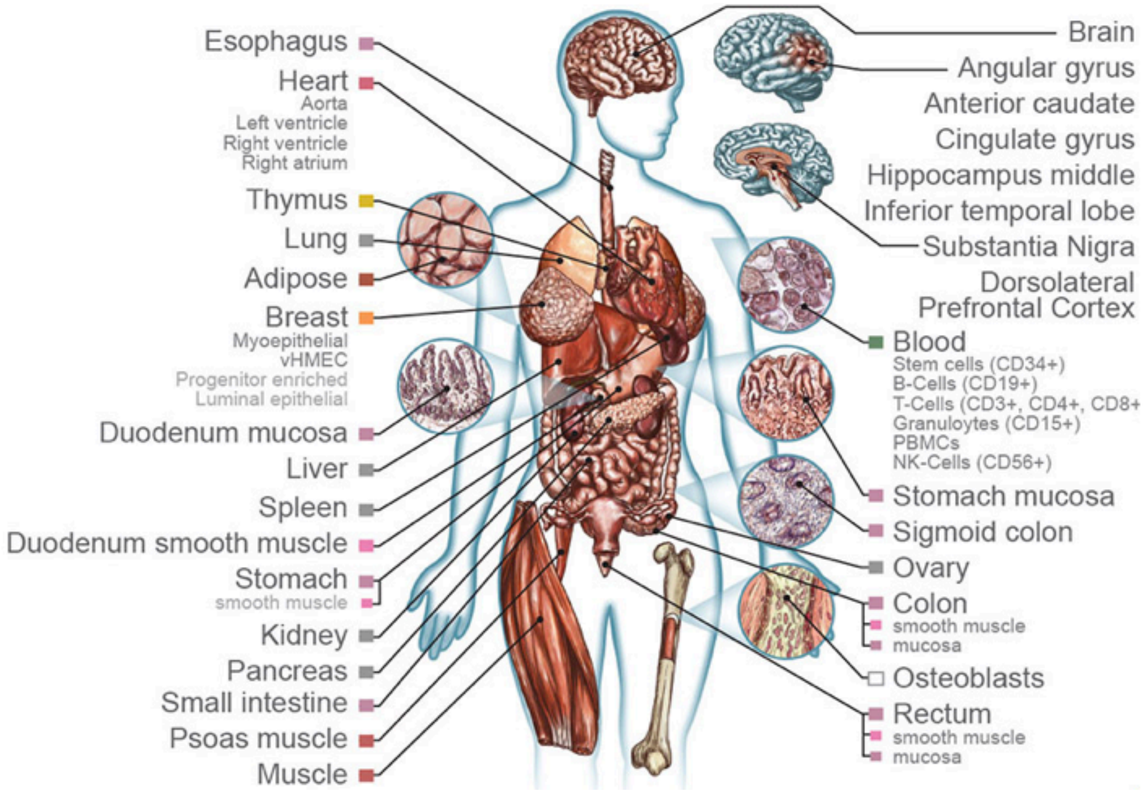
Decoding single-cell spatial genomics and epigenomics with AI/ML

Jian Ma

Ray and Stephanie Lane Professor of Computational Biology
Ray and Stephanie Lane Computational Biology Department
Director, Center for AI-Driven Biomedical Research
School of Computer Science
Carnegie Mellon University

September 4, 2025 | 02541 / 02741 / 11481 / 11781

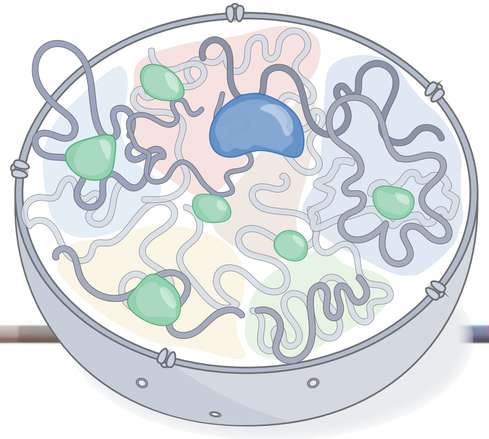
Genome function in different cell types

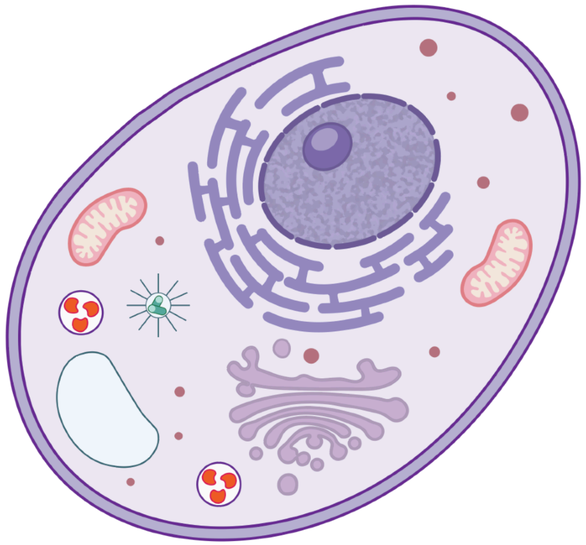


- Numerous different cell types with distinct functions in our body
- Same genome but different gene regulation and epigenome
- Development of single cell technology provides closer look at cell types and cell states

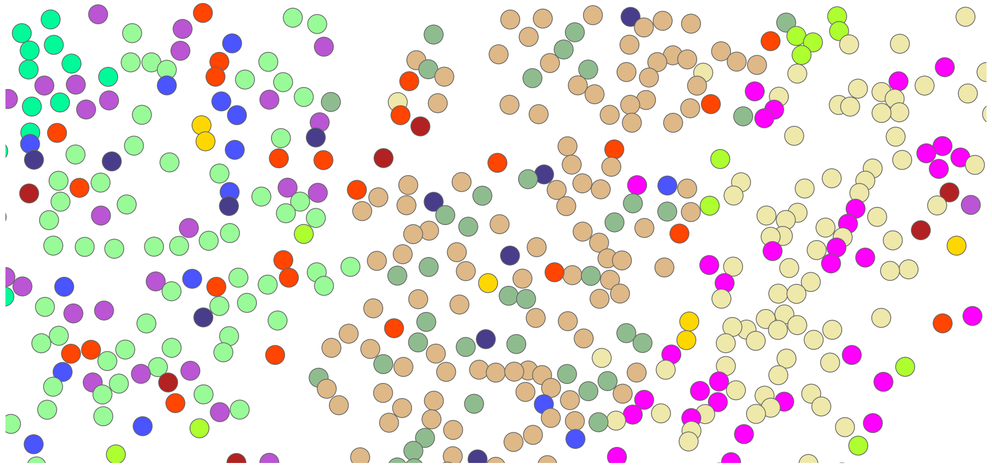
Yao et al. *Nature* 2023

Multiscale cellular structure and function



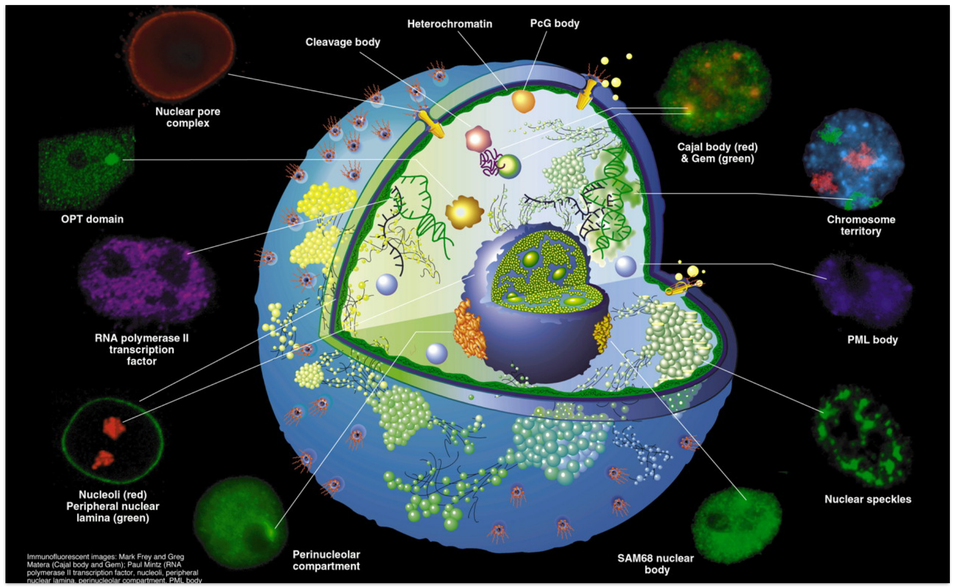


- Single-cell **3D epigenome** and gene regulation
- Cellular **spatial organization** and **interaction**
- graph / hypergraph neural network, self-supervised, latent embedding, metagene, attention, (foundation model)

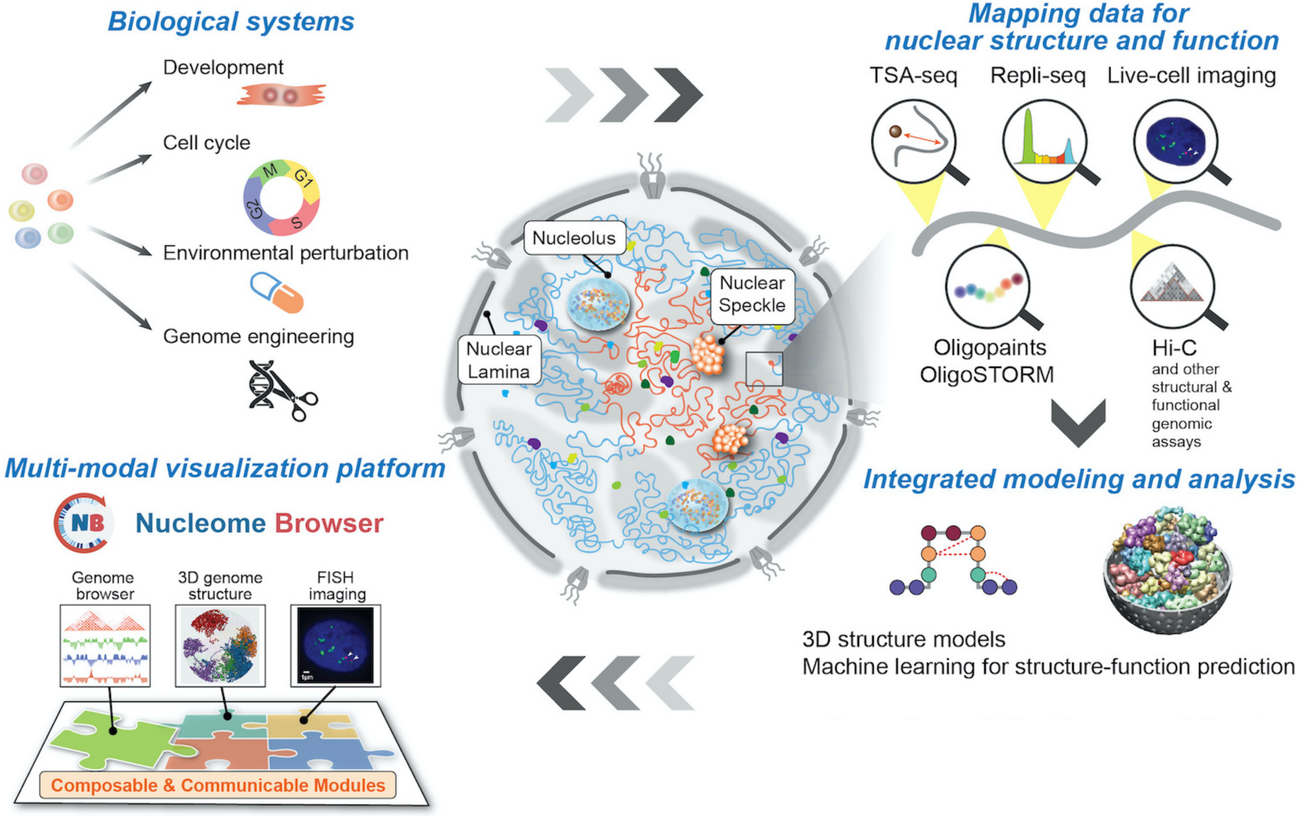


Visualization: Ruochi Zhang & Yang Zhang

Large-scale genome organization in the nucleus



Spector, *J of Cell Science* 2001

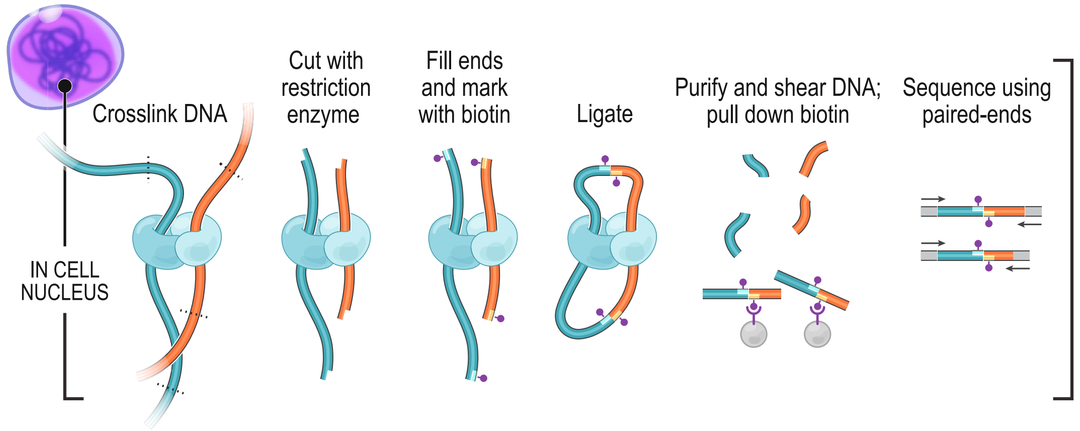


Multiscale Analyses of 4D Nucleome Structure and Function by Comprehensive Multimodal Data Integration

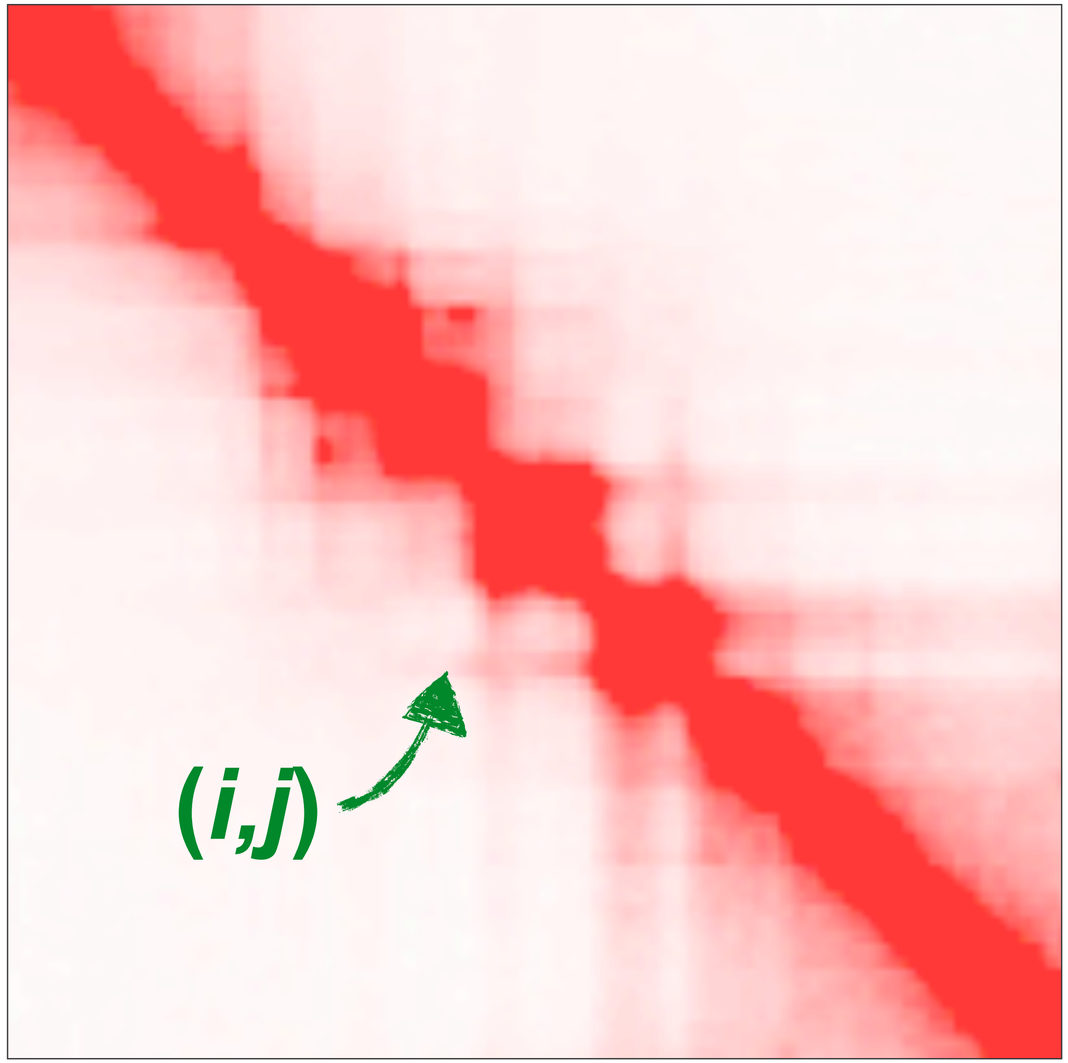
4D Nucleome

Carnegie Mellon University · ILLINOIS University of Illinois at Urbana-Champaign · Allen Institute for Cell Science · Harvard Medical School · UCLA · NIH National Cancer Institute · University of Dundee · San Diego Biomedical Research Institute · Brown

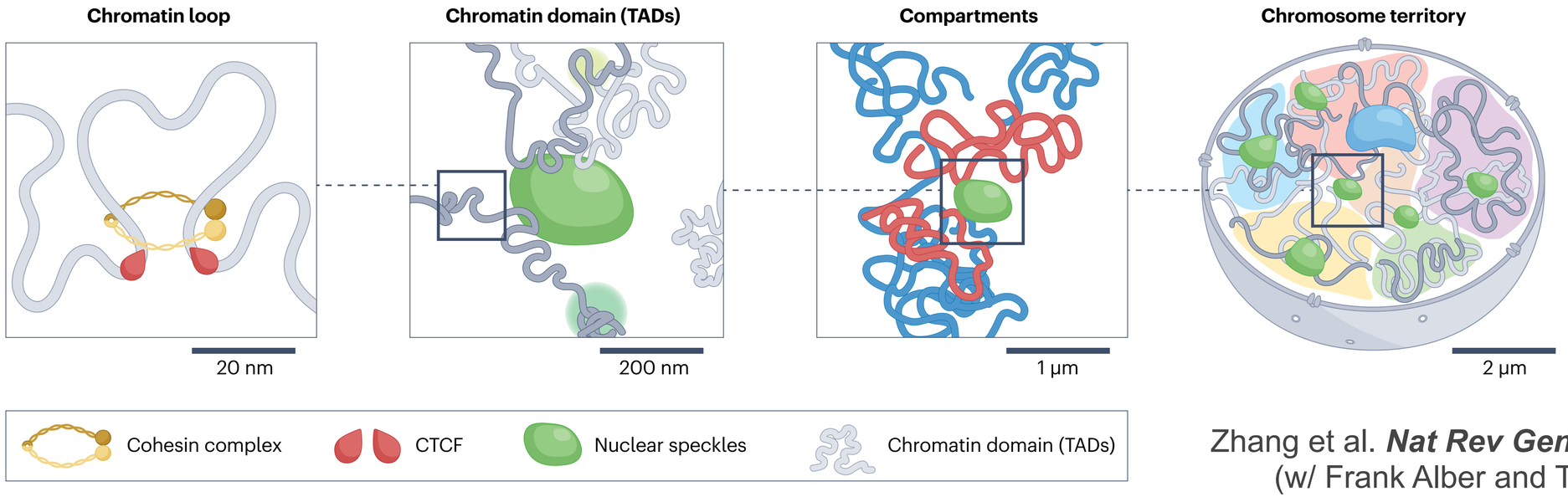
Hi-C contact map



Hi-C

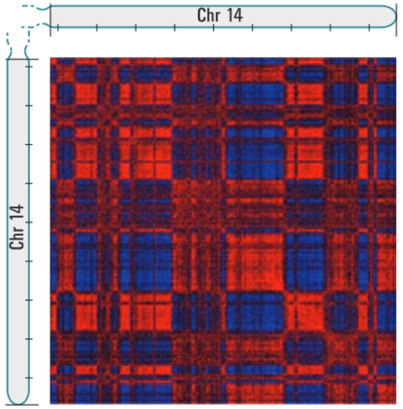


Multiscale 3D genome organization



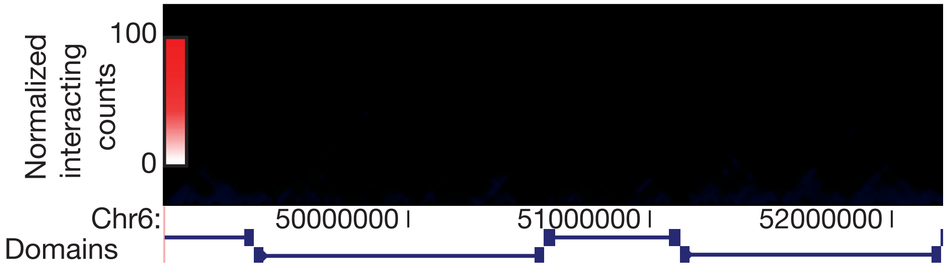
Zhang et al. ***Nat Rev Genetics*** 2024
(w/ Frank Alber and Tom Misteli)

A/B Compartments



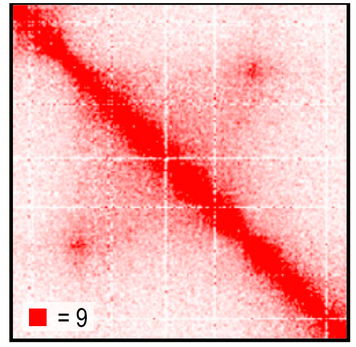
Lieberman-Aiden et al.
Science 2009

TADs



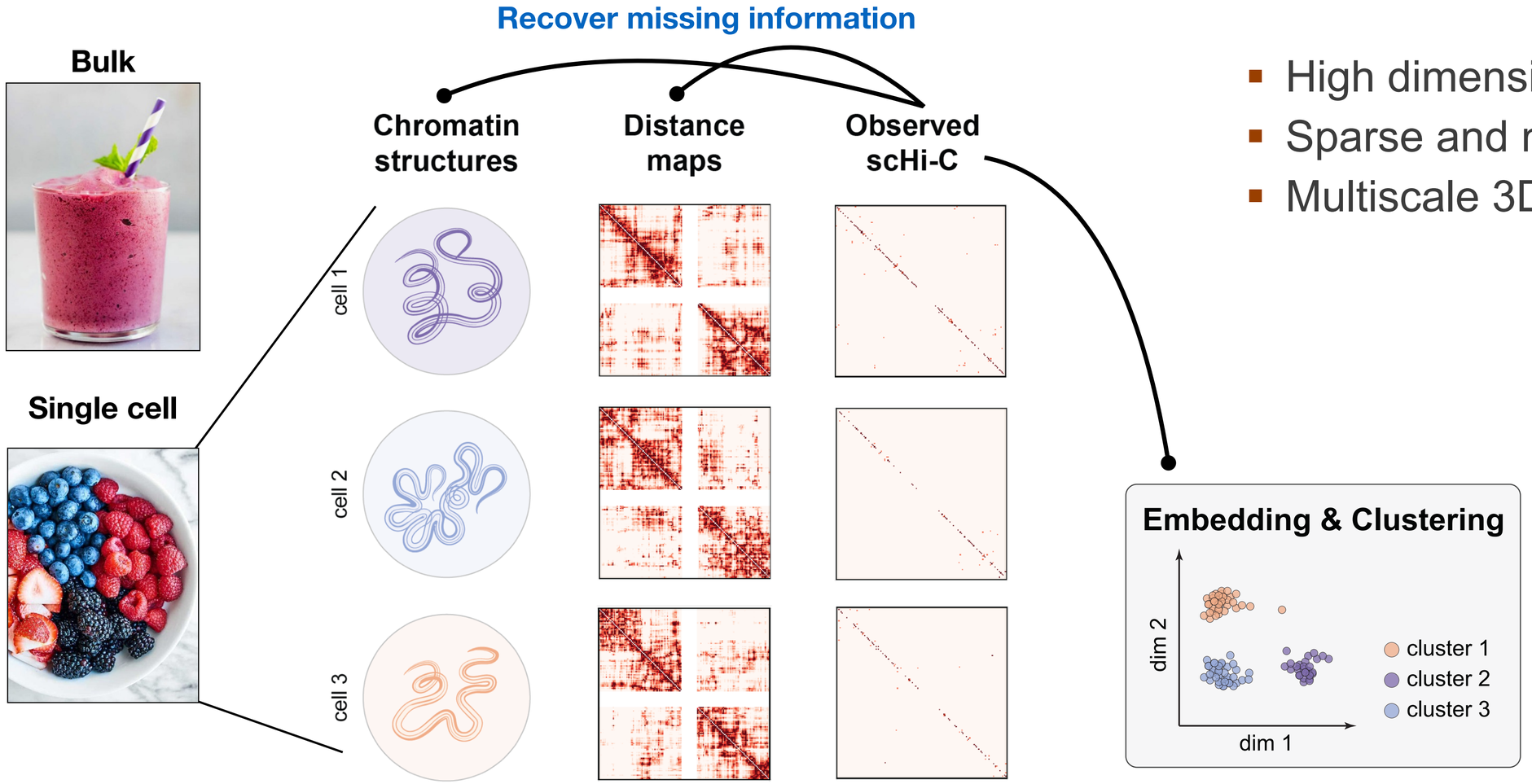
Dixon et al. *Nature* 2012
Nora et al. *Nature* 2012

loops



Rao et al. *Cell* 2014
Jin et al. *Nature* 2013
Fullwood et al. *Nature* 2009

Challenges for single-cell 3D genome analysis using scHi-C data



- High dimensionality
- Sparse and noisy data
- Multiscale 3D genome features

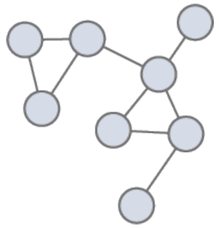
- What does 3D genome look like in **single cells**?
- How do multiscale 3D features vary in individual cells and different cell types?

Zhou et al. ***Annu Rev Biomed Data Sci*** 2021
Zhang et al. ***Nat Rev Genetics*** 2024

Hypergraphs for higher-order interactions

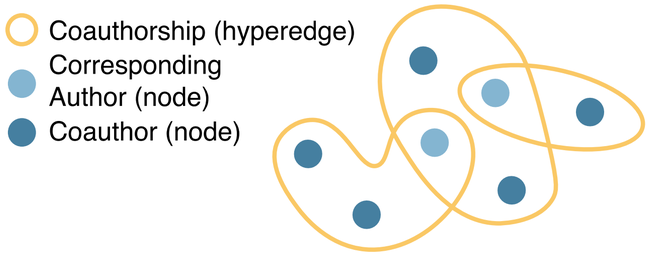
- Hypergraphs are used to represent higher-order interactions
  - Example: events (human, location, activity)

Example of a graph



- A **hypergraph** $H = (V, E)$
  - V / E : the set of nodes / **hyperedges**
  - $e \in E$: a hyperedge connects two or more nodes
  - $\forall e \in E,\ |e| = k$, H is a k-uniform hypergraph

Example of a hypergraph

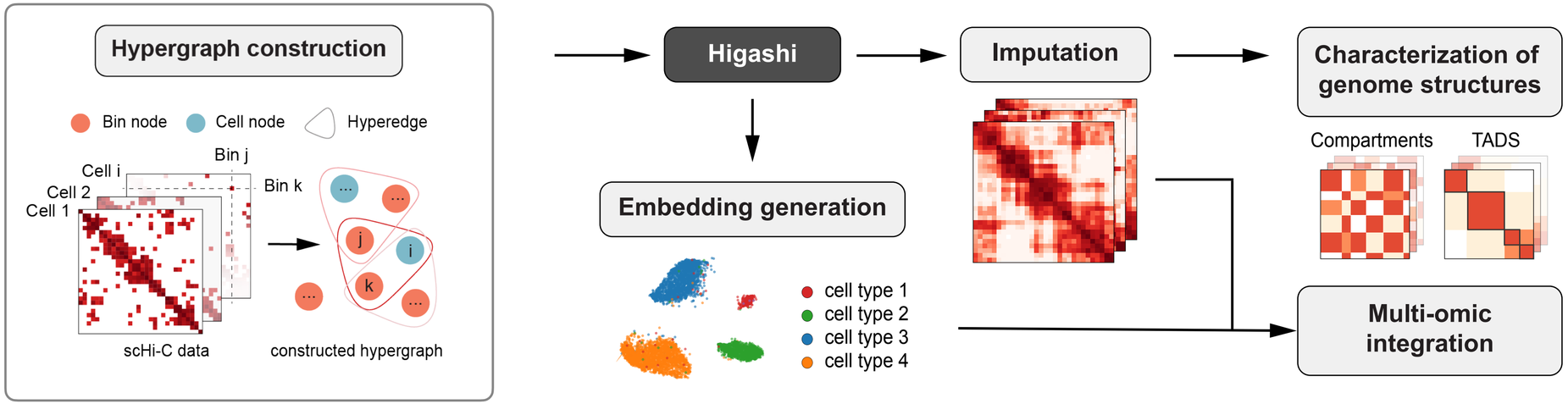


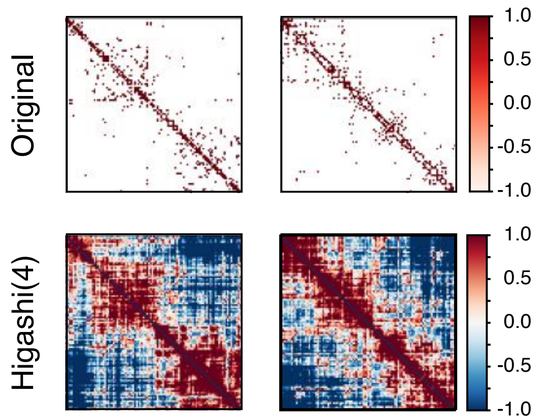
Hyper-SAGNN — Hypergraph representation learning

- Input: a hypergraph with features for each node (non-uniform heterogeneous hypergraph)
- We aim to
  - Learn embeddings for the nodes in the hypergraph
  - Learn to predict the existence of hyperedges given the node embeddings

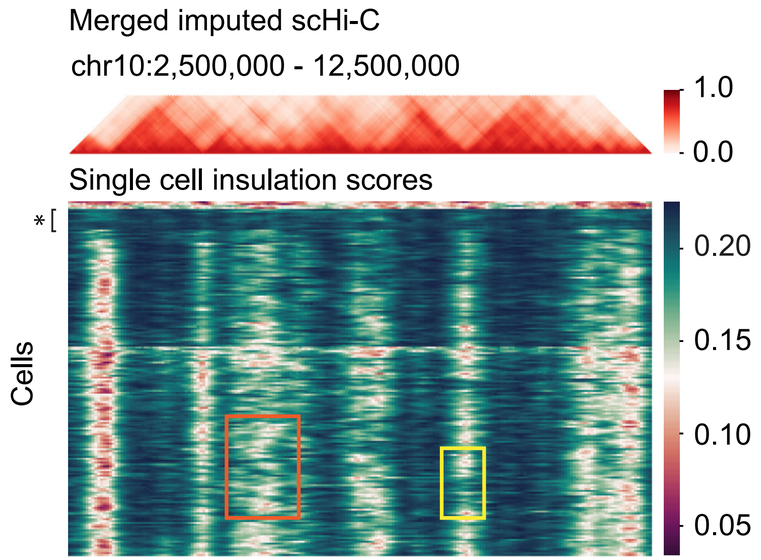
Zhang et al. ***ICLR*** 2020

Higashi — modeling scHi-C data as a hypergraph





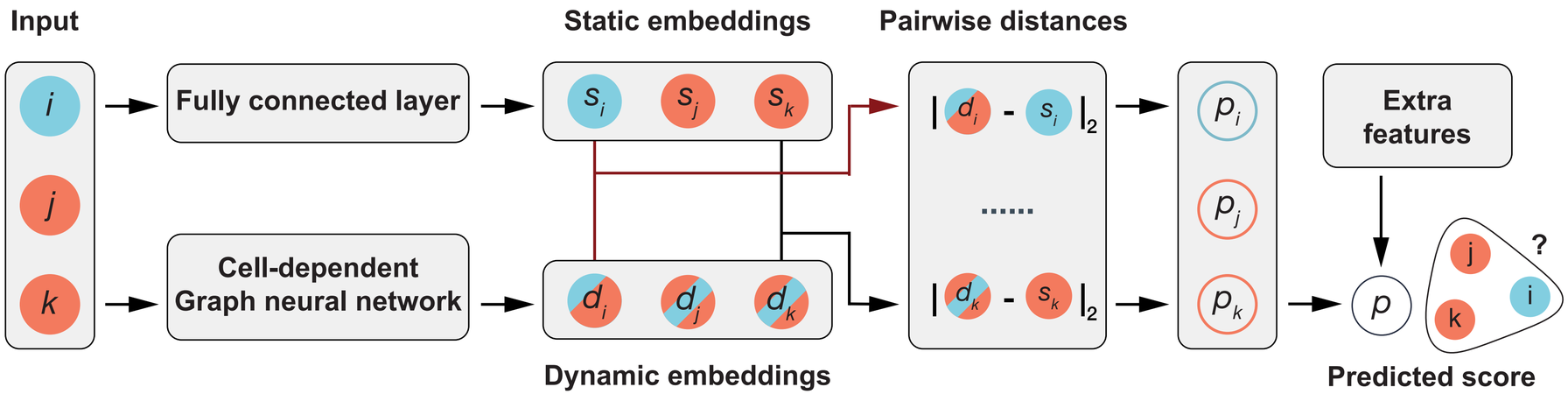
WTC-11 cells



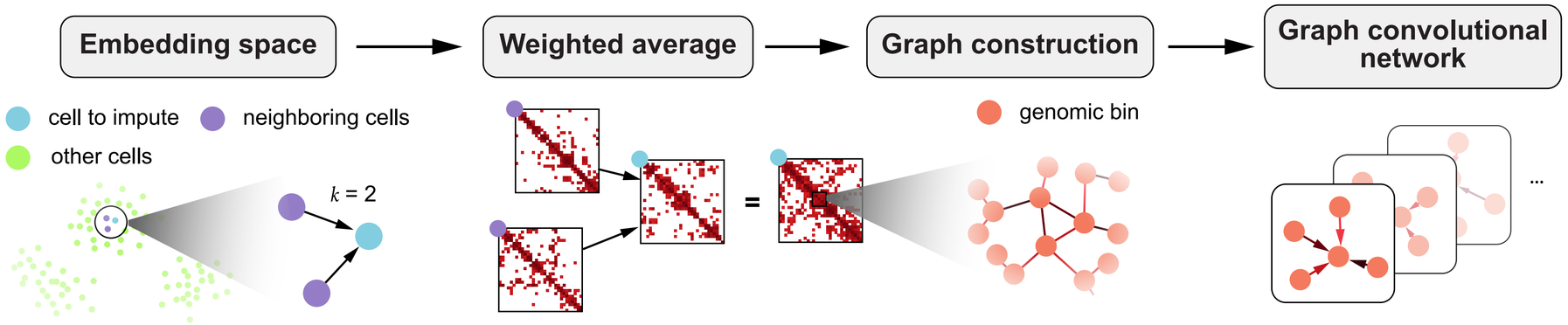
Higashi — Zhang, Zhou, and Ma. ***Nature Biotechnology,*** 2022
(also Fast Higashi — Zhang et al. *Cell Systems* 2023)

Higashi — modeling scHi-C data as a hypergraph

Network structure



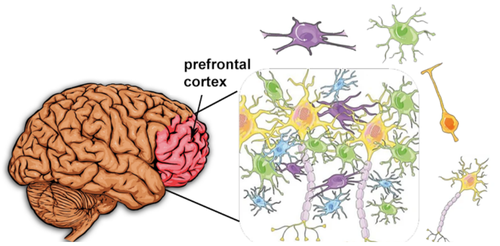
Cell-dependent GNN

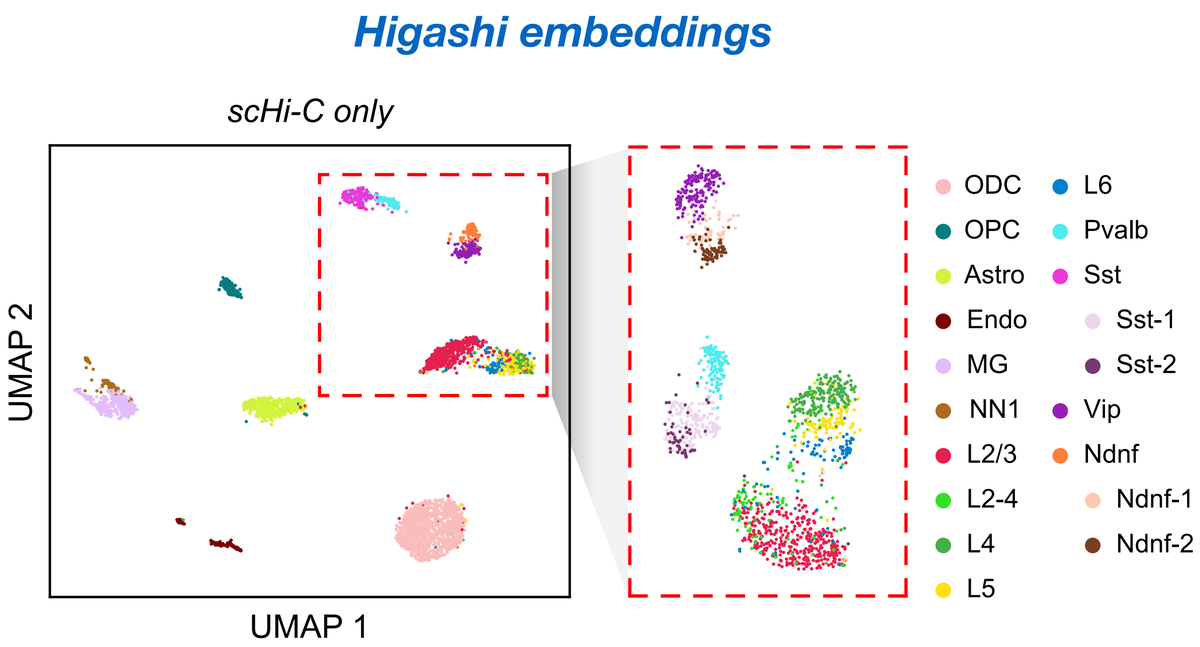


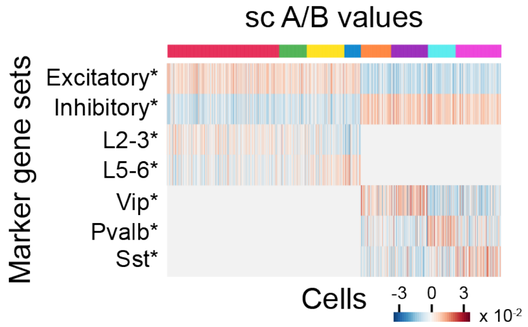
Zhang et al. *Nat Biotechnol,* 2022
Zhang et al. *ICLR* 2020

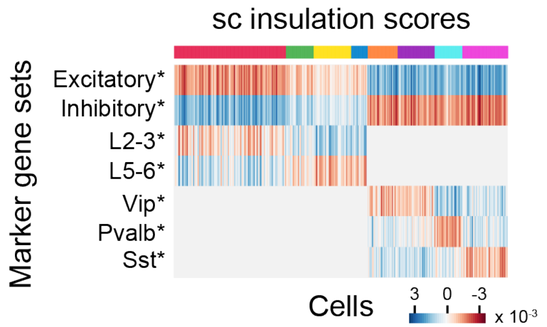
Higashi separates complex cell types in human prefrontal cortex

- Higashi embeddings separate neuron subtypes using only the scHi-C part of sn-m3c-seq (data from Lee et al. *Nat Methods* 2019)
- Cell type-specific 3D chromatin structures near marker genes







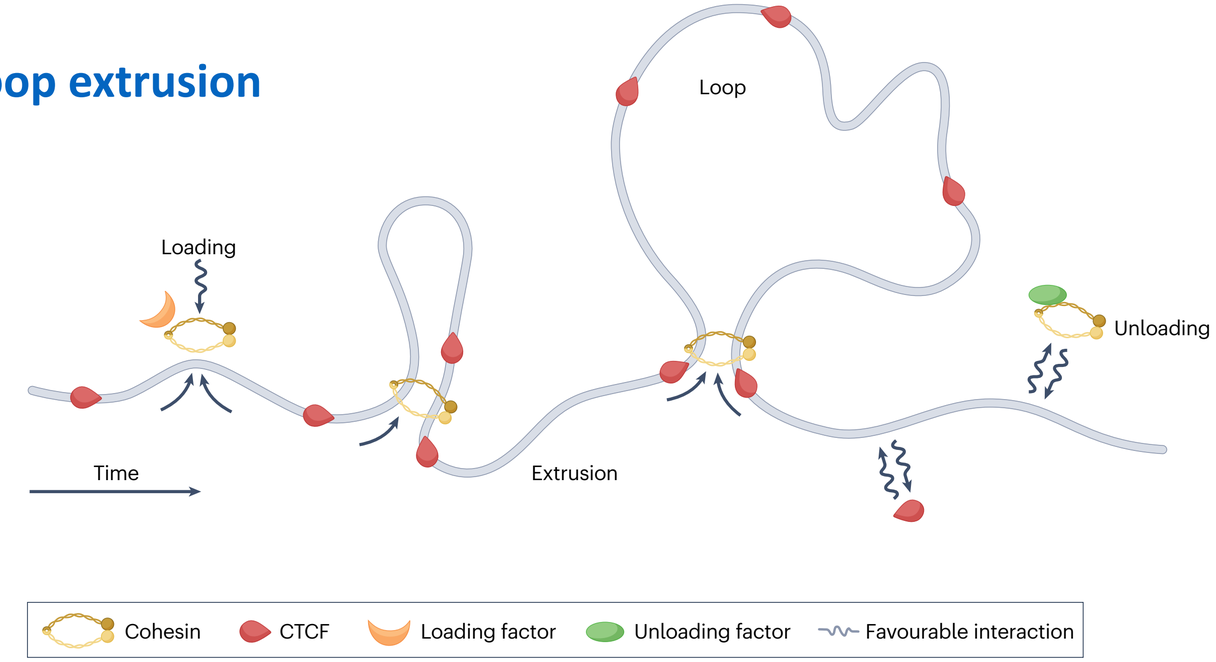


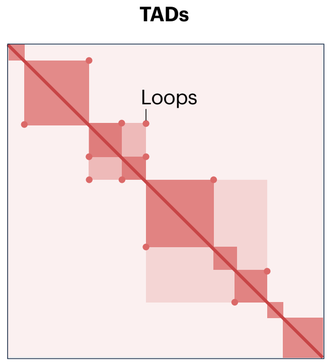
* stands for using other neurons as background

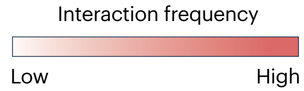
Higashi — Zhang, Zhou, and Ma. ***Nature Biotechnology,*** 2022
(also Fast Higashi — Zhang et al. *Cell Systems* 2023)

Mechanisms of genome folding

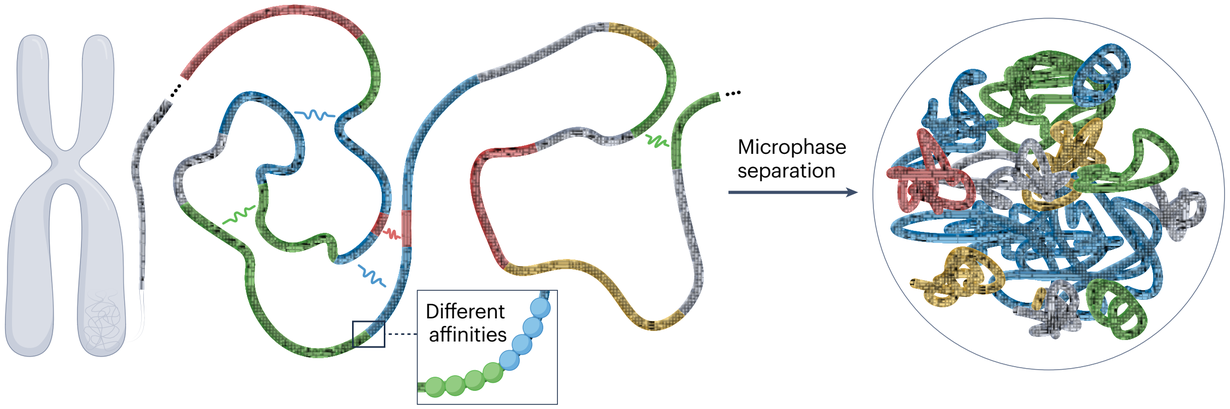
Loop extrusion

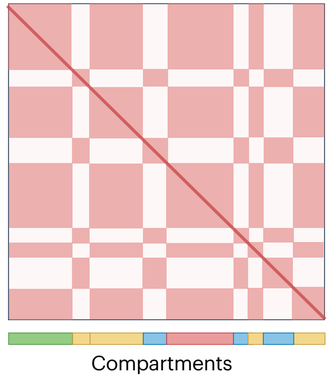






Phase separation

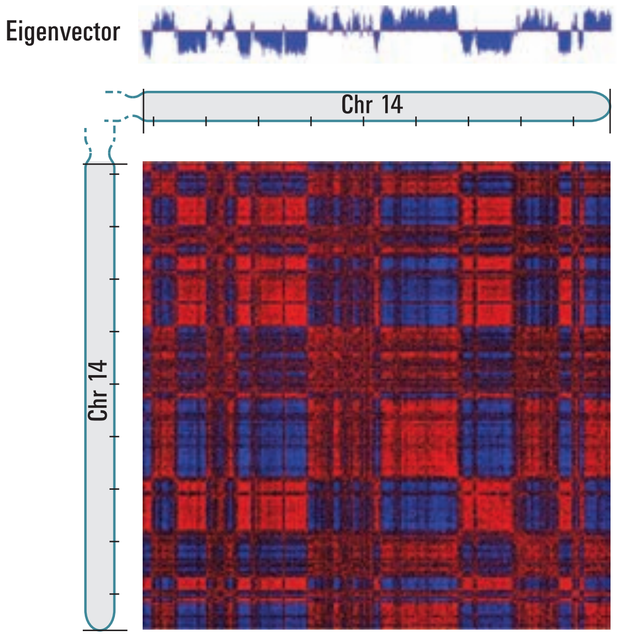




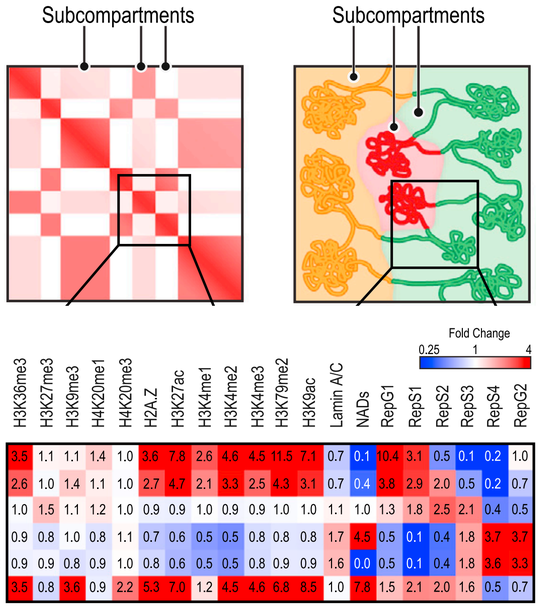
Zhang et al.
Nat Rev Genetics 2024

Patterns of chromosome spatial segregation based on Hi-C data

- Chromosomes are segregated into A and B compartments
- High-coverage Hi-C in GM12878 identified **subcompartments** by clustering inter-chromosomal contact maps

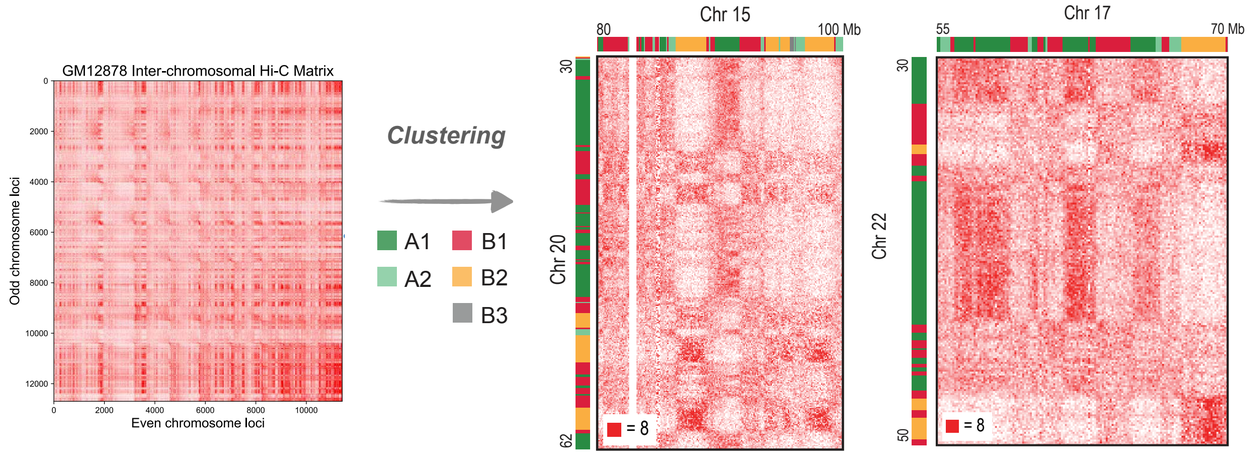


Lieberman-Aiden et al. *Science* 2009



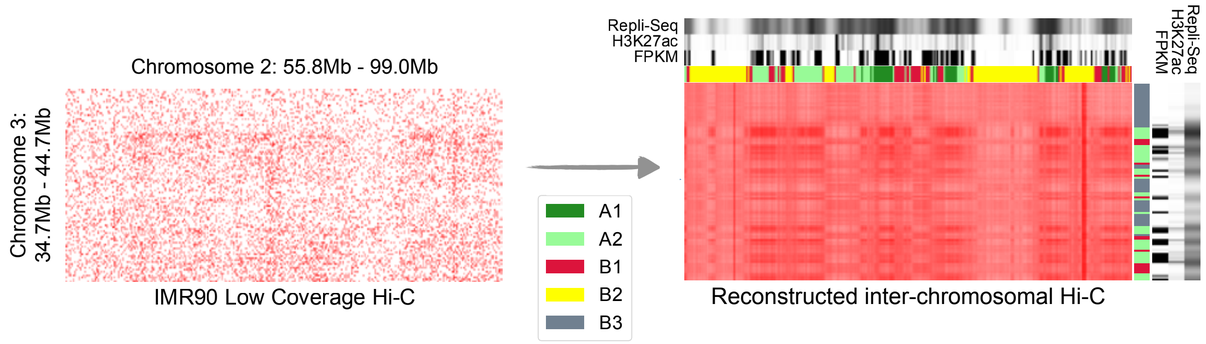
Rao et al. *Cell* 2014

Clustering using high-coverage GM12878 data



Rao et al. *Cell* 2014

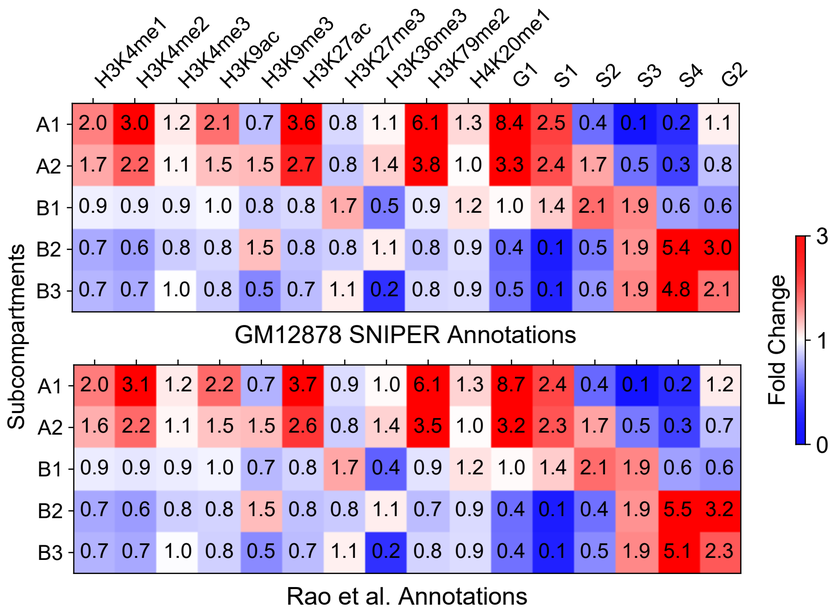
Addressing sparsity of inter-chromosome interactions

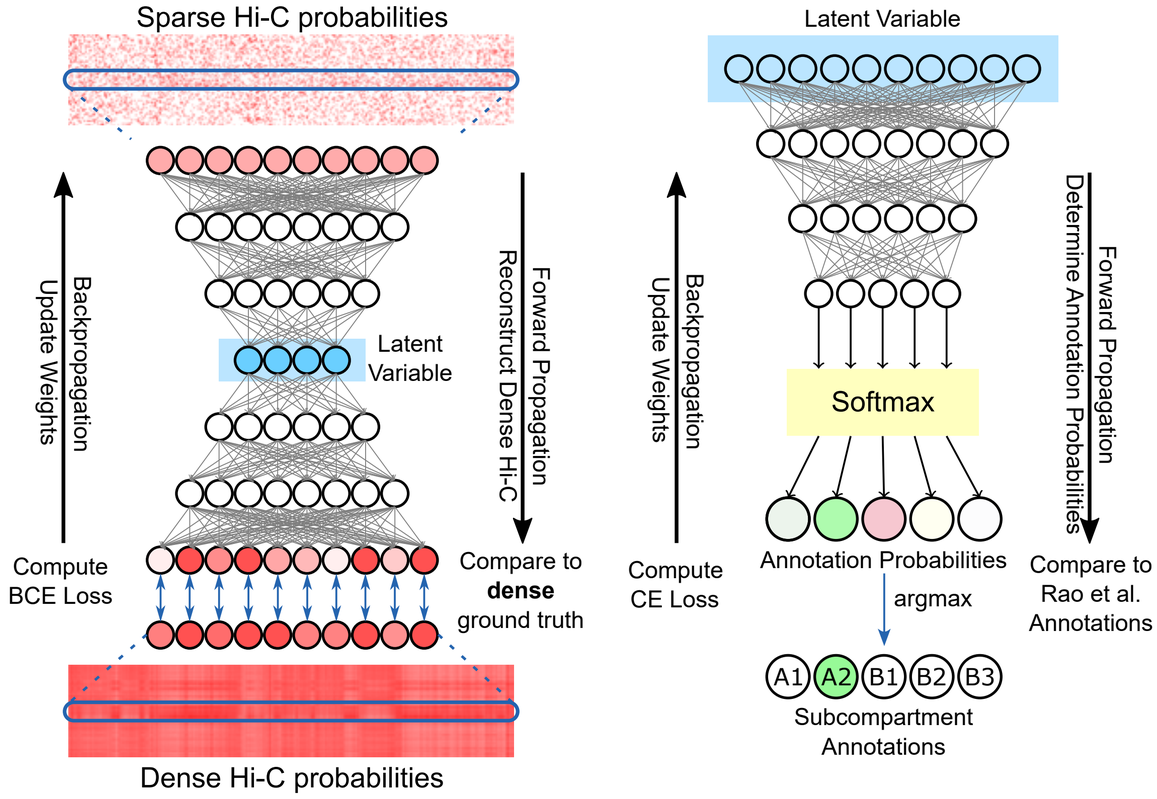


SNIPER — Xiong and Ma. ***Nat Commun*** 2019

SNIPER — inferring Hi-C subcompartments

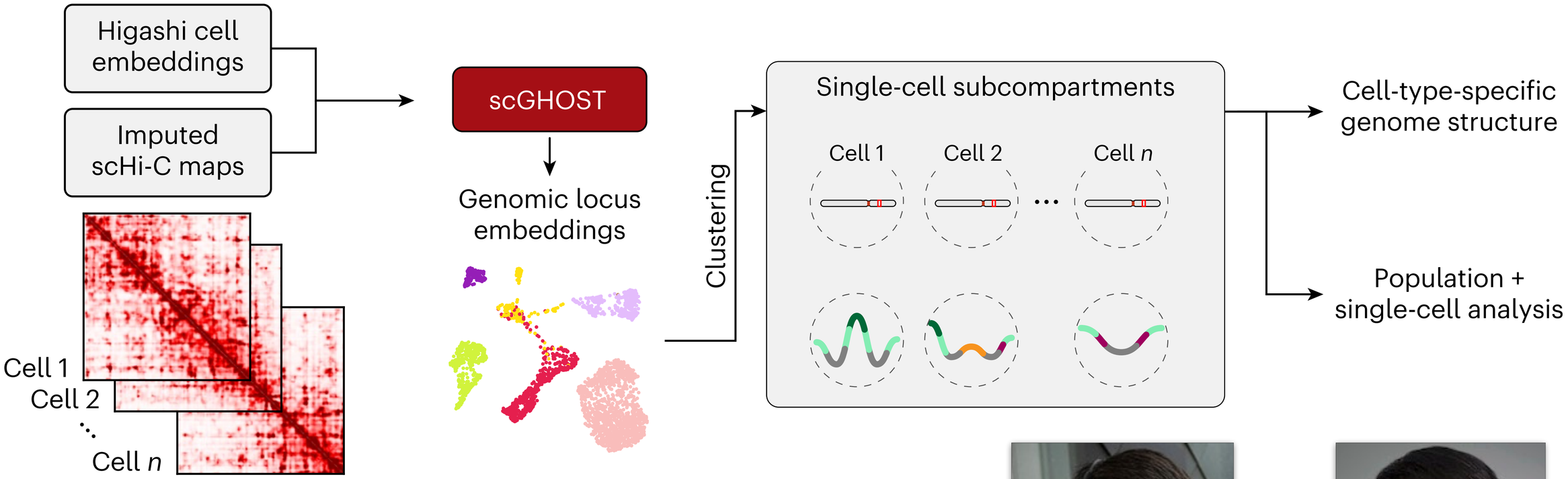
- Autoencoder compresses sparse inter-chromosomal Hi-C contacts into embeddings and imputes missing contacts
- A classifier then uses embedded inter-chromosomal Hi-C data to predict subcompartments





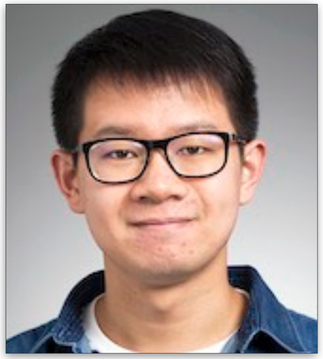
SNIPER — Xiong and Ma. ***Nat Commun*** 2019

scGHOST: Identifying single-cell 3D genome subcompartments



- scHi-C contact maps as **graphs**
- Cell embeddings define k-NN cells. **Genomic locus embeddings** lead to subcompartments
- A unique **random sampling** procedure that filters noise in imputed scHi-C data

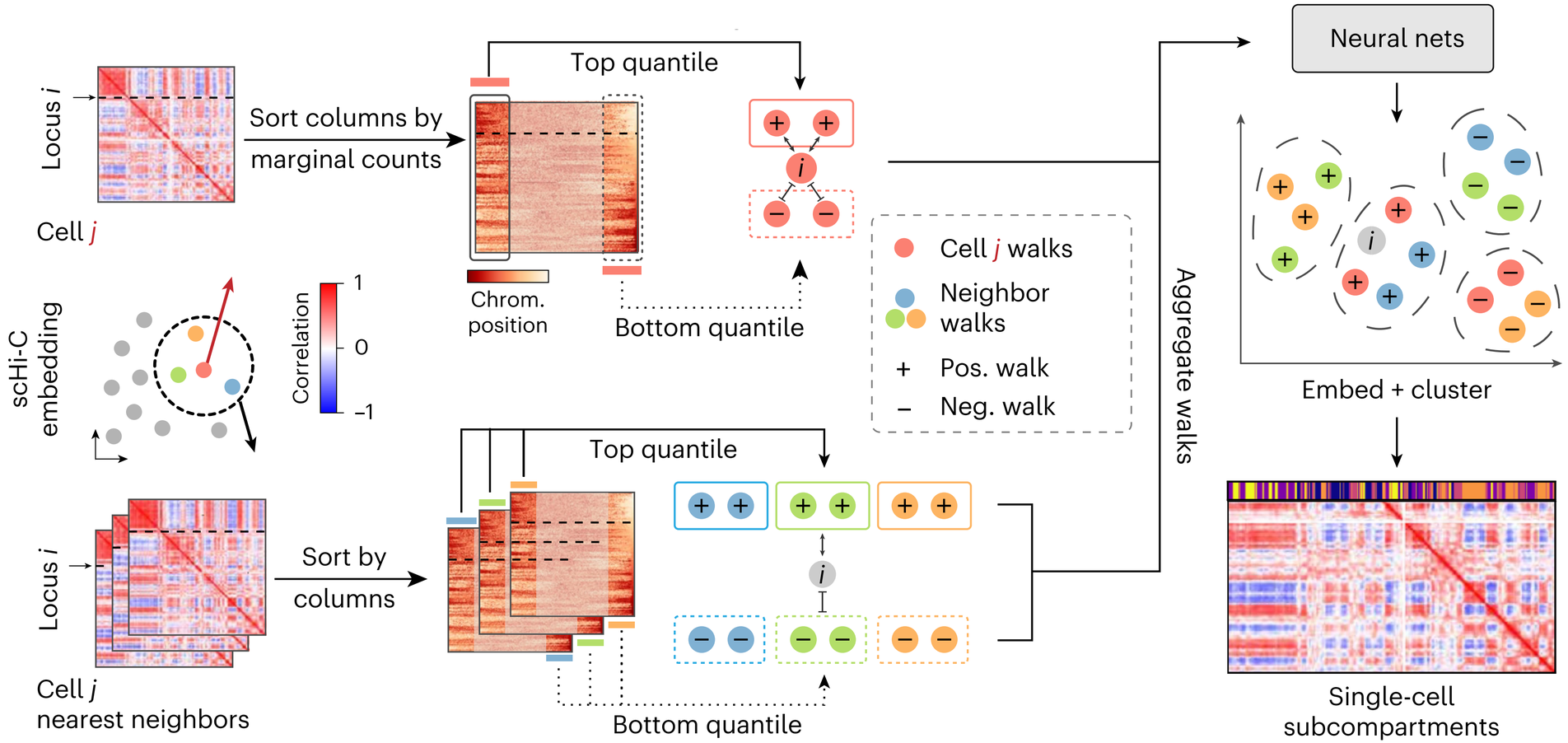
Kyle Xiong



Ruochi Zhang

Xiong #, Zhang #, and Ma. ***Nature Methods,*** 2024

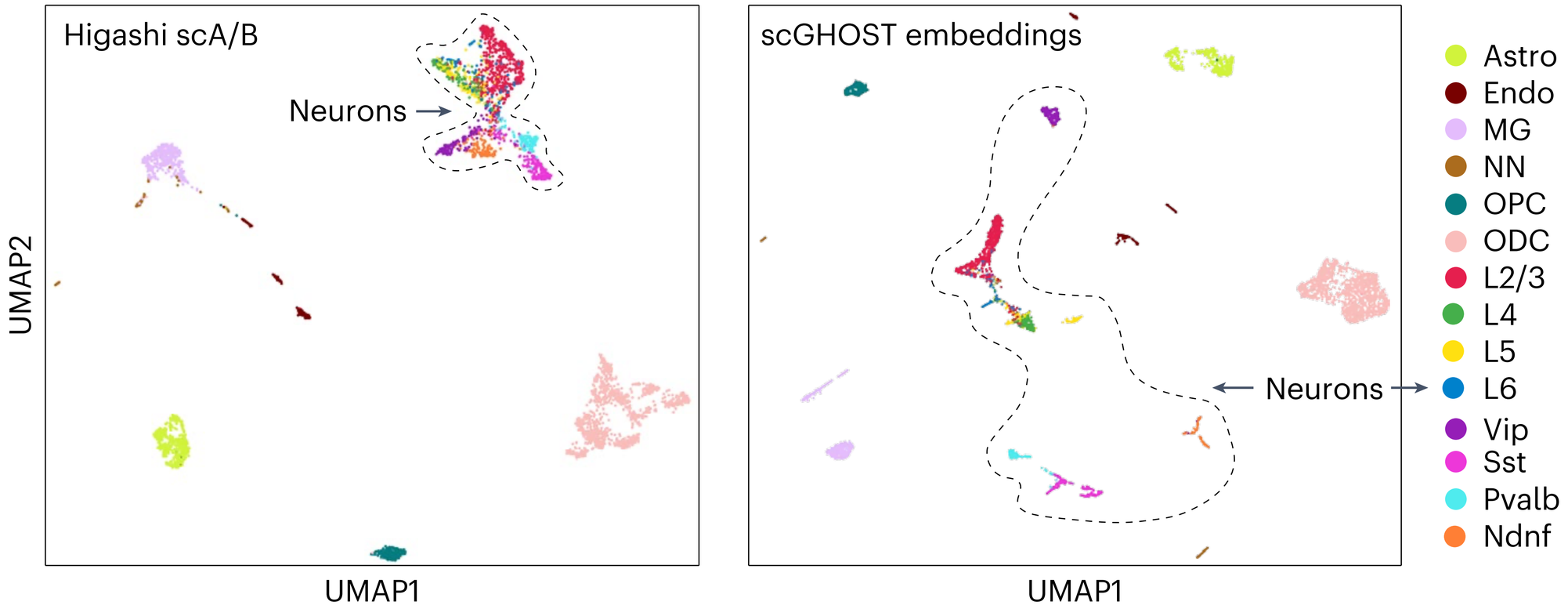
scGHOST: Identifying single-cell 3D genome subcompartments



- Sampling based on both first-order random walks and second-order random walks
- Graph node embedding using NN

Xiong #, Zhang #, and Ma. ***Nature Methods,*** 2024

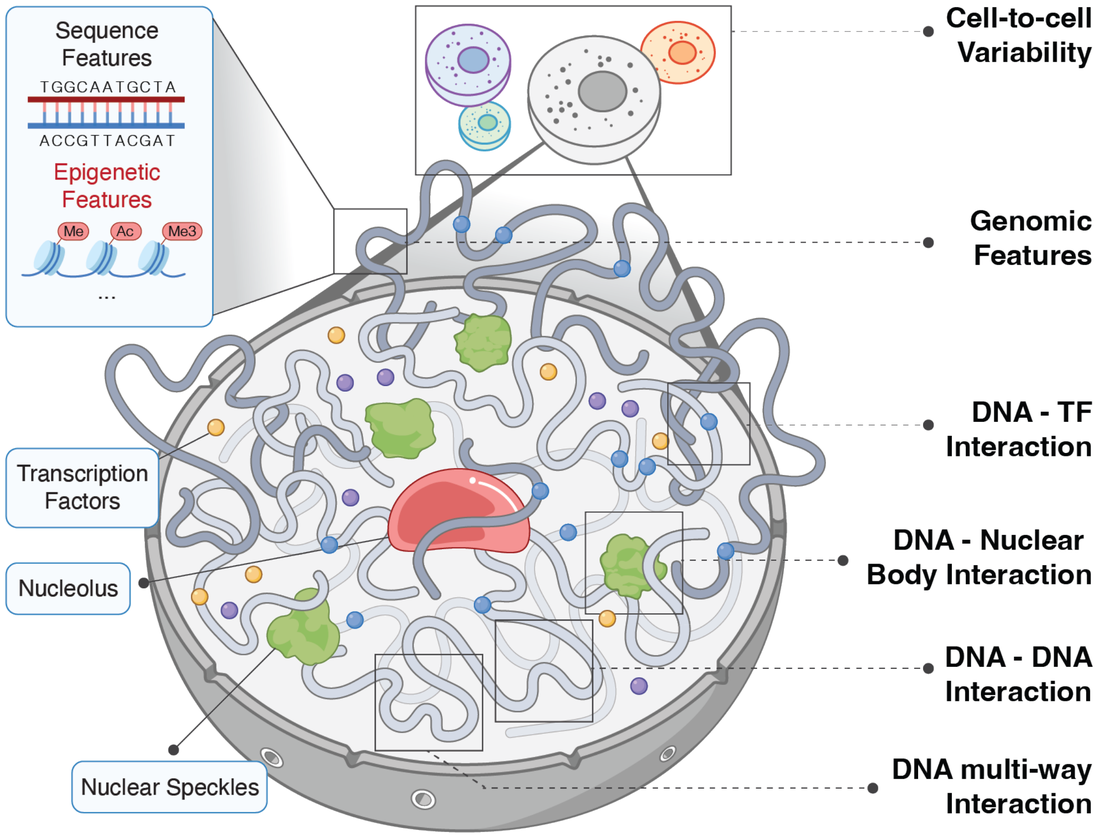
Single-cell subcompartments of human prefrontal cortex

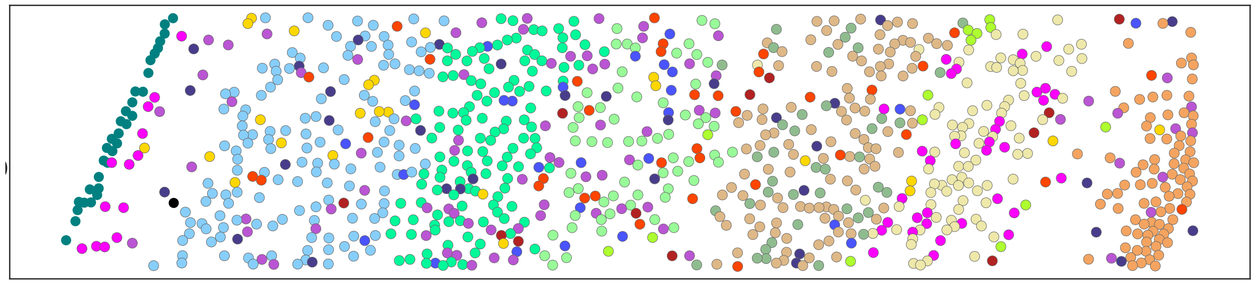


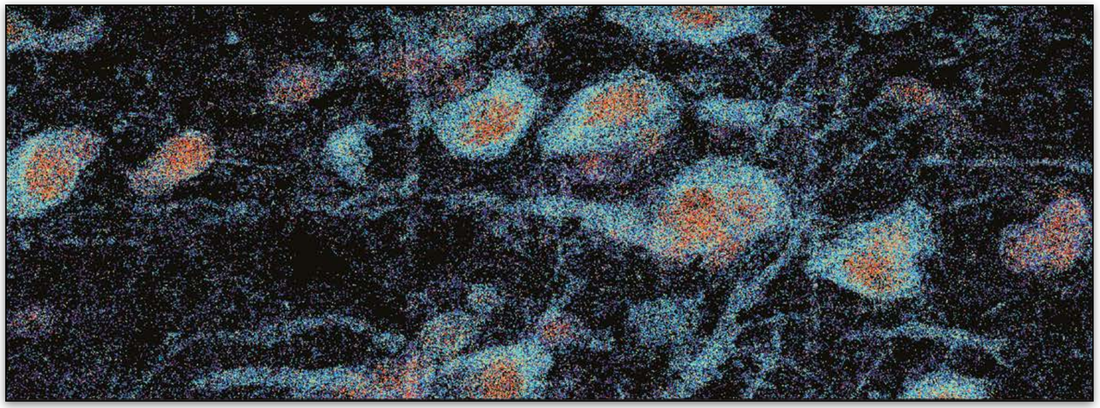
Data from: Lee et al. *Nat Methods* 2019

Xiong #, Zhang #, and Ma. ***Nature Methods,*** 2024

Spatial organization of genomes and cells



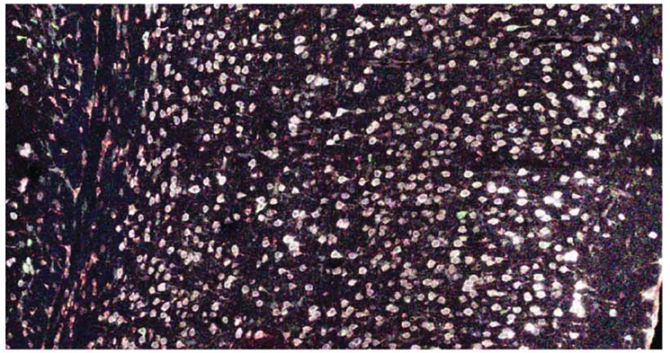




Spatial transcriptomics technologies reveal where in a tissue each gene is expressed

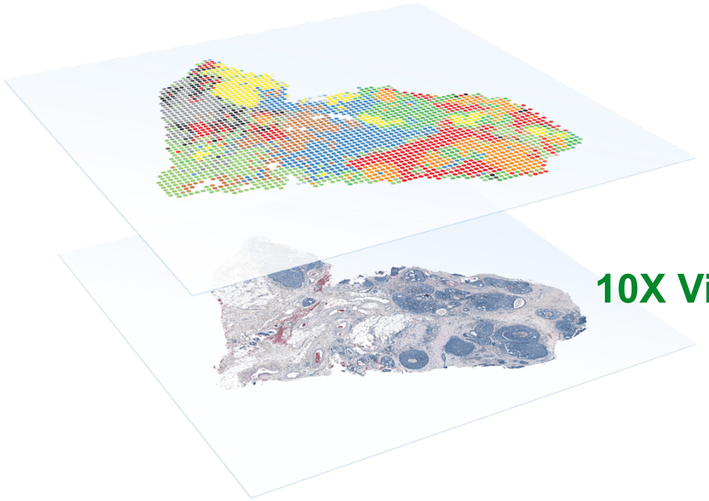
Spatial transcriptomic technologies

- **Imaging-based** — single-cell resolution
  - **STARmap / STARmap PLUS**
    Wang et al. *Science* 2018 / Zeng et al. *Nat Neurosci* 2023
  - **seqFISH+**
    Eng et al. *Nature* 2019
  - **MERFISH / Vizgen MERSCOPE**
    Moffitt et al. *Science* 2018



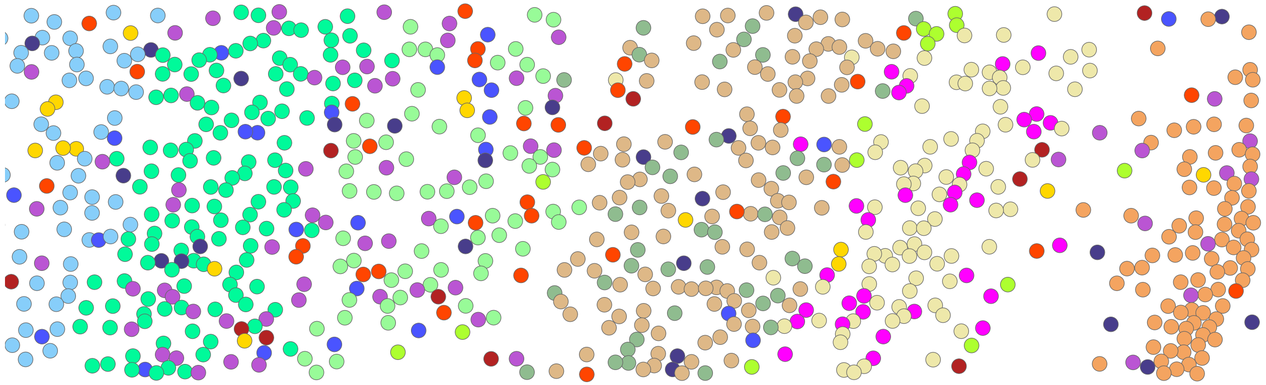
STARmap

- **Sequencing-based** — full transcriptome
  - **Visium**
    Stahl et al. *Science* 2016
  - **Slide-seq / Slide-seq V2**
    Rodriques et al. *Science* 2019 / Stickels et al. *Nat Biotech* 2021
  - **Stereo-seq**
    Chen et al. *Cell* 2022



10X Visium

- **Challenge:** Lack of computational methods that integrate both gene expression and spatial factors to model cell identity



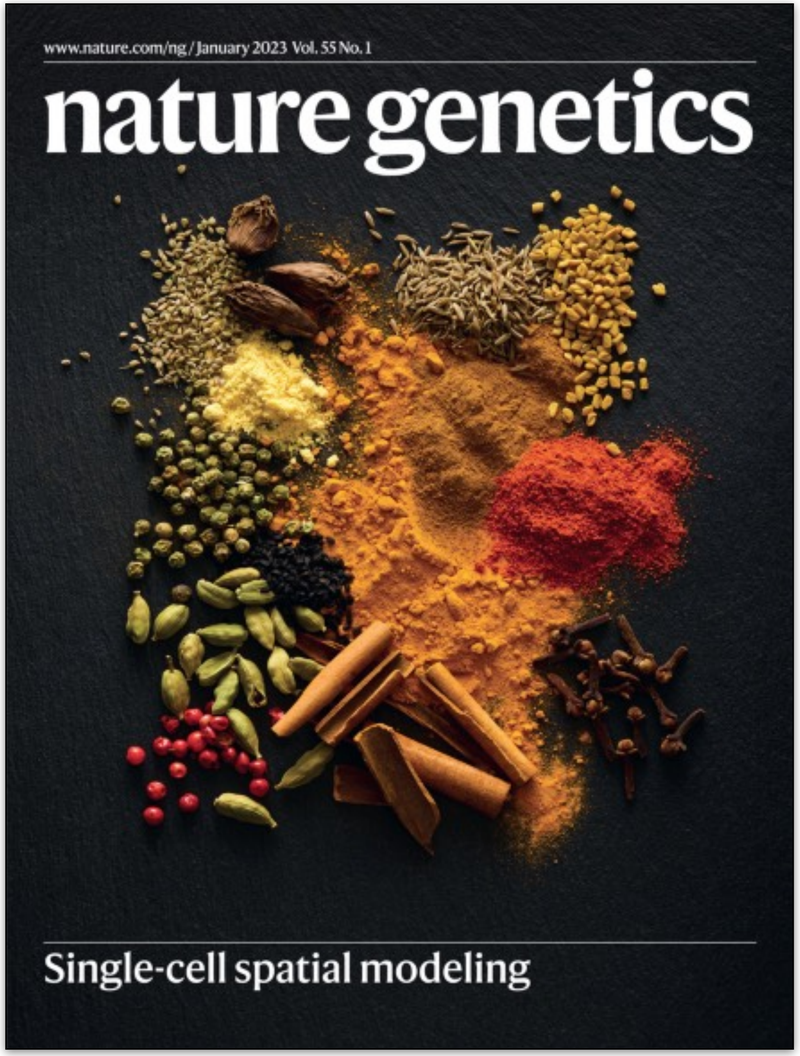
SPICEMIX enables integrative single-cell spatial modeling of cell identity

Chidester B#, Zhou T#, Alam S, and Ma J. ***Nature Genetics,*** 2023

Ben Chidester

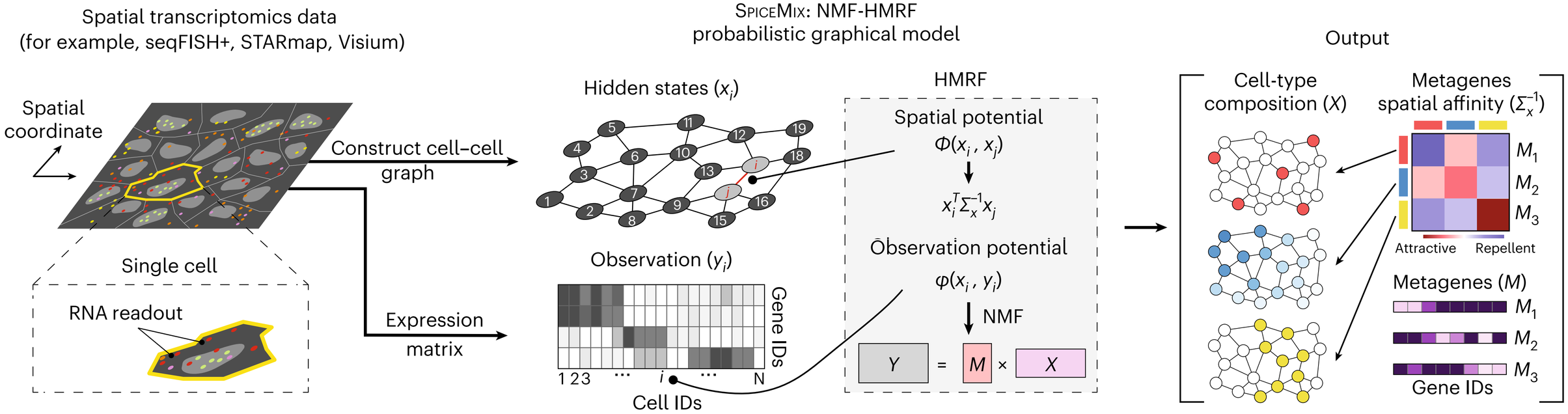
Tianming Zhou

Shahul Alam



January 2023 issue

SPICEMIX — NMF + HMRF



$$y_i = M x_i$$

latent embedding/representation

metagene = spatially variable features

latent states

regularization

$$P(Y, X|\Theta) = \frac{1}{Z(\Theta)} \prod_{(i,j)\in\mathcal{E}} \varphi(x_i, x_j) \prod_{i\in\mathcal{V}} \phi(y_i, x_i)\pi(x_i)$$

observed expression

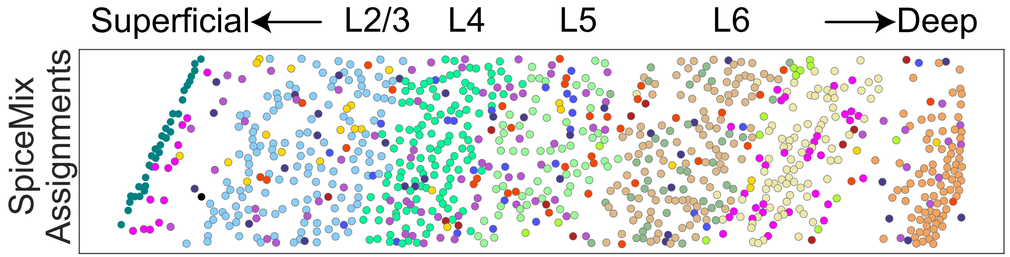
spatial dependency

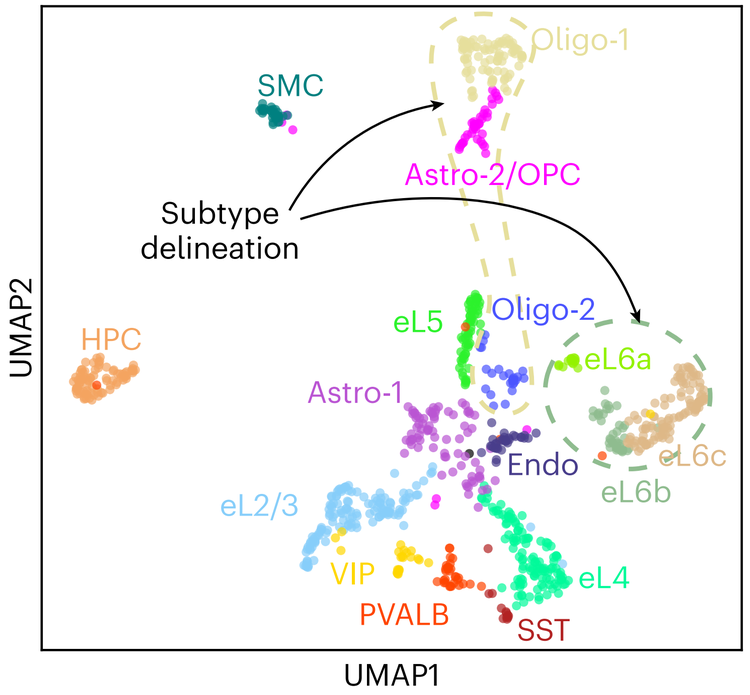
matrix factorization

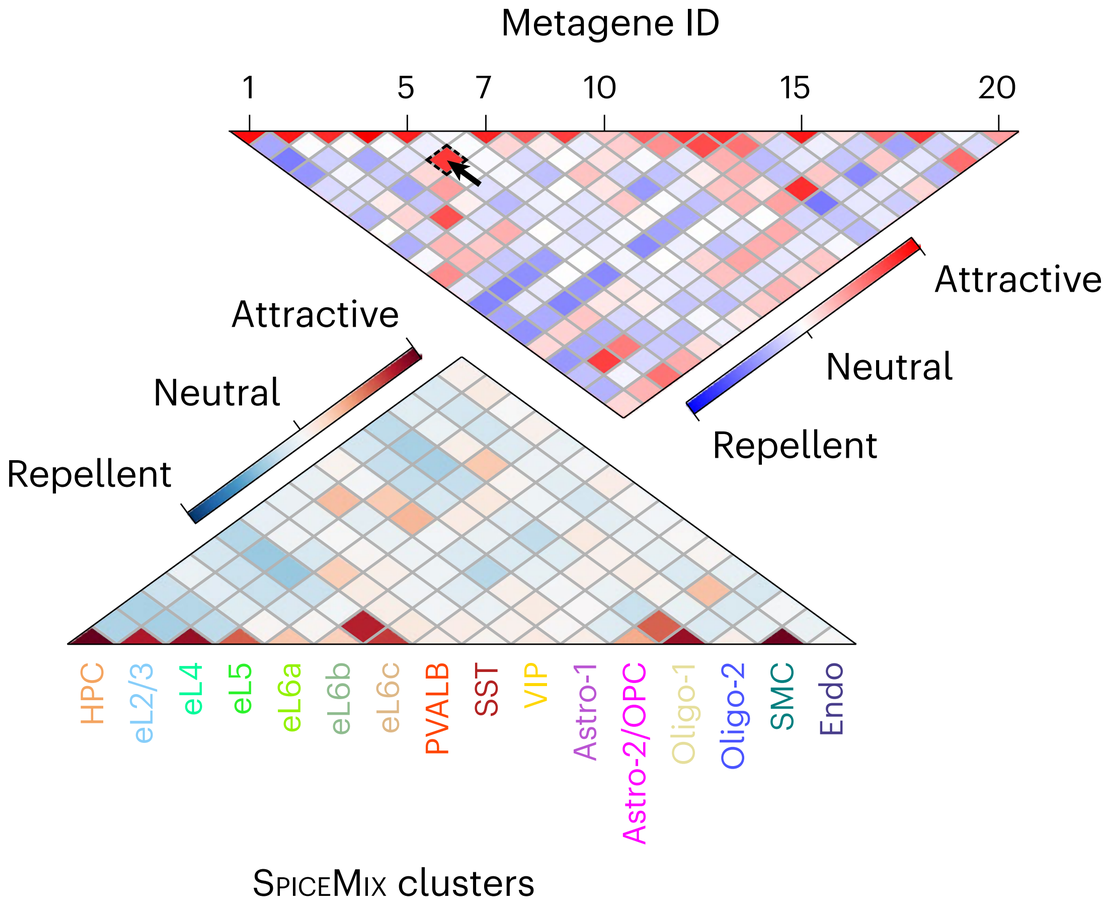
Chidester #, Zhou #, Alam, and Ma. ***Nature Genetics,*** 2023

Application to the STARmap dataset

- Mouse visual cortex — Data from Wang et al. *Science* 2018
- SpiceMix infers rare subtypes and spatially variable metagenes

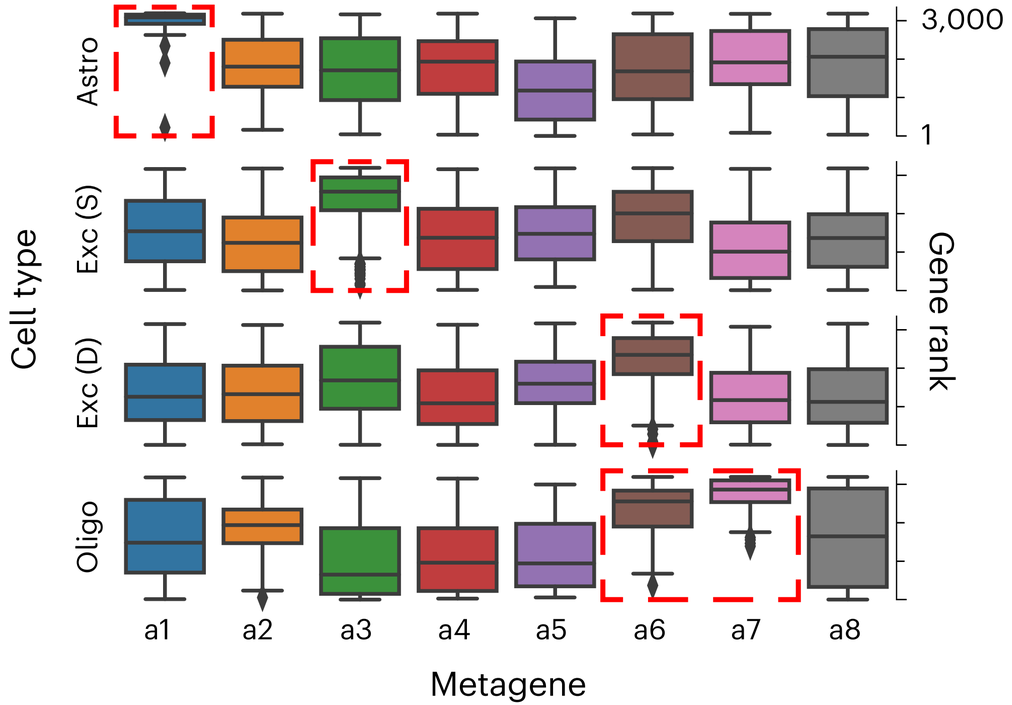


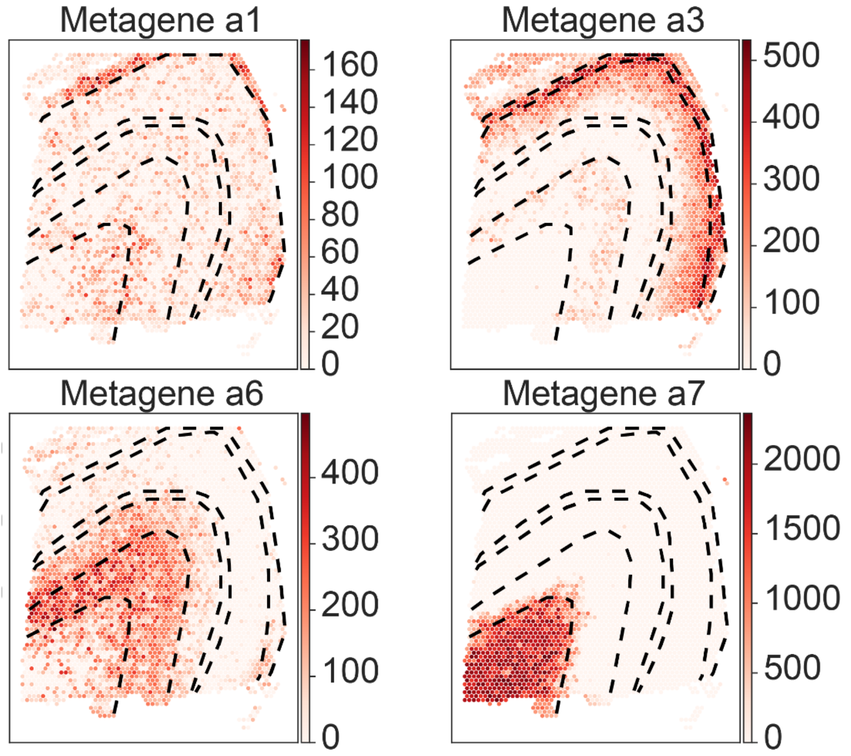




SPICEMIX disentangles cell type composition

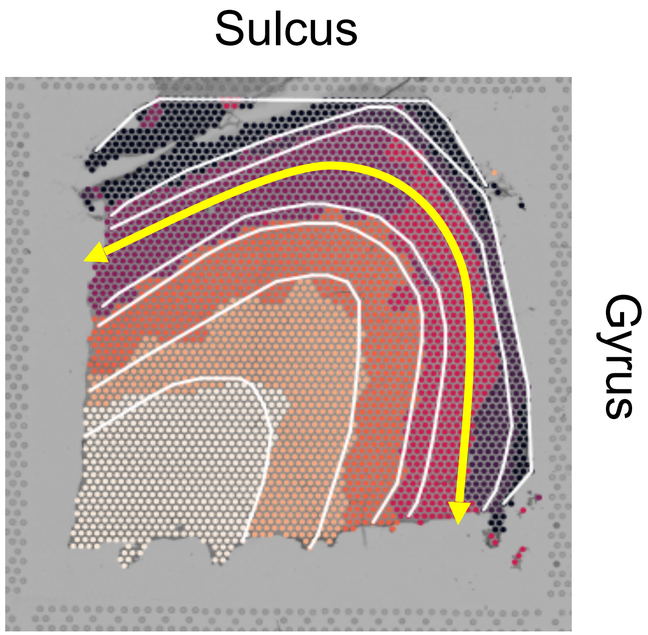
- Human dorsolateral prefrontal cortex (DLPFC)
  - Data from Maynard et al. *Nature Neuroscience*, 2021
- The correspondence between metagenes and layers is not one-to-one
- SPICEMIX captures the continuous gradient along the layer-axis

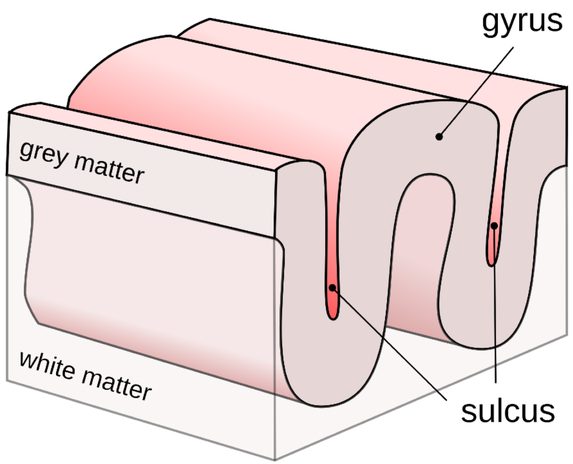




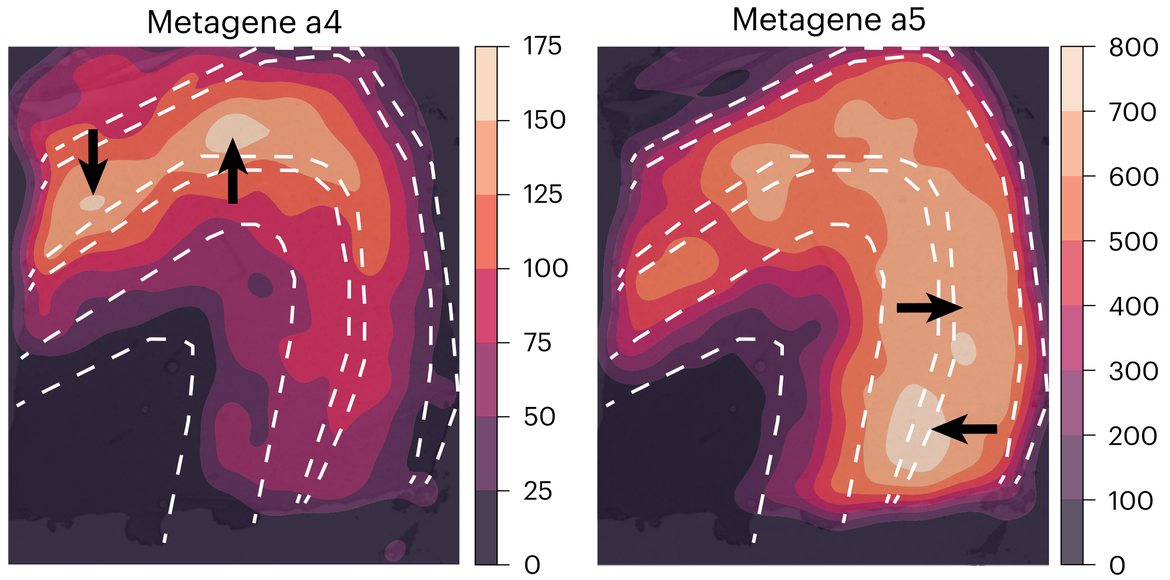
Chidester #, Zhou #, Alam, and Ma. ***Nature Genetics,*** 2023

SPICEMIX identifies the gyro-sulcal gradient





- The sulcal side and gyric side are anatomically different
- Metagenes a4 and a5 identified the gradient along the gyro-sulcal axis, supported by differentially expressed genes

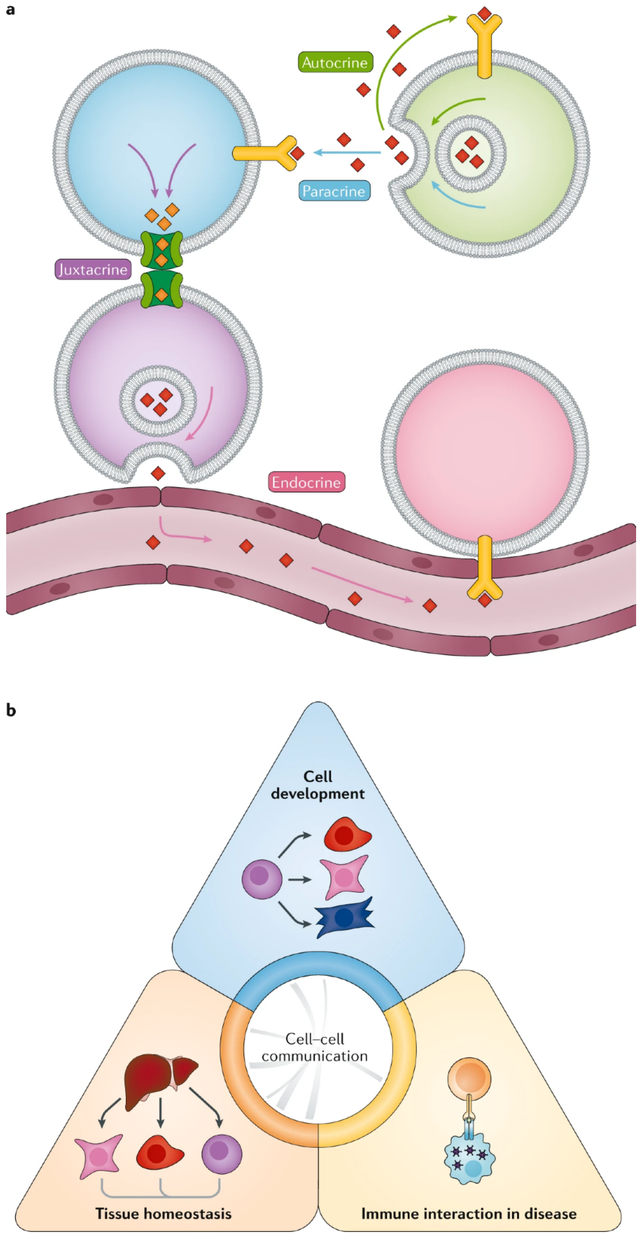


Limitations of analyzing cellular interactions

- We need a method to quantify **cell-cell interactions** and address the following:
  - Cell-cell interactions are specific to cell type/state
  - Cell-cell interactions are specific to spatial domains/context
  - Allowing for *in silico* spatial perturbation — predicting the effect of a changing environment on a cell

- Cell-cell interactions are inherently **multi-scale**

- Current cell-cell interactions databases are incomplete
  - Need ***de novo*** approaches to learn from the data

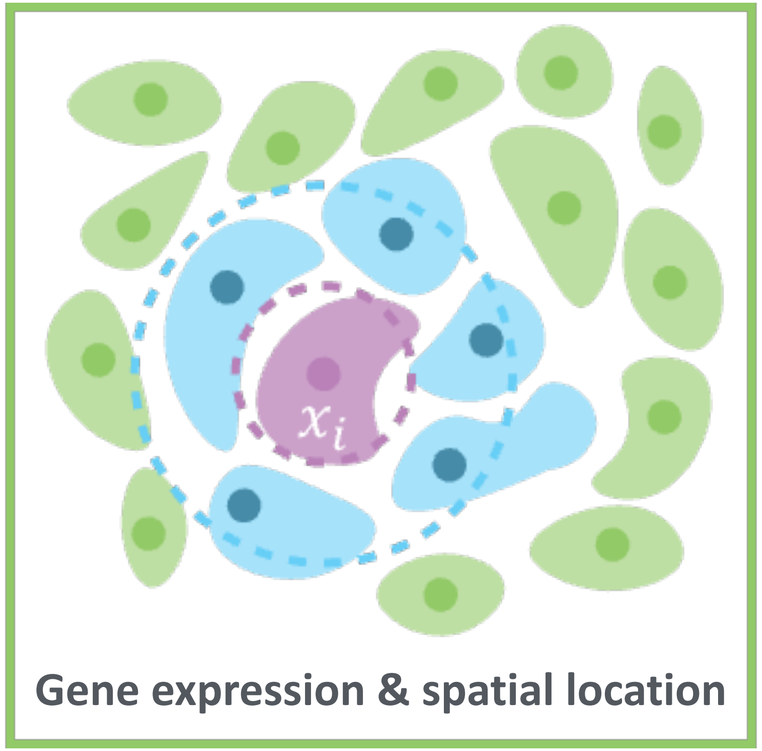


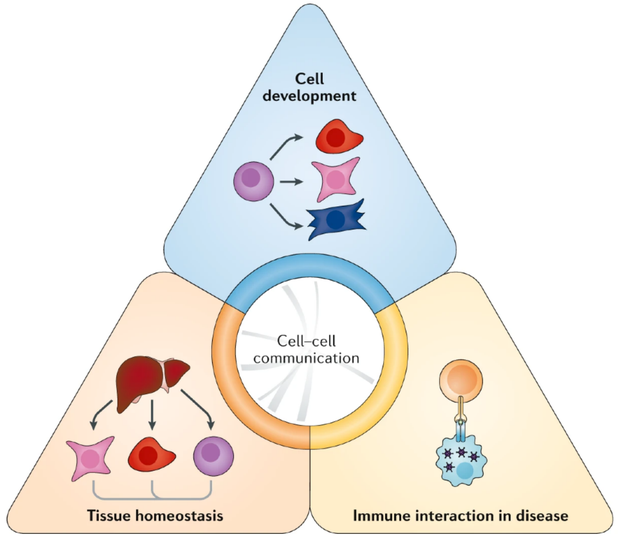
Armingol et al. ***Nat Rev Genet*** 2021

STEAMBOAT: modeling cell-cell interactions

- The molecular profile of cells is a result of superimposing:
  - **intrinsic factors**
  - **interactions at multiple scales**

- **How do we decompose them and model such multiscale interactions?**





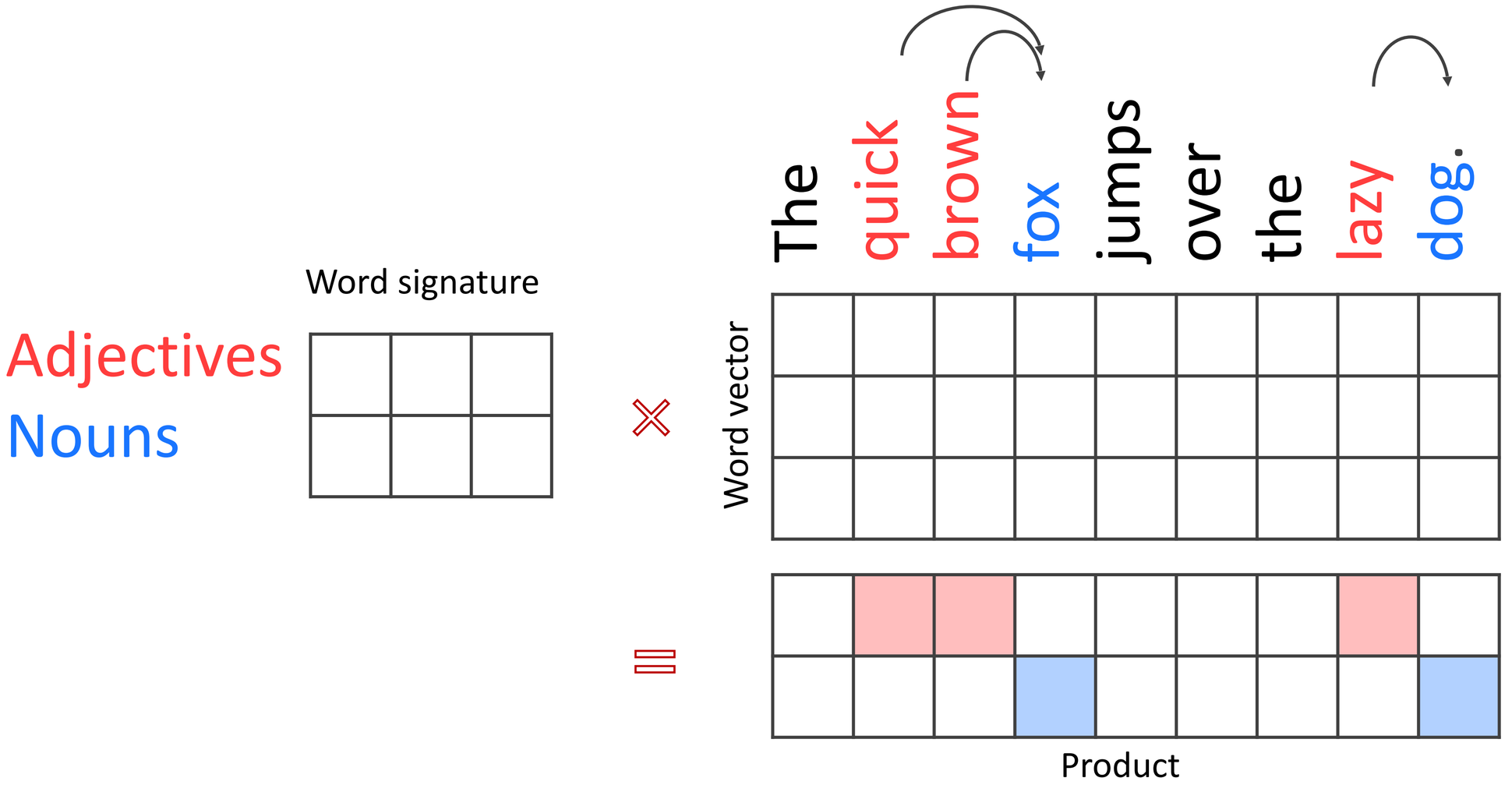
Armingol et al. *Nat Rev Genet* 2021

Attention-based multiscale delineation of cellular interactions in tissues

Shaoheng Liang

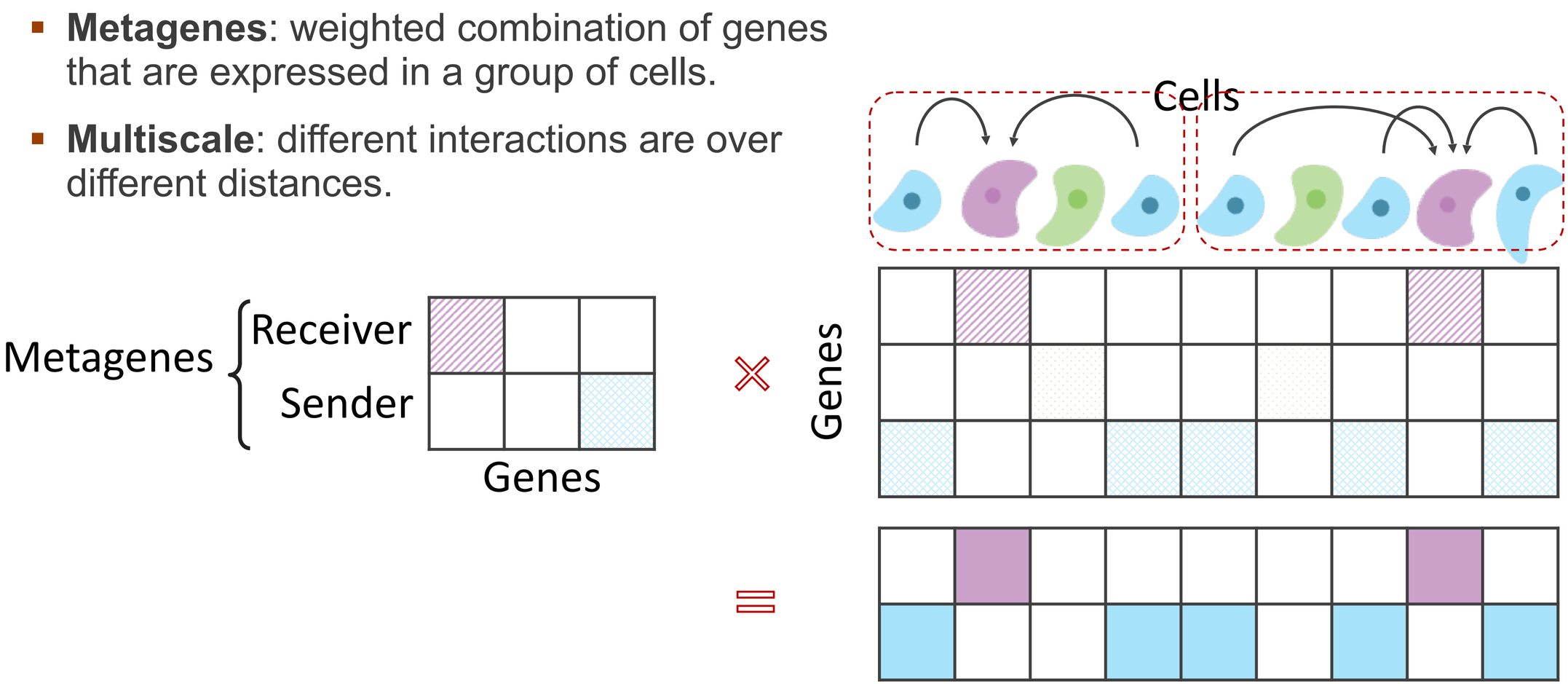
Liang et al. *bioRxiv* 2025

Find words that interact by attention



Cells in a tissue as words in a sentence

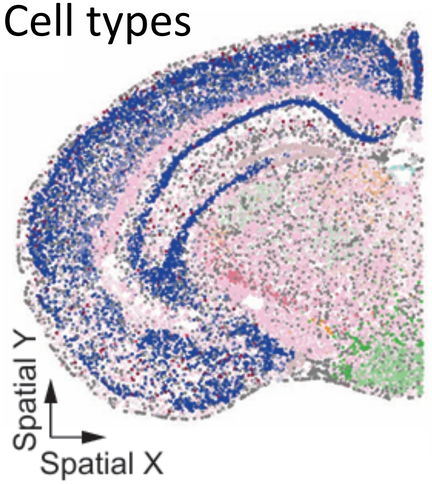
- **Metagenes**: weighted combination of genes that are expressed in a group of cells.
- **Multiscale**: different interactions are over different distances.



Liang et al. *bioRxiv* 2025

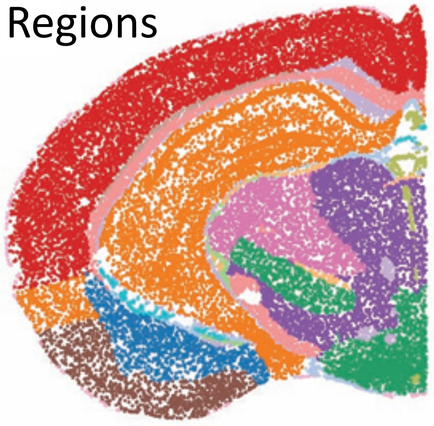
STEAMBOAT identifies underlying factors in mouse brain

Cell types



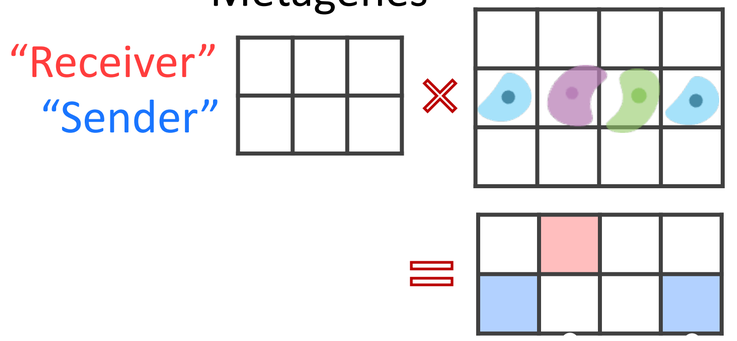
- 01 IT-ET Glut
- 02 NP-CT-L6b Glut
- 03 OB-CR Glut
- 04 DG-IMN Glut
- 05 OB-IMN GABA
- 06 CTX-CGE GABA
- 07 CTX-MGE GABA
- 08 CNU-MGE GABA
- 09 CNU-LGE GABA
- 11 CNU-HYa GABA
- 12 HY GABA
- 13 CNU-HYa Glut
- 14 HY Glut
- 16 HY MM Glut
- 18 TH Glut
- 19 MB Glut
- 20 MB GABA
- 21 MB Dopa
- 23 P Glut
- 25 Pineal Glut
- 29 CB Glut
- 30 Astro-Epen
- 31 OPC-Oligo
- 32 OEC
- 33 Vascular
- 34 Immune

Regions

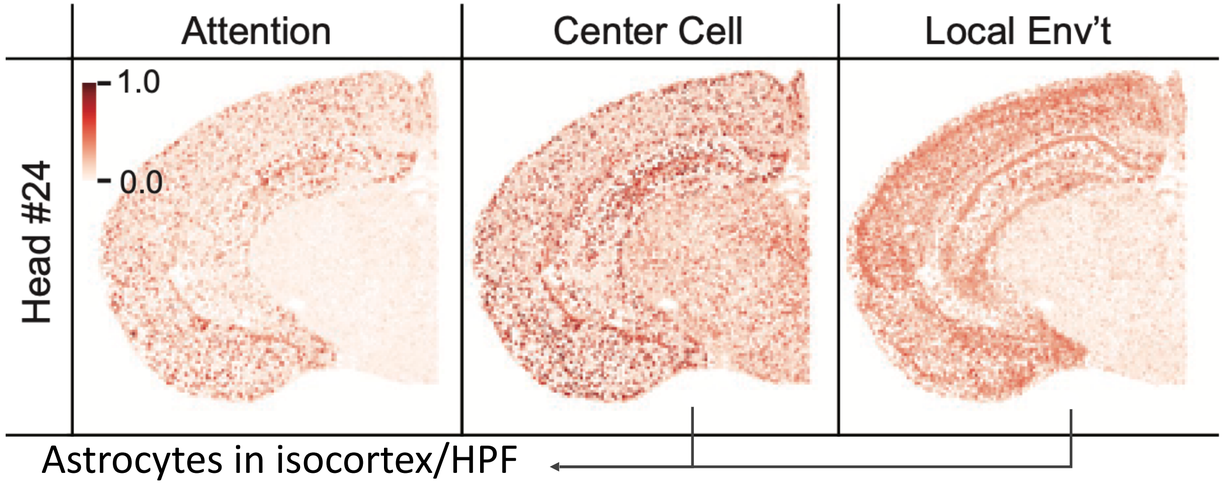


- CTXsp
- HPF
- HY
- Isocortex
- MB
- OLF
- TH
- V3

Metagenes



All astrocytes Neurons in isocortex/HPF



Mouse brain MERFISH data from: Zhang et al. ***Nature,*** 2023

STEAMBOAT unveils spatial features in colorectal cancer

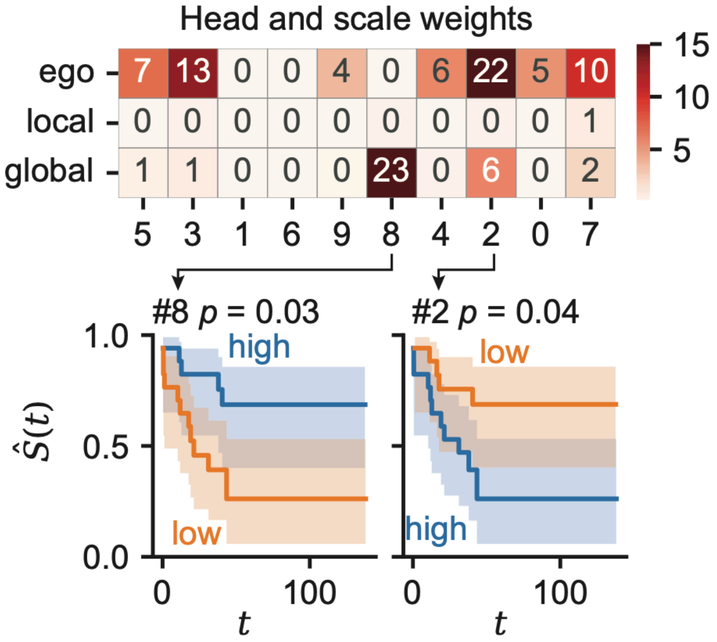
Cell

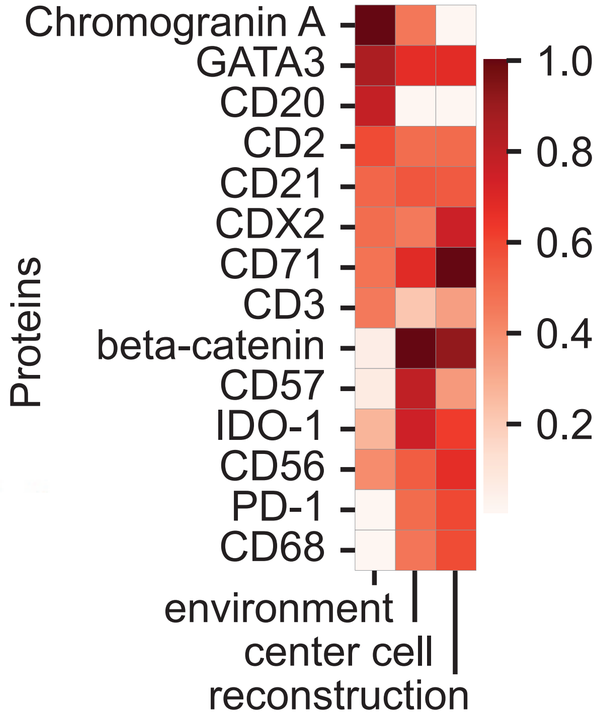
Resource

Coordinated Cellular Neighborhoods Orchestrate Antitumoral Immunity at the Colorectal Cancer Invasive Front

Christian M. Schürch,[1,2,6,*] Salil S. Bhate,[1,2,3,6] Graham L. Barlow,[1,2,6] Darci J. Phillips,[1,2,4,6] Luca Noti,[5] Inti Zlobec,[5] Pauline Chu,[1,2] Sarah Black,[1,2] Janos Demeter,[1] David R. McIlwain,[1,2] Shigemi Kinoshita,[1] Nikolay Samusik,[1] Yury Goltsev,[1,2] and Garry P. Nolan[1,2,7,*]

Weight of three scales for all attention heads

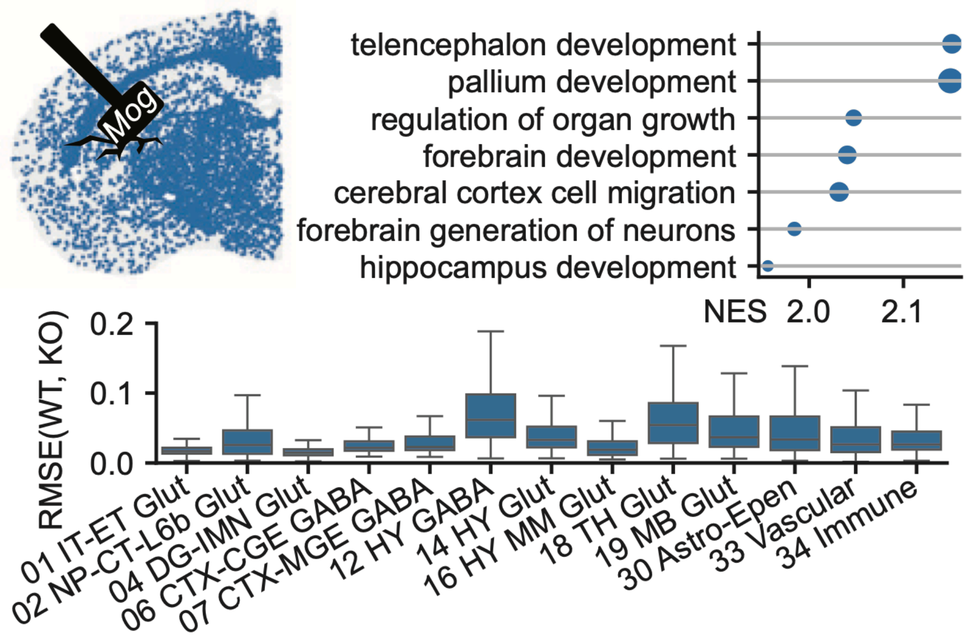




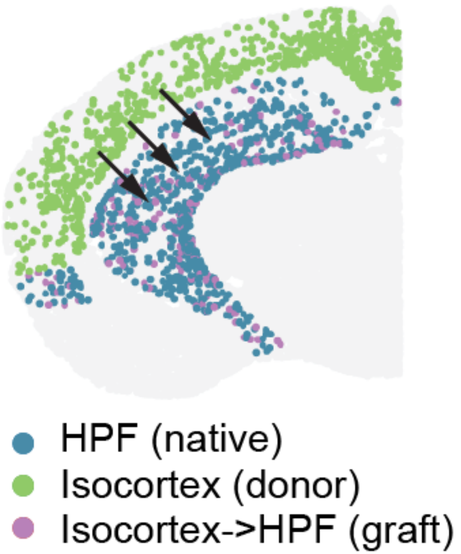
Metagene for head #8

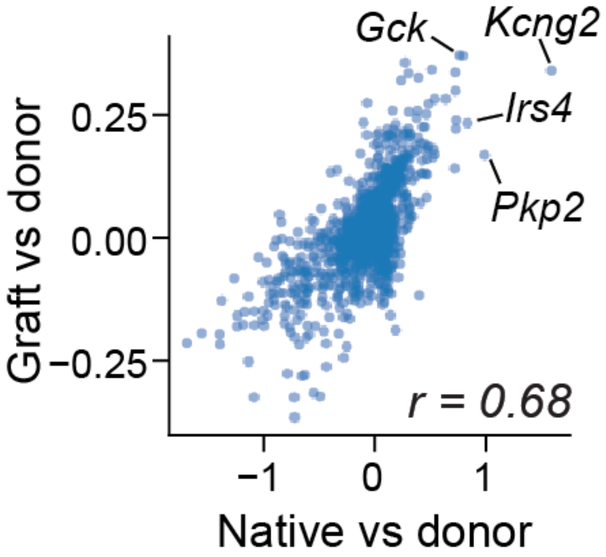
Liang et al. *bioRxiv* 2025

STEAMBOAT enables *in silico* spatial perturbations



- Knock out *Mog* in OPC-Oligo.
- What happens in other cells?



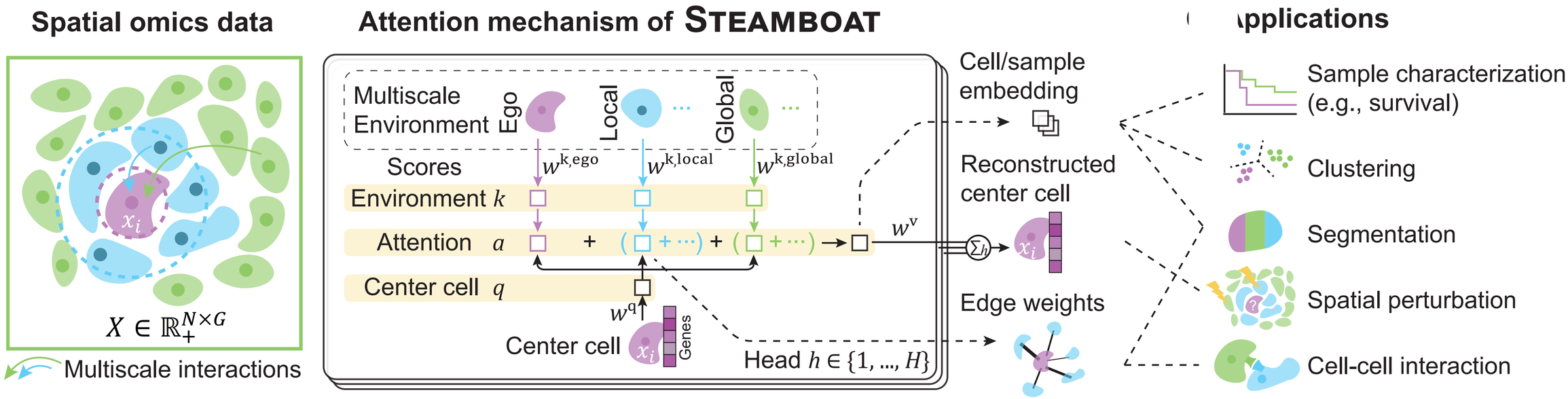


- Move OPC-Oligo from Isocortex to HPF region.
- Gene expression of transplanted cells change towards native ones.

Liang et al. *bioRxiv* 2025

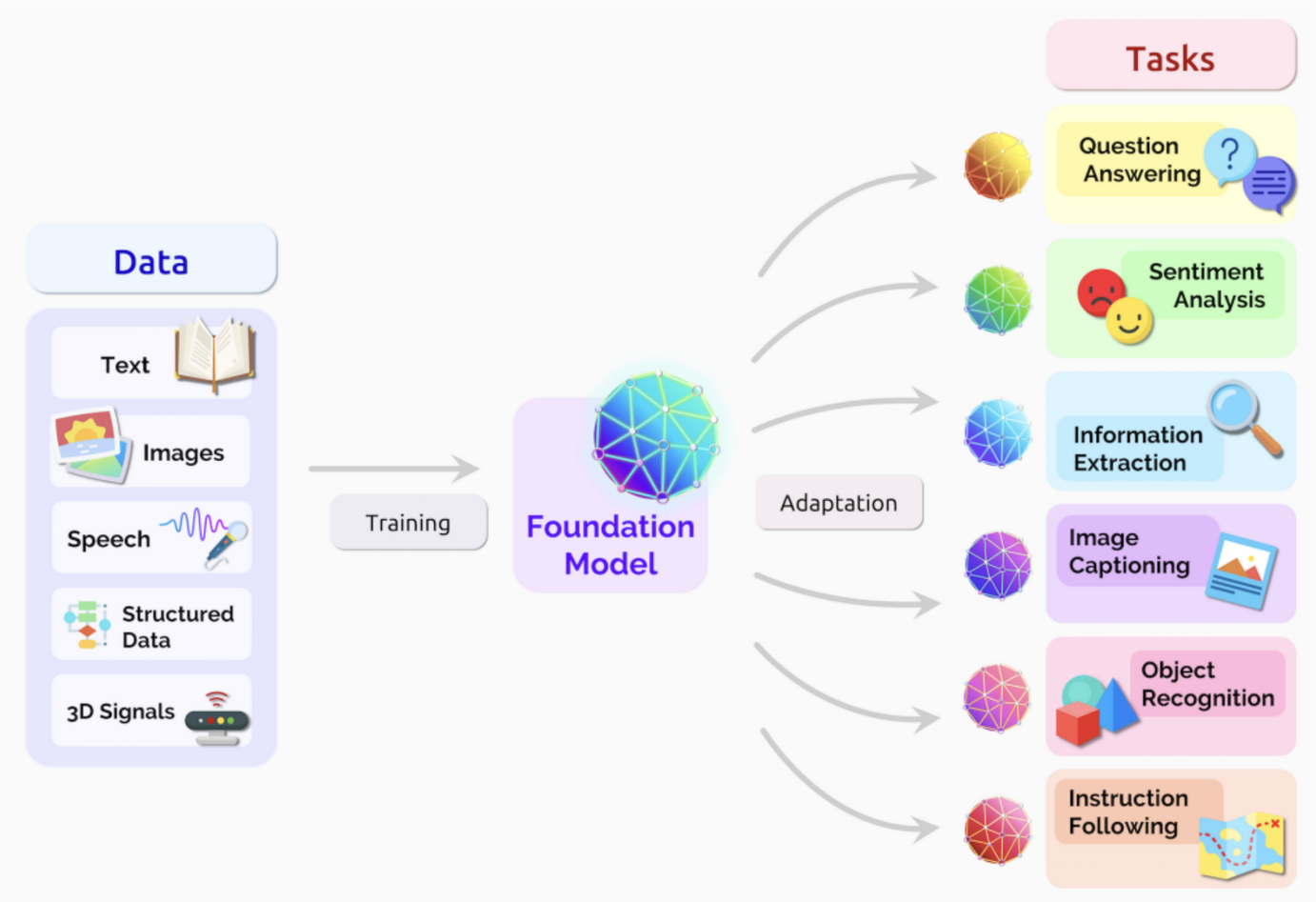
Steamboat for modeling cell interactions in tissues

- Steamboat is a de novo multi-scale cell-cell interaction model.
- Many applications: spatial perturbation; characterize samples; spatial domains & cell types
- The model could serve as the basis for tissue Foundation Model.



Liang et al. ***bioRxiv*** 2025 (*RECOMB* 2025)

What is a Foundation Model?



"On the Opportunities and Risks of Foundation Models"
Bommasani et al. Stanford CRFM 2022

- Foundation models are a replacement for task-specific models

- Large-scale **pretraining** on large unlabeled datasets
- **Finetuning** for diverse downstream tasks

- Self-supervised learning
- Transfer learning

- GPT-4, DALL-E 2, BERT, etc.

Application of (large) language models in genomics

- Large pretrained models can be utilized for finetuning on downstream tasks with limited training data

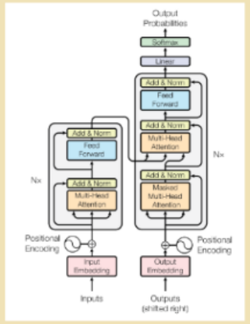
- Data sparsity problem in biology
  - noisy/sparse data
  - incomplete data in biology, e.g., rare disease, precious samples

- Embeddings with more generalized knowledge can help mitigate batch effects

- A few interesting attempts in several direction:
  - Modeling genomic sequences
  - Modeling single cell gene expression data
  - Modeling protein sequence and structure

Architecture of LLMs for genomic sequence

Choose Your Fighter (DNA Language Model):

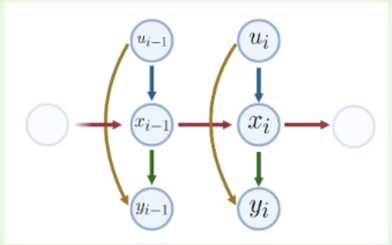


Attention is all you need
(Viswani et al 2015)

Vanilla Attention

- Gave Rise to immense success in vision and NLP

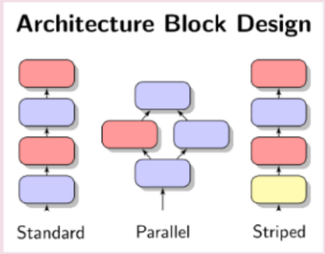
- Pros: Effective, relatively well studied
- Cons: Quadratic Complexity



Subquadratic Attention

- SSMs/Mamba/Hyena/ RetNet/RNNs/RWKV/ Griffin/BASED

- Pros: Subquadratic complexity
- Cons: Approximates Vanilla Attention, have trade-offs

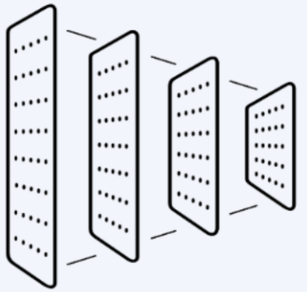


Mechanistic Design of Hybrid Architecture (Poli et al 2024)

Hybrid Attention

- Striped-Hyena
- Striped-Mamba

- Pros: Subquadratic
- Cons: Less well understood, only 2 canonical striped models



Created by Oleksandr Panasovskyi from Noun Project

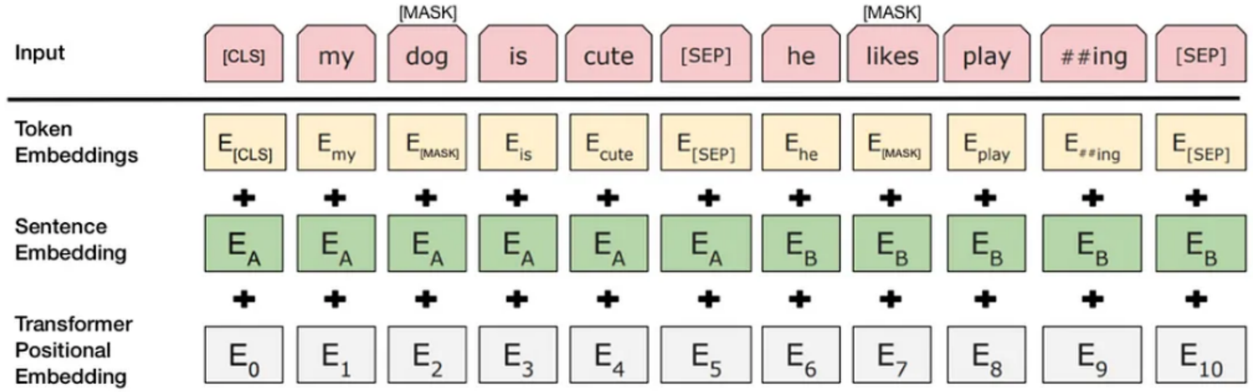
Convolutions

- Dilated Convolutions
- Hyena Hierarchy (global convs)

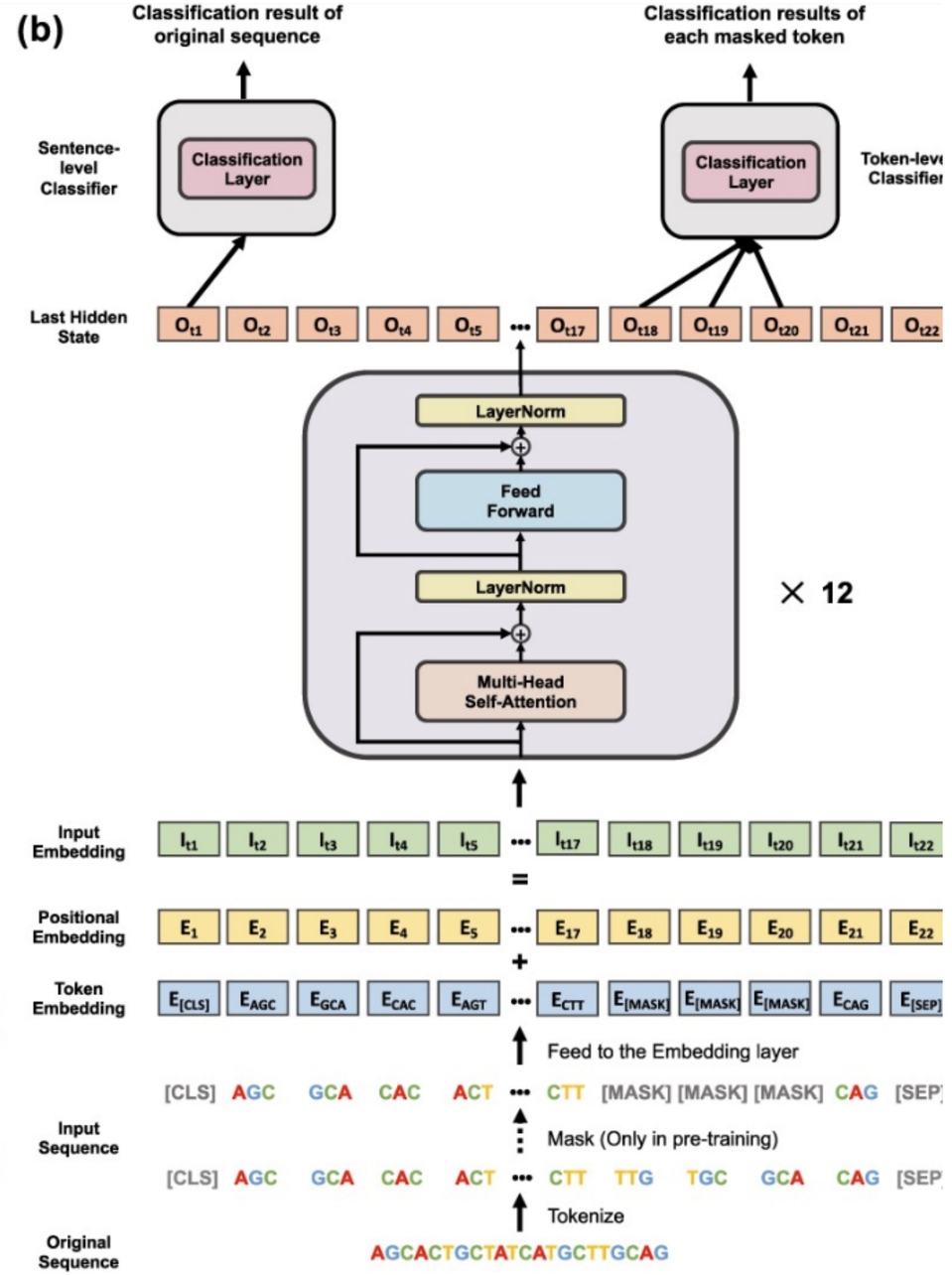
- Pros: Widely used, local convs appear effective for DNA
- Cons: may lack global context, less expressive to attention

Modeling DNA sequences

- DNABERT
  - Pre-trained BERT for DNA sequences based on the human reference genome
  - Overlapping k-mer tokenization
- More recent methods:
  - NT, HyenaDNA, Caduceus, Evo, …



BERT input representation

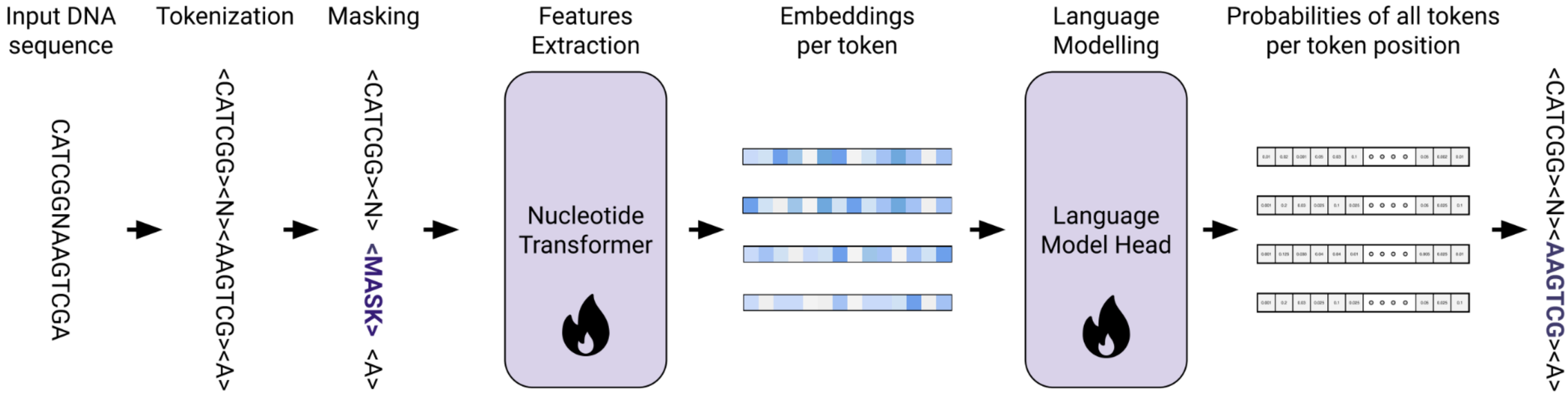


DNABERT input representation
Ji et al. *Bioinformatics* 2021

Nucleotide Transformer

- Pre-trained BERT for DNA sequences on humans, 1000 genomes, and multispecies
- Non-overlapping K-mer tokenization
- Context length of 12K bp

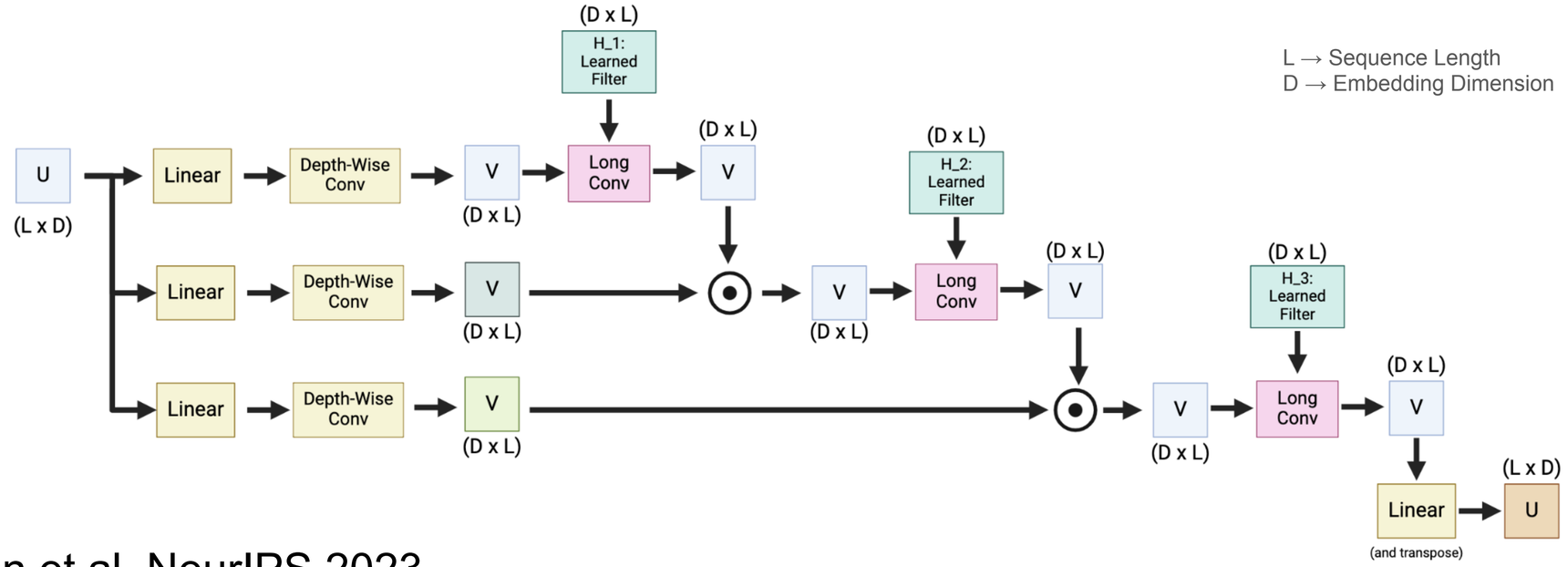
- Downstream prediction tasks:
  - promoter region, TFBS, splice site, functional variants identification



Dalla-Torre et al. Nat Methods 2024

HyenaDNA

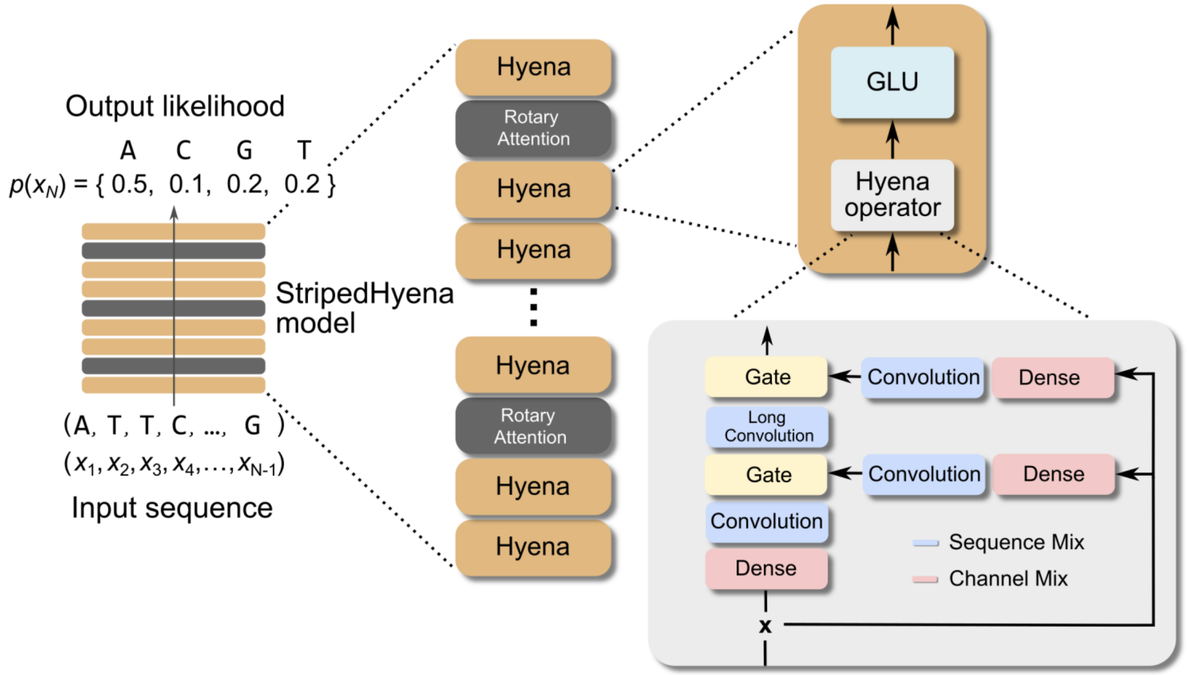
- Pre-trained next token prediction for DNA sequences using a convolution-based architecture
- Tokenization: Nucleotide base-pair resolution
- Advantages: Long context modeling (~1M context length)
- Disadvantages: Not quite clear if this convolutional architecture has the capacity to match transformers

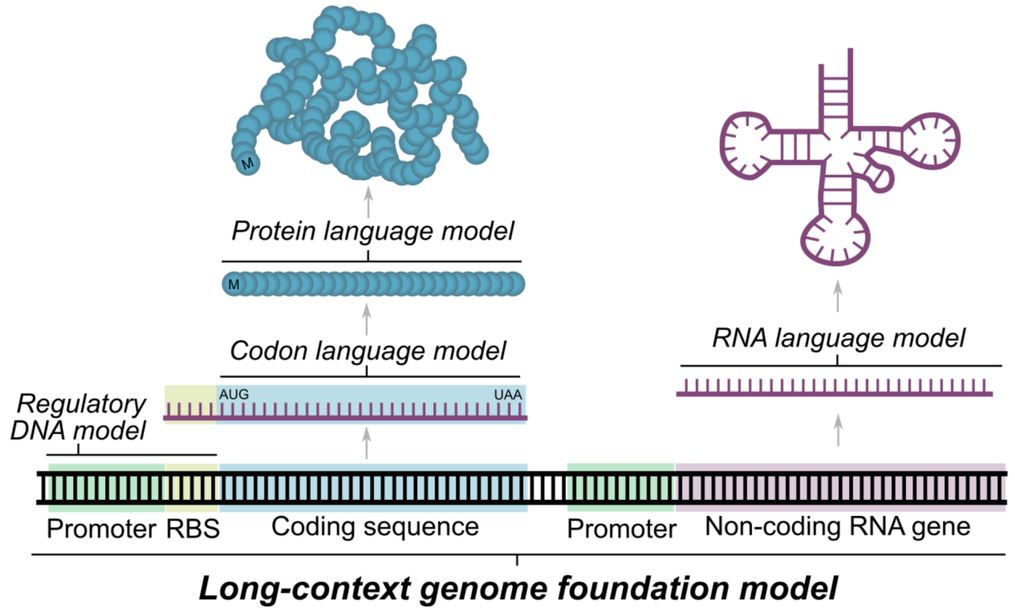


Nguyen et al. NeurIPS 2023

Evo

- Autoregressive (next-token prediction) pretrained on prokaryotic and phage genomes
- Striped Hyena architecture: combination of 29 hyena layers and 3 attention layers
- Demonstrates that aspects of protein and ncRNA can be evaluated through a model trained on DNA sequences





Nguyen et al. Science 2024

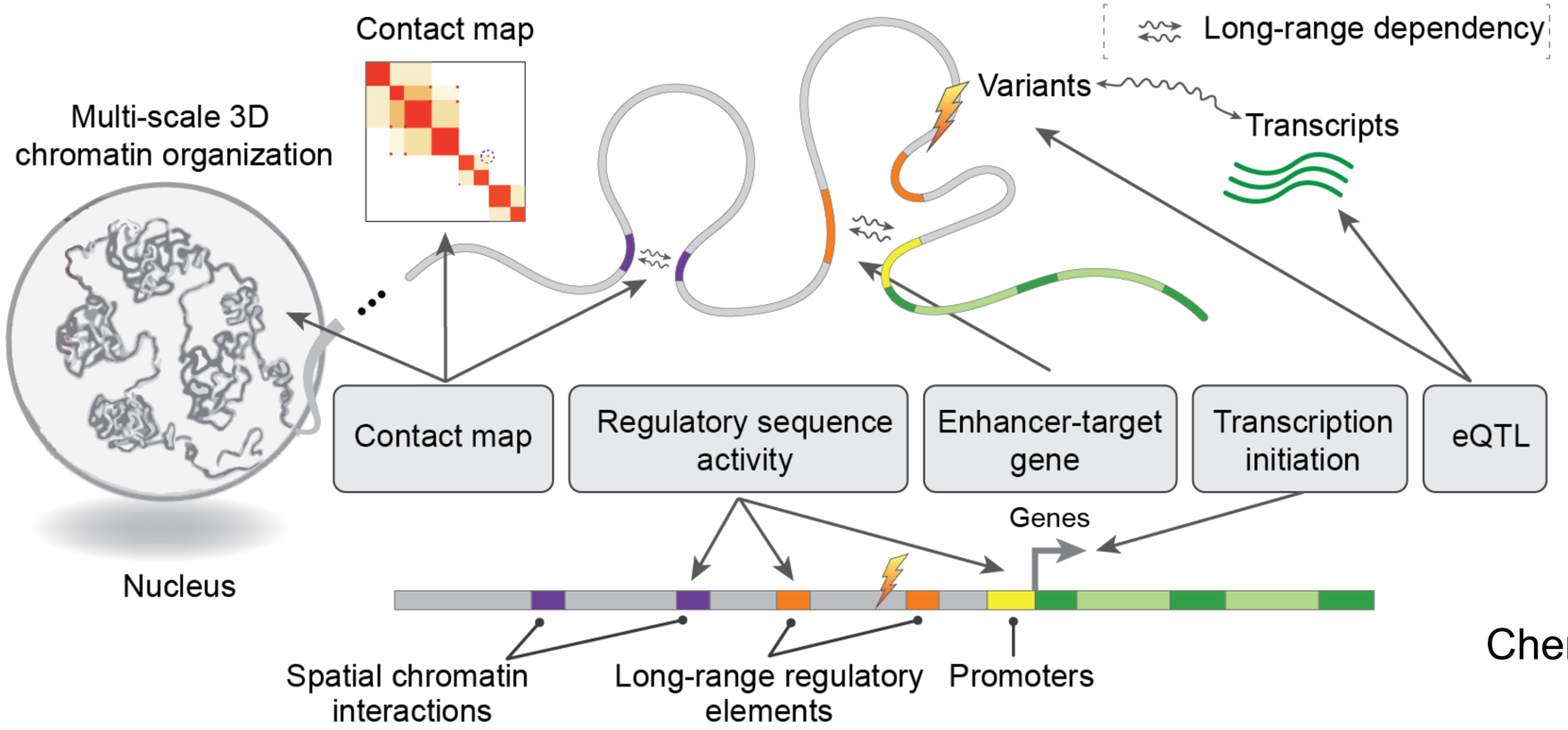
Importance of benchmark datasets

- Motivation: Most previous benchmarks for genomics focus on short-range (input lengths < 1000) classification tasks
- DNALongBench: A benchmark suite for long-range DNA prediction tasks

Wenduo Cheng

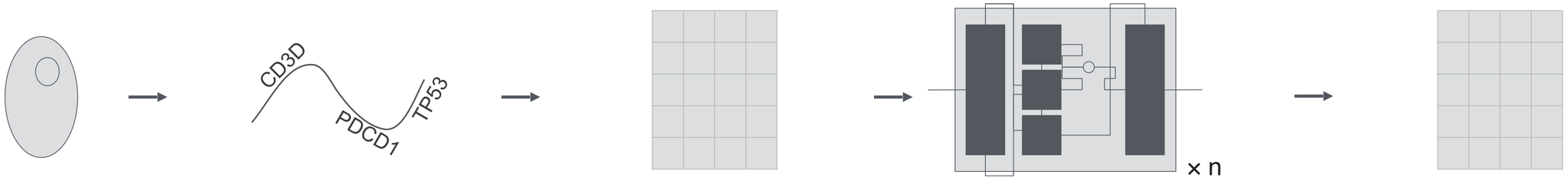
Zhenqiao Song

Lei Li



Cheng et al. bioRxiv 2024

General structure of single-cell FMs



Cell

Sentence of genes

Initial embedding

Transformer (w/ attention)

Final embedding

Expression value embedding

Gene embedding

Pooling

Pooling

Trained Gene embedding

- gene network
- gene dosage sensitivity

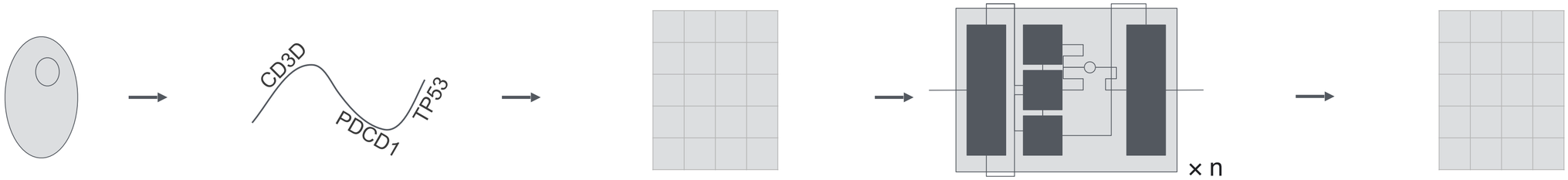
Cell embedding

- cell-type annotation
- batch integration
- perturbation prediction

Sample embedding

- patient outcome

Tokenization for cells



Cell → Sentence of genes → Initial embedding → Transformer (w/ attention) × n → Final embedding

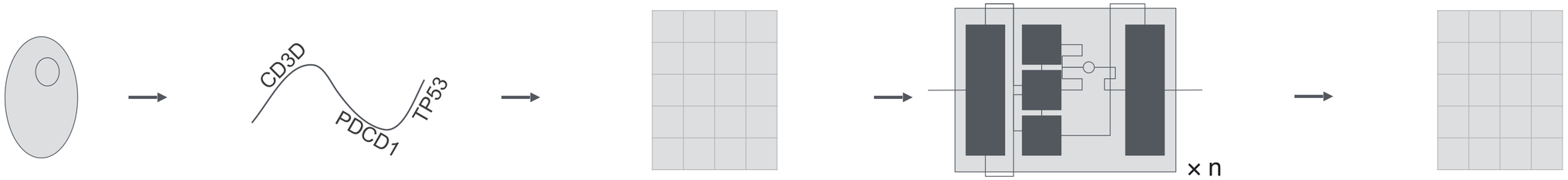
Gene embeddings ordered by their expression value (Geneformer)

Gene embeddings + binned expression value embeddings (scGPT, scBERT)

Gene embeddings + expression value embeddings (scFoundation)

Gene embedding by protein language model (UCE)

Network structures and training strategies



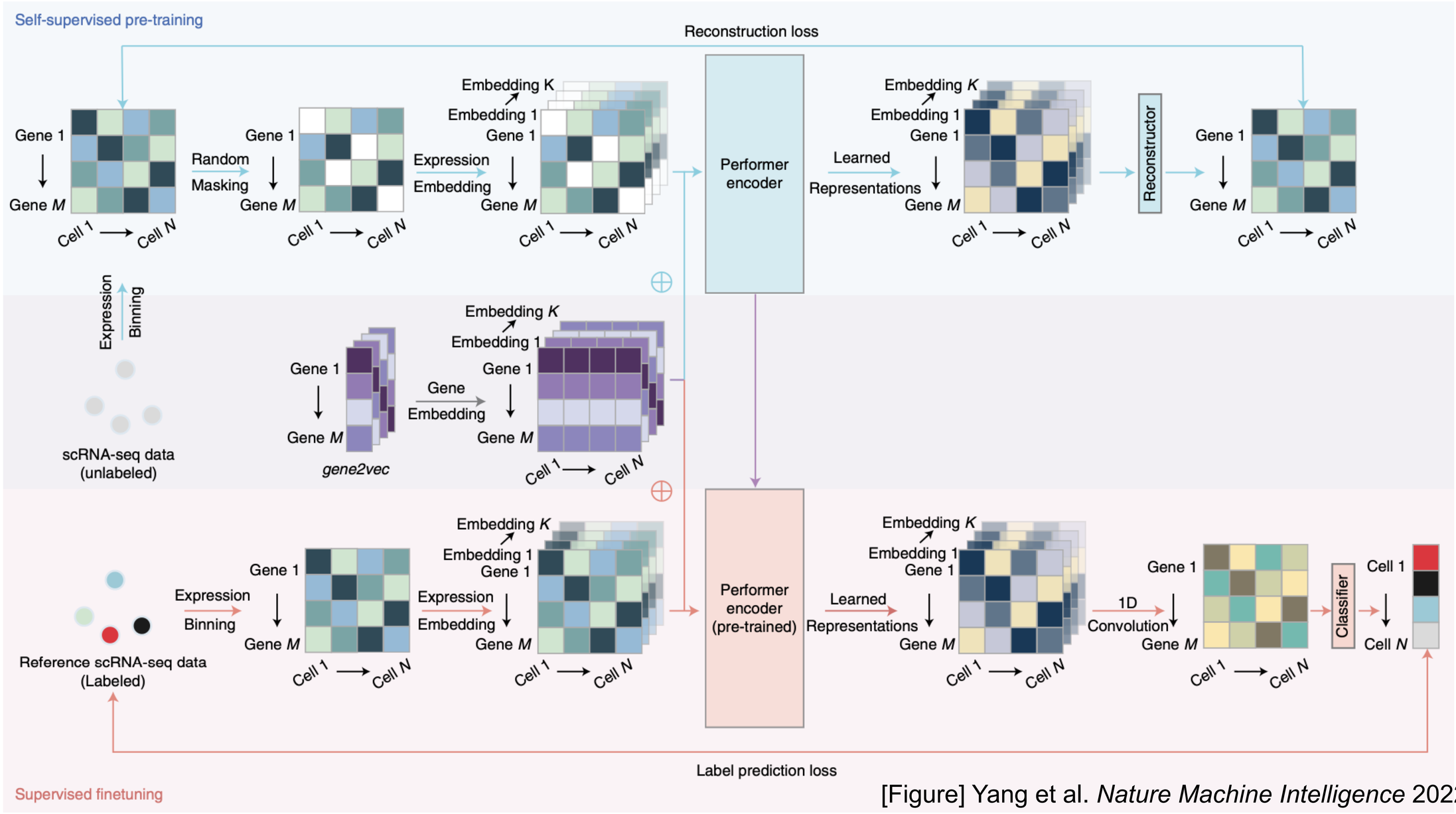
Cell → Sentence of genes → Initial embedding → Transformer (w/ attention) → Final embedding

- Autoregressive pretraining (scGPT)
- Masked language modeling (MLM) pretraining (scBERT, Geneformer, scFoundation)

Timeline of scFMs

	# Parameters	Training data size	Highlights	Paper
2022 **scBERT**	5M	1M	> Scalability: Performer	Yang et al., Nat Mach Intell 2022
Geneformer	40M	30M	> Gene networks inference	Theodoris et al., Nature 2023
scGPT	51M	33M	> Generative pretraining (cell & gene prompt)	Cui et al., Nat Methods 2024
2023 **scFoundation**	100M	50M	> Scalability: reduced input length > Integration: confounding factors regressed out	Hao et al., Nat Methods 2024
UCE	650M +15B pLM (fixed)	46M	> Cross-species integration: utilizes pLM (ESM-2) for gene embedding	Rosen et al., bioRxiv 2023
scMulan	368M	10M	> Multi-tasking: query by prompts > Richer pretraining: metadata	Bian et al., RECOMB 2024
2024 **NicheFormer**	50M	110M	> Integration: dissolved & spatial assays	Schaar et al., bioRxiv 2024

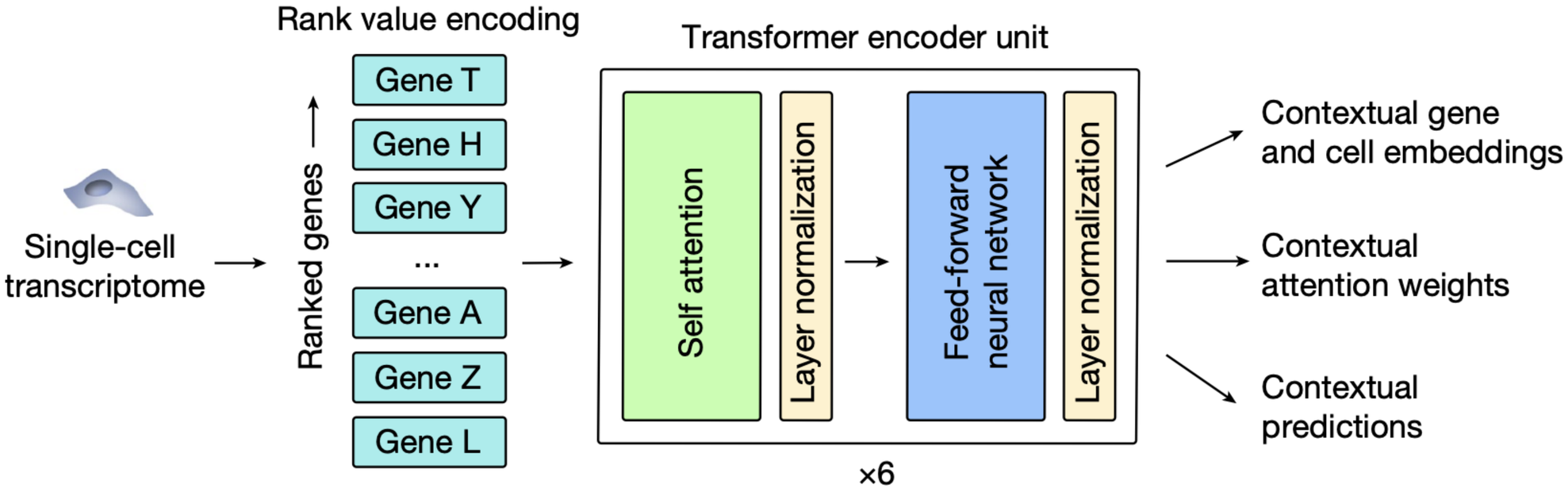
Full scBERT model training scheme



[Figure] Yang et al. *Nature Machine Intelligence* 2022

Geneformer

- Geneformer
  - Pretrained on 30 million scRNA-seq to enable context-specific predictions
  - Discretize gene expression through ranking genes according to their expression
  - Encodes network hierarchy in the attention weights of the model
  - In silico perturbation: remove a gene, compare cell and gene embeddings
- Other recent methods: scGPT, UCE …



Theodoris et al. *Nature* 2023

Leveraging prior knowledge for improved gene embeddings

Gene embeddings can be trained de novo, but prior knowledge may help:

- Gene2vec
  - Distributed representation based on co-expression (used in scBERT)
- GenePT (Chen and Zou)
  - Use GPT-3.5 to generate gene embeddings from gene description.

However, because each gene is treated as a separate entity, knowledge about one gene is not transferable to another. Also, recognizing similarity of genes across species is important for a universal model.

- Universal Cell Embeddings (UCE)
  - Uses protein LLM to embed a sample's genes with protein products
  - Protein products make genes across species more comparable

Promises and challenges

- Foundation models for genomes and cells seem to have potentials
  - Pretraining on large number of cells will discover intrinsic interaction of genes
  - Pretrained models are easily adaptable to multiple tasks to enable biological findings

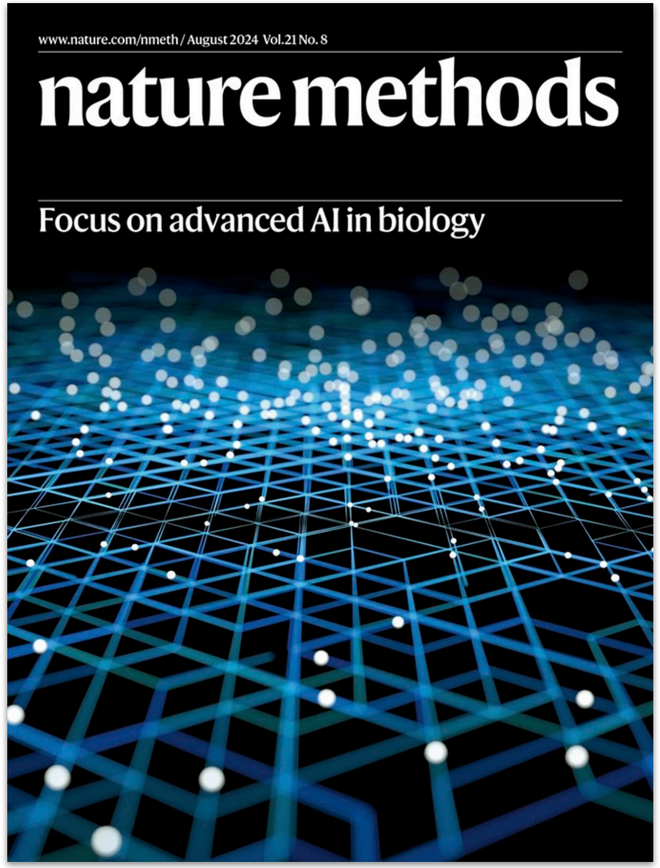
- But biology is complicated and its “language” is likely much harder to model than natural languages.
  - Biological data involve many confounding factors
  - Biological questions are often not mathematically well-defined
  - In this data driven era: “what is the best question to ask”

Questions

- How to better evaluate LLMs?
- How to make LLMs more accessible?
- How to embed cell/gene to better maintain biological contexts?
- How to incorporate prior knowledge into the neural network?
- Do we have enough data available to pretrain LLMs or Foundation Models for various modalities in genomics? (Are we ever going to?)
- DNA and single-cell LLMs have comparable performance compared to existing approaches – need more challenging problems.

- **What are the important problems for LLMs?**

Interpretable AI/ML in the era of LLMs



nature methods

Perspective

https://doi.org/10.1038/s41592-024-02359-7

Applying interpretable machine learning in computational biology—pitfalls, recommendations and opportunities for new developments

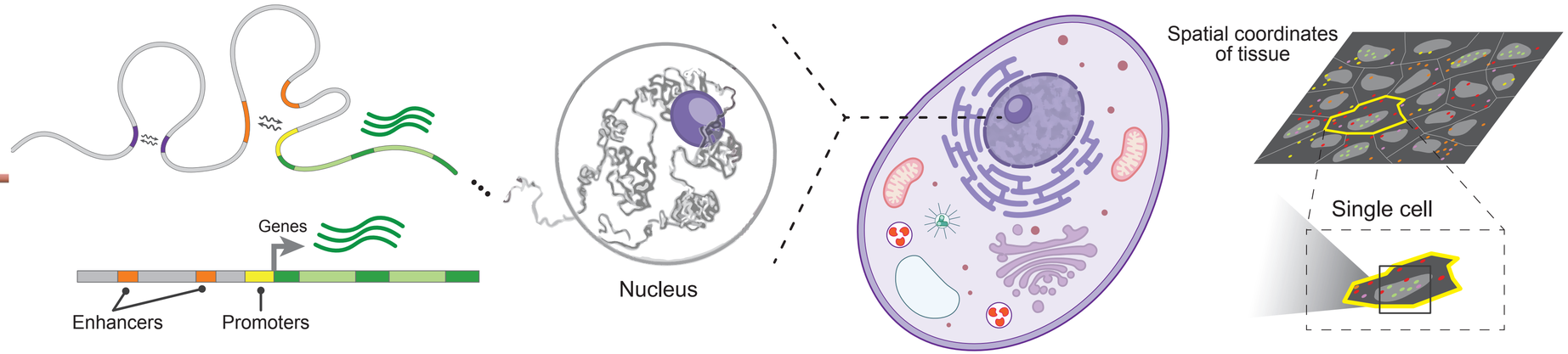
Received: 27 October 2022

Accepted: 24 June 2024

Valerie Chen [1,3], Muyu Yang [2,3], Wenbo Cui [1], Joon Sik Kim [1], Ameet Talwalkar [1] ✉ & Jian Ma [2] ✉

Chen # and Yang # et al. ***Nature Methods,*** 2024

Decoding the “language” of genomes, cells, and tissues



sequence — structure — function

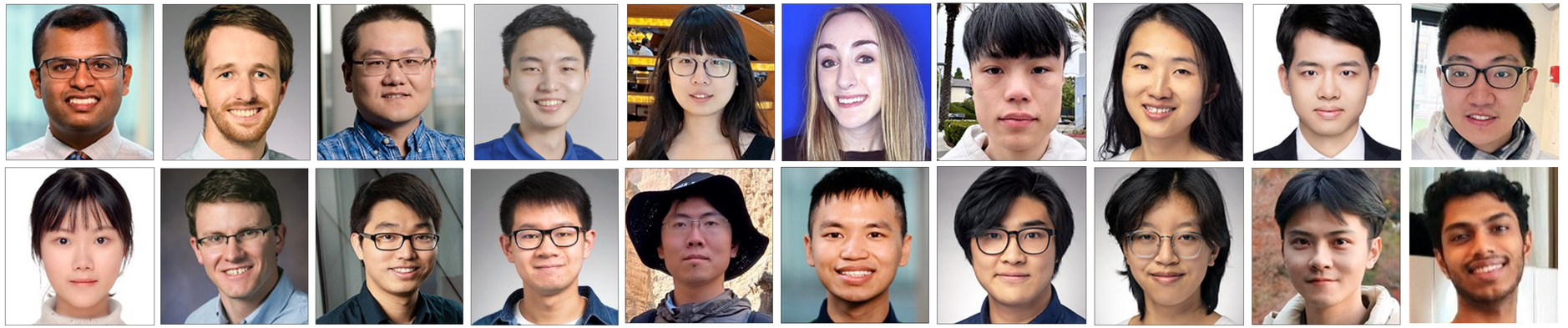
Biology is multiscale. So must our models be.

Living cell was first observed ~350 years ago by

Antonie van Leeuwenhoek

Leeuwenhoek
Microscope
(circa late 1600s)

Acknowledgments



- **4DN UM1 Center**
  Andy Belmont (UIUC)
  Ting Wu (Harvard)
  Frank Alber (UCLA)
  Susanne Rafelski (AICS)
  Nicola Neretti (Brown)
  Dave Gilbert (SDBRI)
  Jason Swedlow (Dundee)
  Tom Misteli (NCI)
- **SenNet Project**
  Nicola Neretti (Brown)
  Steven Wang (Yale)
- **Other Collaborators**
  Zhijun Duan (UW)
  Xiaowei Zhuang (Harvard)
  Mitch Guttman (Caltech)
  Hansruedi Mathys (Pitt)
  Bing Ren (UCSD)
  Jenn Cremins (Penn)
  Fei Chen (Broad / Harvard)
  Jesse Dixon (Salk)
  Ting Wang (WashU)
  Lei Li (CMU)
  Ameet Talwalkar (CMU)
- **Ma lab (CMU)**
  Shahul Alam
  Ellie Haber
  Wendy Yang
  Shuming Liu
  Spencer Krieger
  Remy Liu
  Tianming Zhou
  Ben Chidester
  Kyle Xiong
  Yang Zhang
  Junjie Tang
  Xinyue Lu
  Wenduo Cheng
  Nick Ho
  Ajinkya Deshpande
  Shike Wang
  Xue Er Ding
  Ruochi Zhang