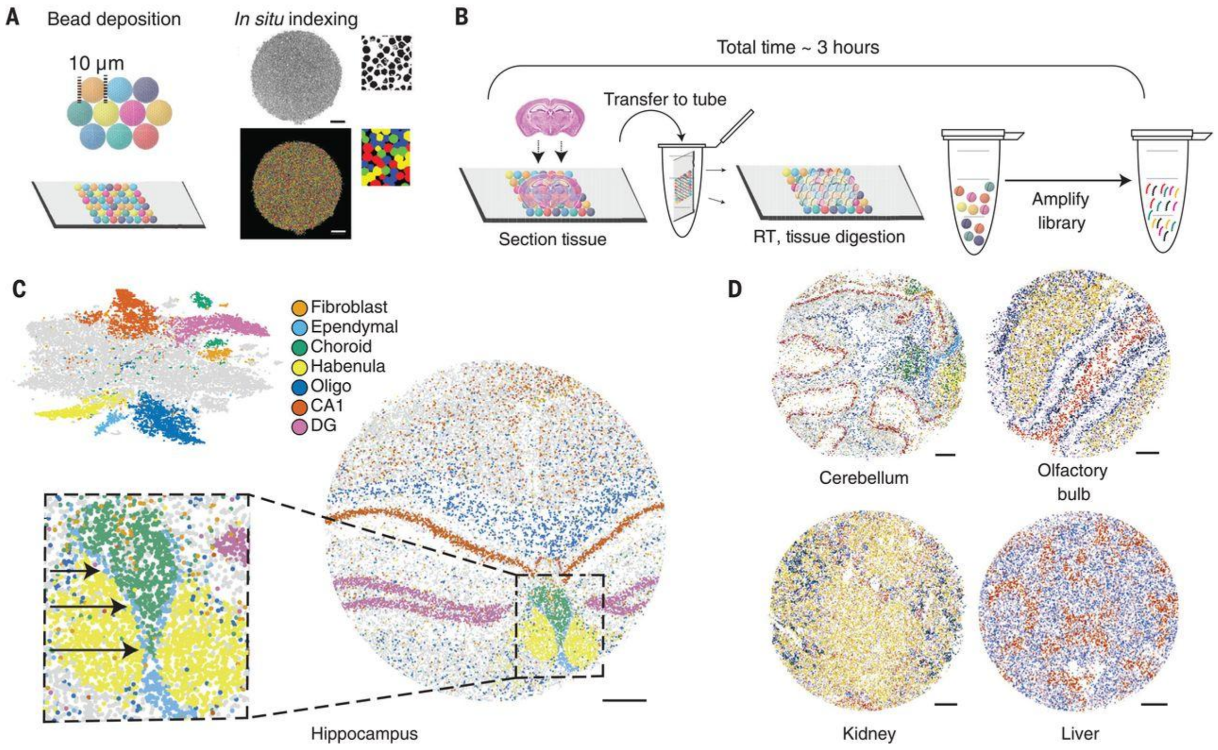
Spatial transcriptomics 2

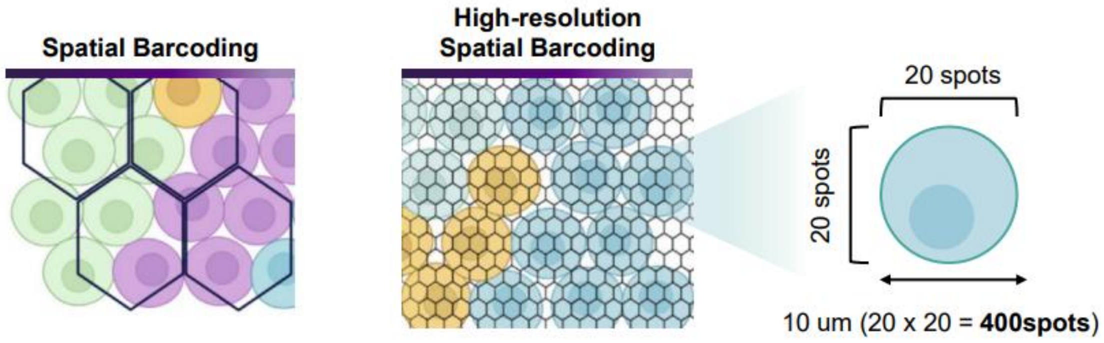
Sequence-based high resolution

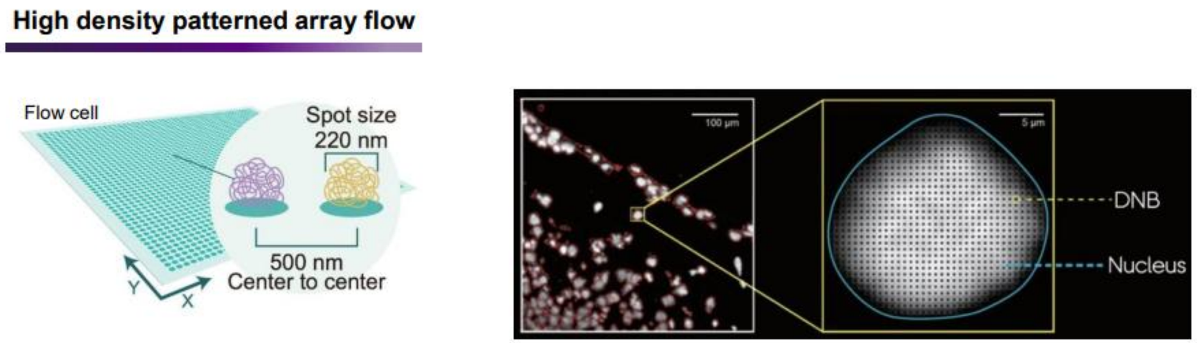
-Visium & GeoMx: full transcriptome → but low spatial resolution

-Probe-base (MerFish, Cosmx, Xenium): high spatial resolution → poor coverage

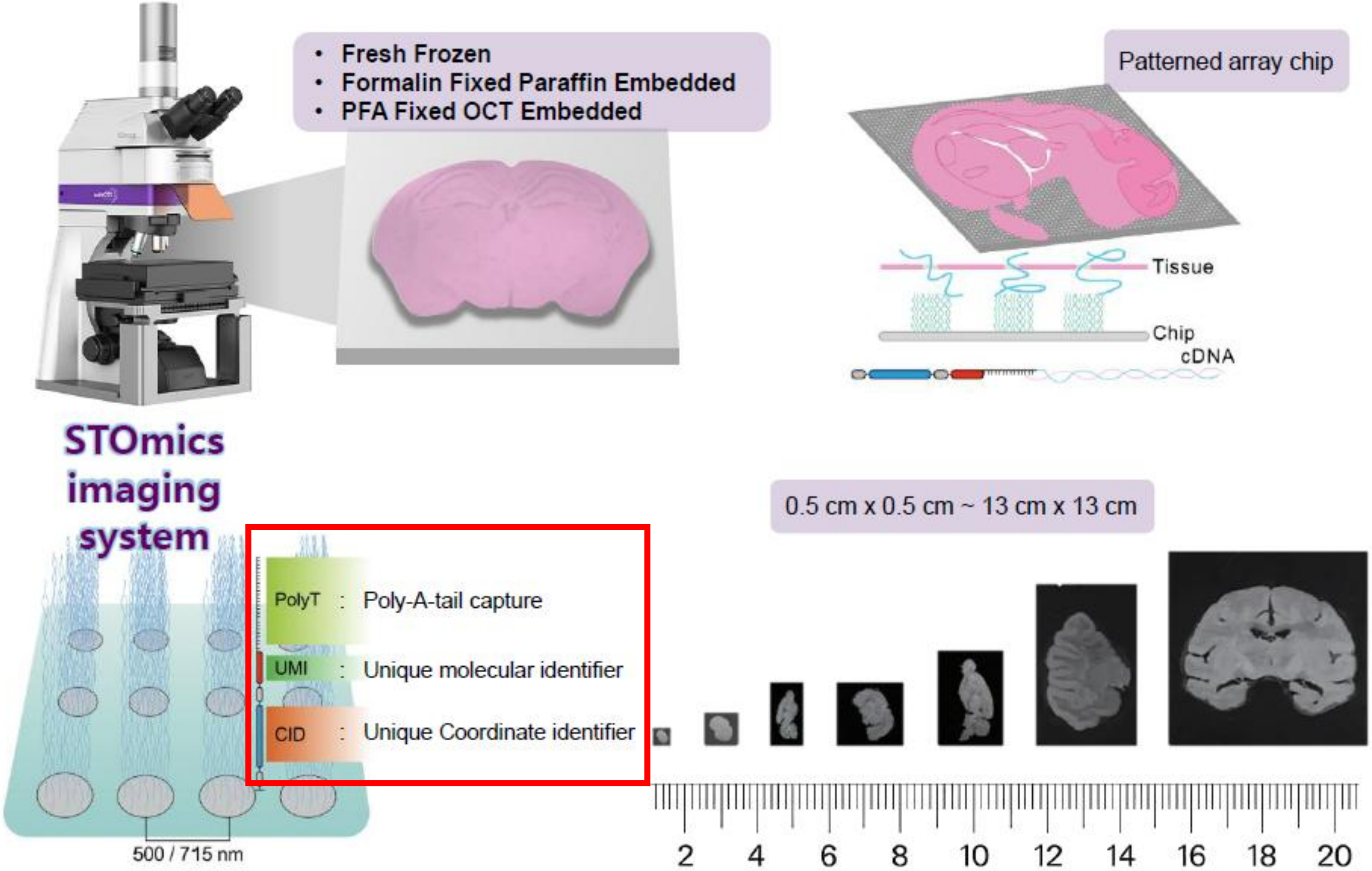
-Slide-seq v2, Stereo-seq → full transcriptome + subcellular level







- Stereo-seq



- Fresh Frozen
- Formalin Fixed Paraffin Embedded
- PFA Fixed OCT Embedded

PolyT : Poly-A-tail capture

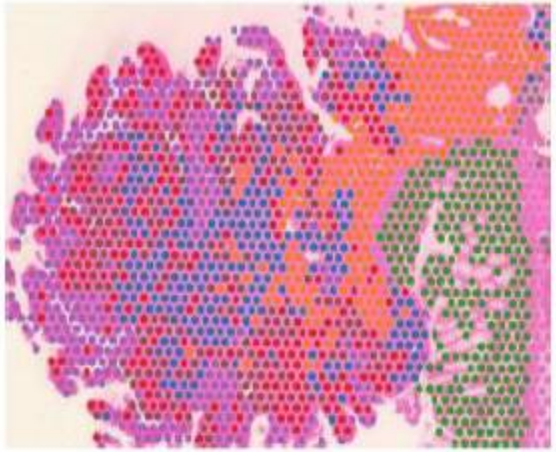
UMI : Unique molecular identifier

CID : Unique Coordinate identifier

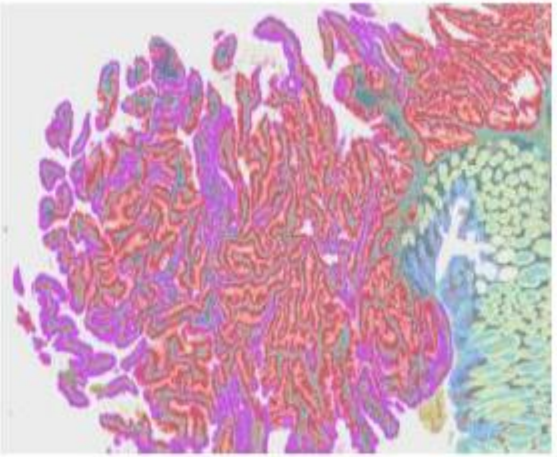
0.5 cm x 0.5 cm ~ 13 cm x 13 cm

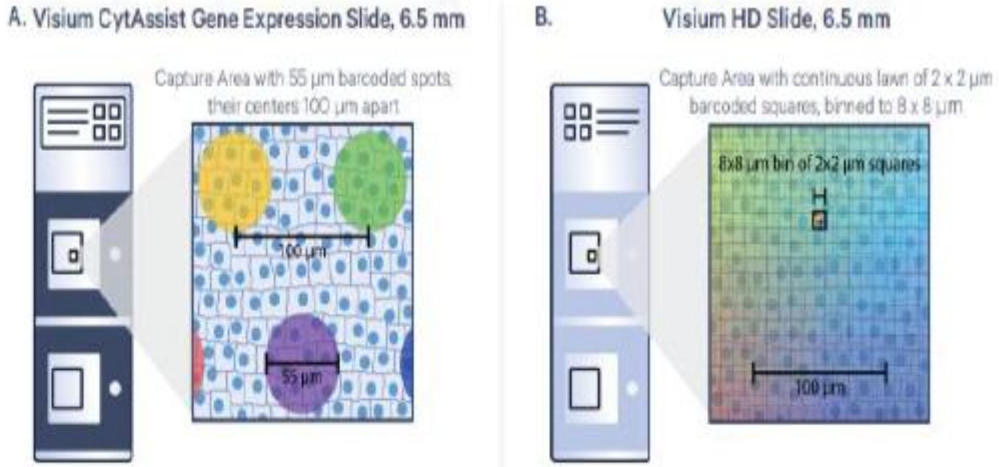
- Stereo-seq

Visium



Visium HD





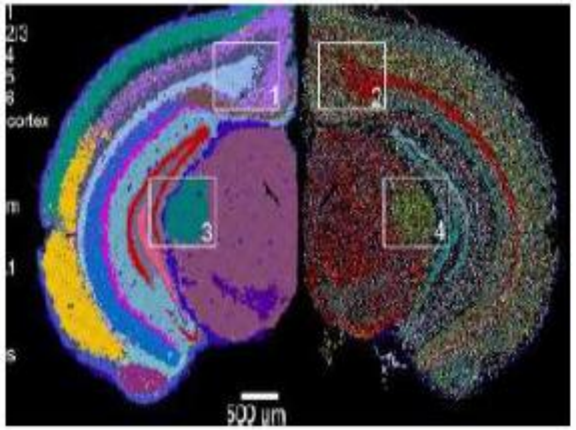
A. Visium CytAssist Gene Expression Slide, 6.5 mm

Capture Area with 55 µm barcoded spots, their centers 100 µm apart

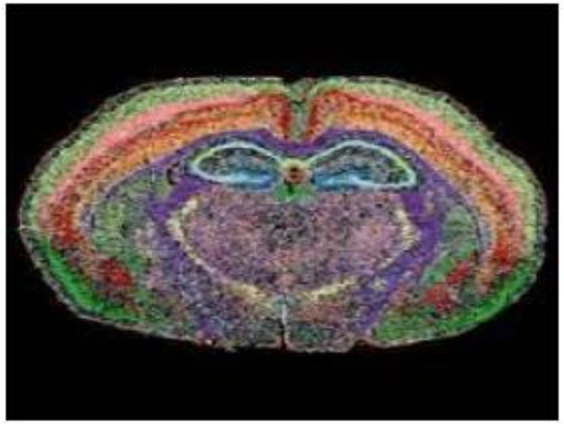
B. Visium HD Slide, 6.5 mm

Capture Area with continuous lawn of 2 x 2 µm barcoded squares, binned to 8 x 8 µm

Stereoseq



Xenium

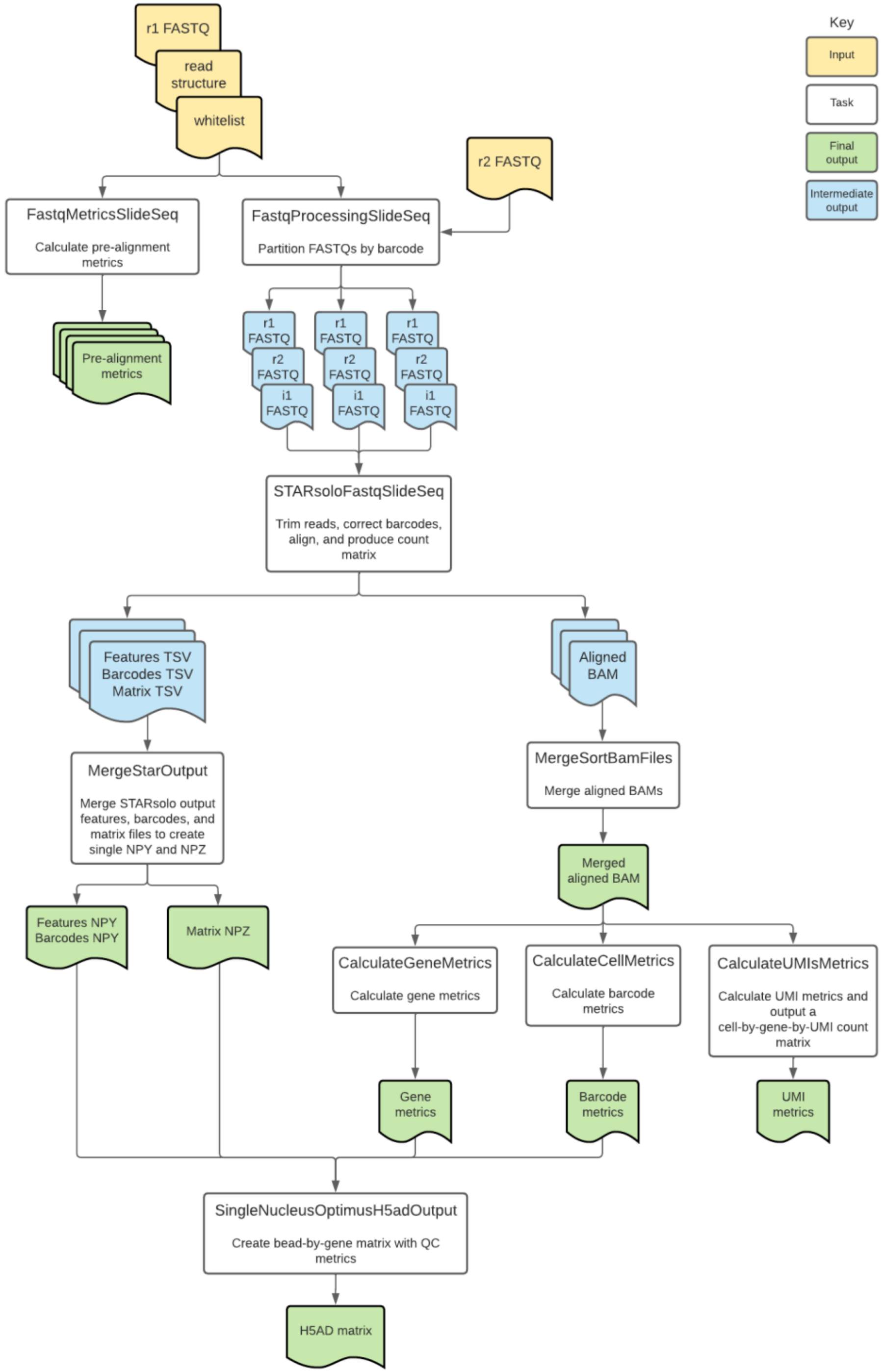


Platform	Spot size (um)	Step size (um)	Resolution (um)
Visium	55	100	100
Visium HD	2	No gap	2
Stereo-seq	0.22	0.5	0.5

- Sequence-based high resolution

-FASTQ (read file + index file: spot coordinate)
-Alignment: STAR
→ Count matrix

-Slide-seq v2: 10um → ~ 1 cell size
-Stereo-seq: 10um: 400 transcriptome bulb)



Key
Input
Task
Final output
Intermediate output

r1 FASTQ
read structure
whitelist
r2 FASTQ

FastqMetricsSlideSeq
Calculate pre-alignment metrics

FastqProcessingSlideSeq
Partition FASTQs by barcode

Pre-alignment metrics

r1 FASTQ
r2 FASTQ
i1 FASTQ

STARsoloFastqSlideSeq
Trim reads, correct barcodes, align, and produce count matrix

Features TSV
Barcodes TSV
Matrix TSV

Aligned BAM

MergeStarOutput
Merge STARsolo output features, barcodes, and matrix files to create single NPY and NPZ

MergeSortBamFiles
Merge aligned BAMs

Merged aligned BAM

Features NPY
Barcodes NPY

Matrix NPZ

CalculateGeneMetrics
Calculate gene metrics

CalculateCellMetrics
Calculate barcode metrics

CalculateUMIsMetrics
Calculate UMI metrics and output a cell-by-gene-by-UMI count matrix

Gene metrics

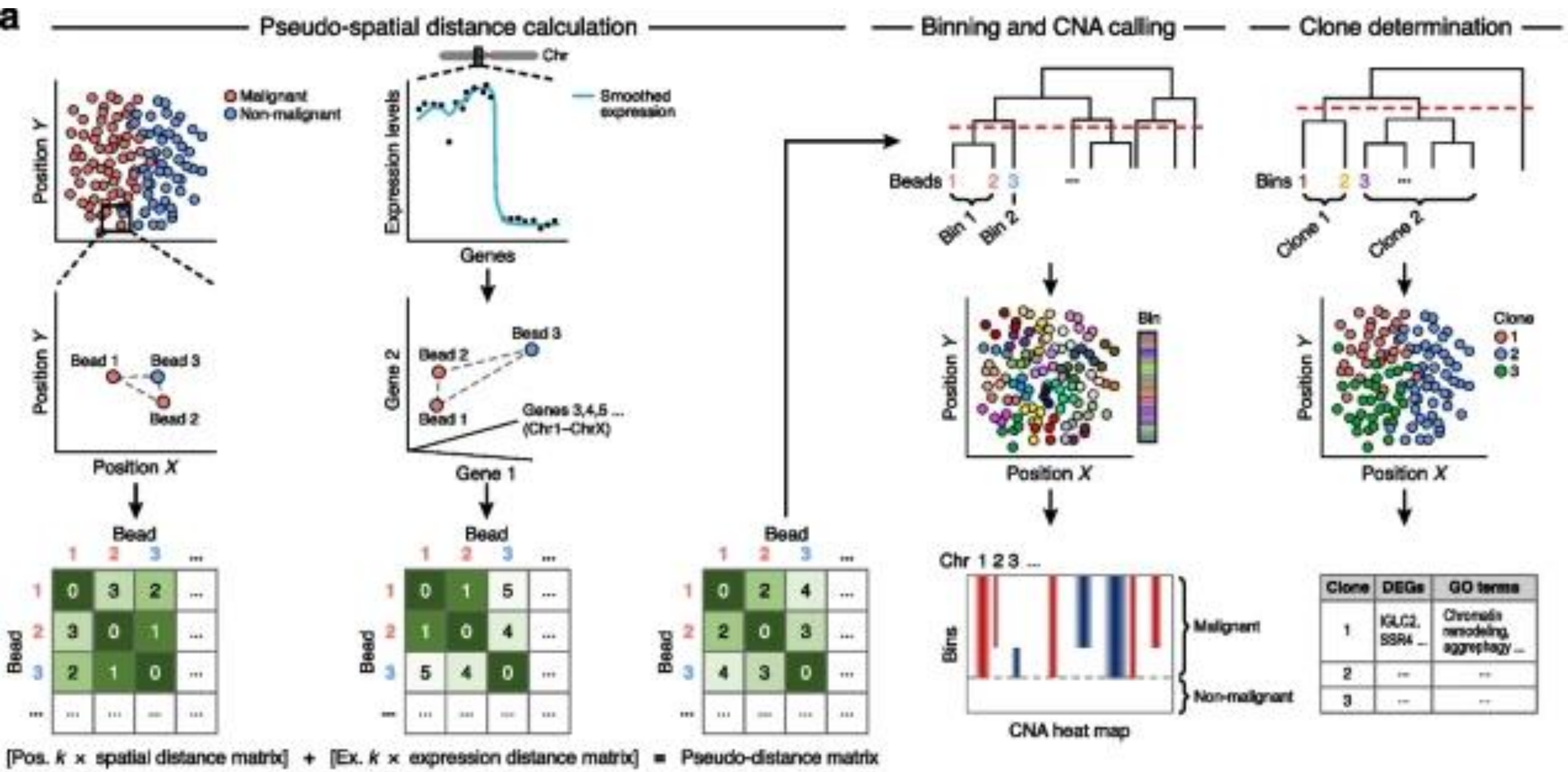
Barcode metrics

UMI metrics

SingleNucleusOptimusH5adOutput
Create bead-by-gene matrix with QC metrics

H5AD matrix

- SlideCNA

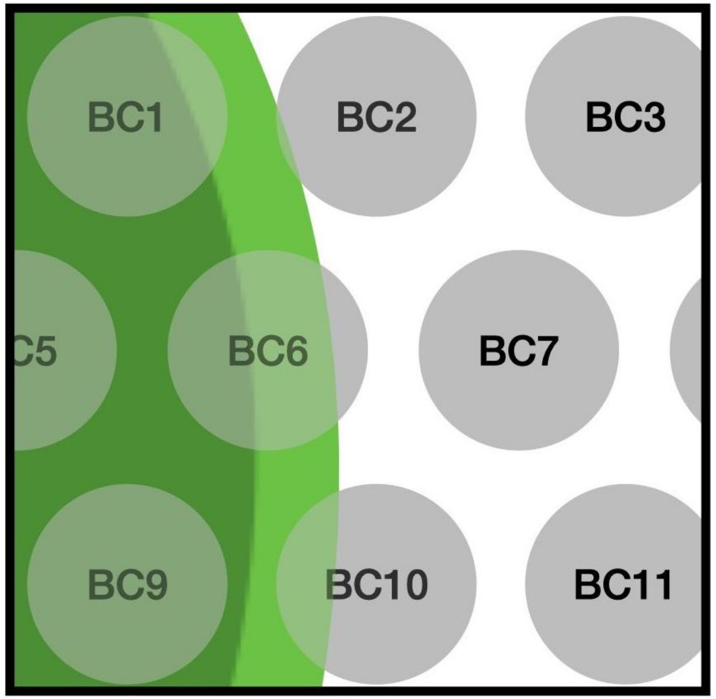


-InferCNV + Spatial

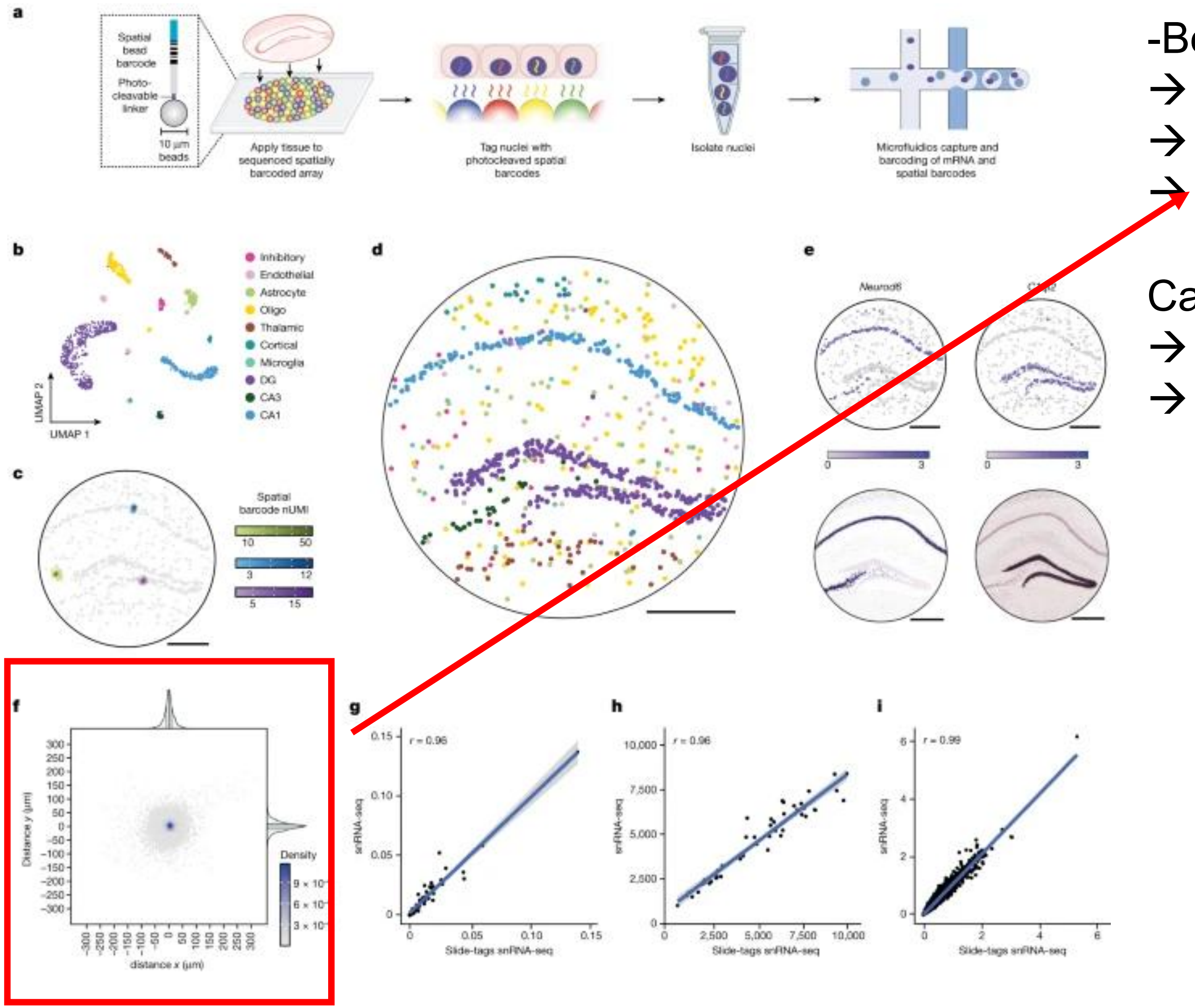
-Sparsity → averaging neighbor spots → InferCNV

- Slide-tag

-Spot-based → anyway, it is not exactly the cell



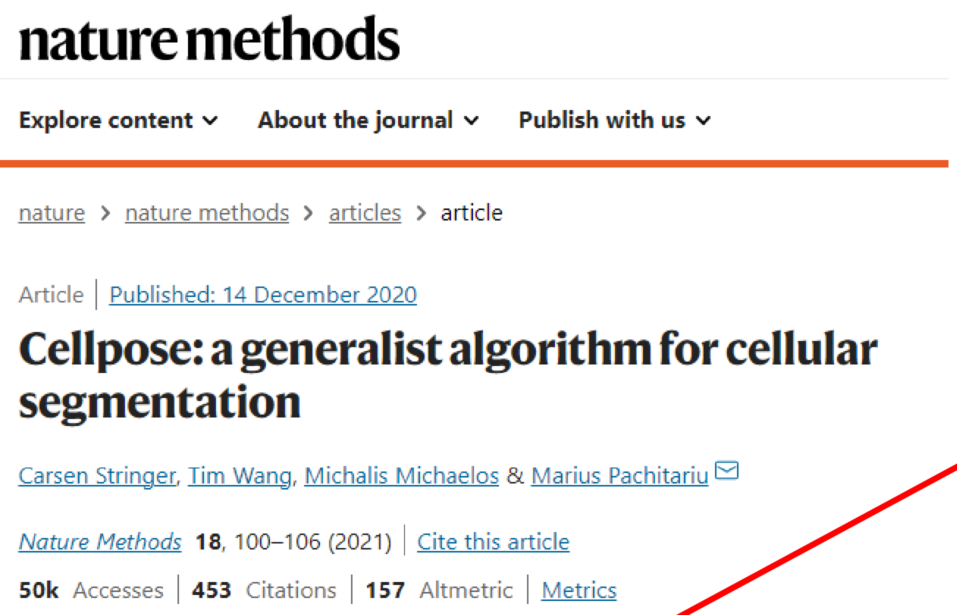
- Slide-tag



-Bead on the Barcode
→ Photoreactive diffusion
→ Each cell (specificity)
→ 10x platform

Caveat !
→ Similar to scRNA-seq
→ Drop-out ↑

• Cell segmentation (Cellpose)



nature methods

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature › nature methods › articles › article

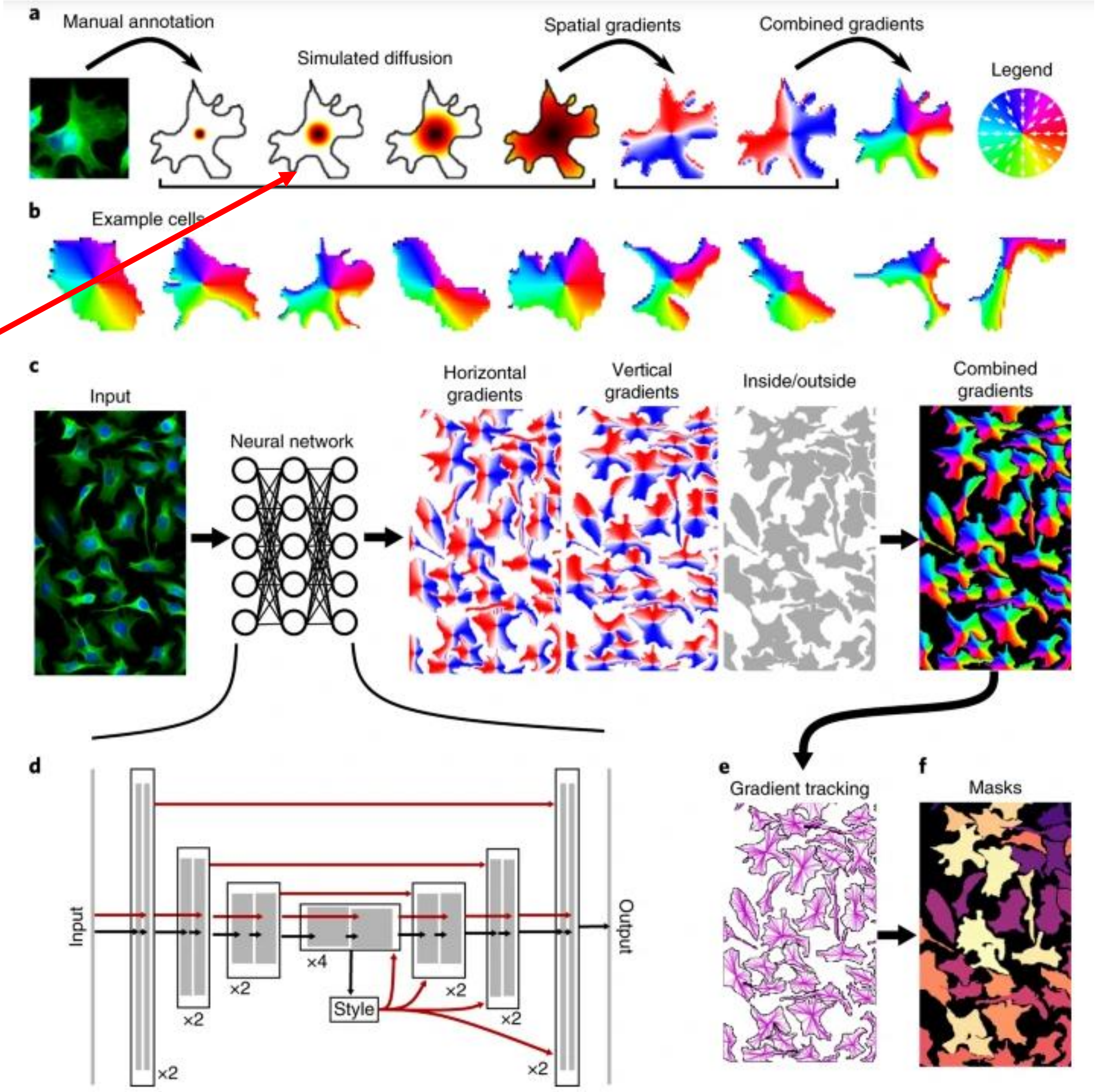
Article | Published: 14 December 2020

Cellpose: a generalist algorithm for cellular segmentation

Carsen Stringer, Tim Wang, Michalis Michaelos & Marius Pachitariu ✉

Nature Methods **18**, 100–106 (2021) | Cite this article

50k Accesses | **453** Citations | **157** Altmetric | Metrics



Training

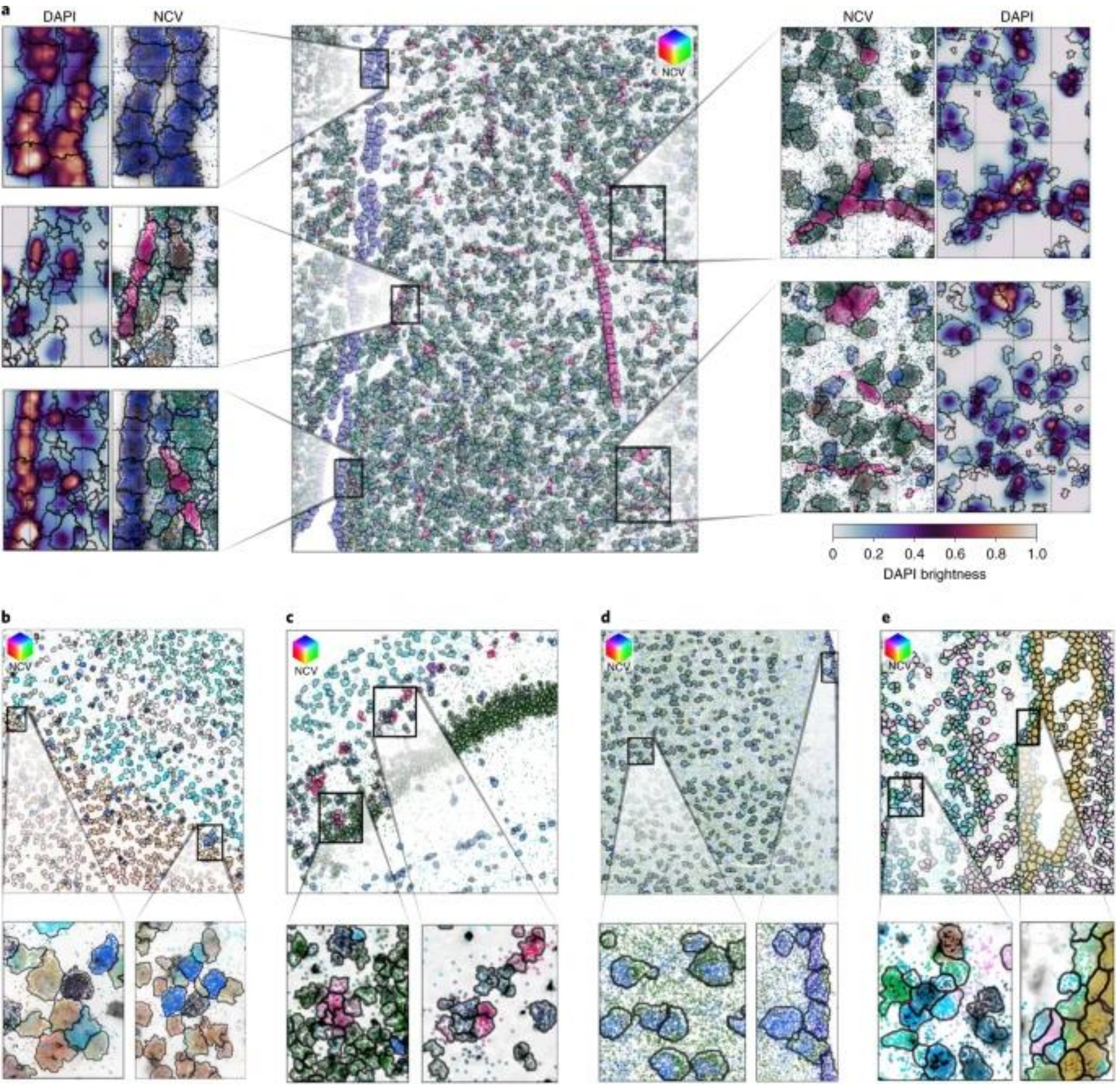
-Draw cell boundary manually

-Gradient diffusion from many angle

-Neural network

Cf) Cosmx

• Cell segmentation (Baysor)



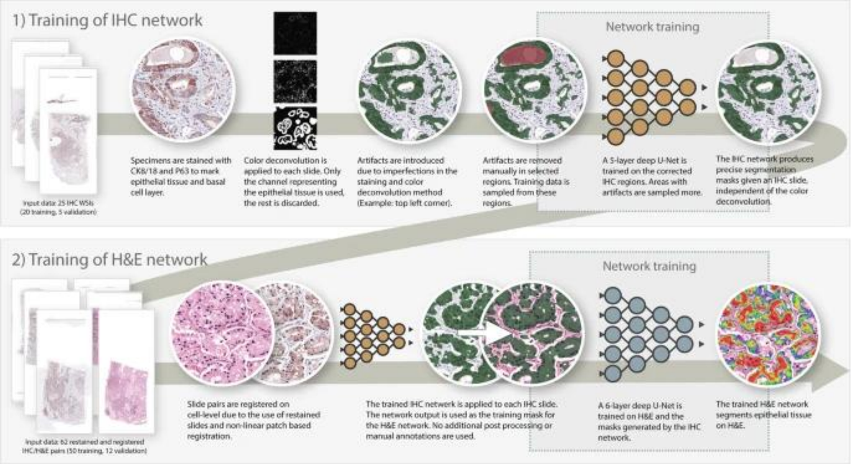
-Location of transcript
+ nucleus or surface marker information (optional)

*assumption
-transcripts are grouped within each cell
-each cell has unique profile
-Unique distribution of transcript in each cell

→ Markov Random Field (MRF) → grouping transcript nearby
→ Bayesian inference → distinguish the group of transcripts into a cell
→ Expectation-maximization: optimization

Cell segmentation in imaging-based spatial transcriptomics

• Cell segmentation (for H&E staining)


1) Training of IHC network

Input data: 25 IHC WSIs (20 training, 5 validation)

Specimens are stained with CK8/18 and P63 to mark epithelial tissue and basal cell layer.

Color deconvolution is applied to each slide. Only the channel representing the epithelial tissue is used, the rest is discarded.

Artifacts are introduced due to imperfections in the staining and color deconvolution method (Example: top left corner).

Artifacts are removed manually in selected regions. Training data is sampled from these regions.

Network training

A 5-layer deep U-Net is trained on the corrected IHC regions. Areas with artifacts are sampled more.

The IHC network produces precise segmentation masks given an IHC slide, independent of the color deconvolution.

2) Training of H&E network

Input data: 62 restained and registered IHC/H&E pairs (50 training, 12 validation)

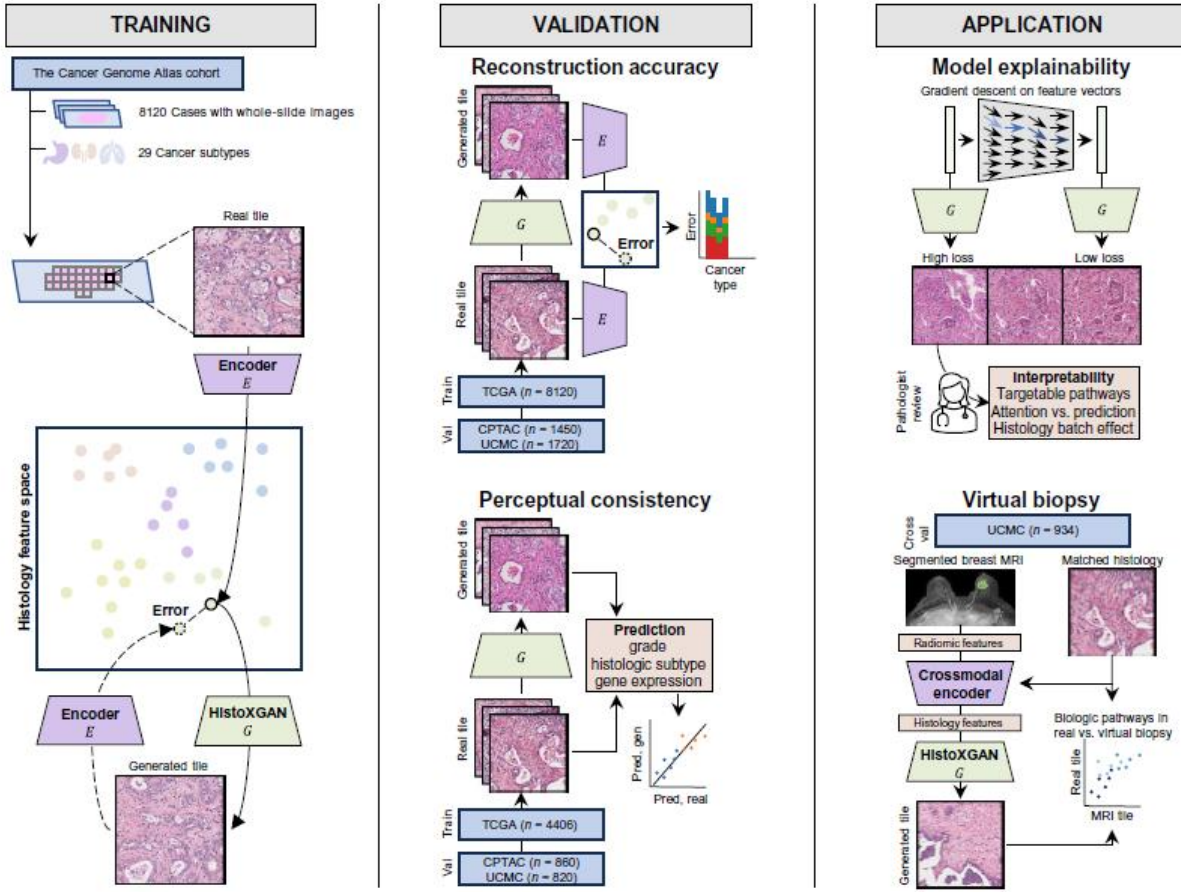
Slide pairs are registered on cell-level due to the use of restained slides and non-linear patch based registration.

The trained IHC network is applied to each IHC slide. The network output is used as the training mask for the H&E network. No additional post processing or manual annotations are used.

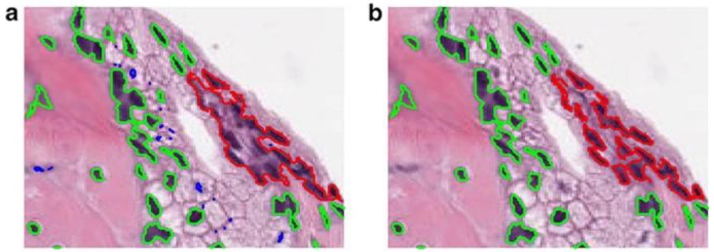
Network training

A 6-layer deep U-Net is trained on H&E and the masks generated by the IHC network.

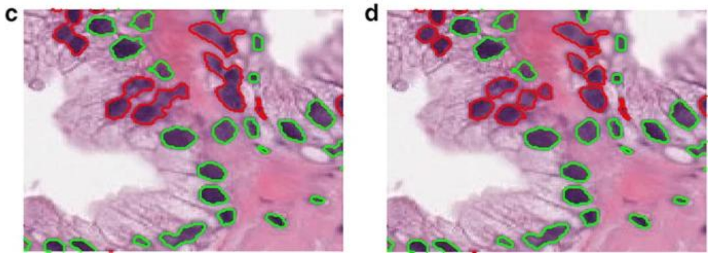
The trained H&E network segments epithelial tissue on H&E.



Area-based correction

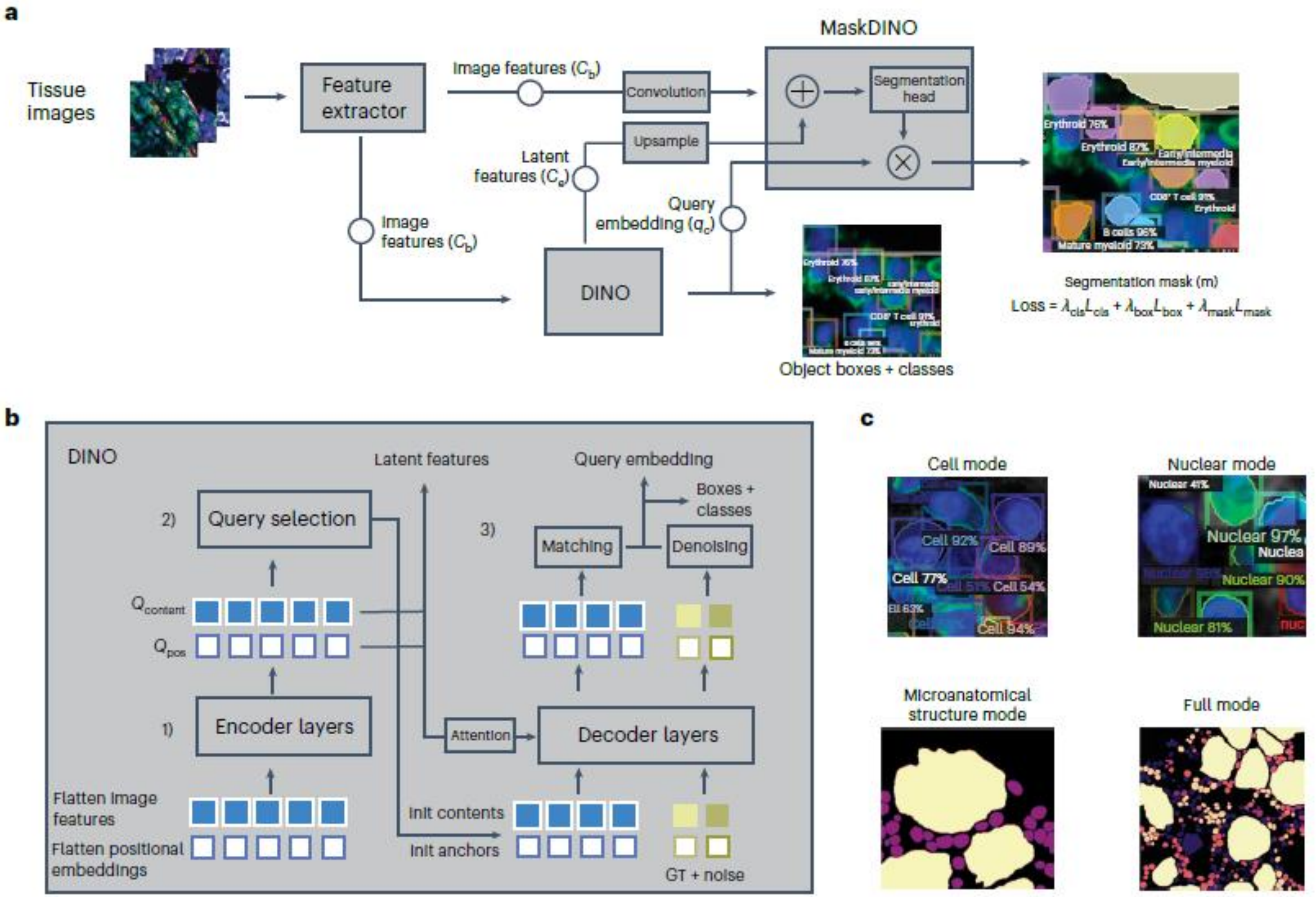


Nuclei separation



Epithelium segmentation using deep learning in H&E-stained prostate specimens with immunohistochemistry as reference standard
Multi-tissue and multi-scale approach for nuclei segmentation in H&E stained images
Generative adversarial networks accurately reconstruct pan-cancer histology from pathologic, genomic, and radiographic latent features

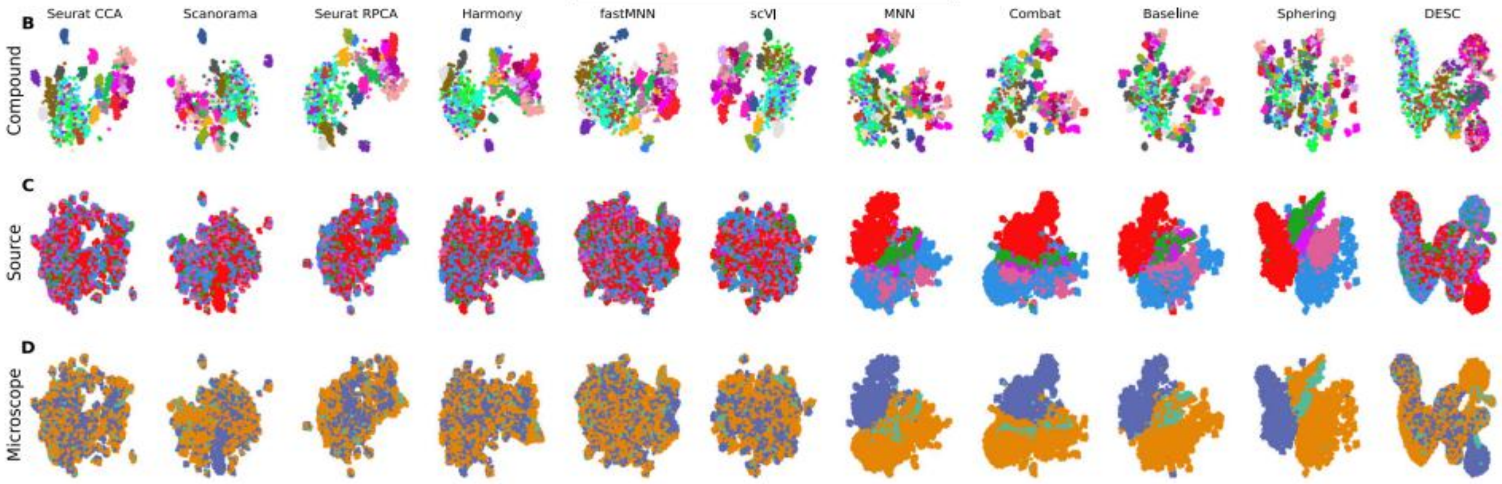
- Cell segmentation (Surface marker-based)



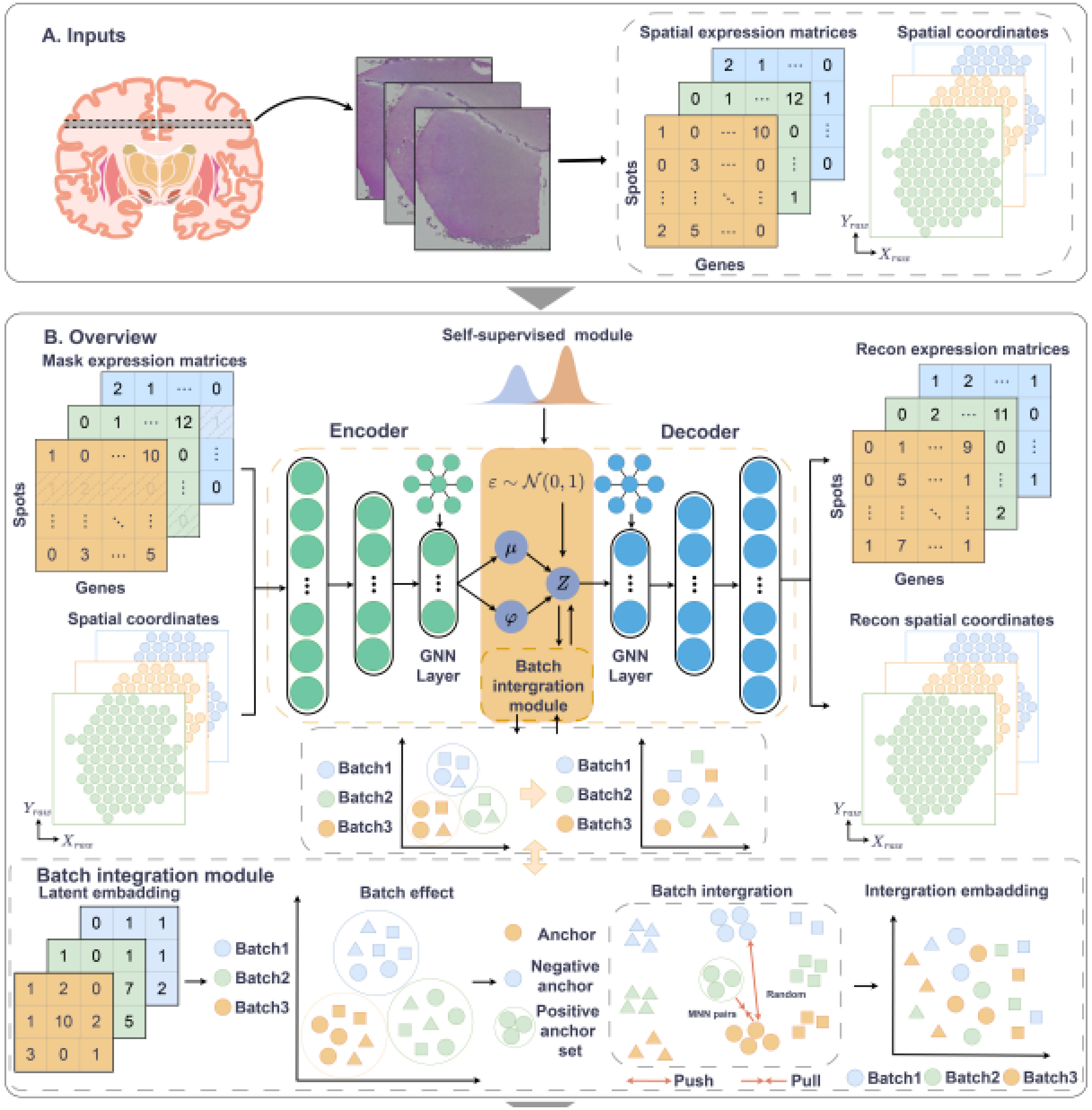
CelloType: a unified model for segmentation and classification of tissue images

- Batch correction

Method	Batch correction: Graph connectivity	Batch correction: KBET	Batch correction: LISI batch	Batch correction: Silhouette batch	Bio metrics: LISI label	Bio metrics: Leiden ARI	Bio metrics: Leiden NMI	Bio metrics: Silhouette label	Bio metrics: mAP (controls)	Bio metrics: mAP (nonrep)	Aggregate score: Batch correction	Aggregate score: Bio metrics	Aggregate score: Overall
Seurat CCA	0.35	0.13	0.49	0.85	1.00	0.01	0.25	0.31	0.27	0.07	0.46	0.32	0.37
Scanorama	0.34	0.14	0.42	0.83	1.00	0.01	0.29	0.27	0.23	0.05	0.43	0.31	0.36
Seurat RPCA	0.35	0.10	0.31	0.84	1.00	0.02	0.28	0.31	0.27	0.06	0.40	0.32	0.35
fastMNN	0.34	0.08	0.45	0.79	1.00	0.01	0.25	0.22	0.26	0.07	0.42	0.30	0.35
Harmony	0.34	0.03	0.25	0.84	1.00	0.02	0.28	0.32	0.26	0.05	0.36	0.32	0.34
scVI	0.34	0.04	0.23	0.77	1.00	0.01	0.28	0.25	0.24	0.05	0.34	0.30	0.32
Sphering	0.34	0.01	0.00	0.84	1.00	0.00	0.21	0.38	0.22	0.03	0.30	0.31	0.30
Combat	0.34	0.01	0.02	0.81	1.00	0.00	0.24	0.31	0.24	0.04	0.30	0.30	0.30
MNN	0.34	0.03	0.01	0.81	1.00	0.00	0.23	0.31	0.22	0.04	0.29	0.30	0.30
Baseline	0.34	0.01	0.01	0.80	1.00	0.00	0.23	0.31	0.21	0.03	0.29	0.30	0.29
DESC	0.33	0.01	0.27	0.58	1.00	0.02	0.34	0.01	0.22	0.04	0.30	0.27	0.28

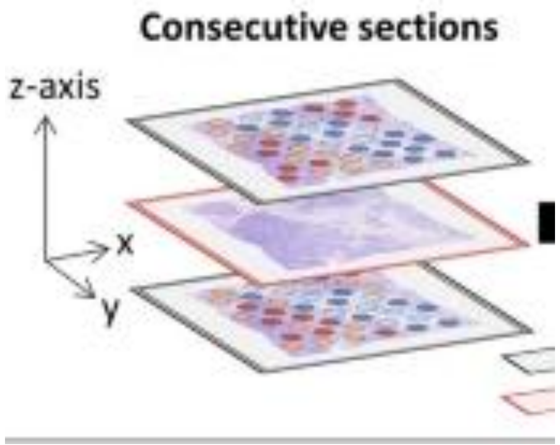


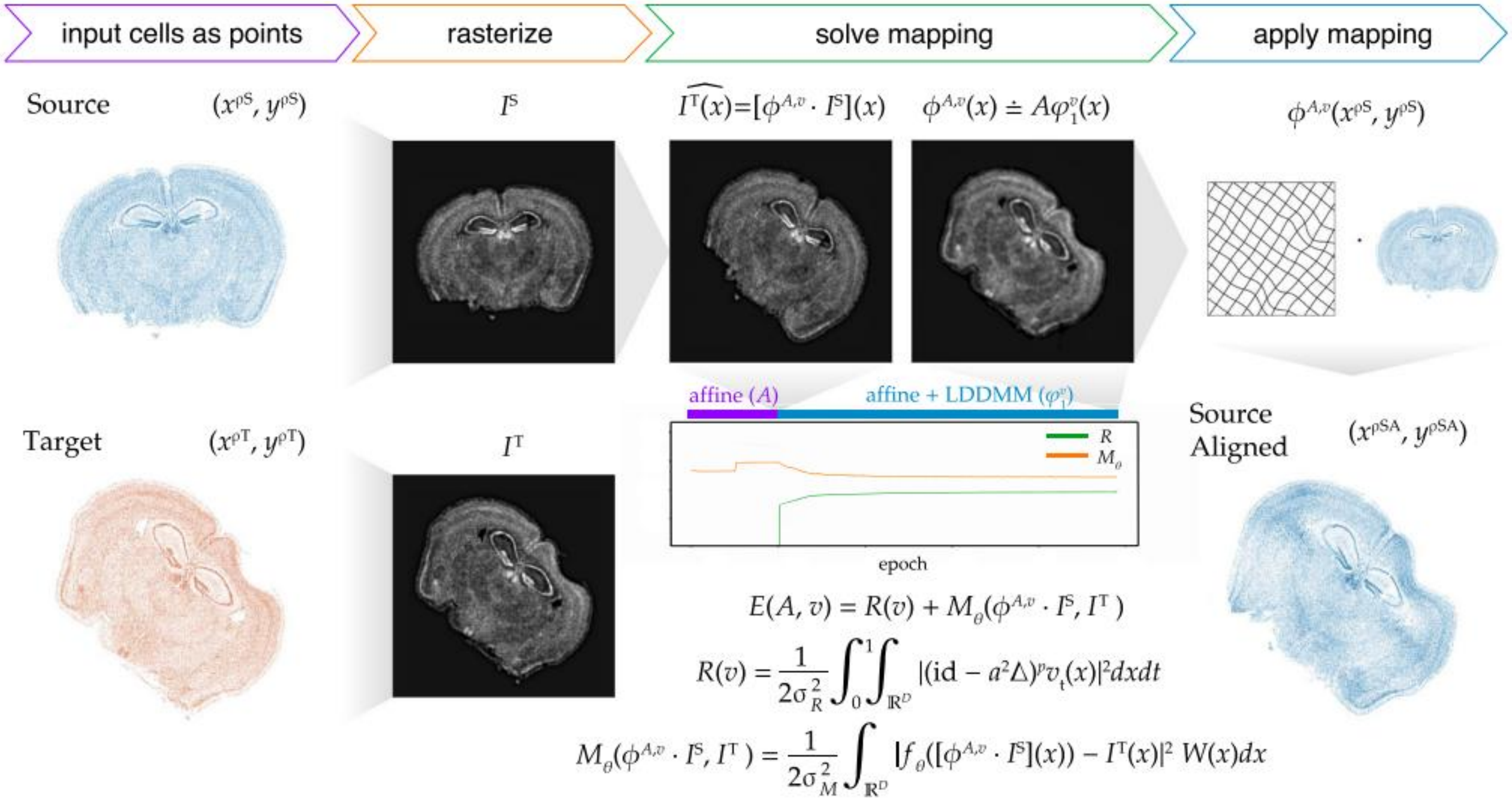
- Batch correction (SpaBatch)



SpaBatch: Batch Alignment of Spatial Transcriptomics Data using Graph Deep Learning

- Consecutive slide (STalign)



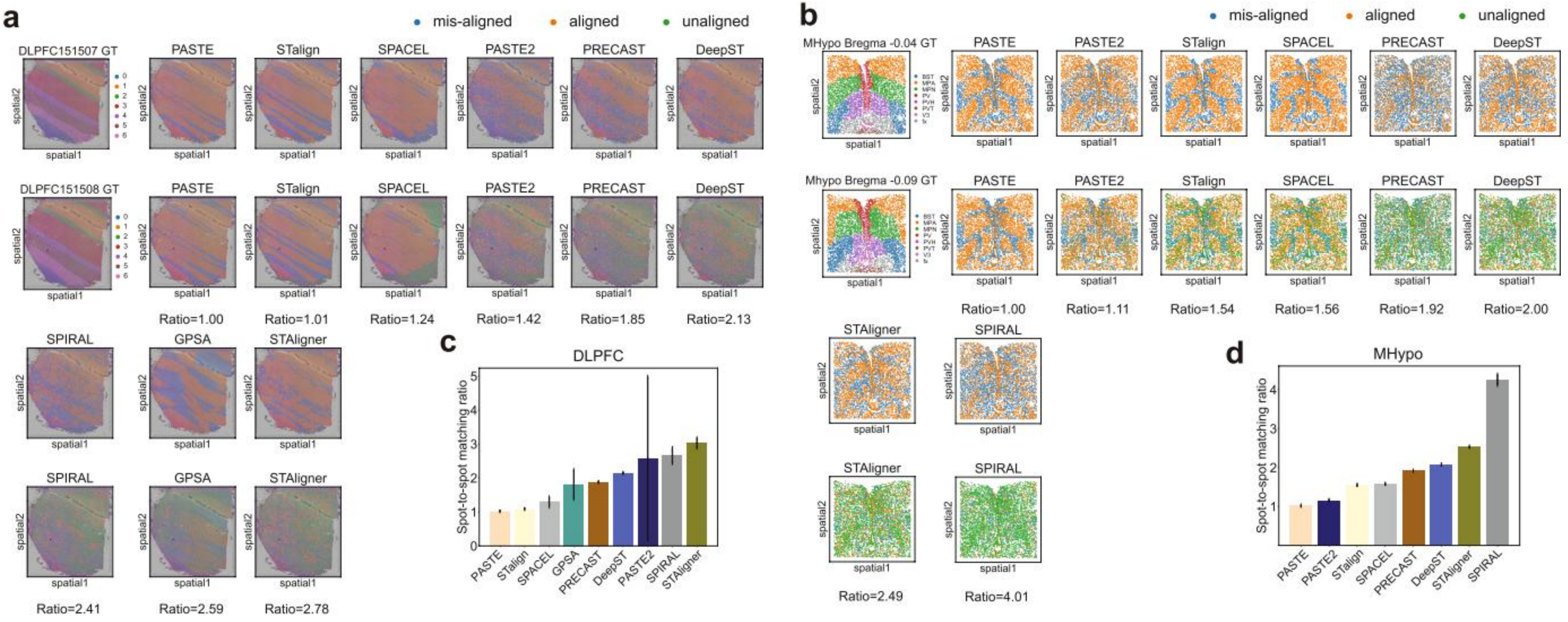


$$E(A, v) = R(v) + M_\theta(\phi^{A,v} \cdot I^S, I^T)$$

$$R(v) = \frac{1}{2\sigma_R^2} \int_0^1 \int_{\mathbb{R}^D} |(\mathrm{id} - a^2\Delta)^p v_t(x)|^2 dx dt$$

$$M_\theta(\phi^{A,v} \cdot I^S, I^T) = \frac{1}{2\sigma_M^2} \int_{\mathbb{R}^D} |f_\theta([\phi^{A,v} \cdot I^S](x)) - I^T(x)|^2 W(x) dx$$

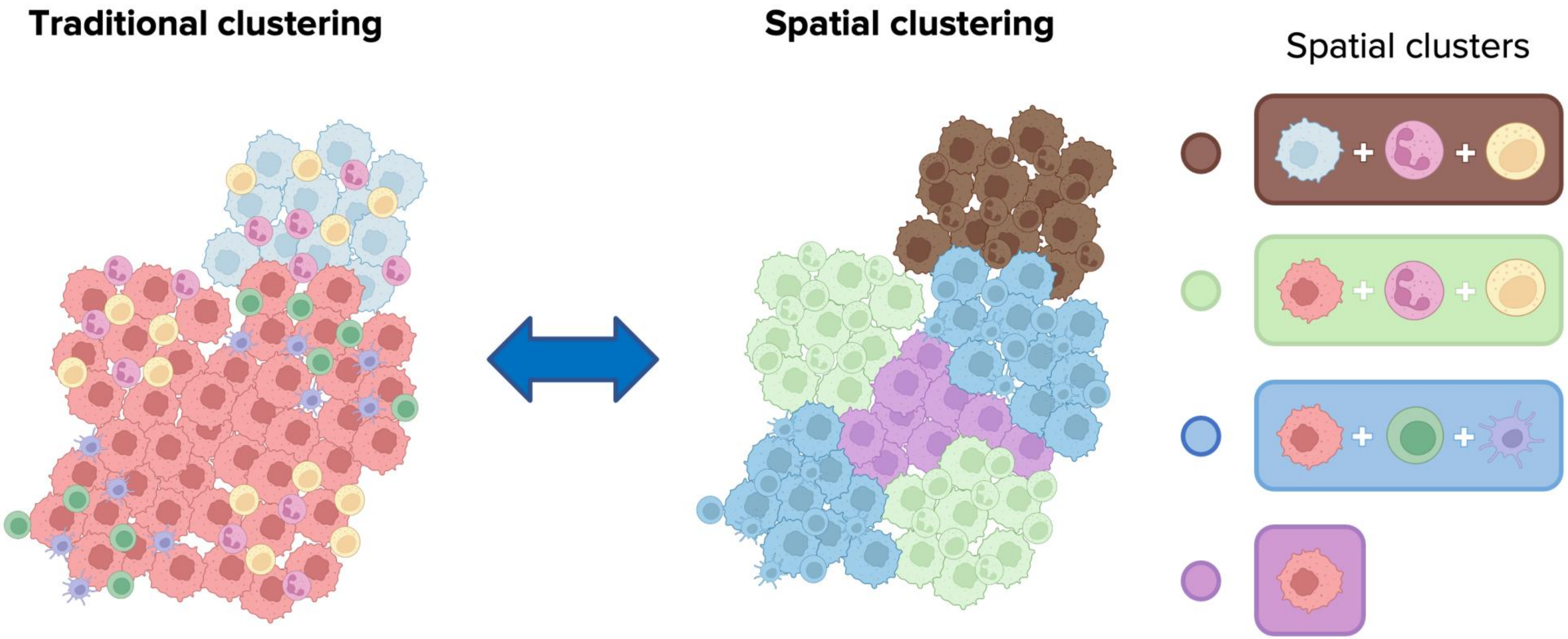
- Consecutive slide



Benchmarking clustering, alignment, and integration methods for spatial transcriptomics

- # Spatial cluster (or Neighborhood analysis)

-Clustering: gene expression only → add spatial information
-Which cells are close to each other?
-Neighborhood analysis → information of local domain
-Spatial cluster → global domain



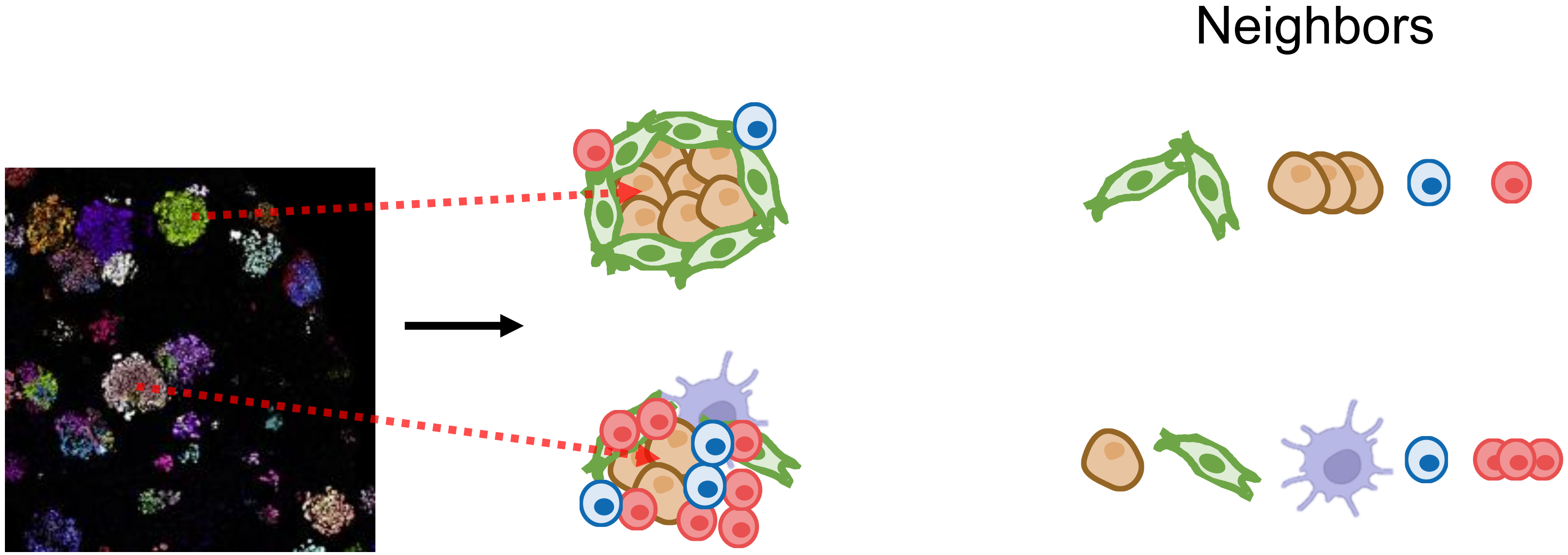
Traditional clustering

Spatial clustering

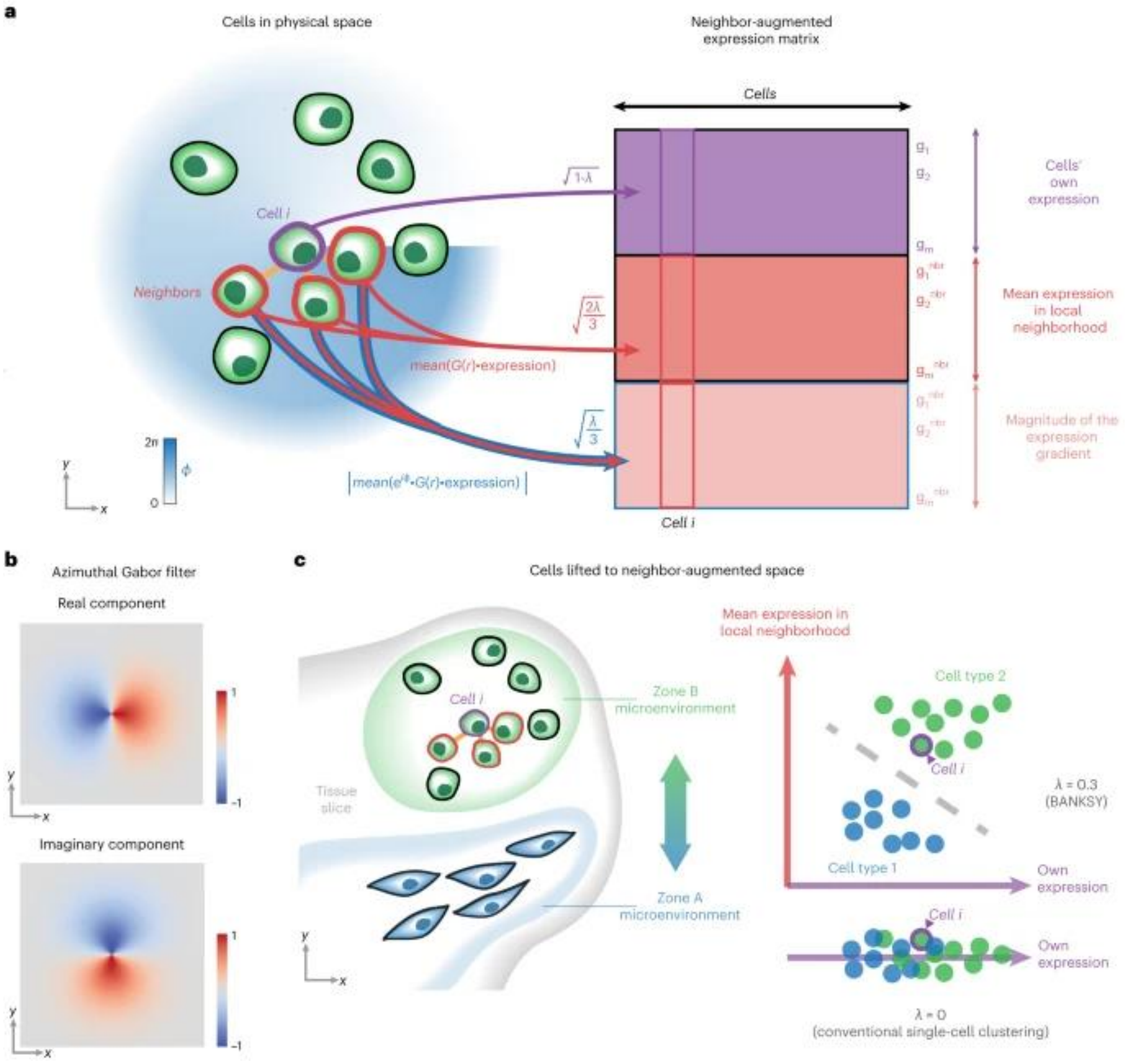
Spatial clusters

• Spatial cluster (KNN-neighbor)

-Provided from Seurat
-Group by similar neighborhood profile
-KNN for each cell
-K-mean clustering by KNN
→ Similar neighbors → same group → spatial cluster



- Spatial cluster (BANKSY)



-Gene expression clustering
+ Mean exp from local neighborhood
+ AGF (azimuthal Gabor filter)
(expression gradient)

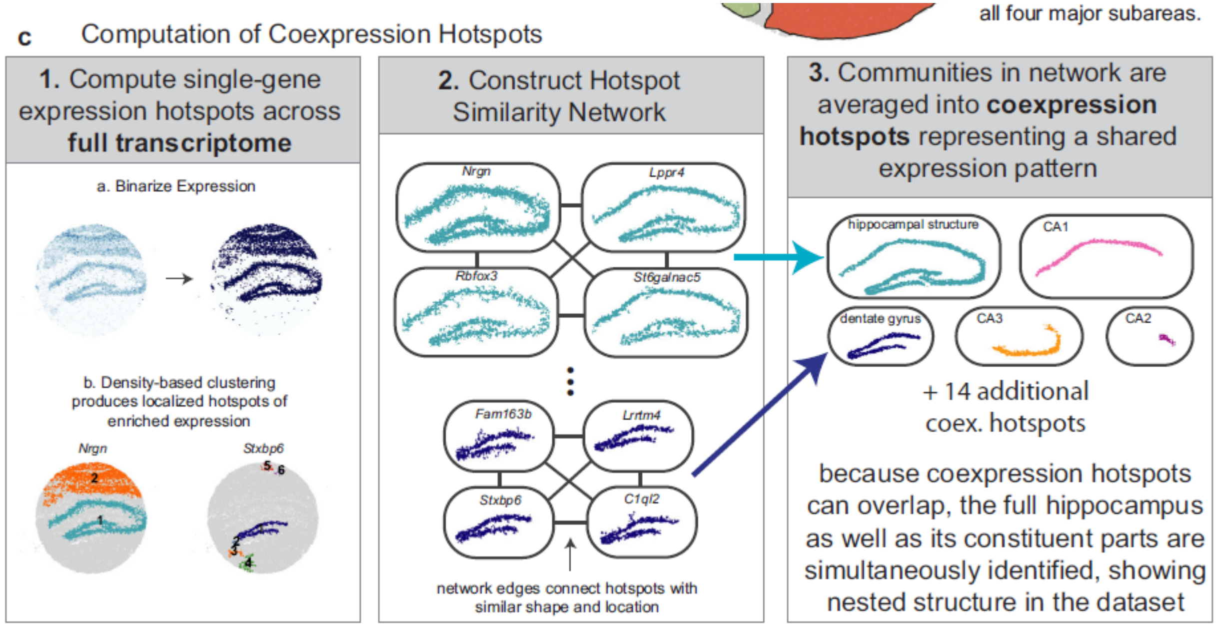
- Spatial cluster (NeST)

nature communications

Article

https://doi.org/10.1038/s41467-023-42343-x

NeST: nested hierarchical structure identification in spatial transcriptomic data



all four major subareas.

c Computation of Coexpression Hotspots

1. Compute single-gene expression hotspots across **full transcriptome**

2. Construct Hotspot Similarity Network

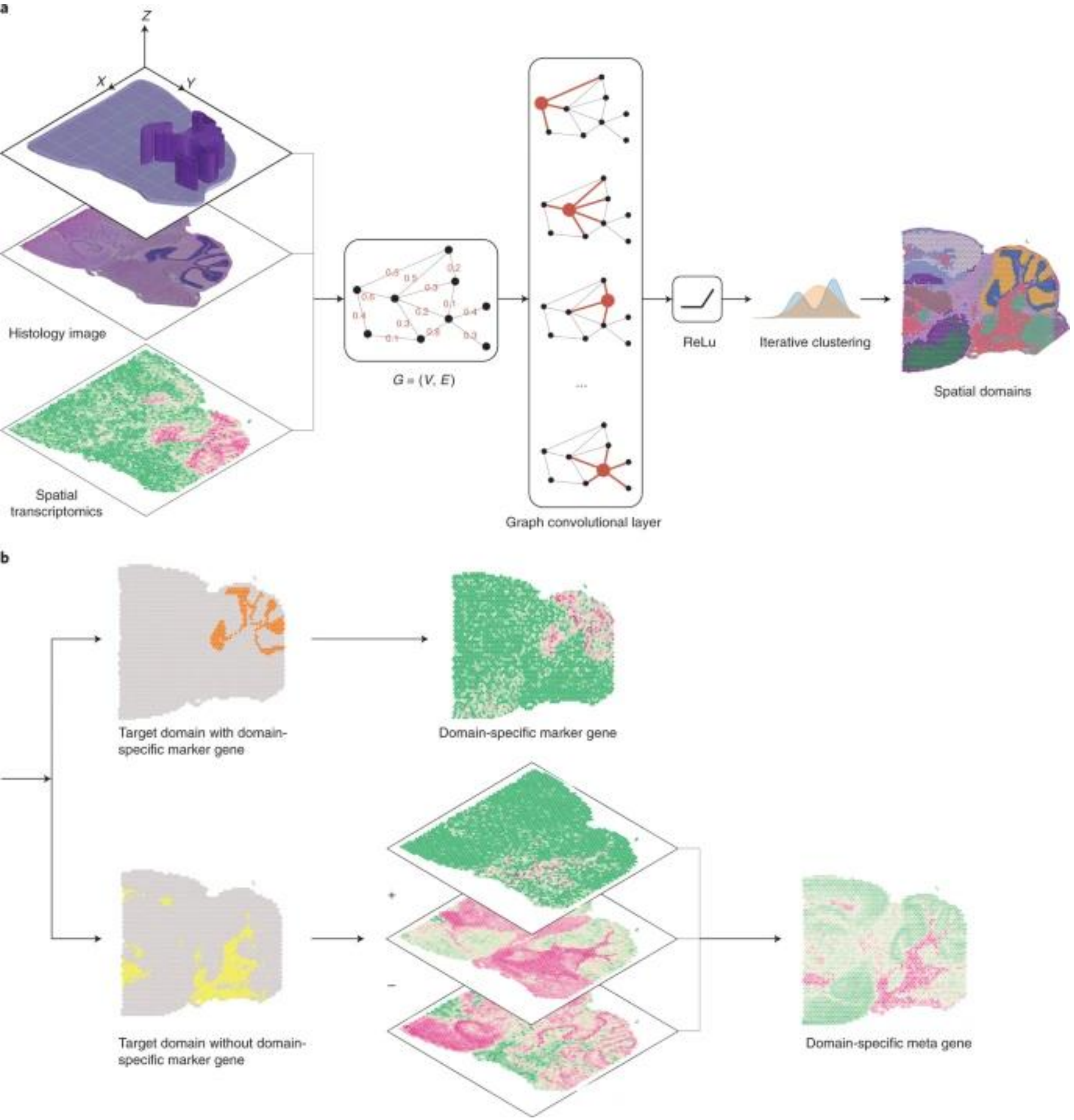
3. Communities in network are averaged into **coexpression hotspots** representing a shared expression pattern

because coexpression hotspots can overlap, the full hippocampus as well as its constituent parts are simultaneously identified, showing nested structure in the dataset

1:Binarize each gene by expression for each spot
2:cluster (DBSCAN) each spot by binarized gene exp → hotspot
3: hotspot overlap between different genes (jaccard) by location
→ Which genes have a high similarity (of hotspot spatial structure)
→ (gene-gene) Network → leiden clustering
→ community detection (coexpression)
→ Various structure!

- Could be extended to cell-cell interaction score

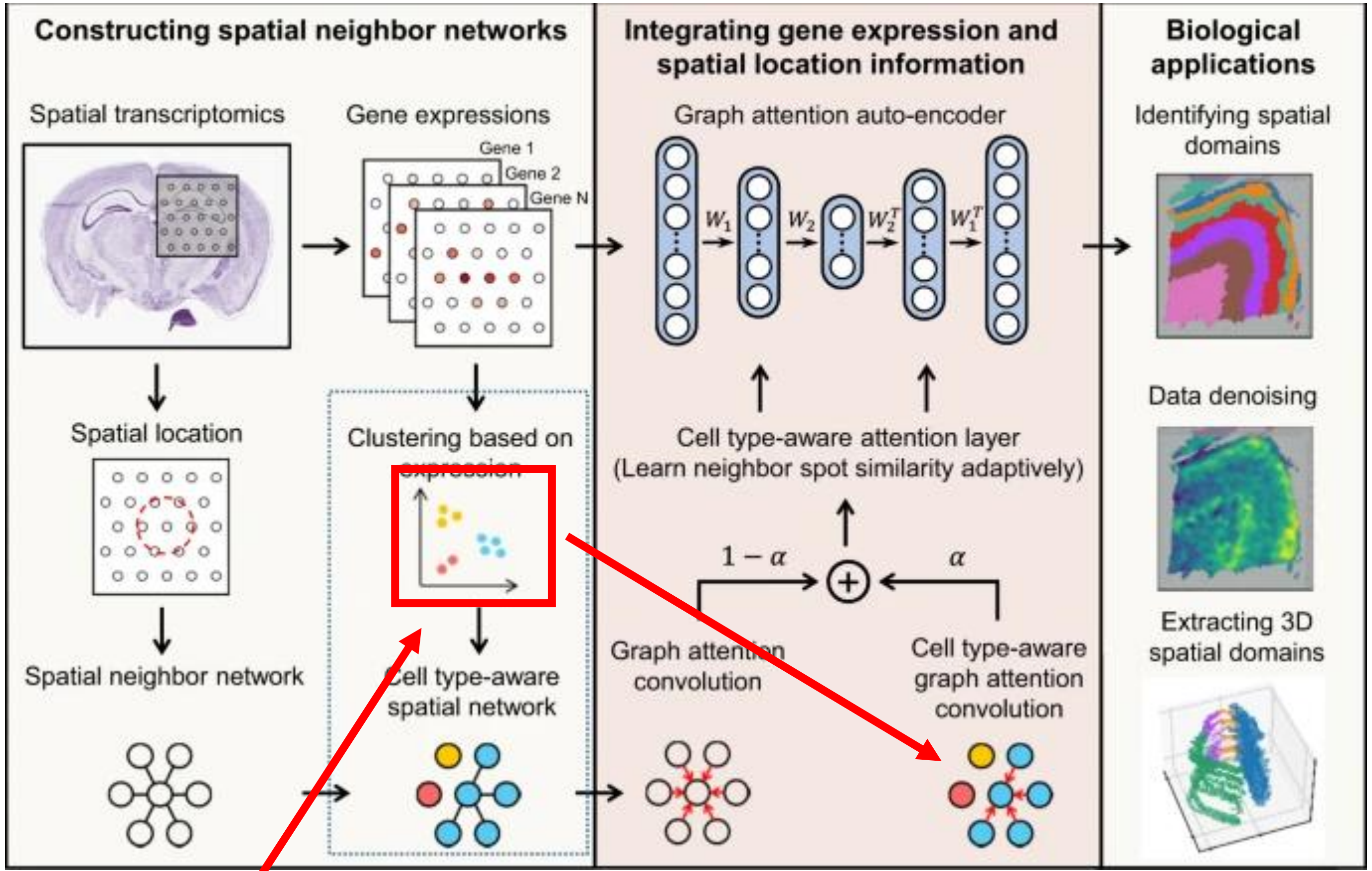
- Spatial cluster (SpaGCN)



-Node: each cell
-Node weight: gene expression
-Graph (edge)→ spatial location

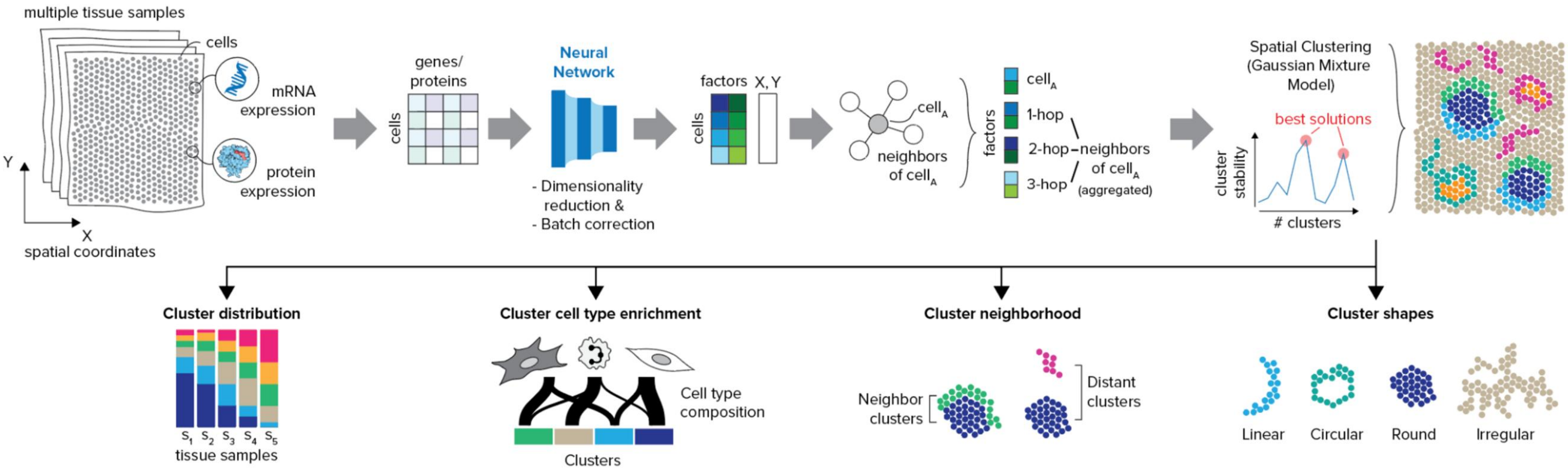
→ Graph convolutional network (GCN)
→ Modified gene expression (aggregated)
→ Clustering

• Spatial cluster (STAGATE)



-VAE-based
-Gene expression → GAN by local neighbor
*use pre-clustering result
→ Use latent space for clustering (ex: Louvain)

• Spatial cluster (CellCharter)

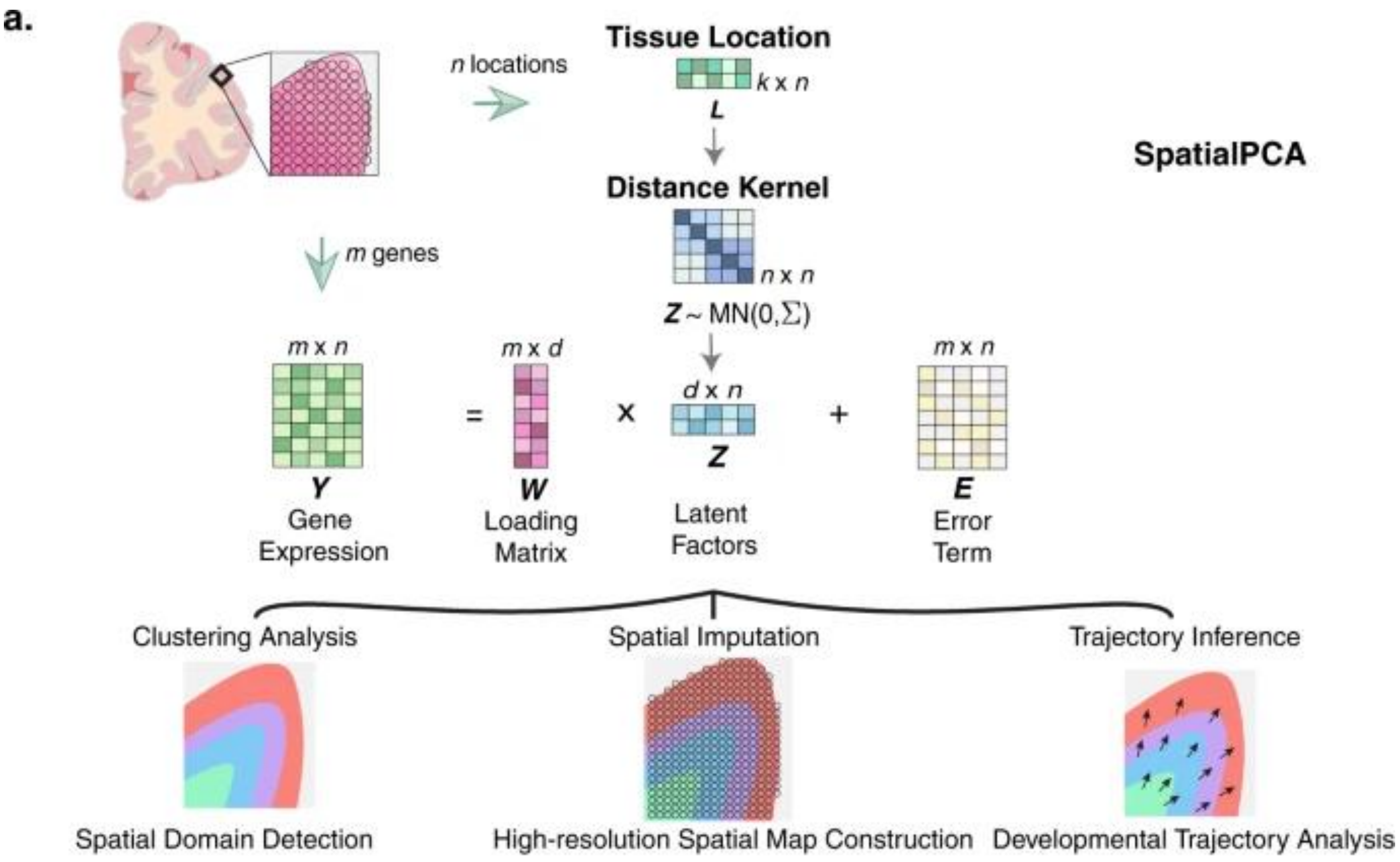


-VAE-based

-Gene expression → weighted-graph by local neighbor

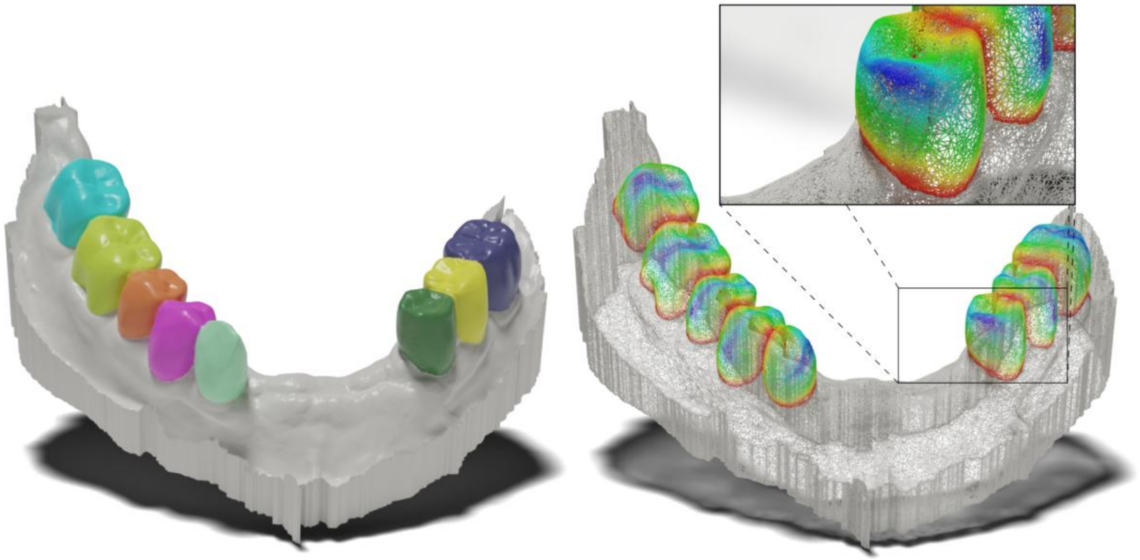
→ Clustering by GMM (Gaussian mixture model)

Spatial cluster (SpatialPCA)

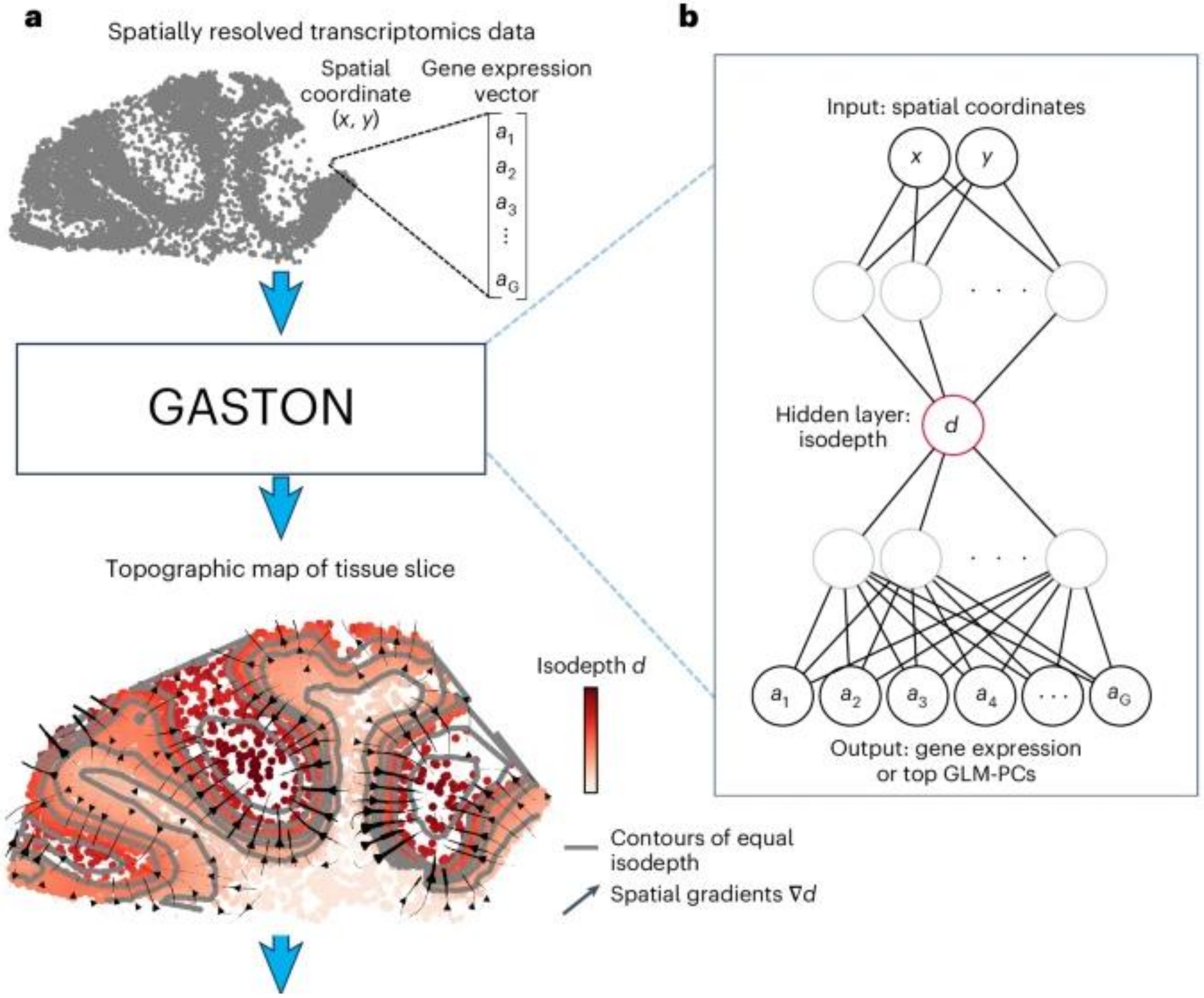


-PCA-based → Singular value decomposition
Tissue location → distance kernel (spatially correlated information)
Gene exp = W * Z + E (error)
Z: sampling from N(0, kernel) → latent space → clustering …

- Spatial cluster (GASTON)



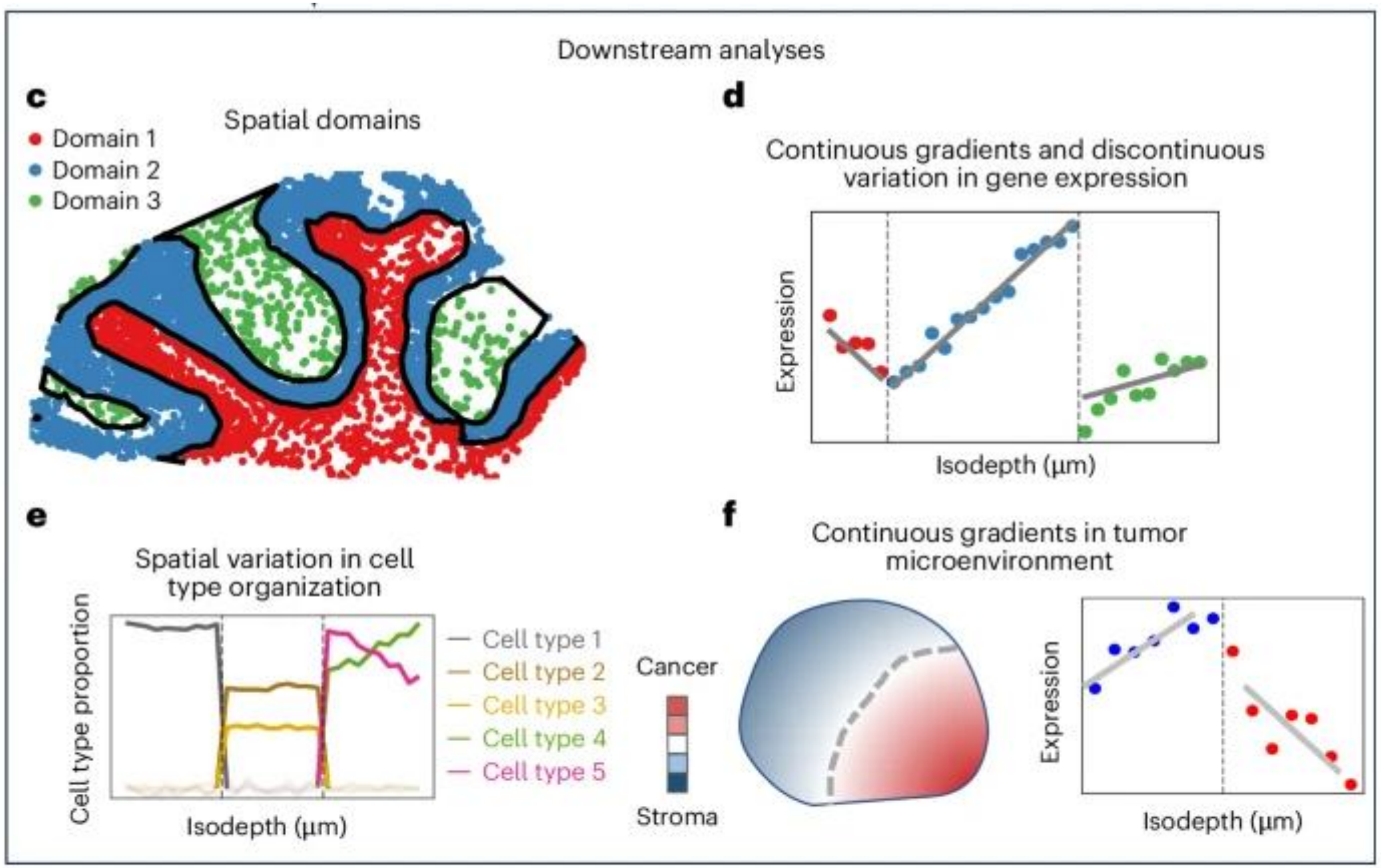
(**a**) Generated segmentation regions and wireframe with colored edge distance for tooth regions



-Topological gradient

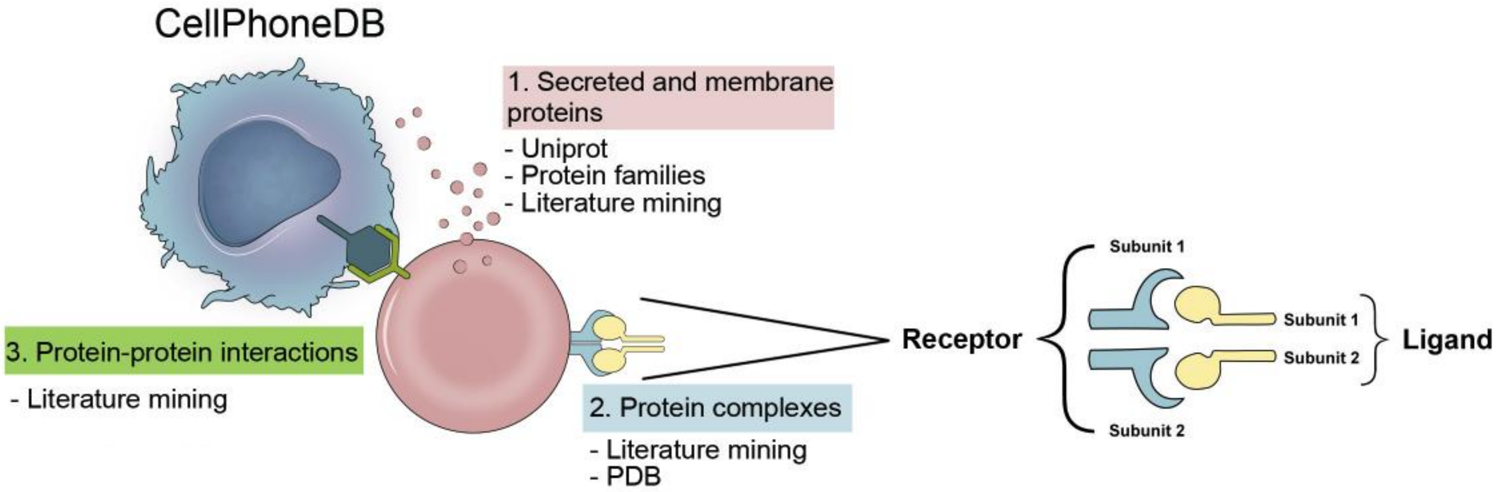
→ Gene expression difference → gradient

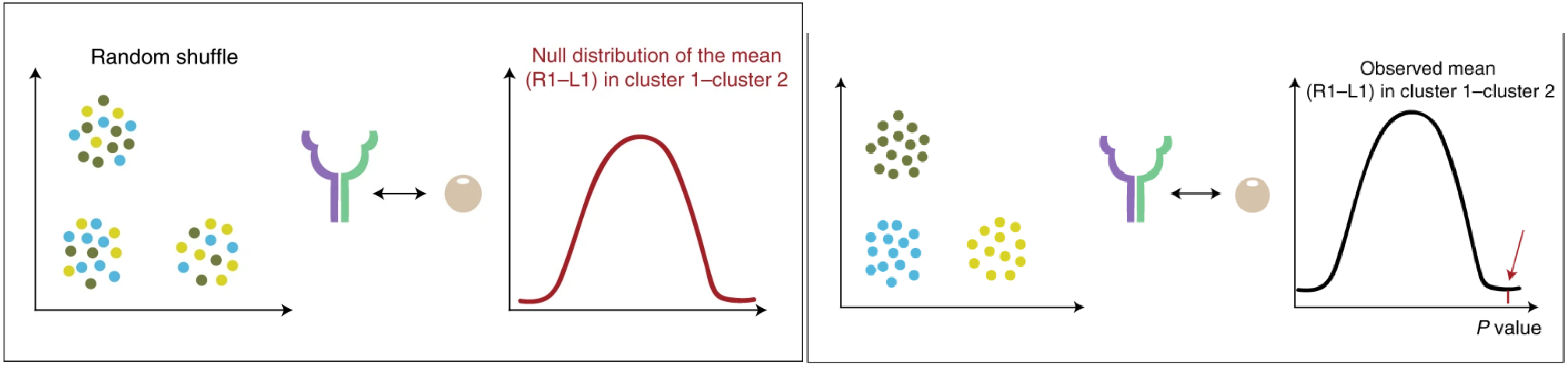
- Spatial cluster (GASTON)



Cell-Cell interaction

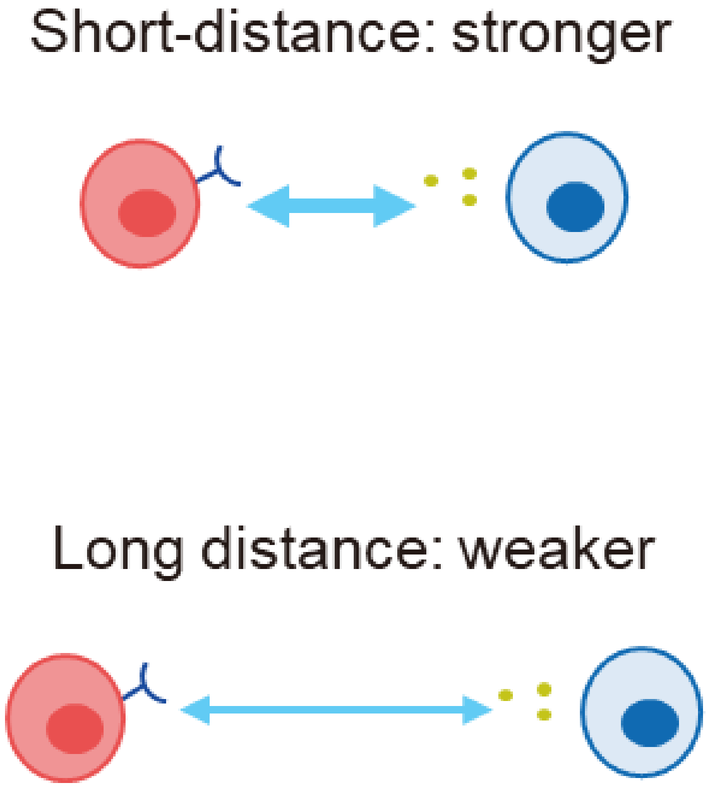
-scRNA-seq: based on gene expression





- Cell-Cell interaction

-Spatial constraint → distance should be a “cost”
Why?
→ Signaling strength ~ Binding affinity
→ Binding affinity ~ ligand concentration
→ Ligand concentration (at the receptor) ~ Diffusion ~ distance



- COMMOT


nature COMMUNICATIONS

ARTICLE

Check for updates

https://doi.org/10.1038/s41467-020-15968-5 OPEN

Inferring spatial and signaling relationships between cells from single cell transcriptomic data

Zixuan Cang[1,3] & Qing Nie[1,2,3]


nature methods

Article https://doi.org/10.103

Screening cell–cell communication in spatial transcriptomics via collective optimal transport

*Optimal transport-based

-C(x,y): cost function to move x → y

-KL-divergence: compare the distribution between x and y

-Wasserstein distance between distributions

$$\min_{T} \int_{X \times Y} c(x,y)\, dT(x,y)$$

- COMMOT

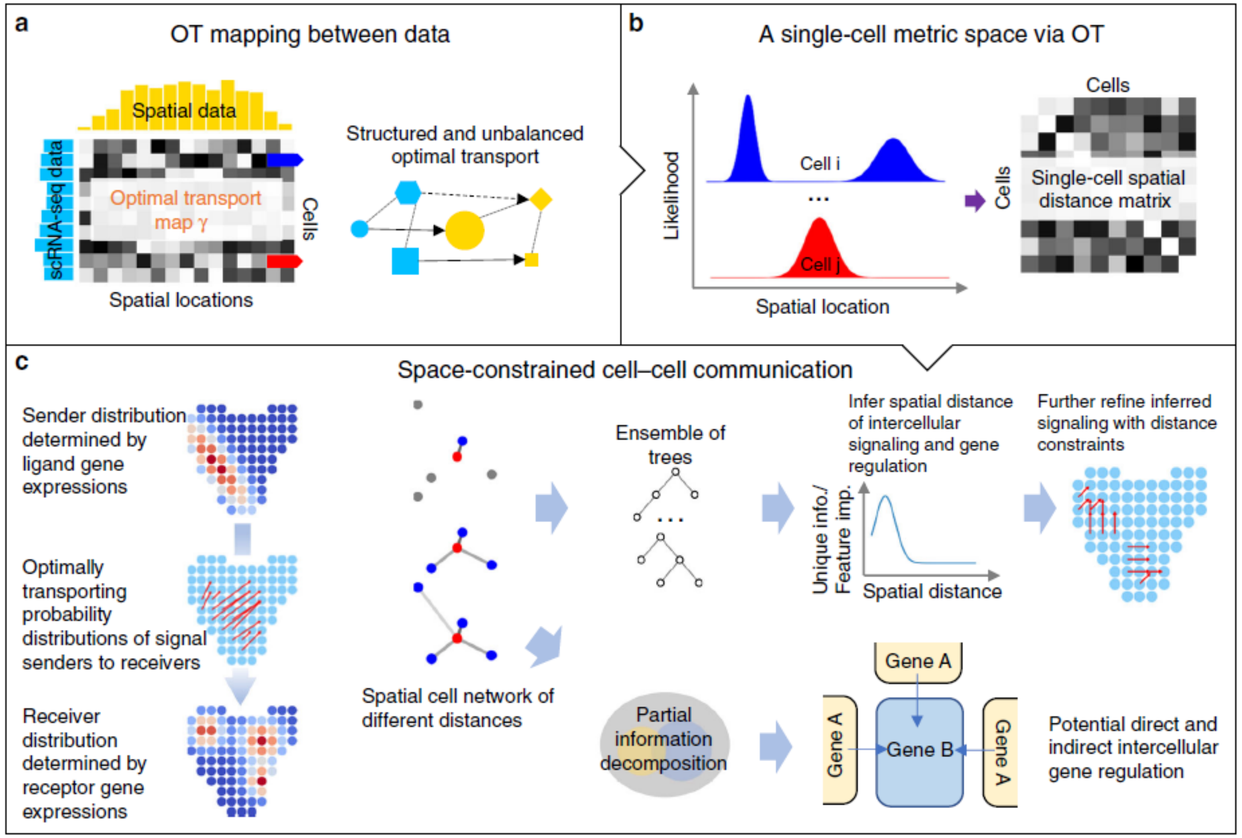


Fig. 1 Overview of SpaOTsc. a The unbalanced transport relaxes the mass conservation constraint (e.g. lines between circles), and the structured transport utilizes additional information (e.g. dotted links) to refine the mapping (e.g. blue hexagon). **b** Cell–cell distance is inferred by computing optimal transport distance of the spatial probability distributions of cells (rows of $\boldsymbol{\gamma}$ in **a**). **c** Calculated cell–cell distance, along with partial information decomposition and random forest models, is used to infer spatial distance of signaling and then construct space-constrained cell–cell communications and identify potential intercellular regulation between genes.

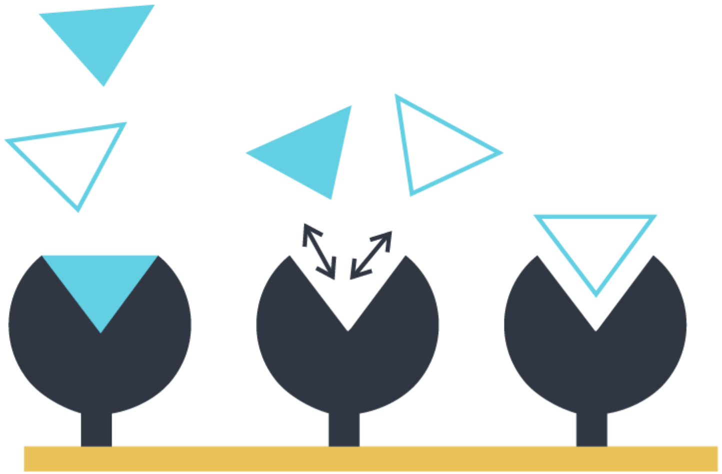
$$\underset{\boldsymbol{\gamma}\in\mathbb{R}_+^{n\times m}}{\operatorname{argmin}}\Big[(1-\alpha)<\boldsymbol{\gamma},\mathbf{M}>_{\mathrm{F}} + \rho\big(\mathrm{KL}(\boldsymbol{\gamma}\mathbf{1}^m|\boldsymbol{\omega}_1)+\mathrm{KL}(\boldsymbol{\gamma}^T\mathbf{1}^n|\boldsymbol{\omega}_2)\big) + \alpha\sum_{i,j,k,l}L(\mathbf{D}_{\mathrm{sc}}(i,k),\mathbf{D}_{\mathrm{spa}}(j,l))\gamma_{i,j}\gamma_{k,l}\Big]$$

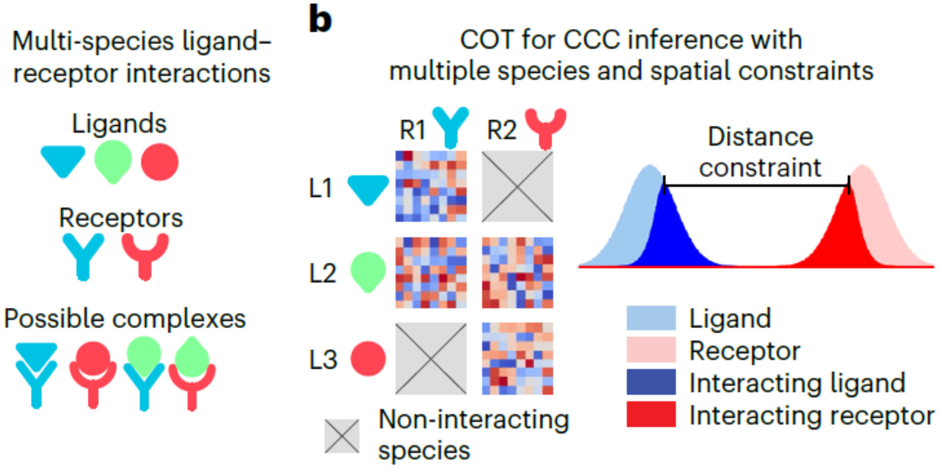
*minimize cost function
(ligand → receptor)

$$\underset{\boldsymbol{\gamma}\in\mathbb{R}_+^{n\times n}}{\operatorname{argmin}}<\boldsymbol{\gamma},\widehat{\mathbf{D}}_{\mathrm{sc}}>_{\mathrm{F}} + \rho\big(\mathrm{KL}(\boldsymbol{\gamma}\mathbf{1}^n|\boldsymbol{\omega}_{\mathrm{L}})+\mathrm{KL}(\boldsymbol{\gamma}^T\mathbf{1}^n|\boldsymbol{\omega}_{\mathrm{D}})\big).$$

-D: distance
-KL: penalty if the distribution between ligand and receptor are too different

- COMMOT





-Competition between different ligands

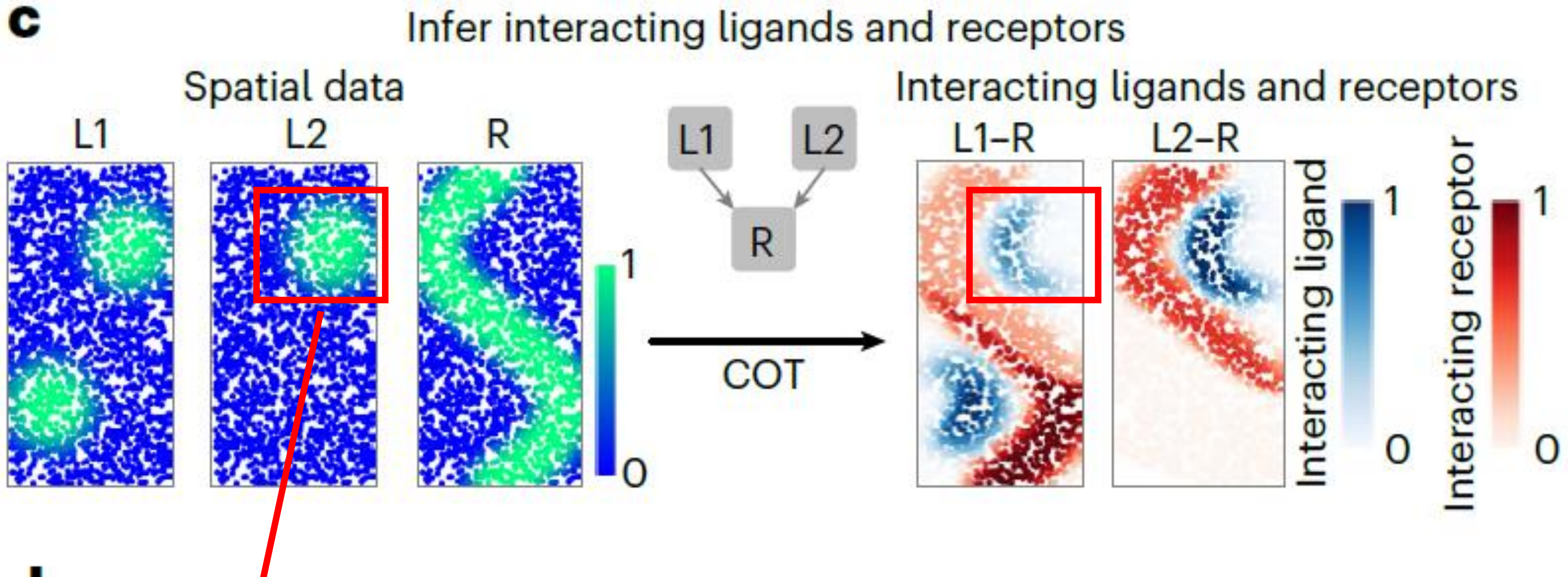
$$\min_{\boldsymbol{P}\in\Gamma} \sum_{(i,j)\in I} \langle \boldsymbol{P}_{i,j,\cdot,\cdot}, \boldsymbol{C}_{(i,j)} \rangle_F + \sum_i F(\mu_i) + \sum_j F(\nu_j),$$

$$\Gamma = \left\{ \boldsymbol{P} \in \mathbb{R}_+^{n_l \times n_r \times n_s \times n_s} : \boldsymbol{P}_{i,j,\cdot,\cdot} = \boldsymbol{0} \text{ for } (i,j) \notin I, \sum_{j,l} \boldsymbol{P}_{i,j,k,l} \le \boldsymbol{X}^L_{i,k}, \sum_{i,k} \boldsymbol{P}_{i,j,k,l} \le \boldsymbol{X}^R_{j,l} \right\}$$

$$\mu_i(k) = \boldsymbol{X}^L_{i,k} - \sum_{j,l} \boldsymbol{P}_{i,j,k,l},\ \nu_j(l) = \boldsymbol{X}^R_{j,l} - \sum_{i,k} \boldsymbol{P}_{i,j,k,l}$$

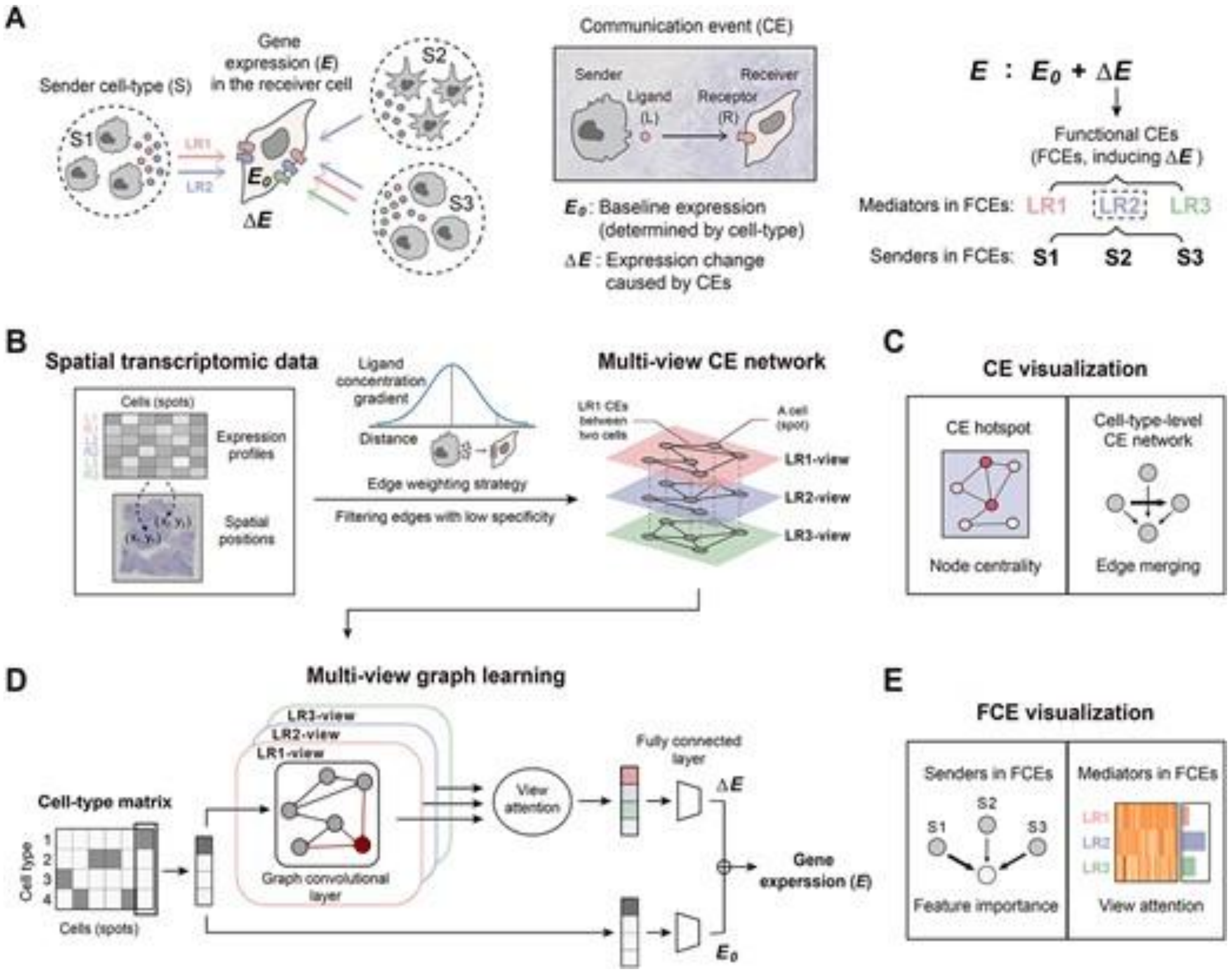
-Optimal transport plan: multi ligand-receptor coupling (C == Dsc)
F: i,k (ligand) – coupling (P:J,L) → untransported term
Total sum **receiver** P < **ligand** K spot

- COMMOT



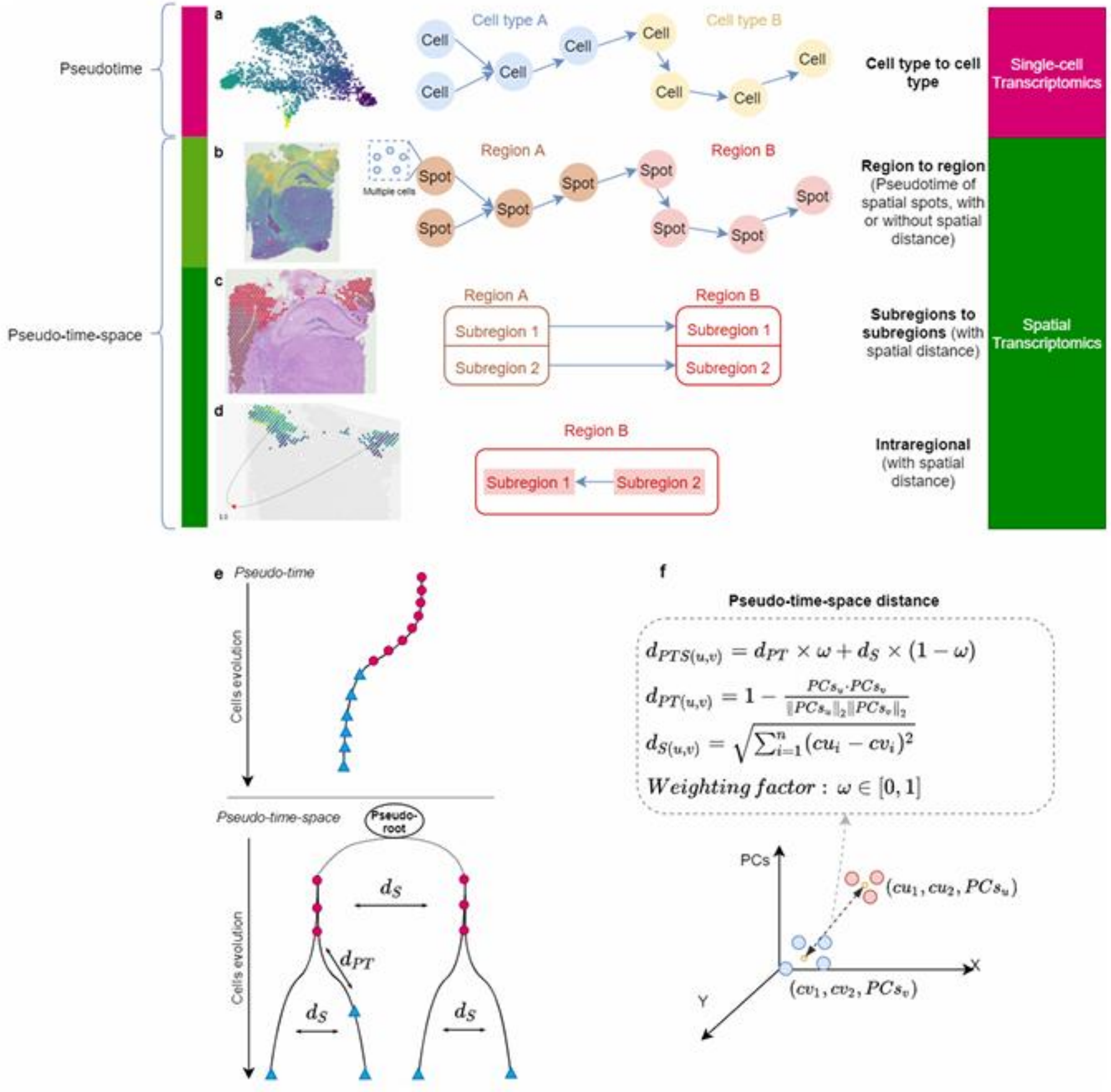
Competition with L1

- HoloNet



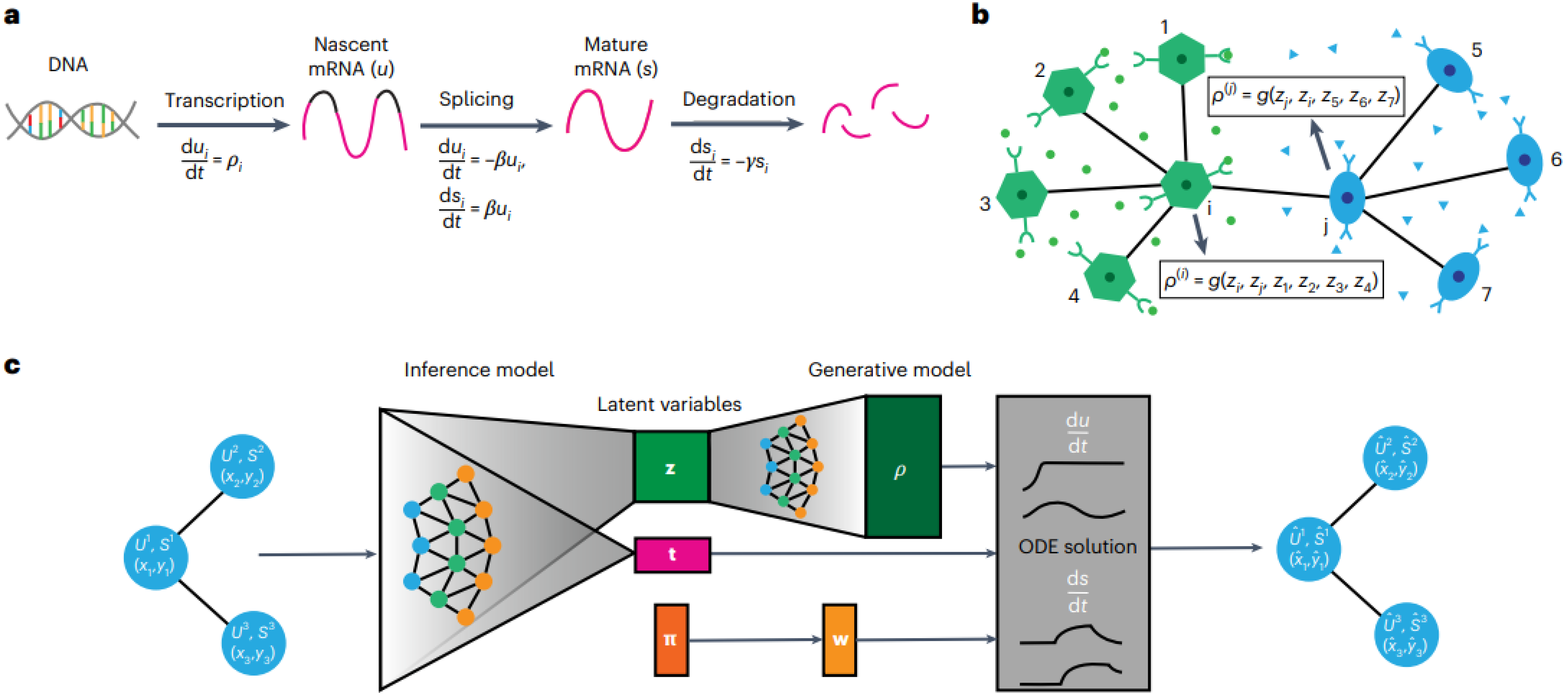
-Multi-view graph model
-Each graph (GCN): cell-cell interaction
-Multi-view: Different interactions
-which has higher weight

Trajectory analysis (stLearn)



-Gene expression-based trajectory method
-Root cell (spot): user-defined or CytoTrace
-PAGA graph by gene exp + proximity

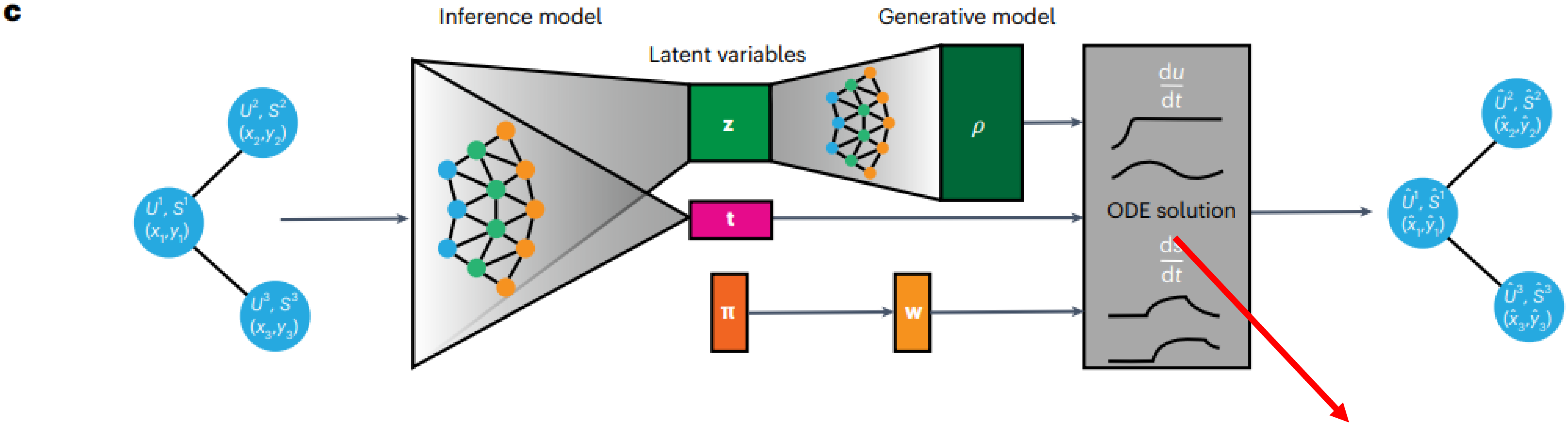
- Trajectory analysis (TopoVelo)



-RNA-velocity-based (splicing ratio)

-Borrow information from neighboring cells (influence each other)

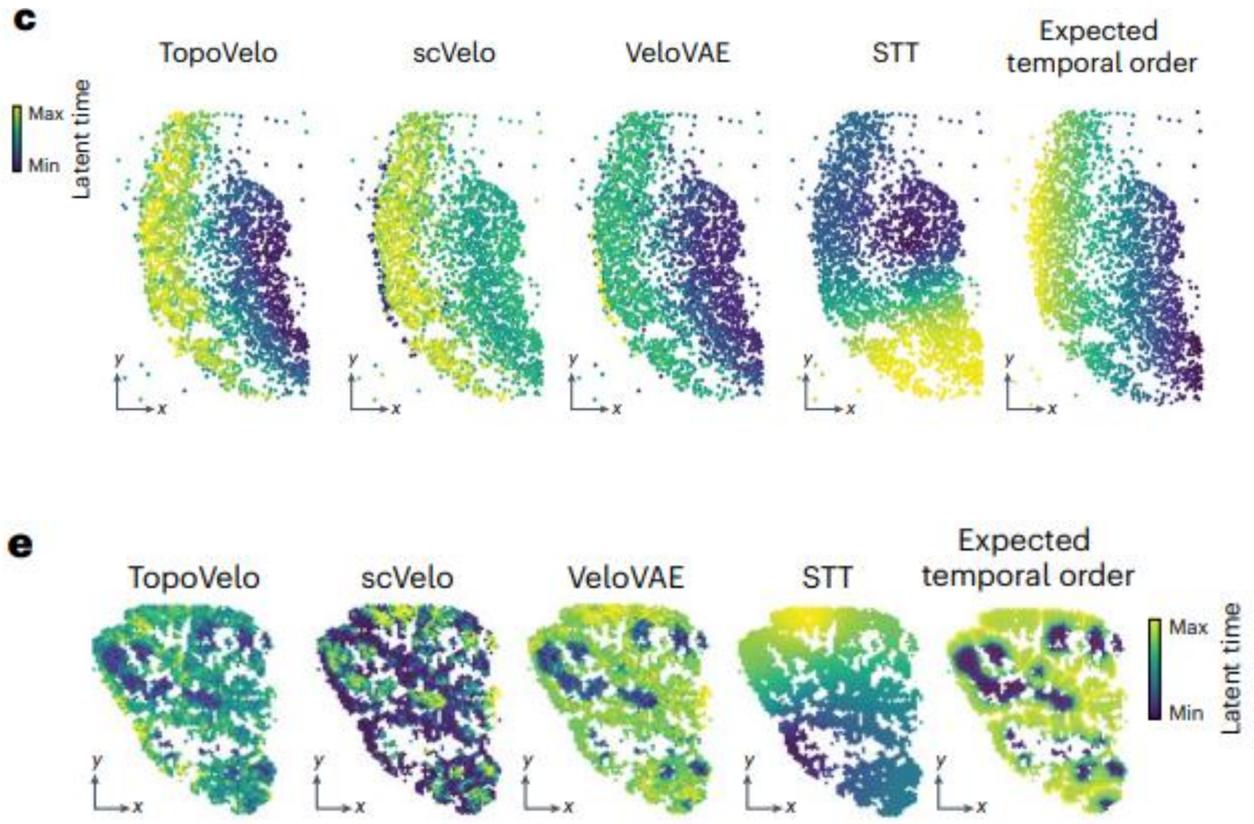
• Trajectory analysis (TopoVelo)



-VAE architecture
-Each cell (splicing ratio; u & s)
-GNN (graph input); edge: spatial proximity
-ρ: transcriptional rate
-t: time (we want this)
-ω: gene phase (→ for splicing level)
→ training: reconstruct “u & s” by ρ, t, ω

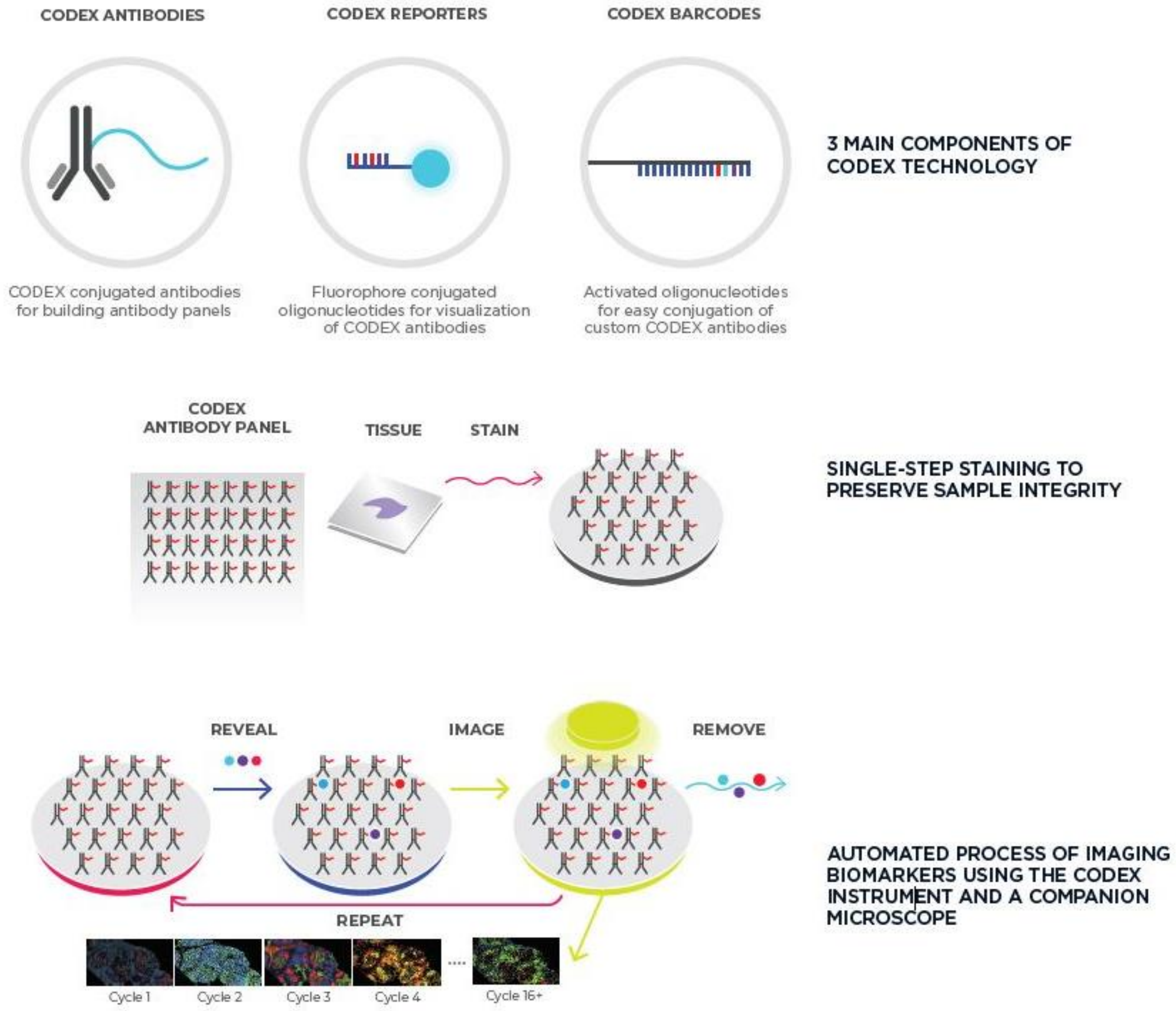
-Differentiation equation for splicing

- Trajectory analysis (TopoVelo)



- Other modalities

-CODEX (Akoya) → ~100 protein



CODEX conjugated antibodies for building antibody panels

Fluorophore conjugated oligonucleotides for visualization of CODEX antibodies

Activated oligonucleotides for easy conjugation of custom CODEX antibodies

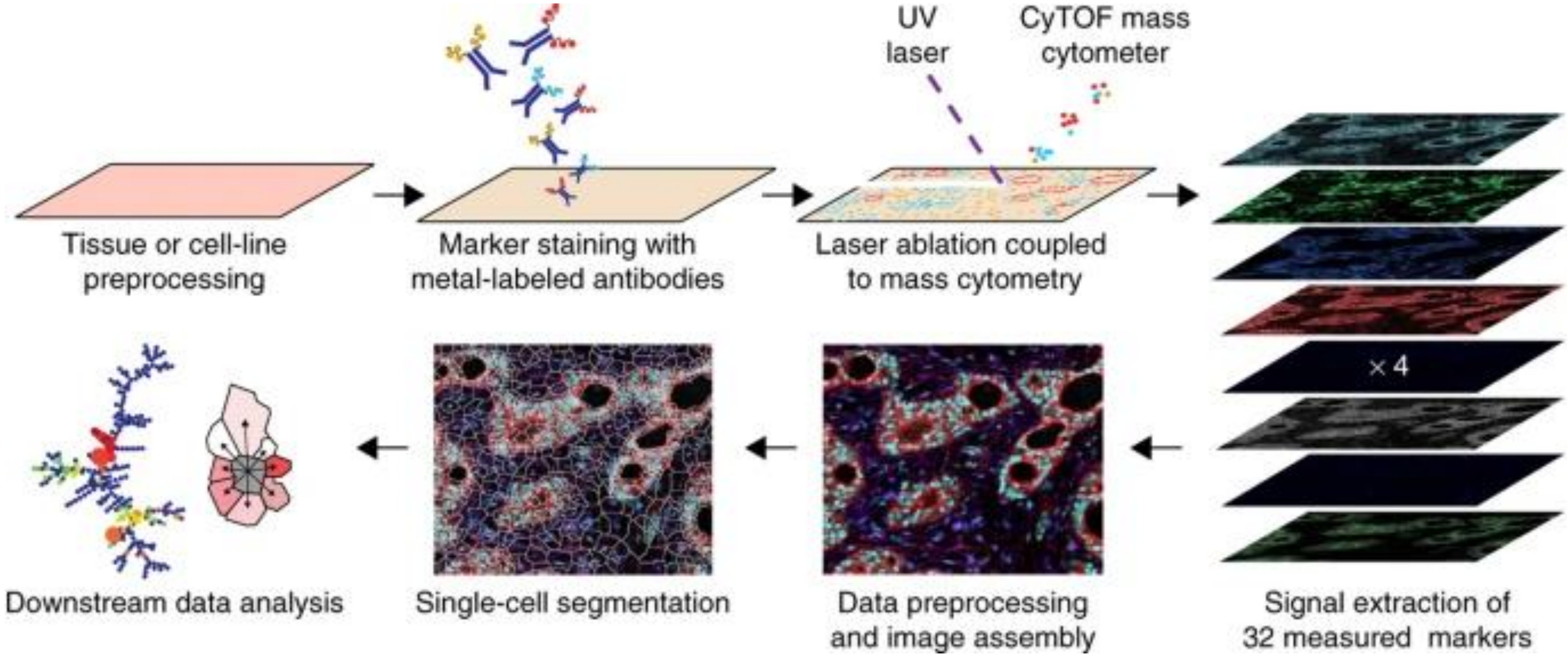
3 MAIN COMPONENTS OF CODEX TECHNOLOGY

SINGLE-STEP STAINING TO PRESERVE SAMPLE INTEGRITY

AUTOMATED PROCESS OF IMAGING BIOMARKERS USING THE CODEX INSTRUMENT AND A COMPANION MICROSCOPE

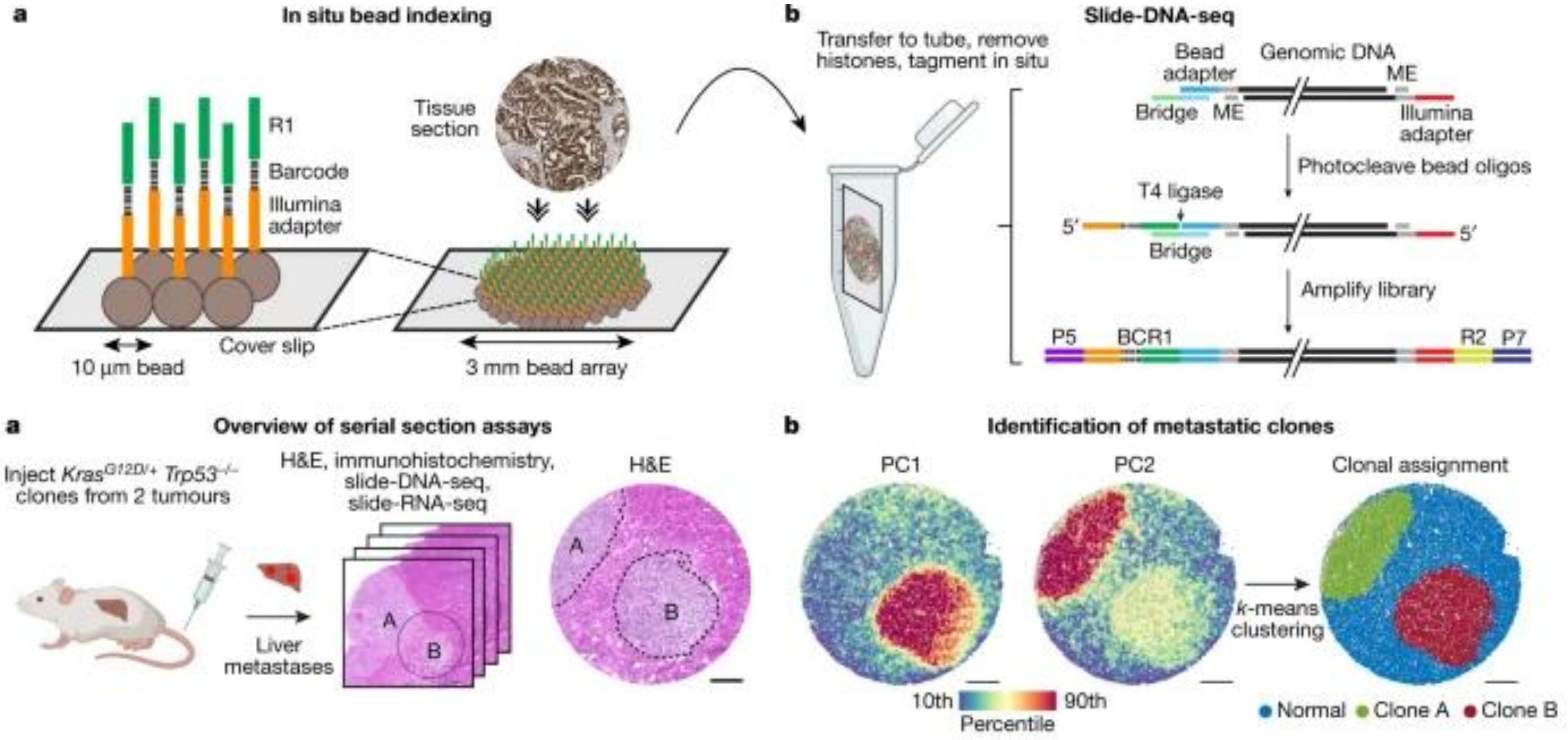
- Other modalities

-Spatial CyTOF (Mass spectrometry)
-Protein (~40 ??)
-Antibody-tagged metal → photo cleavage → TOF



• Other modalities

-Spatial genomics
-DNA → CNA, SNV (clonotyping) → evolutionary analysis

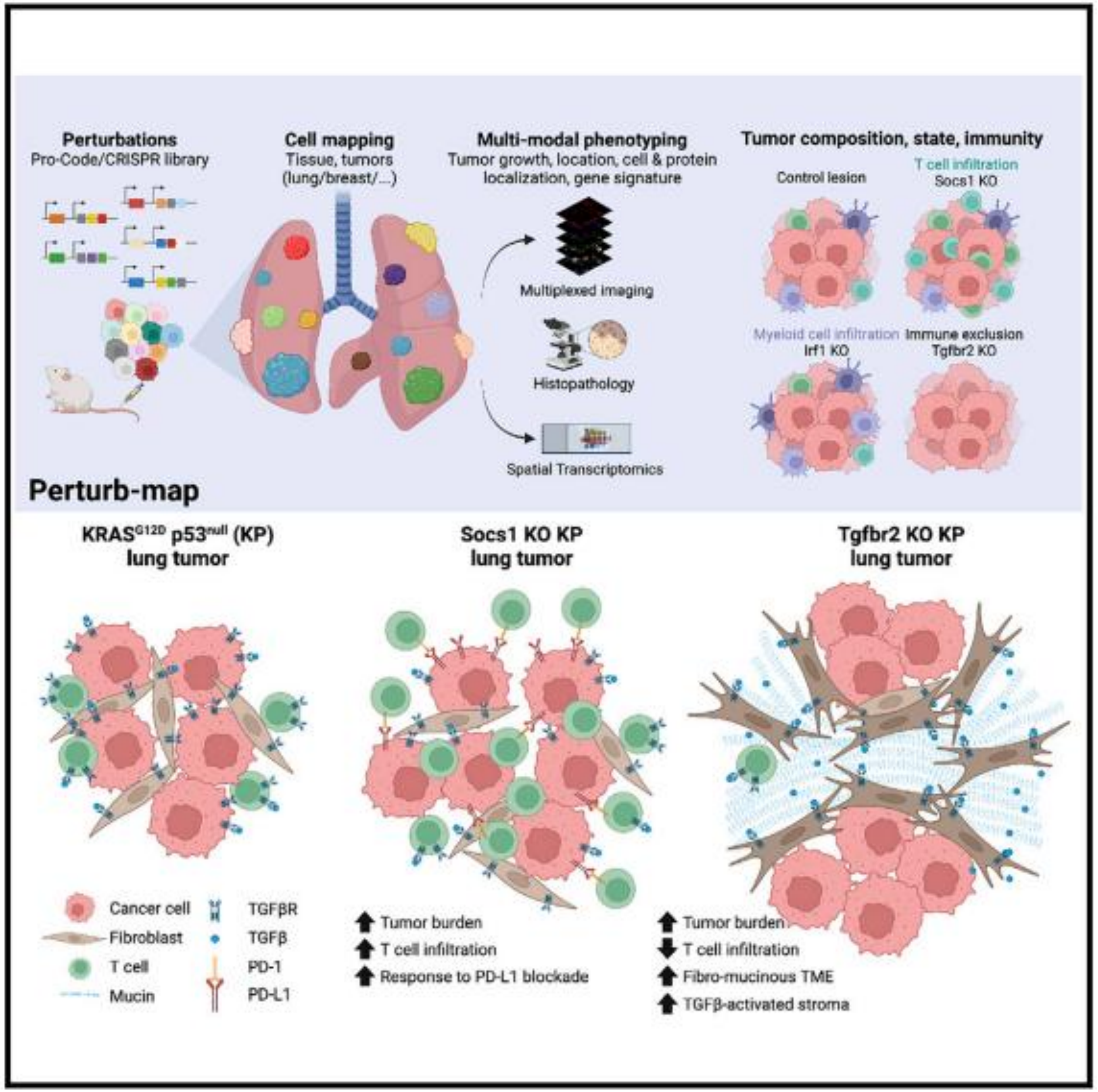


Spatial genomics enables multi-modal study of clonal heterogeneity in tissues

• Other modalities

-Perturbation
-CRISPR-screening → with spatial barcode
-Neighborhood analysis of different gene KO

TGFBR KO of cancer
→ Activated stroma cell + low T cell infiltration



Spatial CRISPR genomics identifies regulators of the tumor microenvironment