focused approach provides a scalable locally-deployable solution for healthcare institutions to review any medical content, and can be implemented across different foundation LLM models.

Keywords
Multi-Agent Workflow, Large Language Models, Patient Education, Clinical Informatics, Chronic Kidney Disease

---

Title: Multi-Agent Workflow, Large Language Models, Patient Education, Clinical Informatics, Chronic Kidney Disease
Author list: Emily Tang[1,2], Jake Y. Chen[3]
Detailed Affiliations
[1]North Carolina School of Science and Math, Durham, NC, USA; [2]Alphamind Club, LLC, USA;[3]Systems Pharmacology AI Research Center and Department of Biomedical Informatics and Data Science, School of Medicine, the University of Alabama at Birmingham, Birmingham, AL, USA

Abstract
Chimeric antigen receptor natural killer (CAR-NK) therapy is an emerging immunotherapy that directs NK cells to specific tumor cells. While it lowers the risk of cytokine release syndrome and graft-versus-host disease, challenges such as limited cell persistence and tumor immune es-cape remain. Of the two NK cell subtypes, CD56 bright and dim, the dim subset is more abun-dant and exhibits cytotoxic and infiltrative activity, while the bright subset mainly mediates cyto-kine secretion. This study investigates the correlation between CD56 dim NK cells and glioma patient survival, as well as NK-tumor interactions, to identify strategies for improving CAR-NK efficacy.
Methods: The correlation between CD56 dim NK cells and glioma patient survival was assessed using the Chinese Glioma Genome Atlas (CGGA) and the University of Alabama at Birmingham CANcer data analysis Portal (UALCAN), focusing on the signature genes NCAM1 (CD56) and FCGR3A. scRNA-seq data from normal brain and glioblastoma tissues were obtained from the CELLxGENE portal. Gene expression in NK cells, endothelial cells, and astrocytes was com-pared between normal and tumor tissues using the built-in Differential Expression tool, with re-sults analyzed by log-fold change/effect size and Welch t-test. For stromal genes found to be up-regulated, correlations with patient survival were evaluated using TIMER2.0 and CGGA portals.
Results: Analyses from CGGA and UALCAN showed a negative correlation between CD56 dim NK marker genes and glioma survival ($p < 0.05$). CELLxGENE analysis revealed that glioblastoma-infiltrating NK cells expressed higher levels of CD74 (1.91), SRGN (1.58), FTL (1.50), CD14 (1.50), FCER1G (1.36), TYROBP (1.25), HLA-A (1.17), and B2M (0.86), indicating roles in an-tigen presentation and cytotoxic

granule release. Glioblastoma endothelial cells upregulated in-flammatory and stress-response genes including NFKBIA (3.58), CEBPD (1.74), JUNB (1.58),
SAT1 (1.46), FOS (1.38), and vascular remodeling genes ZFP36 (1.81), SPP1 (1.58), and CTSB (1.42). Activation of the NF-κB pathway, marked by NFKBIA and downstream effector CEBPD, suggests immune suppression and NK evasion. Tumor astrocytes upregulated MHC I/II and in-flammatory genes while retaining neurotransmission-related gene expression. CEBPD expres-sion negatively correlated with survival ($p < 0.05$).
Conclusion: Although NK cells are activated in glioblastomas, stromal cells in the tumor microenvironment promote immunosuppression through vascular remodeling and inflammatory signaling, particu-larly via the NF-κB pathway. These mechanisms may underlie the observed negative correlation between CD56 dim NK cells and glioma survival. Further scRNA-seq and spatial transcriptomics studies are needed to identify key mediators of NK-tumor interactions and targets to enhance CAR-NK therapy.

Keywords
CD56 dim NK, glioma, tumor microenvironment, computational

Title: Understanding the Autonomy and Limitations of Web Agents Author list
Author list: Keefe Yang[1]

Detailed Affiliations
[1]Information Technology-Software Configuration & Development-Ex, Elil Lilly

Abstract
With the rapid introduction and rise of large language models like GPT-4, AI has revolutionized the tech landscape. It has fostered the development of software programs known as web agents. Traditional web agents differ from the modern web agents we know today. Modern web agents differ from traditional web agents in that they use LLMs to be able to execute multiple tasks, while traditional web agents are hard coded to only complete specific tasks. Additionally, by connecting to APIs like Selenium and BeautifulSoup, modern web agents are able to parse code and execute tasks autonomously. They can save time and effort by accessing websites to take pictures, summarize information, and fill out forms. However, traditional web agents still face challenges when it comes to processing data. Data is required as input for web agents to scrape information and make real time decisions, and it often needs to be highly structured. For example, it isn't sufficient to prompt a web agent to go to Wikipedia to output a summary for a person; you must provide the exact Wikipedia page. From my own experience, giving too general prompts that you would typically type into search engines yields no information or the wrong information. Moreover, many traditional web agents that are accessible to the public are still in the early stages of development. They often use reactive, instead of adaptive reasoning, and require human intervention to correct errors. Another issue I have faced is that web agents create and use entirely separate environments from your computer, leading to complications with login pages in which you already have saved passwords. Additionally, a regularly occurring bottleneck has been asking a web agent to click or fill out multiple elements on a webpage. CSS selectors are often required to move through a webpage, and when pages fail to load properly, selectors may change, leading to errors. From my use of LangChain, I have experienced just how time-consuming debugging can be, but I have also witnessed its potential capabilities when it comes to executing tasks autonomously. Modern web agents are constantly improving through their engagement with LLMs,

allowing them to accept queries in natural language and handle multiple tasks, both of which traditional web agents cannot do. While they clearly have their limitations, web agents will play a key role in redefining the digital space.

Title: Using Genomics Data for Basket Trial Design in Rare Diseases
Author list: Sungrim Moon[1,2], Jessica Maine[1], Ewy Mathé[1], Qian Zhu[1]

Detailed Affiliations
[1]Division of Pre-Clinical Innovation, National Center for Advancing Translational Sciences, National Institutes of Health, Rockville, MD, USA; 2Office of Data Science Strategy, National Institutes of Health, Bethesda, MD, USA.

Abstract
Gaining insight into the underlying molecular etiologies of rare diseases can aid cross-disease research, provide relevant information for the design of basket trials, and identify drug repurposing opportunities. In our preliminary study, we identified 37 rare disease clusters out of 3,242 rare diseases based on common genetic causes and pathophysiological mechanisms (1). However, these clusters were too broad for basket trial applications. In this study, we refined these clusters by collecting allelic variant data from the Online Mendelian Inheritance in Man (OMIM)1, along with corresponding Sorting Intolerant From Tolerant (SIFT) scores for single nucleotide polymorphisms (SNPs) and transcript-level data from Ensembl2, validated using the Medical Genomics Japan Variant Database (MGeND)3. To assess the functional impact of gene mutations, we used SIFT scores at the transcript level and calculated a weighted ratio of deleterious to total cases, defined as (deleterious cases × 3 + tolerated cases × 2 + cases with no SIFT score but with transcript-level information) / (deleterious cases + tolerated cases + cases with no SIFT score). We generated an imputed matrix by extracting deleterious levels from genetic and mutation data for each rare disease and identified shared mutations across diseases. Then, we applied Density-Based Spatial Clustering of Applications with Noise (DBSCAN) to the imputed matrix, creating sub-clusters within the preliminary 37 clusters. Our results are consistent with published studies on the proposed basket trial design (2, 3). For instance, they align with findings on a subcluster of NLRP3 gene mutations, which encode cryopyrin—a protein involved in immune function. These NLRP3 gene mutations are associated with diseases such as Neonatal-Onset Multisystem Inflammatory Disease, Familial Cold Autoinflammatory Syndrome, and Muckle-Wells Syndrome. This alignment suggests the algorithm may reflect known molecular relationships.
Acknowledgment: This project was partially supported by the intramural programs (ZIA TR000410-05) at NCATS/NIH. S.M. was partially supported by the Office of Data Science Strategy ODSS/NIH.

Keywords
Rare diseases, Genomics Data, Basket Trial Design, Clusters, Gene mutations

---

Title: Post-Myocardial Infarction Transcriptional Dynamics and Intercellular Communication in the Mouse Heart
Author list: Pankaj Singh Dholaniya[1], Helena Islam[1], Syed Alvi[1], Muhamad Mergeye[1], Onur Kanisicak[1,2], Mahmood Khan[1,2,3,*].

Detailed Affiliations
[1]Division of Basic and Translational Research, Department of Emergency Medicine, The Ohio State University, Columbus, OH;[2]Davis Heart and Lung Research Institute, The Ohio State University, Columbus, OH;[3]Department of Physiology and Cell Biology, The Ohio State University, Columbus, OH.

Abstract
Myocardial infarction (MI) is a major contributor to global cardiovascular disease and death. It results in decreased blood flow to the heart, leading to oxygen deprivation and impaired diastolic and systolic function, which increases the risk of arrhythmias. Various cardiac cell types respond to this stress to maintain heart function, but the precise mechanisms underlying these responses remain unclear. Characterizing transcriptional alterations in specific cell types during the early and late chronic phases following MI is critical for identifying new therapeutic targets. To systematically characterize cell-type-specific gene expression profiles we performed single-nuclei RNA sequencing (snRNA-seq) on left ventricular tissue from mouse hearts at baseline (Day 0) and post-MI (Week 1 and Week 4) to assess gene expression changes across different cardiac cell types. Given their critical involvement in contractile performance and their vulnerability to post-MI chronic stress, we focused on cardiomyocytes (CMs) to investigate transcriptional state transitions. A deep learning-based approach utilizing an encoder-decoder algorithm was employed to identify key gene expression modulations associated with these transitions. Additionally, we inferred alterations in cell-cell communication networks using CellChat to understand post-MI intercellular interactions. Our analysis revealed significant transcriptional shifts in different cell types following MI. CMs exhibited distinct gene expression modulations associated with excitation-contraction coupling and calcium handling. Fibroblasts (FBs) also displayed dynamic transcriptional changes, suggesting their involvement in the post-MI healing process. Cell-cell communication analysis identified key signaling pathways altered after MI, highlighting potential mechanisms of cellular crosstalk in the injured heart. This study provides a comprehensive transcriptional landscape of cardiac cell populations, especially CMs and FBs following MI, revealing key molecular changes and intercellular interactions. The findings from this study enhance the understanding of post-MI gene expression changes in these cardiac cells, aiding in identifying novel therapeutic targets.

Keywords
Myocardial infarction, Heart failure, snRNA Sequencing, cell-cell communication, Deep learning

---

Title: Characterizing Aberrant Brain Network Dynamics in Schizophrenia via Entropy-Synchronization Interplay
Author list: Natalia Maksymchuk[1], Robyn L. Miller[1] and Vince D. Calhoun[1]

Detailed Affiliations
[1]Tri-Institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Georgia State University, Georgia Institute of Technology and Emory University, Atlanta, GA USA

Abstract

The human brain is a complex adaptive system that constantly balances the integration and segregation of information across brain networks. This balance is reflected in its dynamic transitions between synchronized and desynchronized network states. A theoretical framework capturing this duality is the concept of chimera states, where coherent and incoherent activity patterns coexist within a network of coupled oscillators with intrinsic heterogeneity. Such hybrid dynamics are thought to underlie flexible cognitive functioning but may be disrupted in psychiatric disorders such as schizophrenia (SZ), which is characterized by altered functional connectivity and impaired information processing.
We investigated the relationship between network synchronization and dynamic inter-network connectivity entropy (DICE) in SZ and matched healthy control (HC) groups. Using resting-state fMRI data from 311 individuals (151 individuals with SZ, and 160 HC), we extracted 53 intrinsic brain networks via group independent component analysis (ICA) and computed time-resolved functional connectivity using a sliding window approach. We quantified synchronization and DICE for each network over time, integrating methods from network neuroscience and information theory.
Our results show that SZ patients exhibit significantly reduced average synchronization between most individual brain networks and the rest of the brain. Both synchronization and DICE fluctuated across time exhibiting a strong negative relationship. We revealed that DICE dynamics of SZ patients are more loosely and less sensitively coupled to synchronization, especially in subcortical, cerebellar, and cognitive control domains, reflecting a compromised integration–segregation mechanism. Furthermore, we found that stronger DICE-synchronization correlations in the cognitive control, auditory, and sensorimotor networks were associated with better cognitive performance, particularly in visual learning and attention/vigilance.

Importantly, the state-level analysis revealed that SZ participants had significantly higher occupancy in connectivity states characterized by high entropy and low synchronization, while controls more often occupied low entropy and high synchronization states. SZ patients showed a marked inability to transition into low entropy and high synchronization states – connectivity configurations that likely reflect efficient and ordered brain function.
These findings suggest that disrupted synchronization-entropy dynamics may underlie cognitive and functional impairments in schizophrenia. Reduced flexibility in transitioning to highly synchronized and low-entropy states may reflect a fundamental deficit in the SZ brain's capacity
to enter and maintain efficient global configurations. Our study underscores the importance of entropy-synchronization interplay as a potential biomarker of network-level dysfunction in psychiatric illness and offers a novel lens for examining the brain dynamics in health and disease.

Keywords
schizophrenia, entropy, network synchronization, mental health, biomarkers, fMRI

Title: MolTIF: A Molecular Transformer for Interpretable Fragment-level Representation Learning
Author list: Linqing Mo[1], Bin Chen[2], Jiayu Zhou[3]

Detailed Affiliations
[1]Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, USA;
[2]Department of Pediatrics and Human Development, Michigan State University, East Lansing, MI, USA;
[3]School of Information, University of Michigan, Ann Arbor, MI, USA.

Abstract
Fragment-based molecular pretraining with graph neural networks (GNNs) offers a chemically meaningful and data-efficient approach to molecular representation learning. Leveraging functional fragments helps capture meaningful substructures and improve generalization, with self-supervised pretraining reducing the need for labeled data. However, existing approaches lack mechanisms to model inter-fragment dependencies or provide interpretability during prediction. In this work, we introduce MolTIF, a multi-token Transformer framework that treats each molecular fragment as an individual token and employs self-attention to capture complex relationships among fragments. To further enhance representation quality, we propose a dual-level contrastive learning strategy that aligns fragment and atom-level embeddings at a local level, while enforcing consistency between atom-based and fragment-based molecular views globally. Our approach achieves state-of-the-art performance on multiple molecular property prediction benchmarks and offers fine-grained interpretability by quantifying fragment-level contributions during inference, which could aid domain experts in distilling complex relationships between molecular features and labels from a prediction task.

Keywords
Fragment-based Molecular Pretraining, Contrastive Learning, Molecular Representation, Model Interpretability

---

Title: A Multimodal Vision Transformer using Fundus and OCT Images for Interpretable Classifications of Diabetic Retinopathy

Author list: Shivum Telang[1], Wei Chen[2]
Detailed Affiliations
[1]Department of Pediatrics, University of Pittsburgh School of Medicine, Pittsburgh, PA; [2]North Allegheny Senior High School, Pittsburgh, PA; [3]Department of Biostatistics and Health Data Science, University of Pittsburgh School of Public Health, Pittsburgh, PA

Abstract
Diabetic Retinopathy (DR) is a leading cause of vision loss worldwide, requiring early detection to preserve sight. Limited access to physicians often leaves DR undiagnosed. To address this, AI models leverage lesion segmentation for interpretability, but manually annotating lesions is impractical for clinical use. More importantly, physicians require a model that explains why a classification was made rather than just highlighting lesion locations. Furthermore, current models are one-dimensional, relying on a single imaging modality and achieving limited effectiveness. This study introduces RetGEN, a self-supervised learning-based framework that enhances DR classification through a multimodal vision-transformer architecture with a multimodal contrastive loss function. By integrating 2 types of retinal scans (OCT and fundus), RetGEN improves classification accuracy while providing explainable insights for ophthalmologists. For interpretability, the model generates paired Grad-CAM heat-maps showcasing individual neuron weights across OCT images, visually highlighting the regions contributing to DR severity classification. Trained on a single-modality dataset of 3,000 fundus images and 1,000 OCT images for the multimodal vision transformer we use 125 paired images collected from ophthalmologists in the area, RetGEN outperforms other state-of-the-art models on DR severity classification, delivering more accurate, interpretable, and

clinically meaningful assessments of DR severity. This innovative methodology addresses key limitations in current DR diagnostics, offering a practical and comprehensive tool for improving patient outcomes.

Keywords
Vision Transformer, Contrastive Loss Function, Weight Matrices, Positional Encoding

---

Title: Distinct DNA Methylation Patterns in Alzheimer's Disease Brain Tissue
Author list: Raymond Cheng[1,2,3] *, Jingmin Shu[1,2 *,] Hai Chen[1,2], Li Liu[1,2]

Detailed Affiliations
[1]College of Health Solutions, Arizona State University, Phoenix, AZ 85004, USA.[2] Center for Personalized Diagnostics, Biodesign Institute, Arizona State University, Tempe, AZ 85281, USA.[3]Case Western Reserve University, Cleveland, OH 44106, USA

Abstract
Alzheimer's disease (AD) is a complex neurodegenerative disorder shaped by dysregulated molecular and cellular processes. Although DNA methylation plays a critical role in gene regulation, few consistent methylation markers have been identified in AD. We hypothesize that inconsistencies arise from unrecognized molecular subgroups.
We analyzed DNA methylation profiles from 55 AD patients and 24 cognitively normal controls to identify AD-associated methylation changes while accounting for inter- and intra-group variance. Unsupervised clustering revealed three distinct methylation subgroups: Cluster 2 and Cluster 3 showed predominantly hypo- and hypermethylation, respectively, while Cluster 1 displayed a transitional pattern. Cluster membership explained a moderate proportion of methylation variance (median 12.6%), while cell-type composition accounted for 6.8%. Notably, 45.3% of the variance initially attributed to clusters was explained by cell types, underscoring the importance of adjusting for cell-type composition to accurately interpret cluster-specific epigenetic signals.Using regression models incorporating both main effects and interactions, we identified 703 differentially methylated probes (adj. $p < 0.1$). Clustering improved model-based diagnostic accuracy from ~50% to ~70% across three independent validation and testing datasets. Although overlap-based reproducibility did not improve with clustering, we identified 369 additional biomarkers. Pathway enrichment analysis revealed shared pathways between clustering and non-clustering approaches, including key AD-related pathways. We also identified cluster-specific pathways, suggesting distinct biological mechanisms.
To assess potential genetic and clinical influences on clustering, we examined a fifth dataset. The protective APOE ε2 allele was enriched in the hypomethylated cluster, which also showed altered AD pathology severity. In conclusion, DNA methylation–based subgroups in AD reflect underlying heterogeneity driven by cell composition, epigenetic regulation, genetic interactions, and disease stage. These subgroups offer a framework for improving biomarker discovery and therapeutic targeting in AD.

Keywords
Alzheimer's disease, DNA methylation, Epigenetic regulation, Biomarker discovery, Cell-type composition, APOE ε2 allele.

Title: Robust Non-Negative Matrix Factorization Deconvolution for Developmental Bulk Tissue RNA-seq Data

Author list: Suxian Zhou[1], Su Xu[1], Xue Wang[4], Duan Chen[1], Jun-Tao Guo[3], Shaoyu Li[1,2,*]

Detailed Affiliations

[1]Department of Mathematics and Statistics, UNC Charlotte, NC; [2]School of Data Science, UNC Charlotte, NC; [3]Department of Bioinformatics and Genomics, UNC Charlotte, NC; 4Mayo Clinic, Department of Quantitative Health Sciences, Jacksonville, Florida.

Abstract

Accurate cell-type or stage deconvolution of bulk RNA sequencing (RNA-seq) data is critical for understanding cellular heterogeneity in complex tissues, particularly during dynamic developmental processes. Traditional reference-free methods often suffer from limited interpretability and high sensitivity to noise. To overcome these challenges, we propose an enhanced framework-Geometric Structure Guided Non-negative Matrix Factorization Plus (GSNMF+)-which extends geometric structured NMF for robust deconvolution of developmental bulk RNA-seq data. GSNMF+ incorporates artificially generated pseudo-bulk RNA-seq as an augmentation to strengthen the linear relationship between cellular composition and the expression of marker genes, thereby improving both robustness and interpretability. Additionally, it includes a component annotation module to facilitate biological interpretation of the inferred factors. Applied to several simulated datasets and realistic Plasmodium bulk RNA-seq data, our approach demonstrates strong performance in recovering underlying stage compositions. Compared to existing deconvolution methods, GSNMF+ also exhibits greater stability and reliability across varied datasets.

Keywords

Nonnegative matrix factorization, reference-free deconvolution, Bulk RNA sequencing

---

Title: RNA Splicing Events in Circulation Distinguish Individuals With and Without New-onset Type 1 Diabetes

Author list: Bobbie-Jo M Webb-Robertson[1*], Wenting Wu[2,3*,] Javier E Flores1, Lisa M Bramer[1], Farooq Syed[2], Sarah A Tersey[4], Sarah C May[4], Emily K Sims[2], Carmella Evans-Molina[2,5], and Raghavendra G Mirmira[4] *These authors contributed equall

Detailed Affiliations

1Biological Sciences Division, Pacific Northwest National Lab, Richland, WA, USA; 2Center for Diabetes and Metabolic Diseases, Indiana University School of Medicine, Indianapolis, IN, USA; 3Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, USA; 4Diabetes Research and Training Center and the Department of Medicine, The University of Chicago, Chicago, IL, USA; 5Roudebush Veteran's Affairs Medical Center, Indianapolis, IN, USA

Abstract

Alterations in RNA splicing may influence protein isoform diversity that contributes to or reflects the pathophysiology of certain diseases. Whereas specific RNA splicing events in pancreatic islets have been

investigated in models of inflammation in vitro, how RNA splicing in the circulation correlates with or is reflective of T1D disease pathophysiology in humans remains unexplored.
To investigate whether alternative RNA splicing differs between individuals with and without new-onset type 1 diabetes (T1D), we performed deep whole-blood RNA sequencing on two independent cohorts. The training cohort included 12 individuals with new-onset T1D and 12 age- and sex-matched controls (mean depth: 185 million reads), and the validation cohort is kept the same design (mean depth: 136 million reads). Alternative splicing was assessed using rMATS 4.1.1 across five event types: skipped exons (SE), mutually exclusive exons (MXE), retained introns (RI), alternative 5' splice sites (A5SS), and alternative 3' splice sites (A3SS). Normalized exon inclusion levels were used as features in a Random Forest (RF) classifier. For each FDR and tree number combination, we performed 3-fold cross-validation, repeating this process 25 times to assess model robustness. The average AUC across folds was used to evaluate classification performance.
Distinct patterns of RNA splicing differentiated participants with T1D from unaffected controls. In the training cohort, 7,915 splicing events (FDR < 0.05) across 4,192 genes significantly distinguished T1D patients from controls. Across all types of events in the training data, the average AUC exceeded 0.9, indicating a high predictive performance. Notably, retained introns (RI) provided the strongest predictive signal, with AUCs of 0.901 (training) and 0.868 (validation). Adding other splicing event types to the RI-based model did not enhance classification accuracy, suggesting that RI events alone carried the greatest predictive capability. Gene ontology enrichment of RI-associated transcripts revealed upregulation of systemic antiviral response pathways in T1D, suggesting that abnormal intron retention may lead to immunogenic transcript production, potentially triggering autoimmune processes.
Alternative RNA splicing events in whole blood are significantly enriched in individuals with new-onset T1D and can effectively distinguish them from unaffected controls. Moreover, RNA splicing profiles, especially retained intron events offer potentials for uncovering novel mechanisms underlying T1D pathogenesis

Keywords
Type 1 diabetes, machine learning, alternative RNA splicing, biomarkers

---

Title: Multimodal Single-Cell and Computational Deconvolution Analysis Uncovers Altered Immune Cell Subsets and Their Role in Early-Onset Type 1 Diabetes

Author list: Carmella Evans-Molina[1,2,3,4], Tingbo Guo[5], Cameron R.Rostron[1,2,3], Raghavendra, G.Mirmira[6], Yunlong Liu[7], Jia Shen[7,8], Chi Zhang[5], Wenting Wu[1],

Detailed Affiliations
[1]Center for Diabetes and Metabolic Diseases, Indiana University School of Medicine, Indianapolis, IN, USA; [2]Department of Pediatrics, Indiana University School of Medicine, Indianapolis, IN, USA; [3]Herman B Wells Center for Pediatric Research, Indiana University School of Medicine, Indianapolis, IN, USA; [4]Richard L. Roudebush VA Medical Center, Indiana University School of Informatics and Computing, Indianapolis, IN, USA; [5]Department of Biomedical Engineering and Knight Cancer Institute, Oregon Health & Science University, Portland, OR, USA;[6]Kovler Diabetes Center and Department of Medicine, The University of Chicago, Chicago, IL, USA;[7]Department of Medical and Molecular Genetics, Indiana

University School of Medicine, Indianapolis, IN, USA; 8Medical Sciences Program, Indiana University School of Medicine, Bloomington, IN, USA

Abstract

Single-cell RNA sequencing (scRNA-seq) was conducted to profile 108,795 peripheral blood mononuclear cells (PBMCs) obtained from 5 youth within 48 hrs of Stage 3 Type 1 diabetes (T1D) onset and 5 age- and sex-matched healthy controls (HC) and identified 31 distinct immune cell clusters. We apply our procedure for another CITE-seq datasets of 335,381 human pancreatic lymph node (pLN) cells obtained from 19 human donors with panels extending to 163 antibodies to contract a multimodal atlas of the cross-tissue immune system.

Firstly, focusing on cell-specific transcriptional changes, PBMCs profiles showed that Natural Killer (NK) cells had the largest number of differentially expressed genes (n=363). Strikingly, multi-modal cell-specific gene expression revealed the most robust similarities between NK cells in the circulation and the pLN ($r = 0.97$, $P = 1.38\times10\text{-}5$ by fold change (FC) ≥ 1.19 and FDR≤ 0.05 threshold), followed by CD4+ TCM cells. Two major NK cell types, CD56brightCD16lo and CD56dimCD16hi, key players in immune surveillance and cytotoxicity, exhibit altered subset compositions and functional shifts in individuals with recently diagnosed T1D. Protein disulfide isomerase family A member 3 (PDIA3) gene has been found upregulated (T1D vs HC) in NK cell across two tissues, and statistically significantly upregulated (average FC= 1.53) by multiple NK cytokines, especially type I interferons. By SCENIC prediction into upstream regulator, PDIA3 is significantly enriched of IRF1. To validate these findings, a CD56brightCD16lo cell line (NK-92) was treated with type 1 interferon. RT-qPCR analysis showed increased expression of IRF1 (FC= 3.48, adjusted $P = 0.008$) and markers of activation and maturity.

Secondly, using this scRNA-seq reference dataset derived signatures, we ran computational deconvolution algorithms CIBERSORTx and our in-house reference-free method ICTD to deconvolute cell proportions using public clinical trial data testing the anti-CD20 monoclonal antibody rituximab (n=37) vs. placebo (n=17). Strong concordance was observed between the methods' estimates ($P < 2.2\times10\text{-}16$). Rituximab responders and non-responders showed distinct cell composition patterns, notably in $CD4^+$ TCM, Treg, and neutrophil populations. Lastly, we present a user-friendly R Shiny application designed for intuitive visualizing scRNA-seq gene expression data.

Single-cell profiling reveals an increasing trend in NK subsets with heightened cytotoxic activity, which may contribute to pancreatic islet destruction, highlighting a potential link between NK cell dynamics and T1D disease progression. This work demonstrates the power of multimodal scRNA-seq integration and computational deconvolution in linking cell-specific immune dynamics to clinical outcomes and suggest new computational frameworks for biomarker discovery and therapy stratification in autoimmune diseases.

Keywords

Single-Cell RNA-seq, CITE-seq, Deconvolution, Multimodal, Type 1 Diabetes, Natural Killer (NK) cells

---

Title: A Spatial Transcriptomics Analysis: Benchmarking and Pipeline

Author list: Gillespie J[1,2], Xie J[1,3,4], Jung KJ[5], Pietrzak M[1], Hardiman G[6], Song MA[7], Chung D[1,3,4]

Detailed Affiliations

[1]Department of Biomedical Informatics, The Ohio State University, Columbus, 43210, Ohio, U.S.A.[2]Comprehensive Cancer Center, The Ohio State University, Columbus, 43210, Ohio, U.S.A.[3]The

Interdisciplinary Ph.D. program in Biostatistics, The Ohio State University, Columbus, 43210, Ohio, U.S.A.[4]Pelotonia Institute for Immuno-Oncology, The James Comprehensive Cancer Center, The Ohio State University, Columbus, 43210, Ohio, U.S.A.[5]Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, U.S.A.[6]Faculty of Medicine, Health and Life Sciences, School of Biological Sciences & Institute for Global Food Security, Queens University Belfast, Stranmillis Road, Belfast, BT9 5AG, UK. G.Hardiman@qub.ac.uk[7] Division of Environmental Health Sciences, College of Public Health, Ohio State University, Columbus, OH, U.S.A.

Abstract
Spatial transcriptomics (ST) profiles close-to-cell-level gene expressions with spatial coordinates, which allow for the discovery of detailed RNA localization, study development, investigation of the tumor microenvironment, and creation of a tissue atlas. ST analysis consists of many steps, four of which are frequently performed: tissue architecture identification, spatially variable gene detection, cell-cell communication , and deconvolution. A large number of ST analysis software is available with little information on which may be better suited for particular datasets or compute environments. While individual papers exist for each software demonstrating their abilities, the datasets analyzed, hardware utilized, and quality metrics vary widely across publications. Benchmarking studies compare software head to head under controlled conditions for each task and this provides a more accurate assessment. However, again, as each benchmarking study uses different metrics for evaluating software, a direct comparison is not possible. To address this challenge, we implemented a meta-review of multiple benchmarking papers to make better recommendations and assist in making a more informed software choice. Several factors are considered for recommendations, although not in equal priority, including: accuracy, runtime, system requirements, programming language, and compatibility with the Visium 10X platform, a popular choice for ST sequencing. Based on this meta-review, we suggest a potential pipeline for spatial transcriptomics data analysis, with a focus on the 10X Visium platform.

Keywords
spatial transcriptomics, 10X Visium, tissue architecture identification, spatially variable gene detection, cell-cell communication identification, and deconvolution analysis

---

Title: Characterizing mRNA Localization in Polarized Neuronal Compartments With Spatial Transcriptomics

Author list: Chenyang Yuan[1,2], Krupa Patel[1], Hongshun Shi[1,3,4], Hsiao-Lin V. Wang[1,5],Feng Wang[1], Ronghua Li[1,5], Yangping Li[1], Victor G. Corces[1,5], Hailing Shi[1,4,5], Sulagna Das[1,6], Jindan Yu[1,3,4], Peng Jin[1,5], Bing Yao[1 *], Jian Hu[1,2 *]

Detailed Affiliations
[1]Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, USA; [2]Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA, USA; [3]Department of Urology, Emory University School of Medicine, Atlanta, GA, USA; [4]Winship Cancer Institute, Emory University School of Medicine, Atlanta, GA, USA; [5]Emory Center for Neurodegenerative Diseases, Emory University School of Medicine, Atlanta, GA, USA; [6]Department of Cell Biology, Emory University School of Medicine, Atlanta, GA, USA.

Abstract
Background: Neurons are highly polarized cells with distinct distal compartments—such as synapses, dendrites, and axons—that enable connectivity and communication. Thousands of mRNAs have been shown to localize to these distal sites, supporting local protein synthesis critical for their plasticity and function. The diversity and regulation of these mRNAs are closely linked to neuronal function and are implicated in various neurodegenerative diseases. However, conventional RNA sequencing methods, such as single-cell/nucleus RNA sequencing, are limited to capturing nuclear and somatic transcripts, leaving the mRNA content of distal compartments unexplored. Recent advances in in situ spatial transcriptomics (iST) allow subcellular measurement of gene expression by capturing individual mRNAs in their spatial context. This resolution creates new opportunities to study compartmentalized transcripts. As most analytical pipelines still focus on mRNAs located in the cell body, the primary challenge lies not in the technology itself, but in the development of appropriate analytical methods.
Methods: We present mcDETECT, a machine learning framework for studying mRNA localization in polarized neuronal compartments using iST data. mcDETECT first applies density-based clustering to identify RNA granules within distal subcellular regions. It then integrates information from nearby RNA granules to reconstruct compartment-specific expression profiles for individual neurons. These profiles reveal neuronal cell states that complement conventional cell type classifications based on nuclear mRNA.
Results: We applied mcDETECT to mouse brain datasets generated using various state-of-the-art iST platforms, including Xenium 5K, MERSCOPE, and CosMx. mcDETECT successfully identified RNA granules within polarized neuronal compartments and further classified them into distinct subtypes, each associated with specific functional roles across brain regions. Leveraging compartment-specific expression profiles, mcDETECT consistently uncovered previously unrecognized neuronal states across all different brain regions. In an Alzheimer's disease (AD) mouse model, mcDETECT revealed alterations in both compartmentalized RNA patterns and neuronal states prior to observable neuronal loss, potentially serving as molecular markers for early AD diagnosis and therapeutic intervention.
Conclusion: mcDETECT is the first method to systematically characterize mRNAs for polarized neuronal compartments, offering novel molecular insights and revealing potential therapeutic targets within these distal regions for neurodegenerative diseases.

Keywords
spatial transcriptomics, subcellular compartment, RNA granule, Alzheimer's disease, machine learning

---

Title: MERGE: Multi-faceted Hierarchical Graph-based GNN for Gene Expression Prediction from Whole Slide Histopathology Images

Author list: Aniruddha Ganguly[1], Debolina Chatterjee[2], Wentao Huang[1], Jie Zhang[2], Alisa Yurovsky[3], Travis Steele Johnson[2], Chao Chen[1]

Detailed Affiliations
[1]Stony Brook University, NY, USA;[2]Indiana University School of Medicine, IN, USA

Abstract
Recent advances in Spatial Transcriptomics (ST) pair histology images with spatially resolved gene expression profiles, enabling predictions of gene expression across different tissue locations based on image

patches. This opens up new possibilities for enhancing whole slide image (WSI) prediction tasks with localized gene expression. However, existing methods fail to fully leverage the interactions between different tissue locations, which are crucial for accurate joint prediction. To address this, we introduce MERGE (Multi-faceted hiErarchical gRaph for Gene Expressions), which combines a multi-faceted hierarchical graph construction strategy with graph neural networks (GNN) to improve gene expression predictions from WSIs. By clustering tissue image patches based on both spatial and morphological features, and incorporating intra- and inter-cluster edges, our approach fosters interactions between distant tissue locations during GNN learning. As an additional contribution, we evaluate different data smoothing techniques that are necessary to mitigate artifacts in ST data, often caused by technical imperfections. We advocate for adopting gene-aware smoothing methods that are more biologically justified. Experimental results on gene expression prediction show that our GNN method outperforms state-of-the-art techniques across multiple metrics. Our paper was recently accepted for publication at IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR) 2025 and was presented as a poster in the conference. The full manuscript can be accessed from ArXiv since the proceedings of the conference have not been published at the time of this submission.

Keywords
Spatial Transcriptomics, Gene Expression Prediction and Smoothing

---

Title: Analyzing Soil Microbial diversity in an Organically Managed Fields Amended with Dredged Material in Northeast Ohio.
Author list: Shikshya Gautam1, Zhaohui Xu1, Angélica Vázquez-Ortega2

Detailed Affiliations
[1]Department of Biological Sciences, 2School of Earth Environment and Society, 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 of open water disposal utilizing this dredged material as a soil amendment could provide a sustainable approach to managing DM disposal issues. Previous research has shown the DM can improve soil physio chemical properties in the greenhouse setting; current research will focus on investigating the impact of dredged material (DM) on the microbial community in an organic field in northwest Ohio. This study aimed to characterize the taxonomic diversity of microbial communities in the fields amended with DM. Using 16S rRNA gene sequencing analyzed through QIIME2, we profiled the microbial community structure. The study has three organically managed fields. Field 1 and field 3 received 40 tons per acre and 80 tons per acre of DM, respectively, and field 2 received no DM (control treatment). Alpha diversity metrics (Shannon index) did not show significant differences among the fields. ($p > 0.05$). 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. The study further suggests studying the relation

between specific microbial taxa and functional pathways involved in nitrogen metabolism, carbon cycling, and stress response which will provides valuable insights into the complex interactions between microbial diversity and function in soil environments and highlights the importance of microbial community management for agricultural productivity and environmental health.

Keywords
Dredged Sediment, Lake Erie, Soil Health, Soil microbiome, Microbial diversity.

Title: AI-Powered Computational Platform for Corynebacterium glutamicum Strain Optimization: Predicting Mutational Outcomes for Enhanced L-Glutamate Production
Author list: Rahma Hussein

Detailed Affiliations
Prototyping Research Lab, Faculty of Biotechnology, Modern Science and Arts University (MSA), Giza, Egypt.

Abstract
Abstract: This study introduces an AI-driven bioinformatics framework to enhance strain engineering by predicting the impact of targeted DNA mutations on Corynebacterium glutamicum L-glutamate productivity, significantly reducing ineffective laboratory experiments. Unlike traditional random mutagenesis, which is time-consuming and imprecise, our opensource platform integrates a reinforcement learning-based deep neural network, trained on real genomic (FASTA/GFF) and phenotypic data, to forecast mutation outcomes with high accuracy. The system identifies both beneficial effects (e.g., increased productivity) and deleterious effects (e.g., 0.8x productivity), enabling researchers to bypass non-viable mutations early. Key components include: (1) a real-time biological simulator modeling bacterial behavior under complex industrial conditions (temperature 10-60°C, pH 3-10, varying antibiotic concentrations), incorporating Perlin noise algorithms for realistic bacterial movement and resource competition; (2) an interactive gene editing interface supporting single nucleotide polymorphisms (SNPs), insertions, deletions, and frameshifts, providing instant phenotypic feedback; and (3) a predictive database of 58 annotated genes (e.g., gadABC for acid resistance, atp operon for energy metabolism), linking mutations to phenotypic traits. The framework evaluates seven critical parameters: L-glutamate yield, growth rate, energy efficiency, pH tolerance, thermal stability, antibiotic resistance, and genetic stability, delivering quantitative metrics (e.g., productivity in g/L, survival rate). Validation studies demonstrate significant reductions in deleterious mutations, streamlining strain development. This platform advances precision strain engineering, with applications in sustainable biofuel production, pharmaceutical manufacturing, and environmental remediation, and offers adaptability for other microbial organisms

Keywords
precision strain engineering, Corynebacterium glutamicum, deleterious mutation prediction, reinforcement learning, dynamic simulation, industrial biotechnology

---

Title: MPET: Dissecting Intracellular Transport Mechanisms Underlying Surface Protein Dysregulation in Disease via Single-Cell Multi-Omics

Author list: Rekha Mudappathi[1,2], Vaishali Bhardwaj[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

Abstract

Cell surface proteins mediate essential immune functions, including antigen presentation, cytokine signaling, and cell-cell interactions, and are frequently targeted in diagnostic and therapeutic applications. However, surface protein abundance often diverges from the transcriptional levels of their coding genes, highlighting the importance of regulatory mechanisms beyond the transcriptional level.

To address this, we developed MPET (Modeling Protein Expression and Transport), a computational framework that leverages Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-seq) data to model the role of intracellular protein transport (ICT) in regulating surface protein expression and disease phenotypes. MPET decomposes the ICT regulatory network into trios of surface protein, its coding gene, and an ICT gene. Using mixed-effects regression and both single- and multi-exposure mediation models, MPET identifies transport circuits where ICT gene activity influences phenotypes by modulating surface protein transport independently of gene transcription.

We applied MPET to single-cell CITE-seq data from peripheral blood monocytes in COVID-19 patients and uncovered widespread rewiring of ICT pathways across disease states. MPET revealed that increased surface expression of CD69 in severe cases occurred without corresponding changes in its mRNA levels, implicating enhanced ICT activity. Conversely, reduced expression of HLA-DR, a critical antigen presentation molecule was linked to overactivation of endocytic pathways and disrupted ICT function.

Moreover, the findings highlight the multifaceted roles of ICT genes such as CLU, CD74, and ACTB, many of which are involved not only in general intracellular transport, but also in immune signaling and regulation. These results underscore their dual contributions to both cellular logistics and immune response, suggesting that immune dysfunction in severe COVID-19 may stem from the combined effects of transport disruption and inflammatory imbalance.

By integrating multi-omic information from CITE-seq at single-cell resolution, MPET provides a mechanistic bridge between gene regulation and protein-level phenotypes. Its modular design supports broad application across disease contexts where surface protein dysregulation plays a role, including cancer and neurodegeneration. Our results demonstrate the power of MPET to uncover regulatory programs beyond transcription that shape immune responses and disease trajectories, offering new insights into therapeutic intervention strategies.

Keywords

CITE-seq, surface protein expression, intracellular transport, mediation analysis, single-cell multi-omics, COVID-19, post-transcriptional regulation, mixed-effects modeling, immune dysregulation.

Title: Toward the best generalizable performance of machine learning in modeling omic data

Author list: Fei Deng[1], Yongfeng Zhang[2], Lanjing Zhang[1, 3, 4]

Detailed Affiliations
[1]Department of Chemical Biology, Ernest Mario School of Pharmacy, Rutgers University, Piscataway, NJ, USA;[2] Department of Computer Sciences, School of Arts & Sciences, Rutgers University, Piscataway, NJ, USA;[3] Department of Pathology, Princeton Medical Center, Plainsboro, NJ, USA;4 Rutgers Cancer Institute of New Jersey, New Brunswick, NJ, USA.

Abstract
There are often performance differences between intra-dataset and cross-dataset tests in machine learning (ML) modeling. However, reducing these differences may also reduce ML performances. It is a challenging dilemma for developing models that excel in intra-dataset testing and are also generalizable to cross-dataset testing. Therefore, we aimed to understand and improve consistency and performance of ML in intra-dataset and cross-dataset tests. Cases of lung adenocarcinoma in the TCGA (n=286) and Singapore Oncology Database (n=167) were subject to ML modelling with support vector machine (SVM), XGBoost (XGB), least absolute shrinkage and selection operator (LASSO), multilayer perceptron (MLP), and logistic regression (LR). Multivariable analyses were carried out to identify the factors independently associated with ML performance, that were in our study ML model, parameter configurations, normalization of transcriptomic data, and feature selection with various differentiated expressed and non-differentially expressed genes combinations on both intra-dataset and cross-dataset testing. We then seek the models with best consistency and performance, using three criteria, namely, a top 10% performer, least statistical indifference in ML performance of intra-dataset and cross-dataset testing and with ML performance significantly different from other models as the group. Specifically, we developed a robust screening framework that integrates statistical evaluation and SHAP-based interpretability to identify "high-quality model parameter combinations" that meet the three criteria. Interestingly, ML models based on LASSO and LR were most often chosen among all ML classifiers. The results indicate that relying solely on average accuracy may obscure consistency issues, while incorporating score differences, mean thresholds, and standard deviation analysis can effectively prevent the selection of overfitting models. This framework may be applied to downstream feature selection, model deployment, patient prognostication and other processes.

Keywords
Machine learning, cross-dataset, generalizable performance, omic

---

Title: Normalization and selecting non-differentially expressed genes improve machine learning modelling of cross-platform transcriptomic data
Author list: Fei Deng[1], Catherine H Feng[1,2], Nan Gao[3,4] ,Lanjing Zhang[1, 4, 5, 6]

Detailed Affiliations
[1]Department of Chemical Biology, Ernest Mario School of Pharmacy, Rutgers University, Piscataway, NJ, USA.;[2] Harvard University, Cambridge, MA, USA;[3] Department of Biological Sciences, School of Arts & Sciences, Rutgers University, Newark, NJ, USA;[4] Department of Pharmacology, Physiology, and Neuroscience, New Jersey Medical School, Rutgers University, Newark, NJ, USA;[5] Department of

Pathology, Princeton Medical Center, Plainsboro, NJ, USA;[6] Rutgers Cancer Institute of New Jersey, New Brunswick, NJ, USA.

Abstract
Normalization is a critical step in quantitative analyses of biological processes. Recent works show that cross-platform integration and normalization enable machine learning (ML) training on RNA microarray and RNA-seq data, but no independent datasets were used in their studies. Therefore, it is unclear how to improve ML modelling performance on independent RNA array and RNA-seq based datasets. Inspired by the house-keeping genes that are commonly used in experimental biology, this study tests the hypothesis that non-differentially expressed genes (NDEG) may improve normalization of transcriptomic data and subsequently cross-platform modelling performance of ML models. Microarray and RNA-seq datasets of the TCGA breast cancer were used as independent training and test datasets, respectively, to classify the molecular subtypes of breast cancer. NDEG ($p>0.85$) and DEG ($p<0.05$) were selected based on the p values of ANOVA analysis and used for subsequent data normalization and classification, respectively. Models trained based on data from one platform were used for testing on the other platform. Our data show that NDEG and DEG gene selection could effectively improve the model classification performance. Normalization methods based on parametric statistical analysis were inferior to those based on nonparametric statistics. In this study, the LOG_QN and LOG_QNZ normalization methods combined with the neural network classification model seem to achieve better performance. Therefore, NDEG-based normalization appears useful for cross-platform testing on completely independent datasets. However, more studies are required to examine whether NDEG-based normalization can improve ML classification performance in other datasets and other omic data types.

Keywords
Machine learning, Feature Selection, Normalization, Transcriptomics, Breast Cancer

---

Title: Improving Differential Expression Analysis of sncRNAs through Family-Level Integration in Paired Samples

Author list: Hukam C. Rawal[1] , Qi Chen[2,3], Tong Zhou[1]

Detailed Affiliations
[1]Department of Physiology and Cell Biology, University of Nevada, Reno School of Medicine, Reno, NV 89557, USA; [2]Molecular Medicine Program, Division of Urology, Department of Surgery, University of Utah School of Medicine, Salt Lake City, UT 84132, USA; [3]Department of Human Genetics, University of Utah School of Medicine, Salt Lake City, UT 84132, USA.

Abstract
Recent techniques like PANDORA-seq have revealed that microRNAs - the most studied small non-coding RNAs (sncRNAs) - constitute only a minor fraction of total sncRNAs, with tRNA-derived (tsRNAs) and rRNA-derived (rsRNAs) species often dominating. These noncanonical sncRNAs, derived from longer RNAs such as tRNAs, rRNAs, and Y RNAs, play important roles in cellular processes, yet their analysis is hindered by sequence heterogeneity and technical noise. A key challenge is integrating multiple sncRNA species from the same parental RNA without losing biological relevance or statistical power. Existing

approaches either sum reads at the family level, sacrificing sequence-specific resolution, or analyze individual species separately, introducing variability due to low counts - limitations that are especially problematic in small-sample or single-subject studies. FUSION_ps (Family-level Unique Small RNA Integration for paired-sample analysis) addresses this by quantifying individual sncRNA species before aggregating them into RNA families for differential expression analysis. This method preserves sequence-level detail while enhancing statistical robustness, making it well-suited for paired-sample and '1-on-1' comparisons. Using the Wilcoxon signed-rank test, FUSION_ps evaluates family-level expression changes by assessing consistent directional shifts among related sncRNA species. The tool was tested on two publicly available, independent lung adenocarcinoma (LUAD) datasets from Cohort1 (19 paired samples) and Cohort2 (48 pairs), following preprocessing with Cutadapt and annotation via the SPORTS pipeline. FUSION_ps identified consistent, patient-specific dysregulation patterns - such as upregulation of the GtsRNA-Gly-CCC family in tumors - with strong correlation across cohorts, demonstrating its ability to uncover meaningful changes missed by traditional group-based analyses. Unlike standard differential expression tools such as DESeq2 and edgeR, which face limitations with sncRNA-specific challenges and small sample sizes, FUSION_ps balances fine-grained resolution with statistical power, making it especially effective for personalized medicine applications such as monitoring treatment responses. FUSION_ps uniquely combines high-resolution quantification with family-level integration, providing a biologically relevant and statistically robust framework. This approach advances sncRNA analysis and supports progress in precision diagnostics and therapeutic development.

Keywords
small non-coding RNAs, differential expression, dysregulation, paired-sample, tsRNA, rsRNA.

Title: AI-Powered Analysis Reveals GATA4's Role in Controlling Cardiac Fibroblast Migration and Fibrosis Post-Myocardial Infarction
Author list: Elena T Zhu1,2, Zhentao Zhang1

Detailed Affiliations
[1]Department of Surgery, Davis Heart and Lung Research Institute, The Ohio State University, Columbus, Ohio 43210, USA; [2]Upper Arlington High School, Upper Arlington, Ohio 43221, USA

Abstract
Cardiac fibrosis, an adaptive yet pathological response to cardiac injury such as myocardial infarction (MI), is driven by the transformation of quiescent cardiac fibroblasts (CFs) into extracellular matrix-producing myofibroblasts (MFs). GATA binding protein 4 (Gata4), a critical transcription factor in heart development, is highly expressed in Gata4$^+$ non-myocytes (CFs, marked by Pdgfr-α but not endothelial marker CD31) following MI.
In vivo, using myofibroblast-specific lineage tracing mouse models (Postn-MerCreMer(iCre):R26tdTomato (tdT) control mice and PostniCre:Gata4fl/fl:R26tdT) and left anterior descending (LAD) coronary artery ligation, we found that Gata4 knockout hearts exhibited significantly greater production of fibrotic proteins (e.g., Collagen I, α-SMA), higher Trichrome staining signals, larger infarct sizes (over 20% larger than controls, $p<0.005$), and reduced contractile function (over 15% reduction, $p<0.05$), demonstrating Gata4's critical role in suppressing post-injury fibrosis.

In vitro, CRISPR/Cas9-mediated Gata4 knockout (Gata4-/-) CFs spontaneously displayed elevated fibrogenic gene expression (e.g., Col1a1, Col1a2, Col3a1, Tgfb1) and enhanced migratory behavior, which we analyzed using artificial intelligence (AI)-powered live cell imaging with Nikon NIS-Elements software. Our AI-based path analysis revealed that Gata4-/- fibroblasts exhibited 32% longer average path lengths (633 μm, $p<0.0001$), 27.5% higher path speeds (0.013 μm/s, $p<0.0001$), 38.7% longer line lengths ($p<0.0001$), and 22.6% higher line speeds ($p<0.0002$) compared to controls, alongside 35.7% larger cell sizes ($p<0.0001$). These AI-driven insights indicate increased migratory capacity, cellular activation, and fibrogenic potential due to Gata4 deficiency, highlighting Gata4's novel regulatory function in fibroblast dynamics.
In conclusion, this study demonstrates that Gata4 is essential for regulating cardiac fibrosis by suppressing fibroblast migration and fibrogenic activity, with AI-powered analysis providing unprecedented precision in quantifying these cellular changes, positioning Gata4 as a promising therapeutic target for cardiovascular disease.

Keywords
GATA4, Cardiac Fibrosis, Fibroblast Migration, AI-Powered Analysis

---

Title: Mechanistic annotation of GWAS loci for circulating fatty acids by single-cell omics and CRISPR screens
Author list: Huifang Xu[1], Haifeng Zhang1, Ge Yu[1], Yitang Sun[1], Elijah Sterling[1,2], Saurav Choudhary[3], Pengpeng Bi[1], Kaixiong Ye[1,3]

Detailed Affiliations
[1]Department of Genetics, University of Georgia, Athens, Georgia, USA;[2]Regenerative Bioscience Center, University of Georgia, 3Institute of Bioinformatics, University of Georgia, Athens, Georgia, USA

Abstract
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, the underlying biological mechanism linking these identified loci to FA metabolism remains largely unclear. Here, we integrate GWAS with single-cell multi-omics and single-cell CRISPR screens to systematically uncover the cellular and molecular mechanisms underlying FA-associated genetic loci. We colocalized GWAS signals for 19 FA traits with six types of multi-omics quantitative trait loci (QTL), including gene expression, protein abundance, DNA methylation, splicing, histone modification, and chromatin accessibility, to identify intermediate molecular phenotypes that mediate the associations between the genetic loci and 19 FA traits. We found that 35% of GWAS loci overlapped with QTL signals for at least one molecular phenotype. Notably, a locus (around genes GSTT1/2/2B) associated with total fatty acids, the percentage of omega-6 polyunsaturated fatty acids (PUFA) in total FAs, and total monounsaturated fatty acids overlapped with QTL signals across all six molecular phenotypes. We analyzed single-cell RNA-seq data of over 100,000 cells from liver tissue to explore the cellular context. We discovered that hepatocyte cell populations, particularly those located in the periportal region, are enriched for genes associated with FA traits. To explore the regulatory function of FA-associated loci, we conducted a single-cell CRISPR screen in over

200,000 HepG2 liver cells, targeting 360 candidate regulatory elements (CREs) from fine-mapped FA trait variants. We identified target genes in cis for 298 CREs, providing a direct map of the regulatory relationship. Our integrative analysis reveals the molecular and cellular mechanisms regulating circulating fatty acid levels, providing mechanistic insights into the genetic architecture of fatty acid metabolism.

Keywords
Fatty acids, GWAS, single-cell multi-omics, single-cell CRISPR screen, Mechanistic annotation

---

Title: Genetic and molecular characterization of long COVID-19-associated Alzheimer's disease and related dementia using multi-omics approaches

Author list: Aishwarya Deengar[1,2,3,] Noah Lorincz-Comi[1,3,4] & Feixiong Cheng[1,3,4]

Detailed Affiliations
[1]Cleveland Clinic Genome Center, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44195, USA; [2] Case Western Reserve University, School of Medicine, Cleveland, OH 44106, USA; [3]Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, OH 44195, USA; [4]Department of Molecular Medicine, Cleveland Clinic Lerner College of Medicine, Case Western Reserve University, Cleveland, OH 44195, USA

Abstract
Recent studies report a bi-directional epidemiological association between Coronavirus disease 2019 (COVID-19) and Alzheimer's disease (AD) incidence [1]. Long COVID (L-COVID), which is a symptomatic state that persists long after SARS-CoV-2 infection, has been hypothesized as a risk factor for AD onset or disease progression because of the neurological damage in confers. However, genetic susceptibility and underlying molecular mechanisms underlying long COVID remain understudied. Further, the extent to which L-COVID and AD share genetic risks and biological pathways will offer effective prevention and treatment approaches in the future [3,4]. We first performed bi-directional univariable Mendelian Randomization (MR) between AD and multiple COVID-related traits included L-COVID, severe COVID, COVID infection, and hospitalized COVID (S/I/H-COVID) using population-scale gene-based association test statistics from genome-wide association studies (GWAS). For each gene, we next applied the Cauchy combination meta-analysis test [2] to the set of COVID phenotypes (MetaCOVID) and tested for their joint association with AD. For genes with evidence of association with MetaCOVID and AD, we tested for over-expression in microglia from brain tissue using single-cell RNA sequence data from a public L-COVID cohort and multiple integrated AD cohorts. Our results do not suggest a causal effect of AD on L-COVID (inverse variance-weighted $P>0.05$) or L-COVID on AD ($P>0.05$). However, seven genetic loci were jointly associated with MetaCOVID and AD at the level Bonferroni significance and included well-known AD-associated genes such as CRHR1, KANSL1, and IL10RB. These genes are broadly involved in genomic pathways related to immune activity/modulation, cell-cycle regulation, and histone acetylation. Using transcriptional profiles, we additionally found that 16 genes including BPTF, CHD3, KANSL1, HLA-B, SLC11A1, and DGKQ were upregulated in both AD and L-COVID cases compared to controls. These genes are key regulators of inflammation, synaptic signaling, cellular stress, and neuronal function and represent promising candidates for future research into the shared biological

pathways and offer potential drug targets for therapeutic development in long COVID19 associated AD and dementia.

Keywords
Alzheimer's Disease, Long COVID

Title: In Silico Design of a Population-Specific mRNA Vaccine Targeting MUC1 for Colorectal Cancer: Focus on Iranian HLA Diversity

Author list: Zarrin Minuchehr[1], Sara Farahbakhsh[2], Raha Mahdavi karimi[3], Hanita Kouzegar[3], Atrin Tofighi[3], Amitis Masumian[3]

Detailed Affiliations
[1]PhD, Head of Department of National Institute of Genetic Engineering and Biotechnology (NIGEB), Tehran, Iran;[2]PhD graduated in biotechnology, Bu-Ali Sina University, Faculty of Agriculture, Department of Biotechnology, Hamedan, Iran;[3]High School Student (Grade 9), Salam Eslam Private High School, Tehran, Iran

Abstract
This study details the computational design of an mRNA vaccine targeting MUC1, a tumor-associated antigen overexpressed in 80% of colorectal cancer (CRC) cases, with specific optimization for the Iranian population[2,9]. Through comprehensive bioinformatics analysis [1,13], we identified the 9-mer peptide epitope SVSDVPFPF (residues 1140-1148 of MUC1) as the optimal candidate based on its exceptional binding affinity to prevalent Iranian HLA class I alleles (IC50 = 23 nM for HLA-B*35:01)[7], cross-reactivity with HLA class II molecules (percentile rank <1 for HLA-DRB1*11:01)[5], and robust immunogenic potential (VaxiJen score = 0.69)[6]. The epitope demonstrated superior antigen processing characteristics, with a proteasomal cleavage score of 1.07 and TAP transport efficiency of 1.2, yielding a total processing score of 0.91. Population coverage analysis using IEDB tools confirmed 100% coverage (95% CI: 99.2-100%) across Iranian subpopulations[11], achieved through strategic selection of high-frequency HLA alleles including HLA-A*24:02 (18.7%), HLA-B*35:01 (17.2%), and HLA-C*04:01 (15.9%). The vaccine construct incorporated multiple optimized elements: a 15-nt 5' UTR derived from clinically validated Moderna platforms, GCCGCCACC Kozak sequence for enhanced translation initiation, IgE signal peptide (MDWTWILFLVAAATRVHS) for proper MHC class I presentation, and a 100-nt poly(A) tail flanked by 3' UTR stability elements[3,10]. Secondary structure prediction using RNAfold 2.6.0 revealed excellent thermodynamic stability (minimum free energy = -52.78 kcal/mol at 37°C) with 62.19% ensemble diversity, suggesting favorable translation efficiency while maintaining structural flexibility for immune recognition[8]. This design offers significant advantages over conventional approaches by targeting consistent MUC1 overexpression rather than patient-specific mutations, enabling broader applicability while maintaining tumor specificity[12]. The vaccine's modular architecture permits future incorporation of additional tumor-associated antigens, and its population-specific HLA targeting addresses regional disparities in cancer immunotherapy access[4]. Future directions include preclinical validation in humanized mouse models expressing Iranian HLA alleles and investigation of synergistic effects with PD-1 inhibitors to counteract CRC's immunosuppressive microenvironment[14]. This work establishes a robust framework for developing cost-effective, population-tailored mRNA vaccines against

solid tumors, with particular relevance for regions where personalized neoantigen vaccines remain logistically challenging.

Keywords
mRNA vaccine, Colorectal cancer , MUC1, HLA diversity, Immunoinformatics Epitope prediction

---

Title: Brain Tumor Detection And Classification Using Machine Learning Algorithms
Author list: Raza Tasawar

Detailed Affiliations
[1]Department of Computer Science, City University of Science & Information Technology, Peshawar, Pakistan

Abstract
Brain tumors pose a significant health challenge worldwide, necessitating accurate and efficient detection for timely intervention. This research focuses on leveraging machine learning techniques for brain tumor detection, specifically utilizing Support Vector Machines (SVM) and Logistic Regression algorithms. The primary objective of this study is to compare the performance of these two algorithms in classifying three major types of brain tumors: glioma, meningioma, and pituitary adenoma. The research dataset comprises a collection of medical images, each associated with one of the three tumor types. Feature extraction techniques are applied to capture relevant information from these images, transforming them into a format suitable for machine learning analysis. The extracted features are then used as input for both the SVM and Logistic Regression models. The experimental results demonstrate promising accuracy for both algorithms. The SVM model achieves a training score of 0.936 and a testing score of 0.840, while the Logistic Regression model attains a training score of 1.0 and a testing score of 0.807. These scores indicate the models' ability to effectively differentiate between tumor types based on the extracted features. Furthermore, the research contributes to the medical field by providing insights into the performance of machine-learning algorithms for brain tumor classification. The higher training score achieved by the Logistic Regression model suggests that it may have a stronger capability to fit the training data. On the other hand, the SVM model demonstrates relatively robust generalization to unseen data, as indicated by its competitive testing score.

In conclusion, this work underscores the potential of machine learning algorithms, specifically SVM and Logistic Regression, in the accurate detection of glioma, meningioma, and pituitary adenoma brain tumors. The comparative analysis of these algorithms not only sheds light on their classification capabilities but also highlights their suitability for different stages of a diagnostic pipeline. This research contributes to the ongoing efforts to enhance medical imaging analysis and brain tumor diagnosis, ultimately aiding medical professionals in making informed decisions and improving patient outcomes.

Keywords
Meningioma tumor, Glioma tumor, Pituitary tumor, and No tumor, Machine learning, Support Vector Machine, Logistic Regression

Title: Meningioma tumor, Glioma tumor, Pituitary tumor, and No tumor, Machine learning, Support Vector Machine, Logistic Regression
Author list: Faraz Rabbani[1]

Detailed Affiliations
[1]Chief Executive Officer, Peek, San Francisco, CA, USA

Abstract
We present PRISM (Perturbation Response Integration of Single-cell Measurements), a comprehensive collection of over 70 harmonized public single-cell datasets comprising more than 10 million cells with standardized metadata annotations. This resource integrates diverse perturbation experiments across multiple organisms (primarily human and mouse), cell types (including cancer, stem, T cells, and neural cells), and CRISPR modalities (KO, CRISPRi, CRISPRa). The PRISM collection addresses a critical challenge in computational biology: the lack of standardized, large-scale perturbation data necessary for building predictive models of cellular behavior.
We implemented rigorous preprocessing protocols to ensure consistency across datasets, including the standardization of gene expression values, the harmonization of metadata, and quality control filtering. All datasets are provided in .h5ad format with uniform annotation structures to facilitate seamless integration.
Leveraging this curated resource, we developed a preliminary AI model using conditional flow matching techniques and gene embeddings (GenePT) to predict cellular responses to genetic perturbations. Our model architecture employs neural ordinary differential equations to simulate the trajectory from unperturbed to perturbed states, conditioned on gene embeddings that
capture the biological context of the perturbation target. Validation across diverse cell types demonstrates promising accuracy in predicting expression changes following perturbation.
This work represents a significant step toward building a virtual cell - a comprehensive computational model capable of simulating cellular responses to arbitrary perturbations. By integrating data across diverse experimental conditions and developing predictive models, we establish a foundation for in silico prediction of cellular behavior. Such capabilities have profound implications for understanding gene regulatory networks, drug development, and personalized medicine approaches.
The PRISM dataset and model code are publicly available to foster community-driven advancement in this field. Future work will focus on expanding dataset coverage, refining model architectures to better capture cell type-specific responses, and developing interfaces to make virtual cell predictions accessible to researchers without computational expertise.

Keywords
Perturbation, Single-cell, CRISPR, Harmonized datasets, Generative modeling, Virtual cell

---

Title: Multi-omic quantitative trait loci link tandem repeat size variation to gene regulation in human brain
Author list: Ya Cui1,2, Frederick J. Arnold2, Leslie M. Thompson3, Albert R. La Spada2, Wei Li1

Detailed Affiliations
[1]Department of Biological Chemistry, University of California, Irvine, Irvine, CA, USA;[2]Departments of Pathology & Laboratory Medicine, Neurology, Biological Chemistry, and Neurobiology & Behavior, University of California Irvine, Irvine, CA, USA;[3]Departments of Psychiatry and Human Behavior, Neurobiology and Behavior, and Biological Chemistry, University of California, Irvine, Irvine, CA, USA

Abstract
Tandem repeat (TR) size variation is implicated in ~50 neurological disorders, yet its impact on gene regulation in the human brain remains largely unknown. In the present study, we quantified the impact of TR size variation on brain gene regulation across distinct molecular phenotypes, based on 4,412 multi-omics samples from 1,597 donors, including 1,586 newly sequenced ones. We identified ~2.2 million TR molecular quantitative trait loci (TR-xQTLs), linking ~139,000 unique TRs to nearby molecular phenotypes, including many known disease-risk TRs, such as the G2C4 expansion in C9orf72 associated with amyotrophic lateral sclerosis. Fine-mapping revealed ~18,700 TRs as potential causal variants. Our in vitro experiments further confirmed the causal and independent regulatory effects of three TRs. Additional colocalization analysis indicated the potential causal role of TR variation in brain-related phenotypes, highlighted by a 3′-UTR TR in NUDT14 linked to cortical surface area and a TG repeat in PLEKHA1, associated with Alzheimer's disease.

Keywords
Tandem repeats, QTL, Brain, ALS, Alzheimer's disease

Title: Tokenvizz: GraphRAG-Inspired Tokenization Tool for Genomic Data Discovery and Visualization
Author list: Çerağ Oğuztüzün[1], Zhenxiang Gao[1], Jing Li[1*],Mehmet Koyutürk[†],Rong Xu[‡]

Detailed Affiliations
[1]Case Western Reserve University Cleveland, Ohio, USA

Abstract
Interpreting complex genomic relationships and predicting functional interactions remain key challenges in biomedical research. Traditional sequence-based methods often lack interpretability whichlimitstheexplorationofgenomiclanguagemodelpredictions. To address this gap, we introduce Tokenvizz, a GraphRAG-inspired tool that transforms genomic sequences into intuitive graph representations, where DNA tokens become nodes connected by edges weighted by attention scores derived from genomic language models. This novel approach translates genomic sequences into structured graph visualizations that reveal latent token relationships that are difficult to interpret through purely sequential methods. Tokenvizz provides an integrated pipeline that includes data preprocessing, graph construction fromtokenizedsequences,andaninteractive web-based visualization interface. Users can dynamically adjust edge weight thresholds, perform position-based searches, and examine contextual sequence information interactively. This facilitates intuitive, multi-resolution analysis of genomic sequences and enhances the interpretability and exploratory capabilities of genomic language models. To validate Tokenvizz, we applied its graph representations to promoter-enhancer interaction prediction using a Graph Convolutional Network (GCN) on six datasets from the GUE+ benchmark. Tokenvizz consistently outperformed existing

sequential deep learning models such as DNABERT2 and Nucleotide Transformer, demonstrating the utility of attention-derived graph structures for genomic prediction tasks. Byeffectively bridging attention-based genomic language modeling andinteractive graphvisualization, Tokenvizz offers researchers a visualization tool for exploratory genomic analyses. Future work will explore integrating external genomic annotation databases to further strengthen its interpretability and utility for genomics research.

Keywords
genomic language models, graph visualization, DNA sequence analysis, attention mechanism, regulatory elements, genomic interpretation

---

Title: Pre-symptomatic detection and alarming of COVID using smartwatch data

Author list: Tejaswini Mishra[1], Meng Wang1,2, Ahmed A. Metwally[1], Gireesh K. Bogu[1], Andrew W. Brooks[1], Amir Bahmani[1], Arash Alavi[1], Alessandra Celli[1], Emily Higgs[1], Orit Dagan-Rosenfeld[1], Bethany Fay[1], Susan Kirkpatrick[1], Ryan Kellogg[1], Michelle Gibson[1], Tao Wang[1], Erika M. Hunting[1], Petra Mamic[1], Ariel B. Ganz[1], Benjamin Rolnik[1], Ekanath Srihari Rangan[1],Qiwen Wang1, Kexin Cha[1], Peter Knowles[1], Rajat Bhasin[1], Shrinivas Panchamukhi[1], Diego Celis[1], Tagore Aditya[1], Alexander Honkala[1], Arshdeep Chauhan[1], Jessi W. Li[1], Caroline Bejikian[1], Vandhana Krishnan[1], Lettie McGuire[1], Amir A. Alavi[1], Xiao Li3, Michael P. Snyder[1]

Detailed Affiliations
[1]Department of Genetics, Stanford University, Stanford, CA, USA; [2]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA; [3]The Center for RNA Science and Therapeutics, Case Western University, Cleveland, OH, USA.

Abstract
We used consumer smartwatches for the pre-symptomatic detection of coronavirus disease 2019 (COVID-19) in the work of Mishra, Wang, et.al, Nat. Biomed. Eng (2020). Both offline and online anomaly detection algorithms were developed for detecting physiological alterations due to the COVID infection. This work demonstrated the potential ability that activity tracking and health monitoring via consumer wearable devices can be used for the large-scale, real-time detection of respiratory infections, often pre-symptomatically. We further developed a real-time smartwatch-based alerting system for detecting the COVID-19 and other stress events in the work of Alavi, Bogu, Wang, et al., Nat. Med (2022). In this work, the alarming algorithms were developed to detect aberrant physiological and activity signals (heart rates and steps) associated with the onset of early infection. Our work showed that a real-time alerting system can be used for early detection of infection and other stressors and employed on an open-source platform that is scalable to millions of users.

Keywords
COVID, smartwatch, pre-symptomatic detection, real-time alarming

---

Title: HAT: Automated Pathologist-Guided Label Transfer for Multi-Study, Multi-Sample, and Multi-Status Spatial Omics Data

Author list: Jing Huang[1,2], Michael P. Epstein[1,2 *], Jian Hu1,2 *

Detailed Affiliations

[1]Department of Human Genetics, School of Medicine, Emory University, Atlanta, GA, 30322, USA;[2]Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA, 30322, USA.

Abstract

Background: Recent technological advances in spatial omics have enabled the diverse omics profiling while preserving the native tissue contexts. These techniques, including transcriptomics, proteomics, metabolomics, and chromatin accessibility, provide comprehensive molecular measurements of the tissue. Data from these technologies are often complemented by high-resolution histology images of the same tissue section to examine the cell morphology. With the growing accessibility and affordability of spatial omics techniques, many studies are generating large datasets comprising multiple sections collected from multiple samples, profiling a wide range of disease conditions and developmental stages. While the data richness enhances biological insights, it also presents significant challenges for the joint analysis of multi-sample spatial omics data. One critical task in the integrative analysis is to identify the shared tissue regions, which allows for examining region-specific heterogeneity across samples over different disease stages. As visual inspection of biopsy specimens through histopathology remains a gold standard for disease diagnosis, some studies have generated pathologists' annotations of tissue regions based on associated cell morphologies. However, only a limited number of tissue sections are examined and annotated by pathologists due to the substantial manual efforts required. Therefore, analytic methods that can automatically extrapolate pathologists' annotations from limited reference samples to a large number of unannotated samples is in urgent need.

Method: Here, we present Hierarchical Annotation Transfer (HAT), a supervised machine learning method that can transfer pathologists' annotations to multiple spatial omics datasets collected from different samples and studies. By projecting all samples into a common domain space, HAT starts by constructing a hierarchical tree to delineate the organization of tissue regions. Based on the tree structure, it identifies features to separate tissue regions from annotated reference and then performs automatic label transfer by mapping the unlabeled spatial omics data to the tree following a hierarchical order. All the annotated tissue sections are then assembled into a tissue atlas, enabling region-specific comparisons. HAT further provides a score to assess the cross sample heterogeneity level of each tissue region, with which we can pinpoint the most disease-relevant regions worthy of increased attention in disease diagnosis.

Result: The performance of our method has been evaluated on various spatial omics datasets encompassing different platforms, disease conditions and developmental stages, including breast cancer, prostate cancer, mouse brain, chicken heart, and human tonsil. Our method outperforms existing methods, including Seurat V5, SCGP, and STELLAR, by transferring annotations with high accuracy across all assessed datasets. Additionally, it automatically identifies tissue regions most affected by diseases, shedding light on pivotal tissue hallmarks for disease diagnosis.

Keywords

statistical genomics, spatial omics, pathologist-guided label transfer, tumor microenvironment, machine learning.

Title: Emergent mechanics of living cells and its applications
Author list: D.-S. Guan,[1,2] Y.-S. Shen,[1] R. Zhang,[1] P.-B. Huang,[3] P.-Y. Lai,[4] Penger Tong[1],

Detailed Affiliations
[1]Department of Physics, Hong Kong University of Science and Technology, Hong Kong; [2]Institute of Mechanics, Chinese Academy of Sciences, Beijing, China;[3]Division of Life Science, Hong Kong University of Science and Technology, Hong Kong;[4]Department of Physics, National Central University, Taoyuan City 320, Taiwan

Abstract
Living cells exhibit unique mechanical properties that blend fluid- and solid-like behaviors across various spatial and temporal scales. Despite extensive experimental studies revealing a wide range of viscoelastic behaviors, a comprehensive model of cell mechanics remains elusive. For instance, the Young's modulus E of living cells can vary dramatically—by three orders of magnitude (0.1–100 kPa)—depending on the samples and experimental techniques used. Current theoretical models often oversimplify the complexities of living cell mechanics, leading to inconsistent results. In this presentation, I will share findings from atomic force microscopy (AFM) measurements of stress relaxation and force indentation across 10 cell types, including epithelial, muscle, neuronal, blood, and stem cells [1]. Our unified quantitative description of the relaxation modulus E(t) reveals an initial exponential decay followed by a long-time power-law decay, along with a persistent modulus. These components of E(t) provide a detailed mechanical profile linked to the hierarchical structure and active stress of living cells. This work establishes a robust framework for characterizing the mechanical state of living cells and exploring their physiological functions and disease states.

Keywords
Systems Biology, Self-organization in living systems

Poster Session II
August 4th
5:00 PM – 6:00 PM
Room: First floor Atrium

Title: Fully Automated Real-Time Approach for Human Temperature PredictionBased Thermal Skin Face Extraction using Deep Semantic Segmentation
Author list: Adil Al-Azzawi 1*

Detailed Affiliations
[1]University of Diyala, College of Science, Computer Science Departmenet

Abstract
Covid-19 or Coriona virus pandemic is a disease that infects every country on the earth, which was announced in 2020 by the World Health Organization. Early disease intervention improves the effectiveness of medical treatment and enhances the possibility that patients may live longer without needing intensive care from hospitals and other facilities. The amount of life-saving treatment that hospitals can provide is limited, therefore it's important to stay healthy. However, liver function loss in Cov-19 patients is more common than in non-Cov-19 individuals. According to study, Covid-19 pneumonia has many clinical features with other types of pneumonia. As a result, it is possible that Cov-19 illness may eventually infect people, making it a severe condition. One sign of the sickness is a rising body temperature. Non-contact thermal imaging of human temperate is one of the most important methods for obtaining temperature with some lack of performance. However, most of these devices are private and not available for use, and some belong to for-profit institutions. In this paper, a real-time approach for human temperature prediction and tracking is proposed based on extracting the thermal face skin temperature. A deep semantic segmentation approach is proposed to fully automated thermal skin binary mask prediction. The predicted binary masks are then projected on the original thermal video frames to extract the thermal skin face which are used to calculate the average face temperature. The proposed deep semantic segmentation model is fully trained on a thermal skin binary mask that have automatically generated using my first model which is a fully automated unsupervised learning approach for camera calibration and thermal skin binary mask extraction. Speaking Faces is used as the main dataset for this research. The proposed model shows a very high efficiency human temperature tracking using thermal videos only comparing with the Ground Truth.

Keywords
Supervised learning approach, Deep Learning, Semantic Segmentation, COVID-19, Human Temperature Tracking, Machine Learning.

Title: Platform and Preprocessing Effects on DNA Methylation-Based Aging Estimates: A Comparison of Illumina EPIC Arrays
Author list: Carson Richardson[1], Joseph P. McElroy[1], Peter G. Shields[2], Min-Ae Song[3,4]

Detailed Affiliations
[1] Center for Biostatistics, Department of Biomedical Informatics, College of Medicine, The Ohio State University Wexner Medical Center, Columbus, OH, USA;[2] Comprehensive Cancer Center, The Ohio State University and James Cancer Hospital, Columbus, OH, USA;[3] Division of Environmental Health Sciences, College of Public Health, The Ohio State University, Columbus, OH, USA;[4] Center for Tobacco Research, The Ohio State University Comprehensive Cancer Center, Columbus, OH, USA

Abstract
The Illumina Infinium HumanMethylation (HM) BeadChip arrays have been widely used to predict epigenetic clocks, measured by DNA methylation-based age estimates. Given that aging is the most significant risk factor for most cancers, and age-related epigenetic alterations are key contributors to disease pathogenesis, epigenetic aging has drawn significant attention in cancer research. Over the past 18 years, this platform has evolved from earlier generations (HM27K, HM450K, EPICv1) that were used to develop the original clocks to the recently released EPICv2 array. Although DNA methylation profiles show high concordance across array versions, the fidelity and robustness of epigenetic clock estimates, particularly under different raw data pre-processing options, remain understudied across the versions.
Genome-wide methylation was profiled from lung, blood, and nasal tissues of 91 healthy adults (45 never-smokers, 25 smokers, and 21 e-cigarette users). Samples were assayed on EPICv1 (n = 62), EPICv2 (n = 22), or both (n = 7). Raw IDATs were processed to beta values using the SeSAME package and estimates were obtained via Horvath's New DNA-Methylation Age Calculator for Horvath, GrimAge, PhenoAge, and DNAm-telomere length (DNAm-TL) derived methylation aging estimates and their acceleration. Twelve sets of estimates were generated per clock by crossing three probe-handling strategies - (i) version-specific, (ii) collapsed EPIC v1+v2 probes, (iii) EPICv1 probes lifted to v2 (mLiftOver) - with three preprocessing tiers: none, detection-p < 0.01 masking, and masking + NA exclusion. Pearson correlations and coverage statistics evaluated agreement within and between arrays.
EPICv2 retained ≥ 95% of clock CpGs and yielded stronger replicate correlations than EPICv1 for most clocks ($r > 0.80$, PhenoAge $r \approx 0.66$). In EPICv1, Horvath-mAge showed the highest raw concordance ($r = 0.96$); for cross-platform samples, simple probe merging depressed Horvath correlations ($r < 0.70$) whereas mLiftOver partially restored them. Detection-p filtering minimally affected correlations, while additional NA removal modestly improved DNAmTL. Probe exclusion reduced usable CpGs to 44-89%. Within EPICv2, clustering was evident by tissue, but no systematic clustering was evident by sex or smoking status. EPICv2 supports accurate implementation of established epigenetic clocks and adds tissue-discriminatory power, yet clock-specific sensitivity and probe-set harmonization remain critical when integrating legacy EPICv1 data. These findings suggest that platform choice, probe-

handling, and preprocessing decisions can substantially alter epigenetic clock outputs and that these decisions require methodologically consistent consideration to avoid misinterpretation of biological aging signals and downstream cancer-related insights.

Keywords
Illumina EPIC Arrays; Epigenetic Aging; DNA methylation; Data Processing

---

Title: Optimizing bioinformatic workflows for extracting usable gene expression data from clinical RNAseq tumor fusion panels

Author list: Xiaokang Pan[1], Yi Seok Chang[1], Ryan Stevens[1], Ashley Patton[1], Matthew Avenarius[1, 2], Matthew Hunt[1], Nehad Mohamed[1], Daniel Chappell[1], Weiqiang Zhao[1, 2], Dan Jones[1, 2, 3, *]

Detailed Affiliations
[1]James Molecular Laboratory at Polaris, The Ohio State University Wexner Medical Center, Columbus, OH 43240, USA; [2]Department of Pathology, The Ohio State University Wexner Medical Center, Columbus,

OH 43210, USA; [3]The Ohio State University Comprehensive Cancer Center, James Cancer Center and Solove Research Institute, Columbus, OH 43210, USA.

Abstract
Targeted RNA sequencing (RNAseq) panels to detect oncogenic gene fusions have been widely used in clinical molecular laboratories. However, the use of these panels to provide diagnostically useful data for tumor classification in fusion-negative cases has been limited to date. Using expression data in smaller panels requires careful consideration on the bioinformatic methods for sequence read counting, gene normalization, clustering algorithms and data presentation. To define a clinical-grade pipeline for this application, we present comparative data on these methods for differential gene expression (DGE) using ~200 gene RNAseq fusion panels as a model for limited gene sets. We compared six widely used methods of read counting for their speed, precision and accuracy, with the featureCounts method selected as the most rapid and robust. To find the optimal methods for adjusting gene levels from different samples prior to DGE, we compared the effects on gene expression distributions of normalizing using housekeeping, highly expressed, consistently expressed and all genes in two different datasets of carcinomas vs soft tissue tumors (for lineage assignment) and low-grade vs high-grade sarcomas (for tumor grade), respectively. Normalization using 5 genes with highly consistent expression had optimal suitability. To investigate optimal methods for inferring tumor lineage, principal component analysis (PCA), T-SNE and heatmap-clustering were compared using the same datasets. We present a model pipeline and discussion on the methods best suited for extracting usable gene expression data from targeted RNAseq tumor panels to maximize the clinical utility of these assays.

Title: Deep Learning Models for Cell Cycle Phase Prediction from Single-Cell RNA Sequencing Data
Author list: Halima Akhter[1,2], Debra Piktel[2], Laura F. Gibson[3], Gangqing Hu[2*], Donald Adjeroh[1*]

Detailed Affiliations
[1]Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV, USA; [2]Department of Microbiology, Immunology & Cell Biology, School of Medicine, West Virginia University, Morgantown, WV, USA. 3School of Medicine, West Virginia University, Morgantown, WV, USA.

Abstract
Accurate prediction of cell cycle phases is essential for mitigating confounding effects in singlecell RNA sequencing (scRNA-Seq) analysis and for understanding how certain diseases (e.g., cancer) develop and respond to treatment. In this study, we evaluated both traditional machine learning algorithms (including AdaBoost, Random Forest, and LightGBM) and deep learning approaches (including dense neural networks [DNN3 and DNN5], a convolutional neural network [CNN], a hybrid CNN-Dense model, a feature embedding approach, and ensemble models) for cell cycle phase prediction from scRNA-seq data. Models were trained using consensus predictions derived from four cell cycle analysis tools (CellCycleScore, ccAF, Revelio, and Tricycle) applied to unlabeled scRNA-seq data generated from two human leukemia cell lines. Performance was subsequently validated using two public scRNA-seq datasets: one for human embryonic stem cells (GSE64016; 246 cells) and the other for human osteosarcoma cells (GSE146773; 1,151 cells), both with experimentally verified cell cycle phase annotations. Our evaluation revealed that deep learning

models consistently outperformed traditional methods, showing accuracies ranging from 51.01% to 70.85% versus 44.11% to 61.00% on the GSE64016 dataset, and 72.00% to 75.35% versus 44.11% to 75.35% on the GSE146773 dataset. Moreover, we compared the best-performing models (DNN3, Embedding3TML, and Top 3 Decision Fusion (Top-3 D.F.)) with several publicly available tools, including Cyclum, SC1CC, and the four used for generating the training consensus. On the GSE64016 dataset, DNN3 achieved the highest accuracy (70.85%), outperforming all existing tools, including Revelio (68.20%). In the larger GSE146773 dataset, DNN3 and Embedding3TML both attained the top accuracy of 75.35%, followed by Top-3 D.F. (75.00%), while other tools ranged between 38.2%–68.34%. These results underscore the potential of deep learning models and ensemble strategies for robust and accurate cell cycle phase prediction in scRNA-seq data.

Keywords
scRNA-Seq; cell cycle phases; machine learning; deep learning; ensemble learning; performance evaluation

---

Title: LiverHomo: Integrating Single-Cell Transcriptome Landscapes of Human Liver Diseases to Explore Cell-Specific Protein-Protein Interaction Networks

Author list: Ziyu Shi[1], Zhiyuan Song[1], Haiqing Zhao[1]
Detailed Affiliations
[1]Department of Biochemistry and Molecular Biology; Sealy Center for Structural Biology and Molecular Biophysics, University of Texas Medical Branch, Galveston, TX, USA

Abstract
Liver has a unique microenvironment that promotes local and systemic immune tolerance by expressing anti-inflammatory mediators and T cell inhibitory signals. This distinctive "tolerance effect" not only makes the liver vulnerable to infections from pathogens but also poses challenges in understanding the complex mechanisms underlying liver cancer pathogenesis. In this work, we explore the liver's microenvironment and immunoregulatory landscape using state-of-the-art omics data. We introduce LiverHomo, a comprehensive atlas constructed from publicly available single-cell RNA sequencing (scRNA-seq) data, encompassing 1.16 million cells from 320 liver patients. To our knowledge, this is the largest scRNA-seq database focused on human liver diseases. Through rigorous quality control and a unified analytical pipeline, LiverHomo enables diverse analyses—including differential gene expression, functional enrichment, cellular crosstalk, and pseudo-time trajectory mapping—offering deep mechanistic insights into healthy, diseased, and developmental liver states. Moreover, with our recent advances in genome-wide protein-protein interactions (PPIs) studies, PrePPI-AF and ZEPPI, we further identified cell-type–specific PPIs in both normal and diseased liver cells. This integrative approach has the potential to transform our understanding of liver pathology and guide therapeutic innovation. Overall, LiverHomo provides a critical resource for advancing liver biology, disease modeling, and driving progress in precision medicine.

scRNA-seq: To systematically profile the liver cellular ecosystem across diseases, we collected samples from healthy liver, alcohol-associated liver disease (ALD), non-alcoholic fatty liver disease (NAFLD), non-alcoholic steatohepatitis (NASH), viral hepatitis (HBV/HCV), cirrhosis, and hepatocellular carcinoma (HCC), sourced from the Gene Expression Omnibus (GEO). These datasets were curated via a standardized pipeline based on Scanpy, including quality control, batch correction, dimensionality reduction, and

unsupervised clustering. Cell types were annotated by automated classification and manual curations of established marker genes.
Protein-Protein Interaction: PrePPI-AF begins with AlphaFold-predicted monomer structures and systematically generates pairwise intermolecular domain–domain models by identifying structurally similar interactions across species. These models are then evaluated using a Bayesian-trained classifier and the co-evolutionary scoring framework ZEPPI. Together, PrePPI-AF enables the prediction of 1.3 million high-confidence PPIs in humans.
LiverHomo contains 1.16 million single cells collected from 320 patients, making it the largest scRNA-seq resource dedicated to human liver disease to date. Using this dataset, we built cell-type–resolved protein–protein interaction (PPI) networks by combining high-confidence structural predictions from our PrePPI-AF framework with detailed cell-type annotations and
differential-expression profiles. The resulting pipeline is both generalizable and portable, enabling seamless integration with other single-cell datasets to investigate PPI networks across diverse biological contexts.

Keywords
Human liver diseases, single-cell RNA sequencing (scRNA-seq), database, protein-protein interaction (PPI), coevolution

---

Title: EntroPPI: an entropy-aware AI model for quantitative protein-protein interaction predictions
Author list: Zhiyuan Song[1], Ziyu Shi[1], Haiqing Zhao[1]

Detailed Affiliations
[1]Department of Biochemistry and Molecular Biology; Sealy Center for Structural Biology and Molecular Biophysics, University of Texas Medical Branch, Galveston, TX, USA.

Abstract
Protein-protein interactions (PPIs) are fundamental to understanding cellular processes and disease mechanisms, thereby crucial in developing protein or drug therapeutics. Existing AI-based protein and PPI models predominantly rely on protein sequences and static protein structures as their learning resources, neglecting the protein dynamic flexibilities that can count for the entropic contributions to the protein binding. To overcome these limitations, we introduce EntroPPI, a graph-based generative deep-learning model that is designed to address the missing entropy factor in the current protein model design. By representing protein complexes as graphs at both residue and atomic levels, EntroPPI encodes critical interaction features, enabling a graph-based encoder to learn rich, entropy-aware embeddings. Our framework bridges the gap between current sequence or structure-based PPI models and the inherent dynamic nature of protein recognition, advancing the capabilities of AI-based protein modeling.

Keywords
Protein–protein interactions (PPIs), Deep learning, Graph neural networks, Binding affinity prediction

---

Title: Protein–protein interactions (PPIs), Deep learning, Graph neural networks, Binding affinity prediction

Author list: Ruopeng Wu[1, *], Yichao Chen[1,*], Feng-Yu (Leo) Yeh[1], Jie Zheng[1], Nikki Bonevich[1], Alexander D Diehl[2], William D Duncan[3], Laura Barisoni[4], Jimmy Phuong[9], Avi Z Rosenberg[5], Jeff Hodgin[1], Sanjay Jain[6], Ravi Iyengar[7], and the Kidney Precision Medicine Project, Bruce W. Herr II[8], , Yongqun Oliver He[1]

Detailed Affiliations

[1]University of Michigan, Ann Arbor, Michigan, USA; [2]University at Buffalo, State University of New York, Buffalo, NY, USA; [3]University of Florida, Gainesville, FL, USA; [4]Duke University, Durham, North Carolina, USA; [5]Johns Hopkins University School of Medicine, Baltimore, MD, USA; [6]Washington University in St. Louis, St. Louis, MO, USA; [7]Icahn School of Medicine, Mount Sinai, New York, NY, USA; [8]Indiana University, Bloomington, IN, USA; 9University of Washington, Seattle, WA, USA.

Abstract

Precision medicine for kidney disease requires the integration of heterogeneous clinical, pathological, and molecular data to uncover mechanistic insights and enable personalized therapies. The KPMP (Kidney Precision Medicine Project) and HuBMAP (Human Biomolecular Atlas Program) compiled extensive single-cell transcriptomic datasets, providing an opportunity for integrative analysis. By merging these datasets, we aim to compare gene expression across populations with diverse healthy and disease kidney conditions and identify gene markers contributed to the kidney diseases.

An ontology-based approach was used to identify homologous variables and hierarchical cell types for dataset integration, using raw data from KPMP and HuBMAP. There were 35 variables found in KPMP data and 25 variables in HuBMAP data, and 12 of which are homologous for shared comparisons. Cell types used in KPMP and HuBMAP are annotated using the Cell Ontology (CL) at different levels of granularity. For example, 'epithelial cell of proximal tubule' was used in KPMP, and 'epithelial cell of proximal tubule segment 1 (or 2 or 3) was used in HuBMAP. Based on the CL, the deeper cell type class were subsumed as subclasses of the parent cell type, in which variables and cell types can be integrated at a specific granularity based on the ontology.

After quality control analysis, the cell-level gene expression profiles from the integrated KPMP and HuBMAP datasets were systematically analyzed. We studied the expression profiles of those gene markers as annotated by HuBMAP reference kidney study or KPMP kidney disease research communities. Specifically, Acute Kidney Injury (AKI) and Chronic Kidney Disease (CKD) labels were classified using the KPMP community categories (https://www.kpmp.org/for-clinicians) and their profiles compared with the healthy controls, primarily from the HuBMAP dataset. SPP1 emerged as a significantly differentially expressed gene in AKI patient samples compared to Healthy reference cells. Many new statistically significant genes not previously labeled as gene markers were identified. For example, PDE4D and ZBTB20 were highly expressed in both AKI and CKD populations. The roles of these newly identified genes in kidney disease mechanisms deserve further investigation.

Overall, this study demonstrates a novel ontology-guided strategy to integrate two separate big data resources for advanced research and scientific insight exploration.

Keywords

Kidney Precision Medicine Project (KPMP) Human BioMolecular Atlas Program (HuBMAP), Ontology-Based Data Integration, Acute Kidney Injury, Chronic Kidney Injury, Single-Cell Gene Expression

Title: Optimization of Recombinant Biomanufacturing through Mutual Information-Based GRN Topology Discovery

Author list: Ridhi Gutta[1]

Detailed Affiliations
[1]Department of Science, Curabitrix Labs LLC, Ashburn, VA, USA

Abstract
In order to resolve crucial global issues, the widespread application of genetic engineering at an industrial level is key. However, the majority of synthetically engineered strains fail at the industrial level due to disruptions in gene regulation. This stems from a lack of understanding and usage of gene regulatory networks (GRNs), which control cellular processes and metabolism. Effective manipulation of GRNs can improve product yield and functionality significantly. However, current GRN inference tools are extremely slow, inaccurate, and incompatible with industrial scale processes, because of which there are no complete expression based GRNs for any organism. This research proposes a novel computational system, GEMINI, to enable fast GRN inference for integration into industrial scale pipelines. GEMINI consists of two main parts. First, we create a novel mutual information algorithm that replaces traditional sequential inference methods, ensuring compatibility with parallel processing. Second, we integrate a novel GNN architecture based on spectral convolution to efficiently learn global and local regulatory structures. On in silico benchmarks, GEMINI outperforms all industry leaders, achieving a nearly 300% increase in AUPRC compared to the industry leading method, GENIE3. GEMINI also reduced computing time by a factor of 9.5 and was able to perform on a classroom GPU. When applied on a real E. coli dataset, GEMINI not only recovered 98% of existing interactions, but discovered 468 novel candidate interactions, constructing the most complete expression based GRN of E. coli to date, providing a novel biological blueprint for genetic engineers to use at the industrial level.

Keywords
Gene regulatory network (GRN), mutual information, recombinant biomanufacturing, graph neural network (GNN), E. coli

---

Title: Systems-Level Computational Engineering of a Multi-Epitope mRNA Vaccine for Middle East Respiratory Syndrome Coronavirus

Author list: Ridhi Gutta[1]

Detailed Affiliations
[1]Department of Science, Curabitrix Labs LLC, Ashburn, VA, USA

Abstract
Middle East Respiratory Syndrome (MERS) is a deadly respiratory disease caused by a COVID-19 relative known as MERS-CoV, which is associated with an extremely high mortality rate. In the status quo, there

are no approved vaccines or antiviral drugs to prevent MERS-CoV infection. Supportive care is the only method to manage infection and often results in poor outcomes, necessitating the development of novel prophylactic measures. In this study, an immunoinformatics approach was applied to predict a multi-epitope mRNA vaccine candidate targeting the glycoprotein (S protein) of MERS-CoV. NBCI was used to obtain the viral sequences from multiple outbreaks, which were then screened for antigenicity, allergenicity, toxicity, B-cell epitopes, CD8 + T lymphocytes (CTL), and CD4 + T lymphocytes (HTL). These epitopes were used in the construction of the vaccine with the addition of UTR caps and a poly-A tail to the mRNA construct. Molecular docking with TLR-4 and molecular dynamics simulations were employed to validate the stability of the binding complex. Additionally, the secondary and tertiary structures were modeled and optimized. The vaccine's molecular weight was 30380.95 kDA, and its estimated pI was 4.60. The vaccine was also shown to elicit a robust immune response through computer simulation, suggesting that the designed mRNA construct could be an effective and promising vaccine candidate to proceed to laboratory and clinical trials.

Keywords
Middle East Respiratory Syndrome (MERS), MERS-CoV, immunoinformatics, mRNA, vaccine

---

Title: Risk of Adverse Events Associated With GLP-1 Receptor Agonist Use Among COVID-19 and Non-COVID-19 Patients: Evidence From a Large EHR Dataset

Author list: Wei Du[1], Yun Han[2], Jinju Li[1], Jie Zheng[1], Chuan Zhou[1], Lili Zhao[1,3], Yongqun Oliver He[1]

Detailed Affiliations
[1]University of Michigan, Ann Arbor, Michigan, USA; [2]Precision AQ, New York, USA; [3]Northwestern University, Chicago, IN, USA.

Abstract
Background: Glucagon-like peptide-1 receptor agonists (GLP-1 RAs) are widely used in type 2 diabetes for glycemic control and offer additional cardiometabolic benefits (e.g. weight loss). However, concerns remain regarding their potential adverse events (AEs), including gastrointestinal symptoms and risks of pancreatitis, thyroid cancer, and other organ-specific events. Whether these risks differ in patients with COVID-19 infection remains unclear.
Methods: A retrospective cohort study was conducted using National COVID Cohort Collaborative (N3C) data. Adults with either a recorded COVID-19 diagnosis or a first clinical encounter with an index date between January 1, 2020, and September 30, 2024, were included (N = 7,920,739). For those with COVID-19, the index date was the diagnosis date; for others, it was 90 days after the first encounter. A 90-day pre-index period was used to assess GLP-1 RA exposure and covariates. Various AEs were identified. Logistic regression models, with and without an interaction term between GLP-1 RA exposure and COVID-19 status, were used to evaluate associations with AEs, adjusting for relevant covariates (demographics, hypertension, diabetes, obesity, etc).
Results: GLP-1 RA use was associated with increased risks of several AEs, including all thyroid cancer types (adjusted odds ratio (AOR) 1.639, 95% CI [1.420, 1.885]), diabetic retinopathy (AOR 2.472, 95% CI [2.337, 2.614]), gastrointestinal events (AOR 2.044, 95% CI [2.002, 2.087]), ocular events (AOR 2.171, 95% CI [2.099, 2.245]), and respiratory system events (AOR 1.623, 95% CI [1.583, 1.663]). GLP-1 RA

use was associated with reduced risks of pancreatic cancer (AOR 0.507, 95% CI [0.377, 0.668]) and no significant association was observed with pancreatitis (AOR 0.982, 95% CI [0.888, 1.085]). The significant interaction between GLP-1 RA use and COVID-19 diagnosis was observed for gastrointestinal events, ocular events, respiratory system events, suggesting differential effects by COVID-19 status.
Conclusion: GLP-1 RA use was associated with increased risks of several adverse events and a reduced risk of pancreatic cancer. Our results provide real-world evidence, using large-scale electronic health records (EHRs) from the N3C cohort, that supports and strengthens findings from previous studies. The observed interactions between GLP-1 RA use and COVID-19 infection suggest that the effects of GLP-1 RA on gastrointestinal, ocular, and respiratory outcomes may differ by COVID-19 status. These findings highlight the importance of considering COVID-19 history when evaluating the safety profile of GLP-1 RAs.

Keywords
GLP-1, COVID-19, National COVID Cohort Collaborative (N3C), Adverse event, Retrospective cohort study

---

Title: A Machine Learning–Guided Feature Selection Framework for Predictive Mapping of Gut Microbiome–Bone Mineral Density Associations

Author list: Lindong Jiang[1], Martha Isabel Gonzalez Ramirez[1], Zhuoqi Wang[2], Lishu Zhang[2], Bo Tian[3], Xu Lin[4], Xiao Zhang[1], Anqi Liu[1], Yun Gong[1], Chuan Qiu[1], Kuan-Jui Su[1], Zhe Luo[1], Qing Tian[1], Li Wu[1], Shashank Sajjan Mungasavalli Gnanesh[1], Hui Shen[1], Hong-Wen Deng[1]

Detailed Affiliations
[1]Tulane Center of Biomedical Informatics and Genomics, School of Medicine, Tulane University, New Orleans, LA, 70112; [2]Beijing Jiaotong University, No.3 Shangyuancun, Haidian District, Beijing, 100044, P. R. China; [3]Central South University, No.932 South Lushan Road, Changsha Hunan 410083 P.R. China; [4]Southern Medical University, 33 Magang Blvd, Shun De Qu, Fo Shan Shi, Guang Dong Sheng, China, 528303.

Abstract
Understanding the association between the gut microbiome and bone mineral density (BMD) is hindered by the compositional nature of metagenomic relative abundance (mRA) data. To address this, we developed Gradient Boosting Tree–Guided Sequential Binary Partitioning (GB-SBP)—a novel data transformation strategy that converts mRA into mutually independent principal balances (PBs) while capturing non-linear relationships with host BMD. Our analysis utilized data from 2,087 participants in the Louisiana Osteoporosis Study, incorporating both metagenomic profiles and clinical covariates (mean age: 48 ± 14 years; 65% female; 69% White, 28% African American, 3% Asian). We applied GB-SBP to the mRA data and used the resulting PBs in linear regression models to identify microbial associations with BMD at multiple anatomical sites (femoral neck, total hip, lumbar spine, and one-third radius). Microbial species with high inclusion frequencies among significant PBs were selected as a denoised subset for downstream prediction tasks. Using machine learning models (XGBoost and Random Regression Forest), we demonstrated that augmenting clinical features with the mRA of the denoised microbiome subset significantly improved femoral neck BMD prediction (reduced RMSE and increased R^2). The GB-SBP-

selected species subsets also achieved comparable or superior performance to models using the full set of species. Furthermore, integrating species subsets relevant to multiple BMD sites enhanced predictive performance, suggesting shared microbial signatures across skeletal regions. To interpret model predictions, we computed Shapley Additive Explanation (SHAP) values, revealing several microbial species (e.g., Roseburia inulinivorans, Dorea longicatena) with higher predictive contributions than some clinical variables such as gender. SHAP-based principal component analysis followed by Leiden clustering uncovered three distinct subgroups of subjects with differing feature importance profiles. Notably, the associations between key microbial species and femoral neck BMD remained consistent across clusters, reinforcing the robustness of our findings.

Keywords
Machine learning, Metagenomics, Bone mineral density.

Title: Imputing Metagenomic Hi-C Contacts Facilitates the Integrative Contig Binning Through Constrained Random Walk with Restart

Author list: Yuxuan Du[1,2], Wenxuan Zuo[1], Fengzhu Sun[1]

Detailed Affiliations
[1]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA, USA; [2]Department of Electrical and Computer Engineering, University of Texas at San Antonio, TX, USA.

Abstract
Metagenomic Hi-C (metaHi-C) has shown remarkable potential for retrieving high-quality metagenome-assembled genomes from complex microbial communities. Nevertheless, existing metaHi-C-based contig binning methods solely rely on Hi-C interactions between contigs, disregarding crucial biological information such as the presence of single-copy marker genes. To overcome this limitation, we introduce ImputeCC, an integrative contig binning tool optimized for metaHi-C datasets. ImputeCC integrates both Hi-C interactions and the discriminative power of single-copy marker genes to group marker-gene-containing contigs into preliminary bins. It also introduces a novel constrained random walk with restart algorithm to enhance Hi-C connectivity among contigs. Comprehensive assessments using both mock and real metaHi-C datasets from diverse environments demonstrate that ImputeCC consistently outperforms other Hi-C-based contig binning tools. A genus-level analysis of the sheep gut microbiota reconstructed by ImputeCC underlines its capability to recover key species from dominant genera and identify previously unknown genera.

Keywords
Metagenomics; Contig Binning; Hi-C

Title: RAZOR: a Database of PCR Primers Targeting Human Respiratory Viruses

Author list: Hunter Mathias Gill [1], Quoseena Mir [1, 2], Rajneesh Srivastava, Ph.D. [1, 3], Sarath Chandra Janga, Ph.D. [1, 4, 5]

Detailed Affiliations
[1] Department of Biomedical Engineering and Informatics, Luddy School of Informatics, Computing and Engineering, Indiana University Indianapolis, 535 West Michigan Street, Indianapolis, Indiana, 46202 2 Department of Microbiology and Immunology, Indiana University School of Medicine, 635 Barnhill Drive, Indianapolis, Indiana, 46202 3 McGowan Institute for Regenerative Medicine, University of Pittsburgh, 450 Technology Drive, Pittsburgh, Pennsylvania, 15219 4 Department of Medical and Molecular Genetics, Indiana University School of Medicine, Medical Research and Library Building, 975 West Walnut Street, Indianapolis, Indiana, 46202 5 Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, 5021 Health Information and Translational Sciences (HITS), 410 West 10th Street, Indianapolis, Indiana, 46202

Abstract
Respiratory viruses like SARS-CoV-2, Influenza A, and others represent a considerable threat to public health, infecting millions of people annually. Previous respiratory virus outbreaks have demonstrated the value of polymerase chain reaction (PCR) testing as a gold standard for definitive virus identification; however, existing database resources for viral PCR primer design are either outdated or restricted to a small number of species. To address the need for updated and comprehensive viral PCR resources, we introduce RAZOR, a database of nearly 20,000 primers covering 20 different respiratory virus species. We created genome-wide template sets for each virus and used them to design quantitative PCR (qPCR) and standard PCR primer pairs with Primer3. Detailed primer information, including sequence coordinates, melting and annealing temperatures, GC content, and hairpin structure probabilities, is provided through a user-friendly Integrative Genomics Viewer (IGV) interactive display. Validated primers, including a group of SARS-CoV-2 primers tested by our group, are also showcased in a dedicated section of the IGV. RAZOR stands out as a valuable tool for investigators designing targeted PCR approaches for respiratory virus detection.

Keywords
Polymerase Chain Reaction, PCR, Respiratory Virus, SARS-CoV-2, IGV

Title: Generating Hypothetical Cell Transitions with Generative Models: A Case Study on EMT
Author list: Samia Islam[1], Sudin Bhattacharya[2,3,4,5]

Detailed Affiliations
[1]Department of Computer Science and Engineering, Michigan State University, East Lansing, Michigan, MI, USA; [2]Institute for Quantitative Health Science and Engineering, Michigan State University, East Lansing, Michigan, MI, USA;[3]Department of Biomedical Engineering, Michigan State University, East Lansing, Michigan, MI, USA; [4]Department of Pharmacology and Toxicology, Michigan State University, East Lansing, Michigan, MI, USA; [5]Institute for Integrative Toxicology, Michigan State University, East Lansing, , Michigan, MI, USA.

Abstract

Understanding cellular plasticity through trajectory inference is crucial for uncovering dynamic biological processes such as differentiation and disease progression. In this work, we propose a generative framework using Variational Autoencoders (VAEs) to synthesize cell state transitions between distinct cell types, even those not naturally observed. We first train a VAE on single-cell gene expression data to learn a low-dimensional latent representation of cell states. To generate a synthetic trajectory, we encode the source and target cell types into the latent space and perform linear interpolation between their latent representations. These interpolated points are then decoded back into gene expression space using the VAE decoder, providing a continuous trajectory of synthetic intermediate cell states. As an initial validation, we apply this method to the well-characterized epithelial-to-mesenchymal transition (EMT), a key process in development and cancer metastasis. We analyze gene expression dynamics along the generated trajectory and compare them with known EMT markers and regulatory transcription factors. This validation establishes a foundation for modeling and investigating hypothetical cell state transitions and strategies for cell fate reprogramming.

Keywords
Cell Trajectory Generation, Generative Deep Learning, Epithelial-to-Mesenchymal Transition (EMT)

Title: Logit-Based ROC Curves Offer Flexibility for Applying Large Language Models to Binary Classification Tasks

Author list: Arslan Erdengasileng, PhD[1]; Jennifer A. Lee, MD [1,2]; Faraaz Chekeni, MD [1,2]; Jeffrey Hoffman, MD [1,2]; Steven W. Rust, PhD [1]

Detailed Affiliations
[1]Nationwide Children's Hospital, Columbus, OH, USA; 2The Ohio State University College of Medicine, Columbus, OH, USA

Abstract
Large language models (LLMs), particularly generative decoder-only models like Llama 3, offer promising approaches for binary classification tasks due to their generalizability and reduced annotation requirements. While encoder-only models such as BERT achieve strong performance, their reliance on extensive labeled datasets is a significant burden in clinical environments with limited expert resources.

Generative LLMs leverage inherent model knowledge, allowing classification through prompting without extensive annotated data, but with increased computational demands due to larger model sizes. In clinical and medical binary classification tasks using LLMs, sensitivity and specificity are frequently employed as evaluation metrics. However, comparing different methods becomes challenging because each can produce distinct, sometimes opposite, sensitivity-specificity trade-offs at a single decision threshold. Moreover, LLMs are sensitive to prompting, where minor prompt variations may substantially alter the sensitivity-specificity balance, further complicating the comparison and optimization.

To address these challenges, we introduce a logit-based evaluation method that dynamically extracts token-level logits specifically for standardized and explicitly capitalized class labels ("YES" and "NO"). Applying Softmax normalization to these logits yields continuous classification probabilities, enabling comprehensive sensitivity-specificity analysis through receiver operating characteristic (ROC) curves.

We validated our approach on a clinical radiology classification task involving 1,015 physician-annotated chest X-ray reports labeled as "normal" or "abnormal". For the LLM outputs, we mapped these to "YES"

for "abnormal" and "NO" for "normal". We performed two comparative analyses: (1) evaluating prompt strategies, specifically a simple direct prompt versus a structured Chain-of-Thought (CoT) prompt, and (2) assessing performance across Llama 3 model variants (3B, 8B, 70B) using a fixed prompt.
The prompt comparison demonstrated distinct sensitivity-specificity trade-offs; the simple prompt showed high sensitivity (0.794) but low specificity (0.427), while the structured CoT prompt achieved high specificity (0.983) at reduced sensitivity (0.661). At this single operating point, it is difficult to definitively determine the superior prompt. However, ROC analysis notably revealed the structured CoT prompt significantly outperformed the simpler prompt across all thresholds (AUC: 0.9546 vs. 0.7886). Model comparisons revealed incremental improvements with increased model size: the 70B-parameter model achieved the highest AUC (0.9936), closely followed by the 8B model (0.9914), while the smaller 3B model performed significantly lower (0.7886).
Our logit-based ROC evaluation framework significantly enhances the practical utility of applying generative LLMs to clinical binary classification tasks, enabling tailored sensitivity-specificity optimization and robust comparisons of prompts and model configurations.

Keywords
LLMs, ROC Curve, Binary Classification, Llama 3, Evaluation

Title: VaxChat: An Advanced RAG- and GPT-based Natural Language Querying System for Exploring Vaccine Information from the VIOLIN Vaccine Database
Author list: Matthew Asato[1], Feng-Yu (Leo) Yeh[1], Jie Zheng[1], Yongqun Oliver He[1]

Detailed Affiliations
[1] Unit for Laboratory Animal Medicine, Center for Computational Medicine and Bioinformatics, Department of Learning Health Science, University of Michigan, Ann Arbor, Michigan, USA.

Abstract
The Vaccine Investigation and Online Information Network (VIOLIN) is a relational database containing detailed, data-mined information on vaccines. VIOLIN currently features an accessible querying interface, enabling users to select options from drop-down menus and conduct keyword searches for specific categories of vaccine information. With the current Large Language Models (LLMs) and ChatGPT technologies, it is possible to develop a more generalizable vaccine query program. Towards this goal, we developed VaxChat, a new LLM-based method for querying the vaccine data based on a VIOLIN-based knowledge graph. VaxChat is a Retrieval-Augmented Generation (RAG) system that retrieves relevant information from VIOLIN to answer user queries. RAG supplements LLMs by providing them with real data, reducing the chance of hallucination or poorly constructed responses. VIOLIN vaccine data was first represented in a Neo4j knowledge graph to better represent the complex relationships stored in VIOLIN. Then, VaxChat enables natural language querying by translating user queries into Cypher, a Neo4j querying language (similar to SQL), to search the backend knowledge graph. The current VaxChat system utilizes Chat-GPT 4o Mini; however, the program is built to easily change to different models, such as open-source models like Llama. The flexibility of model selection enables rapid adjustments, allowing VaxChat to capitalize on the latest advancements in LLM models. The VaxChat user interface allows users to chat, ask questions, and receive answers in an easy-to-use and reliable manner. For transparency, users can see

the Cypher query that the model generated and the retrieved data. As a demonstration, a user can ask VaxChat questions like "What vaccines target Brucella?", which is converted into Cypher queries like "MATCH (v:Vaccine)-[:TARGETS_PATHOGEN]->(p:Pathogen) WHERE p.NAME CONTAINS 'Brucella' RETURN v.NAME". The retrieved results will be a list of vaccine names that target Brucella, then, using this retrieved data, VaxChat will create a response to the original user query. VaxChat differs from VIOLIN's querying interface in that VIOLIN uses keyword search and key-value pairs in dropdown menus, whilst VaxChat is strictly natural language, providing VaxChat the benefit of being more accessible and easier to use. In conclusion, VaxChat utilizes GPT and RAG to automatically query data stored in the VIOLIN-based knowledge graph. In the future, VaxChat will be updated to incorporate more knowledge from different resources such as the Vaccine Ontology, PubMed literature, and ClinicalTrials database, to support more complex queries and better serve vaccine research.

Keywords
Large Language Models, Retrieval Augmented Generation, Knowledge Graph, Vaccines, ChatGPT

Title: Biomarker Discovery in Tourette's Disorder
Author list: Subramanian Krishnamurthy[1,2], Jay A. Tischfield[1,2], Gary A. Heiman[1,2], Jinchuan Xing[1,2]

Detailed Affiliations
[1]Department of Genetics, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA;[2]Human Genome Institute of New Jersey, Rutgers, The State University of New Jersey

Abstract
Tourette's disorder (TD) is a childhood onset neuro-developmental disorder (NDD), Characterized by Motor and Vocal tics. It has a large genetic component. A variety of DNA-based methods have been applied to identify genetic markers of TD, but the connection to RNA-based gene expression has not been fully exploited, though it has been used in other neuro developmental and neuro degenerative disorders.
We used a cohort of cases (n=21) with TD and controls (n=21) without TD from the Tourette's International Collaborative Genetics Study (TICGenetics). One subject was selected from each family and subjects on medications with potentially confounding effect were excluded. This study utilized bulk RNA sequencing to identify expressed changes in whole blood RNA in TD. Multiple analytical strategies were employed to narrow differentially expressed RNA targets to a small set of potential biomarkers of TD.

We identified 19 genes that had statistically significant fold change > 2. 11 were up-regulated and 8 were down regulated. Ten of the genes are known to be associated with brain development and neurological disorders. For example, NRCAM (Neuronal Cell Adhesion Molecule) is known to be associated with neuro development disorders, LRRN3 (Leucine Rich Repeat Neuronal 3) with Autism, CCR8 (C-C Motif Chemokine Receptor 8) with Multiple Sclerosis and 7 others with Epilepsy Schizophrenia.
Data from this preliminary study suggests these may be promising targets for diagnostics and therapeutics in TD.

Keywords

Tourette's Disorder, RNA sequencing, Transcriptome, Biomarker Discovery, Neuro Developmental Disorder, Diagnostics

---

Title: Differential Effects of Physical Activity Intensity on Chronic Disease Risk: Insights from the All of Us Research Program

Author list: Eric Wang [1], Lynda Faye Bonewald [2,3], Gang Peng [3,4]

Detailed Affiliations

[1]The Elmore Family School of Electrical and Computer Engineering, Purdue University, Indianapolis, IN, USA; [2] Department of Anatomy, Cell Biology & Physiology, Indiana University School of Medicine, Indianapolis, IN, USA; [3] Indiana Center for Musculoskeletal Health, Indiana University School of Medicine, Indianapolis, IN, USA; [4] Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, USA.

Abstract

While physical activity is generally associated with improved health, its effects can vary depending on intensity and types of exercise. In this study, we analyzed data from 14,754 participants in the All of Us Research Program - a national initiative supporting precision medicine through the integration of lifestyle, environmental, and biological data - to examine how different levels of physical activity impact chronic disease risk.

Participants (mean age 49.8 years; 71% female; 79% White; mean BMI 29.9 kg/m²) were clustered into four groups – sedentary (Cluster S), household activity (Cluster H), light activity (Cluster L), and active activity (Cluster A) - based on their Fitbit average daily minutes of light, fairly active, and very active movements. Disease outcomes were identified via electronic health records at least six months after the activity monitoring period.

All three clusters (Cluster L, H, and A) demonstrated lower risks than the sedentary group (Cluster 4) for metabolic conditions, though the extent and nature of these reductions varied.

· Cluster A showed the most comprehensive reductions, with the lowest risk for obesity and type 2 diabetes, highlighting the metabolic benefits of high-intensity activity, though differences in other disease domains were less pronounced.

· Cluster L also exhibited notable reductions, particularly in metabolic conditions like type 2 diabetes and morbid obesity along with improvements in renal and infectious diseases such as acute renal failure and sepsis,

· Cluster H showed only mild overall reductions but was distinctively associated with increased risks for pregnancy-related complications, suggesting a higher concentration of reproductive-age individuals.

However, higher physical activity was not universally beneficial. The risk of stress fractures, hallux rigidus, hammer toe, and bunion increased consistently with greater step counts across most clusters, suggesting potential joint overuse or biomechanical strain.

Several dermatologic and sensory conditions also showed intensity-specific patterns:

· Other retinal disorders and seborrheic keratosis showed decreasing hazard across all active clusters, but increased risk in the sedentary group.

· Hemangioma risk was slightly elevated in the sedentary and household clusters, but stable or lower in more active groups.
While physical activity-particularly at higher intensities-offers clear metabolic health benefits, it may also increase risk for select musculoskeletal and dermatologic conditions, depending on the individual's activity profile, age, and sex. These findings emphasize the need for personalized exercise recommendations in public health and clinical care to optimize benefits while mitigating potential risks.

Keywords
Exercise Intensity, Step Count, Chronic Disease, Sex Differences, AoU

Title: THANOS: An AI Pipeline for Engineering Antibodies
Author list: Arnav Solanki[1], Neha S Maurya[1], Wenjin Jim Zheng[1]

Detailed Affiliations
[1]McWilliams School of Bioinformatics, The University of Texas Health Science Center at Houston, Houston, TX, USA

Abstract
The Brain Research Through Advancing Innovative Neurotechnologies (BRAIN) initiative seeks to unravel the complexities of brain cell types, connections, and functions. One powerful approach to studying neural circuits is imaging brain proteins using antibodies. However, generating high-quality antibodies is often slow and expensive. Recent advances in AI tools like AlphaFold3 and RFdiffusion offer a fast, fully digital alternative to traditional experimental screening. This project introduces THANOS (Targeted High-throughput Antibody Notation, Optimization, & Screening), a novel pipeline for rapidly engineering. THANOS was used to design de novo antibodies targeting human proteins by redesigning antigen-binding sites of antibodies using a high-performance GPU server. This case study focuses on Parvalbumin (PVALB), a calcium-binding protein abundant in neural cells. Alphafold3 was used to model the 3D complexes of human PVALB and 12 mouse antibody variable fragments. With these 12 models as initial states, 300 structural variants were generated by using RFdiffusion to diffuse the complementarity determining regions (CDRs) at the binding sites. ProteinMPNN was used to predict optimal sequences that fold these structures, yielding new antibody chain sequences containing new residues in the CDRs. These 300 variants were screened against PVALB using AlphaFold3 to predict their binding. 4 candidates were observed to bind strongly based on low predicted alignment error. These candidates were validated through: 1) Structural inspection using ChimeraX. 2) Molecular dynamics simulations with GROMACS. The best candidate demonstrated a MMGBSA energy of -35 kcal/mol over 100 ns. 3) Solubility checks using Aggrescan3D to confirm the absence of aggregation-prone residues. 4) Sequence alignment to ensure minimal mutations and preserve nativeness. These antibodies are currently undergoing experimental validation. THANOS demonstrates how AI can accelerate antibody engineering from weeks to hours without a wet lab. This pipeline can be applied on any desired target protein (beyond the interest of the BRAIN initiative) for numerous applications such as viral or cancer therapy, and will be invaluable to the fields of immunology and pharmacology.

Keywords

Antibodies, Protein Engineering, AI, Machine Learning, AlphaFold

---

Title: VaxCT: A Web-based Vaccine Clinical Trial Database Integrating Clinical Trials and VIOLIN Vaccine Resources

Author list: Feng-Yu (Leo) Yeh[1], Yongqun He[1]

Detailed Affiliations

[1] Unit for Laboratory Animal Medicine, Center for Computational Medicine and Bioinformatics, Department of Learning Health Science, University of Michigan, Ann Arbor, Michigan, USA.;

Abstract

Vaccine research and development are crucial for global health, yet researchers struggle to navigate the vast, disparate information within clinical trial databases. While ClinicalTrials.gov is a comprehensive repository, it lacks specialized tools for vaccine-centric queries. This presents a need for a platform that integrates this data with rich ontological and life cycle information. The Vaccine Investigation and Online Information Network (VIOLIN) on the other hand provides an extensive vaccine knowledge base, but a direct, queryable link to clinical trial data has been lacking. We developed VaxCT, a novel web application for the integrated analysis of vaccine trial data. Our methodology leverages the ClinicalTrials.gov (AACT) database, a comprehensive relational version of the clinical trial registry hosted in a PostgreSQL database. The VaxCT system's backend is built with Node.js and the Express.js framework, exposing a RESTful API. This API dynamically constructs and executes SQL queries to retrieve study information, performing joins across core AACT tables such as studies, interventions, keywords, and conditions. The initial identification of vaccine-related trials is performed via direct keyword filtering within these SQL queries. Following retrieval from the AACT database, this trial data is integrated with the VIOLIN knowledge base to create an unified dataset, providing researchers a platform for advanced cross-domain querying. Our pipeline processed the entirety of ClinicalTrials.gov, identifying over 10,000+ vaccine-related clinical trials. The VaxCT platform also enables categorized queries, allowing users to instantly collate trials by disease, such as the 1,600 trials for cancer vaccines or 2,000 for influenza vaccines. A key achievement of our system is the semantic enrichment of trial data. By mapping trials to VIOLIN, we annotate them with formal Vaccine Ontology (VO) identifiers, and this critical linkage enables high-impact queries such as experimental data from laboratory animal model studies, vaccine formulation details (e.g., adjuvant), and post-licensure market safety profiles if the vaccine is used in the market. Such whole vaccine life cycle information will support more advanced data analysis and vaccine evaluation, and future vaccine design. VaxCT is an integrated system that unites clinical trial records with a specialized vaccine knowledge base. By establishing interoperability between these critical datasets and leveraging semantic ontology, VaxCT provides a powerful platform for knowledge discovery in translational bioinformatics. This tool facilitates more efficient data exploration for researchers and clinicians, potentially accelerates vaccine research and development, and supports a more informed public health ecosystem.

Keywords

Vaccine, Clinical Trial, Database, SQL, Research and Development

---

Title: OntoChimp: A ChatGPT-powered Tool for Identifying Key Concepts in Ontological Development and Its Evaluation in the Vaccine Domain

Author list: Sam Smith[1], Jie Zheng[1], Feng-Yu Yeh[1], B. Damayanthi Jesudas[2], John Beverley[3,4], William D. Duncan[2], Yongqun He[1]

Detailed Affiliations

[1]University of Michigan Medical School, Ann Arbor, MI, USA; 2University of Florida, Gainesville, FL, USA; 3University at Buffalo, NY, USA; 4Institute for Artificial Intelligence and Data Science, NY, USA

Abstract

Background: Identifying key concepts (KCs) is a critical step in ontology development. Candidate KCs are typically solicited from domain experts, based on their experience and knowledge - a process that can be time-consuming. Here we describe our development of OntoChimp, a ChatGPT-supported system to automatically identify KCs that may be considered as ontological entities in the domain of interest. KCs are the core concepts in the document while the clusters are semantically grouped words that support the KCs. OntoChimp has been applied to identify and evaluate KCs and clusters in the domain of vaccines. OntoChimp Development: OntoChimp was developed using Python, incorporating ChatGPT API (GPT-4-turbo), spaCy SentenceTransformer. The strategy is to first identify a manually curated set of documents in a domain of interest, and analyze these documents in sufficient detail to identify and categorize the key concepts found. Each document is submitted individually to ChatGPT for KC identification using a short prompt introducing the domain, the document text, and instructions for JSON formatting of results. The KCs are then combined into a workbook as well as a Pydantic BaseModel class "TermTable" in which the lemmatized term, head term for multi-word phrases are available for subsequent analysis. Additionally each KC is processed against the BioPortal Annotator API to match with any terms found in established ontologies for the domain, with KCs not found being possible new terms to add to a domain-specific ontology such as Vaccine Ontology (VO). OntoChimp Evaluation: OntoChimp was applied to analyze papers in the domain of vaccinology, and the results were manually evaluated by experts. For example, OntoChimp was used to mine a review paper on vaccine informatics (PMCID: PMC3134832), a review article on vaccine adjuvants (PMCID: PMC10356842), and a research article on a specific vaccine (PMID: 1908158), which resulted in the detection of 71, 64 and 22 KCs, respectively. A combination of the KCs led to the finding of 139 distinct KCs, of which 94 matched terms in the VO using the BioPortal Annotator and manual checking. Furthermore, 11 were found to be vaccine-related terms (e.g., antigenic variability, immunostimulant, and reverse vaccinology) worth being added to the VO as new ontology terms. Conclusion: OntoChimp has demonstrated its effectiveness in automatically identifying KCs in the vaccine domain. The system is being tested in mental health and other research domains.

Keywords

OntoChimp, ChatGPT API, Key concepts, Vaccine Ontology.

Title: Evaluating LLM Agents for Insurance Coverage Workflow Automation

Author list: Junyoung Kim[1], Youssef Moksit[1], Mengshu Nie[1], Cong Liu[1]

Detailed Affiliations

[1]Department of Pediatrics, Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA USA

Abstract
Genetic testing plays a pivotal role in diagnosing rare diseases, guiding targeted therapies, and assessing patient risk profiles. While its clinical use has expanded rapidly, insurance policies related to genetic testing remain difficult to access and interpret. These policies vary substantially across states and payers, are expressed in complex terminology, and are frequently updated. This creates significant administrative barriers and can often lead to high rejection rates in insurance claims. While Large Language Models (LLMs) have transformed biomedical research and clinical decision-making, their use in insurance-related tasks is still emerging. Web-enabled LLM agents can now retrieve real-time information and complete forms autonomously, offering new ways to streamline insurance workflows. In this study, we evaluate these agents with a focus on three core objectives.
First, we assessed the accuracy of retrieving relevant information and policy documents, including identifying in-network insurance payers associated with a selected vendor (i.e., GeneDx) and retrieving their corresponding policy documents. Our results show that web-based LLM agents (GPT-4o-web-preview) achieved a recall rate of 44.11%, while Perplexity achieved a recall rate of only 2.64% in retrieving insurance payers.
Second, we evaluated the agent's ability to answer 13 common insurance policy questions, covering topics such as age requirements, medical necessity, and CPT code criteria. This task was conducted based on 50 synthetic patient scenarios constructed by five genetic tests (WES, WGS, BRCA1/2, CMA, DMD) with four payers (Aetna, BCBS Federal Employee Plan, Cigna, United Healthcare), all with publicly available policies and confirmed in-network status with GeneDx across all states. Web-based LLM agents achieved 57.29% accuracy, while RAG-based agents using a manually collected insurance policy database reached 61.60%. When the correct insurance policy was directly provided, without retrieval, accuracy rose to 76.19%.
Third, we evaluated the agent's ability to automatically complete pre-authorization forms, focusing on payers with publicly accessible templates (Aetna, Cigna, Connecticut Medicaid, Texas Medicaid). We evaluated the generated submissions across multiple dimensions: submission success, field-level accuracy, and the effectiveness of feedback under a multi-agent configuration. In the multi-agent setting, the initial agent achieved 80.9% field-level accuracy. However, introducing an 'LLM-as-denier' agent to critique the initial submission and allowing a refinement agent to resubmit the form degraded accuracy by 61.1%.
This work represents a foundational effort to scale insurance policy reasoning and administrative automation in genomics/genetic service using LLM agents. Our study contributes to the advancement of the role of LLMs in clinical practice.

Keywords
Large Language Model (LLM), Genetic Testing Insurance, Clinical Decision Support, Pre-authorization Automation, Agent Systems, Biomedical Informatics

---

Title: Robust Group PCA for Separable Noise: An Argument for Subject-Level PCA With Whitening
Author list: Samuel Oriola[1], Calvin McCurdy[2], Bradley T. Baker[3], Vince D. Calhoun[4], Rogers F. Silva[5]

Detailed Affiliations
[1]Georgia State University, Atlanta, GA, [2] Tri-Institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Georgia State University Atlanta, GA, [3]GSU/GATech/Emory, Atlanta, GA, [4]Department of Mathematics and Statistics, Georgia State University, Atlanta, Georgia, USA 30302, [5]Department of Electrical and Computer Engineering Georgia Institute of Technology, Atlanta, Georgia, USA 30332

Abstract
Functional magnetic resonance imaging (fMRI) indirectly captures neural function with high spatial resolution and has driven discoveries in brain connectivity across age, gender, mental illnesses, and developmental stages. As fMRI datasets grow in number, aggregation methods such as averaging or low-rank approximations are increasingly more likely to lose subject-specific details, potentially biasing group estimates and misrepresenting individuals, which in turn limits replication and reduces the translational utility of findings. Group principal component analysis (PCA) is the de facto approach for aggregating datasets, but it varies across implementations. Tools like FSL and GIFT support group PCA (gPCA) and have shaped neuroimaging studies for decades. Yet, the impact of individual subject variability on the quality of group-level results remains unquantified. This study aims to identify computational strategies that improve the accuracy and robustness of group-level representations. Three common gPCA implementations are considered: 1) simple concatenation, 2) concatenation with subject sum of squares normalization, and 3) concatenation with subject-level PCA whitening. Simulated scenarios test these methods to identify optimal approaches for group dimensionality reduction while preserving the ground-truth group mean information. Unique to this work, the similarity between individuals is systematically adjusted from high to very low across ground-truth sources, while the additive noise is kept orthogonal to the signal, serving as a fair, non-confounding setting for this initial investigation. In addition, we manipulate two key parameters: 1) the proportion of total variance assigned to signal sources, which represents the variance contribution from relevant signals in the data, and 2) the variance profile, which determines which subjects are emphasized at the group level. We also study the threshold for variability retained during subject-level PCA with whitening. Our results across ten different random seeds demonstrate that the ground-truth mean consistently captures variability well for highly similar sources across subjects, but performs poorly for dissimilar sources. These errors are largest for the most dissimilar subjects, revealing that the true group mean can bias inferences about atypical subjects. Concatenation with sum of square normalization yields variance recovery on par with the ground-truth mean in situations of high and moderate signal-to-noise ratios, while subject-level PCA with whitening and no data reduction matches the ground-truth mean performance even at extreme noise levels. Simple concatenation gave the worst approximation of the ground-truth group mean overall. Future work will explore non-separable noise scenarios and methodology impact on published datasets to establish final recommendations for neuroimaging analyses.

Keywords
Computing; Data analysis; FUNCTIONAL MRI;Modeling; Multivariate; Group PCA

Title: Counselor Agent: A RAG-Based AI Agent for Reliable Genomic Guidance in Clinical Care
Author list: Youssef Mokssit[1], Cong Liu[1], Junyoung Kim[1], Adelina Nie[1]

Detailed Affiliations
[1]Department of Pediatrics, Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA USA

Abstract
In the NICU and other clinical settings, genomic evaluation often requires significant effort in terms of navigating disparate tools and systems and piecing together information from multiple internal and external guidelines and databases. Delays in accessing accurate genomic guidance can hinder timely diagnosis and impact clinical decision-making. To address this challenge, we present an LLM-Agent designed to assist clinicians, genetic counselors, and other healthcare professionals in accessing up-to-date, guideline-based information for rare genetic conditions. Our approach integrates automated knowledge update workflows, diverse PDF document parsing strategies, prompt instructions to enforce tiered knowledge search logic, and prompt optimization based on the TextGrad framework. We introduced an evaluation benchmark dataset consisting of synthetic genetic conditions and real-world guideline-based queries to evaluate the agent's ability to retrieve, reason, synthesize, and present information accurately and more importantly, in alignment with source priority and organisational standard operating procedures (SOPs). To evaluate how document ingestion and parsing quality affect agent performance, we ran a benchmark based on GeneReviews references which are available in both PDF and structured XML formats. The structured XML, a format easily ingested by LLMs, provides a performance ceiling for other parsing methods. Our results show that among the PDF parsing methods tested, Google Document AI performed best, achieving an LLM-judged (gpt-4o) question answering score (0.0 to 1.0) of 0.89 compared to 0.926 for the XML gold standard. To support reliable and protocol-compliant retrieval, we implement a tiered knowledge architecture that prioritizes internal institutional guidelines, followed by authoritative external sources, and finally, fallback mechanisms such as web search when none of the above is available. All genetic materials are parsed, chunked, embedded, and stored in vector databases to enable fast and contextually relevant retrieval. Our evaluation on a benchmark designed to test agent retrieval accuracy and synthesis proficiency yielded an average LLM-judged quality score of 0.58 (scale 0.0–1.0) across all task categories, measuring alignment with gold-standard answers. A more detailed analysis, comparing the agent's behavior to a gold-standard action sequence, showed that it reproduced the prescribed tier-based (SOPs-compliant) search path with 59% exact-match accuracy (EMA). Finally, we observed that prompt optimization techniques such as TextGrad improved the agent's LLM-judged answer accuracy by 18% without requiring any changes to the core architecture or base model. Our study presents a pilot strategy for developing a reliable assistant for genetic providers, while remaining adaptable to incremental improvements and further experimentation with its retrieval architecture, system prompt, and underlying foundational model.

Keywords
Large Language Model (LLM) Agent, Rare Genetic Disease Diagnosis, Genomic Decision Support, Biomedical Informatics, Retrieval-Augmented Generation (RAG), Clinical Guidelines.

Title: Spatially Resolved Single-Cell Analysis Reveals Immune and Structural Disruption in ZIKV-Infected Mouse Brain
Author list: Md Musaddaqul Hasib[1], Wen Meng[2], Hugh Galloway,[1], Shou-Jiang (SJ) Gao[2], Yufei Huang[1]

Detailed Affiliations
[1]Division of Malignant Hematology and Medical Oncology, University of Pittsburgh;[2] Department of Microbiology and Molecular Genetics, University of Pittsburgh

Abstract
Zika virus (ZIKV) infection presents distinct neurological outcomes influenced by strain and day specific dynamics, particularly between the Asian and African strains at day 4 and day 6 post infection. The study utilized spatially resolved single-cell transcriptomics and scRNA-seq to analyze ZIKV-infected mouse brains, revealing that the African strain consistently exhibited higher infection densities than the Asian strain both 4- and 6-day post infection, indicating greater neurovirulence. Upon infection, significant alterations in cell densities, microglia expanded over time, peaking at day 6 post-infection, while excitatory neurons, inhibitory neurons, oligodendrocytes, and astrocytes declined. The inferred infection trajectory indicates that the virus preferentially spreads from the infection origin through anterior cingulate regions located near major fiber tracts(cuado pudamen) to other hemisphere. During this spread ZIKV infects all major brain cell-types without showing any preference. The progression of ZIKV infection was tracked through a pseudo-temporal trajectory, showing that the Asian strain spread more gradually compared to the aggressive spread of the African strain, which was associated with higher infection burdens. Microglial density increased over time in both strains, but the Asian strain showed much faster microglial recruitment and response, leading to better infection containment at day 6 compared to the African strain, which had a less robust immune response. The composition of microglial subtypes shifts with days during infection, with increased disease-associated microglia (DAM) at day 6. The study identified DAM playing a crucial role, with higher DAM recruitment correlated with better infection control, while inflammatory microglia was linked to prolonged neuroinflammation. Infected microglia in African strain-infected samples exhibited extensive transcriptional changes, indicating a dysregulated immune response, phagocytosis at day 6, while Asian strain-infected microglia showed a more controlled activation, enhancing viral clearance without excessive inflammation. We also found that major communication axis “Apoe-(Trem2 + Tyrobp)” between infected cells and microglia is disrupted at day 6 in African-strain infected samples which implies impaired immune response and inflammation regulation. Infection with the African strain also leads to significant cellular imbalances and functional impairments in brain cell types, resulting in severe motor dysfunction and early mortality, while the Asian strain results in milder symptoms and better survival rates. Overall, the research underscores the critical role of microglial dynamics and strain-specific differences in shaping the immune landscape during ZIKV infection, providing insights into potential therapeutic targets for managing viral infections.

Keywords
ZIKV, Spatial transcriptomics, Microglia, Motor dysfunction

Title: From Ambiguity to Precision: Clarifying Cell Identity in C. elegans Embryos via Cell-Cell Interaction-Guided Learning
Author list: Xingyu Chen[1], Michael Q. Zhang[1] *

Detailed Affiliations

[1]Department of Biological Sciences, School of Natural Sciences and Mathematics, University of Texas at Dallas, Richardson, TX, USA

Abstract
The invariant cell lineage of Caenorhabditis elegans has long served as a foundational model for understanding the molecular mechanisms of development (Sulston et al., 1983). In current single-cell RNA-seq datasets, cell annotations often rely on hierarchical clustering guided by known marker genes (Tintori et al., 2016). However, this approach struggles as cell numbers grow and lineage branches diversify. The scarcity of uniquely distinguishing markers leads to annotation ambiguities, particularly among closely related sister cells. For example, cells like ABala and ABalp are frequently merged into a single identity, 'ABalx'. Such ambiguity is pervasive in foundational datasets (Tintori et al., 2016; Packer et al., 2019; Cole et al., 2024), presenting a major barrier to constructing high-resolution models of cell fate specification and morphogenesis. To address this, we present CHACAM (CCI-guided Hierarchical Attention-based Cell Allocation Model), a knowledge-guided computational framework that improves upon conventional hierarchical clustering by incorporating cell–cell interactions (CCI) as biological constraints. As a foundational resource, we built the first comprehensive C. elegans ligand–receptor (L–R) interaction database compatible with the CellPhoneDB format (Efremova et al., 2020). Using this database and high-resolution cell–cell contact data (Cao et al., 2020), we trained a supervised model on confidently labeled cells to learn L–R interaction features as predictive "interaction signatures" that distinguish each cell type's contact partners from non-partners. CHACAM applies these signatures to resolve ambiguous annotations by triangulating from a third-party reference cell with a known contact map. For instance, to resolve the merged label 'ABalx', the model may use the contact profile of MSa, which contacts ABalp but not ABala. Using learned L–R signatures, CHACAM reassigns the ambiguous cell's identity in a way that aligns with both its expression profile and physical interaction context. Cross-validation across multiple third-party reference cells confirmed the robustness of this approach. CHACAM offers a biologically informed, machine learning-based solution to cell identity disambiguation, enhancing legacy C. elegans developmental data. By producing more precise and validated annotations, CHACAM enables more accurate modeling of gene regulatory networks driving early morphogenesis. The methodology is generalizable and holds potential for resolving annotation challenges in other singlecell datasets.

Keywords
C. elegans embryogenesis; single-cell RNA-Seq; cell identity annotation; cell-cell interaction; ligand-receptor signaling; supervised Machine Learning.

Title: Refining Chemotherapy Decisions in Fertility-Preserving and low-risk ER+/HER2− Breast Invasive Ductal Carcinoma Using Machine Learning–Derived Genomic Subgrouping
Author list: Wanru Guo[1,2], Curtis Tatsuoka[1,2]

Detailed Affiliations
[1]Department of Biostatistics, School of Medicine, University of Maryland, Baltimore, Baltimore, MD, USA;
[2] University of Maryland Marlene and Stewart Greenebaum Comprehensive Cancer Center, Baltimore, MD, USA

Abstract

Background
Invasive ductal carcinoma (IDC) is the most common breast cancer subtype. While hormonal therapy is standard for ER+ disease, certain clinical and genomic features may favor the addition of chemotherapy. Using a machine learning approach, we aimed to firstly, identify ER+/HER2− patients under 50 who benefit from combination therapy despite fertility preservation concerns; and secondly, uncover genomic and immune signatures associated with benefit in patients with low NPI that are typically spared chemotherapy.
Methods
We analyzed a longitudinal cohort of 423 IDC patients who received adjuvant treatment post-surgery, with 5-year overall survival (OS) and Recurrence-Free Survival (RFS) as outcomes. A two-stage Virtual Twins (VT) machine learning framework was applied to identify subgroups benefiting from combination therapy (hormonal therapy + chemotherapy ± radiation) versus hormonal therapy alone, focusing on patients <50 years and those with low NPI. In Stage 1, multiple classifiers (random forest, MLP, XGBoost) were compared for OS and RFS prediction; the best-performing model was used to estimate individual treatment effects (ITE) via counterfactuals. In Stage 2, a regression tree was trained on ITE using the top 10 predictive features to identify treatment-sensitive subgroups.
Results
OS and RFS were modeled using 19 clinical and multi-omics features (transcriptomics, CNV, mutations, methylation) in stage 1 of VT, enabling ITE-based subgroup identification via regression trees in stage 2. In ER+/HER2− patients under age 50, those with the greatest overall survival benefit from combination therapy had high NPI (>5.0) and harbored amplifications at 11q13/14 (CCND1, PAK1, RSF1, EMSY), 8p12, 8q, and 20q—indicative of proliferative and endocrine-resistant biology. These tumors also exhibited TRG/TRA deletions and high CD8+ infiltration. Among low NPI (<4.0) patients, the most responsive subgroup showed 17q23 and 20q amplifications, high CD8+ infiltration, and complex genomic instability, including 5q loss, 8q gain, and 10p/12p gains involving mitotic regulators (TTK, AURKB, FOXM1, RAD51AP1), revealing hidden aggressiveness not captured by clinical risk alone. Conversely, patients with genomically quiet tumors—defined by isolated 8p12 amplification, 1q gain, 16q loss, 8q/20q gains—and Luminal A or Claudin-low subtypes exhibited minimal benefit, failing to support the use of combination therapy in this group.
Conclusions
Virtual twin modeling identified ER+/HER2− IDC subgroups—despite clinical low-risk or fertility-preserving status—that derive meaningful benefit from adjuvant combination therapy when harboring high-risk genomic features. These findings support genomics-informed escalation in patients typically spared chemotherapy, highlighting the value of precision oncology in tailoring adjuvant care.

Keywords
Invasive ductal carcinoma, Virtual Twins, Counterfactual modelling, Machine learning, Genomic biomarkers, Treatment heterogeneity

---

Title: Using accessible ML models via PyTorch for early cancer detection
Author list: Edward Yan[1, 2]

Detailed Affiliations

[1]Westlake High School, Austin, TX, USA 2Harvard Computer Society AI Bootcamp, Cambridge, MA, USA

Abstract
Cancer, a group of diseases, is known to have particularly poor prognosis. One of the key elements of cancer that translate to its lower survivability is its lack of early detection. Biopsies for detecting cancer often suffer from long wait times and manual identification, which may hinder prompt treatment. Some cancers may regress from Stage I into Stage IV within a matter of several months. Taking into account times between biopsies and treatment, earlier detection of cancer remains imperative as one step to ensure optimized survivability. Recent investigations have tapped into artificial intelligence (AI) as a tool for earlier detection, particularly with regards to computer vision. Although these algorithms are often complex and proprietary in nature, its effects can be replicated through more simplistic methods. This study aimed to examine the potential of more simplistic algorithms available to the common person, such as via Python. A dataset of 10,000 images pre-classified into malignant and benign breast tumors was obtained from Kaggle. Data was preprocessed via PyTorch and augmented with random transforms including but not limited to rotations, horizontal and vertical reflections, and crops. Augmented and preprocessed data was sent to a machine-learning algorithm using a three-layered convolutional neural network (CNN) developed via PyTorch. End results yielded upwards of 90% training and validation accuracy on breast cancer detection. Using a pre-trained model (ResNet) yielded similar results in accuracy, reaching high points upwards of 90% in both training and validation accuracy, suggesting accessible, simplistic models may also be of use in early cancer detection. Future researchers may wish to explore the role of simplistic AI algorithms in solving cancer, potentially to increase accessibility to hospitals who may not be able to afford expensive, proprietary AI algorithms, as well as generalize it to potentially more aggressive cancers such as brain cancers and lung cancers.

Keywords
Machine learning, breast cancer, PyTorch, accessibility

Title: DMPPred: a tool for identification of antigenic regions responsible for inducing type 1 diabetes mellitus

Author list: Nishant Kumar[1], Sumeet Patiyal[1], Shubham Choudhury[1], Ritu Tomer[1], Anjali Dhall[1] and Gajendra P. S. Raghava[1]

Detailed Affiliations
[1]Department of Computational Biology, Indraprastha Institute of Information Technology, Delhi, Okhla Industrial Estate, Phase III, (Near Govind Puri Metro Station), New Delhi 110020, India

Abstract
There are a number of antigens that induce autoimmune response against β-cells, leading to type 1 diabetes mellitus (T1DM). Recently, several antigen-specific immunotherapies have been developed to treat T1DM. Thus, identification of T1DM associated peptides with antigenic regions or epitopes is important for peptide based-therapeutics (e.g. immunotherapeutic). In this study, for the first time, an attempt has been made to develop a method for predicting, designing, and scanning of T1DM associated peptides with high precision. We analysed 815 T1DM associated peptides and observed that these peptides are not associated with a

specific class of HLA alleles. Thus, HLA binder prediction methods are not suitable for predicting T1DM associated peptides. First, we developed a similarity/alignment based method using Basic Local Alignment Search Tool and achieved a high probability of correct hits with poor coverage. Second, we developed an alignment-free method using machine learning techniques and got a maximum AUROC of 0.89 using dipeptide composition. Finally, we developed a hybrid method that combines the strength of both alignment free and alignment-based methods and achieves maximum area under the receiver operating characteristic of 0.95 with Matthew's correlation coefficient of 0.81 on an independent dataset. We developed a web server 'DMPPred' and stand-alone server for predicting, designing and scanning T1DM associated peptides (https://webs.iiitd.edu.in/raghava/dmppred/).

Keywords
diabetes mellitus, type 1 diabetes, β-cells, BLAST, machine learning, web server

---

Title: Complete sub-lineage-specific reference genomes for refined Mycobacterium tuberculosis genomic analysis

Author list: Sam Modlin[1], Md Saddam Hossain[1], Faramarz Valafar[1]

Detailed Affiliations
[1]Laboratory for Pathogenesis of Clinical Drug Resistance and Persistence, School of Public Health, San Diego State University, San Diego, CA, USA

Abstract
Reference genomes are an essential resource for comparative genomics and contextualizing genomic elements linked to phenotypes of interest in large-scale genetic screens. In M. tuberculosis (Mtb), most genomic analyses rely on a single reference genome, but this practice precludes discovery of variants in regions absent from the reference and can miss map sequencing reads from genomic regions lacking synteny with the reference genome. Here, we provide 8 complete sub-lineage-specific reference genomes and comprehensively define Regions of Difference (RDs) and mutations in antibiotic resistance genes for each sub-lineage. First, we demonstrate that PacBio-only assemblies match or exceed the quality of hybrid assemblies and identify a reduced set variants that differentiate the virulent and avirulent H37 type strains. We demonstrate the accuracy of the new reference genome set by showing improved mapped read percentages compared to traditional reference genome mapping and by recapitulating previously described RDs and genetic background resistance mutations. We then demonstrate the utility of sub-lineage-specific genomes by improved read-mapping completeness from public short-read sequencing data and describe several examples of findings missed in prior analyses using the H37Rv reference genome that are uncovered with sub-lineage-specific reference genomes. Finally, we provide guidance for assessing quality of long-read Mtb genome assemblies with techniques generalizable to other bacterial species. These reference genomes provide a resource for refined comparative genomic analysis of Mtb clinical genomes and can serve as a guide for designing genomic characterization of collections of other bacterial species.

Keywords
Tuberculosis, Structural Variants, Reference Genome, Region of difference

Title: polyAmod: Integrated Single Molecule Profiling of RNA Modifications and Polyadenylation events using Nanopore Direct RNA Sequencing Datasets
Author list: Sahiti Somalraju[1], Sarath Chandra Janga[1,2,3]

Detailed Affiliations
[1]Department of Biomedical Engineering and Informatics, Luddy School of Informatics, Computing and Engineering, Indiana University Indianapolis, Indianapolis, IN, USA; [2]Department of Medical and Molecular Genetics, Indiana University School of Medicine, Indianapolis, IN, USA; [3]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN, USA
Abstract
Post-transcriptional regulation plays a critical role in gene expression, with both poly(A) tail length and chemical modifications shaping RNA stability and translation. Recent studies suggest a mechanistic link between these features, yet most current experimental protocols and computational tools analyze them independently. To address this gap, we developed polyAmod, a modular and reproducible Nextflow pipeline that enables the simultaneous prediction of multiple RNA modifications—including N6-methyladenosine (m6A) and pseudouridine (Ψ)—and poly(A) tail length at the transcript level using Oxford Nanopore Technologies (ONT) direct RNA sequencing data. The pipeline outputs comprehensive annotated BED files, enabling integrative analyses of RNA modifications and polyadenylation dynamics. We applied polyAmod to HEK293, HeLa, and GM12878 cell lines to characterize m6A modification patterns and poly(A) tail length distributions. Across these cell lines, we identified 173 conserved m6A sites, along with numerous cell type–specific sites (560 in HEK293, 347 in HeLa, and 667 in GM12878), highlighting both shared and distinct methylation landscapes. Furthermore, poly(A) tail length distributions differed significantly among the three cell lines (Kruskal–Wallis test, $p < 2.2e-16$). In HEK293 cells specifically, transcripts with a greater number of m6A sites had significantly shorter poly(A) tails (Pearson $R = -0.293$, $p = 0.005$), while m6A sites located closer to the 3' end were associated with longer tails (Pearson $R = 0.293$, $p = 0.004$). These correlations were not observed in HeLa or GM12878 cells. Analysis of additional modifications, including Ψ, is currently underway. Overall, polyAmod expands the capabilities of long-read transcriptomics by providing a unified framework for studying the interplay between RNA modifications and polyadenylation in diverse biological contexts. These findings underscore the importance of developing integrated computational tools capable of profiling multiple aspects of RNA biology simultaneously (Somalraju et al., 2025).

Keywords
Single molecule sequencing, transcriptomics, RNA modifications, poly(A) tail, RNA bioinformatics

---

Title: Single molecule sequencing, transcriptomics, RNA modifications, poly(A) tail, RNA bioinformatics
Author list: Ting Zhang[1], Hui Chen[1], Shuxin Chen[1], Stephen Montgomery[2], Qin Li[3], Lei Li[1]

Detailed Affiliations
[1]Institute of Systems and Physical Biology, Shenzhen Bay Laboratory; Shenzhen, 518055, China; [2]Departments of Pathology and Genetics, Stanford School of Medicine; Stanford, CA 94305, USA;

[3]Department of Genetics, Perelman School of Medicine, University of Pennsylvania; Philadelphia, PA 19104, USA;

Abstract
Genetic variants linked to immune disease risk often exert effects through regulatory mechanisms that vary across cell types and contexts, yet the role of post-transcriptional processes like alternative polyadenylation (APA) remains poorly understood. To address this, we analyzed single-cell RNA-sequencing data from over 7.27 million peripheral blood mononuclear cells across 2,022 individuals. We further developed a new computational method to map cell-type- and context-specific APA variation and identified 10,211 single-cell APA quantitative trait loci (sc-aQTLs) connecting 4,484 independent variants to 5,448 nearby genes, including autoimmune disease genes such as STAT6, which acts through APA independently of gene expression changes. Through colocalization and TWAS, we identified 267 APA-linked putatively causal disease genes, with 80.52% acting through expression-independent regulatory pathways. This sc-aQTL atlas highlights APA as a critical and underrecognized layer of immune disease regulation and a potential source of novel therapeutic targets.

Keywords
Single-cell, GWAS, eQTL

Title: EpiSemoLLM: A Fine-tuned Large Language Model to Predict the Epileptogenic Zone based on Seizure Semiology

Author list: Shihao Yang[1], Neel Fotedar[2], Xinglong Ju[3], Yaxi Luo[4], Danilo Bernardo[5], Vikram R. Rao[5], Xiaochen Xian[6], Hai Sun[7], Ioannis Karakis[8], Josh Laing[9], Felix Rosenow[10], Patrick Kwan[9], Shasha Wu[11], Feng Liu[12]

Detailed Affiliations
[1] Department of Systems Engineering, Stevens Institute of Technology, Hoboken, NJ, USA [2] Department of Neurology, University Hospitals Cleveland Medical Center; School of Medicine, Case Western Reserve University, Cleveland, OH, USA [3] Dixie L. Leavitt School of Business, Southern Utah University, Cedar City, UT, USA [4] Department of Computer Science, Stevens Institute of Technology, Hoboken, NJ, USA [5] Department of Neurology and Weill Institute for Neurosciences, University of California San Francisco, San Francisco, CA, USA [6] H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA [7] Department of Neurosurgery, Rutgers Robert Wood Johnson Medical School, Rutgers University, New Brunswick, NJ, USA [8] Department of Neurology, Emory University, Atlanta, GA, USA; School of Medicine and Department of Neurology, University of Crete School of Medicine, Heraklion, Greece [9] Department of Neuroscience, Monash University, Melbourne, Australia [10] Epilepsy Center Frankfurt Rhine-Main; Department of Neurology, Goethe-University Frankfurt, Frankfurt, Germany [11] Department of Neurology, University of Chicago, Chicago, IL, USA [12] Department of Systems Engineering and Semcer Center for Healthcare Innovation, Stevens Institute of Technology, Hoboken, NJ, USA

Abstract

Significance: Seizure semiology—the study of signs and clinical features during seizures—offers critical insights for localizing the epileptogenic zone (EZ). Due to its descriptive nature, interpreting semiology remains subjective and variable. With advancements in large language models (LLMs), there is potential to improve EZ localization by using LLMs to map semiological descriptions to brain regions. This study presents EpiSemoLLM, the first LLM fine-tuned specifically for seizure semiology interpretation, based on the Mistral-7B foundation model. Method: A total of 1,372 cases, each including seizure semiology and corresponding EZs validated through intracranial EEG and surgical outcomes, were compiled from 392 publications and 590 EHRs from Far Eastern Memorial Hospital (FEMH), Taiwan. This high-quality, domain-specific dataset was used to fine-tune the LLM for improved EZ prediction. EpiSemoLLM's performance was evaluated on 100 benchmark cases, with its outputs compared against those from a panel of five epileptologists. Metrics used were Regional Accuracy Rate (RAR) and Rectified Net Positive Inference Rate (rNPIR). EpiSemoLLM was also benchmarked against its base model (Mistral-7B), ChatGPT, LLaMA variants, and other biomedical LLMs. Results: In zero-shot tests, EpiSemoLLM achieved RARs of 60.71% (frontal), 83.33% (temporal), 63.16% (occipital), 45.83% (parietal), 33.33% (insular), and 28.57% (cingulate), with a mean rNPIR of 0.535. In contrast, epileptologists had RARs of 64.83% (frontal), 52.22% (temporal), 60.00% (occipital), 42.50% (parietal), 42.22% (insular), and 8.57% (cingulate), with a mean rNPIR of 0.460. Notably, EpiSemoLLM outperformed the human panel in temporal, parietal, and cingulate cortex localization and surpassed multiple state-of-the-art LLMs, including Mistral-7B-instruct, GPT-4, LLaMA, and DeepSeek-R1. Conclusion: EpiSemoLLM shows comparable performance to epileptologists in localizing EZs from seizure semiology and surpasses them in specific brain regions, particularly the temporal and insular cortices. Its superior performance over general-purpose and biomedical LLMs underscores the value of fine-tuning with high-quality, domain-specific data. EpiSemoLLM holds promise as a decision-support tool for presurgical evaluation in epilepsy care.

Keywords
Seizure semiology, large language model, seizure onset zone, AI for epilepsy.

Title: TrimNN: Characterizing cellular community motifs for studying multicellular topological organization in complex tissues

Author list: Yang Yu[1], Shuang Wang[2], Dong Xu[1,3], Juexin Wang[4]

Detailed Affiliations
[1]Department of Electrical Engineering and Computer Science, Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA; [2]Department of Computer Science, Luddy School of Informatics, Computing, and Engineering, Indiana University Bloomington, Bloomington, IN 47405, USA; [3]Institute for Data Science and Informatics, University of Missouri, Columbia, MO, 65211, USA; [4]Department of Biomedical Engineering and Informatics, Luddy School of Informatics, Computing, and Engineering, Indiana University Indianapolis, Indianapolis, IN 46202, USA;

Abstract
The spatial organization of cells plays a pivotal role in shaping tissue functions and phenotypes in various biological systems and diseased microenvironments. However, the topological principles governing interactions among cell types within spatial patterns remain poorly understood. Here, we present the

Triangulation cellular community motif Neural Network (TrimNN), a graph-based deep learning framework designed to identify conserved spatial cell organization patterns, termed cellular community (CC) motifs, from spatial transcriptomics and proteomics data. These identified CC motifs are biologically interpretable through a set of downstream analyses, including motif visualization, cellular-level interpretation within cell–cell communication analysis, gene-level interpretation within differentially expressed gene and pathway analysis, and phenotypical analysis within the availability of phenotypical information. TrimNN employs a semi–divide-and-conquer approach to efficiently detect overrepresented topological motifs of varying sizes in a triangulated space, which simplifies the intricate task of occurrence regression by decomposing it into several binary present/absent predictions on small graphs. TrimNN is trained using representative pairs of subgraphs and triangulated cell graphs to estimate overrepresented network motifs. On typical spatial omics samples within thousands of cells in dozens of cell types, TrimNN robustly infers the presence of a large-size network motif in seconds. By uncovering CC motifs, TrimNN reveals key associations between spatially distributed cell-type patterns and diverse phenotypes. These insights provide a foundation for understanding biological and disease mechanisms and offer potential biomarkers for diagnosis and therapeutic interventions. In case studies, TrimNN identifies computationally significant and biologically meaningful CC motifs to differentiate patient survival in colorectal cancer studies and represents pathologically related cell type organization in neurodegenerative diseases and colorectal carcinoma studies.

Keywords
Spatial Omics, Cellular Community motif, Delaunay Triangulation, Graph-based Deep Learning

---

Title: Agentic AI-Driven Adaptive Framework for Omics Data Analysis: An RNA-Seq Perspective
Author list: James Li[1]

Detailed Affiliations
[1]Department of Biostatistics, Bioinformatics & Biomathematics, Georgetown University, Washington, DC, USA

Abstract
Omics data analysis is critical for biomarker discovery and plays a significant role in addressing genetic-related diseases, such as cancer, therefore benefiting pharmaceutical developments and patient care. Despite its potential, omics analysis presents numerous analytical and integrative challenges. Using RNA-Seq data as an illustrative example, standard analytical workflows typically involve rigorous quality assurance (QA/QC), sequence alignment, normalization, differential expression analysis, dimensionality reduction (e.g., tSNE and PCA pre and post differential expression analysis), and subsequent biomarker detection through classification models. Each analytical step needs careful algorithmic selection, for instance, choosing between edgeR or DESeq2 for differential expression, or Random Forest versus Gradient Boosting for classification, posing considerable obstacles and requiring substantial manual intervention. Further complicating matters is the integration of diverse omics datasets, such as miRNA seq, whole genome sequencing, single cell RNA seq, proteomics, metabolomics, and lipidomics, into comprehensive systems biology studies.

To overcome these challenges, we leveraged Agentic AI framework, specifically using OpenAI's GPT-4o as the primary Large Language Model (LLM) to embed autonomous decision-making capabilities within analytical pipeline. As the center of the proposed approach, LLM functions as an intelligent decision-making hub, dynamically evaluating dataset specific characteristics to autonomously select optimal analytical strategies. The implementation utilizes LangChain and LangGraph to orchestrate communication between the LLM and modularized analytical components. Each analytical component is modularized into independent Minimum Viable Products (MVPs), enabling adaptive selection and automated integration. This reduces manual effort, minimizes subjective bias, enhances reproducibility, and increases analytical scalability.

The initial implementation focuses on RNA-Seq data analysis, and multiple MVPs were developed for key analytical components: differential expression analysis using DESeq2 and edgeR, dimensionality reduction using scikit-learn based tSNE and PCA, and biomarker classification using Random Forest and Gradient Boosting models. The LLM driven framework autonomously identified and combined appropriate MVPs through LangChain and LangGraph workflows, achieving analytical results comparable to traditional manually curated methods.

Preliminary findings clearly demonstrate that GPT-4o driven Agentic AI frameworks, supported by LangChain and LangGraph, can effectively streamline complex analytical workflows, which significantly enhances biomarker discovery and pharmaceutical development, and ultimately improves patient outcomes. Future research will comparatively evaluate additional GPT models (e.g., GPT-4 Turbo, GPT-4.1, GPT-3.5 Turbo), open-source models such as LLaMA, and custom fine-tuned models specifically tailored for omics data analysis. Efforts will also focus on broadening the MVP library, expanding applicability to more diverse omics data types, and enhancing multi omics integration to facilitate comprehensive systems biology research.

Keywords
Agentic AI, LLM, Omics, RNA-Seq, Biomarker Discovery, Automation

---

TITLE: MULTISCALE FUNCTIONAL NETWORK CONNECTIVITY REVEALS PROGRESSIVE BRAIN NETWORK DISRUPTION ALONG THE DEMENTIA SPECTRUM.

Author list: Susan Oluwatominiyi Kadri [1,3], Nigar Khasayeva [1,2], Ram Ballem [1,2], Prerana Bajracharya [1,2], Kyle M. Jensen [1,4], Armin Iraji [1,2,3,4]

Detailed Affiliations
[1]Tri-Institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), [2]Computer Science Department, Georgia State University, [3]Mathematics and Statistics Department, Georgia State University, [4] Psychology Department, Georgia State University.

Abstract
Alzheimer's disease (AD) is a progressive neurological disorder that impairs memory, cognition, and daily functioning by disrupting communication between brain regions. Detecting early changes in brain connectivity may offer a valuable opportunity for earlier diagnosis and intervention. This study aimed to identify patterns of brain network disruption across the AD continuum using resting-state functional

magnetic resonance imaging (rsfMRI), a technique that captures spontaneous brain activity when participants are not performing a task. This allows researchers to examine how different brain regions naturally interact over time.
We analyzed rsfMRI data from 2,462 participants enrolled in the Alzheimer's Disease Neuroimaging Initiative (ADNI), including 1,089 cognitively normal individuals (CN), 875 with mild cognitive impairment (MCI), and 322 diagnosed with AD. To investigate how networks across the brain interact, we used multiscale functional network connectivity (msFNC), which evaluates functional interactions across different spatial resolutions. Intrinsic connectivity networks (ICNs) were first identified using spatially constrained independent component analysis (scICA), guided by the NeuroMark 2.2 template. NeuroMark 2.2 provides standardized, reproducible definitions of ICNs across multiple domains, including sensorimotor (SM), higher cognitive (HC), temporoparietal (TP), default mode (DM), subcortical (SC), and cerebellar (CB) systems. msFNC was estimated by calculating Pearson correlation between ICNs time series.
A linear mixed-effects model (LME) was used to assess group-level differences in msFNC, while controlling for age, sex, head motion, scanning site, and subject ID. Our results revealed that connectivity changes emerge early, particularly in the CB–SM, HC–TP, and SC–HC interactions. These domains are associated with functions such as motor coordination, memory, and attention, which are often affected in the early stages of cognitive decline. In the MCI group, significant hypoconnectivity was observed across these circuits, with some focal increases that may reflect early compensatory processes. As AD progresses, we observed broader and more severe disruptions, especially within and between the DM, SC, and HC networks, reflecting breakdowns in self-referential thinking, executive function, and large-scale integration.
These findings support the use of msFNC, combined with NeuroMark 2.2-derived ICNs, as a robust and non-invasive framework for detecting subtle and progressive alterations in brain networks in AD. This approach offers potential for enhancing early detection, improving diagnostic accuracy, and tracking disease progression over time.

Title: PrismNet: A Lightweight Multimodal AI Screening Tool for Skin Lesion Classification
Author list: Nicholas Wei, Austin Xu
Detailed Affiliations
Williamsville East High School, Buffalo, NY*, USA*

Abstract
Background: Melanoma is the most lethal skin cancer, with fatality primarily due to late detection and lack of prompt treatment. Previous AI studies have shown that dermatologist-level accuracy can be achieved under controlled conditions 1-3. However, these systems focus on definitive diagnosis and are resource-intensive with limited accessibility to the general community. Moreover, their reliance on cloud computing raises additional privacy concerns. A widely accessible AI screening aid that can identify early melanomas on a personal device (e.g., a smartphone or computer) and facilitate prompt evaluation by dermatologists is potentially lifesaving. We therefore developed PrismNet, a lightweight AI screening tool capable of running on personal computers. It evaluates suspicious skin lesions in order to identify previously unnoticed cancers and encourage timely referral to dermatological care.

Objective: PrismNet is an AI system for at-home early melanoma screening. Designed for high sensitivity to minimize false negatives, it flags suspicious lesions for prompt dermatological review. Its high-sensitivity enables negative results to be used to rule out benign lesions, while its compact architecture enables fully local operation, safeguarding patient privacy.
Methods: PrismNet uses a single-stage multimodal architecture based on ConvNeXt-Tiny. Dermoscopic close-ups are combined with pre-computed binary lesion masks as a four-channel input. A 13-dimensional clinical metadata vector (age, sex, lesion location) passes through two linear layers to produce gamma and beta parameters for feature-wise linear modulation. This novel framework unifies segmentation masks, metadata fusion via FiLM, and a lightweight ConvNeXt-Tiny backbone. It was trained on >50k images from ISIC (International Skin Imaging Collaboration) datasets with a 90/10 train-test split.
Results: To fulfill various clinical needs, we produced two models: a high-sensitivity model and a balanced model. On an independent validation set of 200 dermoscopic images, PrismNet's high-sensitivity model achieved a 99% sensitivity (95% CI 94.6–99.8) with a 67.5% overall accuracy. This model prioritizes minimal false negatives, tolerating lower accuracy to ensure nearly all cancers are flagged. The second, balanced model reached a 97% sensitivity (95% CI 91.6–98.9) and about 75% accuracy. Both models are compact (~29 MB), allowing real-time inference (~0.5 s per image) on ordinary personal devices.
Conclusion : By combining near-perfect sensitivity with a compact, privacy-preserving design, PrismNet provides a lightweight screening tool for general users to detect melanoma earlier, potentially reducing the incidence of late-stage cancers at diagnosis. Its fully offline operation facilitates compliance with privacy regulations such as HIPAA and GDPR. Future work will strengthen performance by training on larger, more diverse cohorts and validating across populations, while also expanding accessibility to high-risk groups—especially people with high UV exposure in underserved communities.

Title: Machine-learning-facilitated Drug Discovery for Potential Aberrant Connexin-26 Hemichannel Inhibition
Author list: Chenye Xue

Detailed Affiliations
Arcadia High School

Abstract:
Connexins are a family of transmembrane proteins that form gap junctions, enabling intercellular communication and maintaining cellular homeostasis through the passage of ions and small molecules. Among them, Connexin 26 (Cx26), encoded by the GJB2 gene, plays a critical role in the inner ear. Mutations in GJB2 can cause aberrant hemichannel opening, altering pore electrostatics and disrupting normal cell signaling. These gain-of-function mutations are associated with various forms of hereditary hearing loss.
Although Cx26 is a promising therapeutic target, little is known about the molecular features that confer selective inhibition, and the conventional drug development process remains slow and resource-intensive. Machine learning offers a powerful alternative by enabling rapid screening and prediction of potential inhibitors.
This study proposes using machine learning models, implemented through the Weka software, and existing ligand-receptor data for Cx26 inhibitors to identify therapeutic targets for GJB2-related hearing loss. We

propose examining existing ligand-receptor data for both Cx26 inhibitors and compounds targeting other connexins that share conserved binding site sequences, intending to uncover small molecules capable of modulating aberrant Cx26 activity.

Keywords
Connexin 26 (Cx26), GJB2 mutations, hereditary hearing loss, machine learning, gap junctions, pharmacology

---

Title: TransRef-ICA: A Transfer Learning Framework Using Reference-Informed Source Estimation with Application to Dementia

Author list: Abtin Mirzaiesaran1, Vince Calhoun1, Armin Iraji1,

Detailed Affiliations
1Tri-Institutional Center for Translational Research in Neuroimaging and Data Science (Georgia State University, Georgia Institute of Technology, Emory University), Atlanta, GA, USA

Abstract
Alzheimer's disease involves progressive disruptions in brain organization across multiple biological scales. Understanding how alterations in neuroanatomy, functional dynamics, and molecular pathology relate to each is still in general, a challenge. Although methods like joint independent component analysis (jICA) are designed to identify covarying cross-modal patterns, they are limited in their ability to capture how modality-specific features from one domain transfer to or inform representations in another. To address this, we introduce a novel functionally informed multimodal framework that combines biologically grounded priors with advanced data-driven analysis. Specifically, we employ Multivariate-Objective Optimization ICA with Reference (MOO-ICAR) using the NeuroMark 2.2 template, derived from over 100,000 resting-state fMRI datasets, to guide ICA decomposition of T1-weighted structural MRI and Florbetapir PET data gathered from 444 subjects spanning across the Alzheimer's disease continuum.
This approach yields spatially consistent and functionally meaningful structural and molecular components across both modalities. We use ICA loadings, which indicate the contribution of each component to the subject data, to evaluate the diagnostic effect. We also studied co-activation of these component maps for each subject within PET, T1, and between PET and T1. Linear mixed-effects models tested for diagnostic group differences in coupling strength, controlling for age, sex, site, and race.
Of particular interest, we found that in PET data, individuals with mild cognitive impairment (MCI) showed increased coupling between the salience network (TN-SA) and the frontal higher cognition domain (HC-FR), and decreased coupling between TN-SA and the cerebellum (CB), revealing an anterior-posterior decoupling pattern that appear at the MCI stage and notably stable in the progression towards dementia. The stage-specific emergence and subsequent plateau of this pattern highlight its potential as a sensitive early biomarker for prodromal Alzheimer's disease, with possible implications for early diagnosis and targeted intervention.
By jointly examining PET and T1 components within a cross-modal framework, we identified previously inaccessible patterns of component reorganization. In dementia, the T1-defined paralimbic domain showed stronger coupling with PET-based subcortical regions (e.g., extended hippocampus and thalamus) and weaker coupling with frontal and triple-network systems. This domain also exhibited the strongest and most consistent group differences in both T1 coupling and ICA loadings, highlighting it as a key locus of

degeneration. Together, these findings underscore the utility of a unified ICA-based approach for uncovering interpretable multimodal biomarkers and offer fresh insights into evolving structure-function relationships in Alzheimer's disease.

CONFERENCE LOCATION

Pomerene Hall

1760 Neil Ave, Columbus, OH 43210

The Translational Data Analytics Institute (TDAI) is a community of 1000+ faculty, researchers, staff and students from 70+ disciplines working at the forefront of interdisciplinary, big data-enabled science, scholarship and creative expression, with an emphasis on discoveries, solutions and insights for the greater good.

Parking Information

For guests wanting to park onsite, 12th Avenue Garage (340 West 12th Avenue, Columbus, OH 43210) offers visitor parking with a $15 daily maximum ($9 off-peak Max). The 12th Avenue Garage entrance is one block away from Pomerene Hall. https://osu.campusparc.com/find-parking/12th-avenue-garage/

View Map

Airport Information

John Glenn Columbus International Airport: 10 miles from the Conference Location.

Hotel Information

The Blackwell Hotel

Map and Directions

Special Conference Room Rate: $172 + tax and fees, Double or King.

Reservation details: please use this link to make a reservation.

Wifi

The WiFi@OSU wireless network is provided by the Office of Technology and Digital Innovation (OTDI). Guests can use WiFi@OSU for basic Internet access after accepting the university acceptable use policy.

SPECIAL ACKNOWLEDGEMENTS

We thank many people who helped with the peer-review of the manuscripts submitted to the conference. We are grateful for the numerous volunteering help and support from many people. We thank the Managing Director Cathie Smith, Brandon Elmore and Amanda Jovanovich from TDAI for their support and coordination of the events.

MANY THANKS TO OUR SPONSORS!

10x Genomics was founded on the vision that this century will bring advances in biomedicine and transform the way we understand and treat disease. We deliver powerful, reliable tools that fuel scientific discoveries and drive exponential progress to master biology to advance human health. Our end-to-end single cell and spatial solutions include instruments, consumables, and intuitive software, letting you unravel highly intricate biological systems, while bringing into focus the details that matter most.

Volume 118, October 2025 ISSN 1476-9271

Computational Biology and Chemistry

Editors: **Qin Ma, Donald Hamelberg**

www.elsevier.com/locate/cbac

Available online at www.sciencedirect.com

ScienceDirect

Olink's mission is to accelerate proteomics together with the scientific community, to understand real-time biology and gain actionable insights into human health and disease. Our innovative solutions deliver highly sensitive and accurate protein quantification, giving scientists the power to investigate complex biological processes with precision.

One platform. Endless possibilities.

Explore up to 5,400 proteins with high specificity, transparent data, and the flexibility to answer any research question. Meet the next-generation proteomics platform trusted by the scientific community, from small academic research teams through to leading pharma companies.

Singleron

From single cell multi-omics to precision medicine

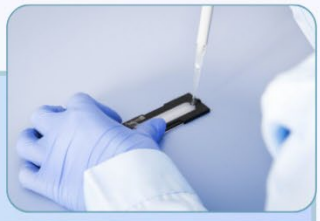
TISSUE PRESERVATION & DISSOCIATION

- Preserve tissue integrity for up to 72 hours - Maintain sample quality during transport or processing delays
- Automated and flexible - Adaptable programs to suit your needs
- Generate clean cell suspensions and isolate nuclei with ease

FLEXIBLE SINGLE CELL EXPERIMENTS

Manual workflow:
Instrument-free option for flexible, low- throughput needs

Automated workflow:
Fully automated cell partitioning & barcoding system generating sequencing-ready libraries



MULTIOMICS KITS

- (Full-length) transcriptome
- Full-length immune profiles
- Targeted variant detection
- Time-resolved transcriptomics
- Cell surface glycosylation
- Combined genome & transcriptome

GET IN TOUCH

Service lab: Singleron Michigan
333 Jackson plaza, Ann Arbor, MI 48103
+1 734-249-0883