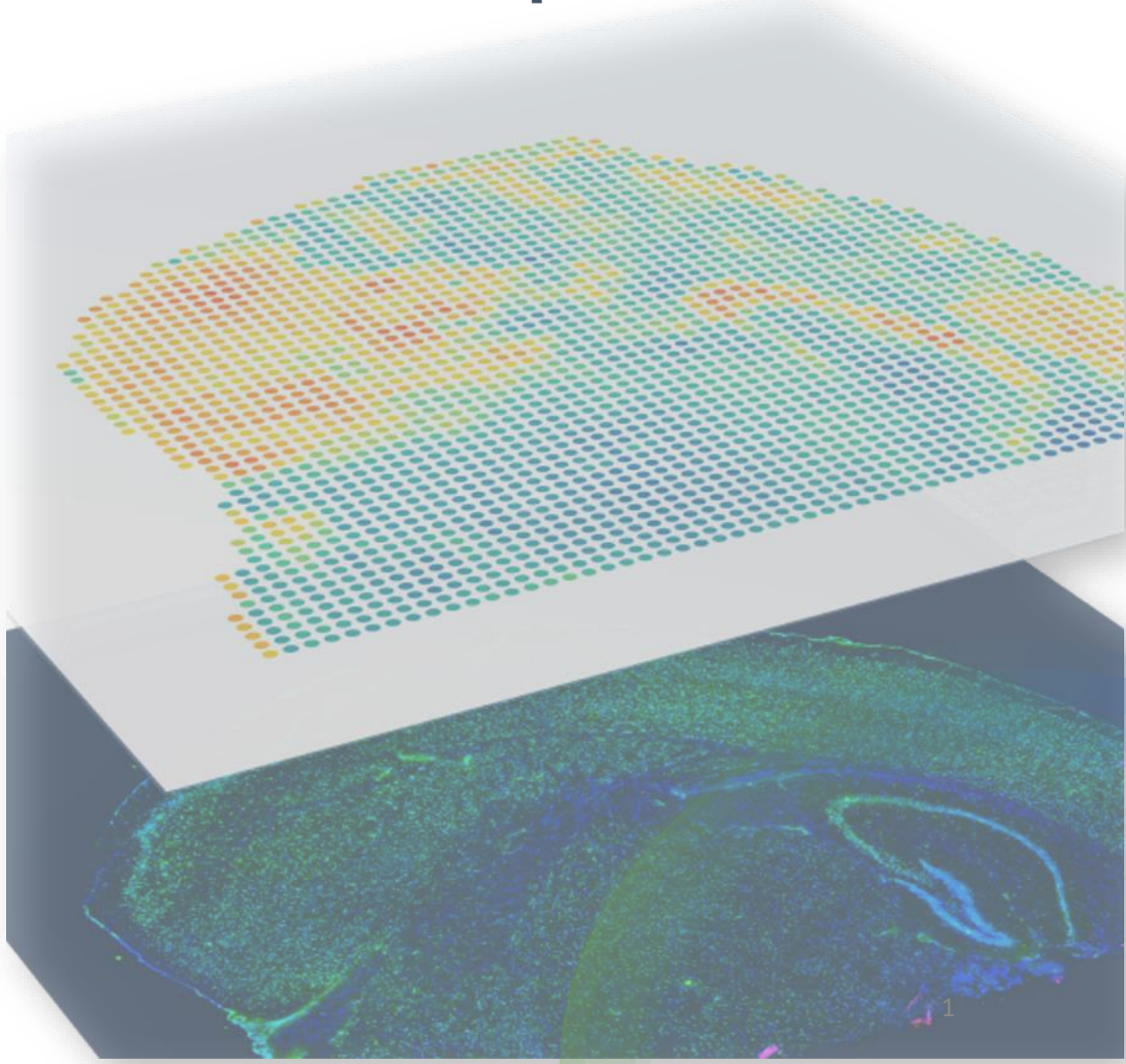
Spatially Resolved Transcriptomics

Data modalities
&
Unsupervised Analysis

Mario Acera Mateos
Cellular System Genomic Lab
macera@carrerasresearch.org

JOSEP CARRERAS LEUKAEMIA RESEARCH INSTITUTE
For a future without leukaemia



Guideline

- Data overview
- Pipeline overview
- Leverage new data modalities
- Spatially aware unsupervised analysis
    - Data normalization
    - Feature selection
    - Model based
  - Benchmarking and further specific results
- Latent representation from reference-based deconvolution
- Bonus: Deep Data Fusion

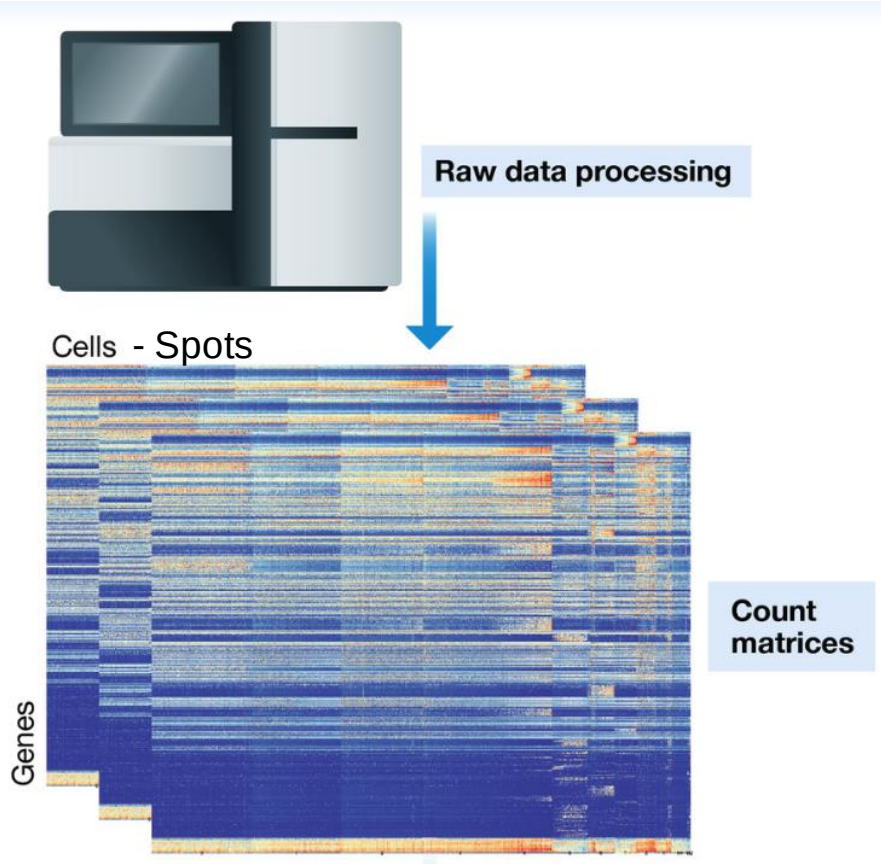
Guideline

- **Data overview**
- Pipeline overview
- Leverage new data modalities
- Spatially aware unsupervised analysis
    - Data normalization
    - Feature selection
    - Model based
  - Benchmarking and further specific results
- Latent representation from reference-based deconvolution
- Bonus: Deep Data Fusion

Data overview

Transcriptomic data

Count matrix



General notation for data science

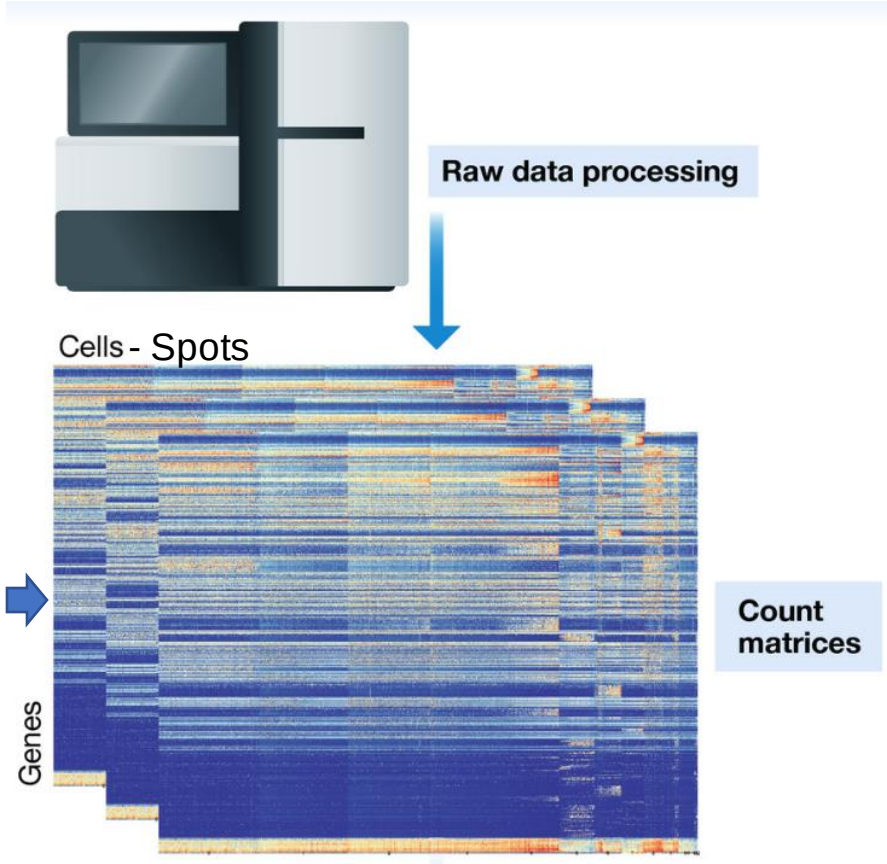
Cells / Spots → Samples

Genes → Features

Data overview

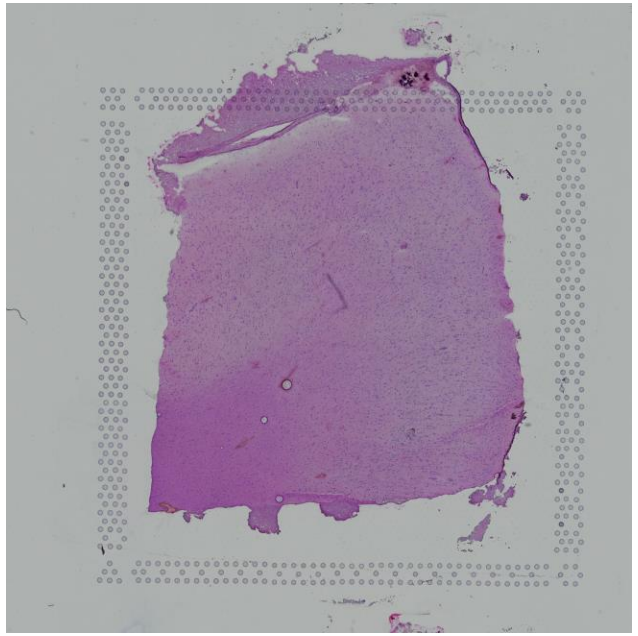
Transcriptomic data

Count matrix

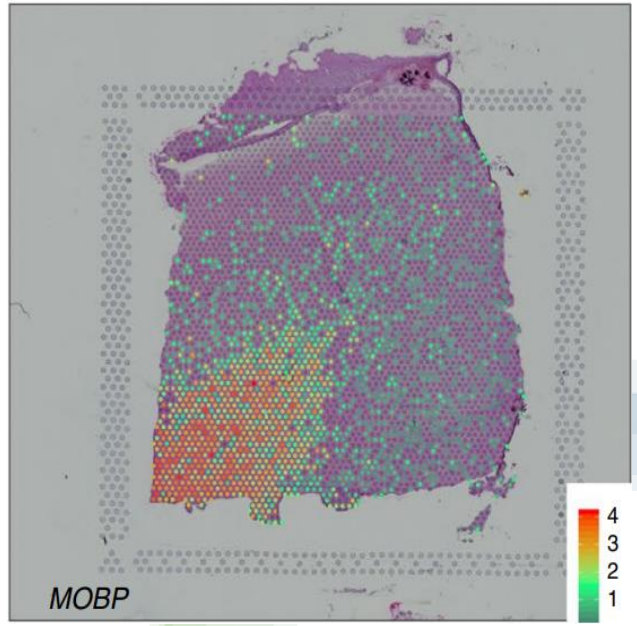


Spatial data

High resolution microscopy image



Spatial location



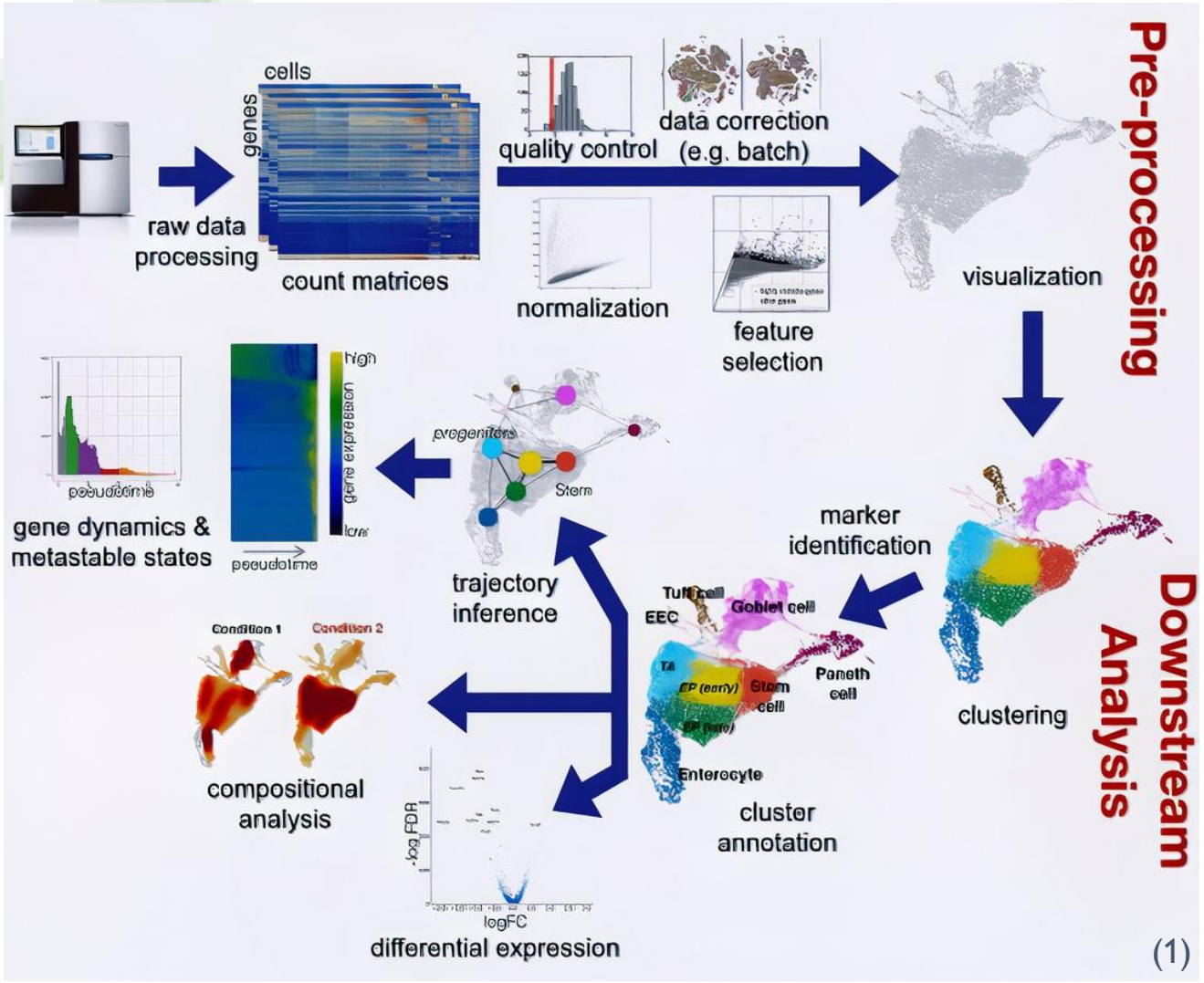
New data modalities!

Guideline

- Data overview
- **Pipeline overview**
- Leverage new data modalities
- Spatially aware unsupervised analysis
    - Data normalization
    - Feature selection
    - Model based
  - Benchmarking and further specific results
- Latent representation from reference-based deconvolution
- Bonus: Deep Data Fusion

Pipeline overview

Unsupervised analysis



Standardized pipeline for scRNA-seq data

This pipeline can be applied on spatially resolved transcriptomics data to do a first exploration of the data.

Our aim shift from identifying and profiling:

Cell types → to → Tissue regions

Guideline

- Data overview
- Pipeline overview
- **Leverage new data modalities**
- Spatially aware unsupervised analysis
    - Data normalization
    - Feature selection
    - Model based
  - Benchmarking and further specific results
- Latent representation from reference-based deconvolution
- Bonus: Deep Data Fusion

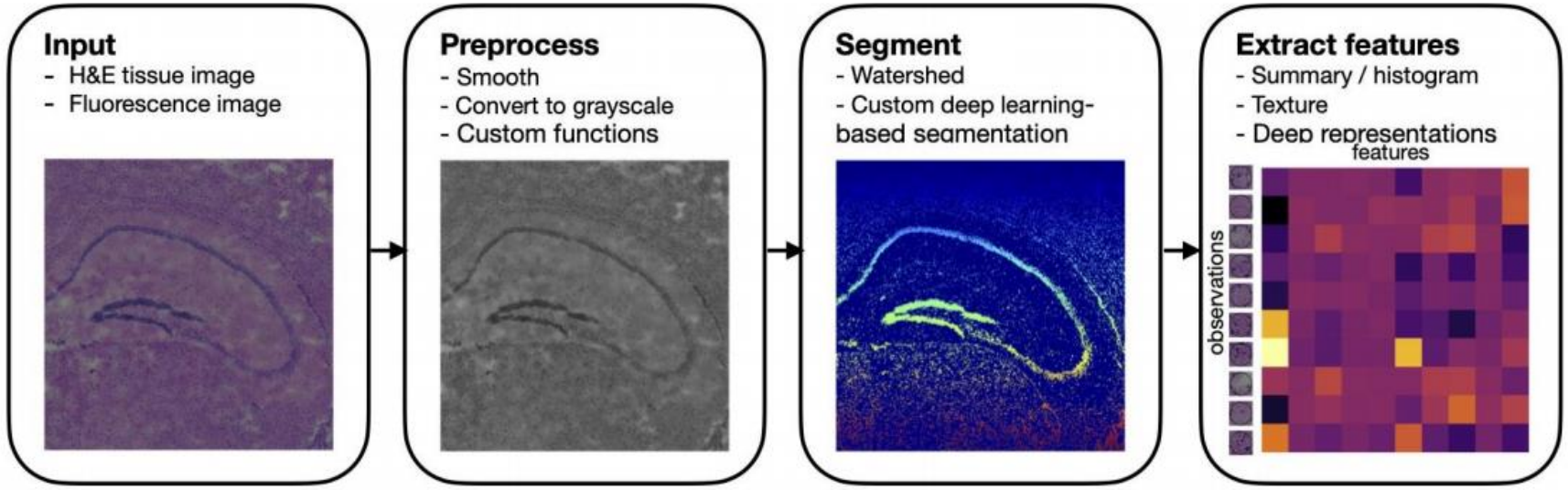
Leverage new data modalities

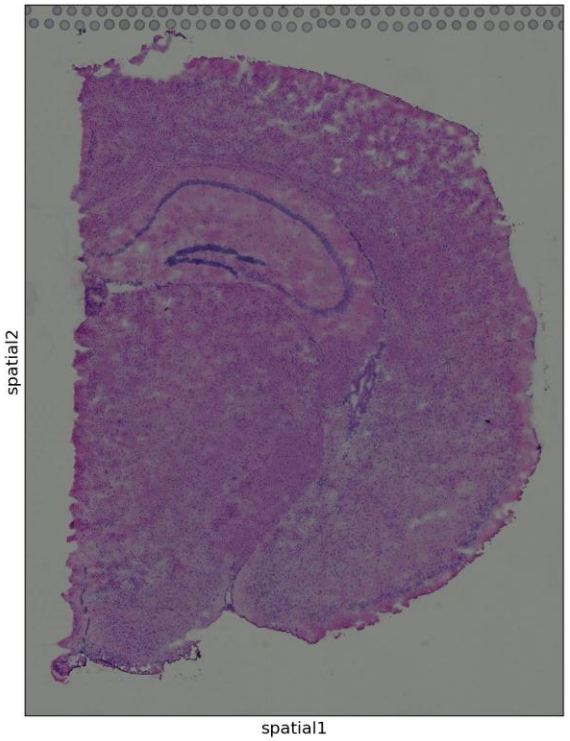
High resolution microscopy image

- Rich source of morphological information

- Useful for visualization and qualitative result assessment

- We can extract image features to complement gene information (transcriptomic features)

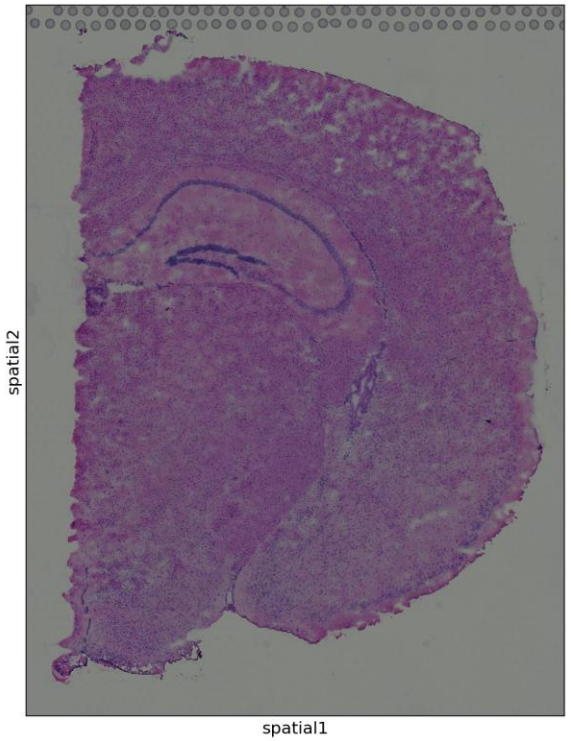




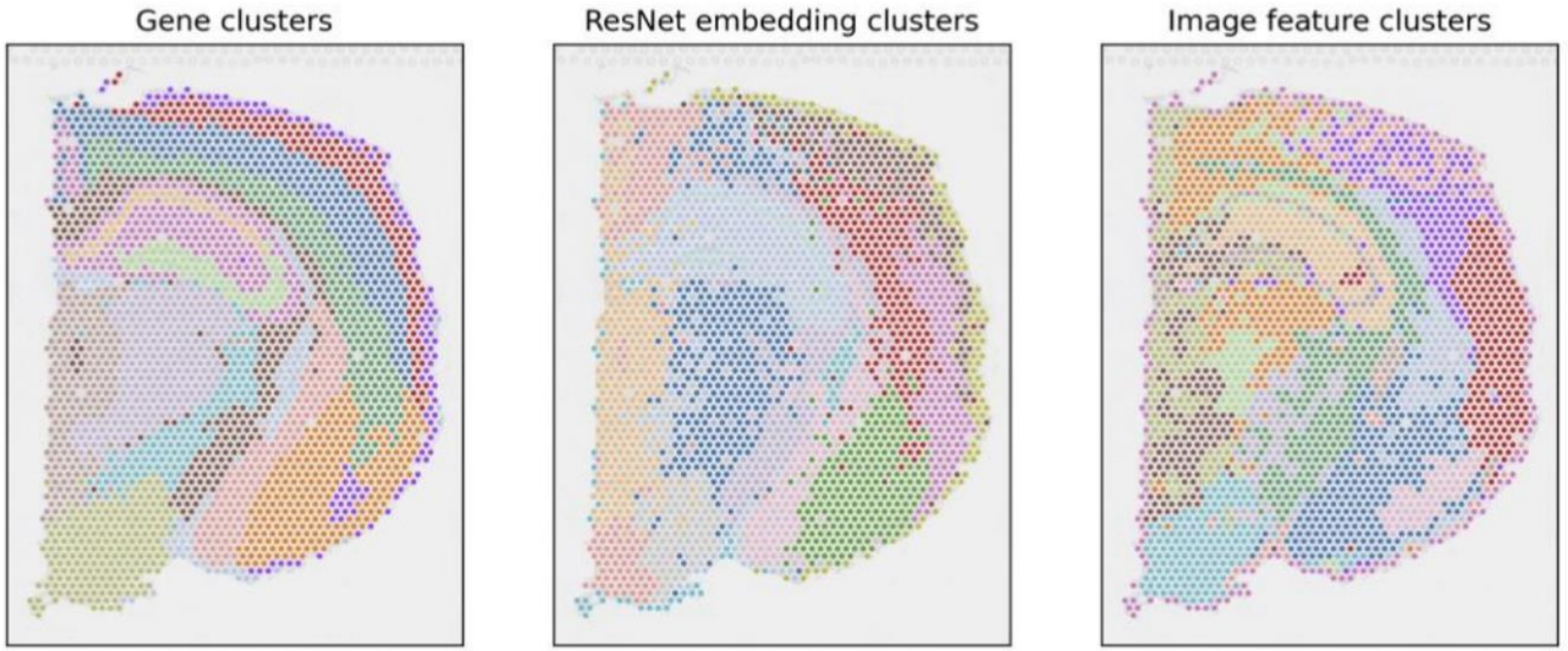
(2) squidpy

Leverage new data modalities

High resolution microscopy image



- **ResNet(DL model) embedding cluster:** Cluster labels obtained from a Deep learning model trained to predict Gene Cluster assignment.
- **Image features cluster:** Clustering based on the intensity mean, standard deviation and 0.1, 0.4 and 0.9 quantiles of the H&E stain at each spot location



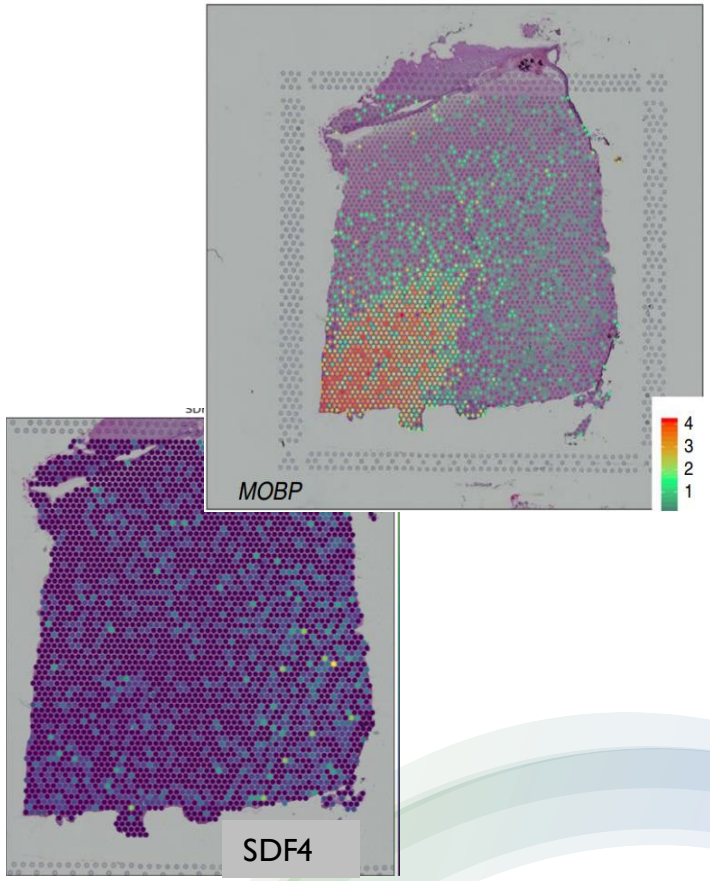
(2) squidpy

Leverage new data modalities

Spatial Location

- **Spatial Statistics** : This field studies entities by using topological, geographic or geometric properties. It offers statistical tests to score the spatial pattern shown by a gene (assess the spatial relevance of each of the features of our samples).

  Can be used for:
  - Feature selection
  - Tissue "markers" exploration



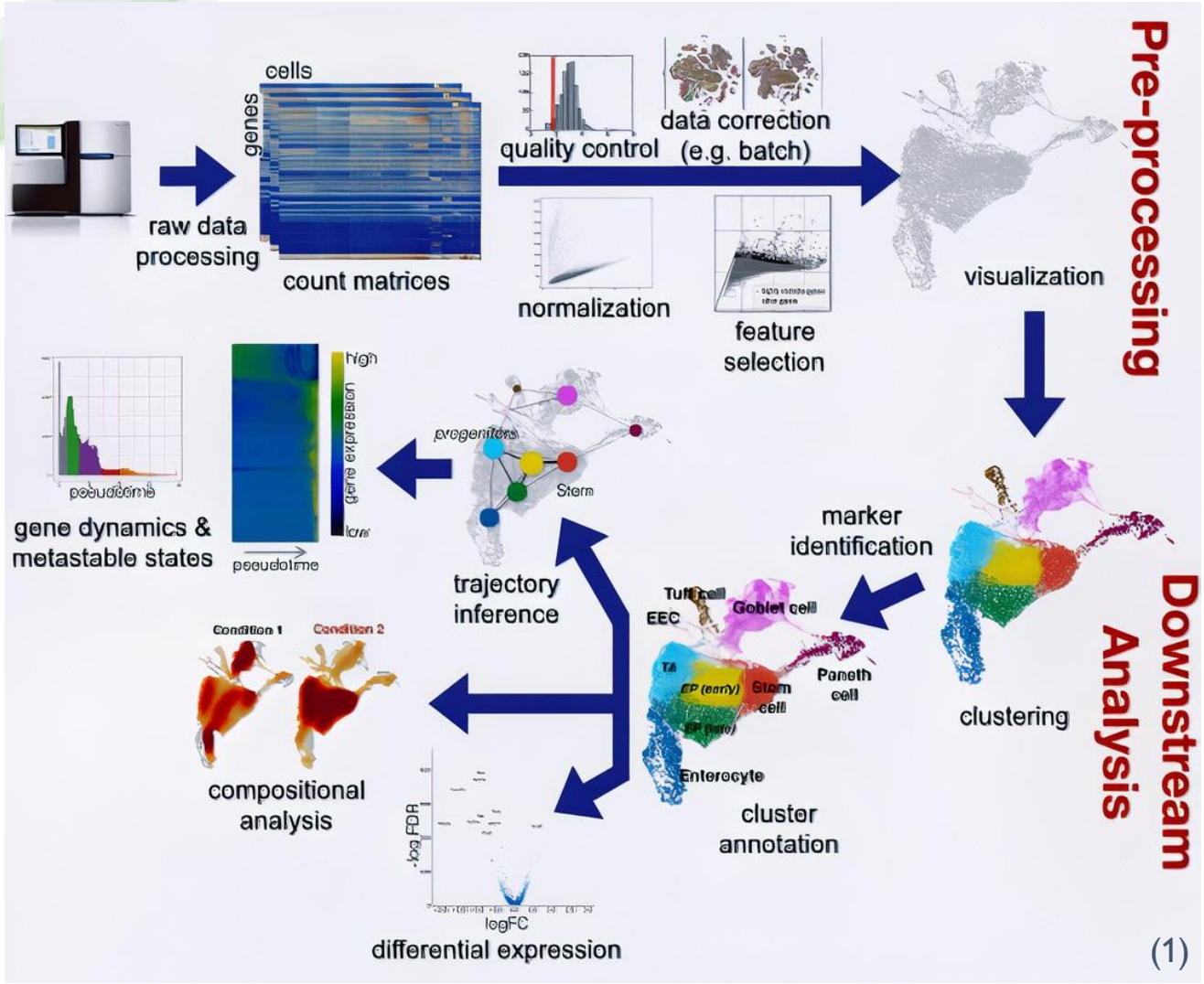
- **Spatial Graphs :** Graphs are incredibly flexible tools. Spatial graphs encode spatial proximity. Can be used for a wide variety of purposes, in preprocessing and downstream analysis.

(2)

Guideline

- Data overview
- Pipeline overview
- Leverage new data modalities
- **Spatially aware unsupervised analysis**
    - Data normalization
    - Feature selection
    - Model based
  - Benchmarking and further specific results
- Latent representation from reference-based deconvolution
- Bonus: Deep Data Fusion

Spatially aware unsupervised analysis

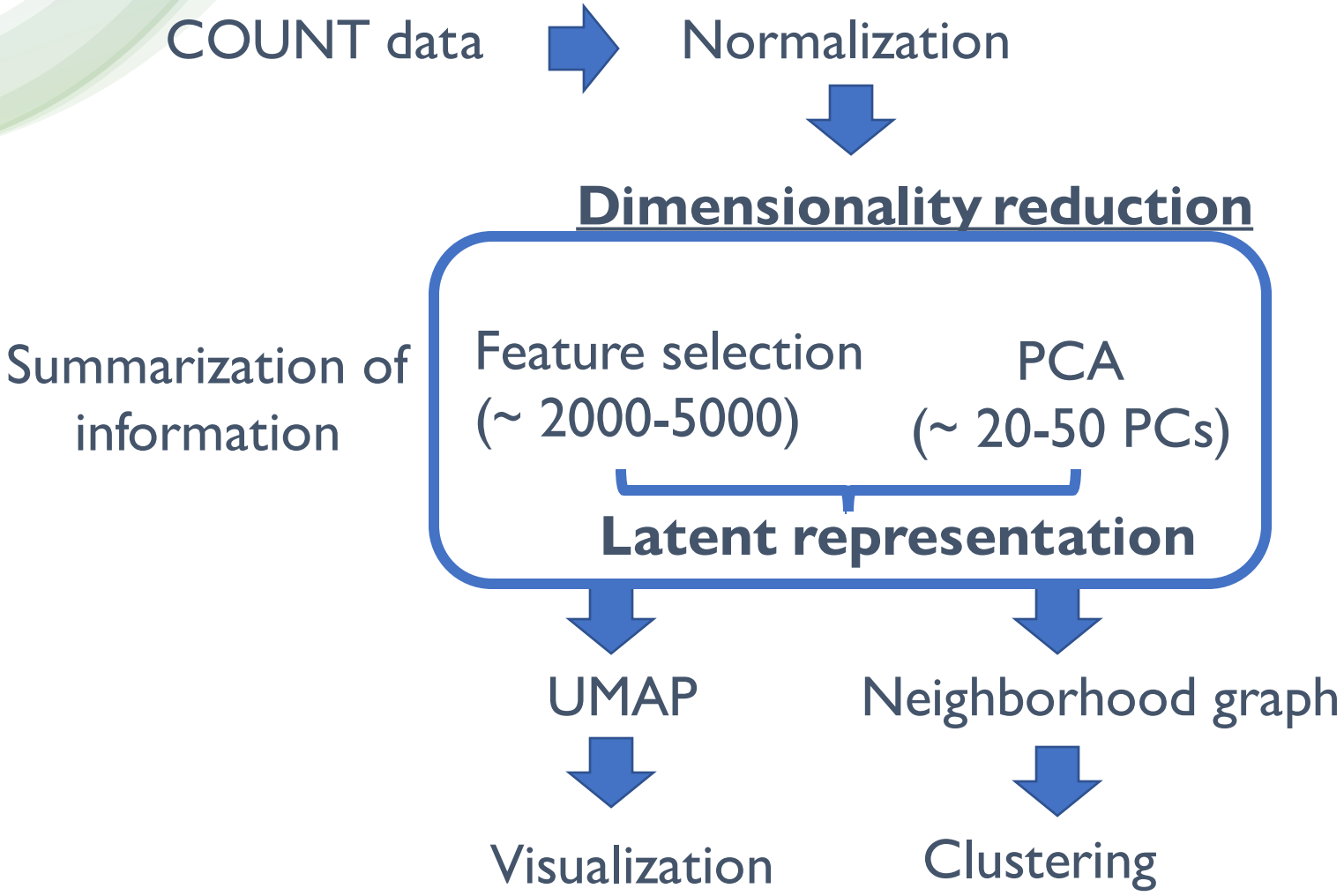


To properly dissect tissue heterogeneity and fully exploit the data:

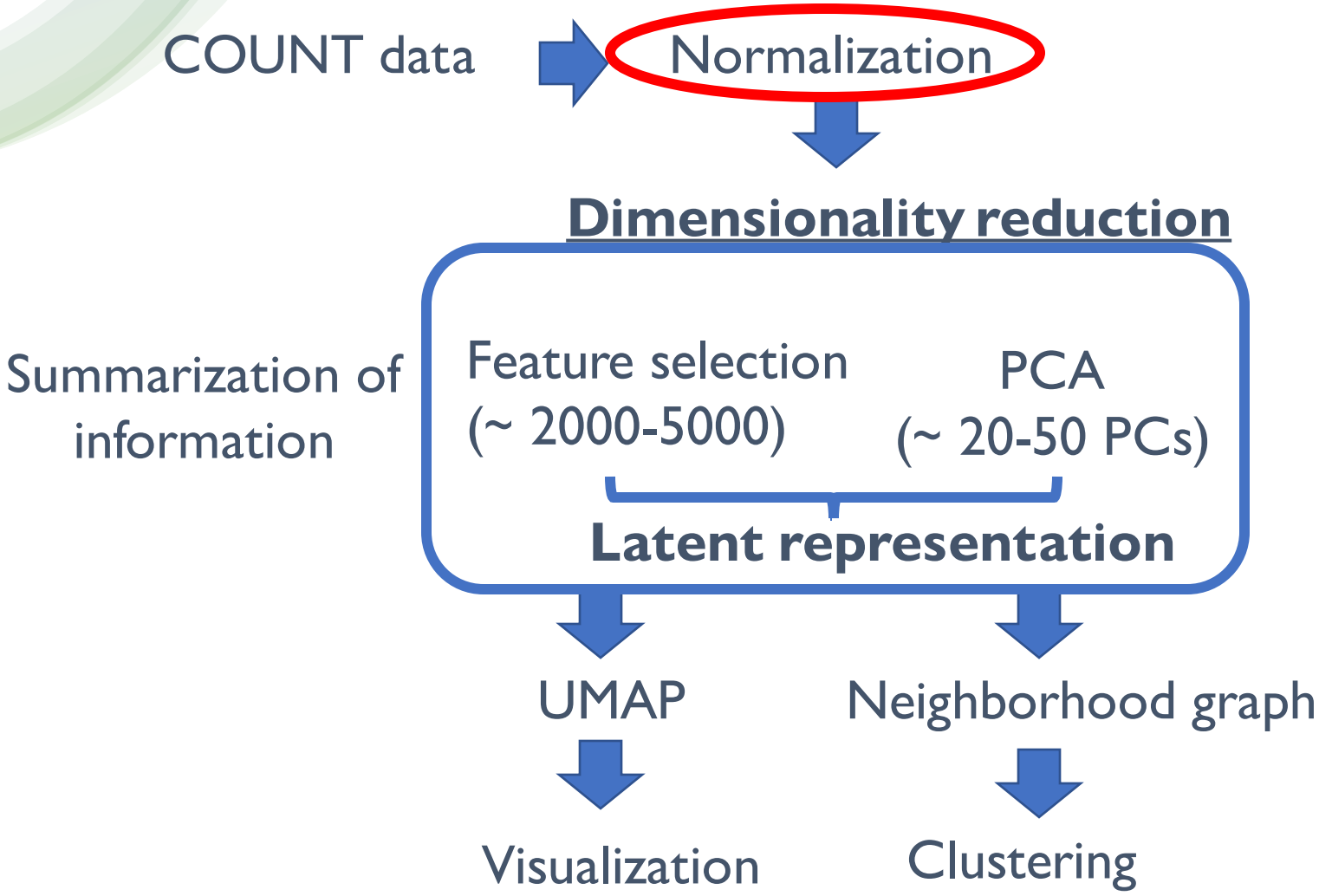
An optimal integration of transcriptomics data and associated spatial information is essential

Let's take a careful look to the pipeline up to the clustering step

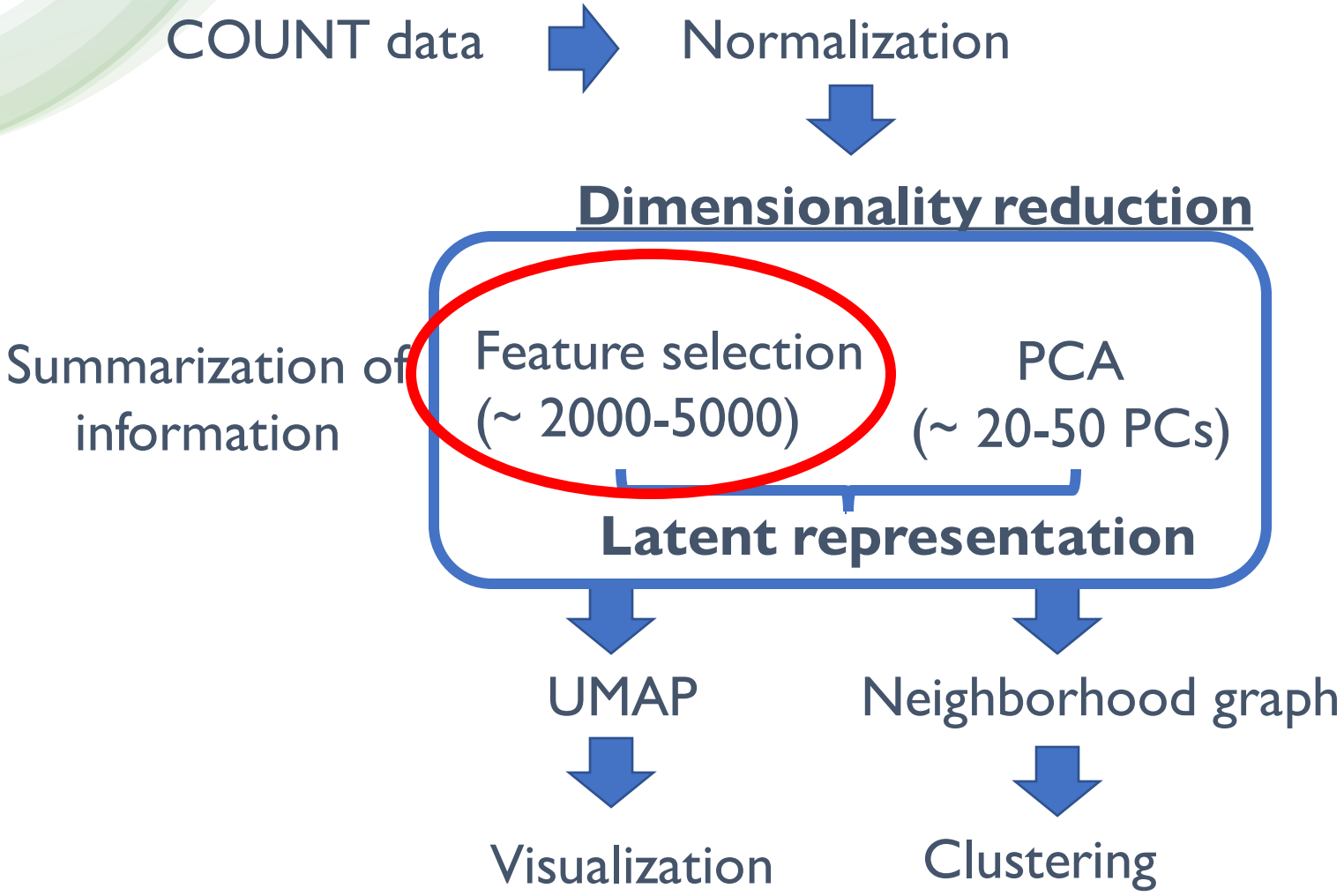
Spatially aware unsupervised analysis



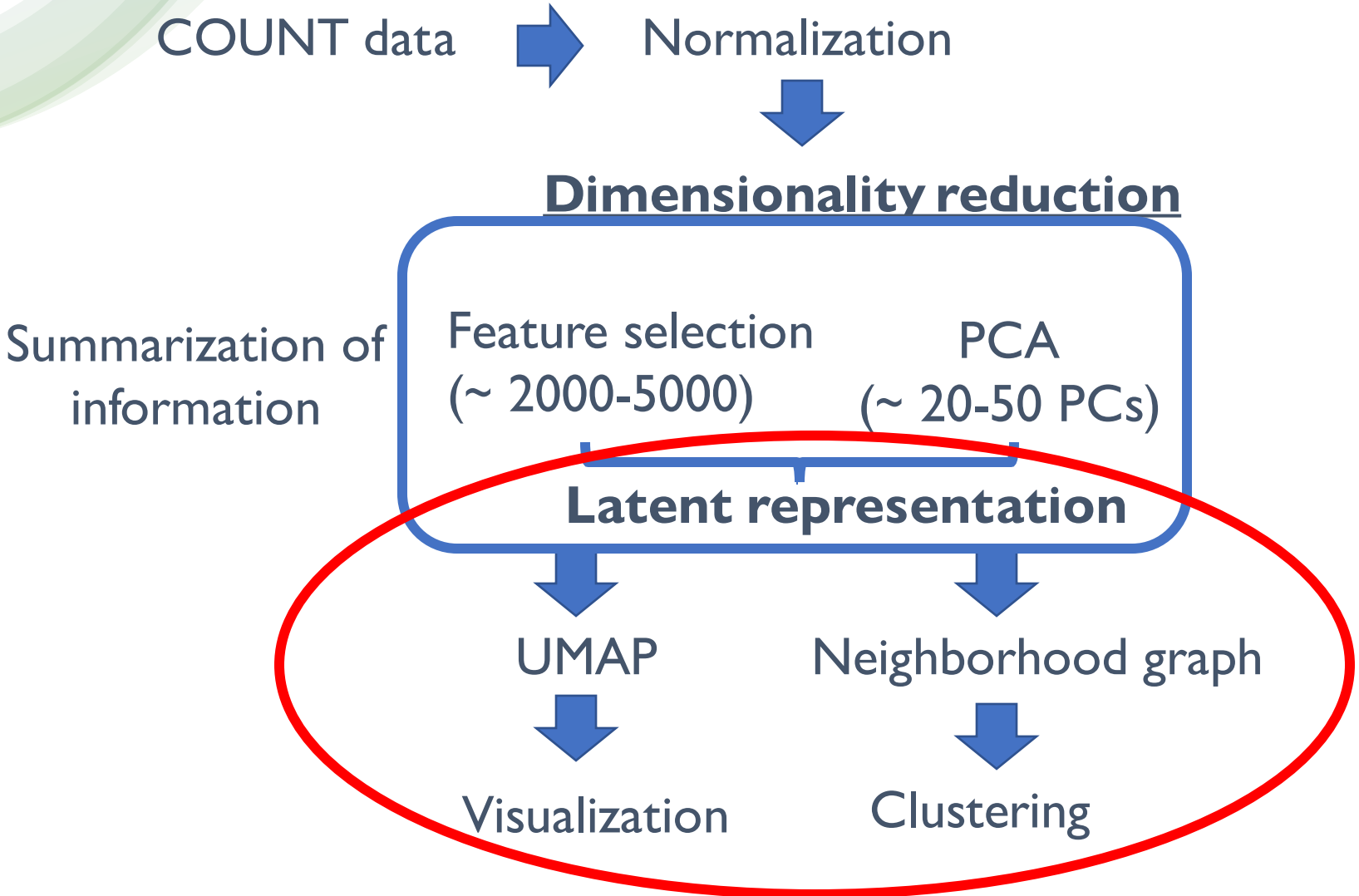
Spatially aware unsupervised analysis



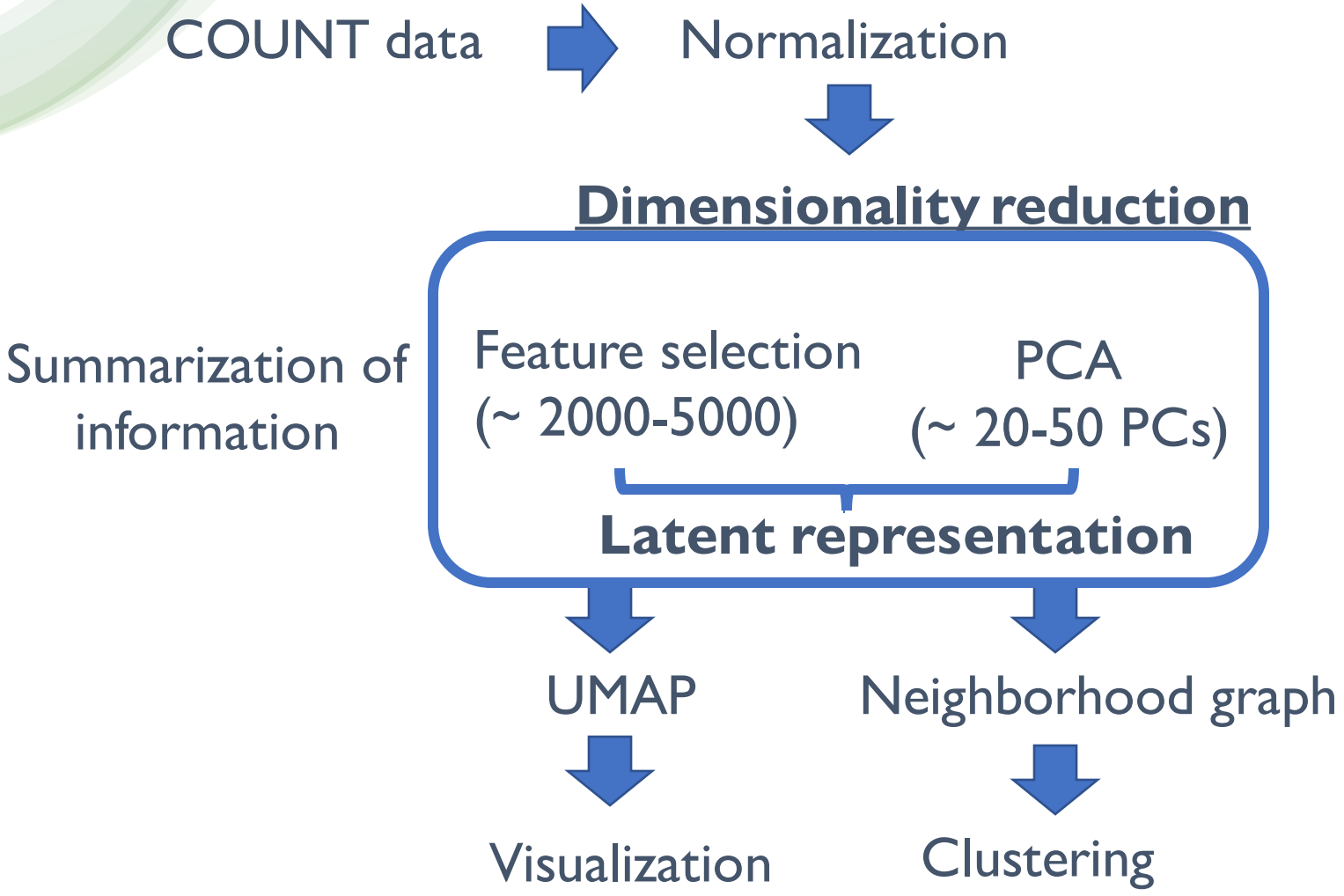
Spatially aware unsupervised analysis



Spatially aware unsupervised analysis



Spatially aware unsupervised analysis



In terms of SUMMARIZATION:
We aim to compress spatial information in the latent representation of the data to **conduct a spatially aware clustering and downstream analysis**

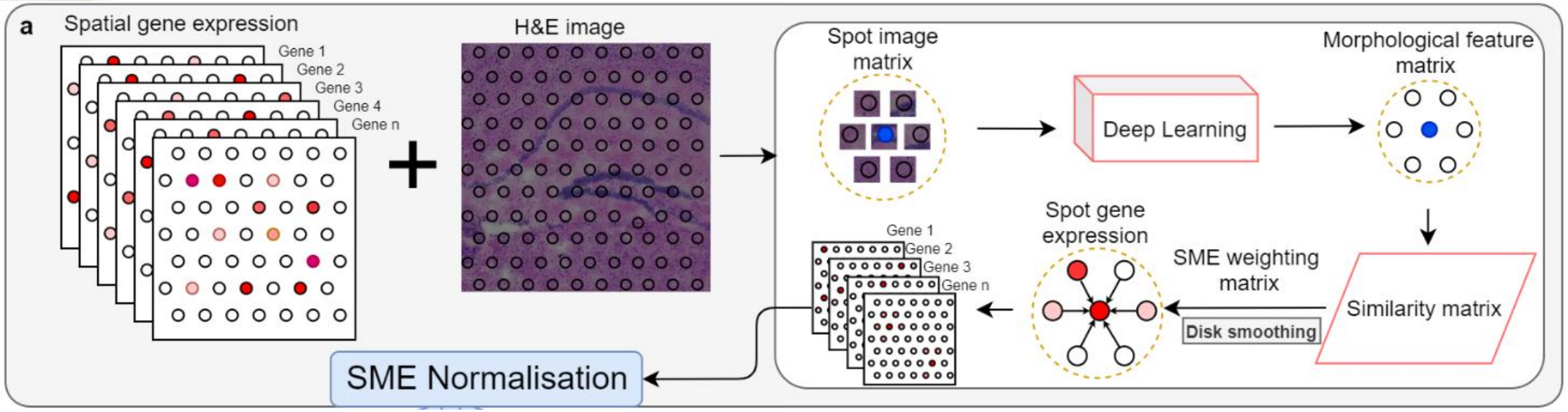
For VISUALIZATION:
Alternatively to the UMAP, we can use SPATIAL COORDINATED to visualize CLUSTER ASSIGMENTS
We can **increase the spatial resolution of the transcriptomic information** with model-based approaches

Guideline

- Data overview
- Pipeline overview
- Leverage new data modalities
- Spatially aware unsupervised analysis
    - **Data normalization**
    - Feature selection
    - Model based
  - Benchmarking and further specific results
- Latent representation from reference-based deconvolution
- Bonus: Deep Data Fusion

Spatially aware normalization

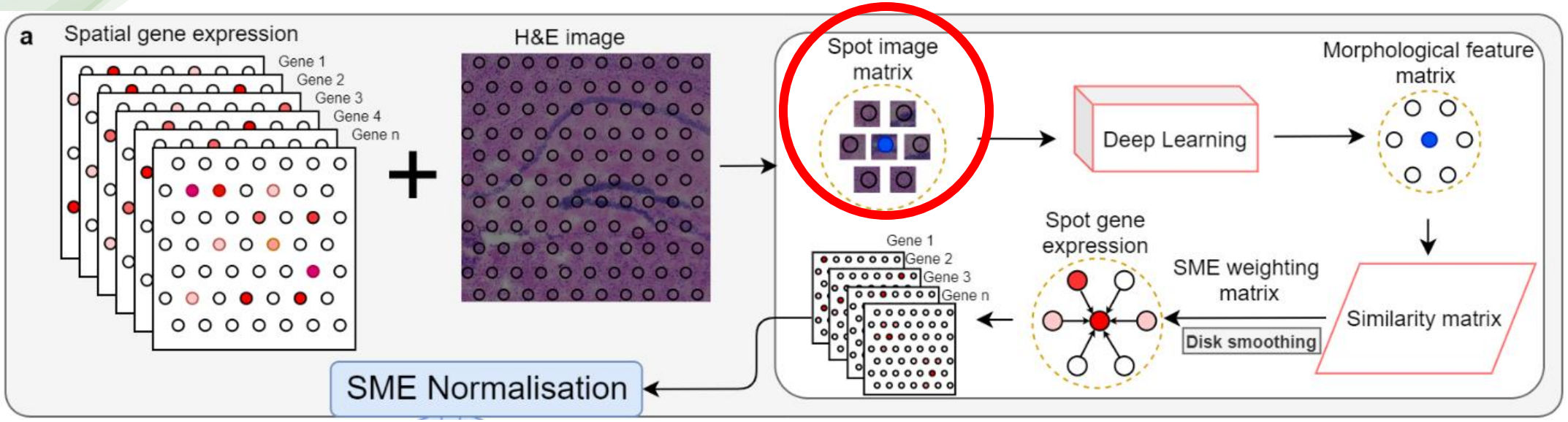
TOOL : **stLearn**[3] **Spatial Morphological gene Expression Normalization**



Spatially aware normalization

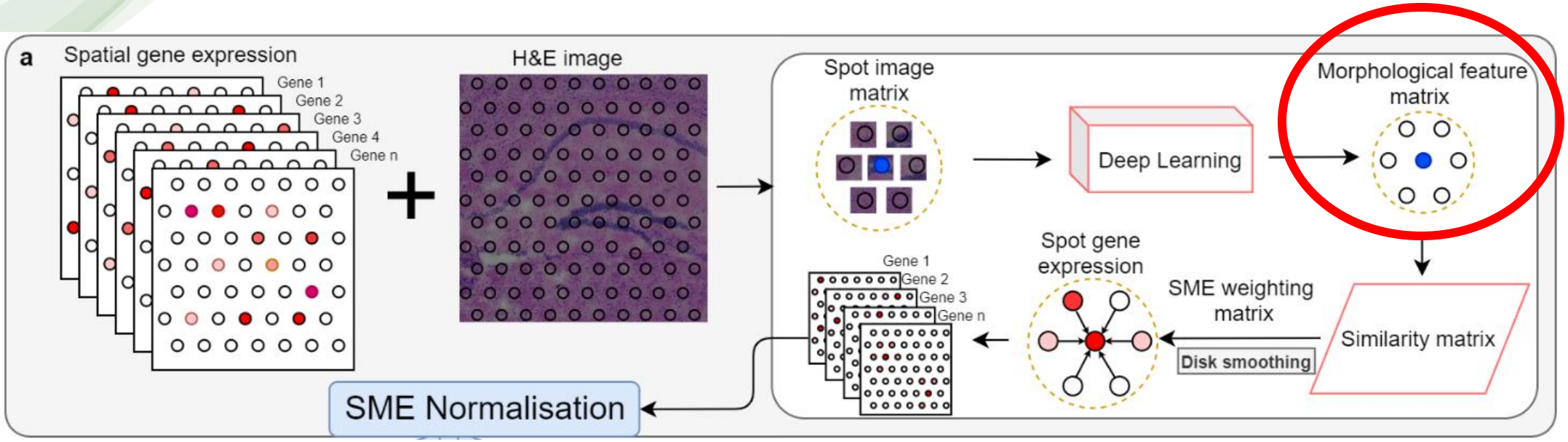
TOOL : **stLearn**[3]

Spatial Morphological gene Expression Normalization



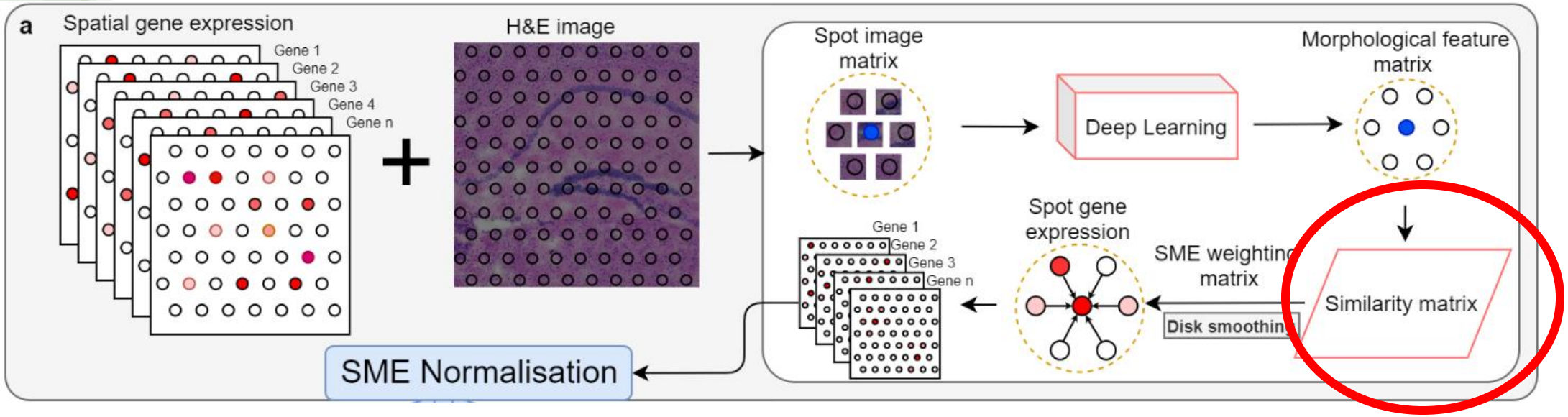
Spatially aware normalization

TOOL : **stLearn**[3] DL models can convert an image into a 2048-dimensional vector. Then we can apply PCA to extract the first 50 PCs as latent features to represent the spot morphology (M)



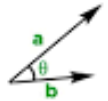
Spatially aware normalization

TOOL : **stLearn**[3] DL models can convert an image into a 2048-dimensional vector. Then we can apply PCA to extract the first 50 PCs as latent features to represent the spot morphology (M)



**Reminder

The Vector Dot Product



$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{b}| \cos\theta$$

$$\cos\theta = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}||\mathbf{b}|}$$

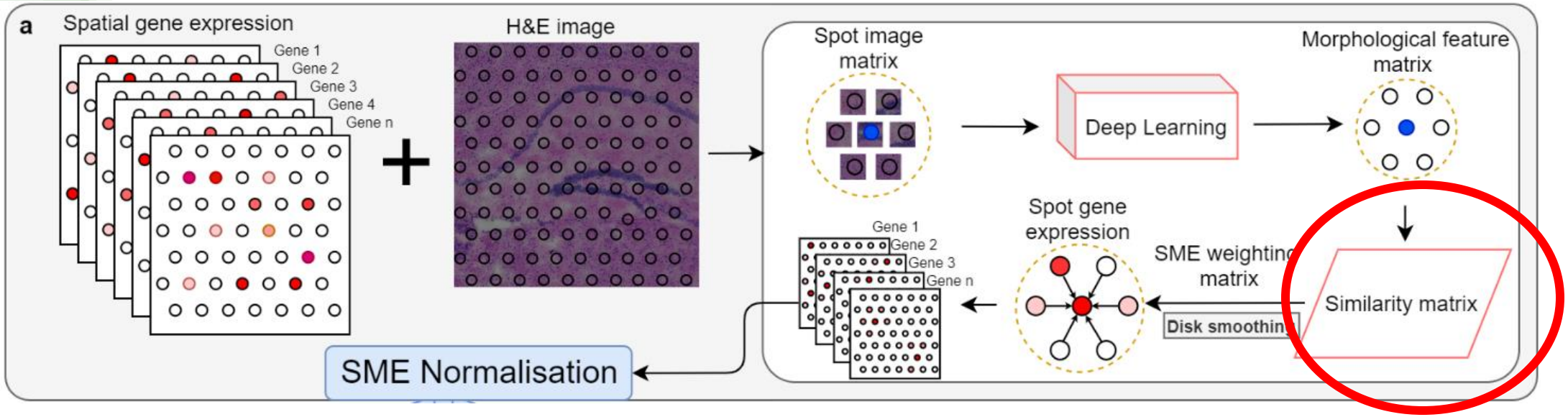
Morphological distance
Similarity metric

$$MD(S_i, S_j) = MD_{ij} = \frac{M_i \cdot M_j}{\|M_i\| \|M_j\|}$$

if $PD_{ij} < r$

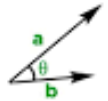
Spatially aware normalization

TOOL : **stLearn**[3] DL models can convert an image into a 2048-dimensional vector. Then we can apply PCA to extract the first 50 PCs as latent features to represent the spot morphology (M)



**Reminder

The Vector Dot Product



$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{b}| \cos\theta$$

$$\cos\theta = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}||\mathbf{b}|}$$

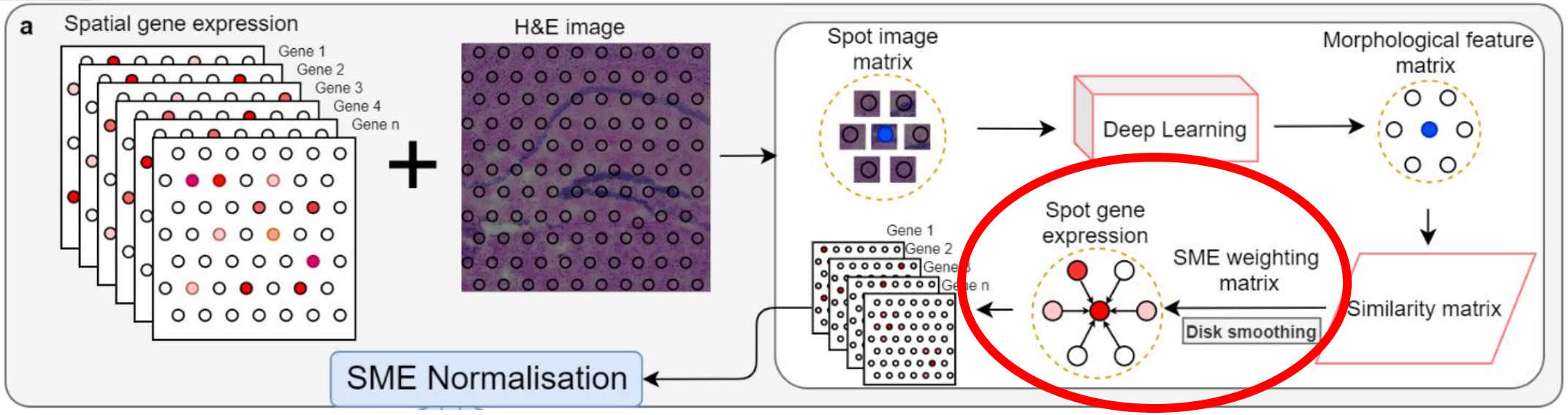
Morphological distance
Similarity metric

$$MD(S_i, S_j) = MD_{ij} = \frac{M_i \cdot M_j}{\|M_i\| \|M_j\|}$$

if $PD_{ij} < r$

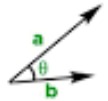
Spatially aware normalization

TOOL : **stLearn**[3] DL models can convert an image into a 2048-dimensional vector. Then we can apply PCA to extract the first 50 PCs as latent features to represent the spot morphology (M)



**Reminder

The Vector Dot Product



$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{b}| \cos\theta$

$\cos\theta = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}||\mathbf{b}|}$

Morphological distance

$$MD(S_i, S_j) = MD_{ij} = \frac{M_i \cdot M_j}{\|M_i\| \|M_j\|}$$

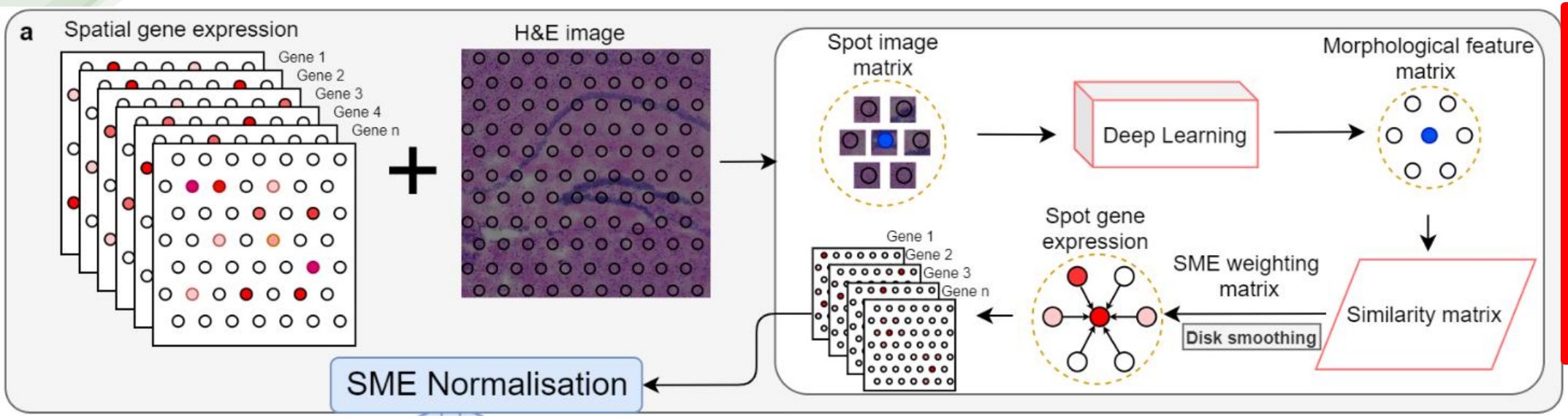
if $PD_{ij} < r$

Spatial Morphological gene Expression Normalizati

$$GE_i' = GE_i + \frac{\sum_{j=1}^{n} GE_j \cdot MD_{ij}}{n}$$

Spatially aware normalization

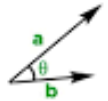
TOOL : **stLearn**[3] DL models can convert an image into a 2048-dimensional vector. Then we can apply PCA to extract the first 50 PCs as latent features to represent the spot morphology (M)



Alternatively, they propose to perform this normalization on the latent features (PCs / UMAP)

**Reminder

The Vector Dot Product



$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{b}| \cos\theta$

$\cos\theta = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}||\mathbf{b}|}$

Morphological distance

$$MD(S_i, S_j) = MD_{ij} = \frac{M_i \cdot M_j}{\|M_i\| \|M_j\|}$$

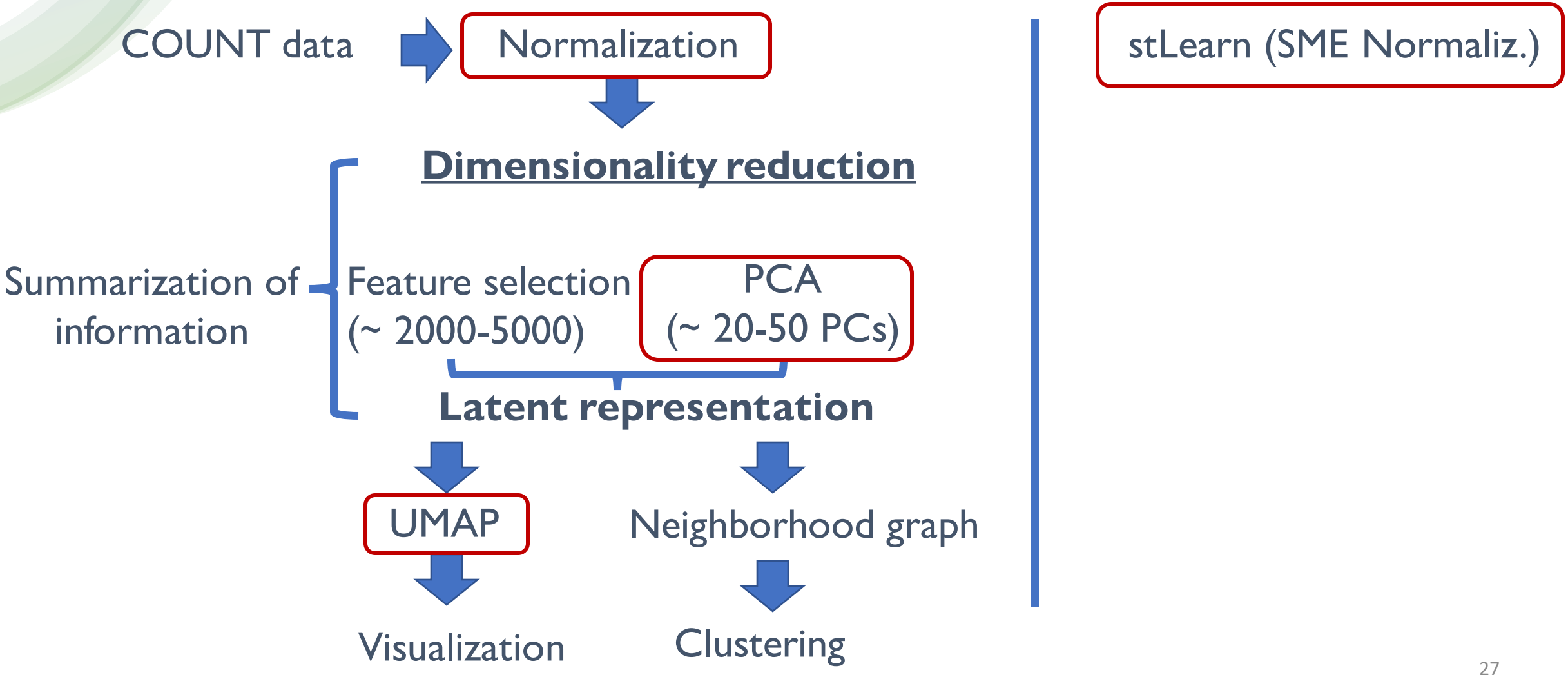
if $PD_{ij} < r$

Spatial Morphological gene Expression Normalizati

$$GE_i' = GE_i + \frac{\sum_{j=1}^{n} GE_j \cdot MD_{ij}}{n}$$

Spatially aware unsupervised analysis

Framed steps are tackled or replaced by the same-colored framed methods

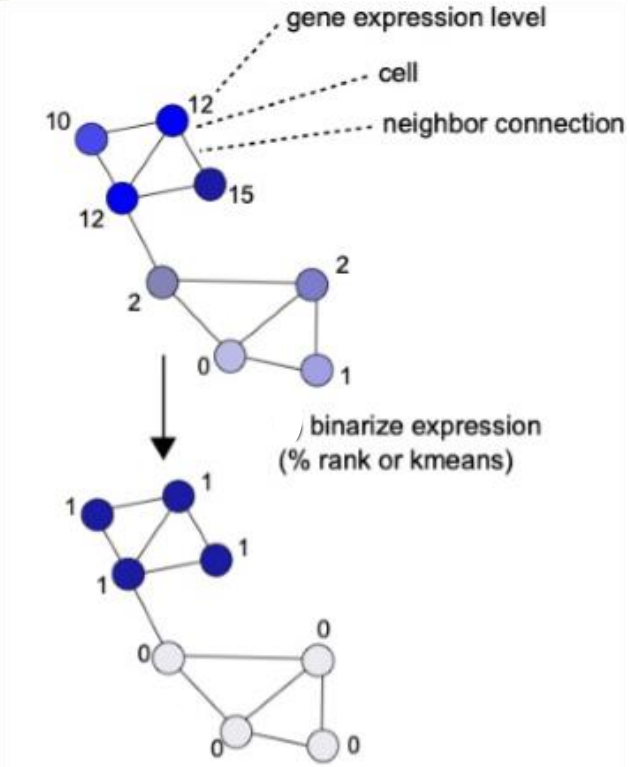


Guideline

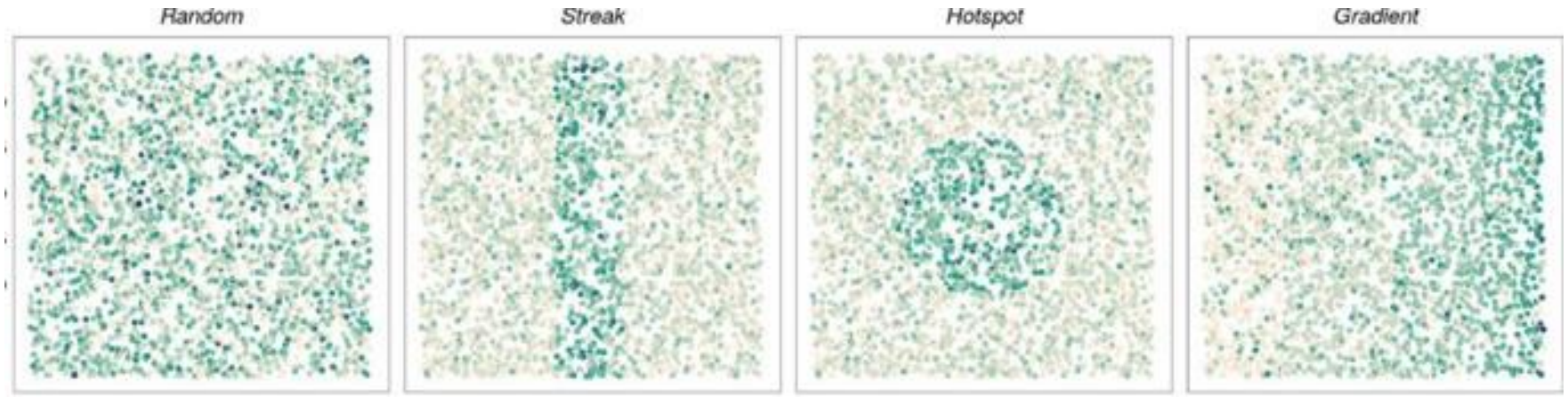
- Data overview
- Pipeline overview
- Leverage new data modalities
- Spatially aware unsupervised analysis
    - Data normalization
    - **Feature selection**
    - Model based
  - Benchmarking and further specific results
- Latent representation from reference-based deconvolution
- Bonus: Deep Data Fusion

Spatially aware feature selection – Spatial Graphs

TOOL : **Giotto** (4) (Binary Spatial extract / BinSpect)



- Binarize the gene expression value (0/1)
- Assess the spatial pattern by checking whether a gene is usually expressed in neighboring cells



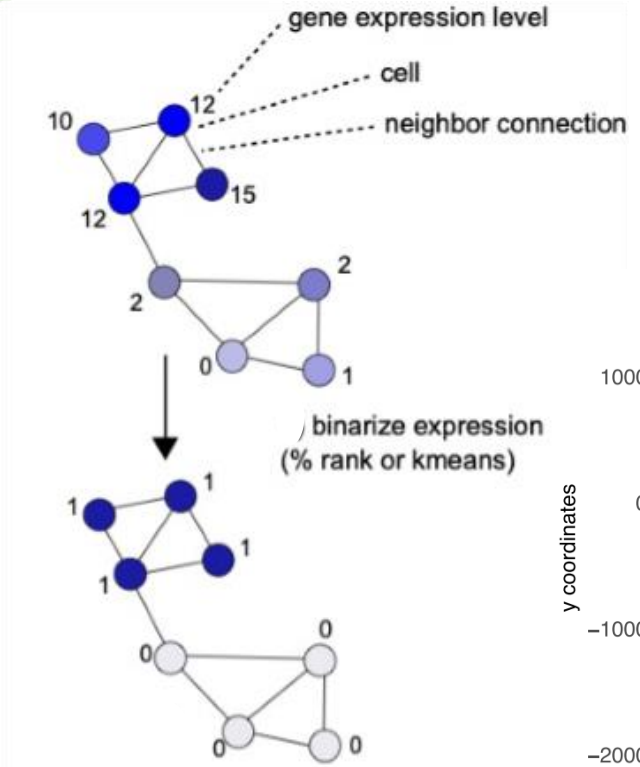
Does not exhibit more potential than Spatial Statistics approaches but is COMPUTATIONALLY MORE EFFICIENT

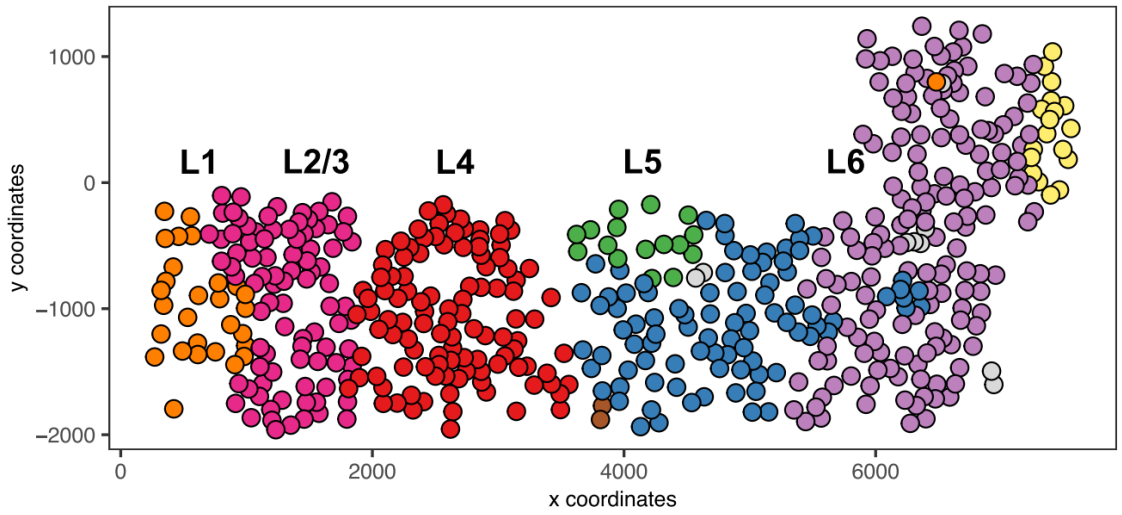
Spatially aware feature selection – Spatial Graphs

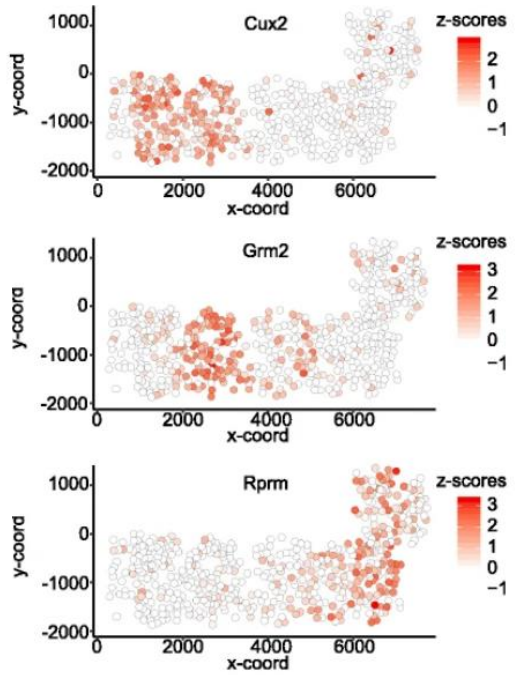
TOOL : **Giotto** [4] (Binary Spatial extract / BinSpect)

- Binarize the gene expression value (0/1)
- Assess the spatial pattern by checking whether a gene is usually expressed in neighboring cells

Suitable for cases such as: seqFISH+ mouse somatosensory cortex (underlying layered structure)







Spatially aware feature selection – Spatial Statistics

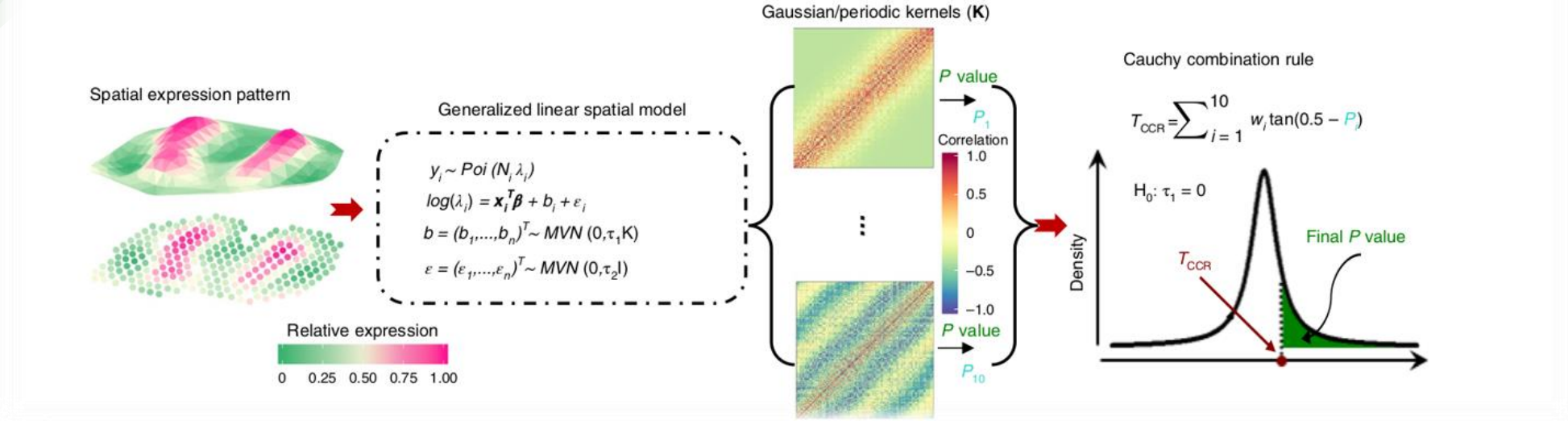
TOOL : **SPARK**[5] (Spatial pattern recognition via kernels)

PREVIOUS APPROACHES		CHALLENGES
SpatialDE[2]	Poor control of type **I** errors* →	Statistical
Based on efficient linear mixed models		
Trendseek[3]	Not Scalable →	Computational
Expensive permutation strategies with non-parametric test statistics		

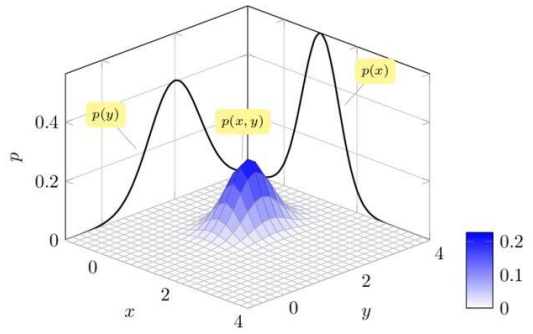
Type **I errors lead to concluding that purely random results are statistically significant*

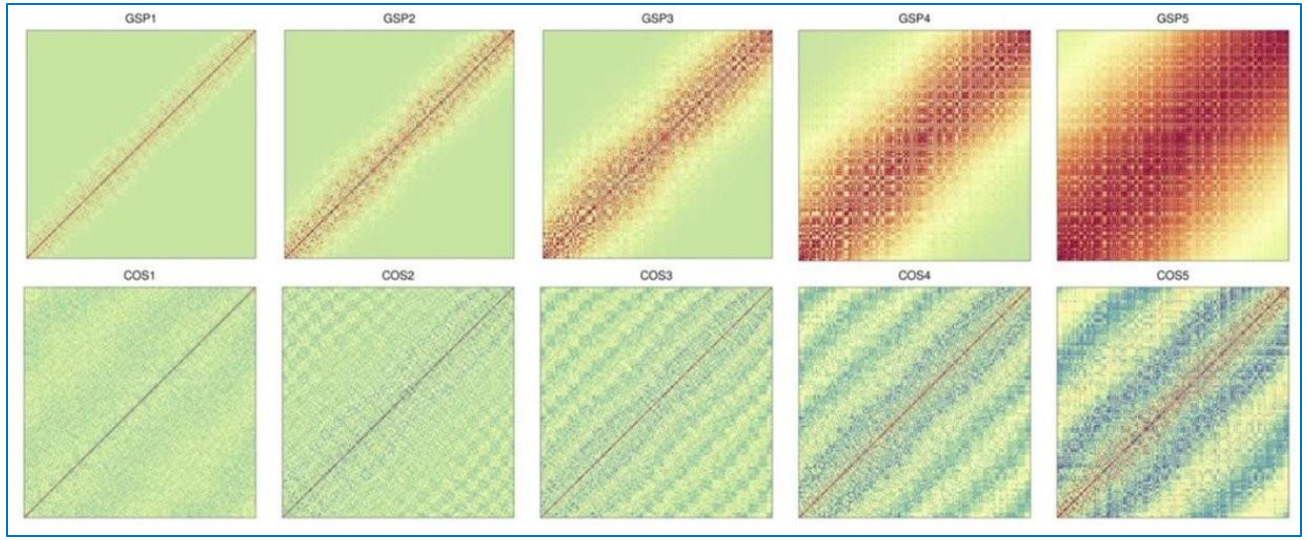
Spatially aware feature selection – Spatial Statistics

TOOL : **SPARK** (5) (Spatial pattern recognition via kernels)



MVN ~ Multivariate normal distribution





Kernels

Gaussian

Periodic

Spatially aware feature selection – Spatial Statistics

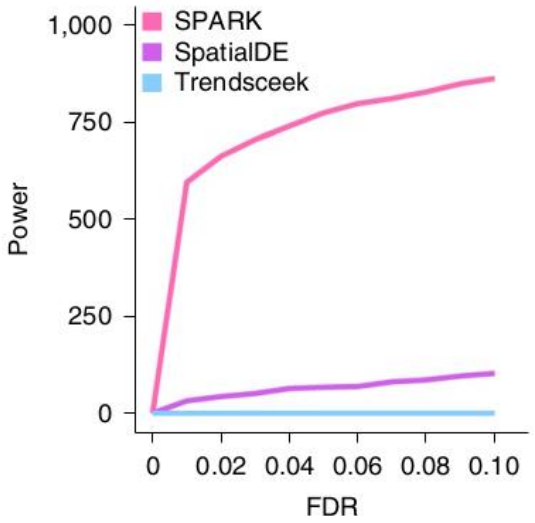
TOOL : **SPARK** (5) (Spatial pattern recognition via kernels)

Olfactory bulb

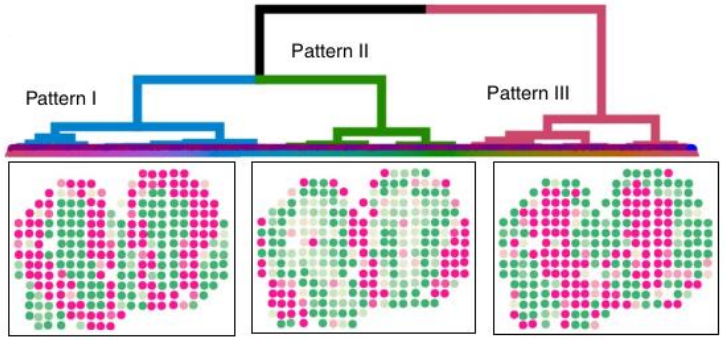
Number of detected genes displaying spatial patterns:

SPARK: 772

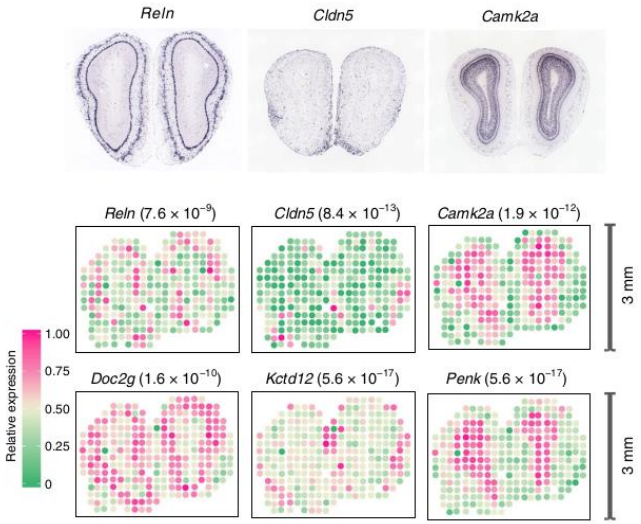
SpatialDE: 67 (62 overlaps)



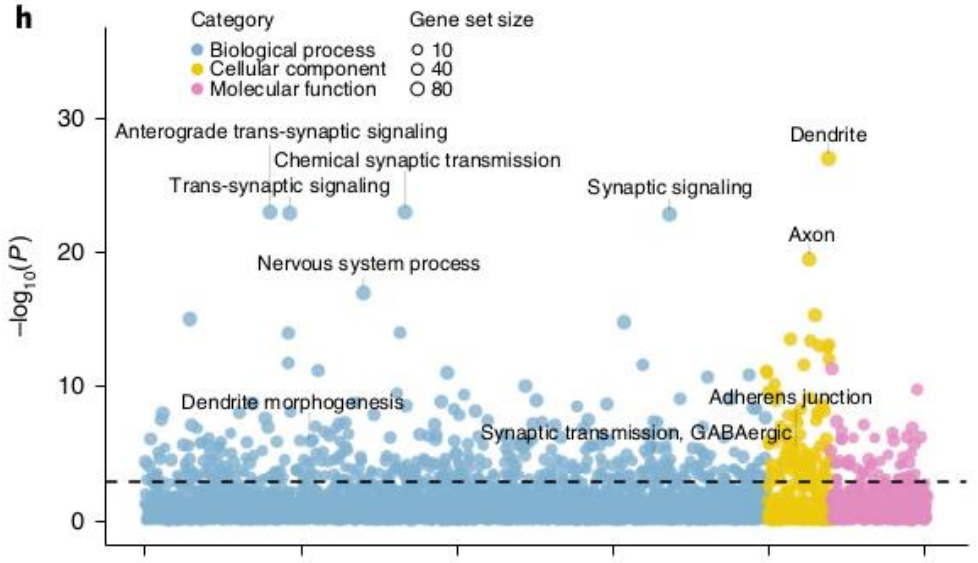
Clustering



Markers



Ontology enrichment analysis



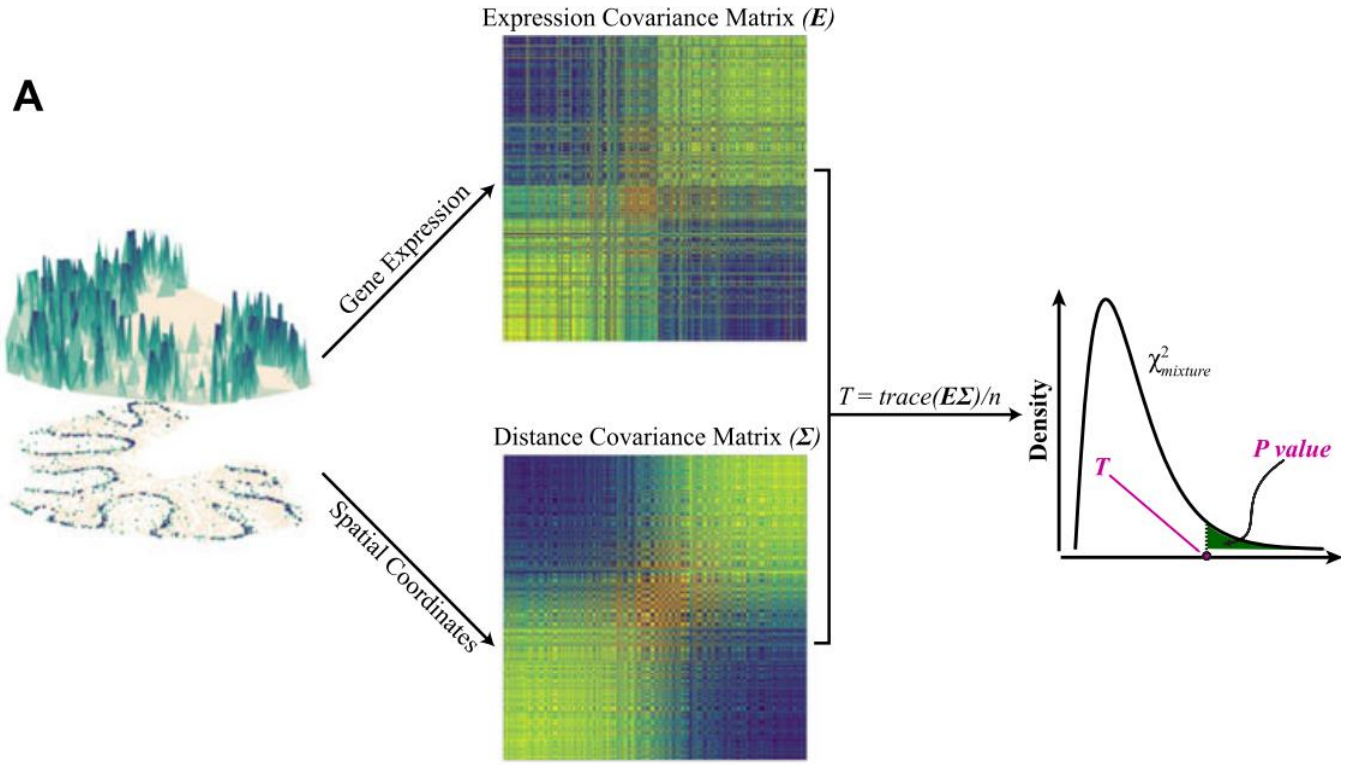
- SYNAPTIC ORGANIZATION
- BULB DEVELOPMENT

Spatially aware feature selection – Spatial Statistics

TOOL : **SPARK-X**[6] (Non-parametric version)

More suitable than SPARK for **sparse, large-scale data**

Scalable for dataset with more than ~10.000 of genes measured on ~10.000 spots.

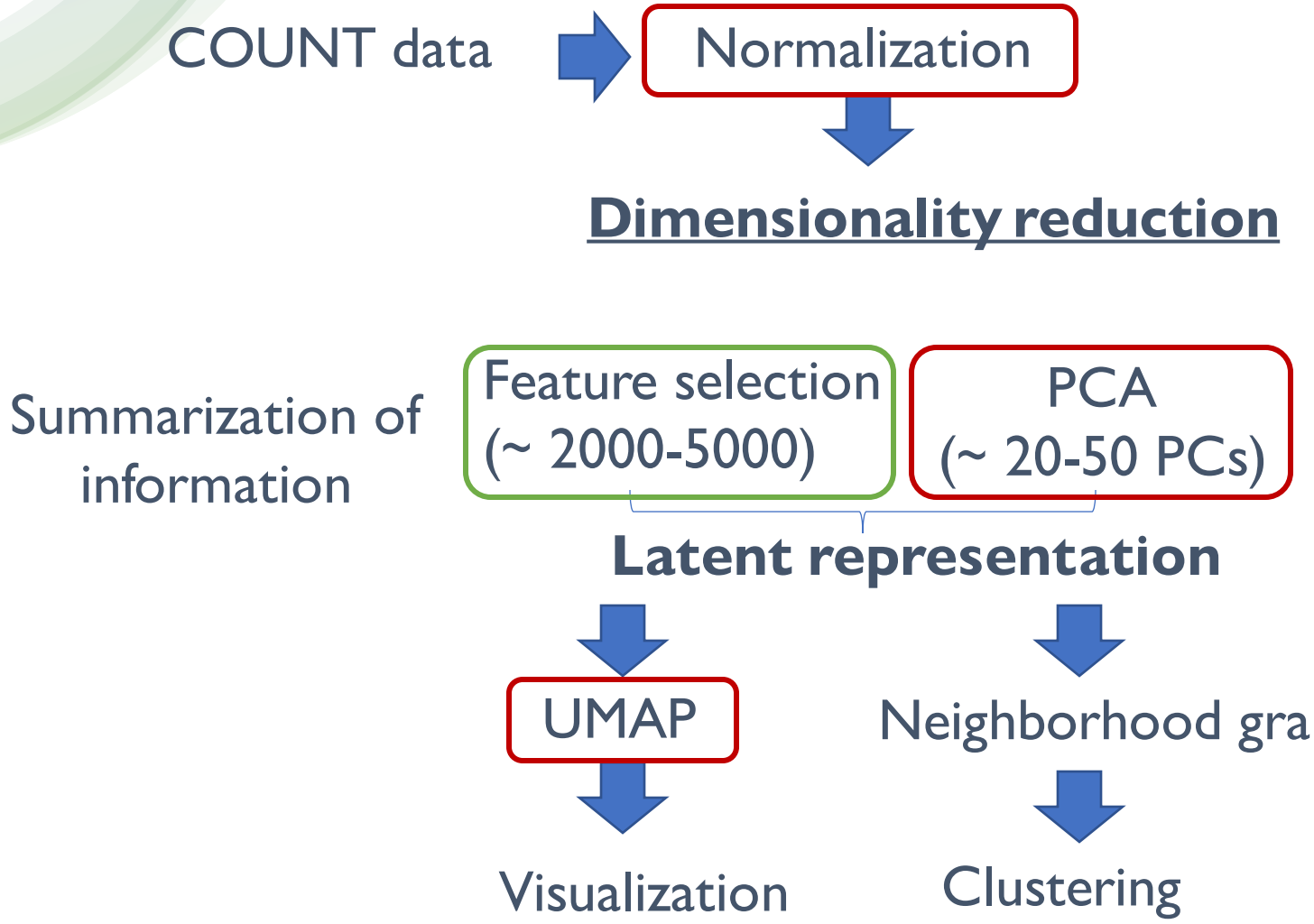


Intuition: if y (gene expression) is independent of S (spatial coordinates), then the spatial distance between two locations ***i*** and ***j*** would also be independent of the gene expression difference between the two locations

2

Spatially aware unsupervised analysis

Framed steps are tackled or replaced by the same-colored framed methods



stLearn (SME Normaliz.)

Spatially aware feature selection

Guideline

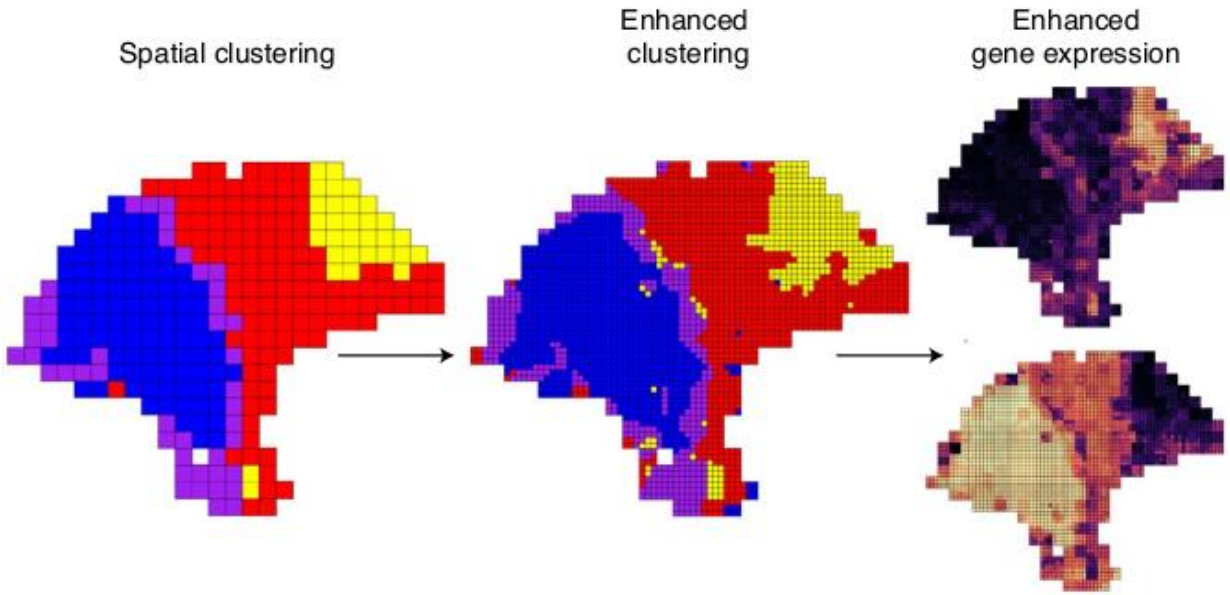
- Data overview
- Pipeline overview
- Leverage new data modalities
- Spatially aware unsupervised analysis
    - Data normalization
    - Feature selection
    - **Model based**
  - Benchmarking and further specific results
- Latent representation from reference-based deconvolution
- Bonus: Deep Data Fusion

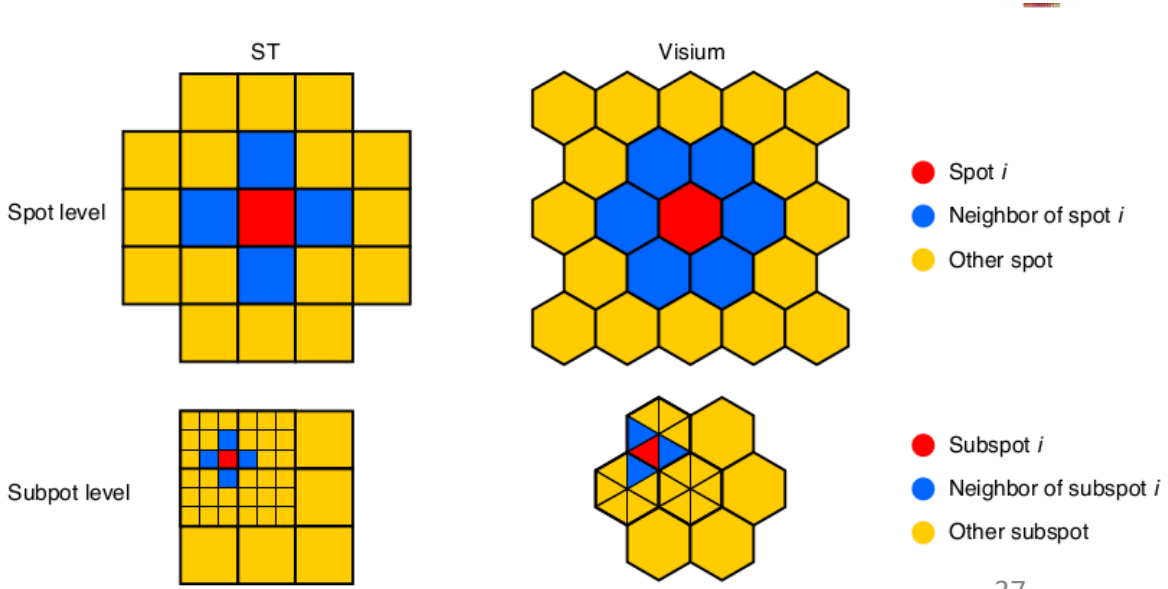
Model based approaches

TOOL : **BayesSpace**(6) Provides the SPOT CLUSTER ASSIGMENTS + SUB-SPOT RESOLUTION

Fully Bayesian Model with Markov Random Field. Inspired on widely used computer vision models for denoising and segmentizing images in a statistical / probabilistic manner

- ***Preprocessing*** : **Normalization + Log-Transformation + top HVG + top PCs (~ 15)**.
- Performance ***relies on empirical knowledge*** for the selection of HVG, PCs, nº of Clusters
- BayesSpace performs iterative clustering, ***CONSTRAINING*** spots to join neighboring clusters (spatial awareness)



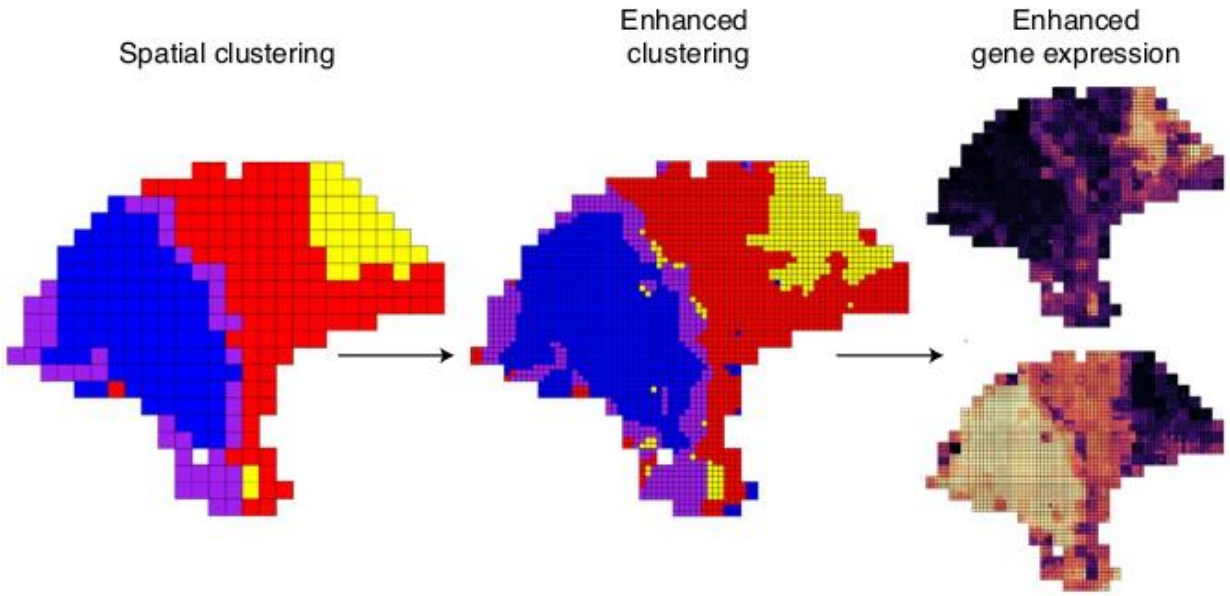


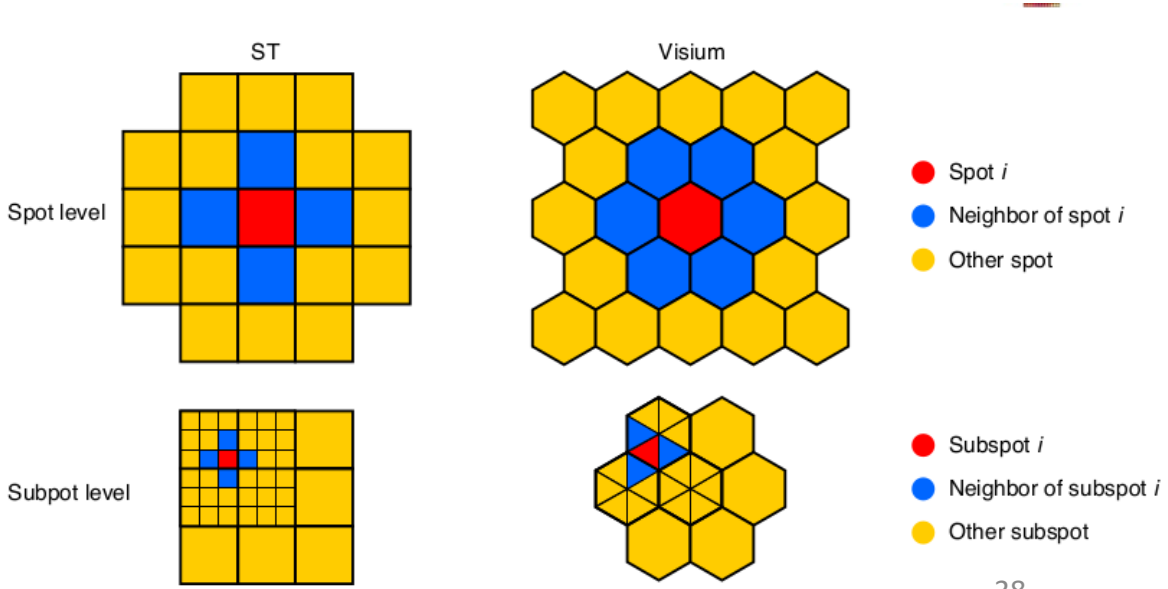
Model based approaches

TOOL : **BayesSpace**(6) Provides the SPOT CLUSTER ASSIGMENTS + SUBS-POT RESOLUTION

if the model **only works with the PCs** of the data:

HOW CAN WE GET SUBSPOT RESOLUTION EXPRESSION MAPS ?





Model based approaches

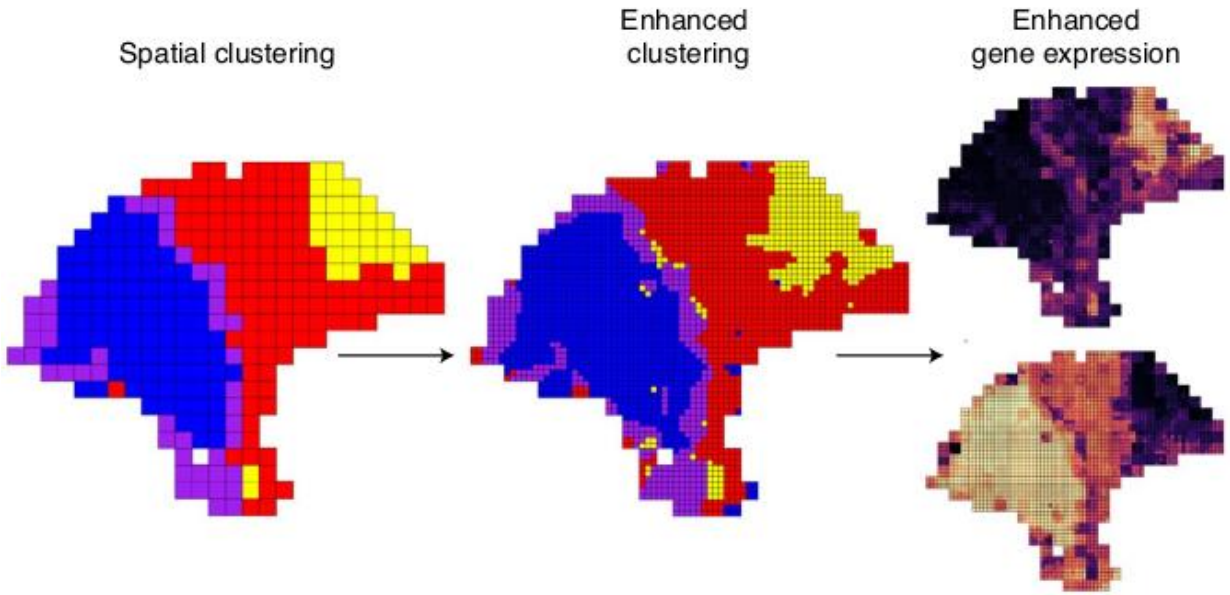
TOOL : **BayesSpace**(6) Provides the SPOT CLUSTER ASSIGMENTS + SUB-SPOT RESOLUTION

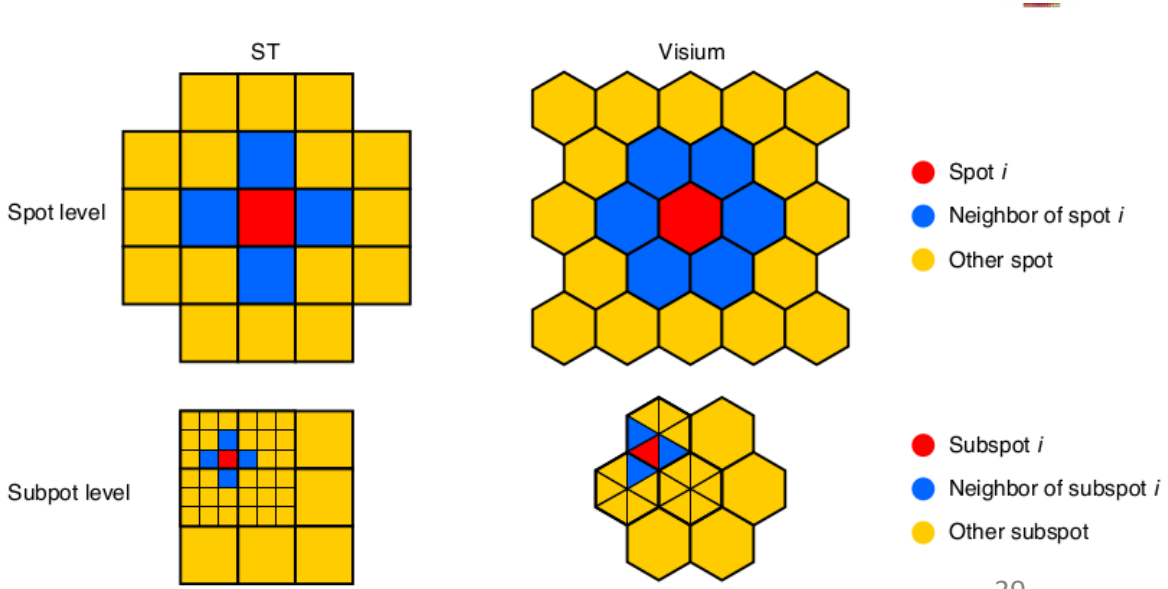
if the model **only works with the PCs** of the data:

HOW CAN WE GET SUBSPOT RESOLUTION EXPRESSION MAPS ?

Need an additional step ➡ Train a model to predict gene expression from PCs on original data

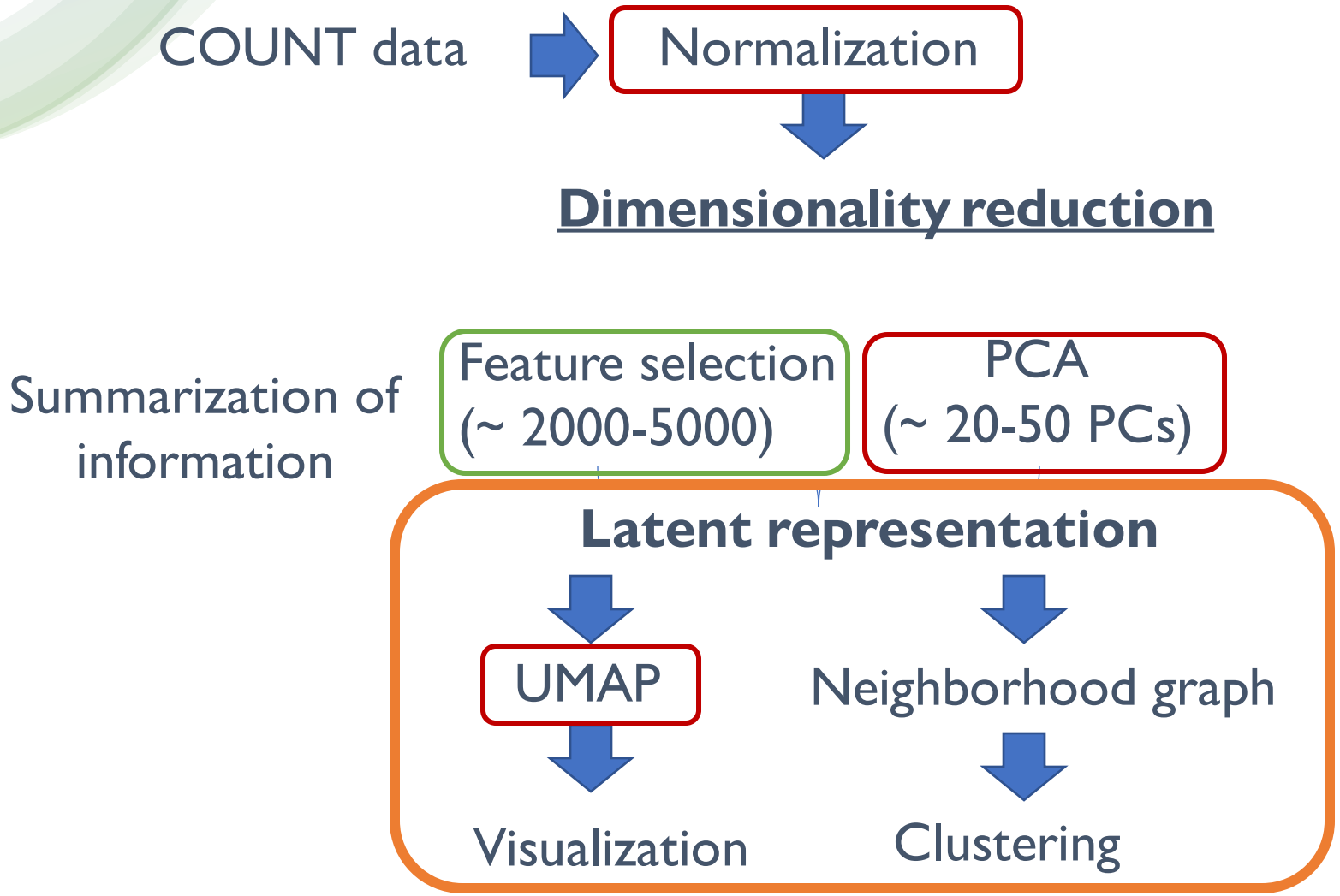
Use this model on the sub-spot PCs values to get Enhanced gene expression maps





Spatially aware unsupervised analysis

Framed steps are tackled or replaced by the same-colored framed methods



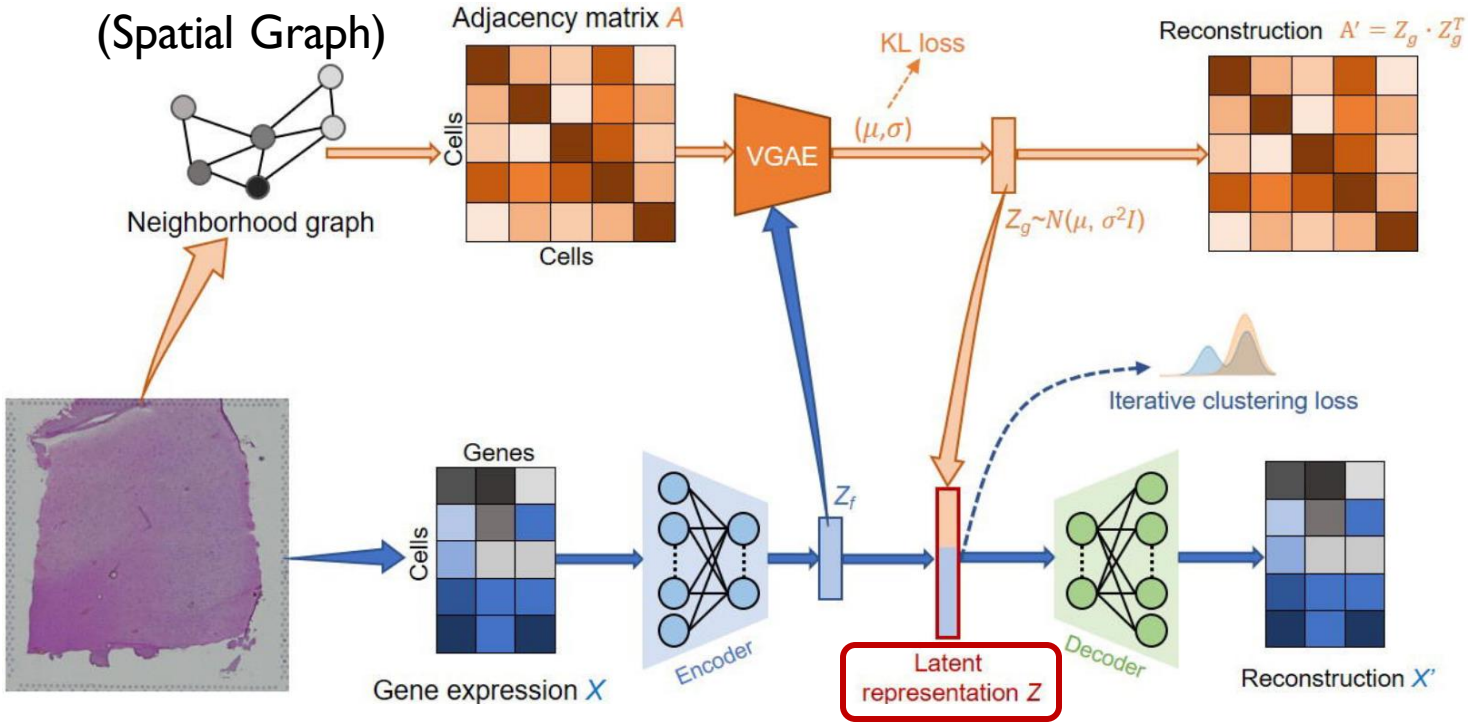
stLearn (SME Normaliz.)

Spatially aware feature selection

BayesSpace

Model based approaches

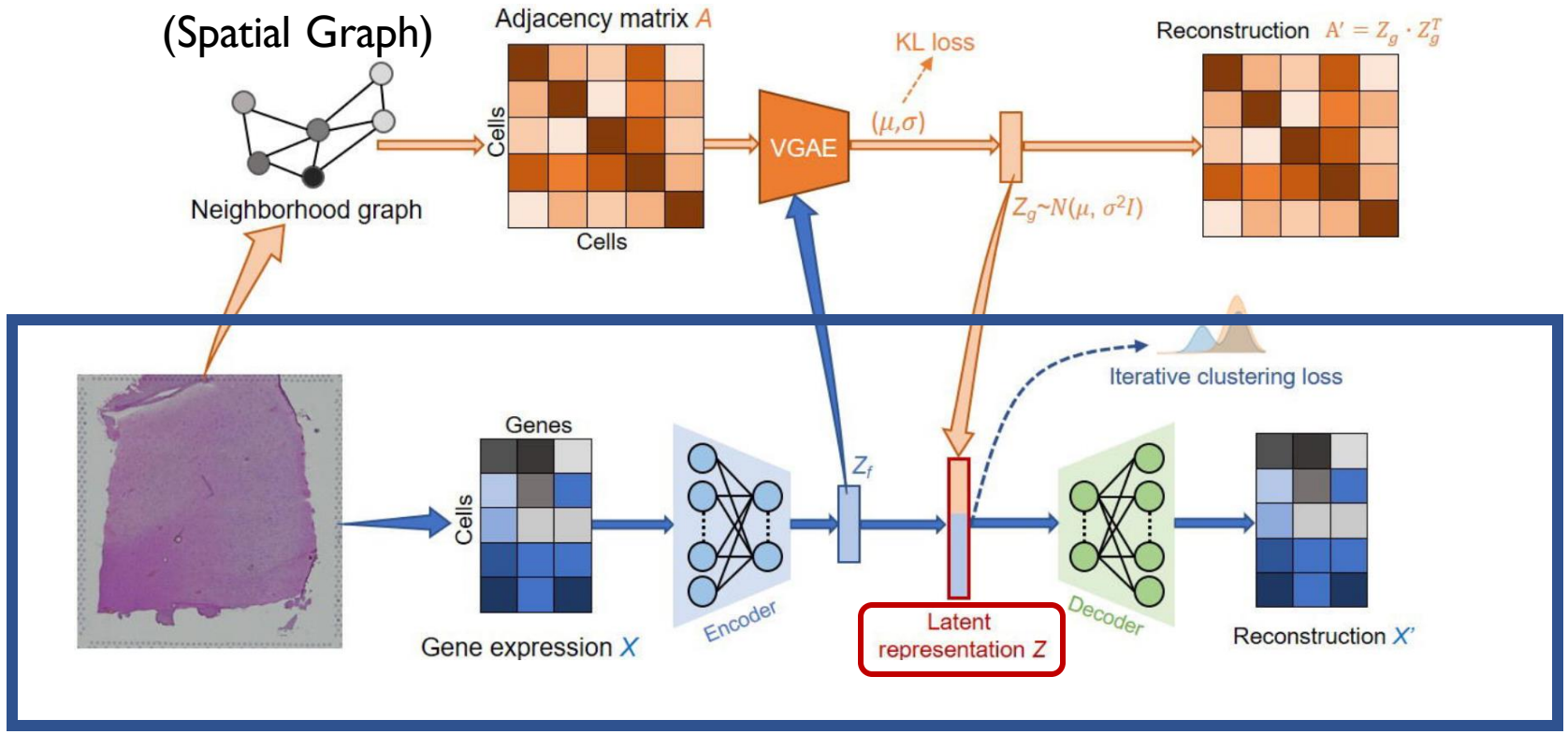
TOOL : Spatially Embedded Dimensionality Reduction (**SEDR**)[7] Provides a Refined Latent Representation



Model based approaches

TOOL : Spatially Embedded Dimensionality Reduction (SEDR)[7] Provides a Refined Latent Representation

Autoencoder

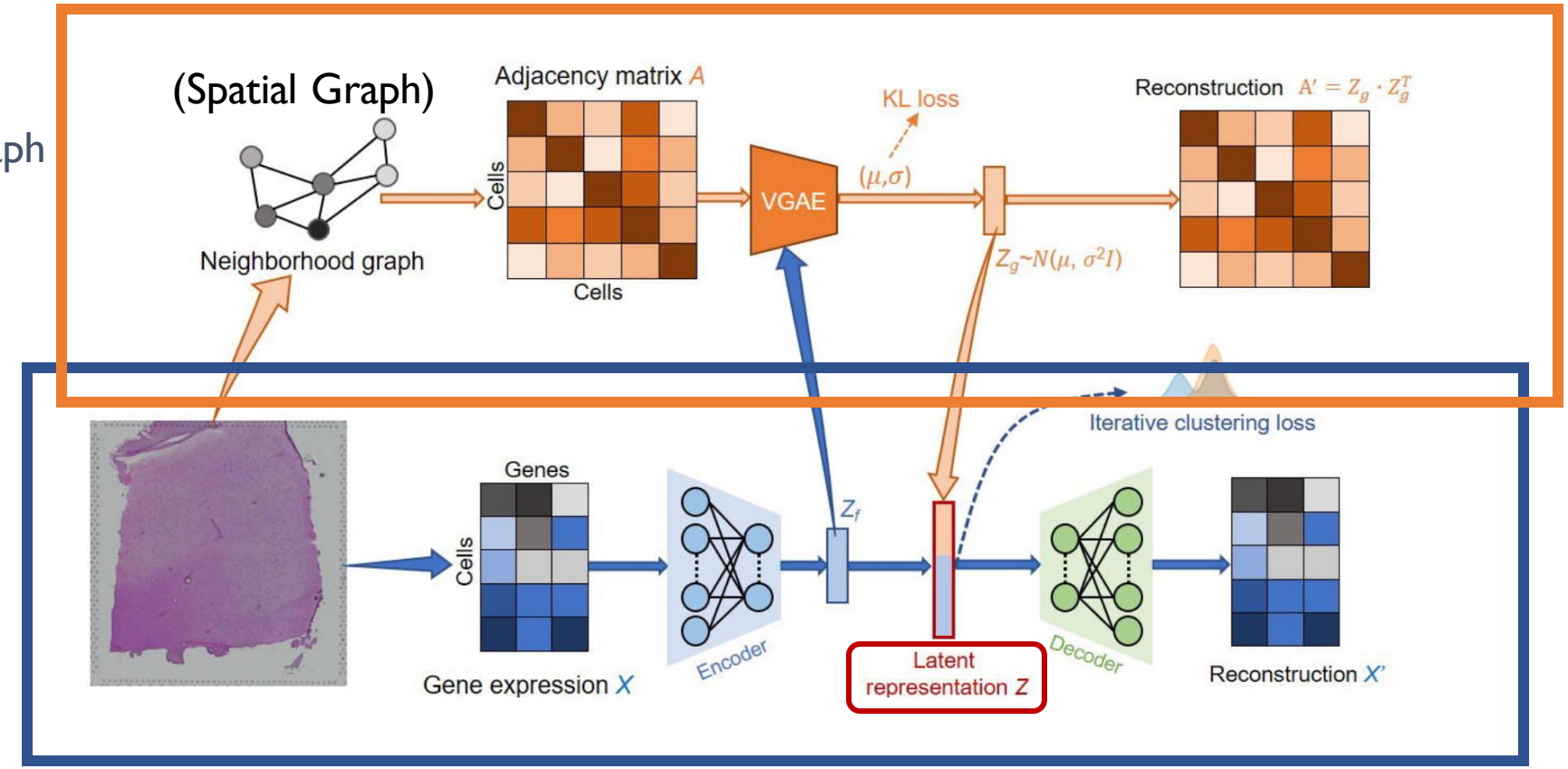


Model based approaches

TOOL : Spatially Embedded Dimensionality Reduction (SEDR)[7] Provides a Refined Latent Representation

Variational graph autoencoder

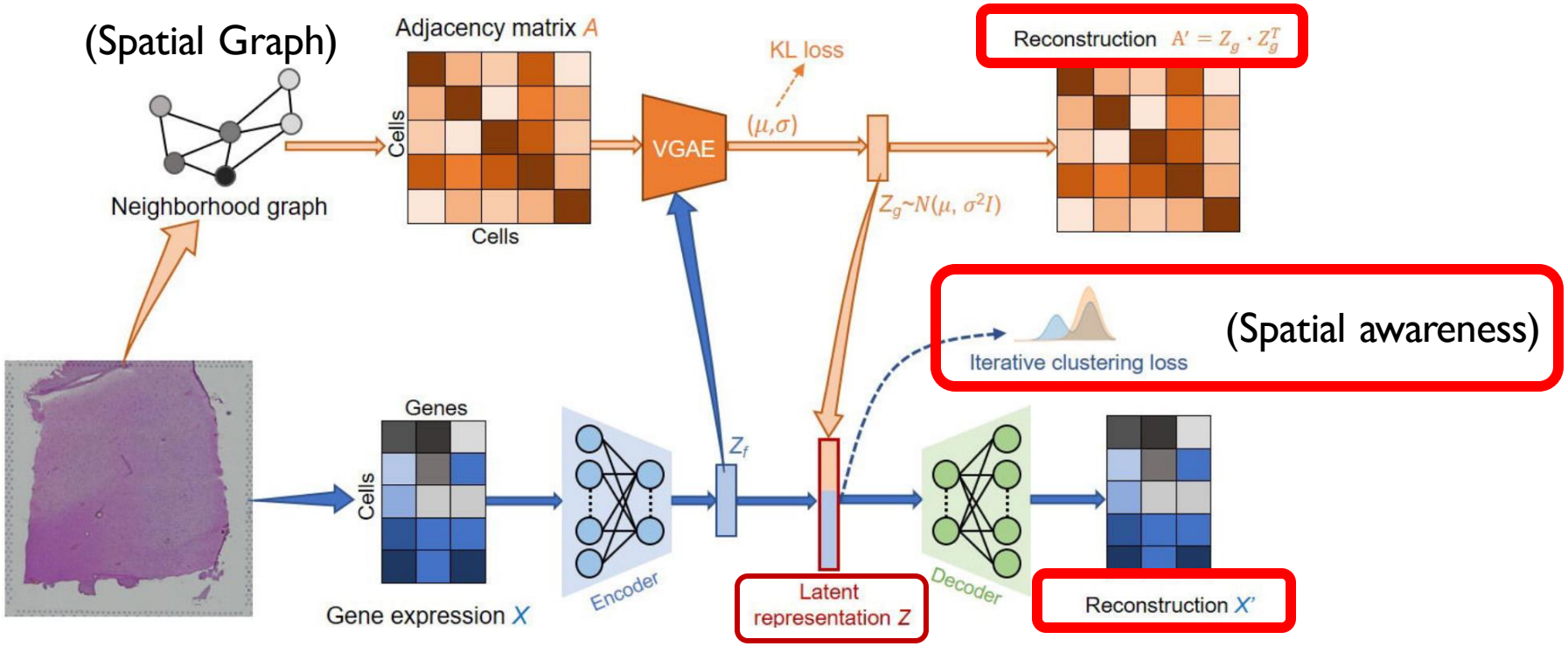
Autoencoder



Model based approaches

TOOL : Spatially Embedded Dimensionality Reduction (SEDR)[7]

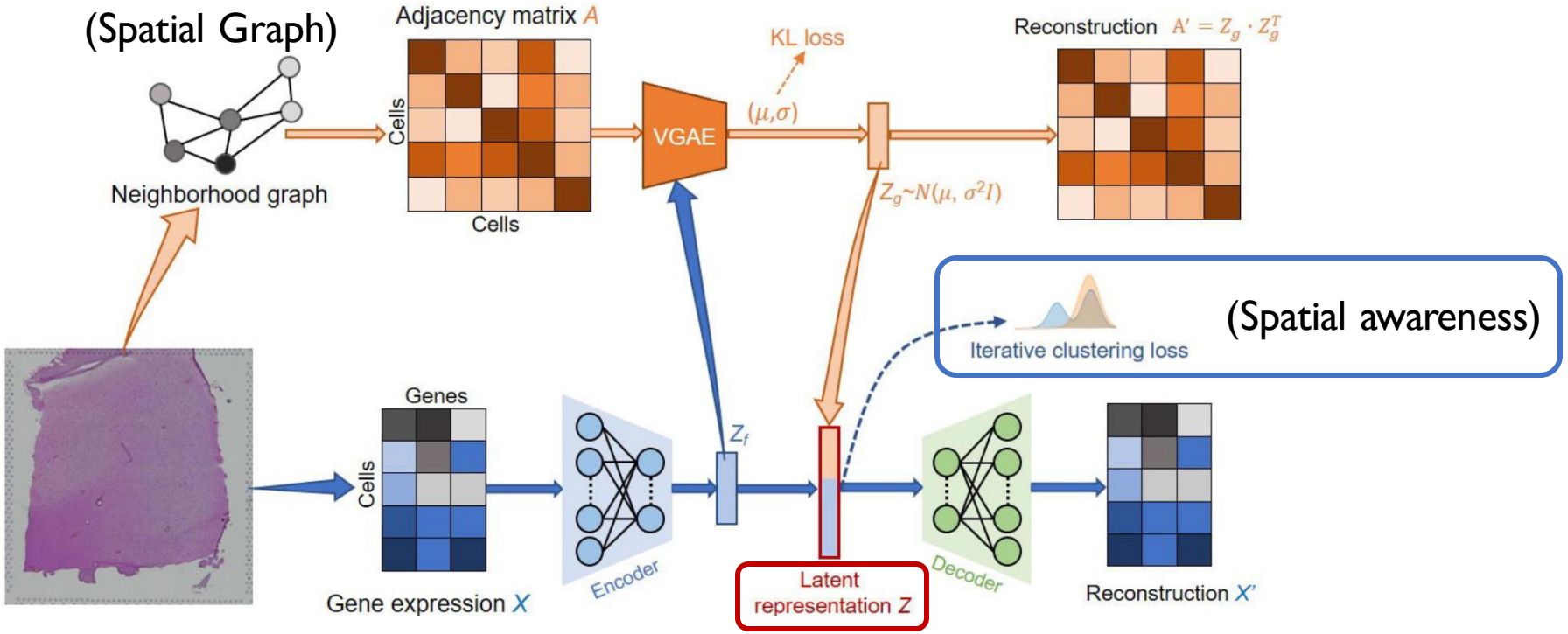
Provides a Refined Latent Representation



Model based approaches

TOOL : Spatially Embedded Dimensionality Reduction (**SEDR**)[7] Provides a Refined Latent Representation

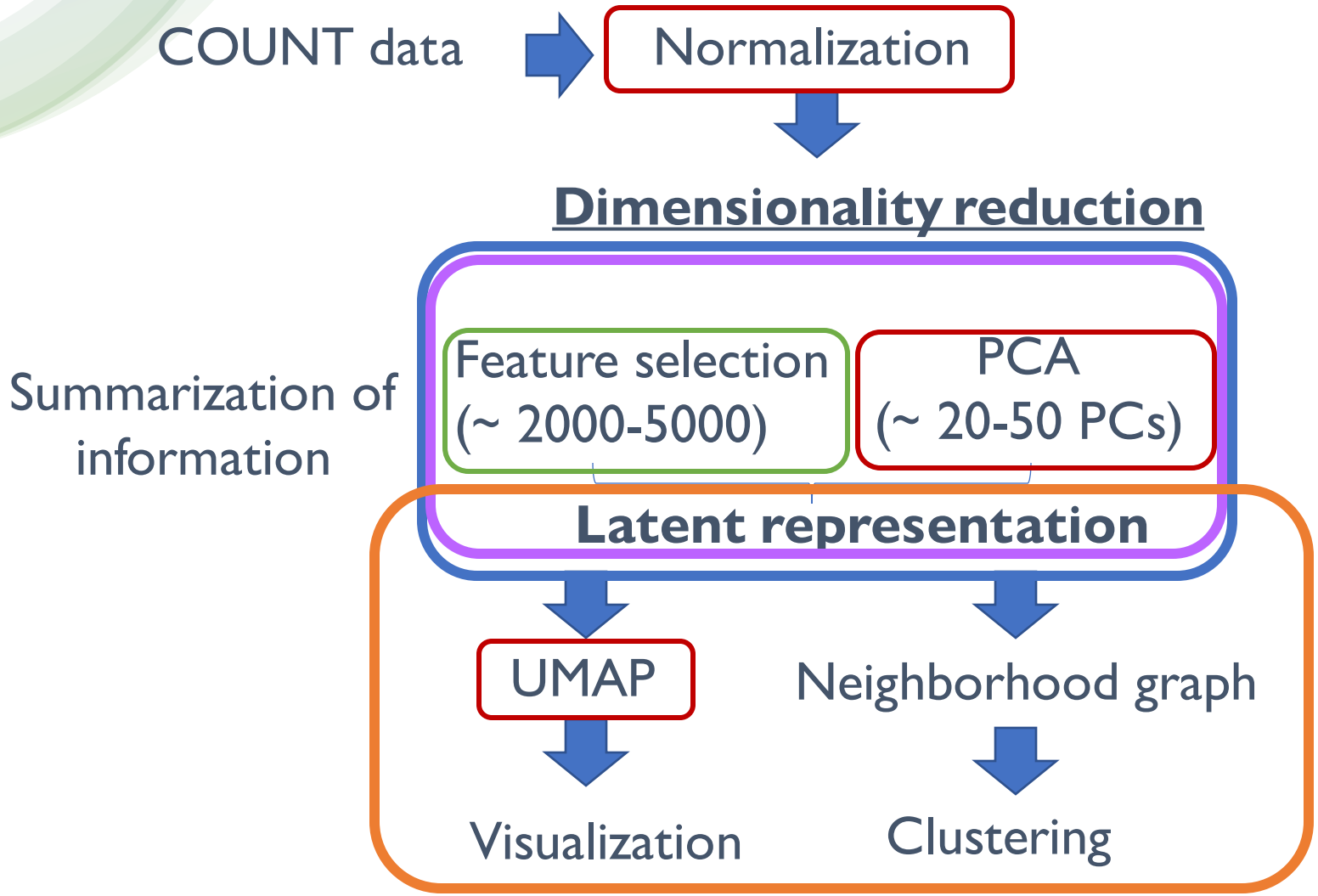
Preprocessing :
Normalize, log-norm. and PCs computation.
Authors recommendation:
~ 300 PCs as input



Latent Representation Z is composed by : - Encoded features of the gene expression +
- Spatial embedded features outputted by the Variational Graph Autoencoder (Spatial Graph)
(encoded features)

Spatially aware unsupervised analysis

Framed steps are tackled or replaced by the same-colored framed methods



stLearn (SME Normaliz.)

Spatially aware feature selection

BayesSpace

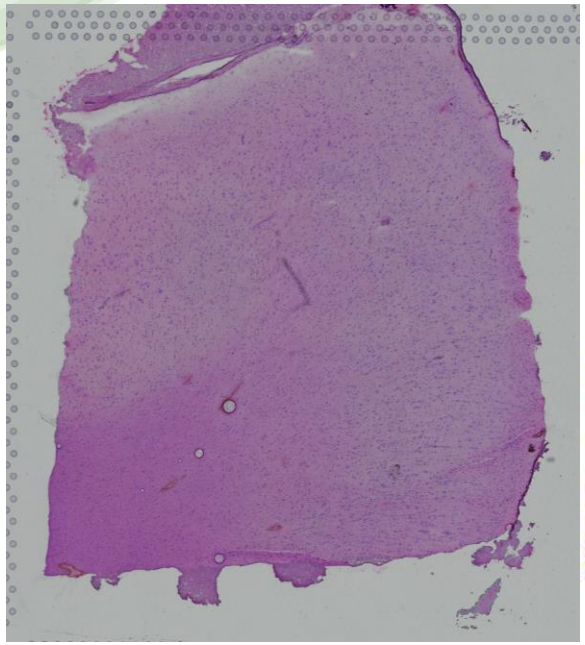
SEDR

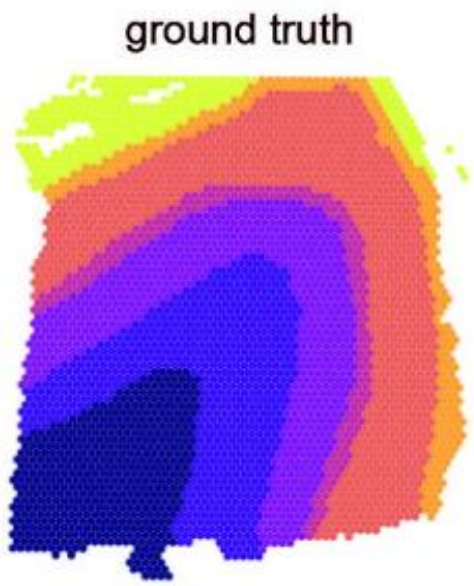
Guideline

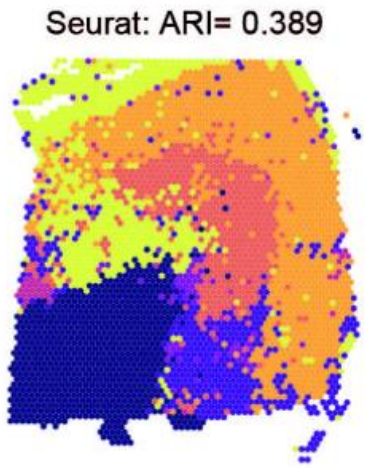
- Data overview
- Pipeline overview
- Leverage new data modalities
- Spatially aware unsupervised analysis
    - Data normalization
    - Feature selection
    - Model based
  - **Benchmarking and further specific results**
- Latent representation from reference-based deconvolution
- Bonus: Deep Data Fusion

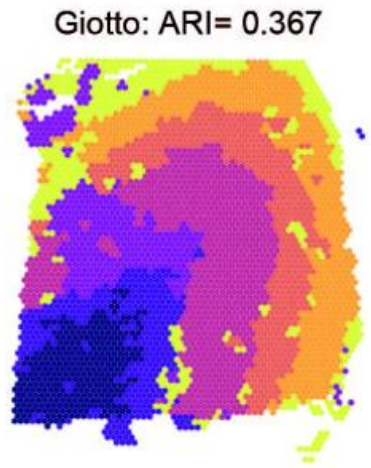
Benchmarking results - DLPFC

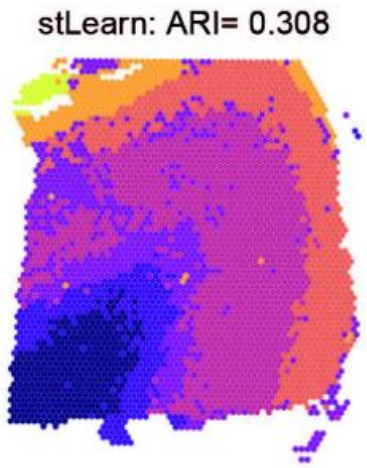
Anatomical structure detection

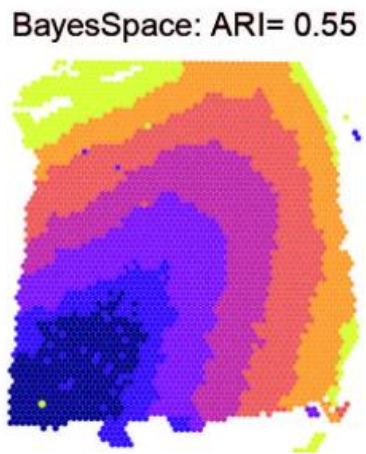


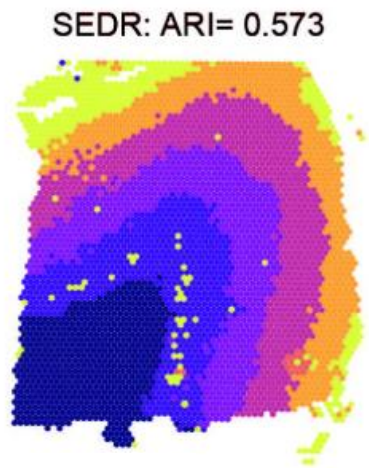




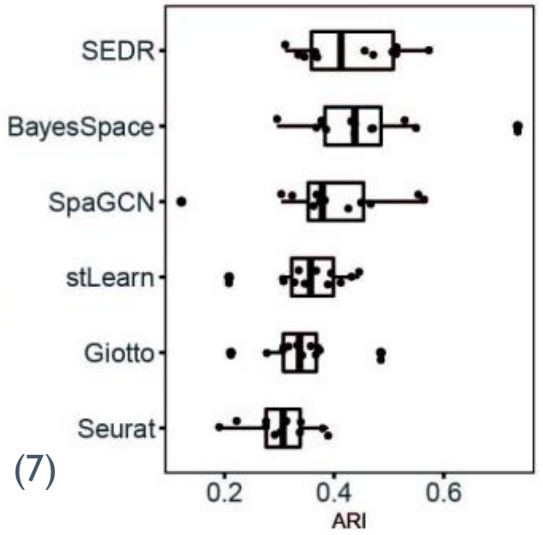








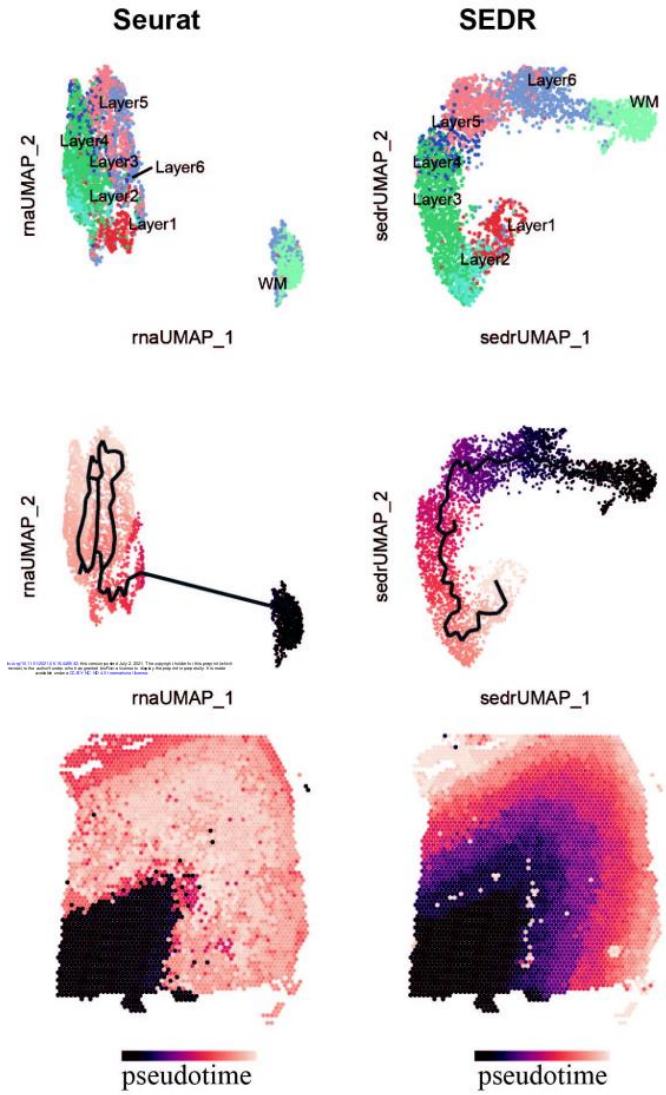
(7)



Benchmarking results - DLPFC

SEDR - Trajectory Inference

Monocle3
Tool for pseudotime estimation



SEDR results reflected the correct "inside-out" developmental ordening of cortical layers

**In tumoral samples pseudo-time can show the tumor progression

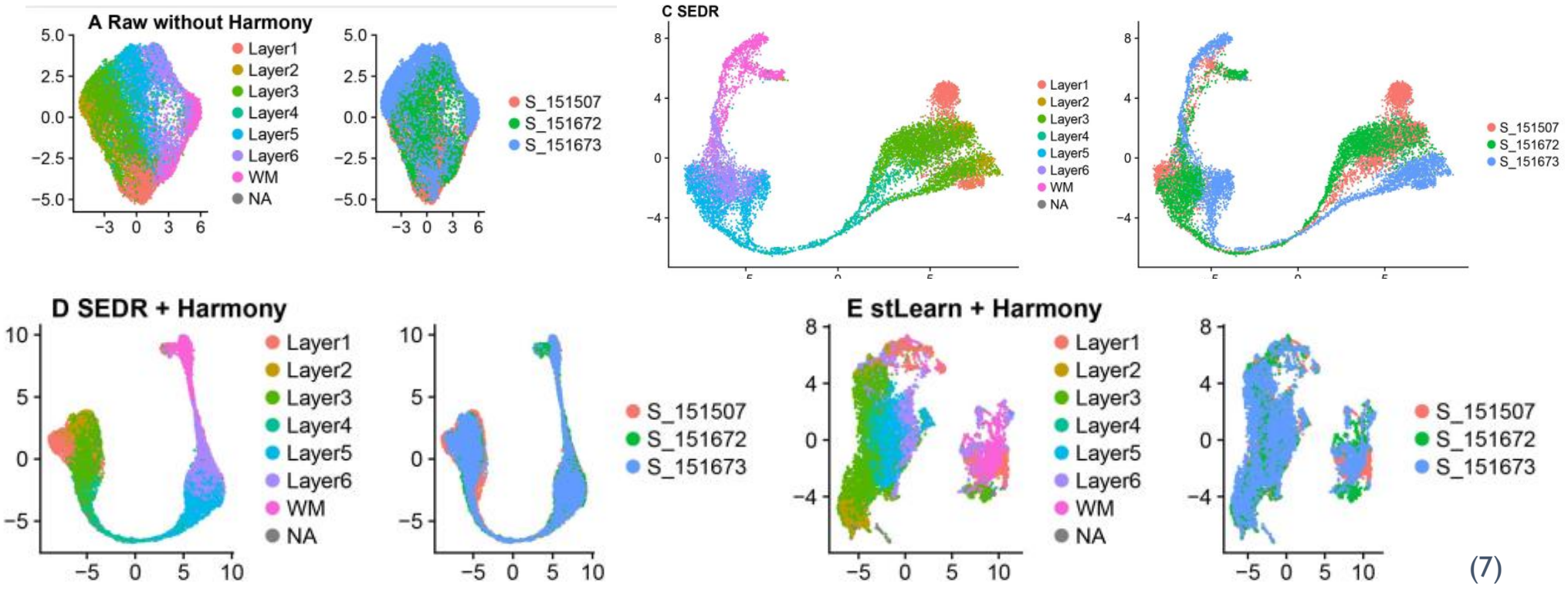
(7)

Benchmarking results - DLPFC

SEDR - Trajectory Inference & Batch effect

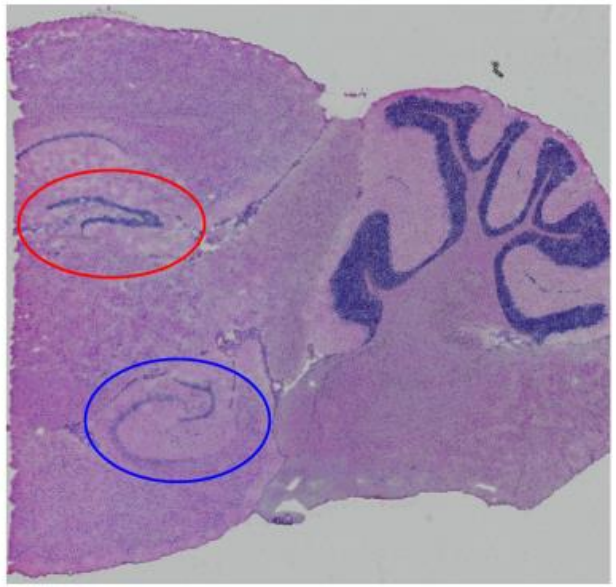
Harmony

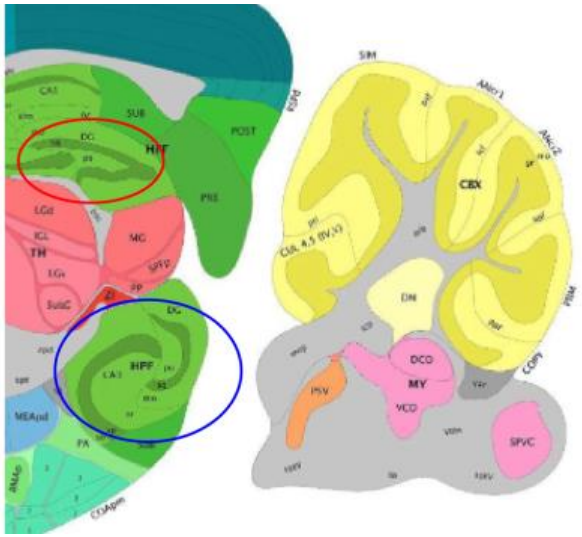
Batch effect removal tool that aligns the samples information in the PC space



(7)

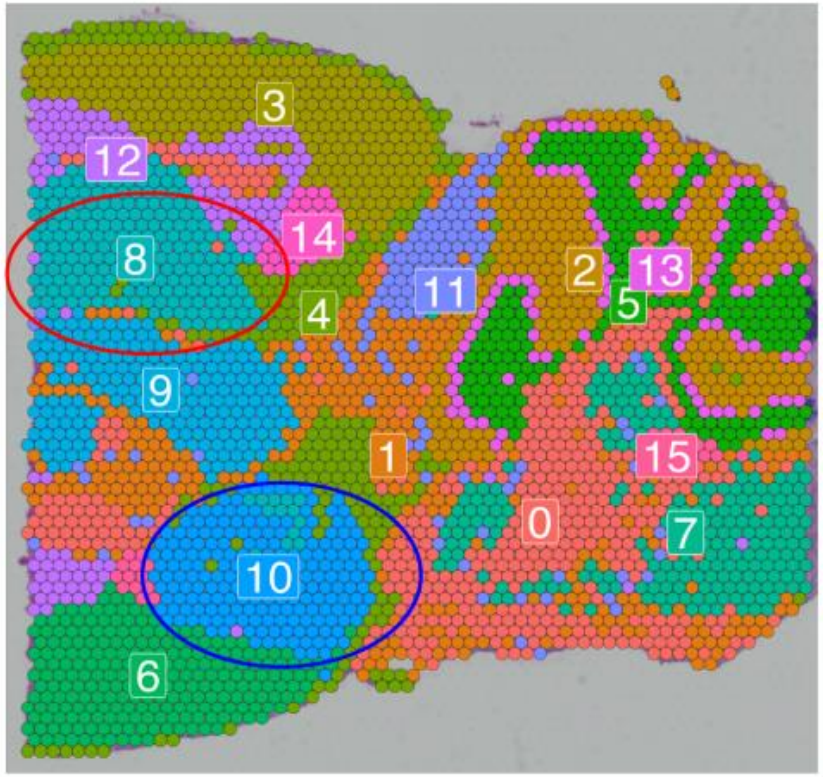
Benchmarking results – Saggital Posterior



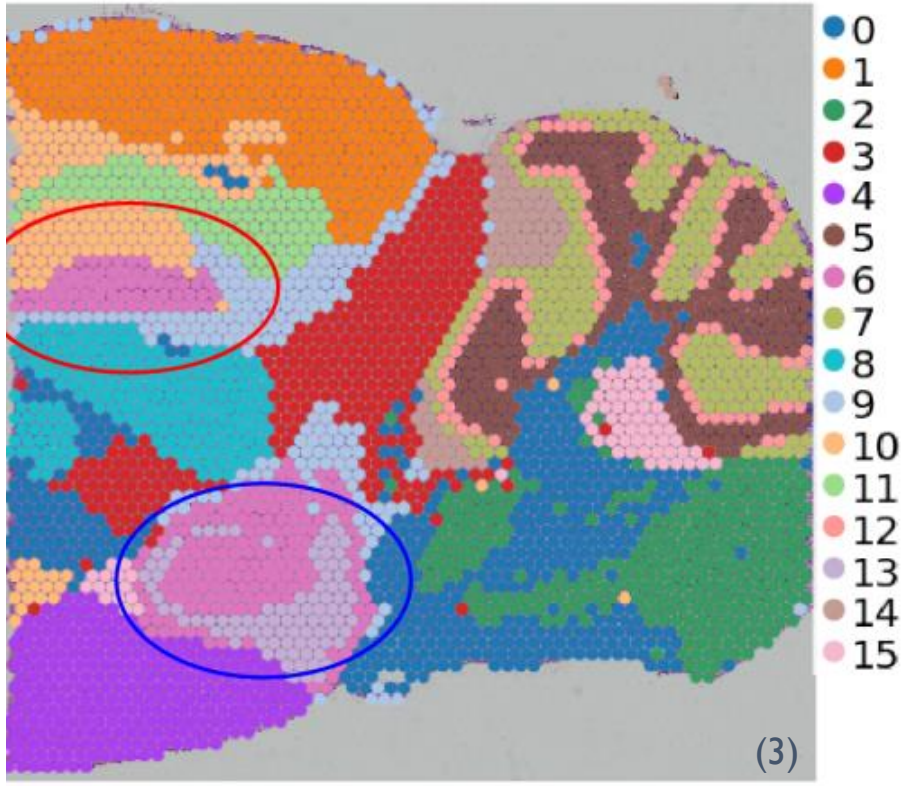


Complex anatomical regions detection – Dentate gyrus

Seurat / Scanpy

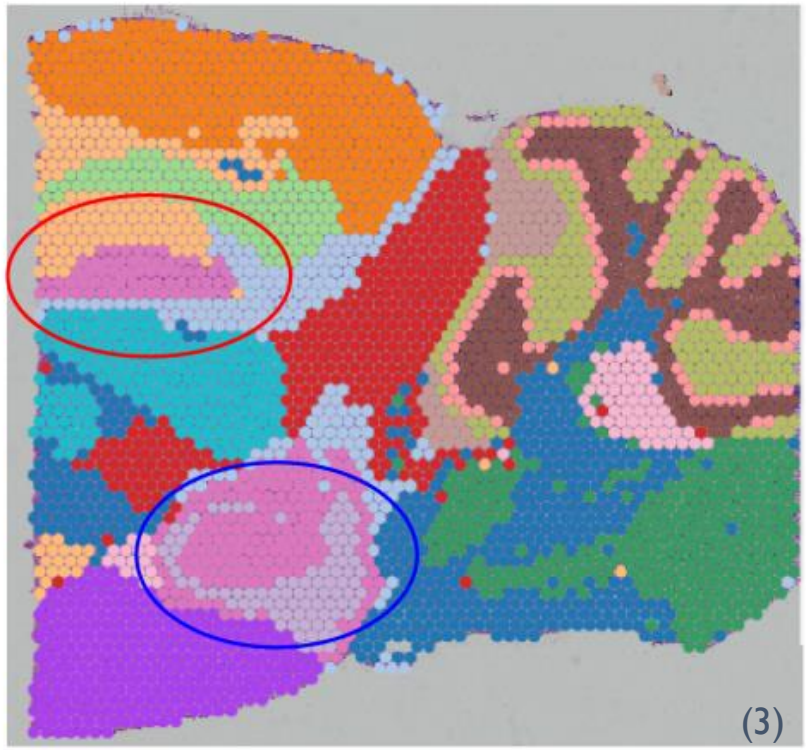


stLearn

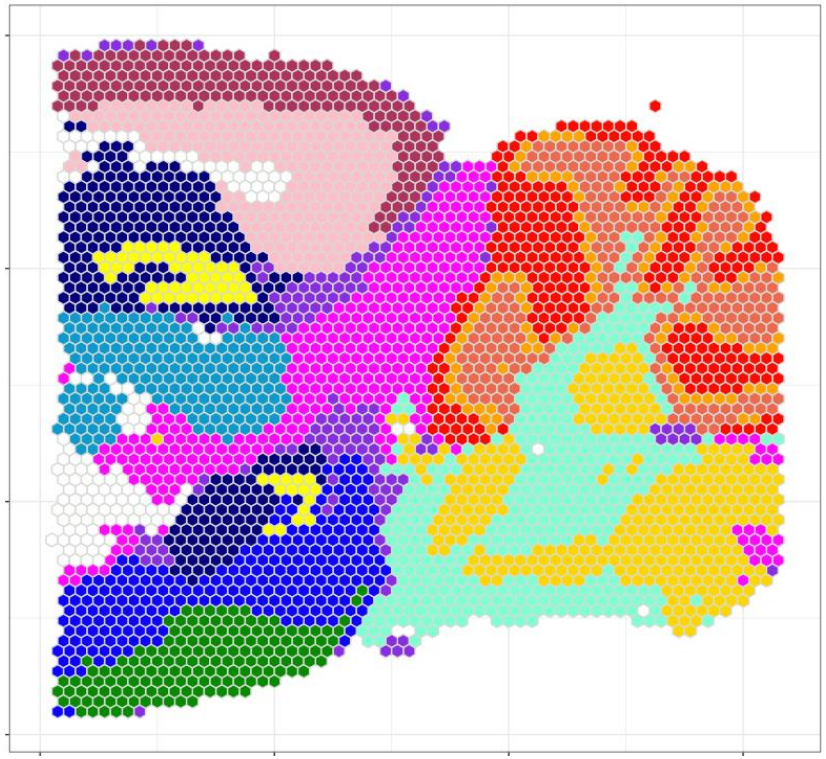


Benchmarking results – Saggital Posterior

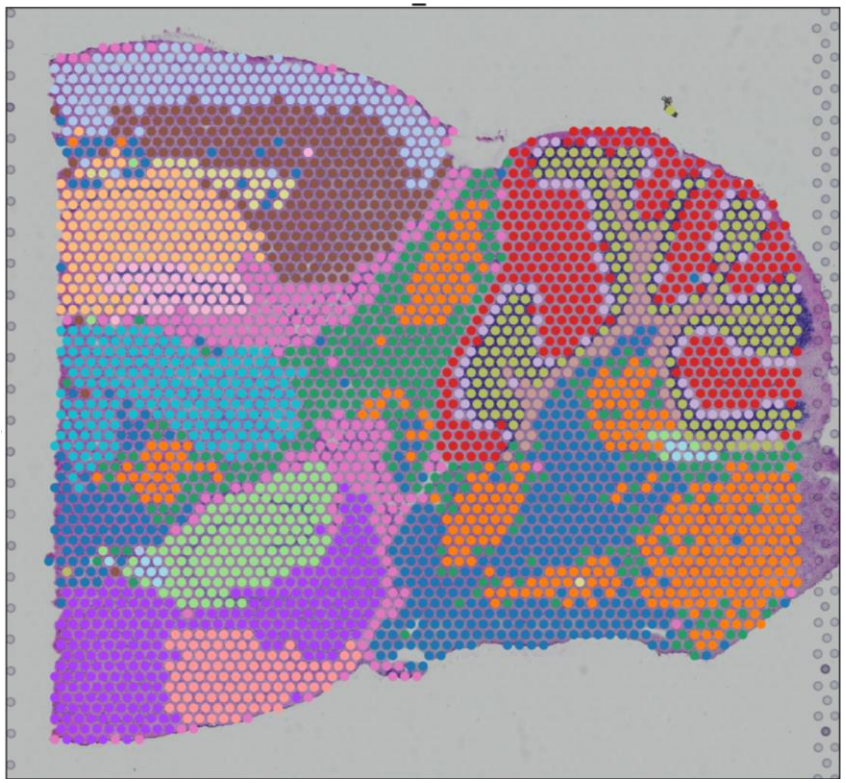
stLearn



BayesSpace

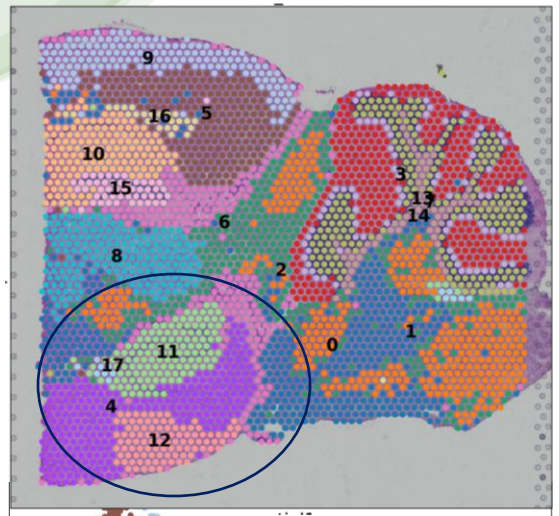


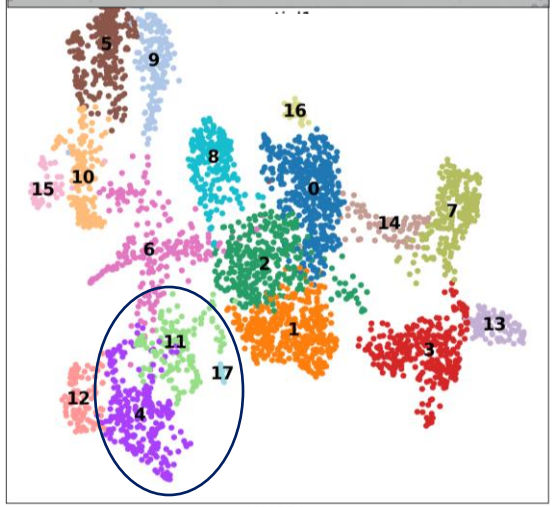
SEDR



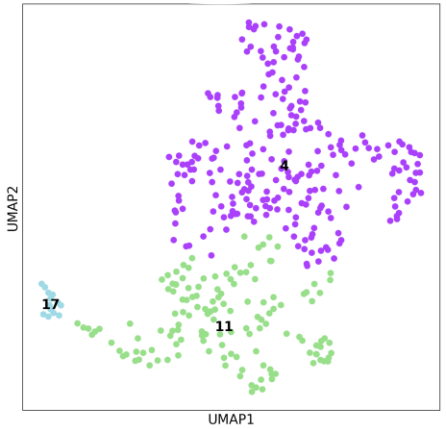
Specific results – SEDR

Allows for user-supervised re-clustering

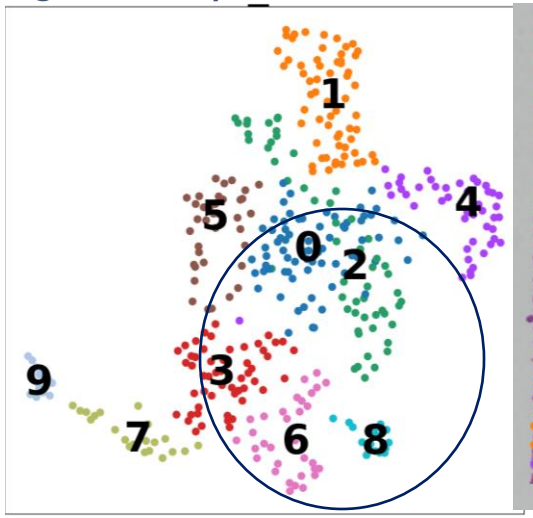


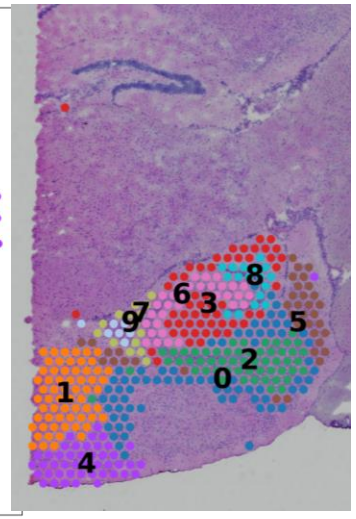


1.- Isolate clusters of interest (4,11,17)

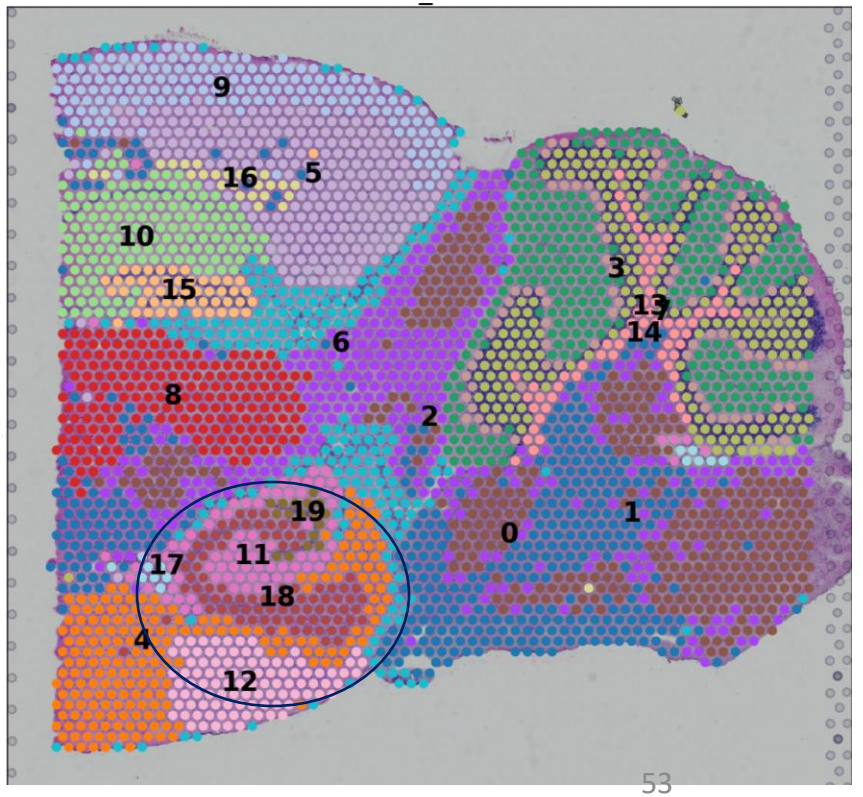


2.- Compute *finer granularity clusters*





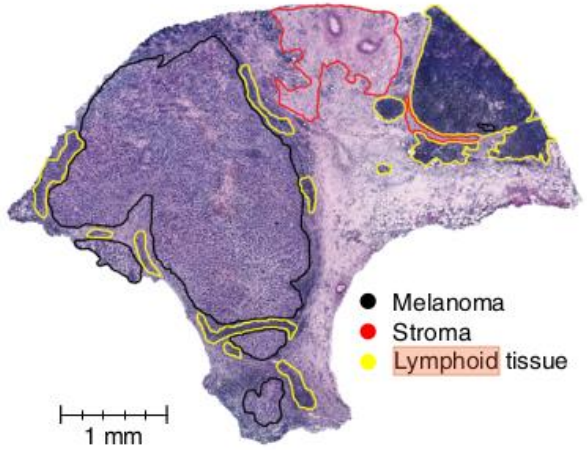
3.- Select finer granularity clusters of interest and merge.
2, 6 ➡ 18 ; *8* ➡ 19

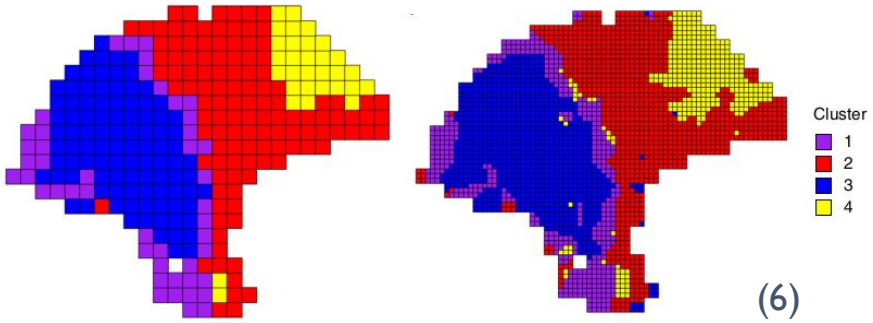


Specific results – BayesSpace

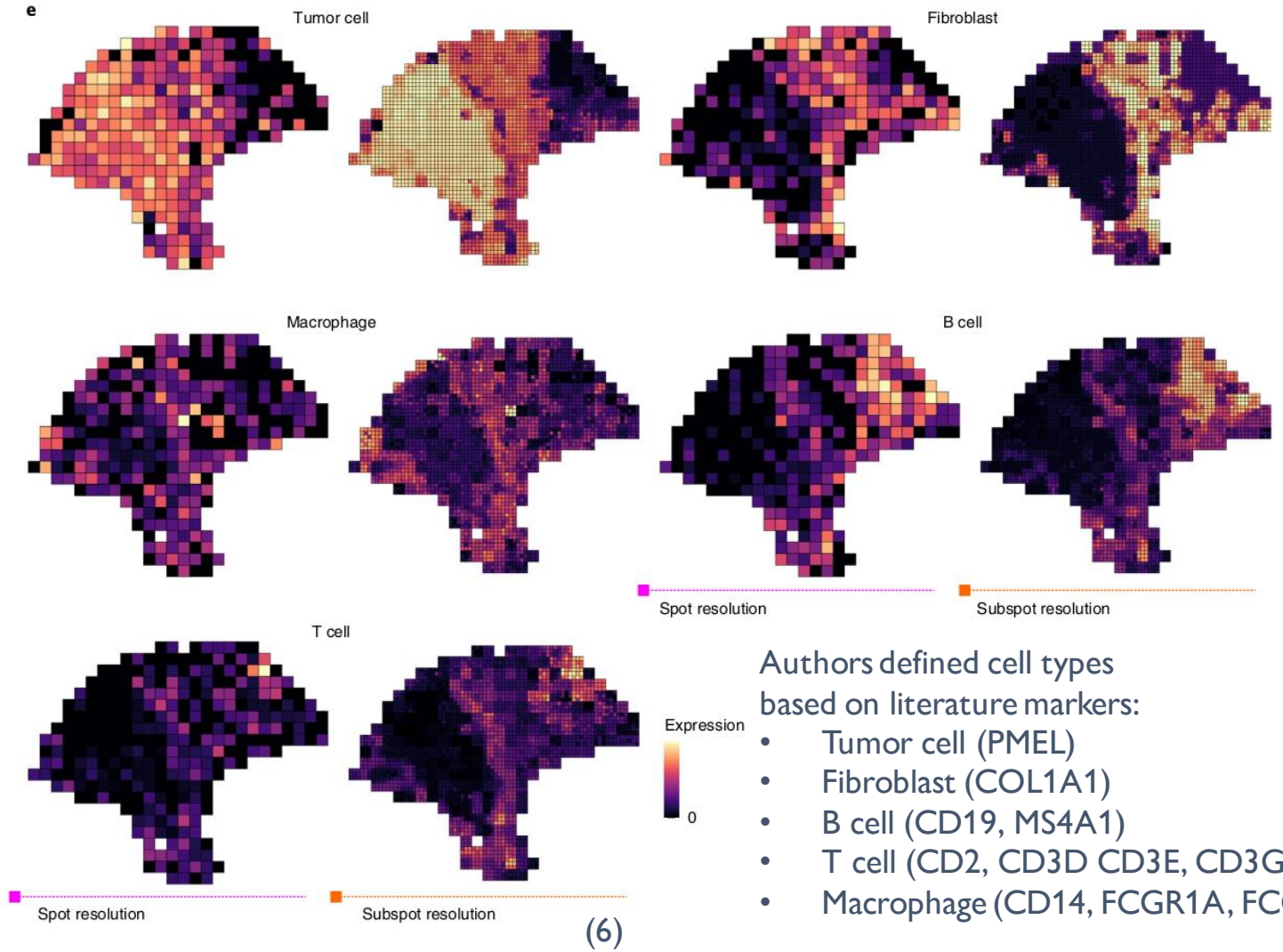
Melanoma sample

Tumor-proximal lymphoid tissue





(6)



(6)

Authors defined cell types based on literature markers:

- Tumor cell (PMEL)
- Fibroblast (COL1A1)
- B cell (CD19, MS4A1)
- T cell (CD2, CD3D CD3E, CD3G, CD7)
- Macrophage (CD14, FCGR1A, FCGR1B)

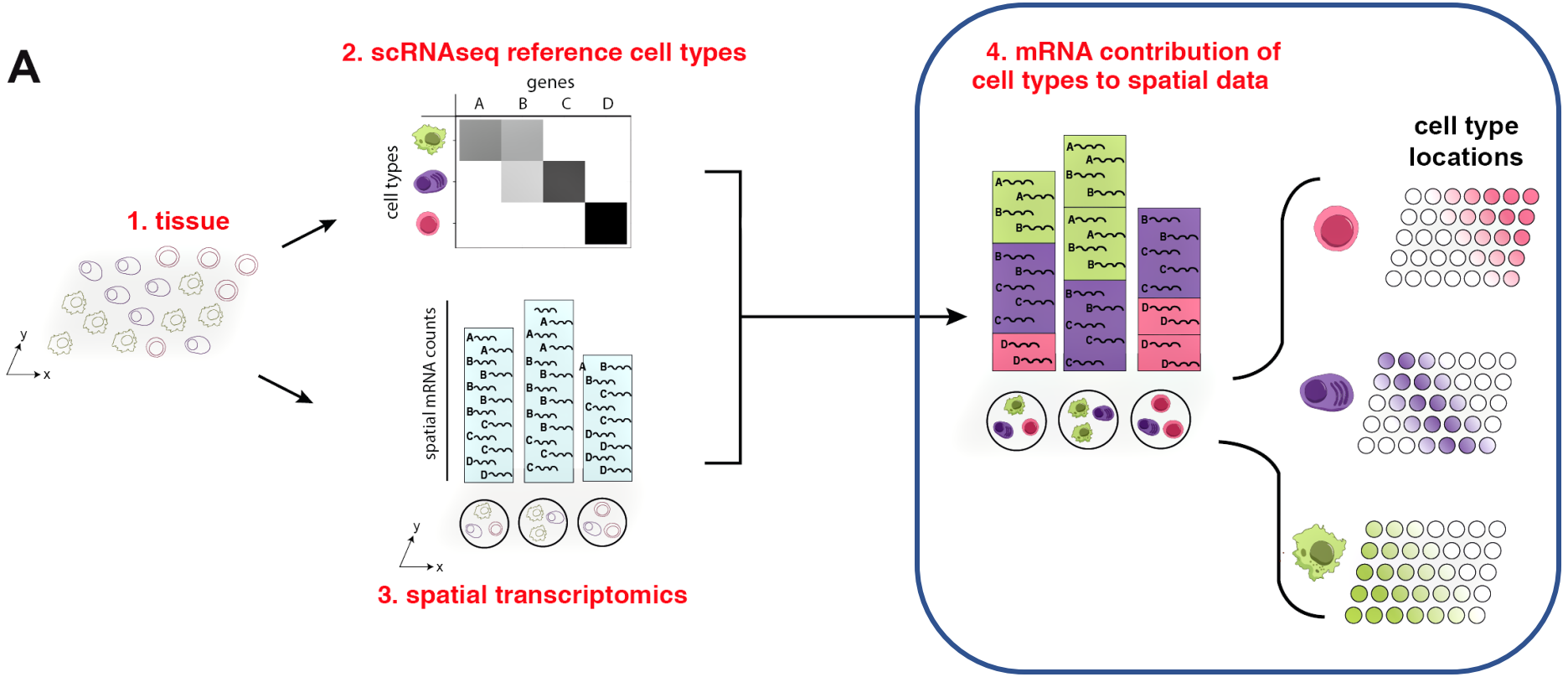
Guideline

- Data overview
- Pipeline overview
- Leverage new data modalities
- Spatially aware unsupervised analysis
    - Data normalization
    - Feature selection
    - Model based
  - Benchmarking and further specific results
- **Latent representation from reference-based deconvo**
- Bonus: Deep Data Fusion

Latent representation from reference-based deconvolution

ST technologies not always provide single cell resolution

Is common to perform spot deconvolution:



Here we obtain a spot representation based on the cell type proportion

Latent representation of the data

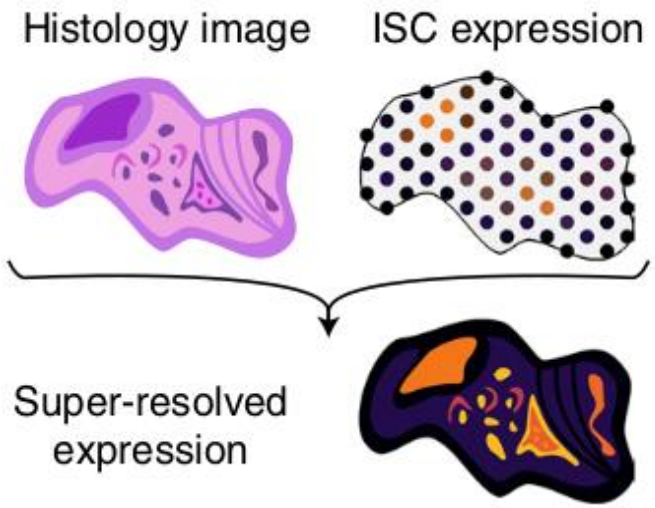
Guideline

- Data overview
- Pipeline overview
- Leverage new data modalities
- Spatially aware unsupervised analysis
    - Data normalization
    - Feature selection
    - Model based
  - Benchmarking and further specific results
- Latent representation from reference-based deconvolution
- **Bonus: Deep Data Fusion**

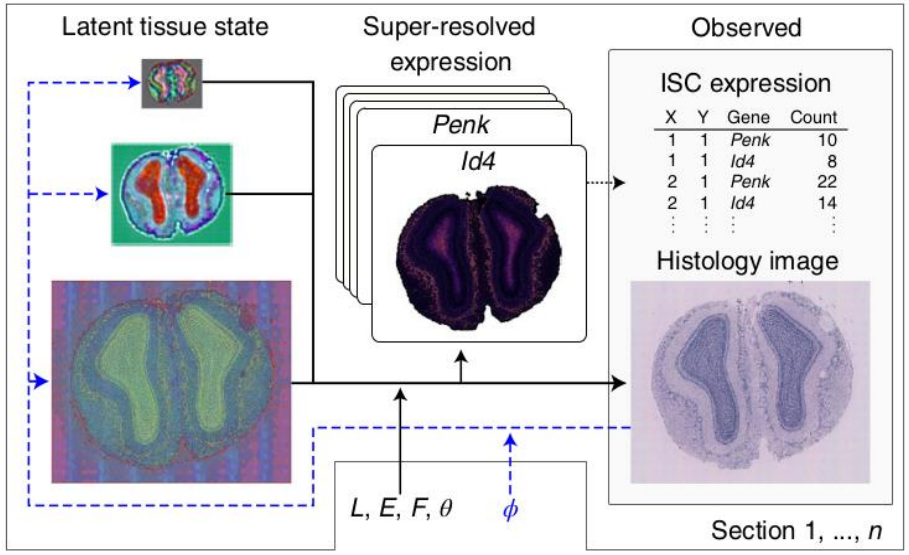
Deep Data Fusion

- Deep generative model that merge ideas from computer vision and generative statistical modeling

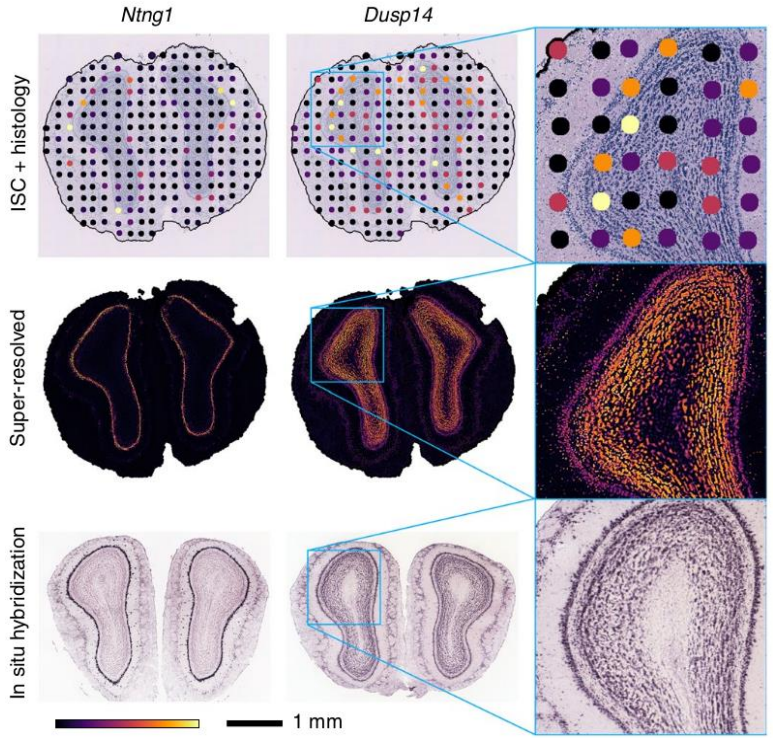
Provides SUPER-RESOLVED gene expression maps



Model scheme



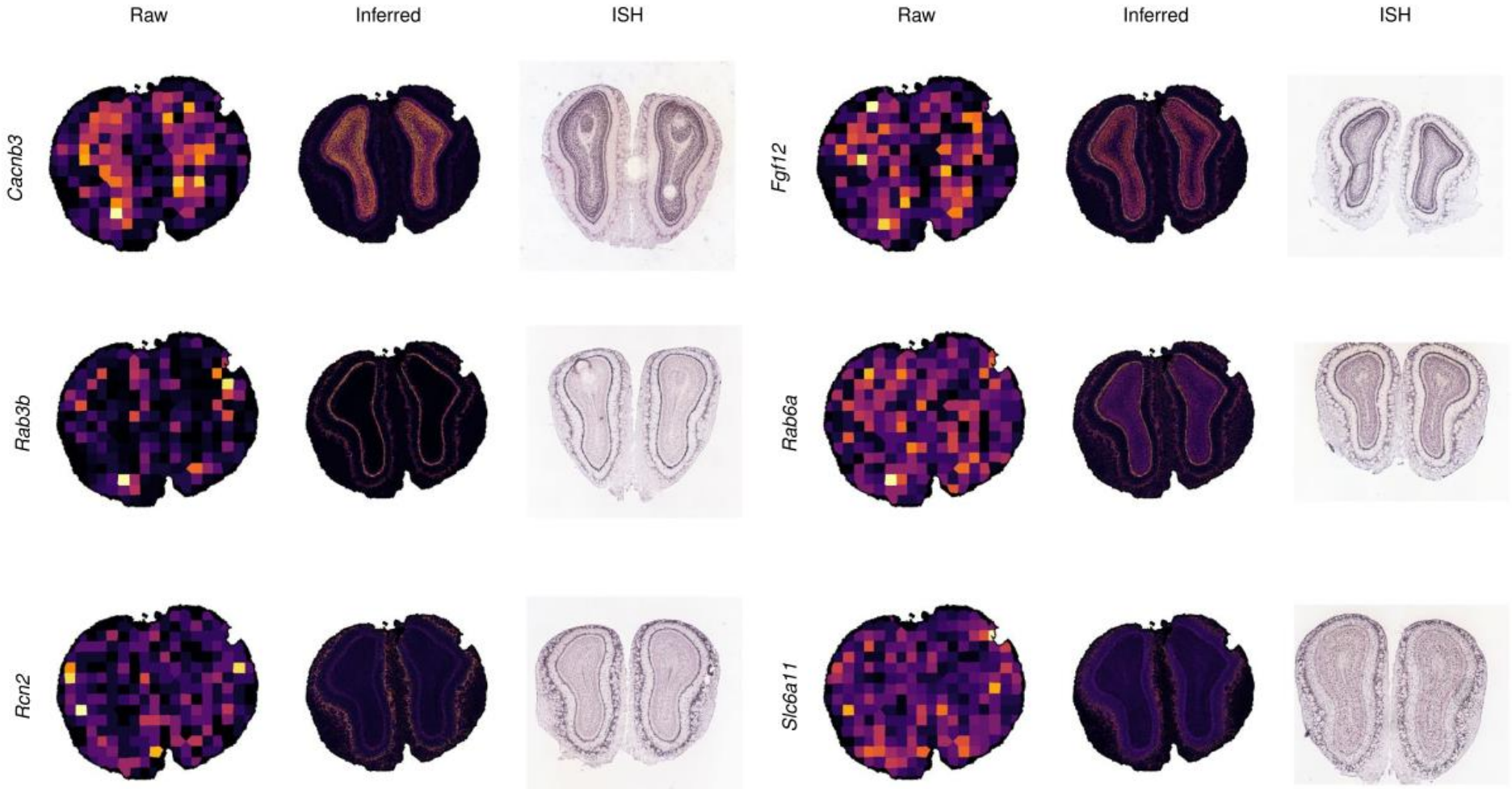
(8)



Spatial approaches

Model based approaches

TOOL : **Deep Data Fusion** Provides SUPER-RESOLVED gene expression maps



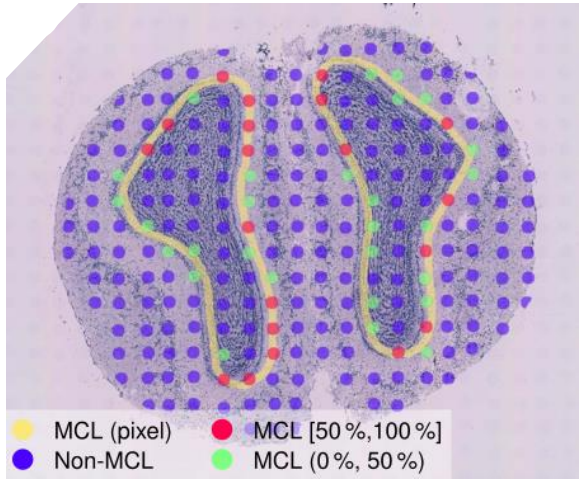
(8)

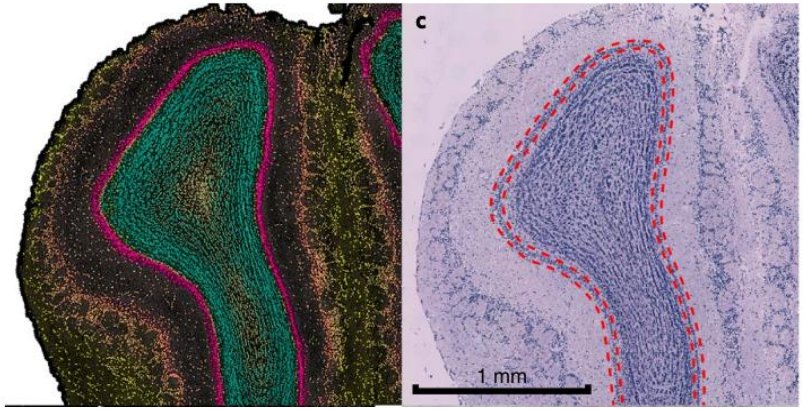
Spatial approaches

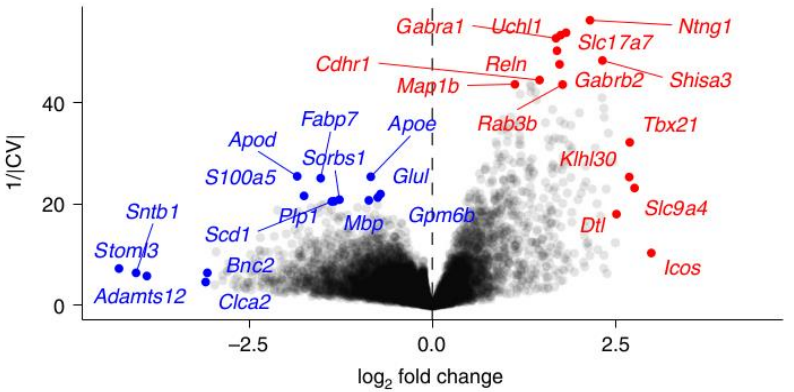
Model based approaches

TOOL : **Deep Data Fusion** Provides SUPER-RESOLVED gene expression maps

- Instead of clustering assigment we have a latent tissue state
- We can still run differential expression analysis: we select measurement locations overlapping with an annotation region (eg. Mitral Cell Layers) . Then, we log-normalize the data and compute differentially expressed genes using the FindMarkers (Seurat) function







(8)

METHODS SUMMARY

	INPUT	OUTPUT
stLearn	RAW ST + Histological Image	Spatial Morphological Gene Expression Normalization
SPARK(-X) // GIOTTO	RAW ST DATA	Spatially Variable Genes
BayesSpace	(~15) Top PCs from HVG	Cluster labels + Sub-spot resolution
SERD	(~300) Top PCs from ALL GENES	Spatially Embedded Latent Representation
Spot deconvolution	RAW ST DATA + Annotated sc-RNAseq reference	Cell type proportion (alternative latent representation)
Deep data fusion	RAW ST + Histological Image	Super-resolved gene expression maps

JOSEP CARRERAS LEUKAEMIA RESEARCH INSTITUTE
For a future without leukaemia

THE END

Thanks for your attention!

Bibliography

- (1) Luecken, M. D., & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular systems biology*, *15*(6), e8746.
- (2) Palla, G., Spitzer, H., Klein, M., Fischer, D., Schaar, A. C., Kuemmerle, L. B., ... & Theis, F. J. (2021). Squidpy: a scalable framework for spatial single cell analysis. *BioRxiv*.
- (3) Pham, D., Tan, X., Xu, J., Grice, L. F., Lam, P. Y., Raghubar, A., ... & Nguyen, Q. (2020). stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. *BioRxiv*.
- (4) Sun, S., Zhu, J., & Zhou, X. (2020). Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nature methods*, *17*(2), 193-200.
- (5) Zhu, J., Sun, S., & Zhou, X. (2021). SPARK-X: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. *Genome Biology*, *22*(1), 1-25.
- (6) Zhao, E., Stone, M. R., Ren, X., Guenthoer, J., Smythe, K. S., Pulliam, T., ... & Gottardo, R. (2021). Spatial transcriptomics at subspot resolution with BayesSpace. *Nature Biotechnology*, 1-10.
- (7) Chen, J., Fu, H., Hang, X. U., Chong, K., Li, M., Ang, K. S., ... & Liu, L. (2021). Unsupervised Spatially Embedded Deep Representation of Spatial Transcriptomics.
- (8) Bergenstråhle, L., He, B., Bergenstråhle, J., Abalo, X., Mirzazadeh, R., Thrane, K., ... & Maaskola, J. (2021). Super-resolved spatial transcriptomics by deep data fusion. *Nature biotechnology*, 1-4.