NIDDK-CR Data Science and Meet the Expert Webinar Series

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

About the Series

- Aims to accelerate data science and AI-driven biomedical research by fostering collaboration between biomedical researchers and experts in the field
- Monthly webinar held on the **last Thursday of each month***

Upcoming Webinars

- Overview of CADR requirements
- Tools for whole-slide imaging pathomics
- Fusion Tool Demo

Learn more about the webinar series, register for future webinars, and access past webinars materials and recordings

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Data Science Centric Challenge Series

Goals of NIDDK-CR Data-science centric challenge series

- Develop tools, approaches, models and/or methods to increase data interoperability and usability for artificial intelligence (AI) and machine learning (ML) applications
- Augment and enhance existing data for future secondary research, including data-driven discovery by AI/ML researchers
- Discover innovative approaches to enhance the utility of datasets for AI/ML applications

Visit our website for more information on our data-centric movement and to learn more about our past data-challenges

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

The Analytics Workbench V1.0

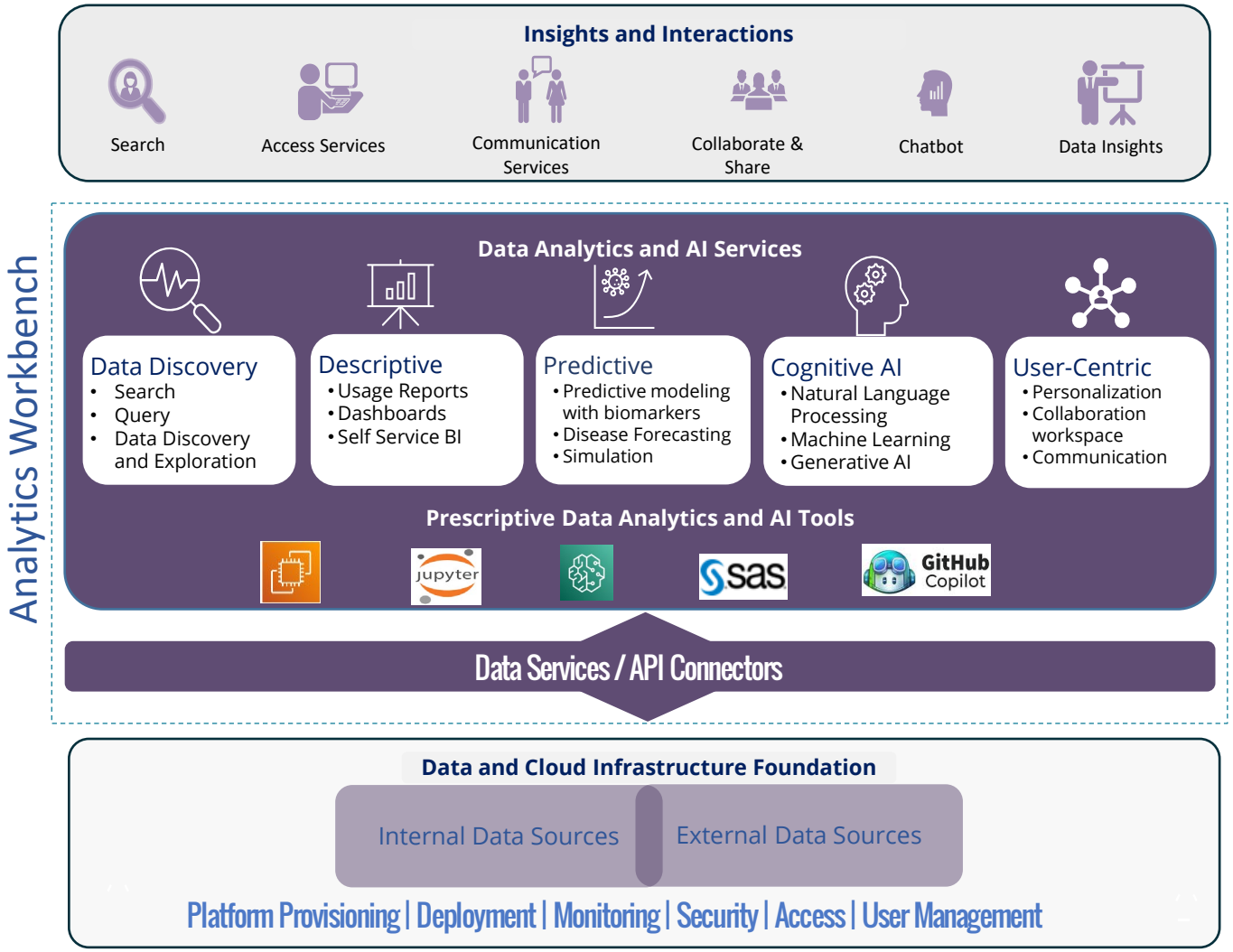
Streamlining *end-to-end data science lifecycle and discovery of data-driven biomedical insights.*

Innovation and ease of use

A cloud-based analytics environment where researchers and data scientists can access a suite of integrated analytics tools and cloud computing resources to participate in data challenges and AI innovation.

Expected Benefits of Analytics Workbench:

- **Promote Collaboration**
- **Support AI Innovation**
- **Minimize Data Movement**
- **Improve User Experience**
- **Discover Data Insights**
- **Advance NIDDK Research Mission**



Analytics Workbench is integrated with the *NIDDK-CR R4R Platform* and *Challenge Management Platform* offerings. More information at https://repository.niddk.nih.gov/home

National Institute of Diabetes and Digestive and Kidney Diseases

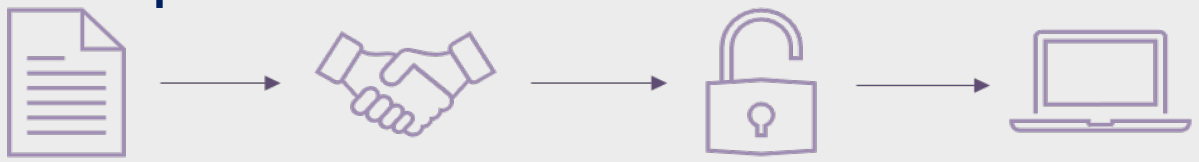
Central Repository

Co-Analysis Pilot

COMING SOON!

A streamlined, joint access pathway unlocking cross-repository data science across metabolic, renal, digestive, and aging research

How the pilot works:



Submit one request → **Single joint DAC review** → **Access both repositories** → **Analyze in secure Workbench**

Cross-Repository Study Collections:

NIDDK CR

LookAHEAD; DPP; DPPOS; DCCT; EDIC; CRIC

NIA AgingResearchBiobank

LIFE; ENERGISE; SWAN; CALERIE; ASPREE; MOST; SOF; STRIDE; Testosterone Trials

- Larger, more diverse combined cohorts; greater statistical power and generalizability
- Well-characterized, longitudinal phenotyping linked to specimens across midlife–older age
- Enables novel cross-domain questions

Built for Co-Analysis, Not Just Co-Access

- Approved investigators can analyze data from both repositories together inside the secure NIDDK CR R4R Analytics Workbench, ***supporting integrated modeling, feature extraction, and AI/ML-driven pattern discovery without moving data out of a compliant environment.***

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Meet the Experts

Dr. Juexin Wang is an assistant professor in the Department of Biomedical Engineering and Informatics at Indiana University Indianapolis. Trained in computer science, Dr. Wang's passion is answering biological and pathological questions using computational approaches. His major research interests are deep learning and machine learning for single-cell and spatial omics.

National Institute of Diabetes and Digestive and Kidney Diseases

Central Repository

Meet the Experts

Dr. Michael Eadon is a versatile physician scientist with expertise in translational molecular analysis, clinical trials, nephrology, and clinical pharmacology. He serves as the David M. and Julie B. DeWitt Professor in Medicine at Indiana University School of Medicine. His primary research efforts focus on the relationship between transcriptomic and genomic markers to drug efficacy and adverse events relevant in renal disease. These efforts include a focus to define the molecular pathogenesis of endothelial dysfunction in diabetic kidney disease.

Empower biological and pathological discovery with machine learning in the age of spatial omics

Juexin Wang, Ph.D.
Department of Biomedical Engineering and Informatics
Luddy School of Informatics, Computing, and Engineering
Indiana University Indianapolis

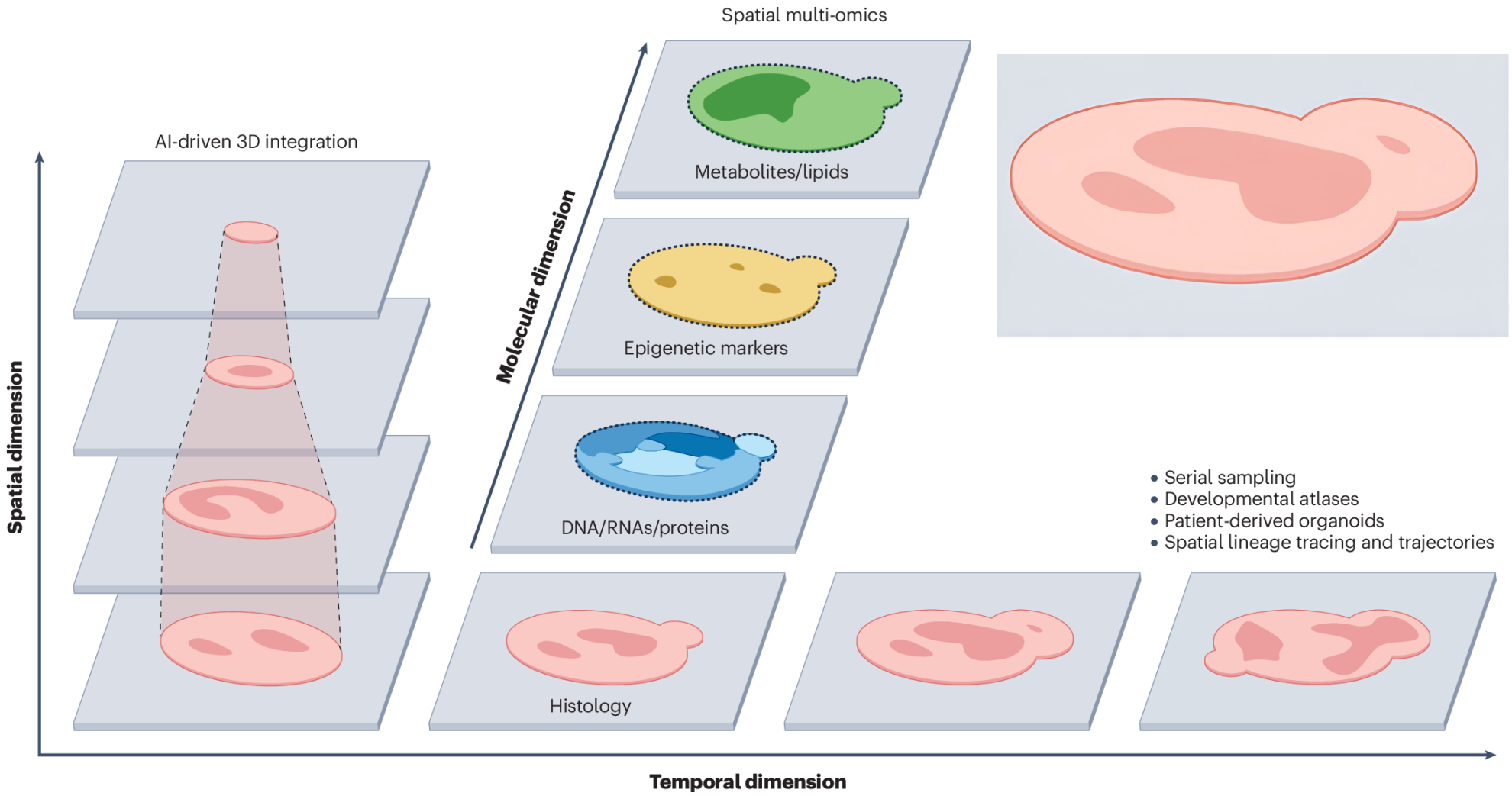
Michael T. Eadon, MD
David M. and Julie B. DeWitt Professor in Nephrology Research
Department of Medicine
Indiana University School of Medicine

08/25/2026

NIDDK-CR Data Science and Meet the Expert Webinar Series

INDIANAPOLIS

Multi-omics in spatial temporal to decode complex biological system



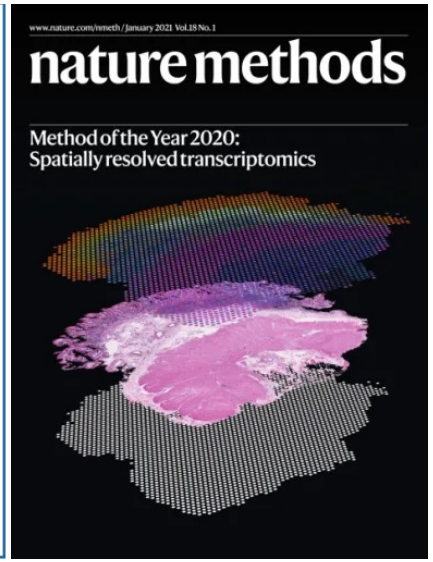
Lázár, E., Lundeberg, J. Spatial architecture of development and disease. *Nat Rev Genet* (2025)

- **Spatial omics** technologies, encompassing both imaging and sequencing-based methodologies, has enabled **a comprehensive understanding of biological and pathological processes from a cellular ecosystem perspective.**

- Data analysis is one of the major challenges

Video by Grok

Method of the year in a row, enables spatial omics atlas in applications



HUMAN CELL ATLAS

HuBMAP

Human Reference Atlas

- 2013 **Single cell sequencing** (Method of the year by Nature Methods)
- 2019 **Single cell multimodal omics** (Method of the year by Nature Methods)
- 2020 **Spatially resolved transcriptomics** (Method of the year by Nature Methods)
- 2022 **Spatial Omics**: one of the top seven technologies to watch (Nature)
- 2024 **Spatial proteomics** (Method of the year by Nature Methods)

HTAN
HUMAN TUMOR ATLAS NETWORK

KIDNEY PRECISION MEDICINE PROJECT

Outline: Can we answer biological and pathological questions using machine learning enabled by spatial omics?

- Q1: How to resolve complex biological heterogeneity in spatial transcriptomics data?
- Q2: How to identify molecular markers that define biologically meaningful spatial organization?
- Q3: How does spatial organization link to tissue functions?
- Discussion and perspective

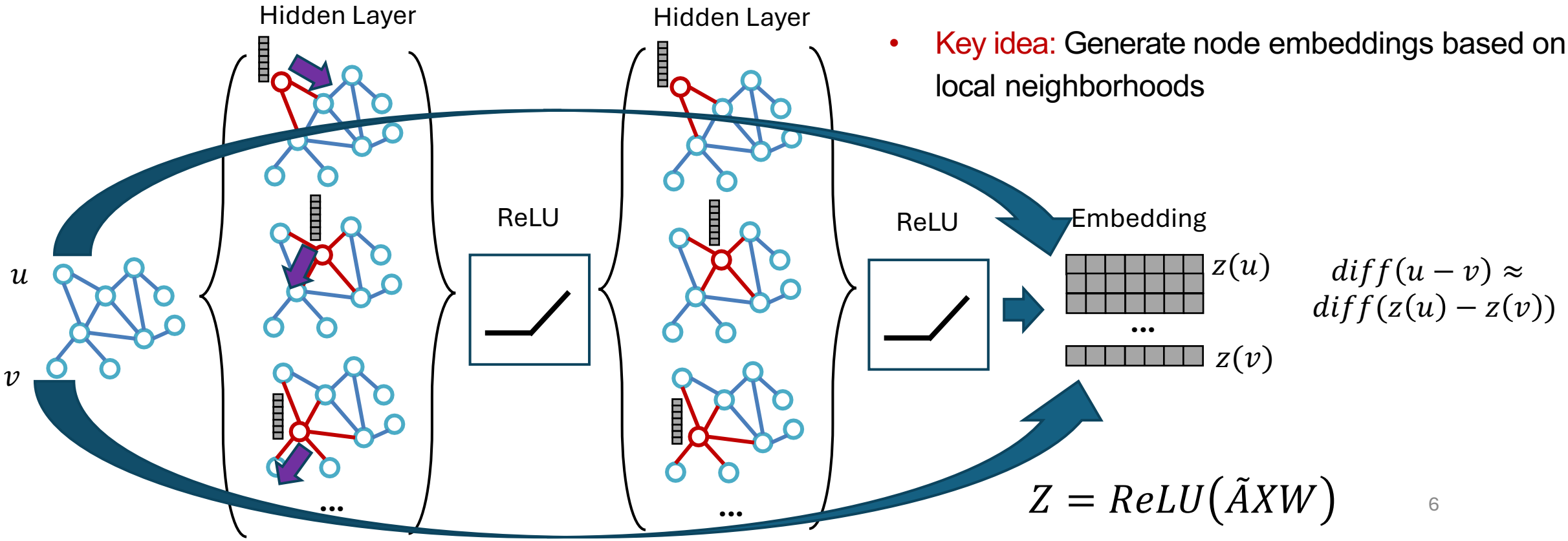
Image credit: Gemini

Outline

- **Q1: How to resolve complex biological heterogeneity in spatial transcriptomics data?**
- Q2: How to identify molecular markers that define biologically meaningful spatial organization?
- Q3: How does spatial organization link to tissue functions?
- Discussion and perspective

Graph modeling is a natural way to model relations between entities

- Graph Neural Networks (**GNN**) learns a task-independent representation of a graph by deconvoluting node relationships through neighbor information propagation in a deep learning architecture.

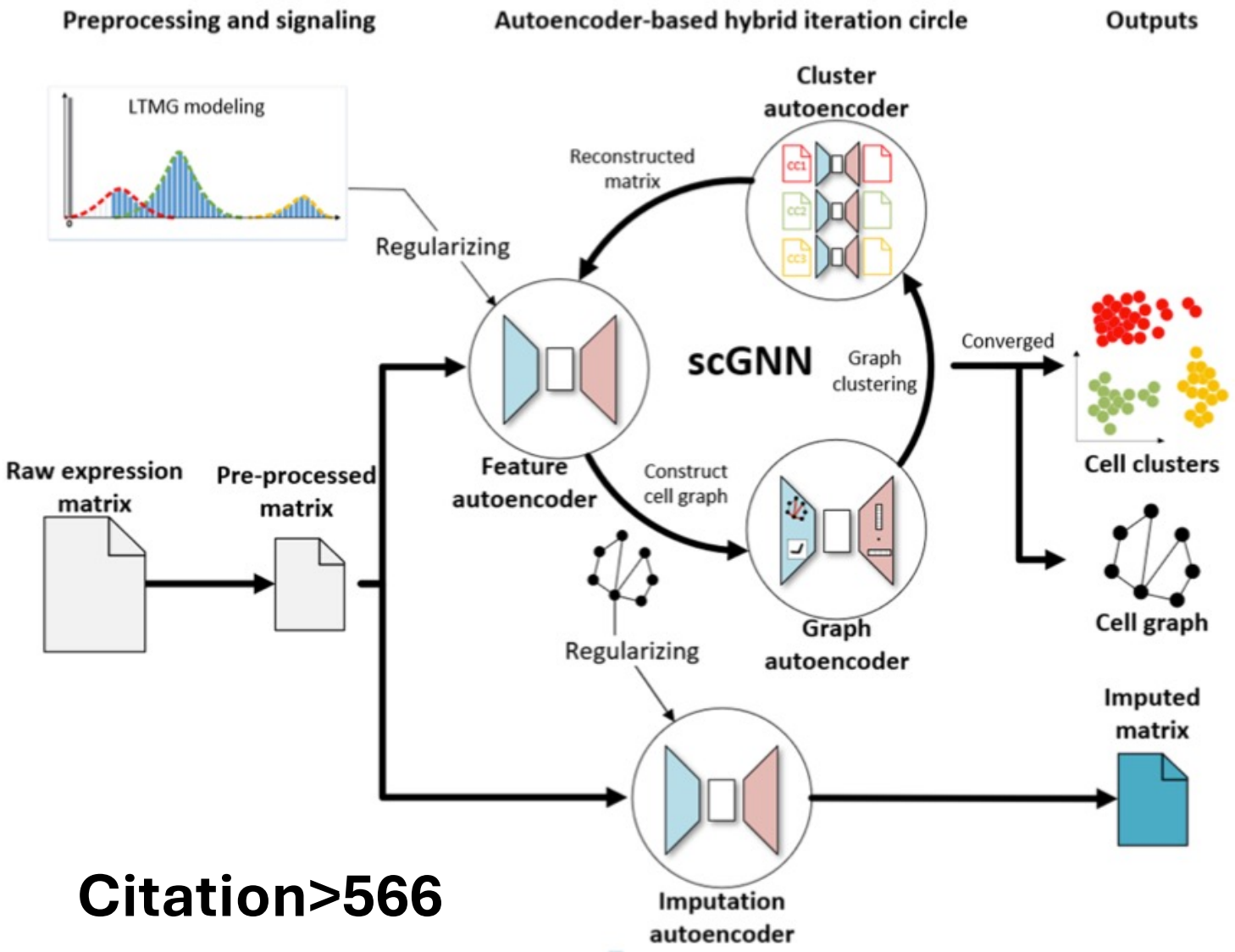


- Key idea: Generate node embeddings based on local neighborhoods

$$diff(u - v) \approx diff(z(u) - z(v))$$

$$Z = ReLU(\tilde{A}XW)$$

scGNN (**single cell** Graph Neural Networks)



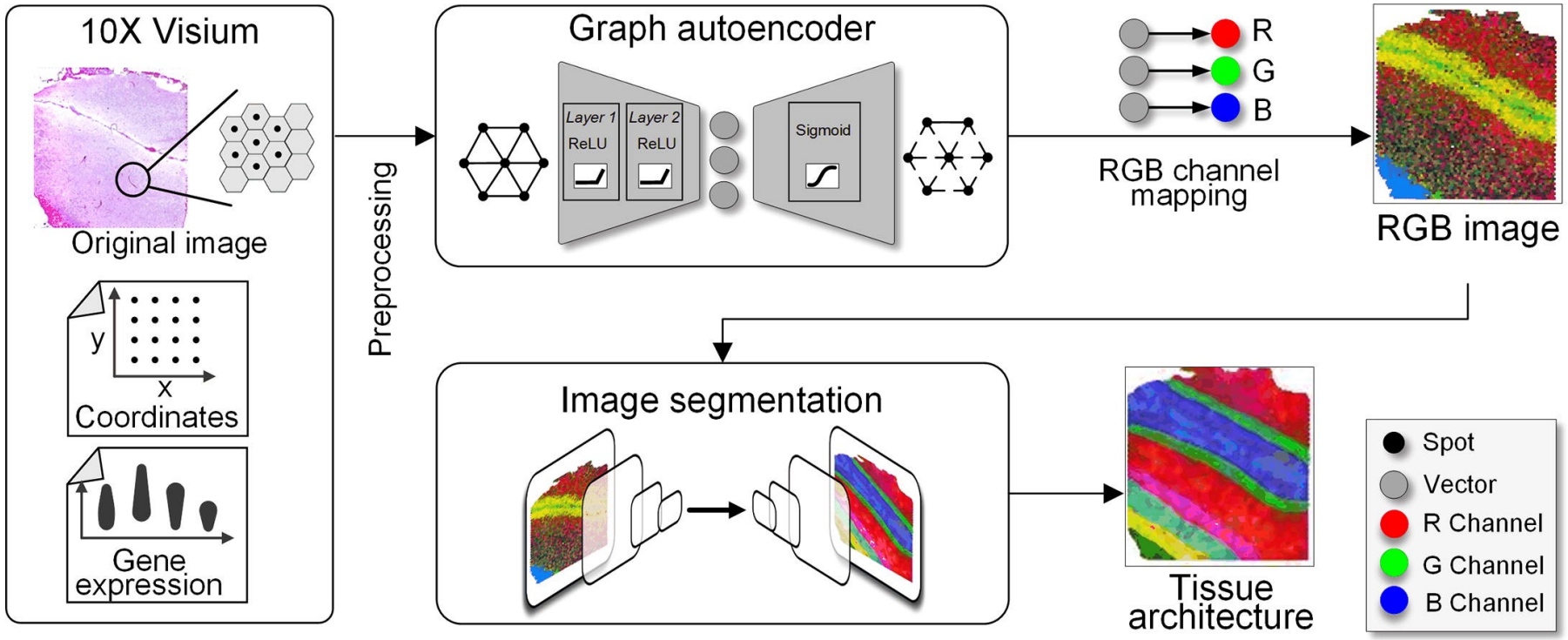
Citation>566

- **First GNN model on single cells**
- Simultaneous cell clustering and gene imputation
- Unsupervised multi-model deep learning framework
- Graph neural network + stacks of autoencoders
- Iterative build the cell graphs

Juexin Wang*, Anjun Ma*, Yuzhou Chang, Jianting Gong, Yuexu Jiang, Ren Qi, Cankun Wang, Hongjun Fu, Qin Ma, and Dong Xu. "scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses." *Nature communications* 12, no. 1 (2021): 1-11.

RESEPT: tissue architecture visualization and identification

- Graph representation learning of spatially resolved transcriptomics data for visualization
- Image segmentation for tissue architecture identification



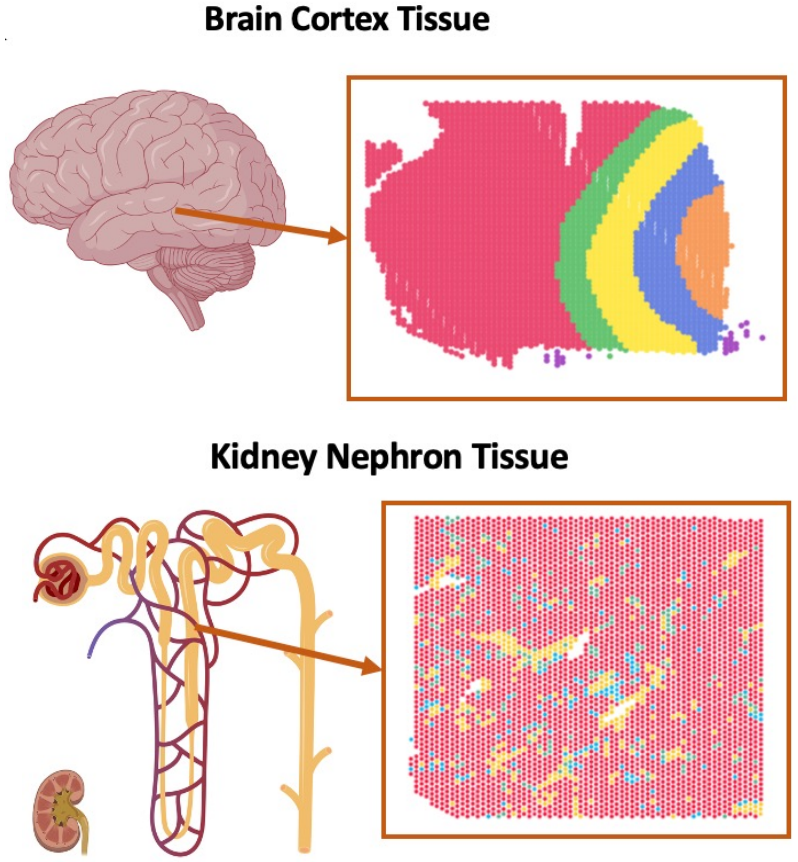
Yuzhou Chang*, Fei He*, **Juexin Wang***, et al. "Define and visualize pathological architectures of human tissues from spatially resolved transcriptomics using deep learning." *Computational and Structural Biotechnology Journal,* 20:4600-4617, 2022

Challenges for kidney SRT

Dr. Michael Eadon
Indiana University

Dr. Qin Ma
Ohio State U

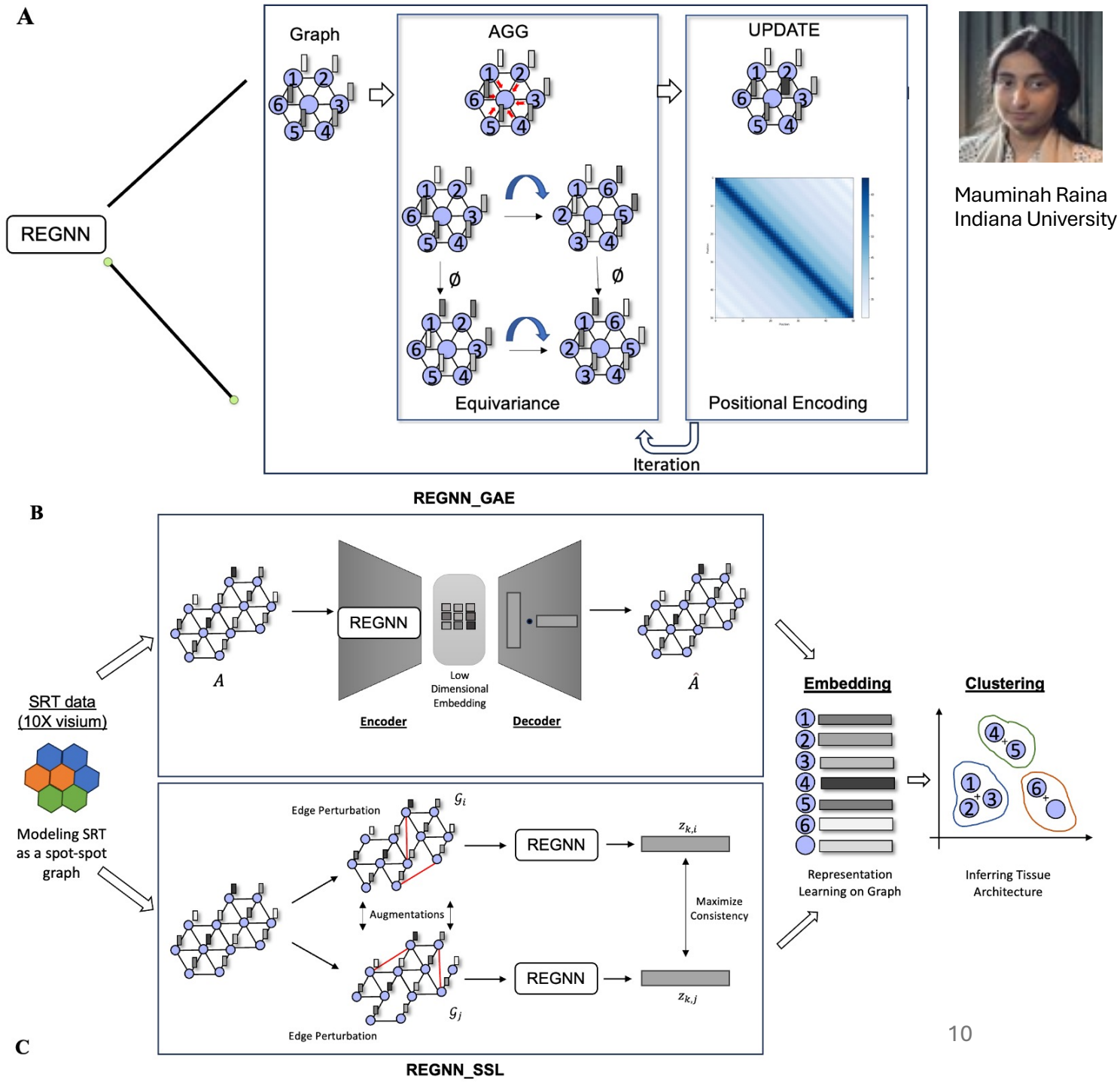
- **The kidney's heterogeneous, sparse, and mosaic-like cell types**
  - Ribbon distributed brain tissues
  - Major kidney cell types and their sub-types are scattered, different cell types being in close approximation to one another.
- **Limited *expressive power* of graph neural networks**
  - Classic GNNs use message passing theoretically limited by the 1-order Weisfeiler-Lehman (1-WL) test
  - Lattice structure topology of the RST spots makes differentiation between nodes difficult in the modeled graphs due to similarity.
- **Efficiently annotating and interpreting the vast omics datasets that are generated in a timely and cost-effective manner**
  - Classical deep learning approaches usually rely on large amounts of well-labeled data to train their models.



Raina, M., Cheng, H., Ferreira, R. M., Stansfield, T., Modak, C., Cheng, Y. H., ... **Eadon, M***, Ma, Q, & **Wang, J***. (2025). Relation equivariant graph neural networks to explore the mosaic-like tissue architecture of kidney diseases on spatially resolved transcriptomics. *Bioinformatics*, 41(6), btaf303.

REGNN (Relation Equivariant Graph Neural Networks)

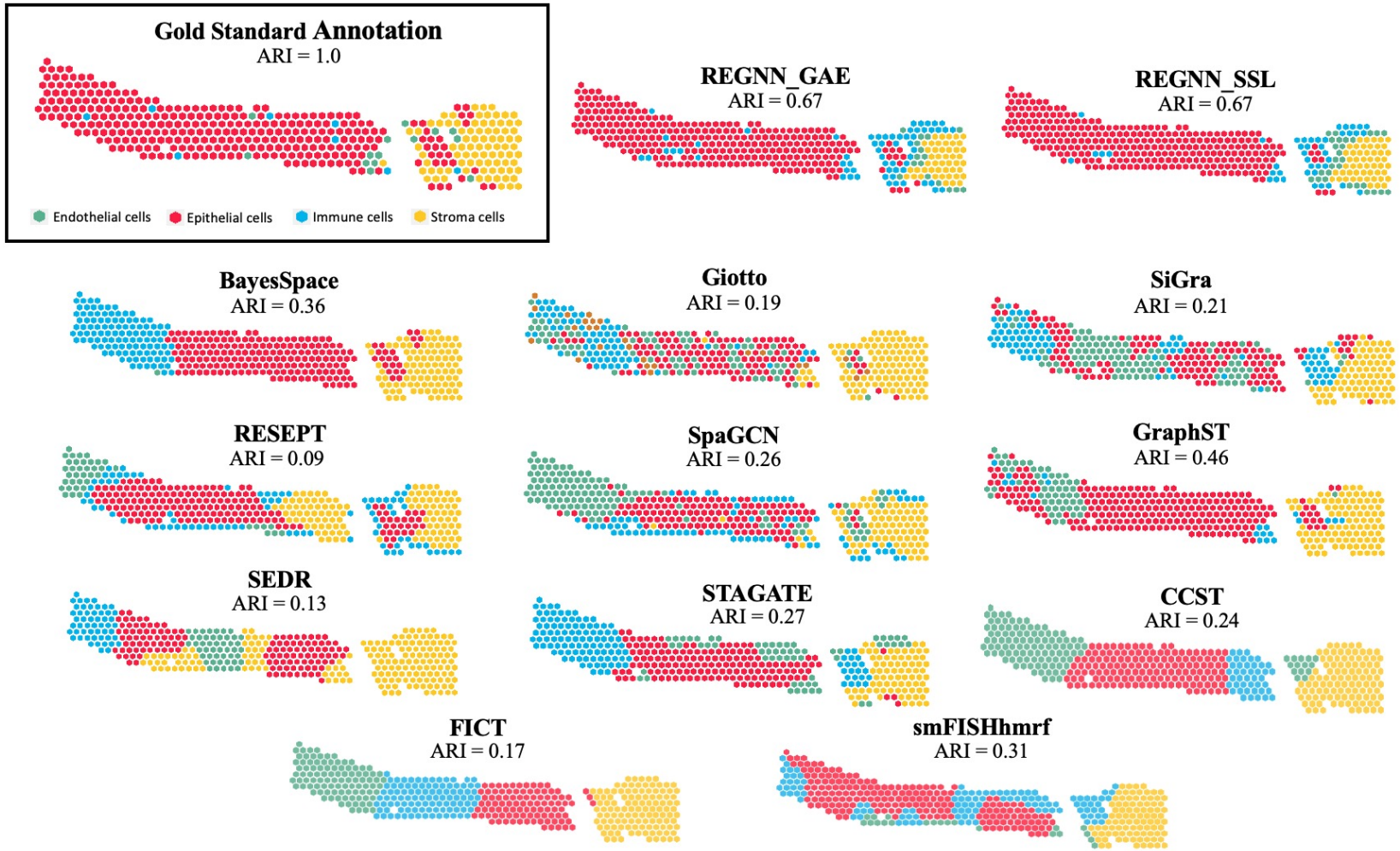
- Goal: **Increasing expressive power** of the graph neural network
- E(n) equivariant GNN
  - Symmetric nature in translation, rotation, reflection, and permutation equivariance of the lattice topology
- Positional Encoding (PE)
  - Mark the unique position of each graph node; aids in identifying the relative spatial relations of the nodes
- Self-supervised learning
  - Learn the random masked nodes/edges/subgraph



Mauminah Raina
Indiana University

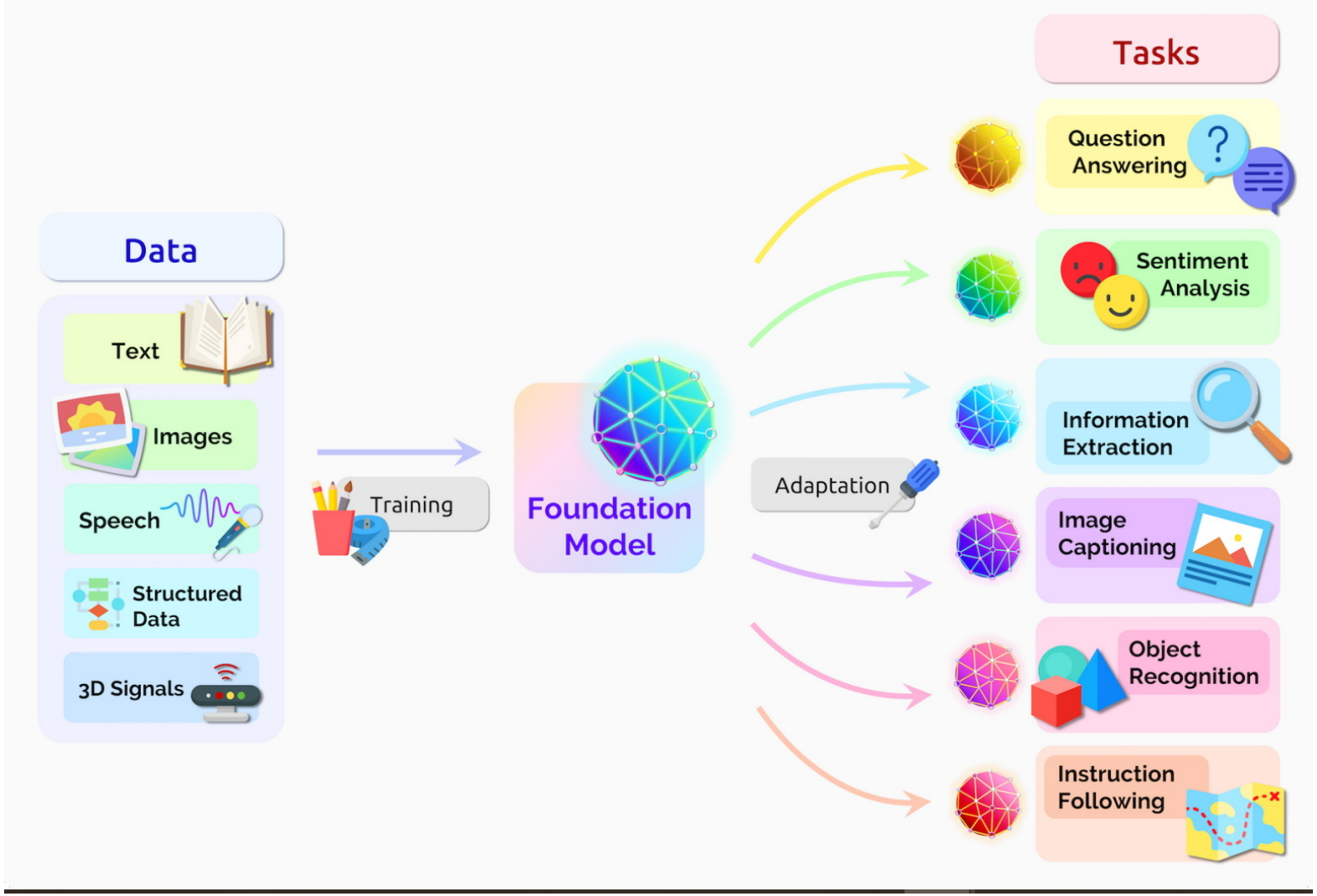
Visualization

- DKD sample V10S14-085_XY04_21-0057



Foundation model

- Any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks
- **A**: Algorithm
- **B**: Business
- **C**: Computing
- **D**: Data



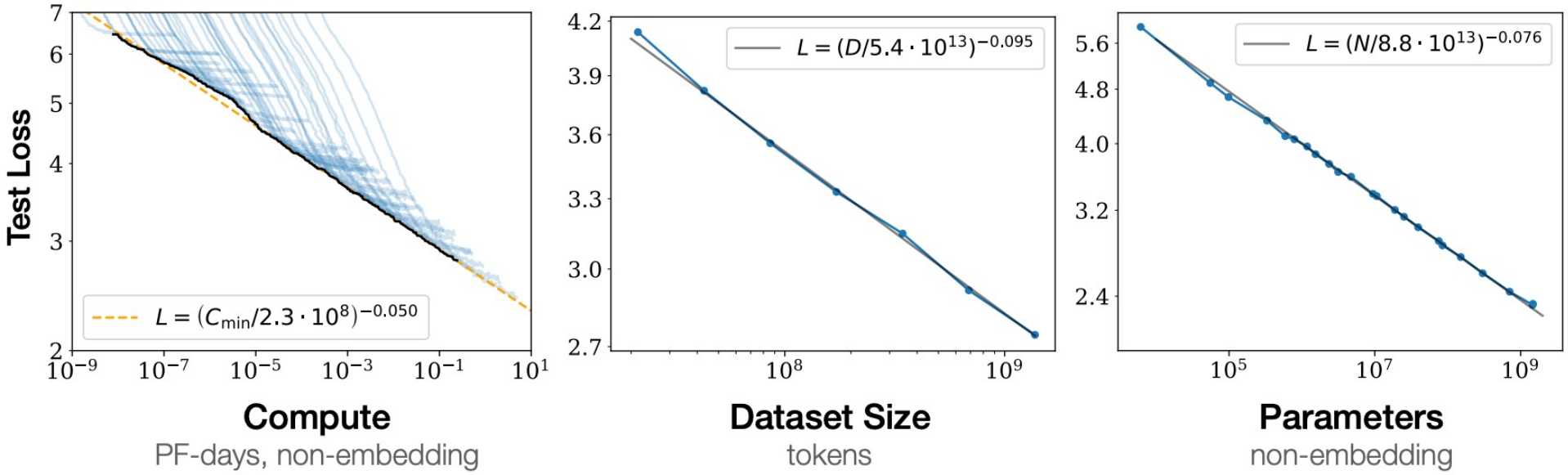
Credit: Dr. Bo Wang at U Toronto

https://blogs.nvidia.com/blog/what-are-foundation-models/

Bommasani, Rishi. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258* (2021).
van Veldhuizen, Vivien, et al. "Foundation Models in Medical Imaging--A Review and Outlook." *arXiv preprint arXiv:2506.09095* (2025).

What makes foundation model successful?
--Scaling law

- Scaling is the successful paradigm in NLP (e.g., GPT) and world models
  - Performance follows **a predictable power-law improvement model that scales with** model parameters, training data, and compute budget, guiding resource allocation.
  - Scaling is one of the most reliable ways to improve general capability without manual feature engineering
  - Different types of scaling support billions to trillions of dollars of stock valuation



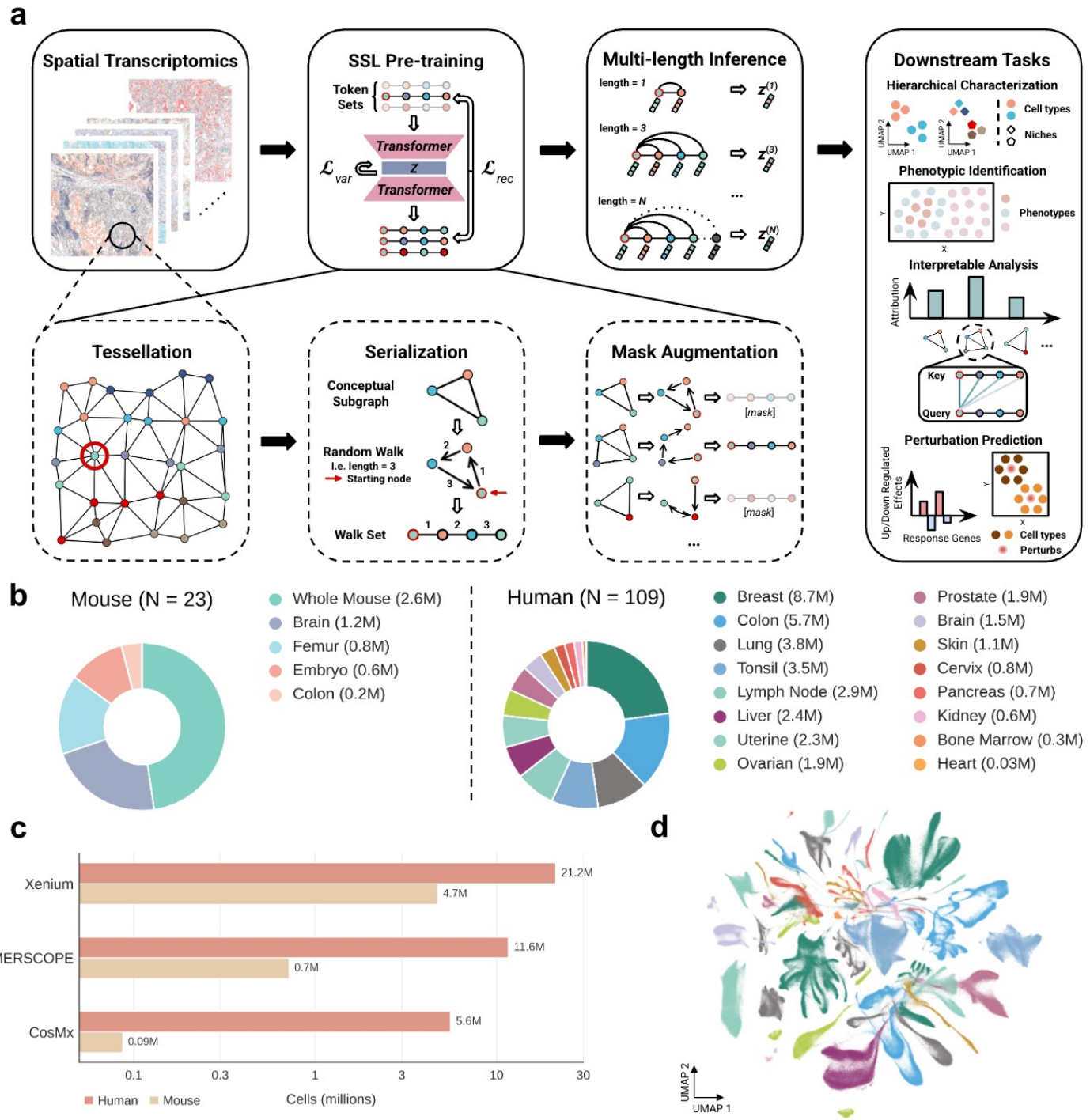
Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

spaGFM: **spa**tial **G**raph **F**oundation **M**odel

- Serializes cellular neighborhoods through **random walks** to learn transformer-compatible representations of tissue organization.
- **Augmented** by multiple walks
- By replacing unscalabe classical message passing with self-supervised neighborhood reconstruction by **M**asked **A**uto**E**ncoder, spaGFM captures higher-order spatial dependencies while enabling model scaling for a large-scale spatial transcriptomics atlas.
- Pretrained on 43.9 million cells from 132 image-based spatial transcriptomics datasets

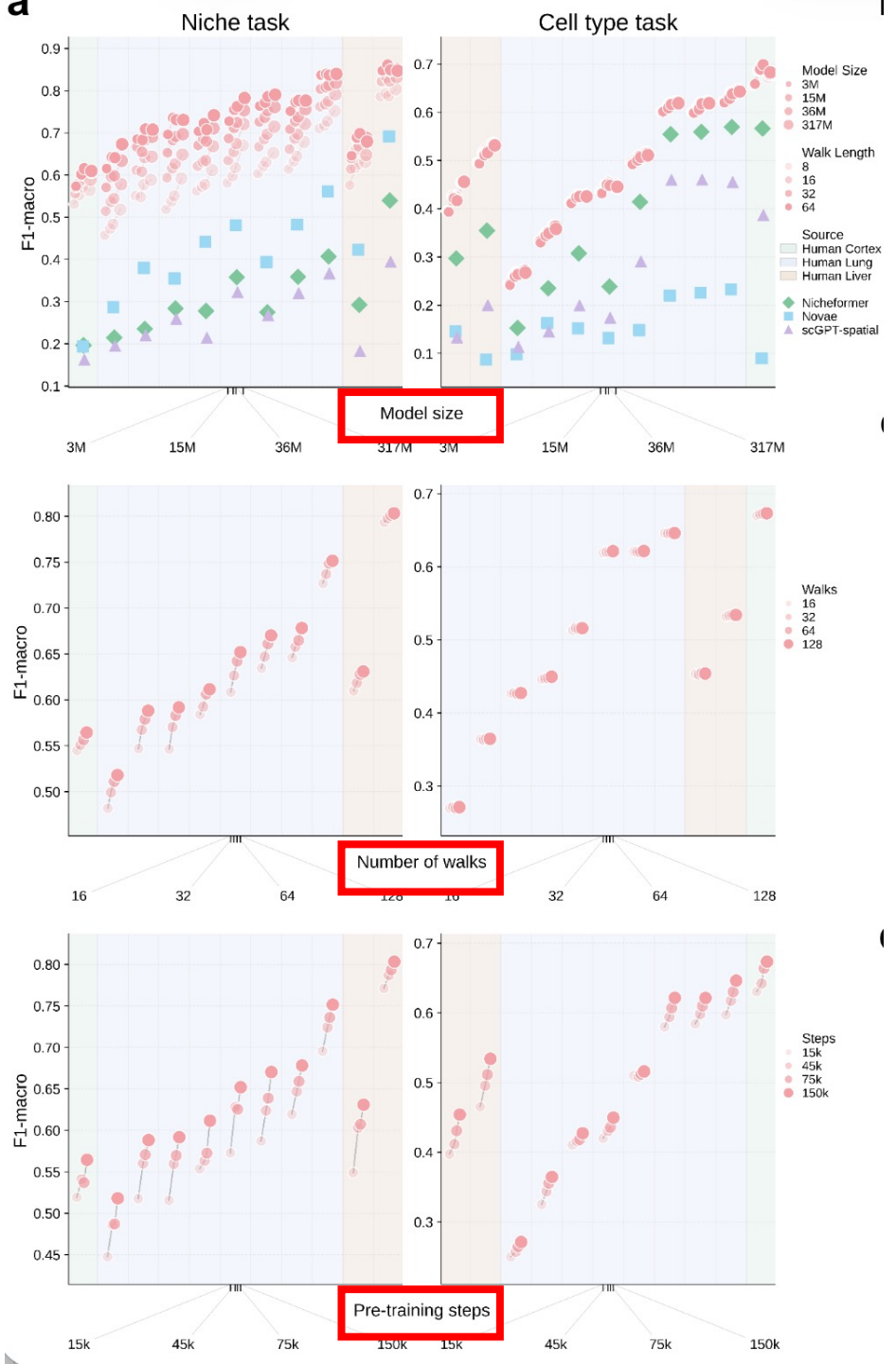
Yu Zhong
Indiana University

Zhong, et al, *In preparation*



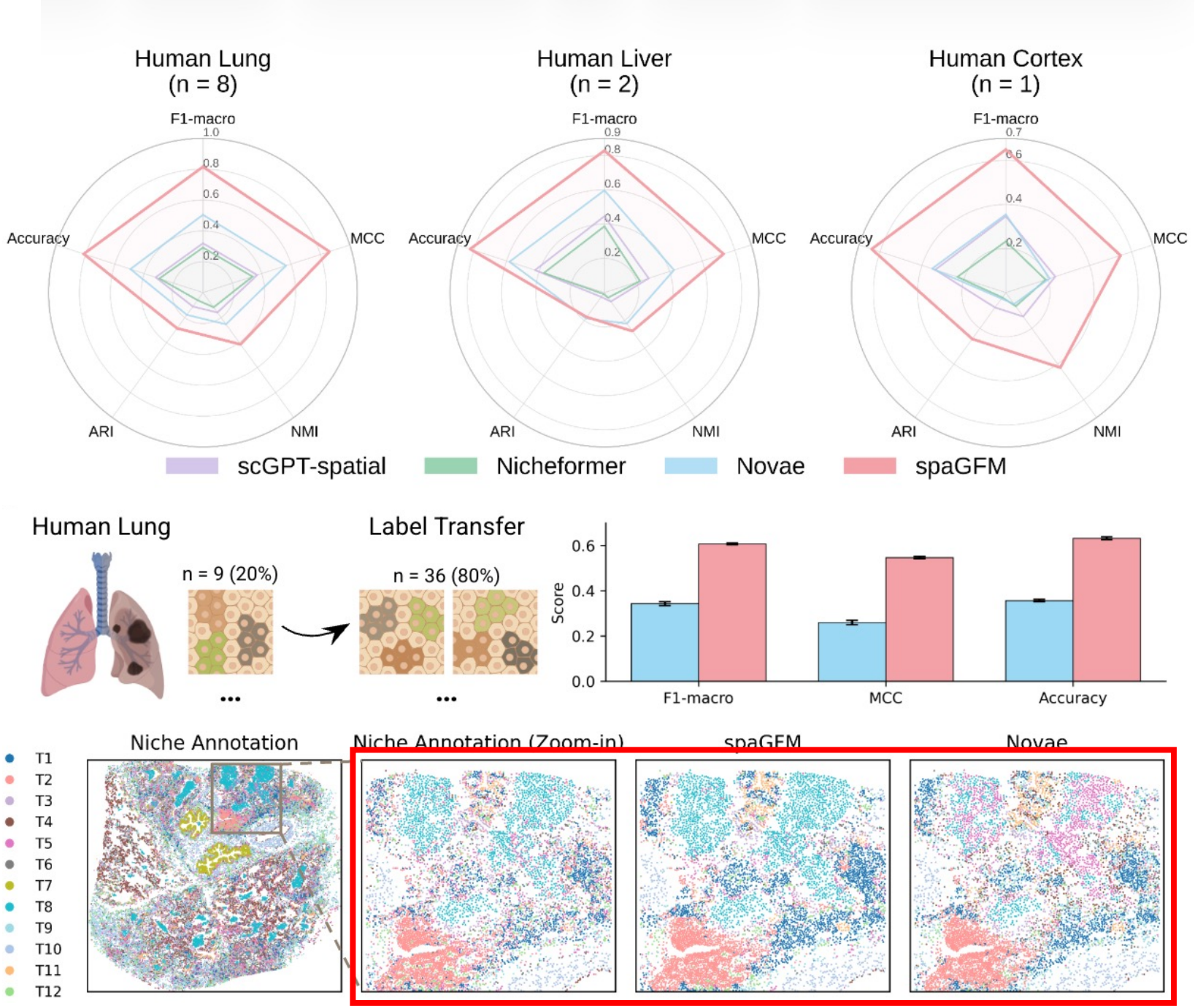
spaGFM is scalable

- **Scalable** with model size, number of walks, and pre-training steps
- Linear probing on held-out datasets including the human cortex, lung, and liver on CosMX with official annotations
- Scale up to 0.3B within available data
- Consistent with the biology that niche is more spatial-aware than cell types



spaGFM outperforms in niche tasks

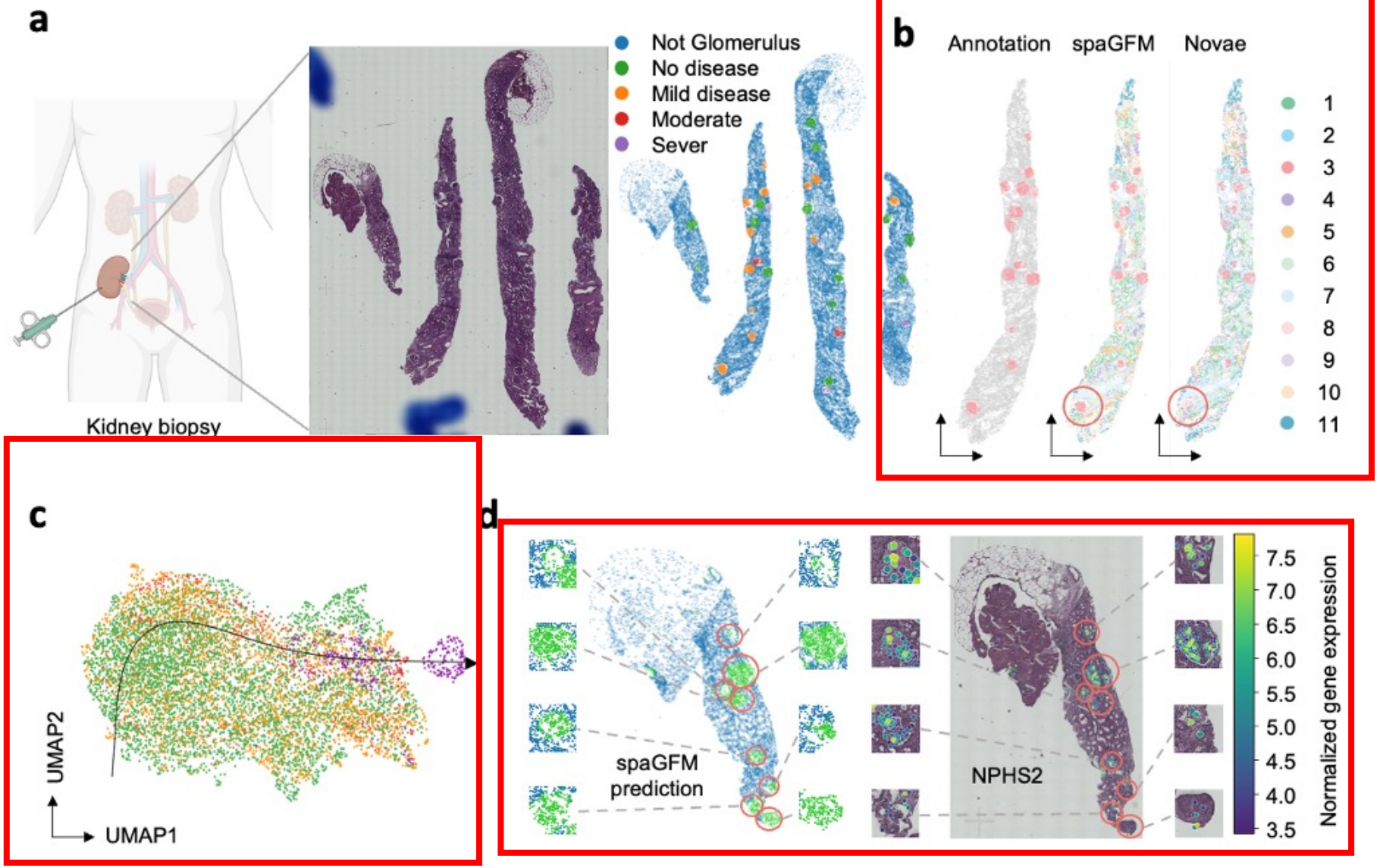
- Zero-shot Label Transfer with K-nearest neighbor
- Outperforms on Unseen human lung niches



Vannan, A., Lyu, R., Williams, A.L. *et al.* Spatial transcriptomics identifies molecular niche dysregulation associated with distal lung remodeling in pulmonary fibrosis. *Nat Genet* **57**, 647–658 (2025). https://huggingface.co/datasets/theislab/CAMEO-Lung

spaGFM characterizes glomerular organization associated with pathological grade in diabetic kidney disease

- Clustering on zero-shot embeddings identified glomeruli
- Trajectories of cell embeddings in glomeruli consistent with severity grade
- NPHS2 confirmed glomeruli identification compensate to H&E annotations



Why triangles work as an inductive bias? (Geometrical perspective)

- Different topological patterns exhibit distinct geometric properties
  - Regular grids are well-suited to **Euclidean geometry** for uniform local connectivity.
  - Cyclic patterns require **spherical geometry** to maintain rotational symmetry and closure.
  - Tree structures rely on **hyperbolic geometry** to preserve exponential volume growth without distortion.
- Part of the reason why the graph foundation model is difficult

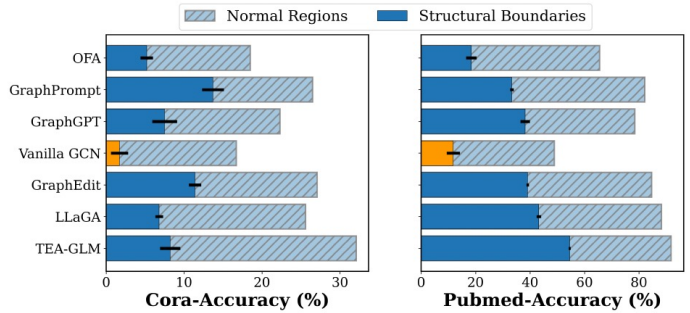
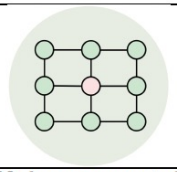
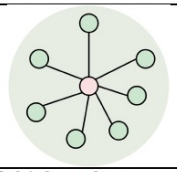
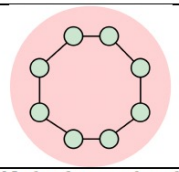
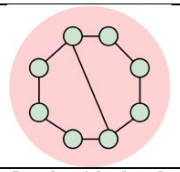
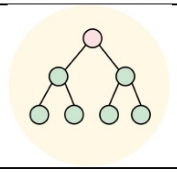
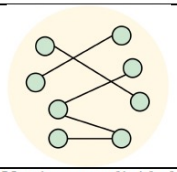
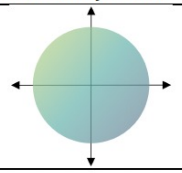
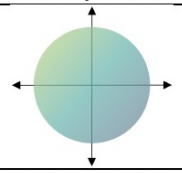
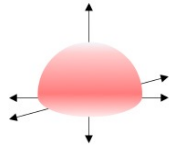
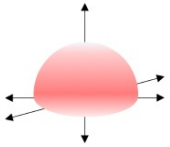
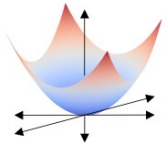
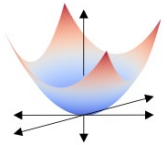


Figure 1: Performance degradation at structural boundaries. We compare zero-shot accuracy on normal regions versus structural boundaries across Cora and Pubmed. All methods exhibit systematic performance drops at boundaries.

Structure Type	Grid	Star	Circle	Chordal Cycle	Tree	Bipartite Graph
Illustration						
Description	Nodes are connected in a grid pattern, and the number of nodes is adjustable.	Multiple nodes connect to a central node, and the number of nodes is adjustable.	Nodes form a closed loop, and the number of nodes is adjustable.	Graph with chords in the cycle, with adjustable nodes and chords.	Nodes form a tree, branching and height are adjustable.	Vertices are divided into two disjoint sets, adjustable nodes.
Geometry						
Geometry	Euclidean	Euclidean	Spherical	Spherical	Hyperbolic	Hyperbolic

Zhang, Heng, et al. "GraphShaper: Geometry-aware Alignment for Improving Transfer Learning in Text-Attributed Graphs." *arXiv preprint arXiv:2510.12085* (2025).

Multi-sample and multi-modal integration with Optimal Transport (OT)

- **Deep learning** approaches aligned learnt embeddings
  - e.g. spatialGlue
- Alternatively, **OT (family)**

  e.g. MOSCOT
  - 1. Mathematical rigor
  - 2. Representation-free, no training, better generalization, global structure-preserving alignment allowing distortion
  - 3. Interpretability within the transport plan

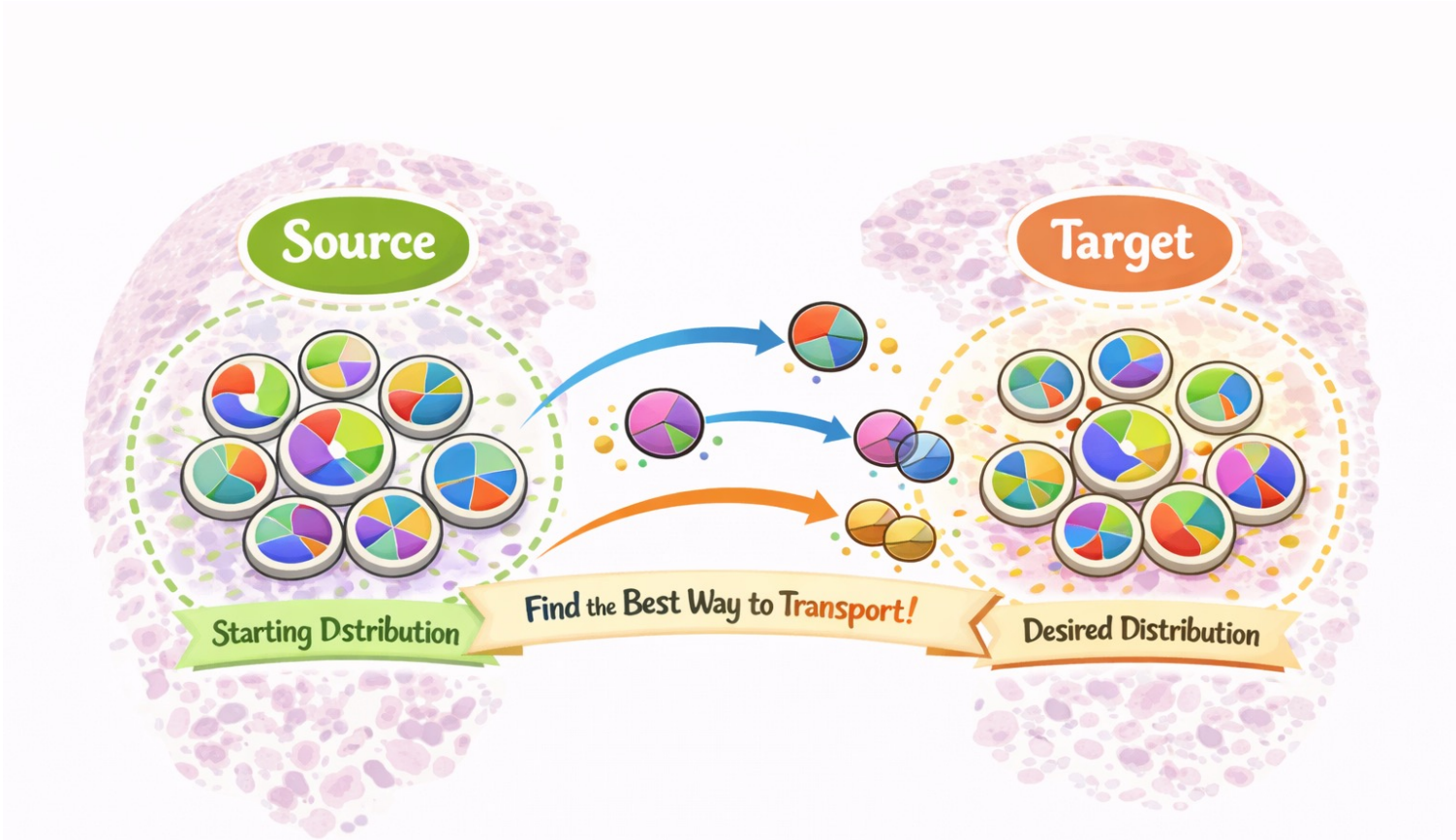
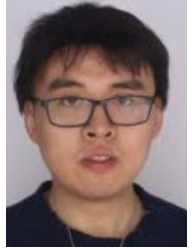


Image from ChatGPT

Long, Yahui, et al. "Deciphering spatial domains from spatial multi-omics with SpatialGlue." Nature Methods 21.9 (2024): 1658-1667.
Klein, D., Palla, G., Lange, M. *et al.* Mapping cells through time and space with moscot.*Nature* **638**, 1065–1075 (2025).

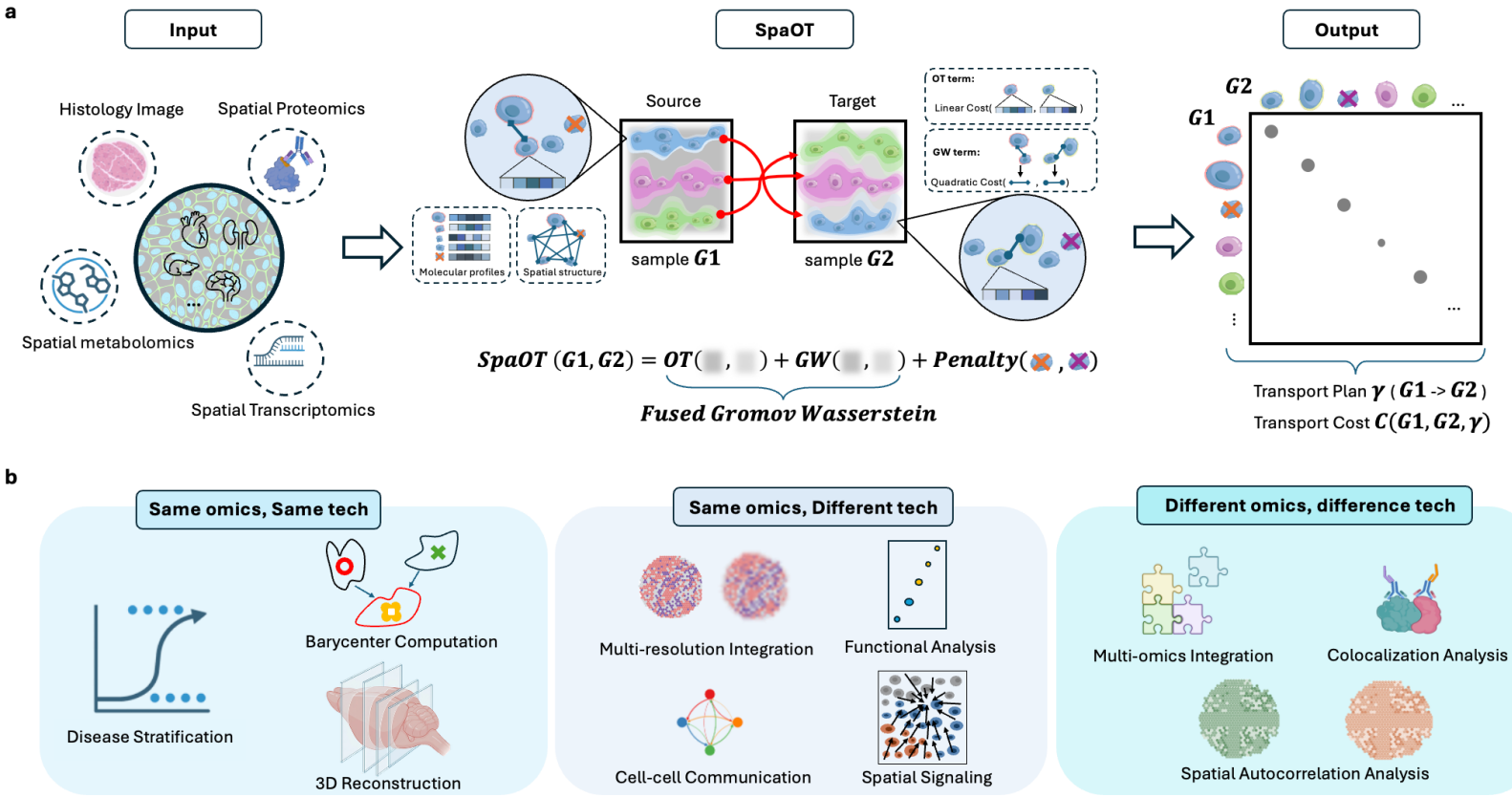
Spatial **O**ptimal **T**ransport (SpaOT)

Shuang Wang
Indiana University



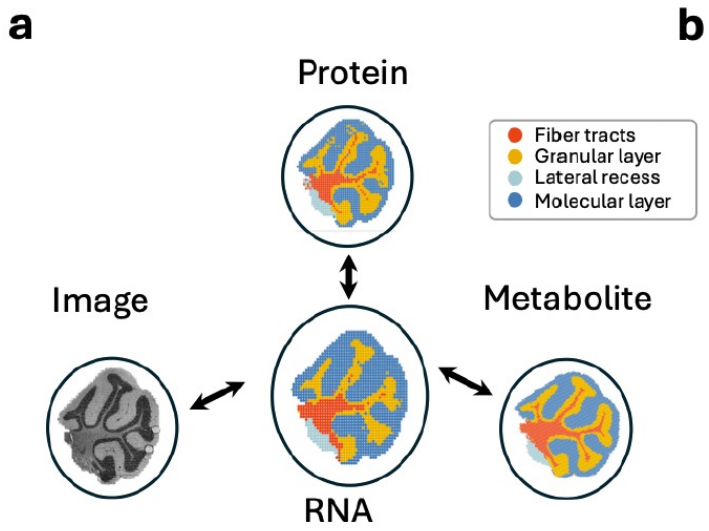
Dr. Yikun Bai
Vanderbilt

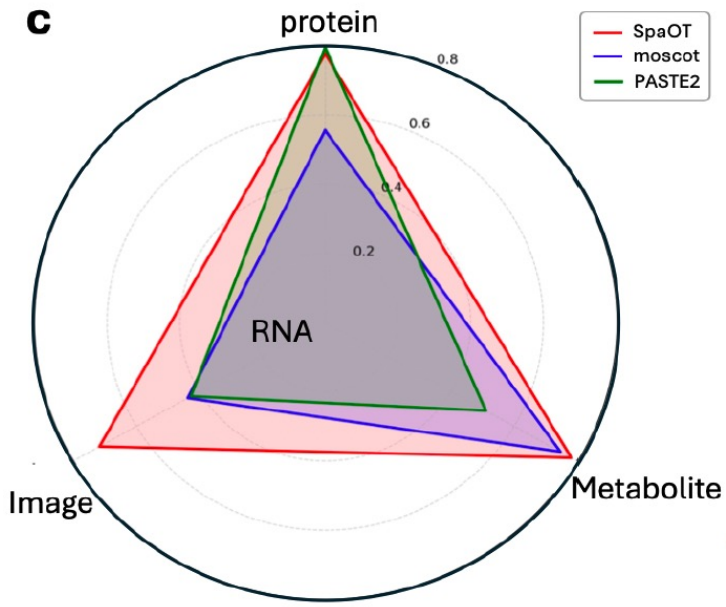
- A spatial multi-omics integration framework based on a theoretically grounded semi-metric formulation of unbalanced optimal transport.
- Utilizing a total-variation-relaxed Fused Partial Gromov-Wasserstein strategy, SpaOT establishes a mathematically consistent distance geometry.
- A general foundation for spatial multi-omics integration and comparative analysis.

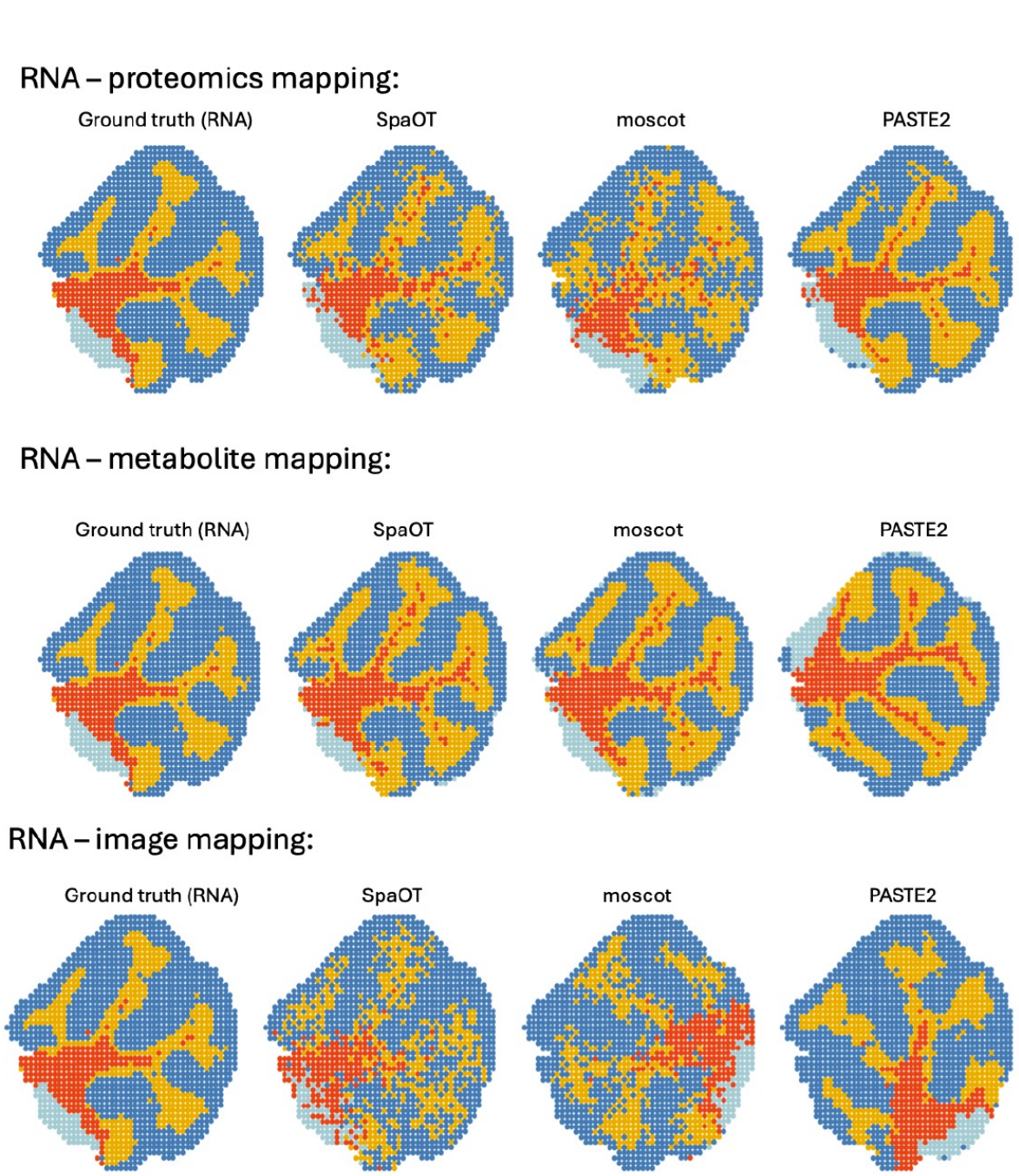


Wang, Bai, ..., **Eadon**, Kolouri, **Wang***, Semi-metric optimal transport enables robust spatial multi-omics integration, *Submitted*

spaOT acutely maps spatial multi-omics samples

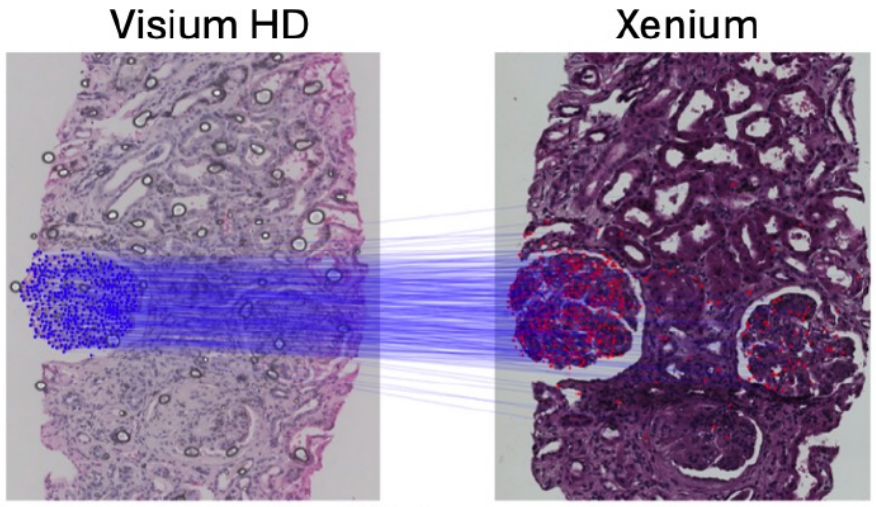


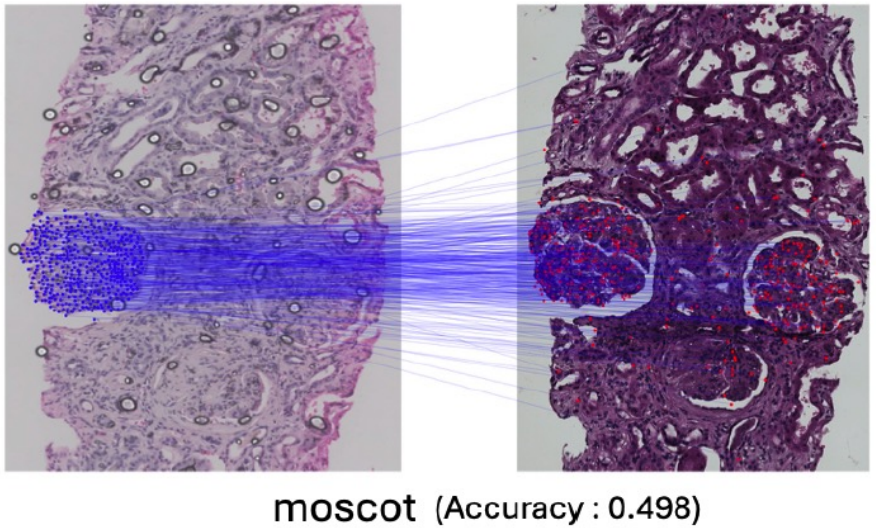


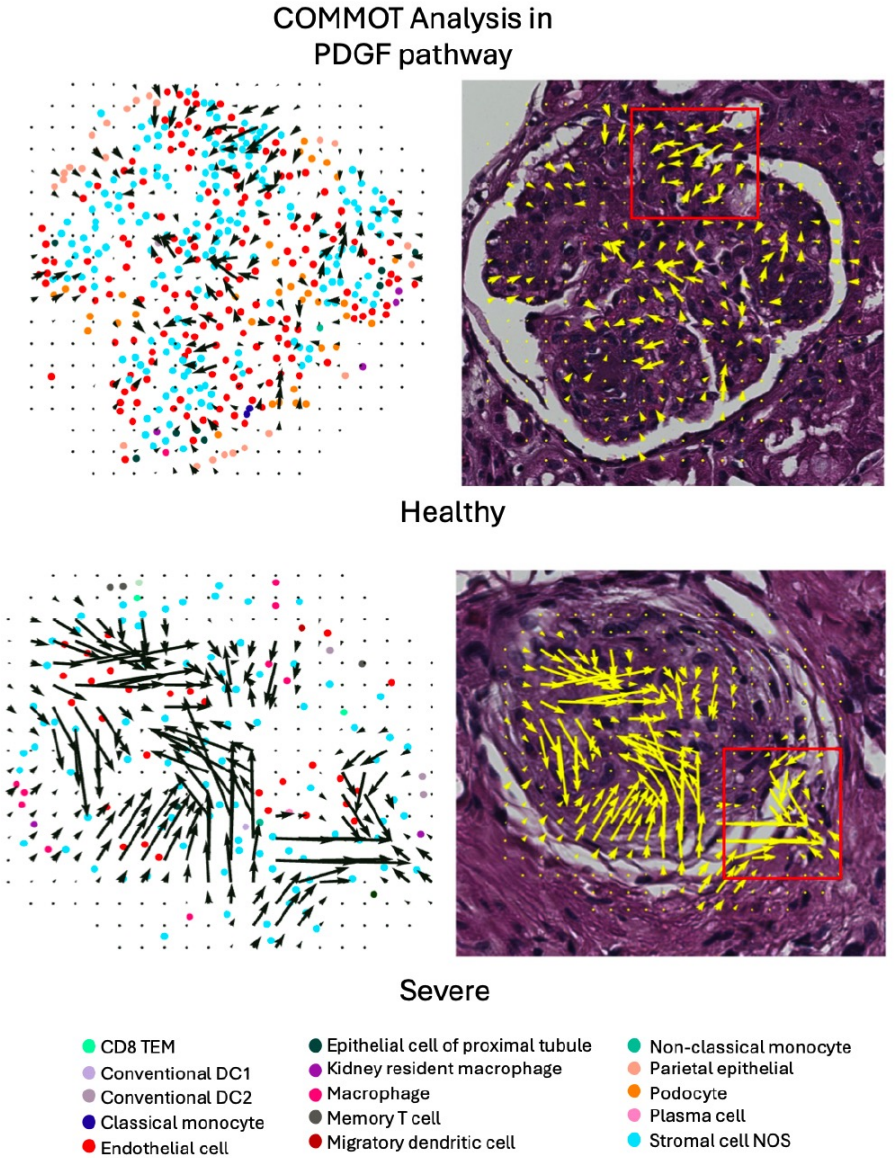


Hu, Beiyu, et al. "High-resolution spatially resolved proteomics of complex tissues based on microfluidics and transfer learning." *Cell* 188.3 (2025): 734-748.

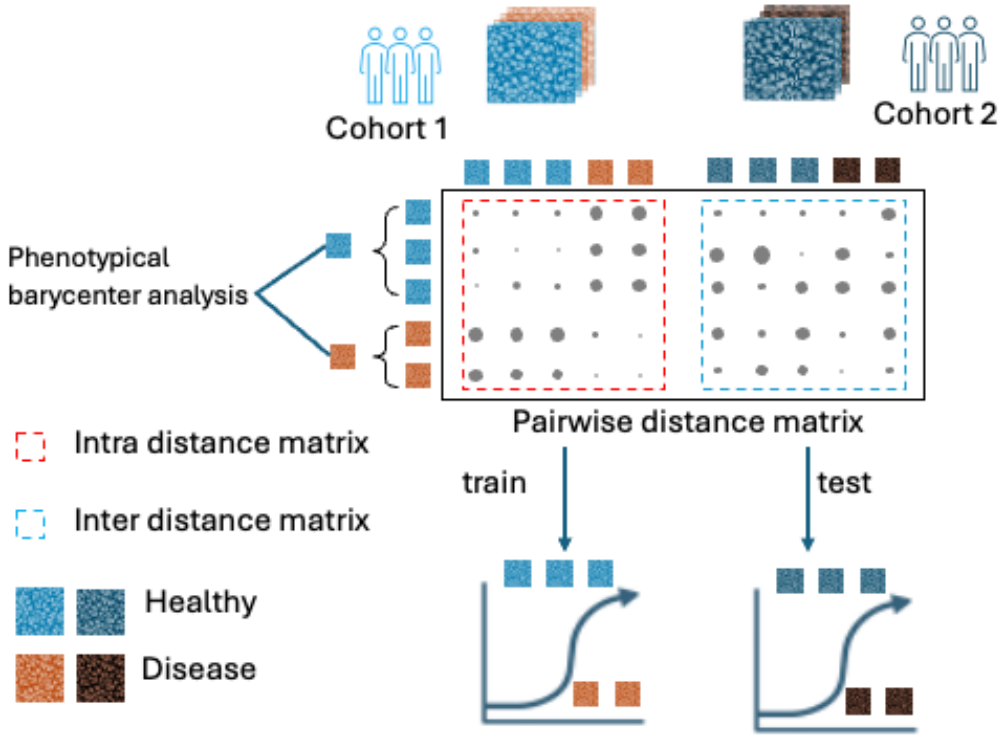
SpaOT-integration enables comprehensive analysis by accurately mapping the Glomerulus between adjacent Xenium and Visium HD





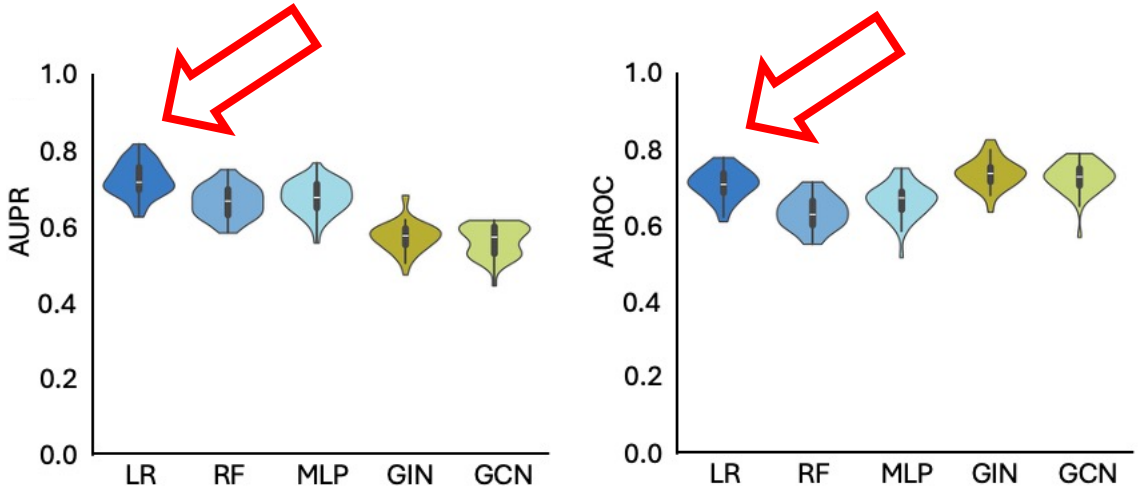


Distance derived from spaOT + Logistic Regression outperforms/competes with Graph Neural Networks stratifying breast cancer samples(**Intra**)

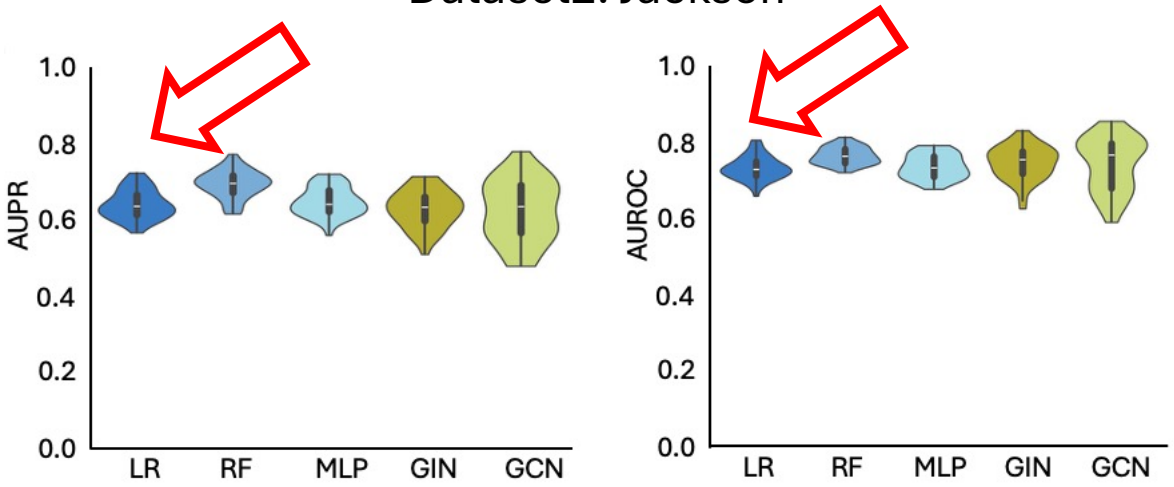


- Metabric: 500 samples (231 grade 1/2 and 269 grade 3)
- Jackson: 559 samples (328 grade 1/2 and 231 grade 3) (source)
- Classification of grade 1/2 vs grade 3
- Use **cross-validation** to get metrics

Dataset1: Metabric



Dataset2: Jackson

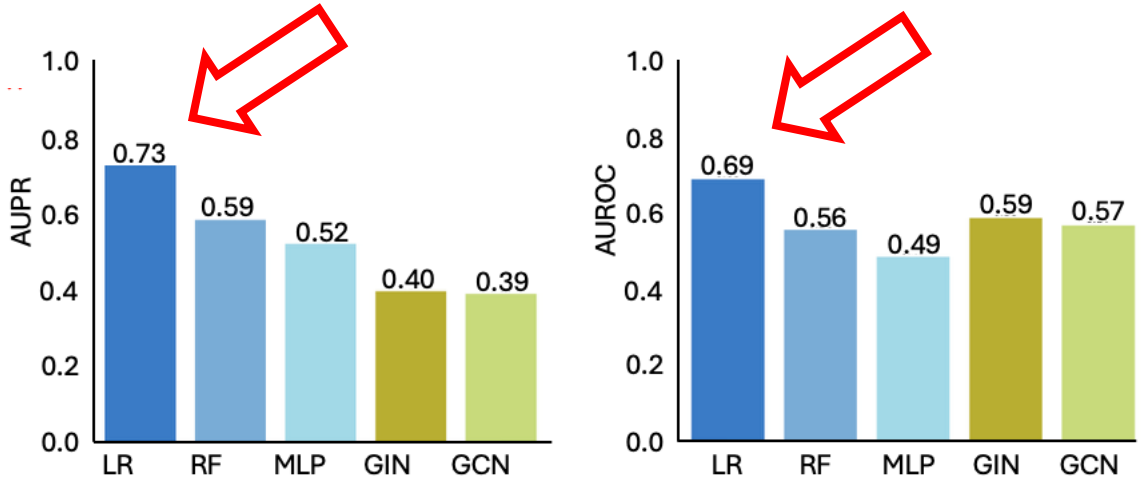


Ali, H. R. et al. Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer. Nat. Cancer 1, 163–175 (2020).
Jackson, H. W. et al. The single-cell pathology landscape of breast cancer. Nature 578, 615–620 (2020).

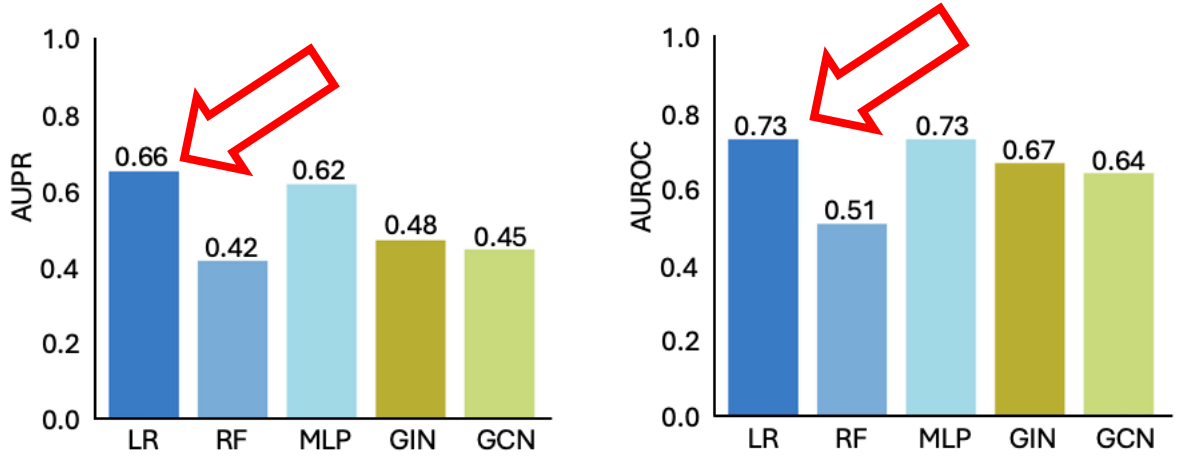
Distance derived from spaOT + Logistics Regression outperforms/competes with Graph Neural Networks stratifying breast cancer (**inter**)

- Highly challenging translational medicine task
- Huge batch effects
- Compared to GNN:
  - No pooling
  - No representation learning

Train Metabric to predict Jackson

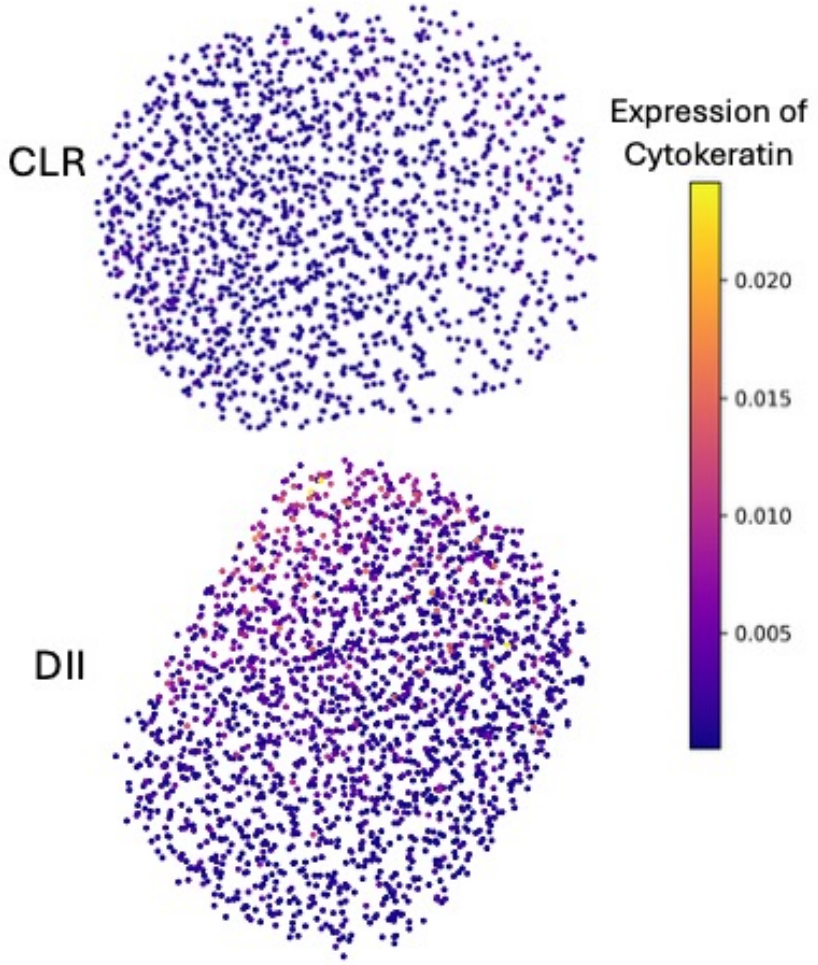


Train Jackson to predict Metabric



Barycenter analysis: mathematically meaningful representative tissues

- Averaged samples



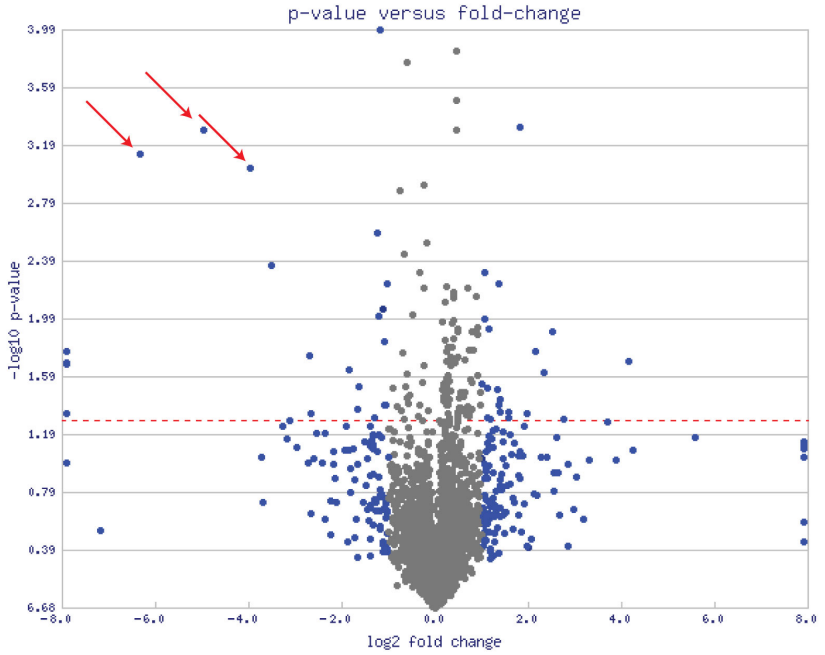
https://learnopencv.com/average-faces-of-fifa-world-cup-2018/

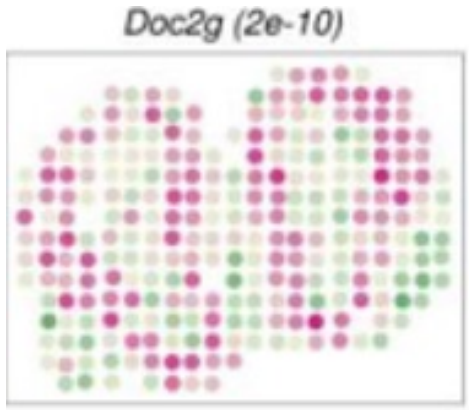
Outline

- Q1: How to resolve complex biological heterogeneity in spatial transcriptomics data?
- **Q2: How to identify molecular markers that define biologically meaningful spatial organization?**
- Q3: How does spatial organization link to tissue functions?
- Discussion and perspective

Transcriptomics: Spatially Variable Genes (SVGs) (spatial context of DEG)

- Differential Expressed Gene -> Spatial Variable Gene
- **Spatially Variable Gene**: Gene differential expressed because of spatial (condition as DEG)
- SVGs are defined as genes with **a highly spatially correlated pattern of expression**, which varies along with the spatial distribution of a tissue structure of interest.
- Spatial expression variation can reflect communication between adjacent cells, position-specific states, or cells that migrate to specific tissue locations to perform their functions.
- Generalize to **Spatially Variable Features (SVFs)** in spatial omics

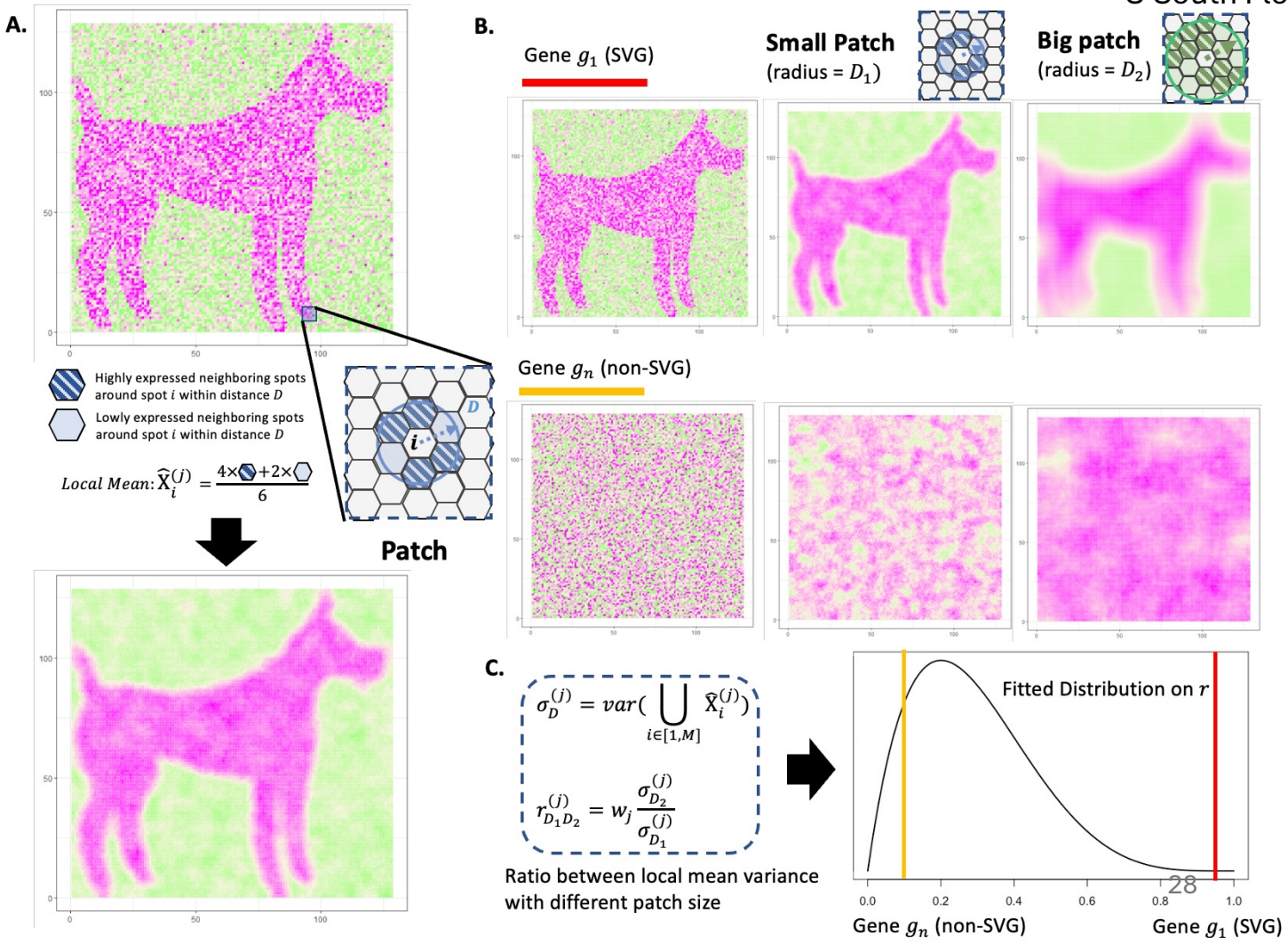




BSP: Dimension-agnostic and granularity-based spatially variable gene identification

Dr. Dong Xu
U South Florida

- BSP (big-small patch), a non-parametric model by comparing gene expression patterns at two spatial granularities to identify SVGs from two or three-dimensional spatial transcriptomics data in a fast and robust manner.
- **Velocity of change in variances of local means in different granularity can be used to distinguish SVGs and non-SVGs**
- Data driven, model free, any dimension
- Highlighted by Editor's choices



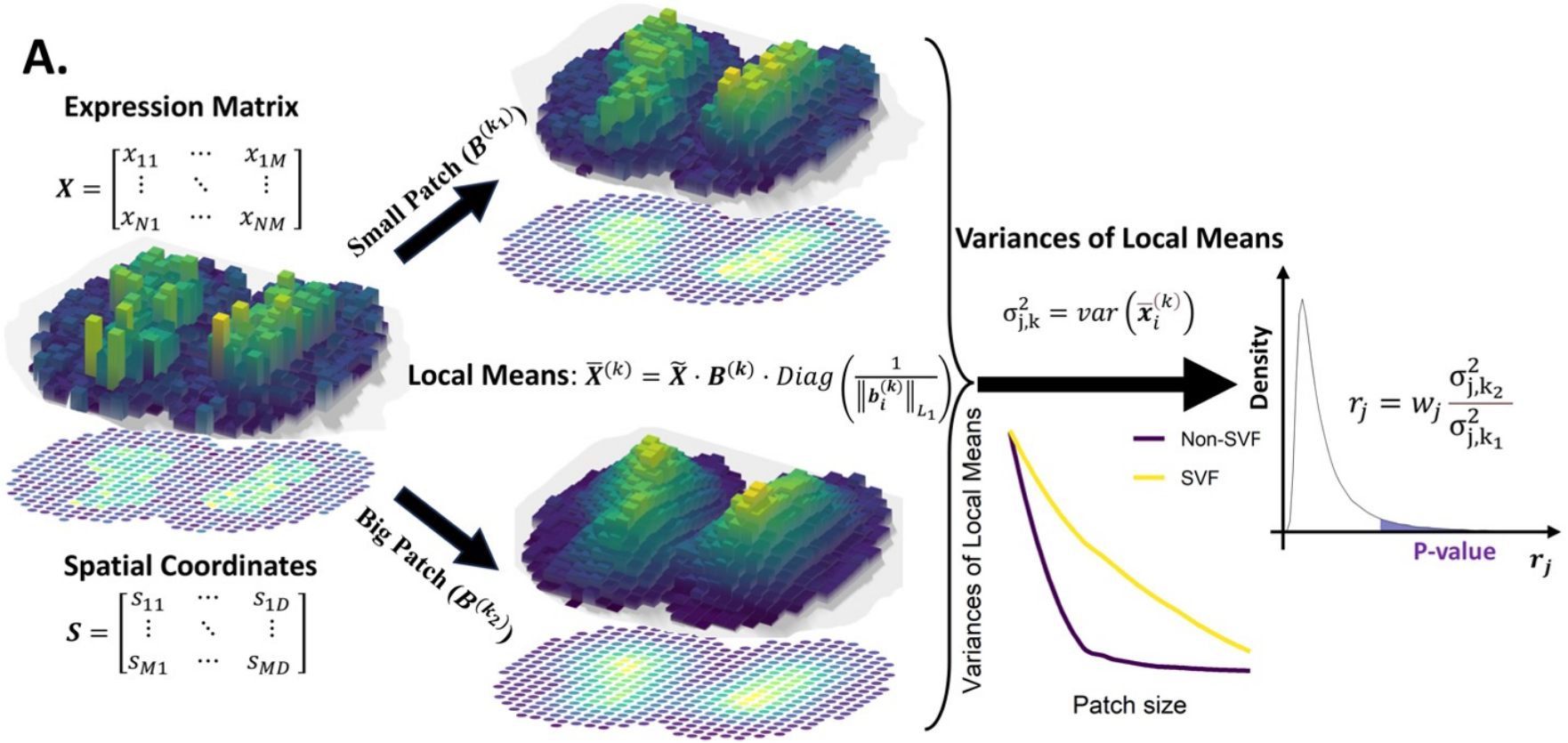
Wang, J.*, Li, J., Kramer, S.T., Su, L., Chang, Y., Xu, C., **Eadon, M.T.**, Kiryluk, K., Ma, Q. and Xu, D., 2023. Dimension-agnostic and granularity-based spatially variable gene identification using BSP. *Nature communications*, *14*(1), p.7367.

scBSP (BSP 2.0): Highly efficient and scalable tool for large-scale high-resolution spatial omics studies

Dr. Jinpu Li
U of Missouri

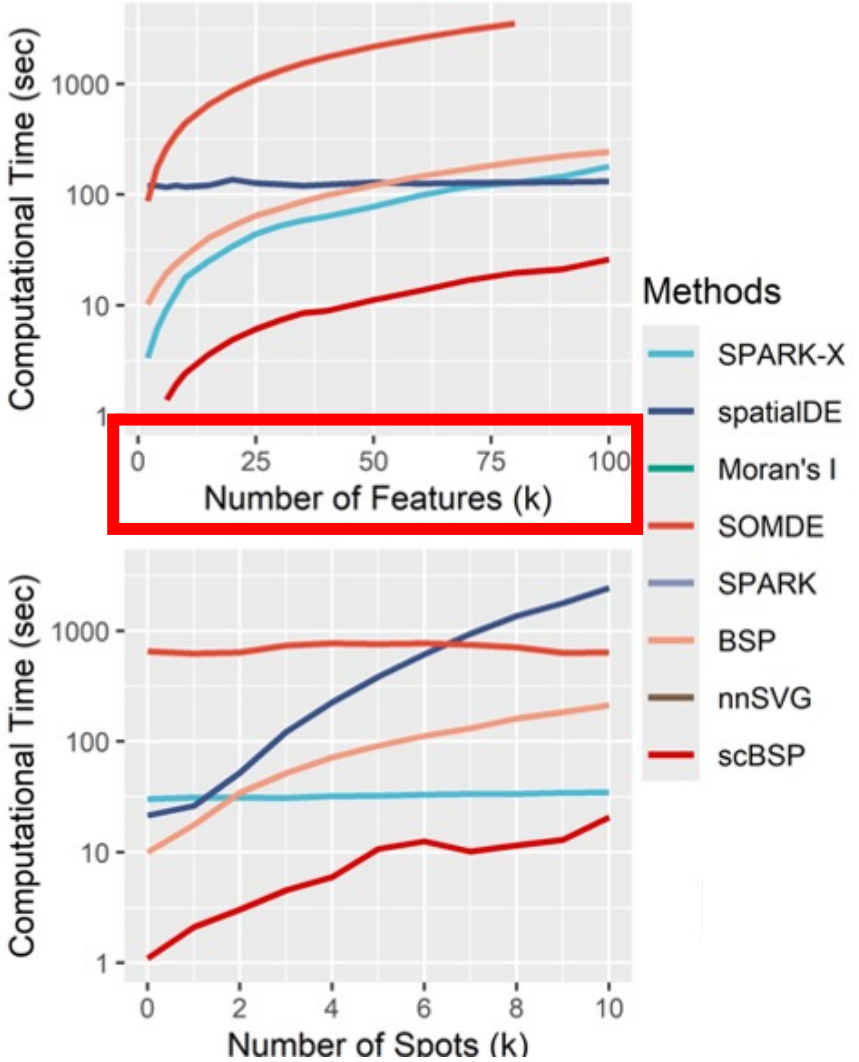
Mauminah Raina
Indiana University

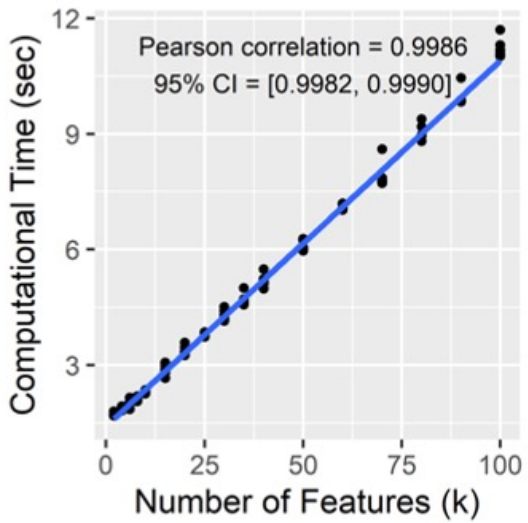
- Using **sparse matrix operation** to accelerate the computing

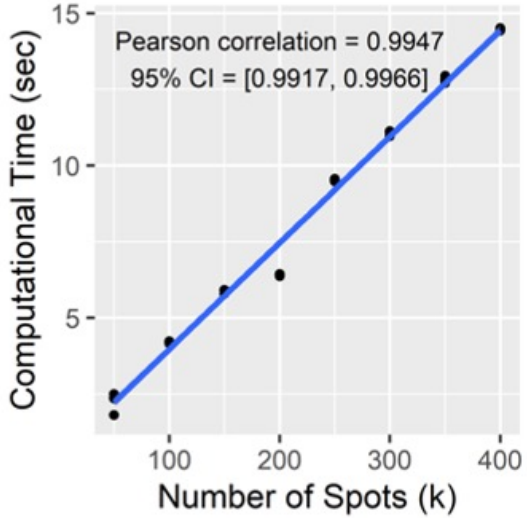


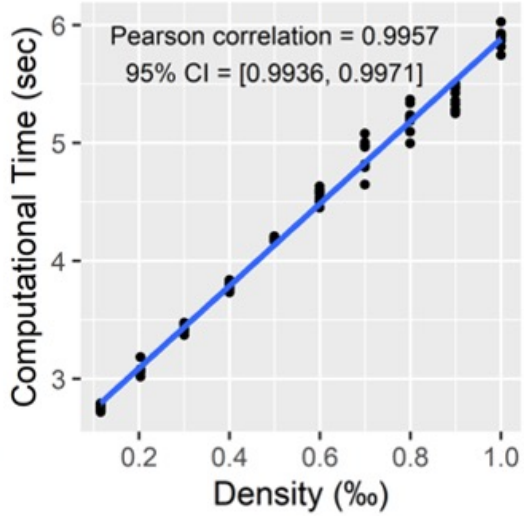
Li, J., Raina, M., Wang, Y., Xu, C., Su, L., Guo, Q., Ferreira, R.M., **Eadon, M.T.**, Ma, Q., **Wang, J*.** and Xu, D., 2025. scBSP: A fast and accurate tool for identifying spatially variable features from high-resolution spatial omics data. *Bioinformatics*, 2025, btaf554

scBSP demonstrates highly computational efficiency and scalability









- Linearity with increasing number of features/spots/densities

Typical usage

- Targeting subcellular Xenium/CosMx/MERSCOPE data

- HDST data:
- ~200,000 spots, 20,000 genes, 0.0003 density

- BSP: 4 hours, 90GB
- scBSP: ~5 sec, 2GB

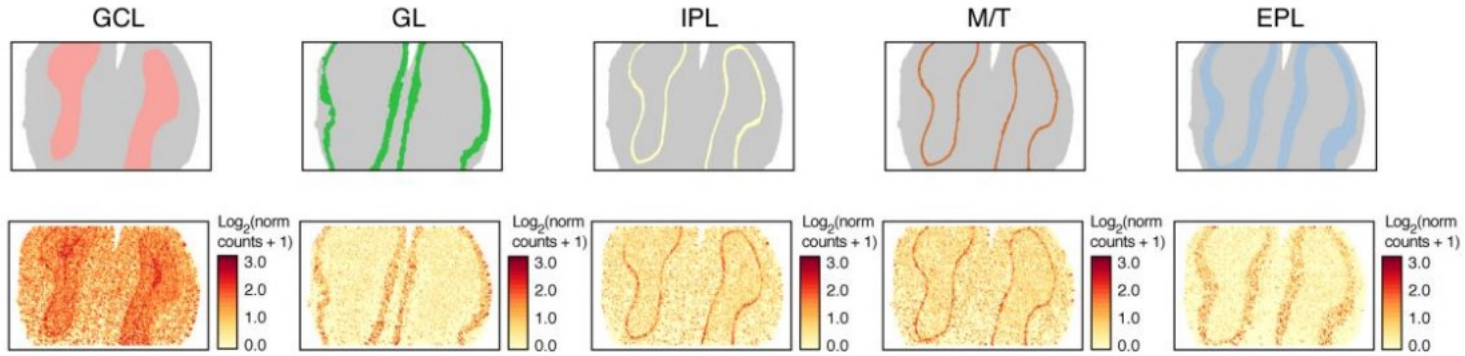
16622 scBSP 11,189

Downloads by period

66	98	504	18K
last day	last week	last month	total

Since its formal launch in May 2024, this tool has been downloaded **~29,000** times from R CRAN and PyPI, averaging more than 1,000 downloads per month Aug 24th, 2026

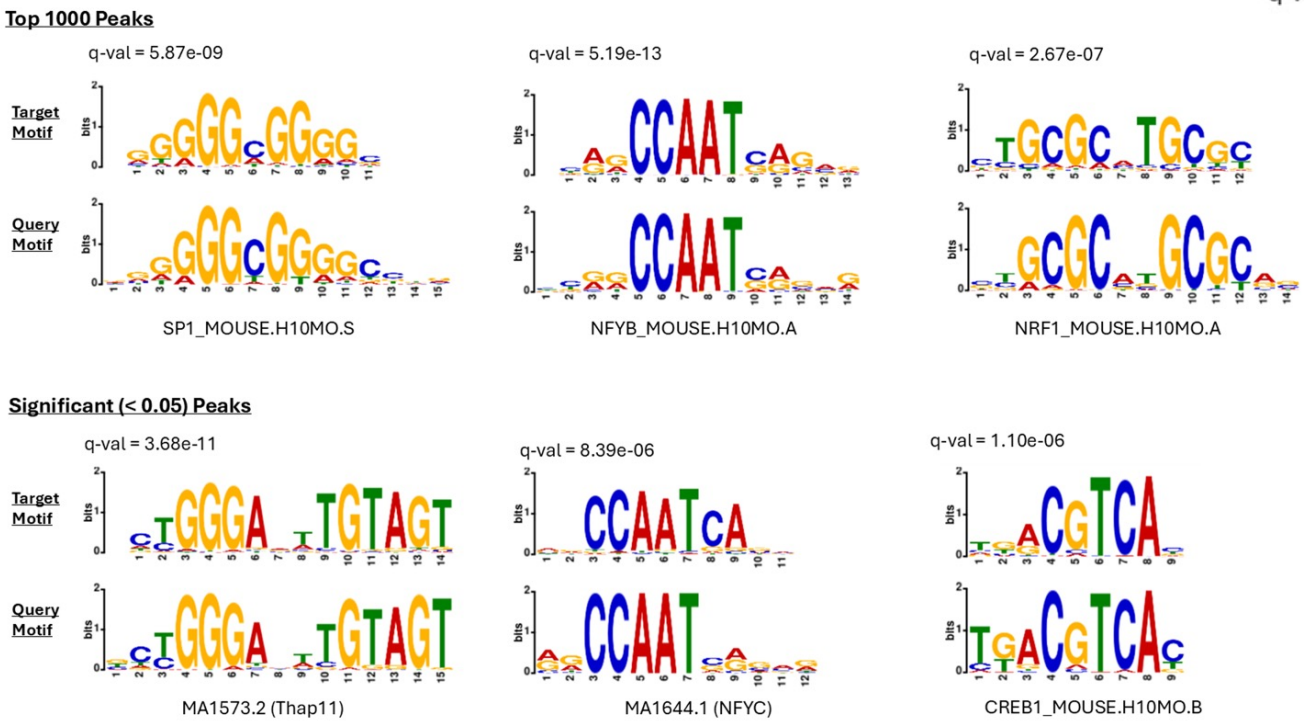
https://www.datasciencemeta.com/rpackages
https://clickpy.clickhouse.com/dashboard/scbsp

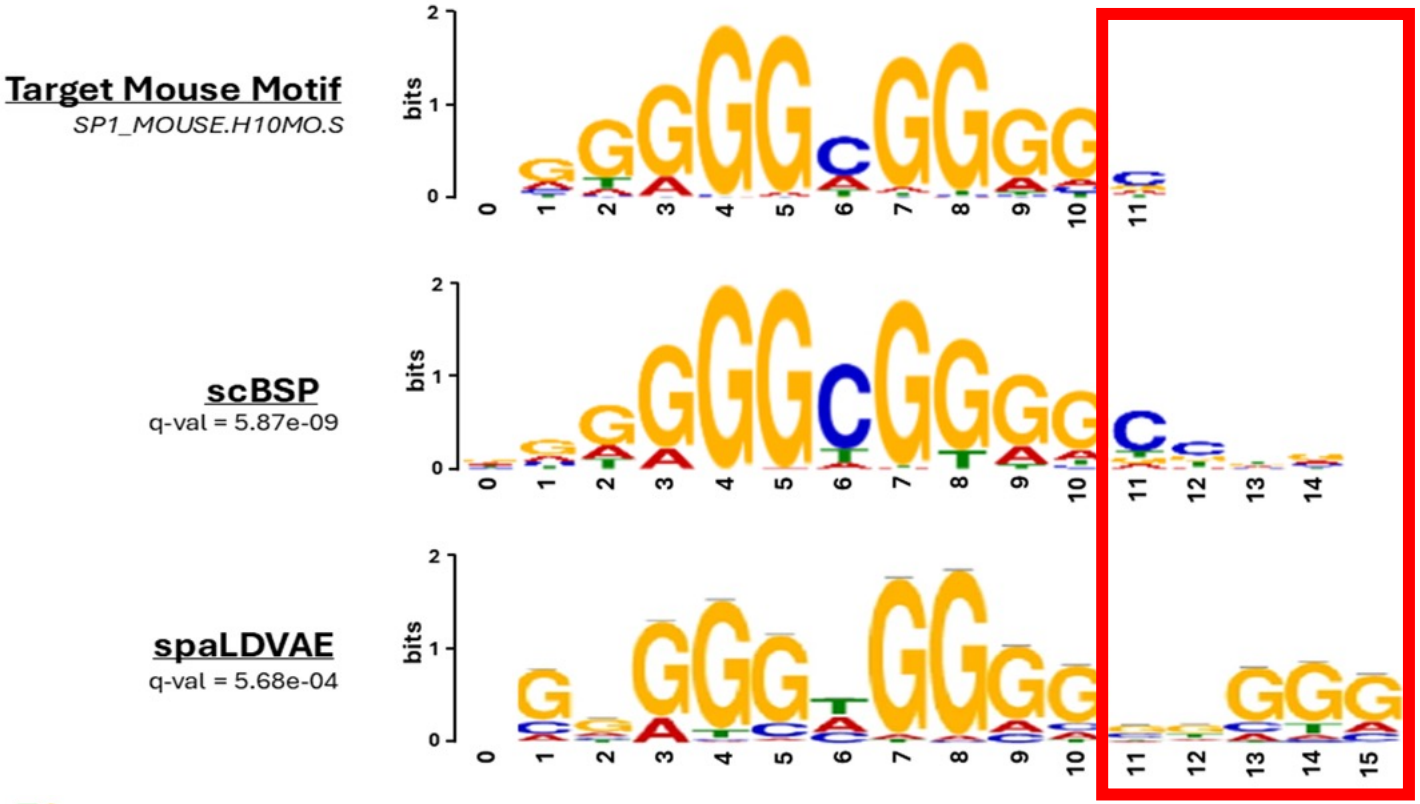


Availability:
R: https://cran.r-project.org/web/packages/scBSP/index.html
Python: https://pypi.org/project/scbsp/

scBSP on spatial ATAC-seq studies, validated by MEME-chip

- Queries motifs are consistent with known motifs



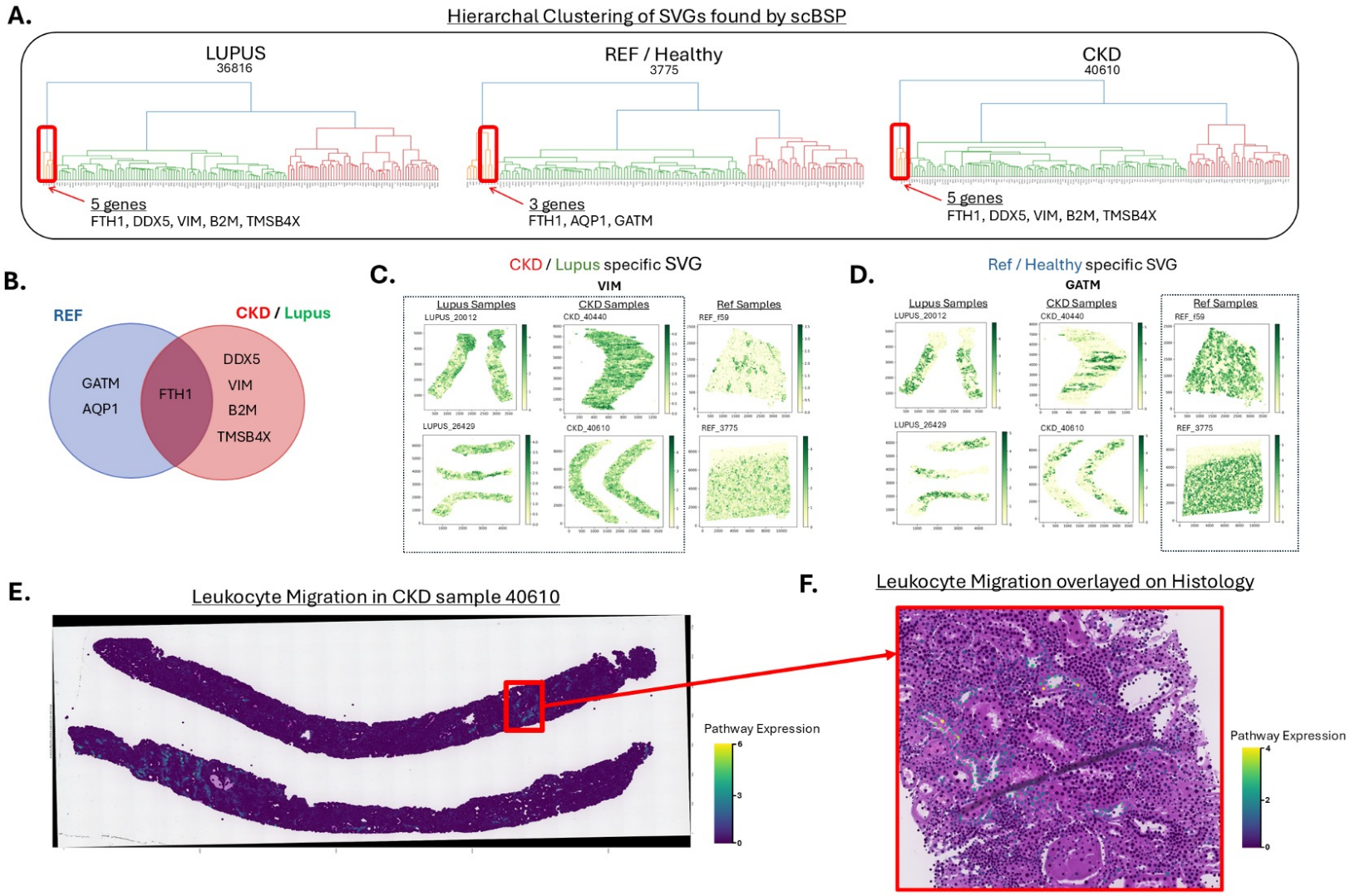


- Identified motifs are much more similar with known motifs than competitive methods

Jiang, F. et al. Simultaneous profiling of spatial gene expression and chromatin accessibility during mouse brain development. Nat. Methods 20, 1048–1057 (2023).

scBSP on 10X Xenium from human kidney

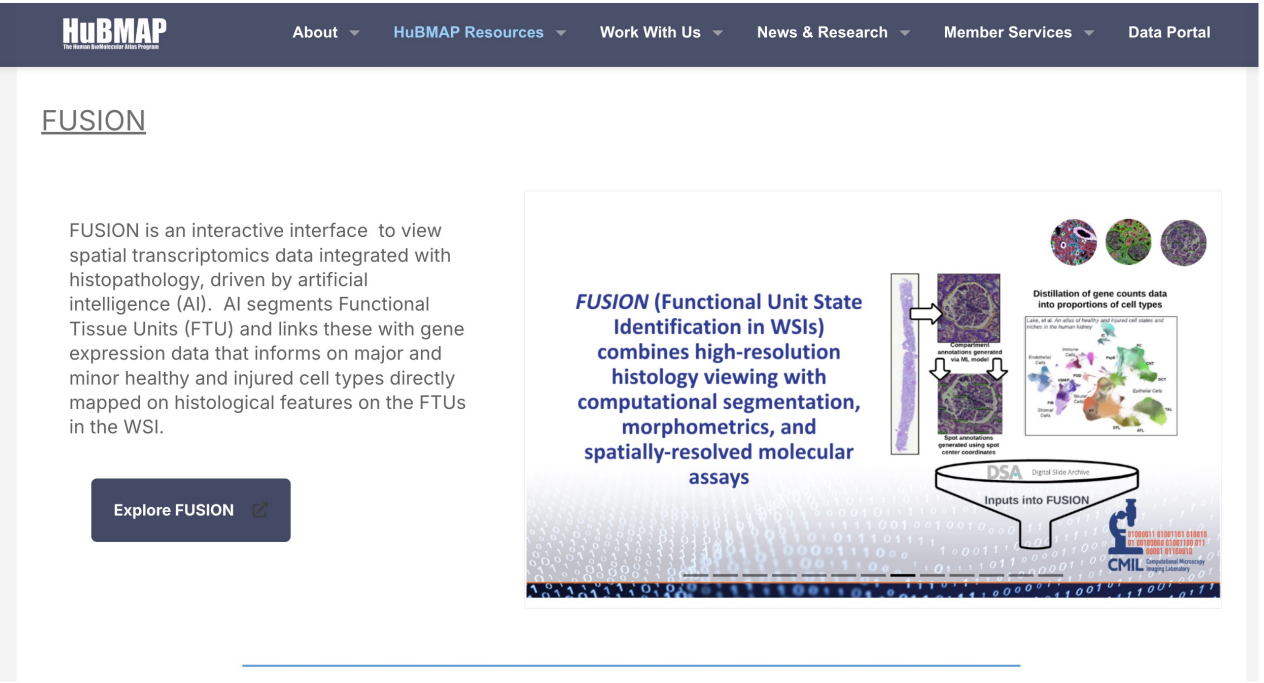
Dr. Michael Eadon
Indiana University



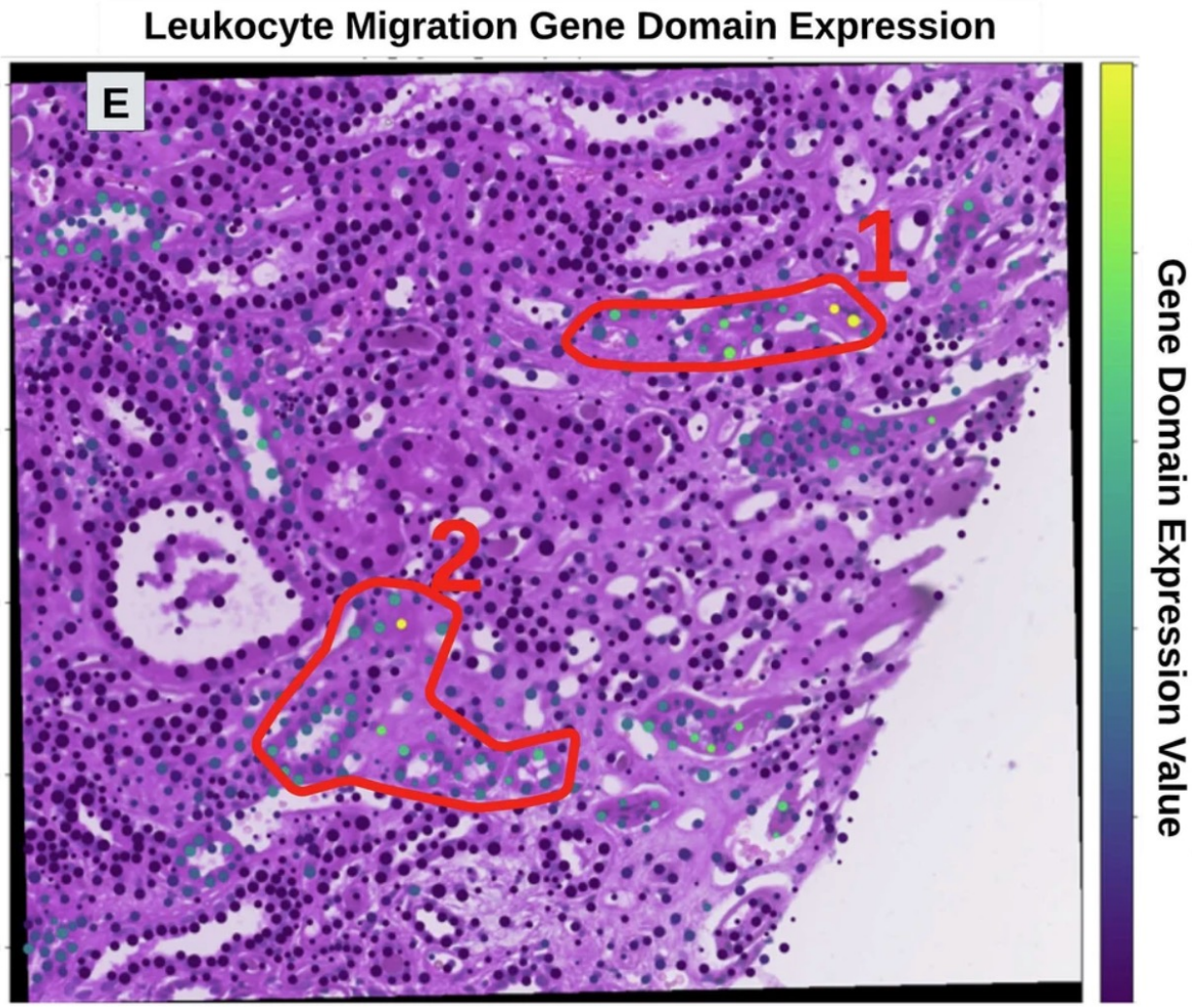
- 4 healthy controls
- 4 lupus nephritis
- 4 chronic kidney disease (CKD)

- 25,000 cells and 300 genes

Application in kidney: Fusion in HuBMAP



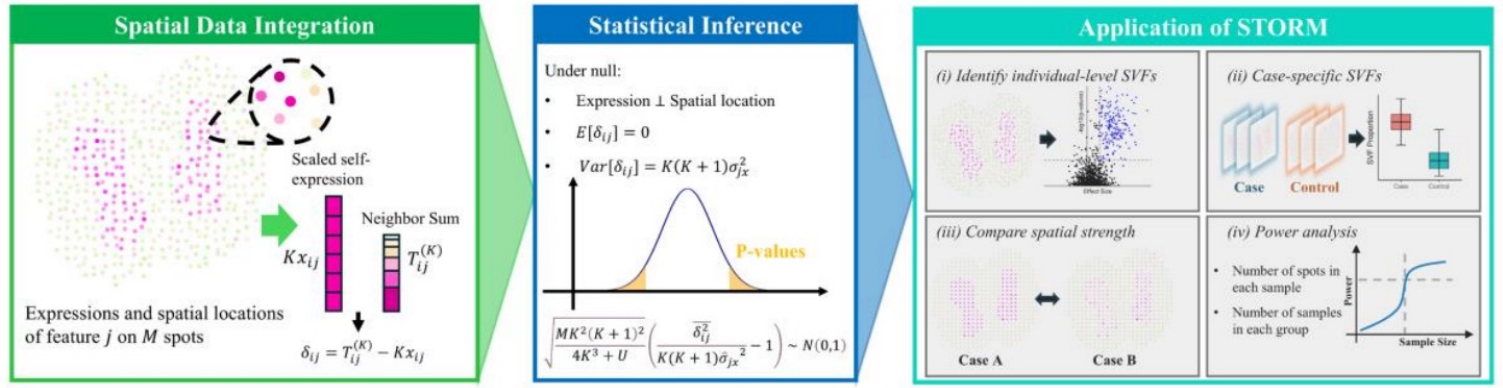
Border, S.P., Ferreira, R.M., Lucarelli, N. *et al. FUSION*: a web-based application for in-depth exploration of multi-omics data with brightfield histology. *Nat Commun* **16**, 8388 (2025)

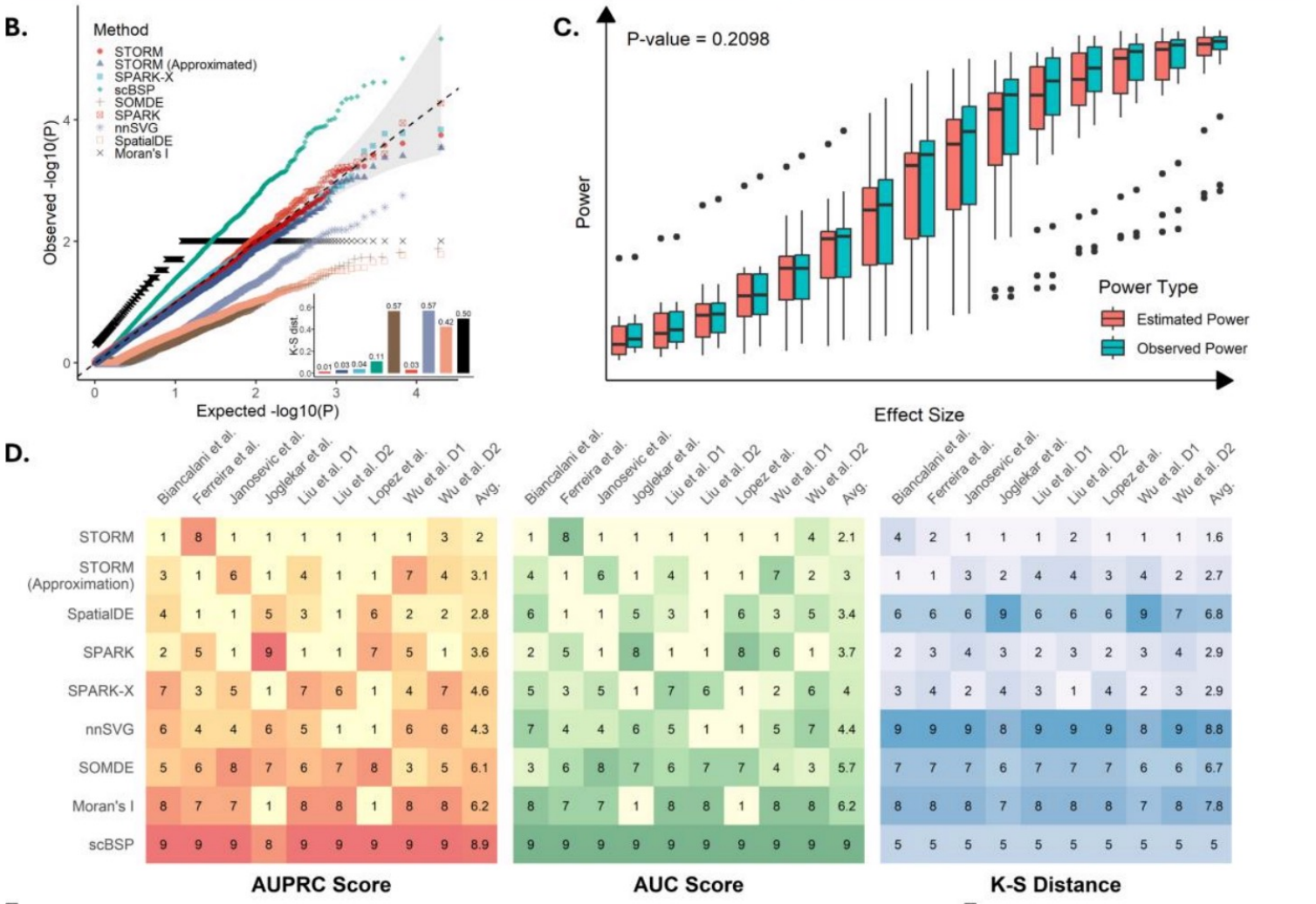


- **Spatially variable gene analysis** highlighting leukocyte migration (Gene Ontology term GO_0050900) pathway expression aligned with the histology image in a chronic kidney disease sample. Region 1 shows immune infiltration overlying interstitial matrix expansion. Region 2 denotes immune infiltration with upregulation of the pathway in epithelial nuclei

STORM, a principled **S**tatistical **TO**ol for spatially **R**esolved **M**ulti-omics

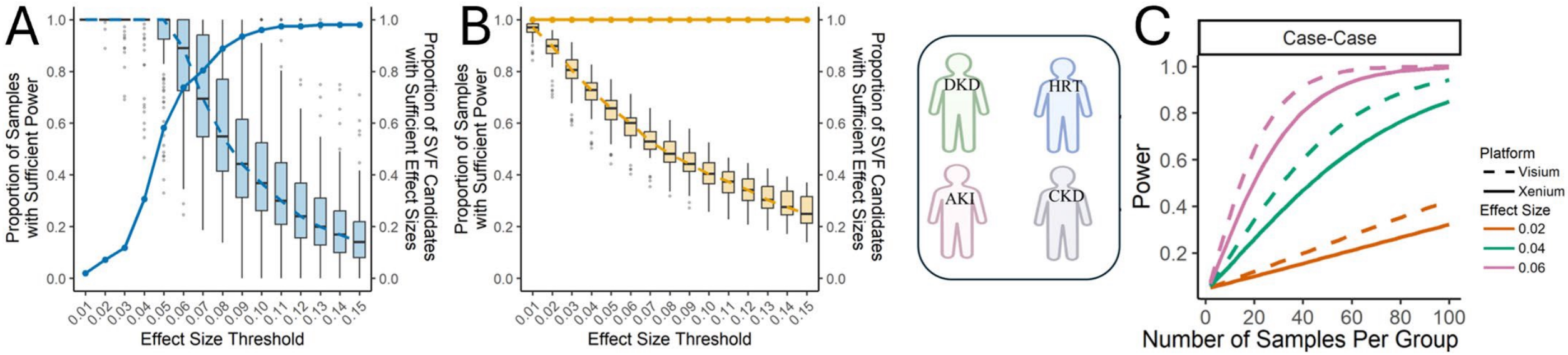
- A robust and efficient **nonparametric test** that quantifies local deviations in molecular feature measurements to detect spatial dependence.
- Estimates an interpretable **spatial effect size**, supports power calculations for both spatial locations and biological replicates, and enables **formal group-level comparisons**.





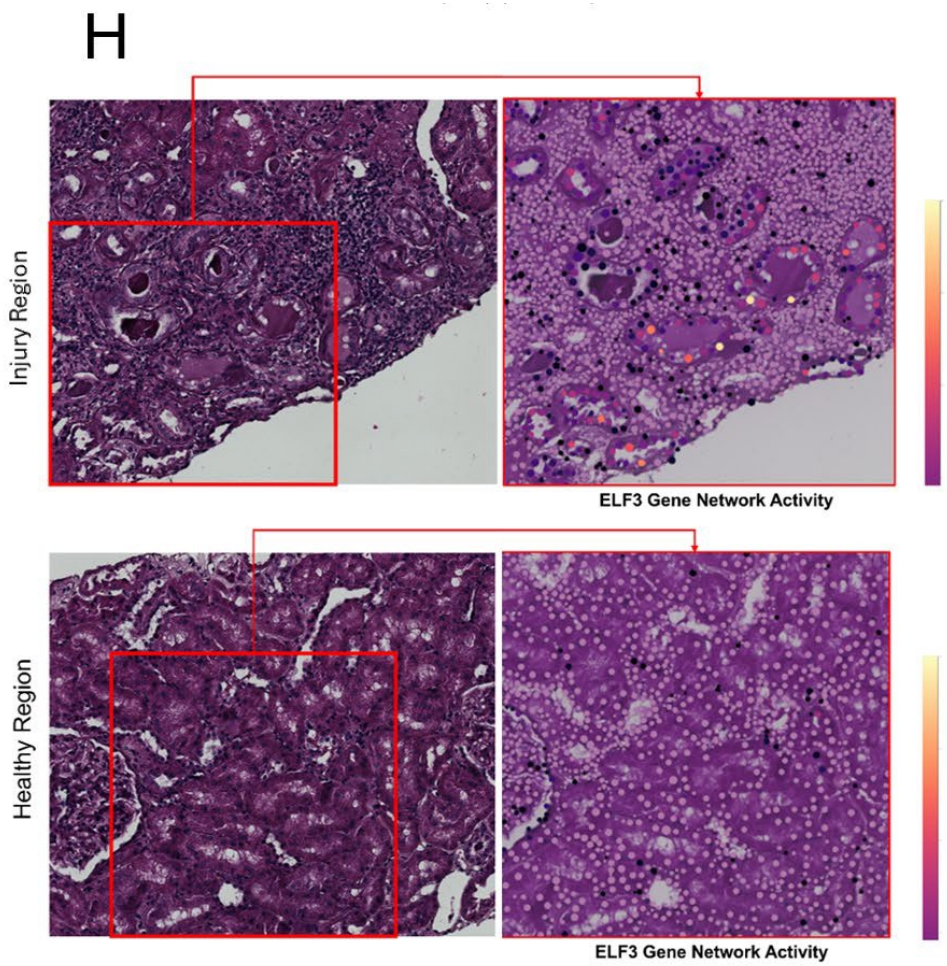
Li, Raina,...,**Wang**, Xu, A Principled Statistical Framework for Analyzing Spatial Patterns in Spatially Resolved Multi-Omics, bioRxiv 2026.08.04.742894

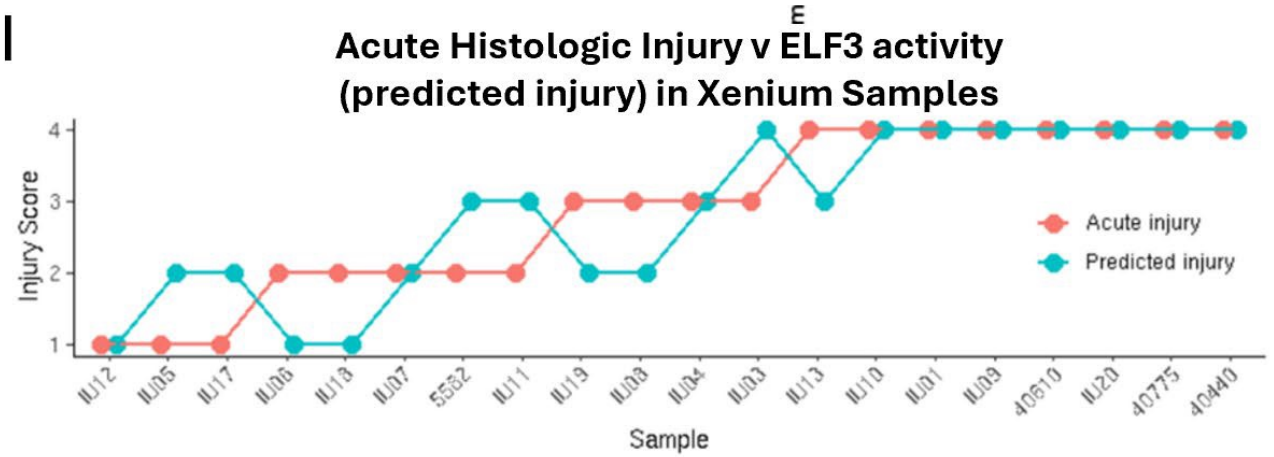
Sample Size Estimates for Disease Sub-phenotyping



Sample Size Estimates (A) Individual-sample power curves showing statistical power as a function of the number of spatial measurements for different effect size thresholds. Dashed line indicates 80% power. (B) Tradeoff between spatial variable feature (SVF) detectability and sample-level feasibility across effect size thresholds. Boxplots show the distribution across KPMP atlas v2 samples of the proportion of SVFs (adjusted p-value < 0.05) exceeding each effect size threshold. Solid Lines indicate the proportion of samples with sufficient spatial resolution to achieve 80% power. (C) Power curves for **group-level** comparisons to distinguish healthy reference tissue (HRT) from disease groups at 3 effect size thresholds. We include Visium with Xenium comparisons in this analysis since both datatypes were used in Atlas v2.

Spatial Variable Feature Example





(H) Inferred TF activity in single cells based on aggregate target gene network expression. ELF3 activity is increased in AKI tubules with casts and epithelial simplification compared to healthy PTs.

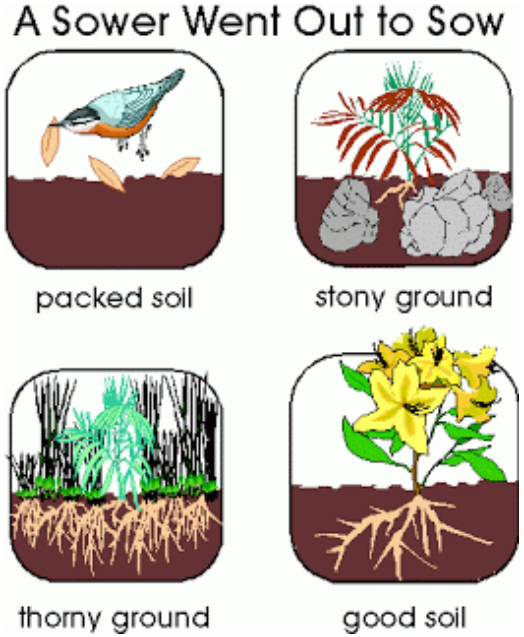
(I) Correlation of pathomic tubular injury on histology with ELF3 TF activity in 20 Xenium samples.

Outline

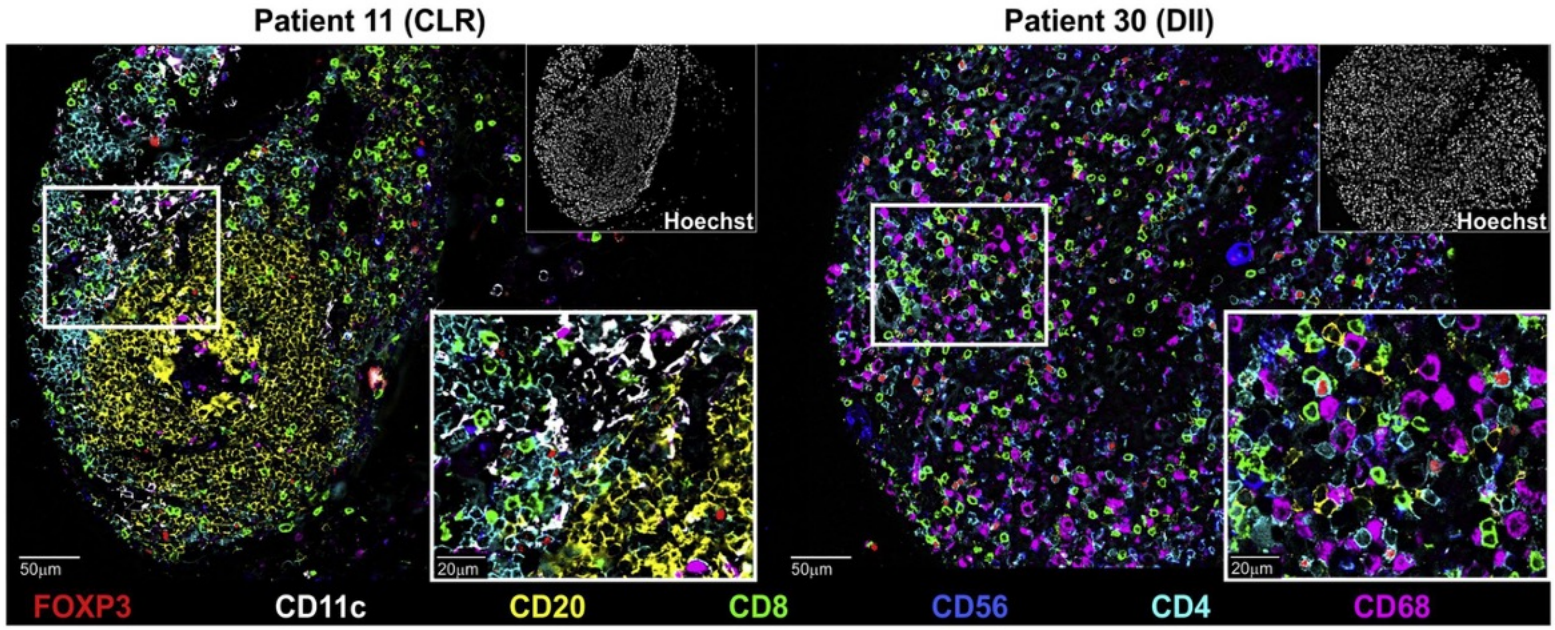
- Q1: How to resolve complex biological heterogeneity in spatial transcriptomics data?
- Q2: How to identify molecular markers that define biologically meaningful spatial organization?
- **Q3: How does spatial organization link to tissue functions?**
- Discussion and perspective

Phenotype=Genotype + Environment

- Severity represented by cell organization, described by environment using spatial omics
  - Cell organization as environment
- Several fundamental biological questions are unknown:
  - How cell organized together in a spatial temporal manner
  - Why cell organized in different organs, diseases, or conditions



The seed and soil hypothesis
Cell-cell interaction changes cell phenotype, not cell type

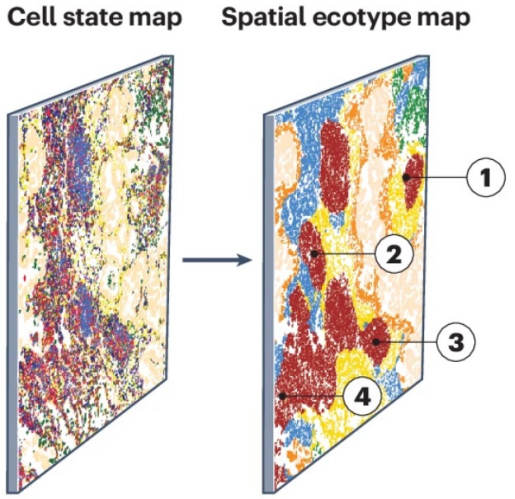


http://www.otherfood-devos.com/2015/06/the-seeda-modern-parable.html

Schürch, et al. "Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front." *Cell* 182.5 (2020): 1341-1359.

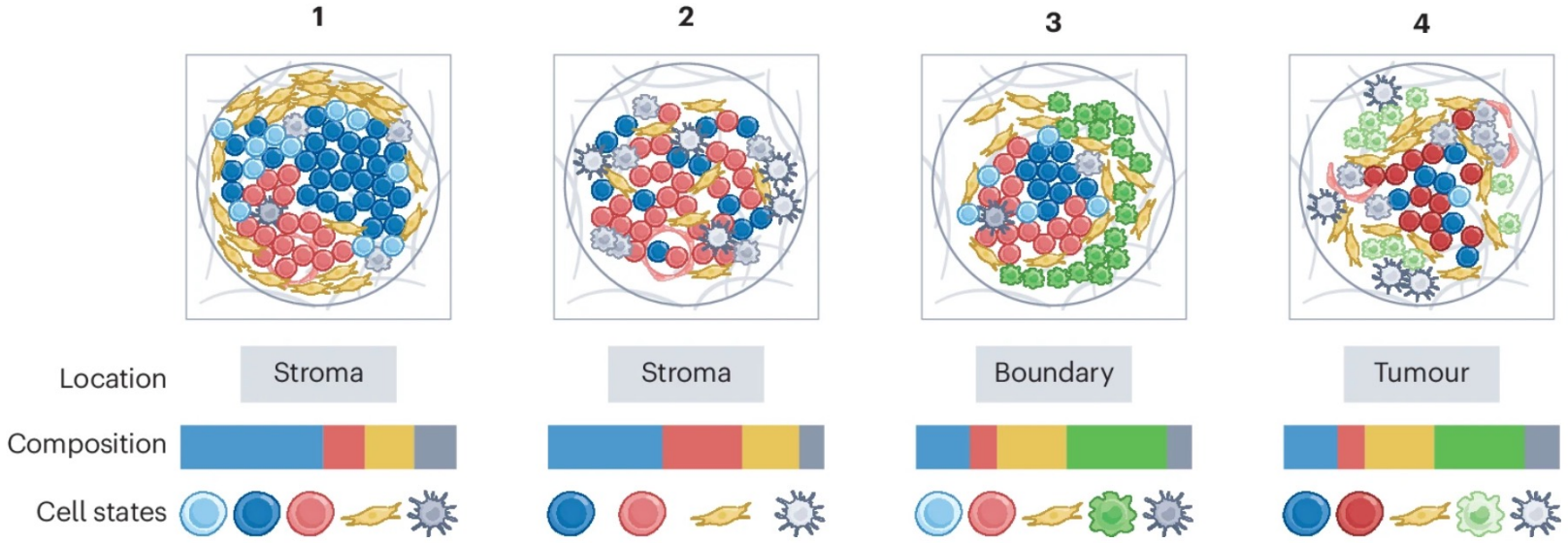
Limitations of cell niches defined by top-down approaches

- **Cell niches** are groups of cells that are spatially co-localized and share similar functional or phenotypic characteristics.
- Modeling cellular communities as a graph
- Clustering on learnt embedding based on the cell composition
  - SpaceGM
  - CytoCommunity
  - BANKSY
  - …
- Pros: Fast
- Cons:
  - **Expressibility**
  - Clustering
  - Interpretability
  - Generalizability



Can we have an alternative bottom-up view?

CNs from four different tertiary lymphoid structures (TLS) located within the tumor or surrounding stroma

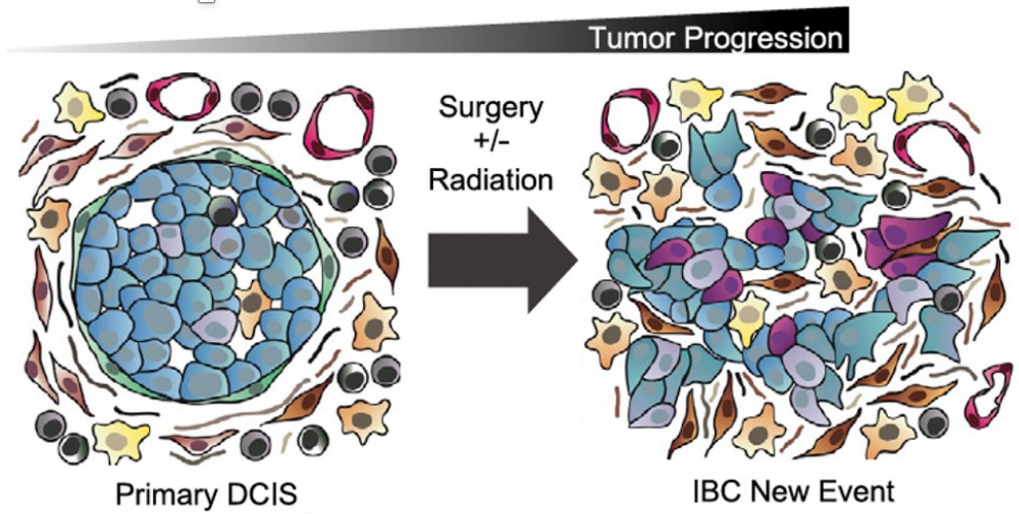


1 and 2 share **similar composition of major cell types**, but their detailed structural organization and cell states are different

Gulati, G.S., D'Silva, J.P., Liu, Y. *et al.* Profiling cell identity and tissue architecture with single-cell and spatial transcriptomics. *Nat Rev Mol Cell Biol* **26**, 11–31 (2025).

Conceptional:
Quantify cell organization as building blocks

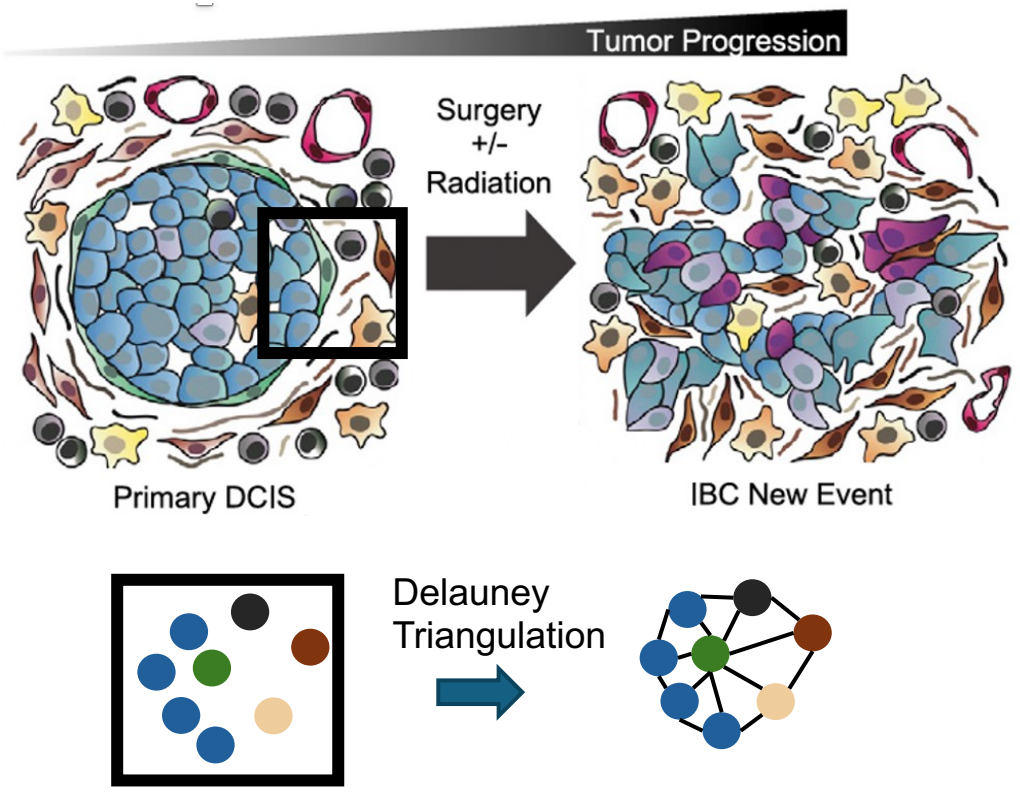
- **Central hypothesis**: There are conservative patterns as building blocks in cell organization that determine cellular function.
  - Topology->gene activity/cell-cell communications->cellular function
- Can we answer these questions?
  - Can we *quantify* the patterns of cell organization?
  - Can we *interpret* the relation between the patterns and biological contexts?
  - Can we *predict* the outcome of disturbing the patterns?
  - Can we *design* patterns for specific outcomes?



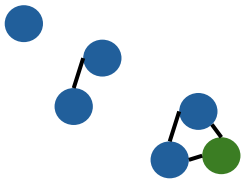
Risom et al., 2022, Cell 185, 299–310

Conceptual to formal:
Cellular Community motifs to quantify cell organization as building blocks

- **Hypothesis**: CC motifs can quantify cell organization in spatial omics data.
  - **Node** as cell, node label as cell type
  - **Edge** as cell relation, spatial or hypothetical relation, no direction
- CC motifs are **conservative** and can be **built up from scratch**.
- CC motifs are related to FTU but not identical.

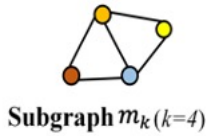


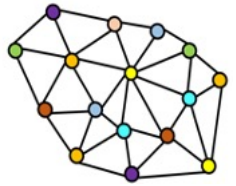
Cells: **size-1** (Cell type composition)
Edges: **size-2** (Cell-cell proximity)
Triangles: **size-3** (Proposed here)



Risom et al., 2022, Cell 185, 299–310

Bottom-up challenges: Finding most frequent size-k motifs

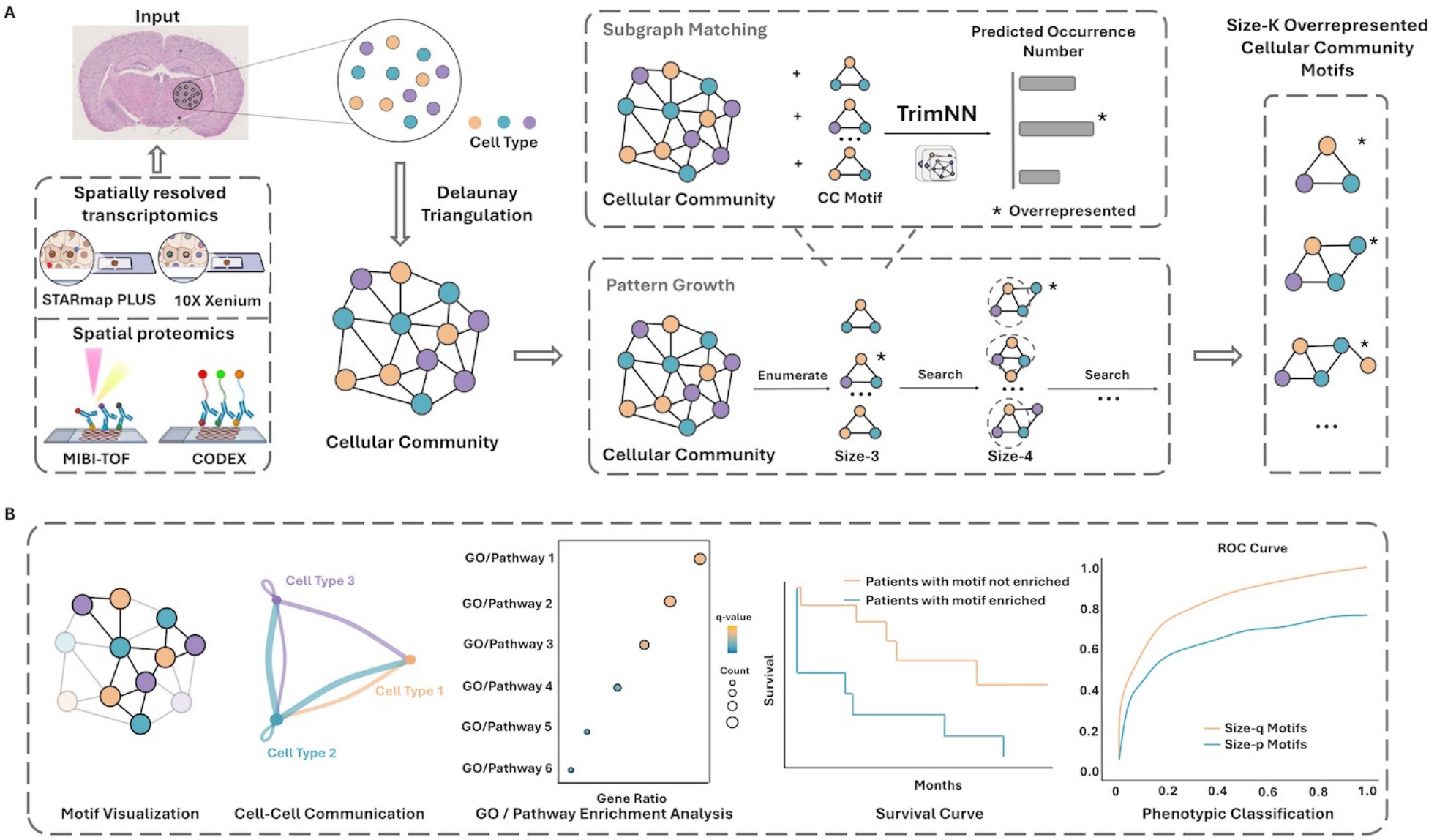




- **Goal:** Identify, among all possible graphs of k nodes, find the query graphs with the highest frequency in the target graph.
- Two sub-challenges:
- **1) Subgraph matching**: given a specific subgraph m_k, *counts* the occurrence of isomorphic subgraphs within the target graph G. $\boldsymbol{F(G, m_k)}$
- **2) Pattern growth**: *enumerates* M_k and identifies the most overrepresented subgraphs based on their occurrences in the **subgraph matching**. $\boldsymbol{F(G, M_k)}$
- **Infeasible** to calculate the exact number of large-size subgraphs
  - Subgraph isomorphism is **NP-complete** + Node type combination
  - Computation time *grows exponentially* as the size of the subgraphs increases
  - Feasible motif size for traditional methods is relatively small (1 to 3)
    - Edge sampling (e.g., MFinder), node sampling (e.g., FANMOD), and global pruning (e.g., Ullmann and VF2)

Can ML help to approximate large-size motifs?

TrimNN: Characterizing cellular community motifs for studying multicellular topological organization in complex tissues



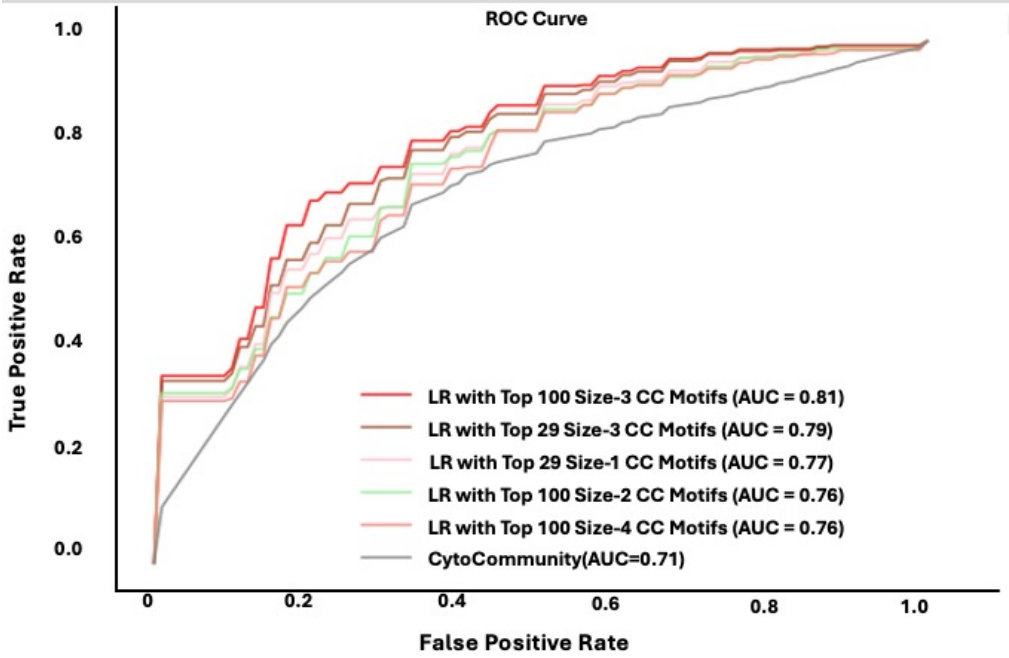
Yang Yu
U of Missouri

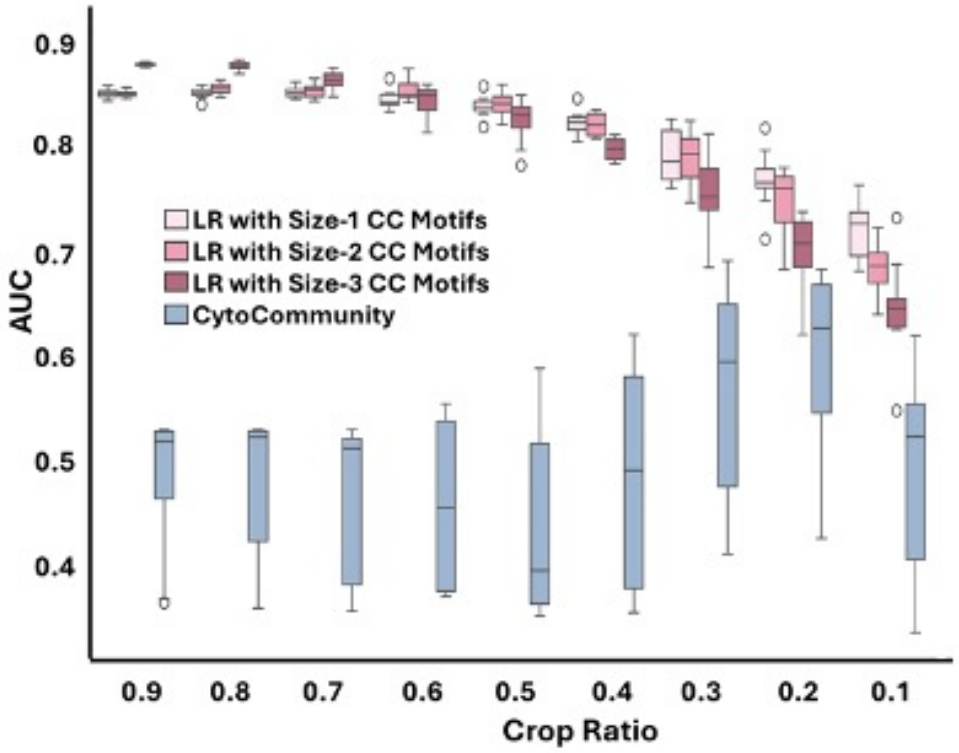
Shuang Wang
Indiana University

Yang Yu, Shuang Wang, Jinpu Li, Meichen Yu, Kyle McCrocklin, Jing-Qiong Kang, Anjun Ma, Qin Ma, Dong Xu, **Juexin Wang***, TrimNN: Characterizing cellular community motifs for studying multicellular topological organization in complex tissues, ***Nat Commun*** 16, 7737 (2025)

CC motifs as features for better **generalizability** cross different samples

- Colorectal cancer using CODEX
  - 17 low-risk (Crohn's-like lymphoid reaction (CLR))
  - 18 high-risk (diffuse inflammatory infiltration (DII))
- Robust in new samples by random cropping
- Similar results in RF, SVM
- Fewer features
- Simpler model
- **Interpretable** model





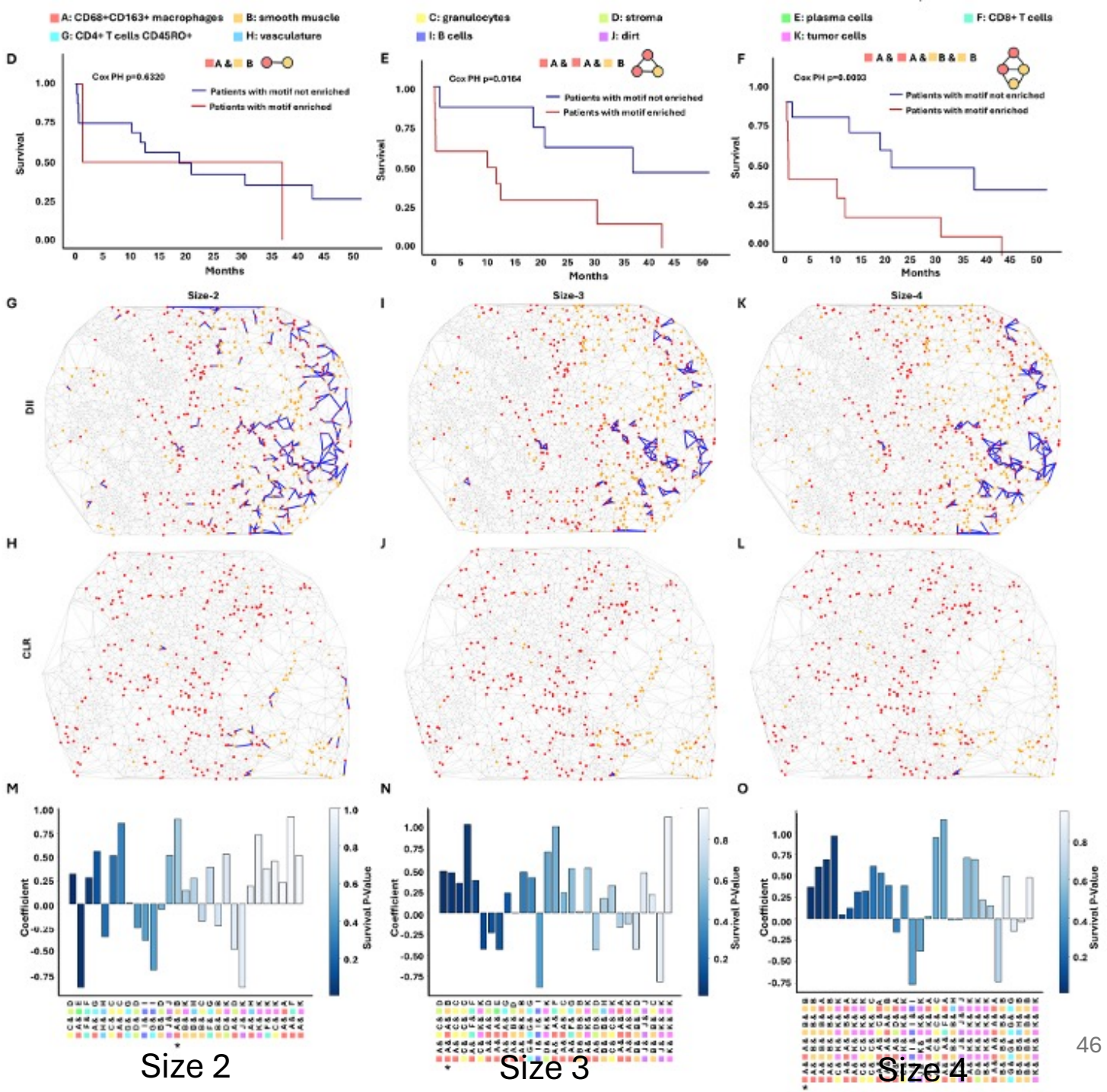
Schürch, Christian M., et al. "Coordinated cellular neighborhoods orchestrate antitumoral immunity at the colorectal cancer invasive front." *Cell* 182.5 (2020): 1341-1359.

- **Clinical impact**

- Enriched CC motifs can differentiate survival in various sizes:

- Size2: Marcophage + Muscle
- Size 3: Marcophage + Macrophage+Muscle
- Size 4: Marcophage+ Marcophage+Muscle +Muscle

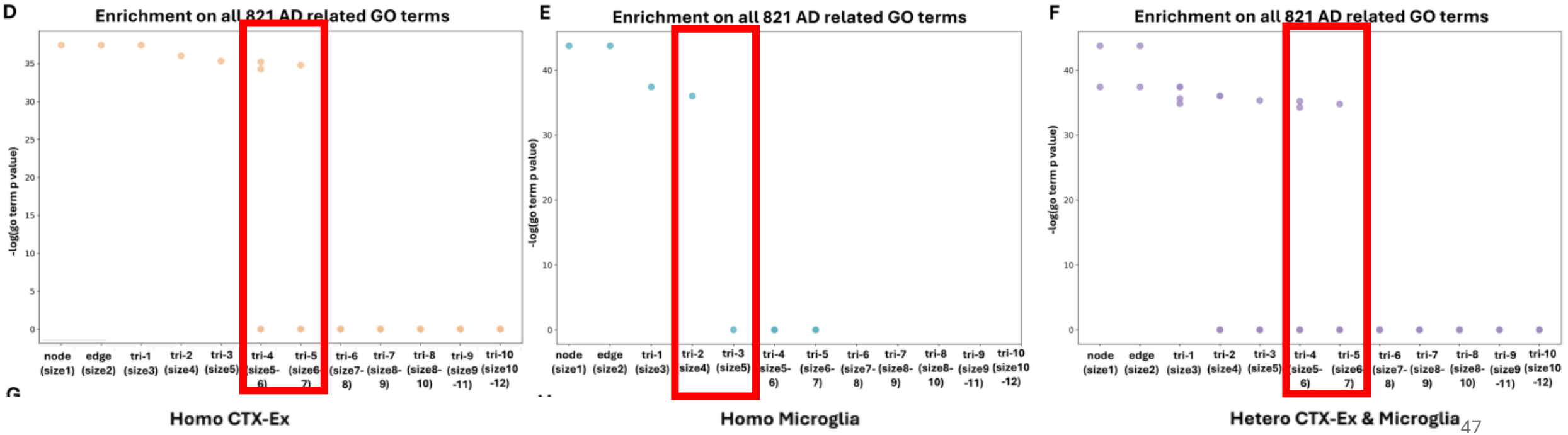
- Interpretability as coefficients from LR



Size 2 Size 3 Size 4

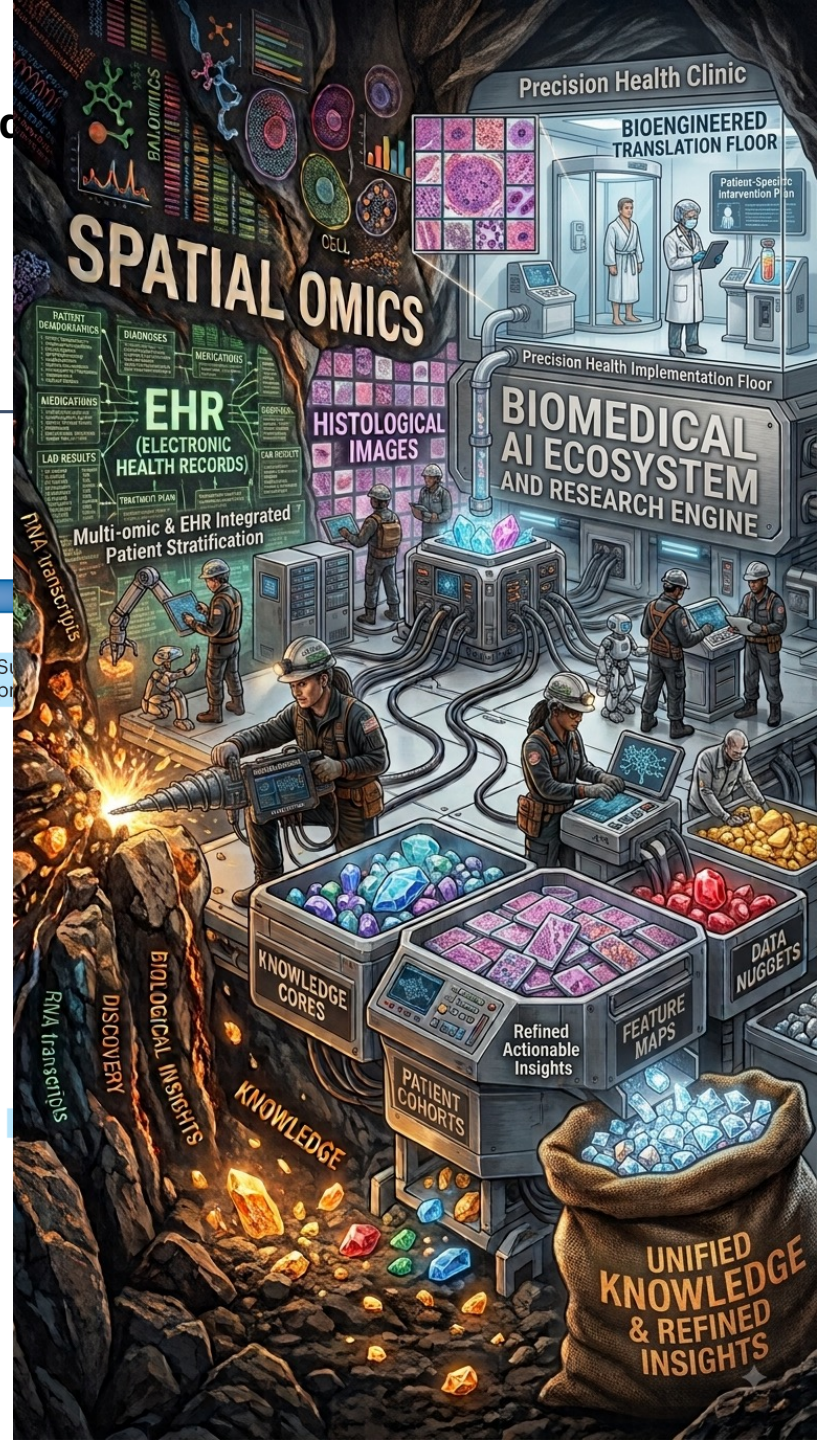
Evaluate which size is more appropriate

- No appropriate statistical-rigor approaches across different sizes
- Surrogating enrichment P-values of GO terms
- Meta-results



Summary

- Data analytics approaches are far behind data generation
- ML and Foundation Models have huge opportunities to explore fundamental biological and pathological questions
- Q1: Heterogeneity modeling
  - REGNN/spaGFM/spaOT
- Q2: Identifying SVGs
  - BSP/scBSP/STORM
- Q3: Relation between structure and function
  - TrimNN
- **More accurate, less consumption, deeper insights**



Bressan, Dario, Giorgia Battistoni, and Gregory J. Hannon. "The
Schäfer, Philipp Sven Lars, et al. "Integrating single-cell multi-or
characterization of the immune system." *Nature Immunology* (20

wangjuex@iu.edu

Acknowledgement

Lab Member

Mauminah Raina
Shuang Wang
Yu Zhong
Kyle McCrocklin
…

ML and Data analysis

Dr. **Dong Xu** at University of South Florida
Dr. **Jinpu Li**
Yang Yu
Dr. Fei He
Dr. **Qin Ma** at Ohio State University
Dr. Anjun Ma
Dr. Yuzhou Chang
Xiaojie Jin
Dr. Soheil Kolouri at Vanderbilt
Dr. **Yikun Bai** (now at Purdue)
Dr. Chao Zhang at BU
Dr. Guangyu Wang at Houston Methodist
Dr. Sengyu Li
Dr. Santo Fortunato at IUB
Dr. Xuhong Zhang at IUB

Kidney Study

Dr. **Michael Eadon** at IU SOM
Dr. Hato Takashi at IU SOM
Dr. Yu Chen at Lilly

Organoid Study

Dr. **Eri Hashino** at IU SOM

Epilepsy Study

Dr. Jingqiong Kang at Vanderbilt

AD Study

Dr. Andrew Saykin at IU SOM
Dr. Kwangsik Nho at IU SOM
Dr. Meichen Yu at IU SOM
Dr. Jingwen Yan at IU Luddy
Dr. Priyanka Baloni at Purdue

Cancer Study

Dr. Alan Wang at IU SOM

Other Studies

Dr. Charlie Dong at IU SOM
Dr. Jianjiong Gao at MSKCC
Dr. Zheng Jiang at Baylor
Dr. Yunlong Liu at IU SOM
Dr. Guoli Dai at IU SOS

Funding:

NIH/NIDDK R01DK138504
NIH/NLM R01LM015252
NIH/NIDCD R01DC022274
NIH/NIDCD R01DC013294

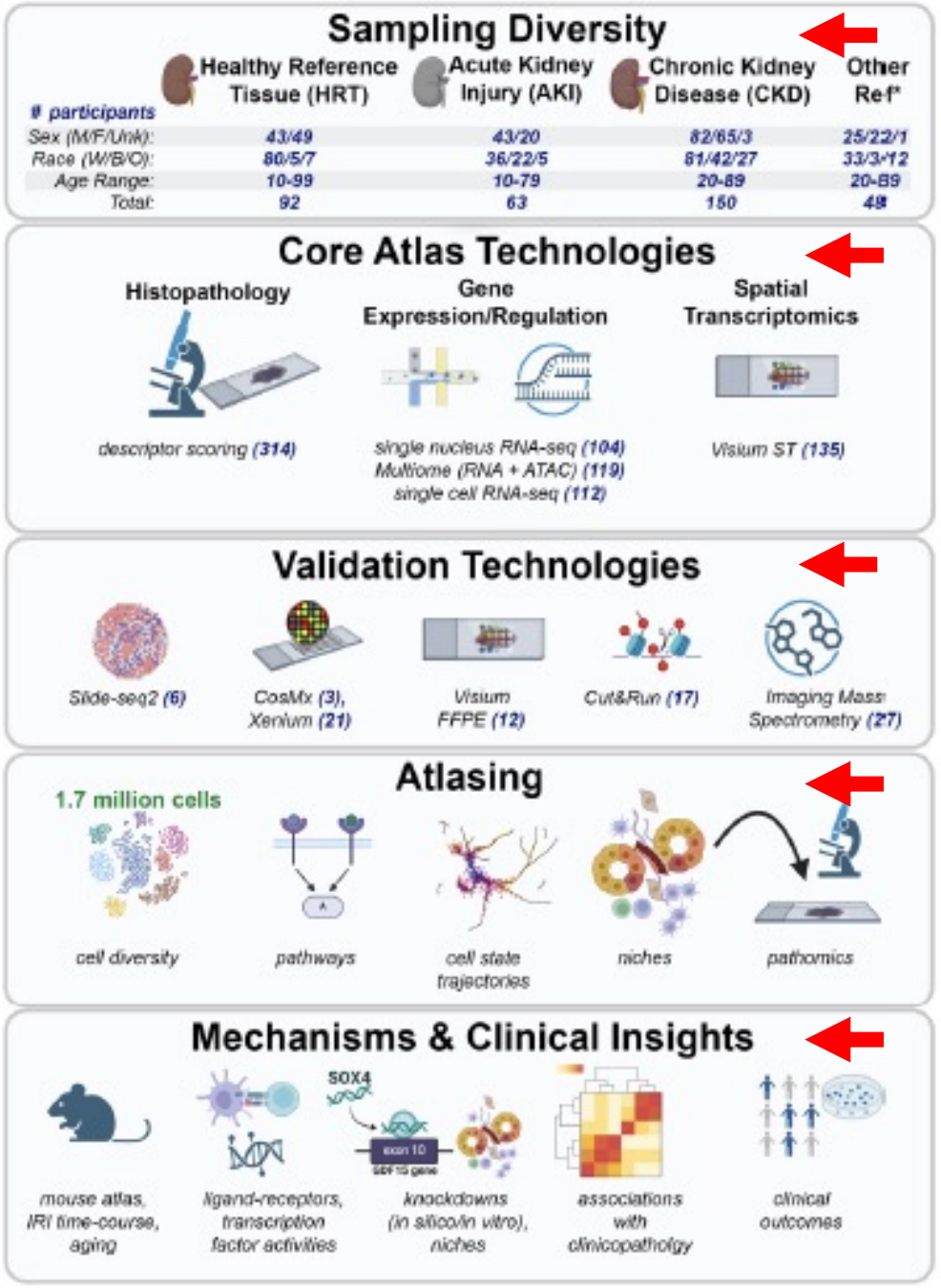
Sanjay Jain MD PhD

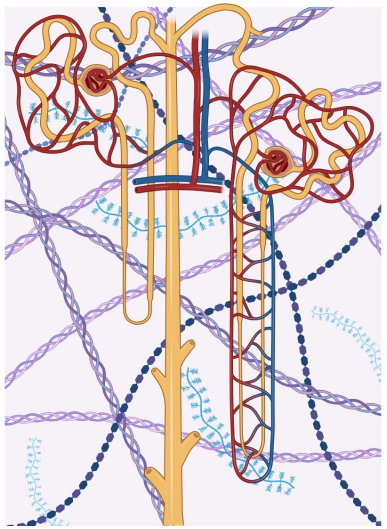
Cellular and Spatial Drivers of Unresolved Injury and Functional Decline in the Human Kidney

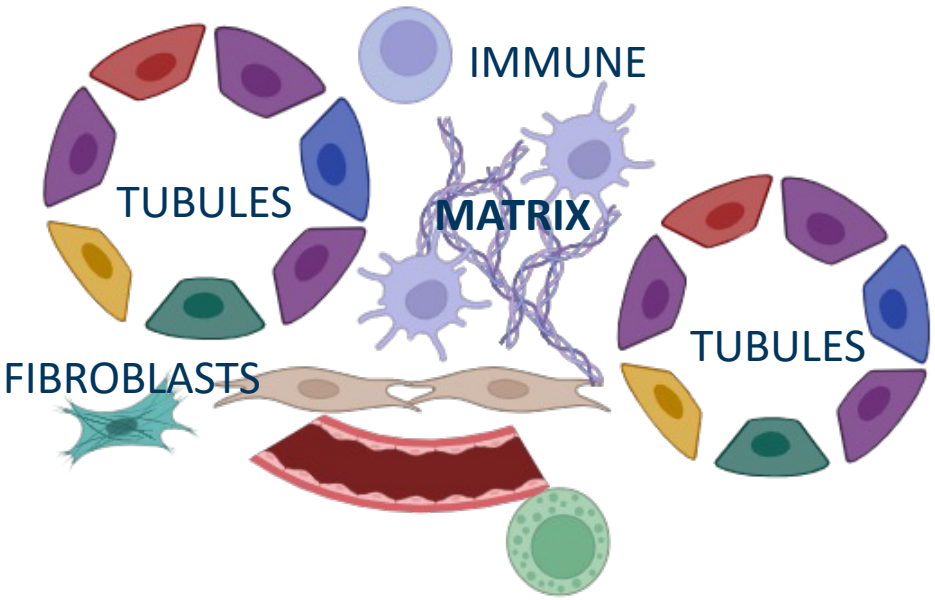
"Atlas 2.0" Publication

INDIANA UNIVERSITY · BROAD INSTITUTE · UC San Diego JACOBS SCHOOL OF ENGINEERING · WashU · PNNL · VANDERBILT UNIVERSITY · UT Health San Antonio · JOHNS HOPKINS MEDICINE · ALTOS · HuBMAP Human BioMolecular Atlas Program · KPMP

OVERVIEW / APPROACH of HKAv2



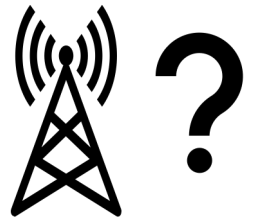




TUBULES – Healthy or Injured (**Kidney Function**)

FIBROBLASTS – Supporting infrastructure, adjust depending on tubular state

IMMUNE – Defend or Attack depending on tubular state

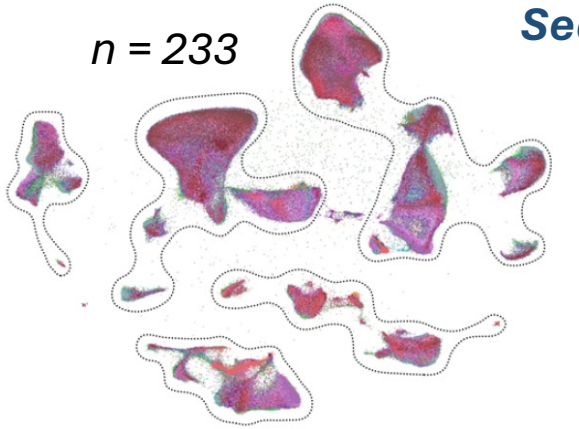


HKAv2

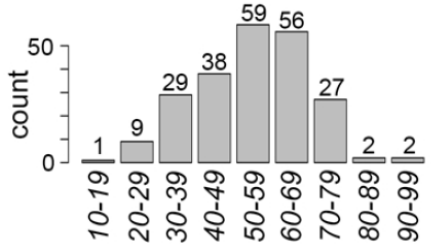
Single Nucleus RNA/Accessible Chromatin Sequencing

n = 233

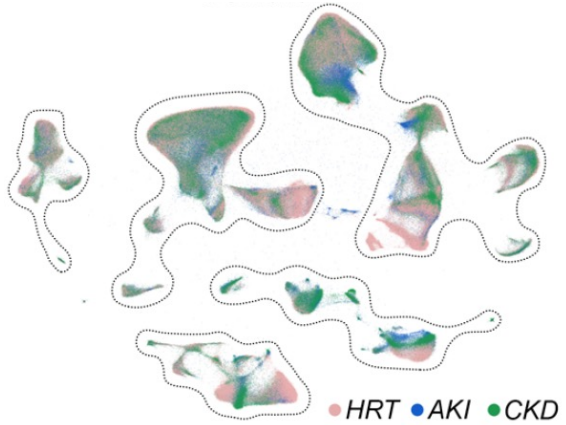
Patient

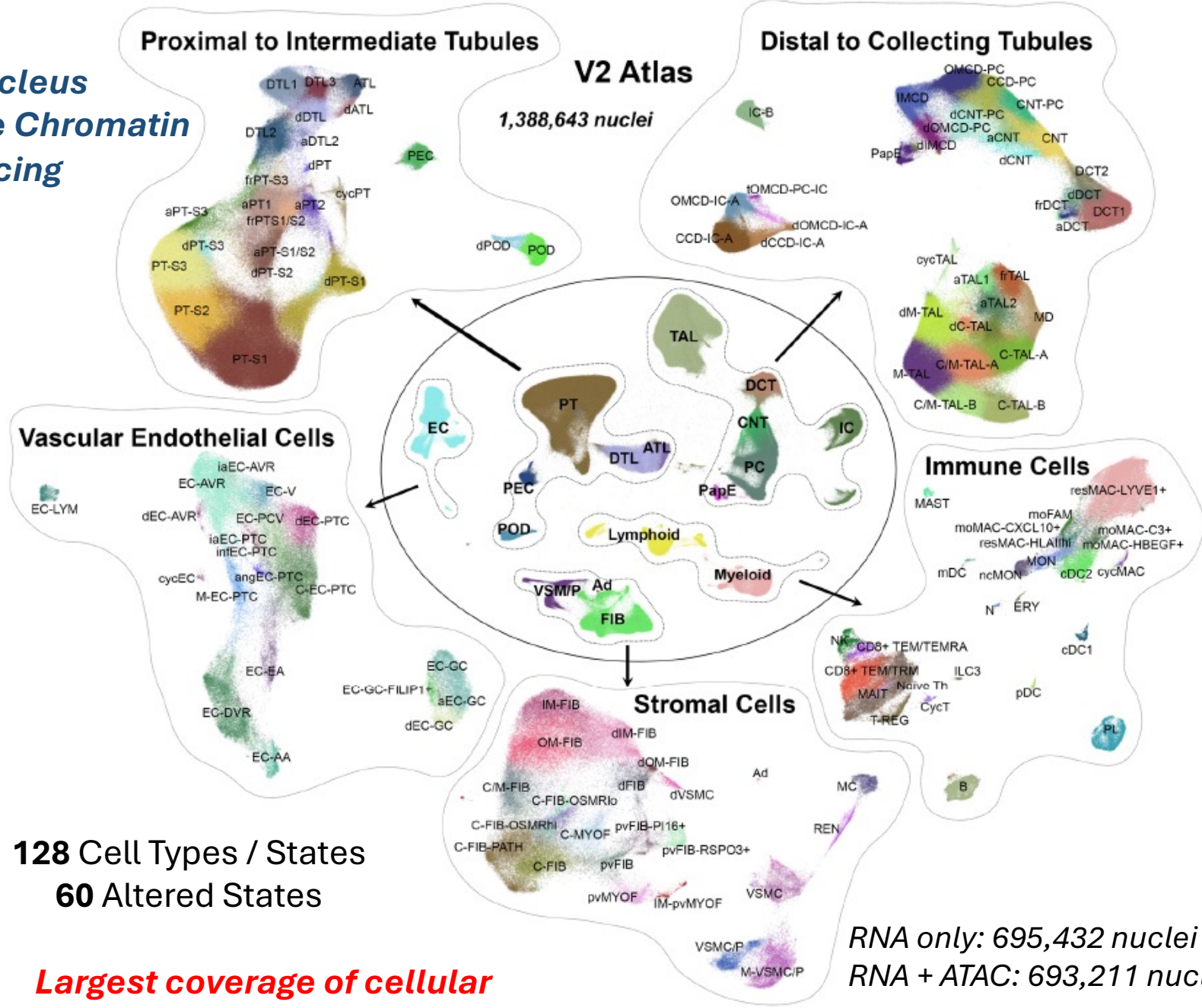


Patient Age



Condition



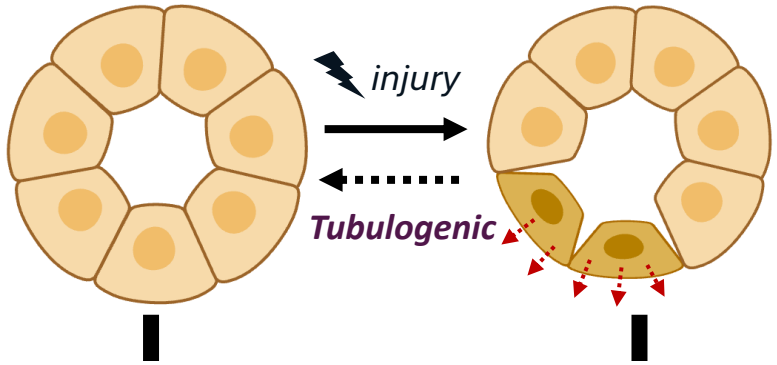


128 Cell Types / States
60 Altered States

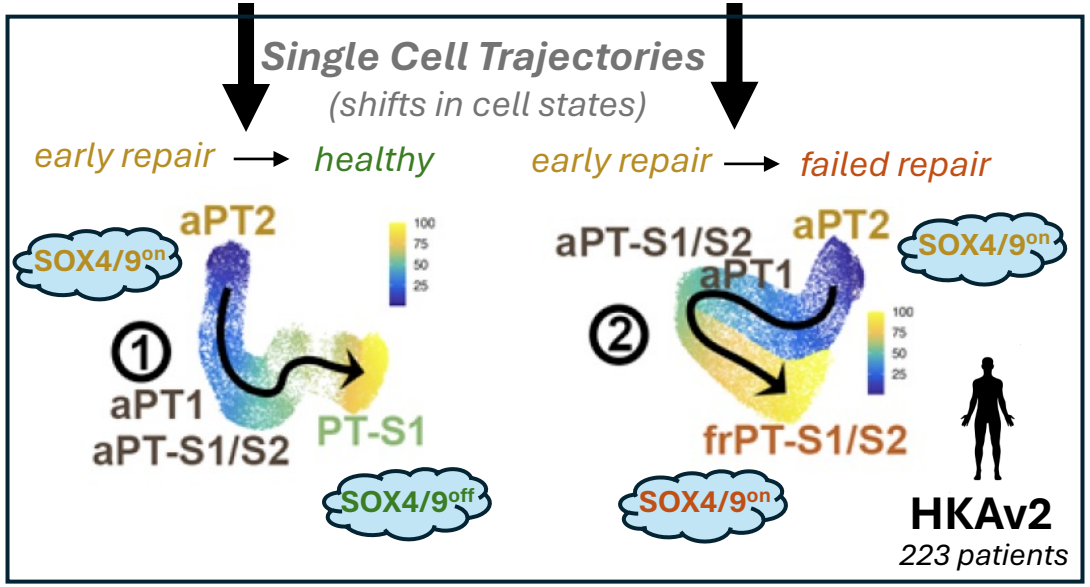
Largest coverage of cellular diversity

RNA only: 695,432 nuclei
RNA + ATAC: 693,211 nuclei

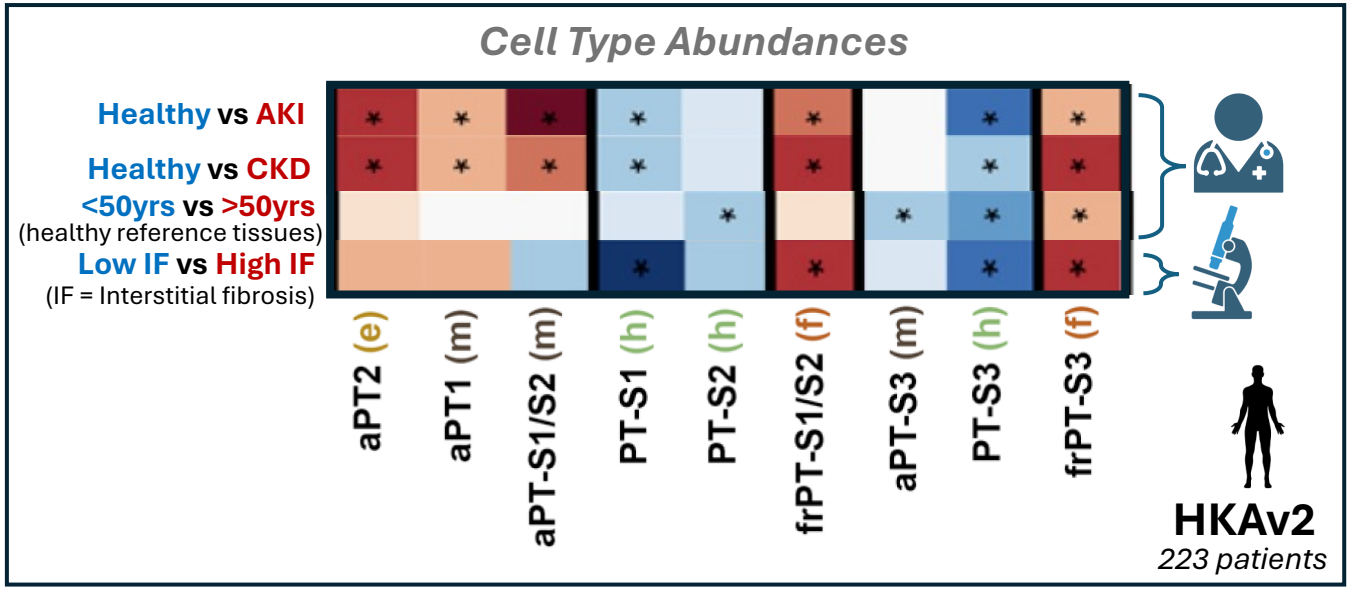
Unresolved Epithelial Repair



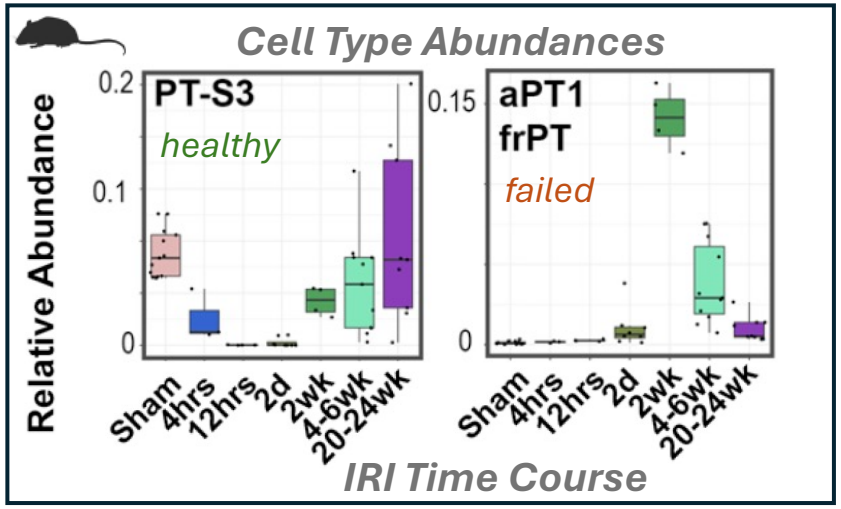
State Shifts in Kidney Injury/Disease



Enriched in Acute and Chronic Disease



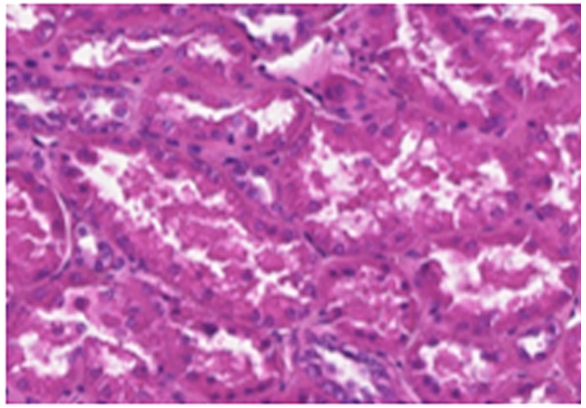
Conserved in Mouse Models



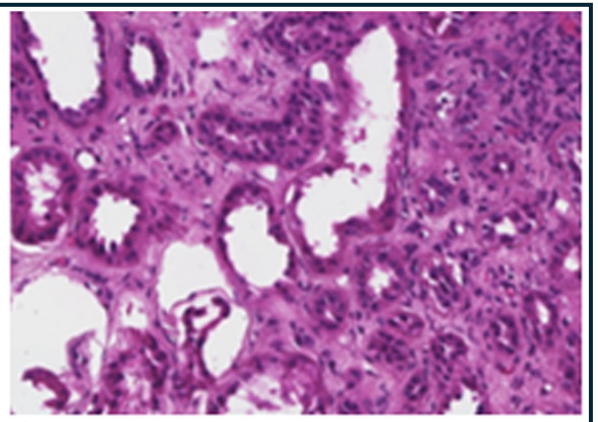
Unresolved (failed) epithelial repair is enriched with aging, CKD and fibrosis

Spatial Localization in the Kidney Atlas v2

Histologic Annotation of Kidney Biopsy Samples



Healthy Kidney



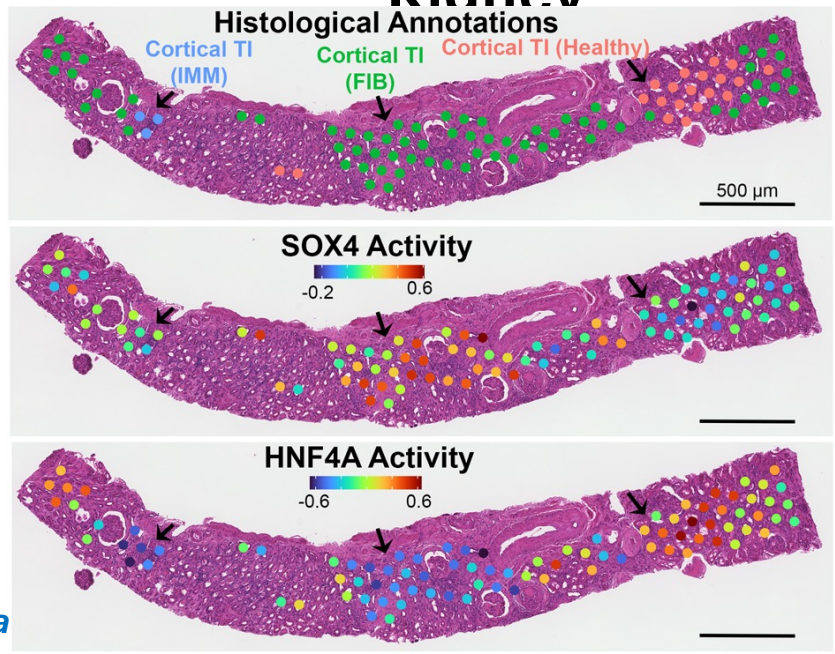
Fibrotic Kidney

How do the molecular signals we identified organize spatially?

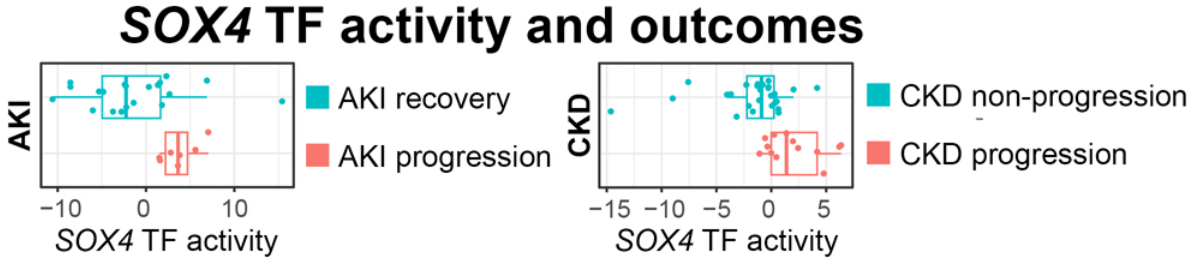
What are the histologic implications for disease progression?

Pathology links to molecular signature

Ricardo Melo Ferreira



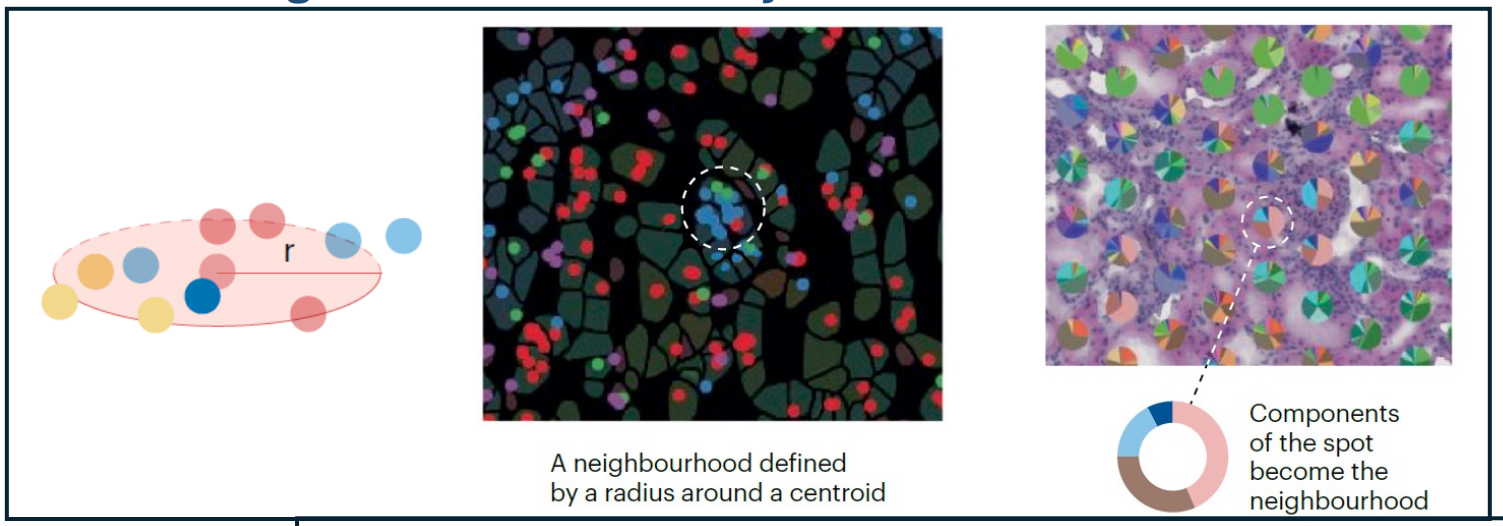
SOX4 Network mapping associated with AKI and CKD outcome in KPMP heroes



Common feature of AKI and CKD

Neighborhoods

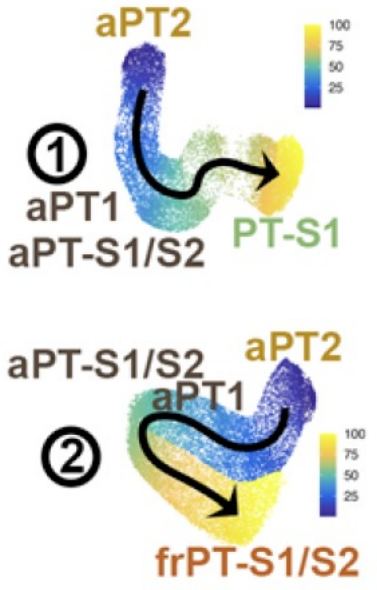
Neighborhoods defined by cells close to each other

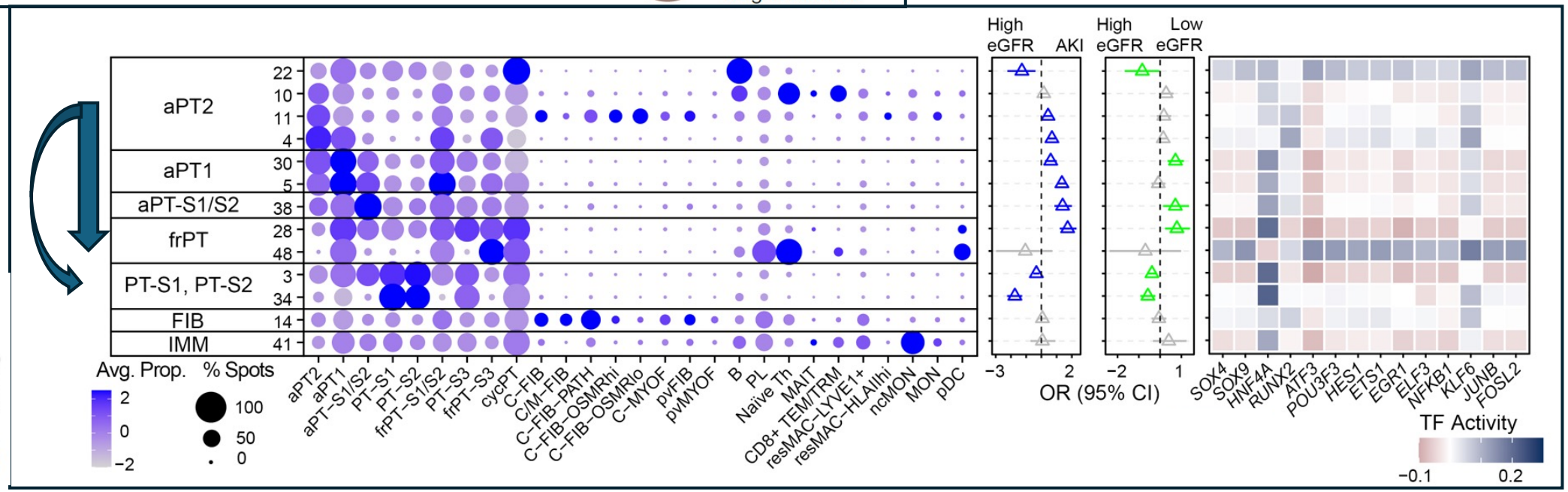


Which groups of cells are more common in disease?

What molecular signals correspond to each group of cells?

Is there a spatial time course?

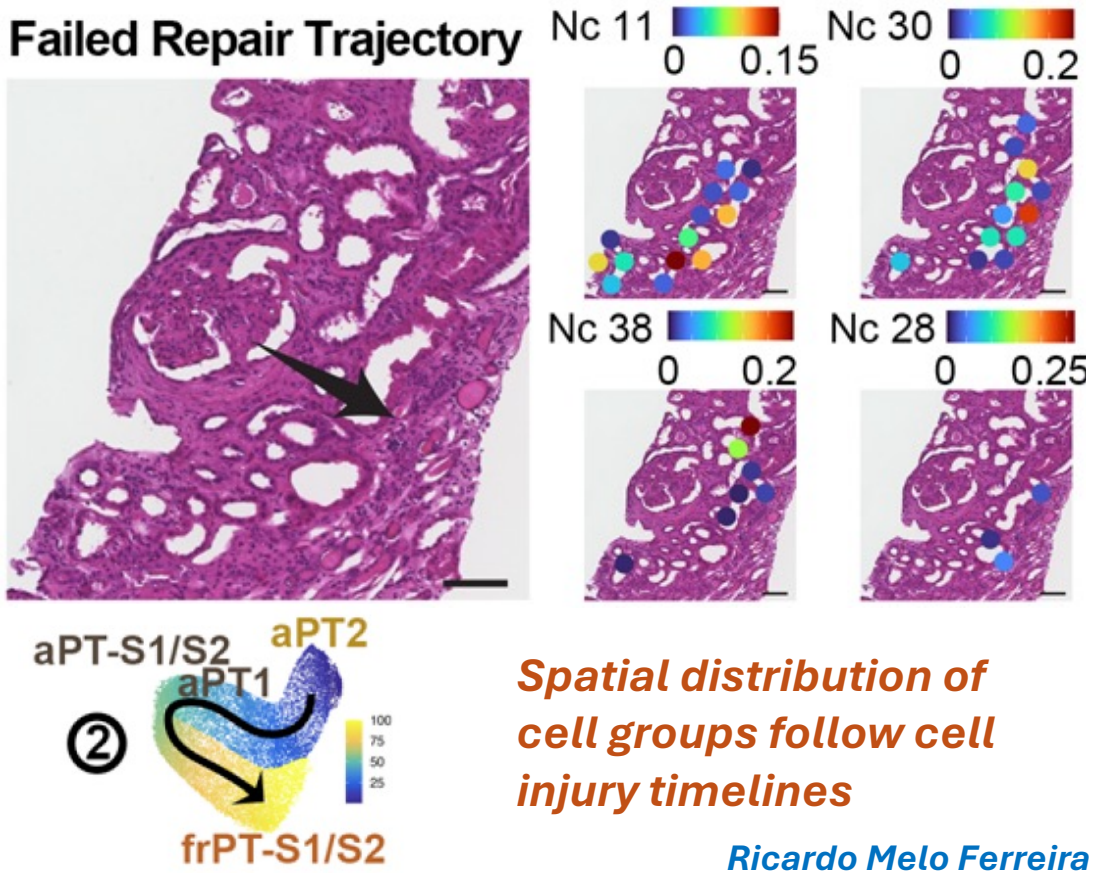




Ricardo Melo Ferreira

Neighborhoods

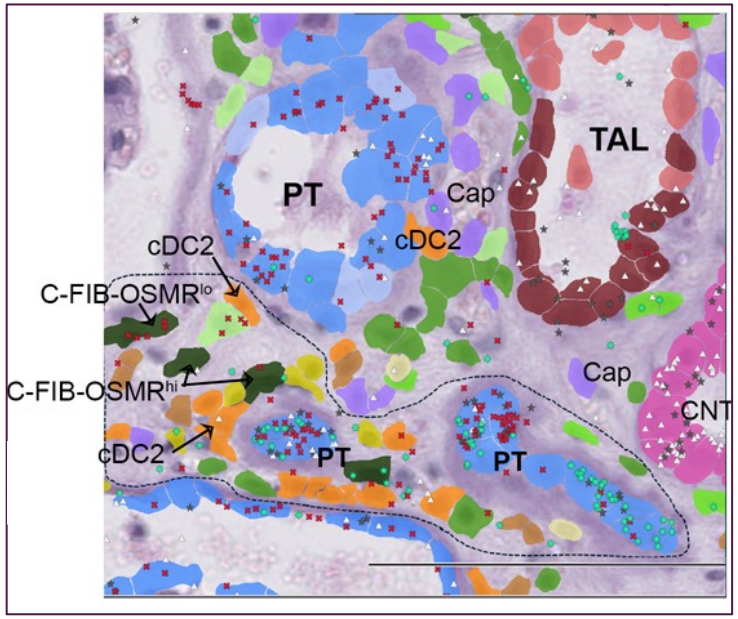
Temporal neighborhoods follow a spatial trajectory



Spatial distribution of cell groups follow cell injury timelines

Ricardo Melo Ferreira

Single Cell Neighborhoods

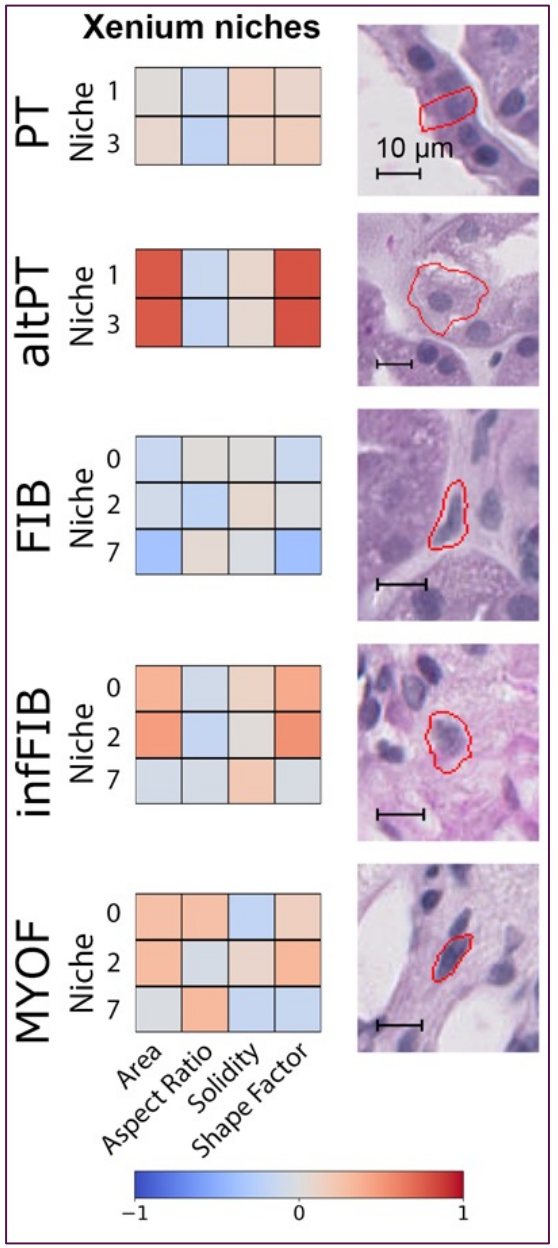


Stephanie Reinert, Asmita Lagwankar

Cells within a neighborhood validated at the single cell level.

Cells expressing SOX4 are seen more frequently in CKD. Healthy and injured cells have different morphologic features

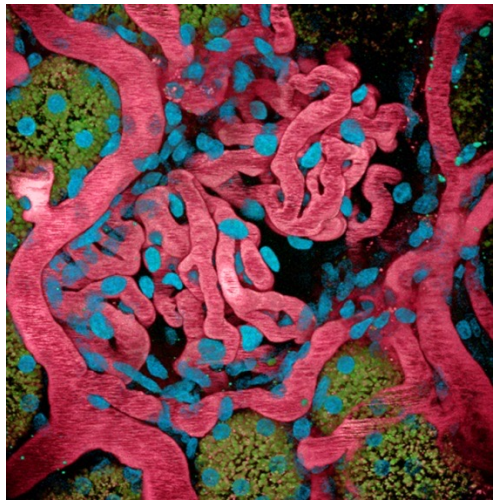
AI Cell Features



Nick Lucarelli, Pinaki Sarder

Acknowledgments

- Pierre Dagher
- Tarek Ashkar (El-Achkar)
- Ricardo Melo Ferreira
- Daria Barwinska
- Debora Gisch
- Mahla Asghari
- Yinghua Cheng
- Jumar Etkins
- Sachiko Koyama
- Gerald So
- Evan Rajadhyaksha
- Soumya Yennapureddy
- Angela Sabo
- Michael Ferkowicz
- M. Ahsan Sohail
- Kelsey McClara
- Abi Colwell
- Carrie Phillips
- Michael Rauchman
- Juexin Wang
- Qin Ma
- Dong Xu
- Sanjay Jain
- Blue Lake
- Chetan Poudel
- Pinaki Sarder
- Bill Bowen
- Matthias Kretzler
- Jeff Hodgin



Indiana O'Brien Center for Advanced Microscopic Analysis
3D Tissue Imaging Core

Funding:

U01 DK114923 (KPMP)
OT2OD033753 (HuBMAP HIVE)
U54 DK134301 (HuBMAP)
U54 DK137328 (O'Brien)
NIDDK P50 DK133943 (PCEN)
NIDDK R01 DK114485
NCCIH R01 AT011463
IIS R01 2306743
NIDDK R01 DK139676
NLM R01 LM013771
NIGMS T32 GM008425