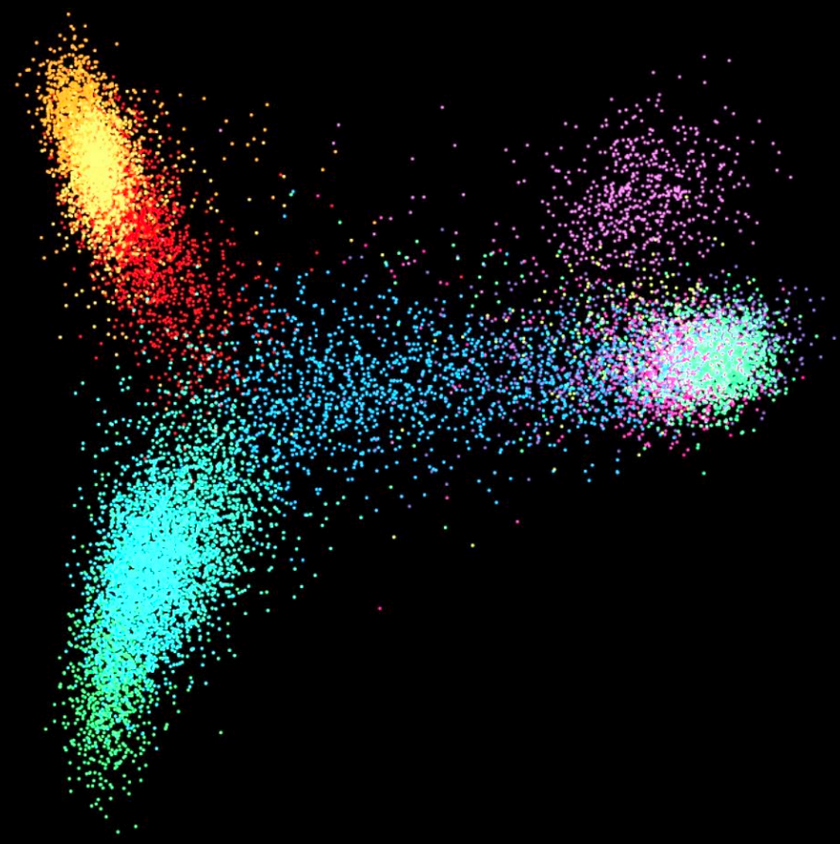
Day 3

Module 1 Final Demo

- 10 min.
- Demo your code on a real dataset (optional task point if you apply it on a new dataset).
- Explain your design decisions.
- Show us the additional features you implemented and exploratory tasks that you did.

Linear representations
(PCA, ICA, …)

Easy comparison & integration across datasets



Nonlinear representations
(t-SNE, UMAP…)

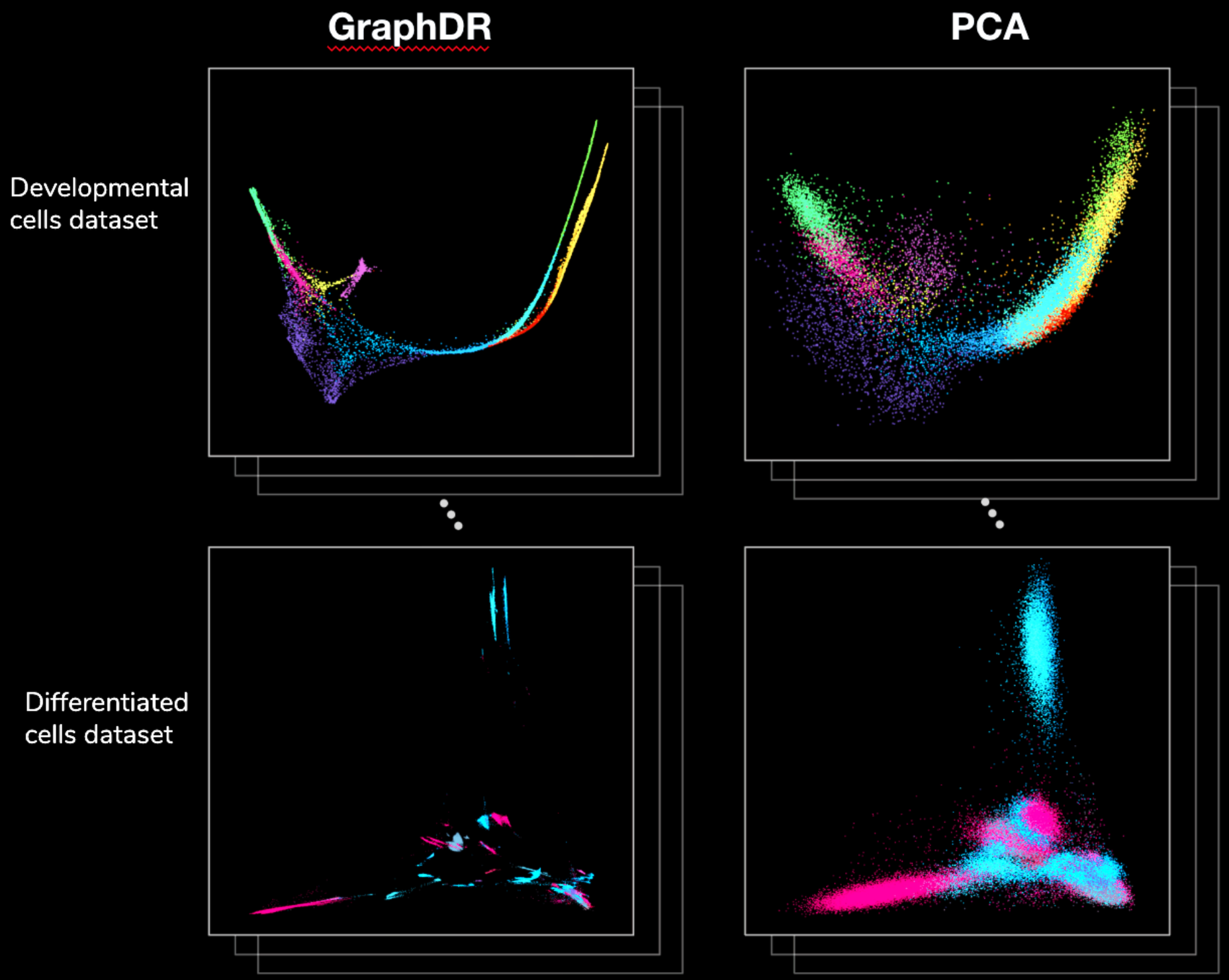
Good representation of cell types / states

Can we combine advantages of linear and nonlinear representations?

Our solution: from linear to **quasilinear** representation

“Quasilinear” approaches:

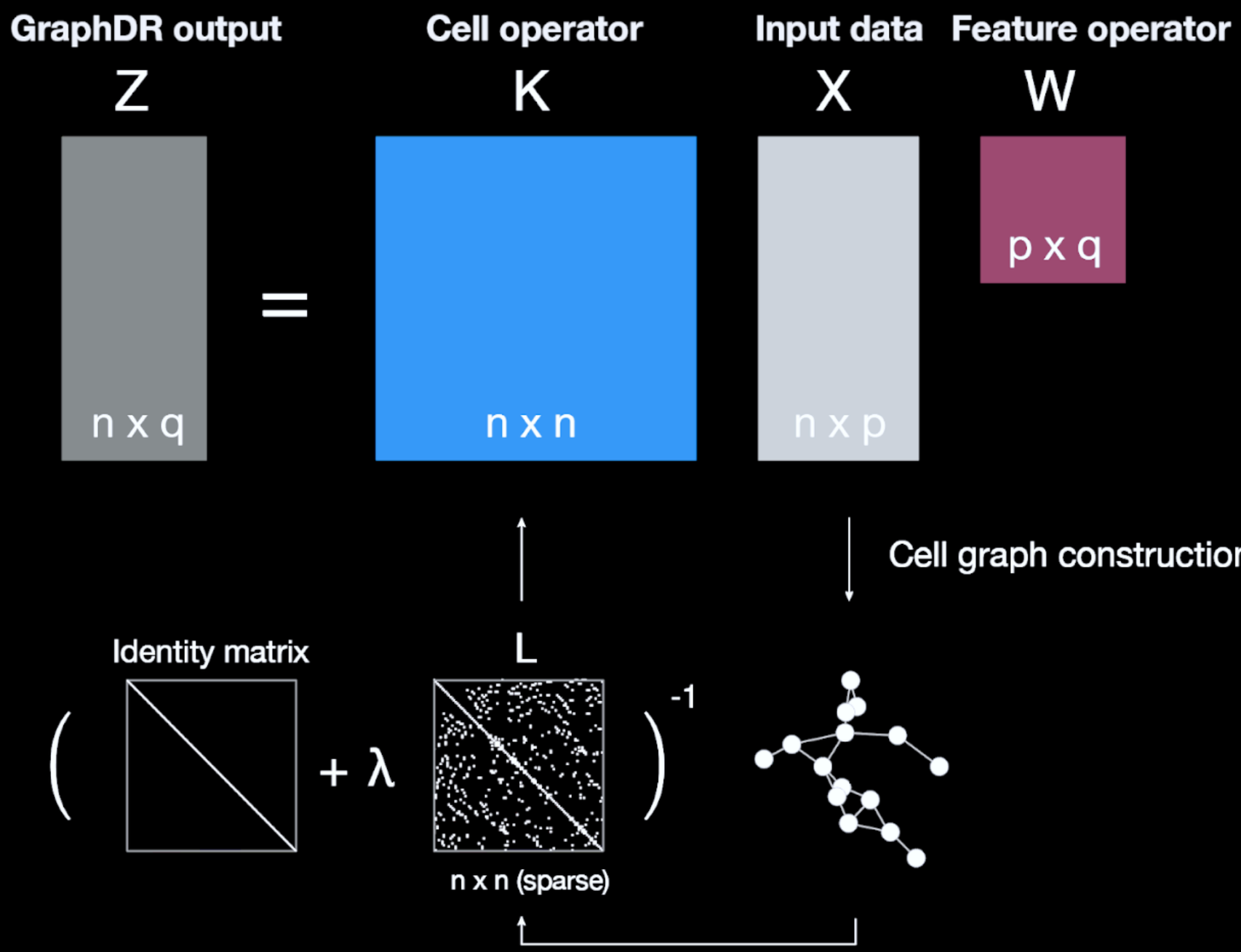
GraphDR - quasilinear visualization and general representation



Data from Hochgerner et al. 2018, Zeisel et al,. 2018

“Quasilinear” approaches:

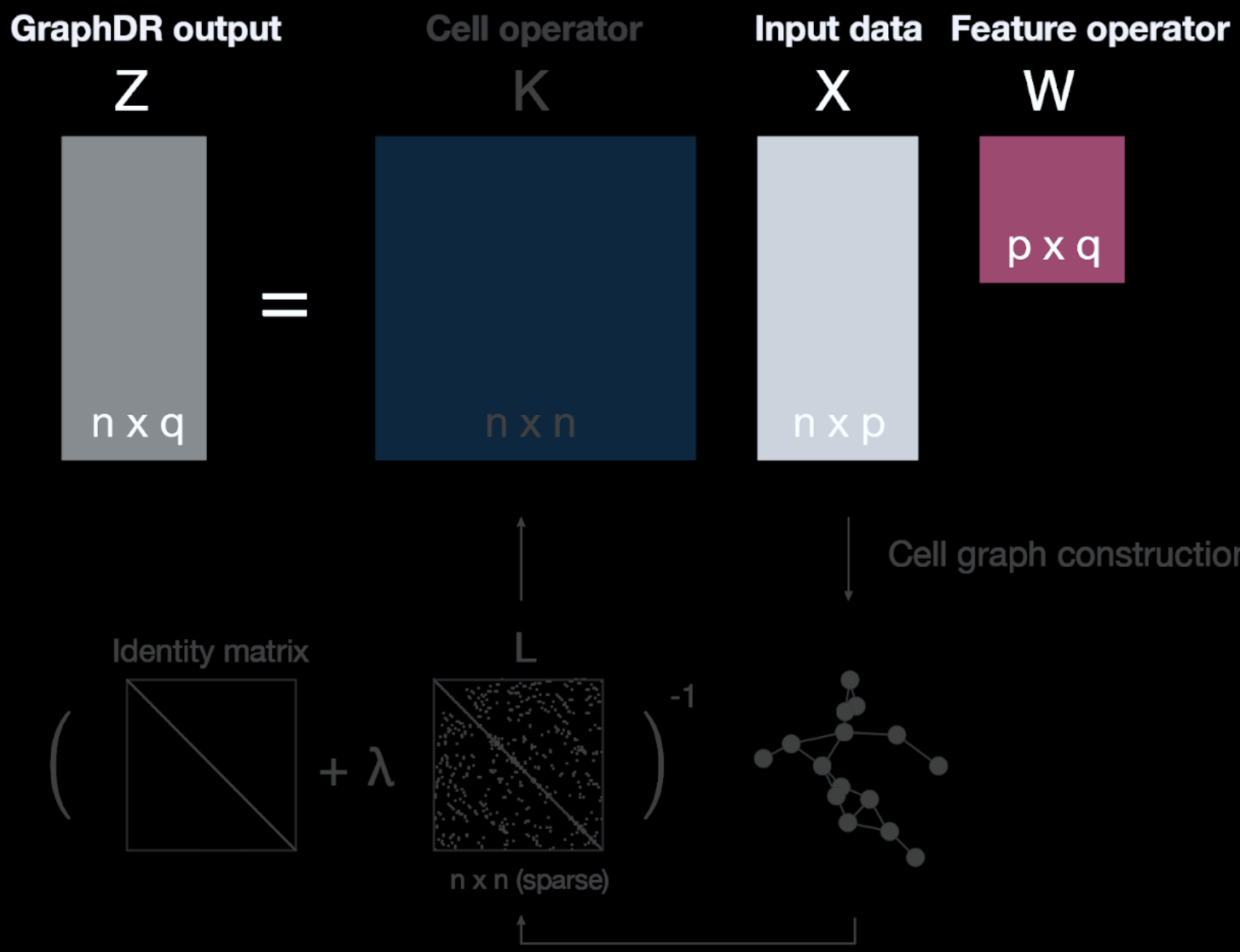
GraphDR - quasilinear visualization and general representation



GraphDR objective: $\underset{W,Z}{minimize} \quad \|X - Z\mathrm{W}^{\mathrm{T}}\|_2^2 + \lambda \sum_{\{i,j\}\in G} G_{ij}\|Z_i - Z_j\|_2^2, \qquad s.t.\ W^TW = I$

“Quasilinear” approaches:

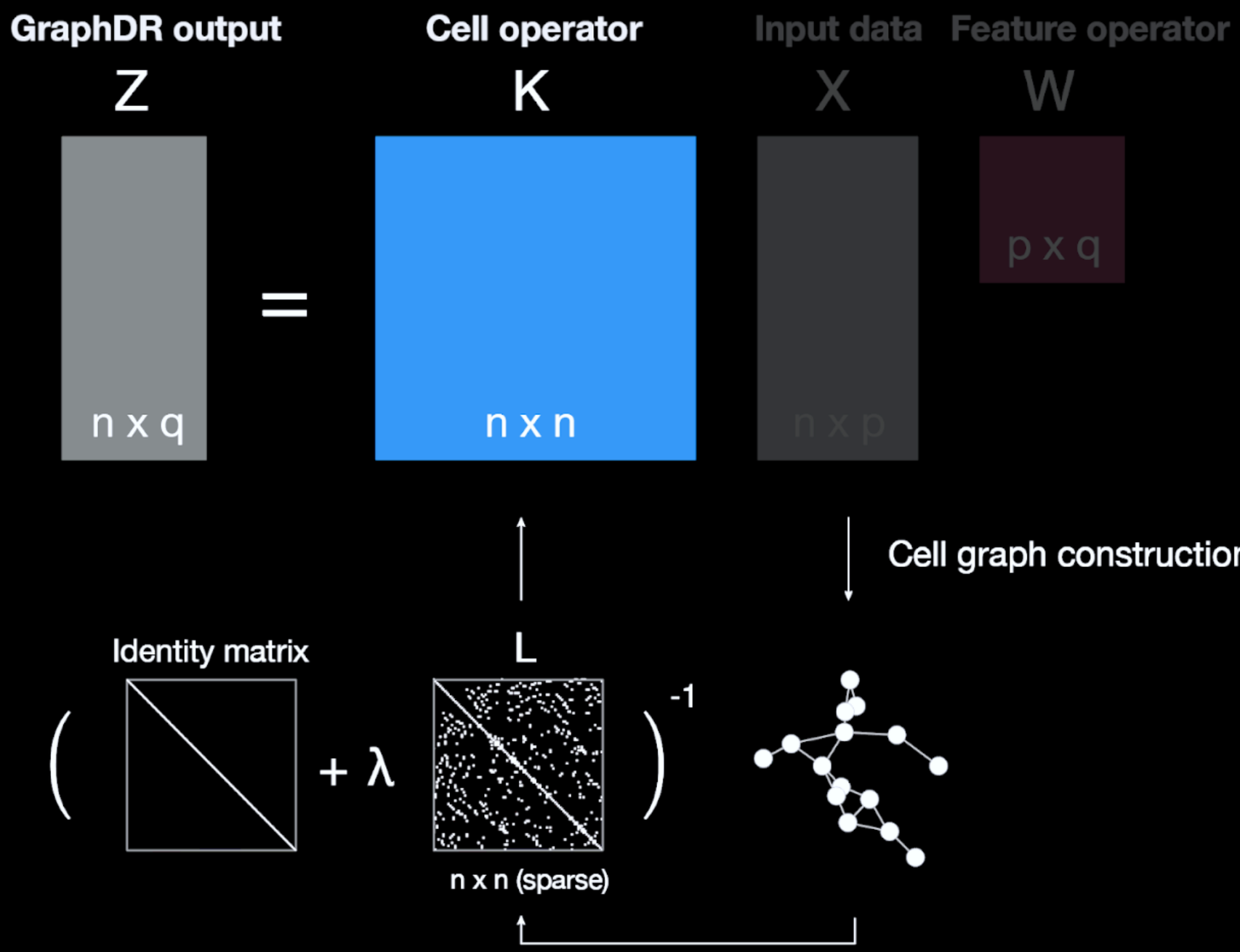
GraphDR - quasilinear visualization and general representation



GraphDR objective function: $\underset{W,Z}{minimize} \quad \|XW - Z\|_2^2 + \lambda \sum_{\{i,j\} \in G} G_{ij} \|Z_i - Z_j\|_2^2, \qquad s.t.\ W^T W = I$

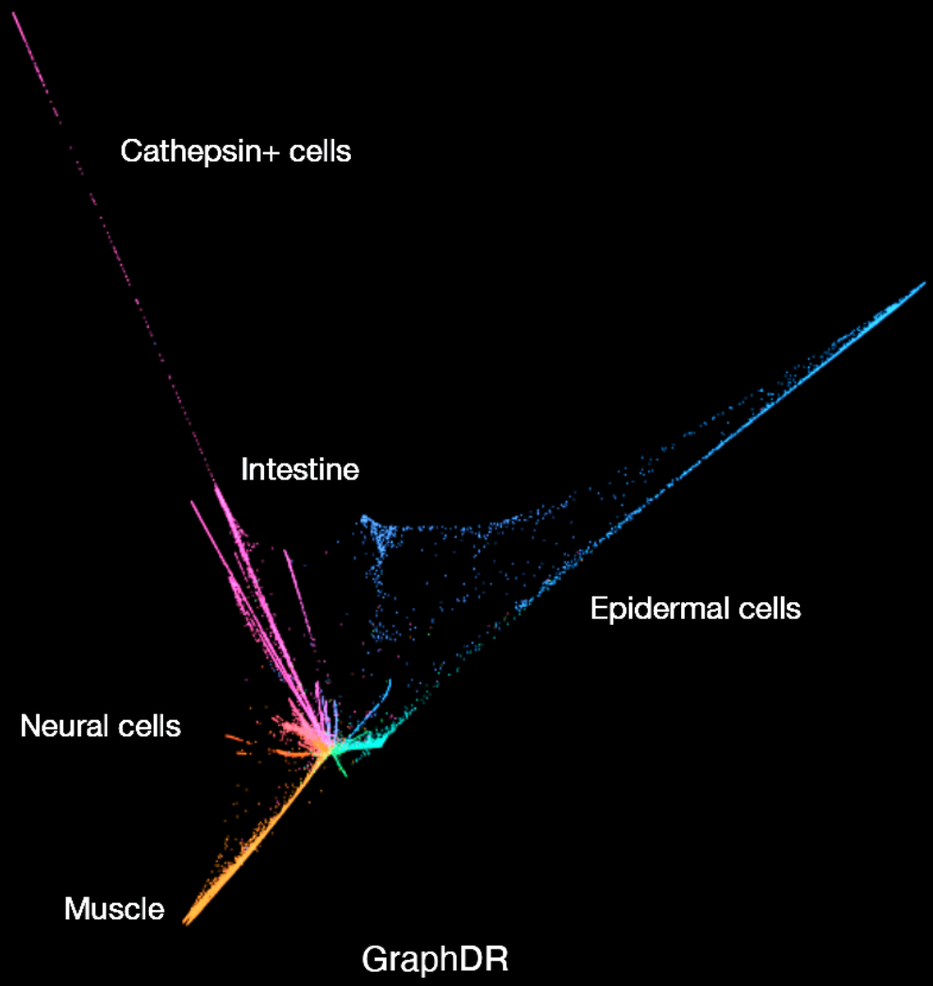
“Quasilinear” approaches:

GraphDR - quasilinear visualization and general representation

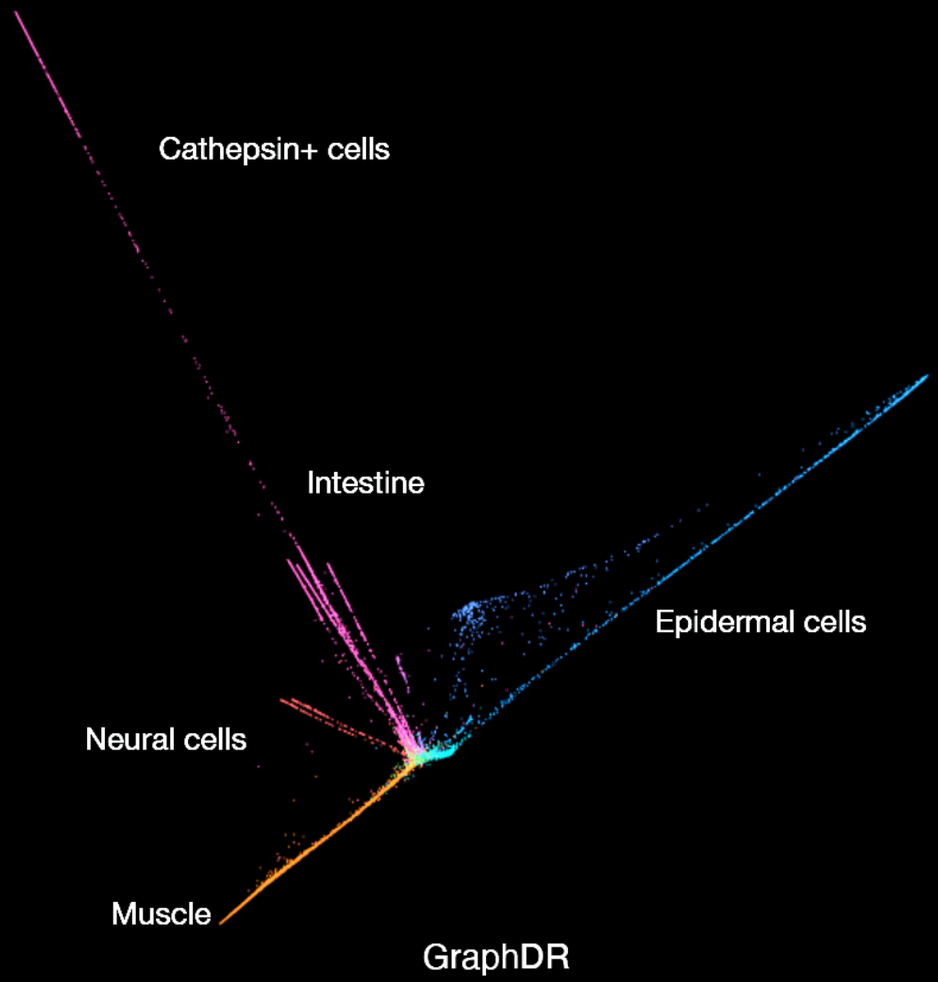


GraphDR objective: $\underset{W,Z}{minimize} \quad \|X - ZW^{\mathrm{T}}\|_2^2 + \lambda \sum_{\{i,j\}\in G} G_{ij}\|Z_i - Z_j\|_2^2, \qquad s.t.\ W^TW = I$

GraphDR representations allow direct comparison across datasets

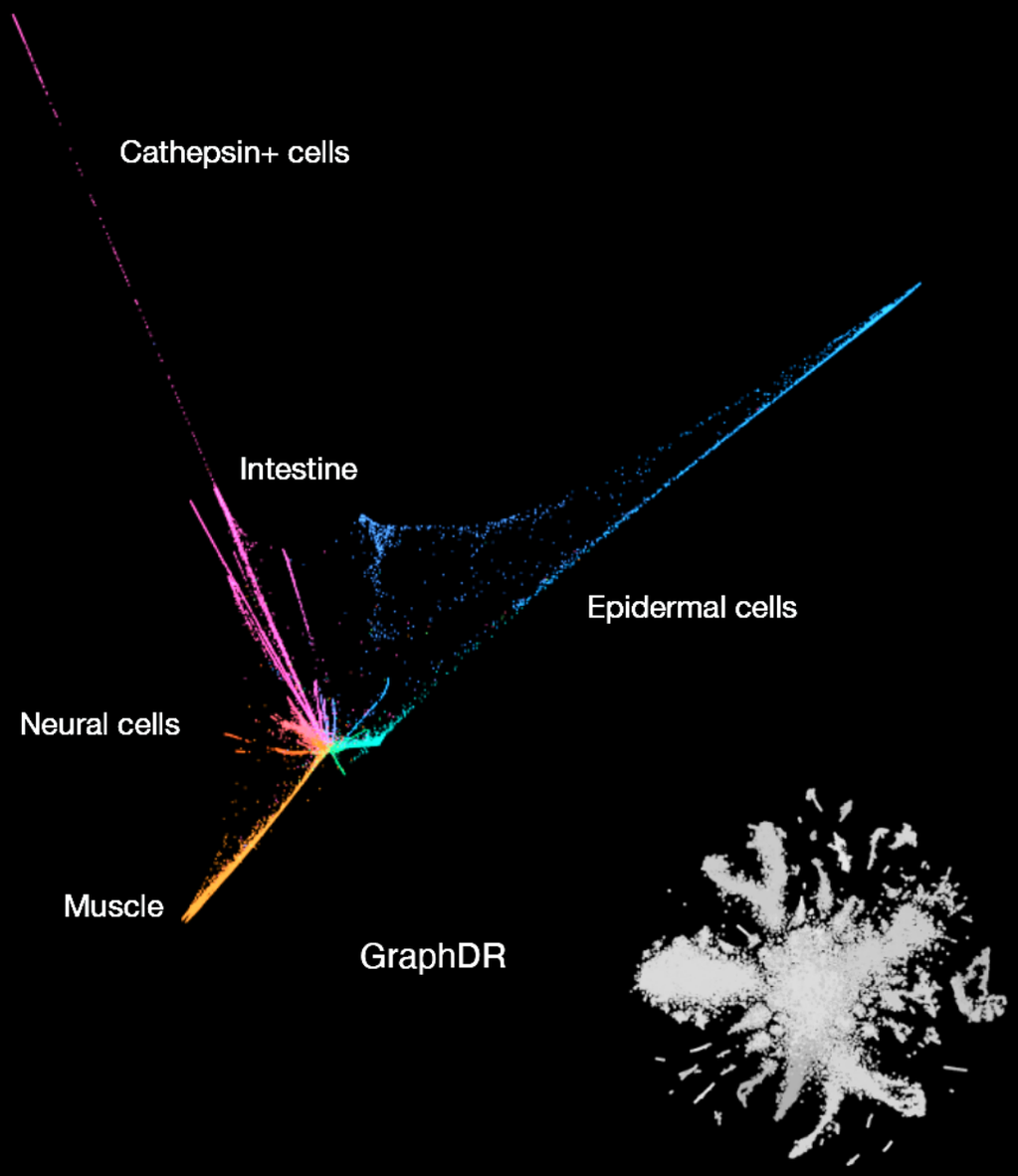


Planarian whole animal (Fincher et. al.)

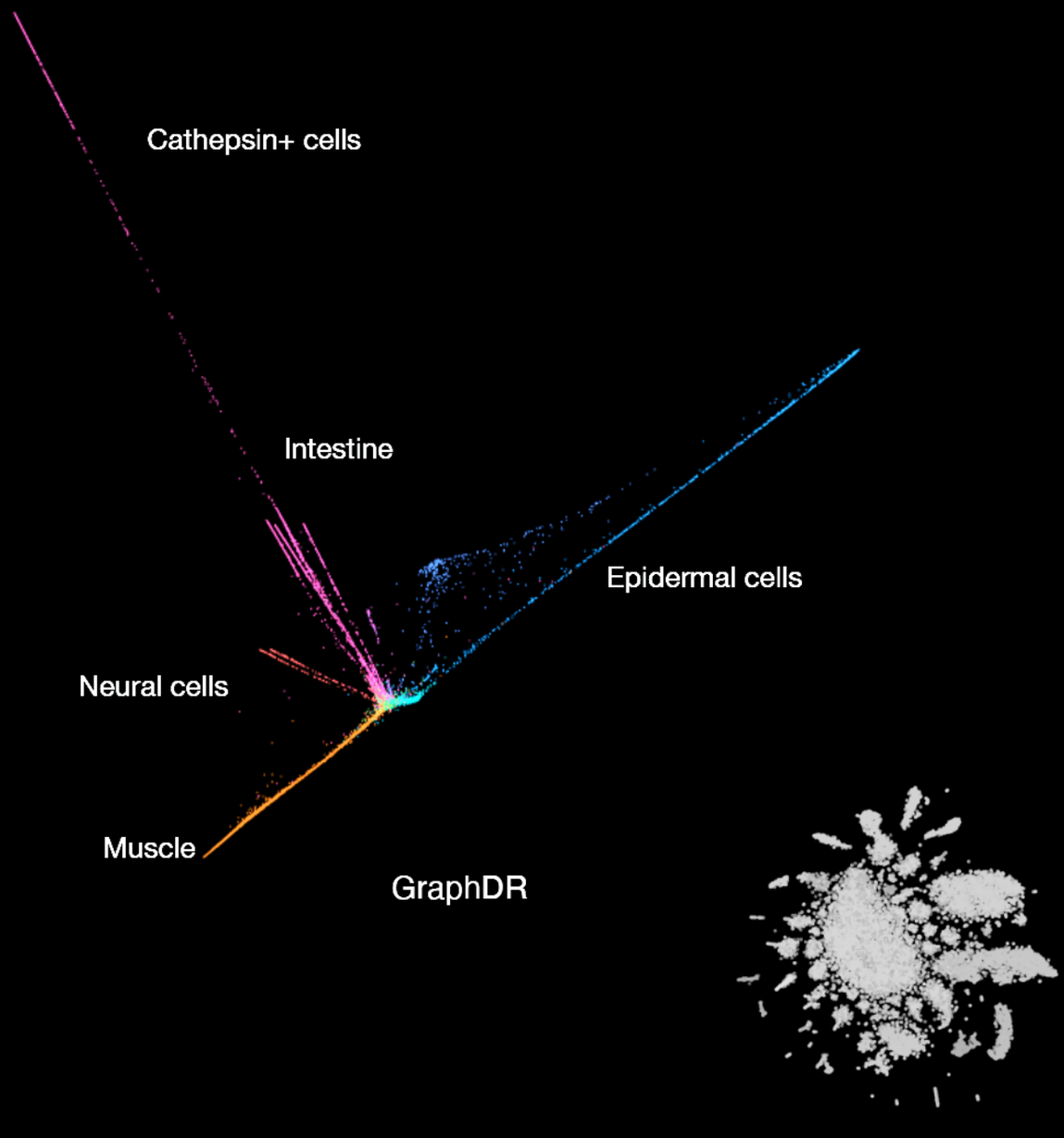


Planarian whole animal (Plass et. al.)

GraphDR representations allow direct comparison across datasets

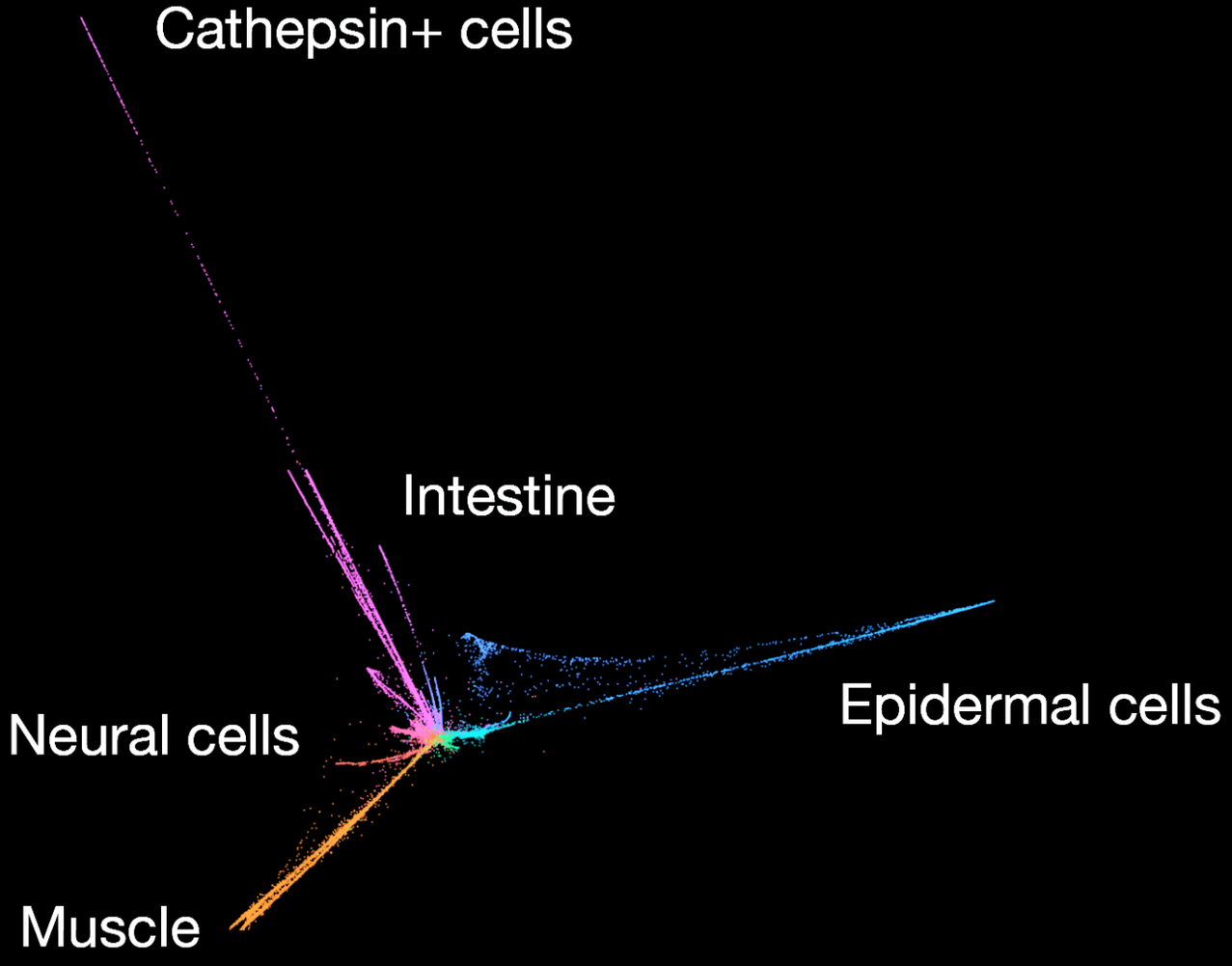


Planarian whole animal (Fincher et. al.)

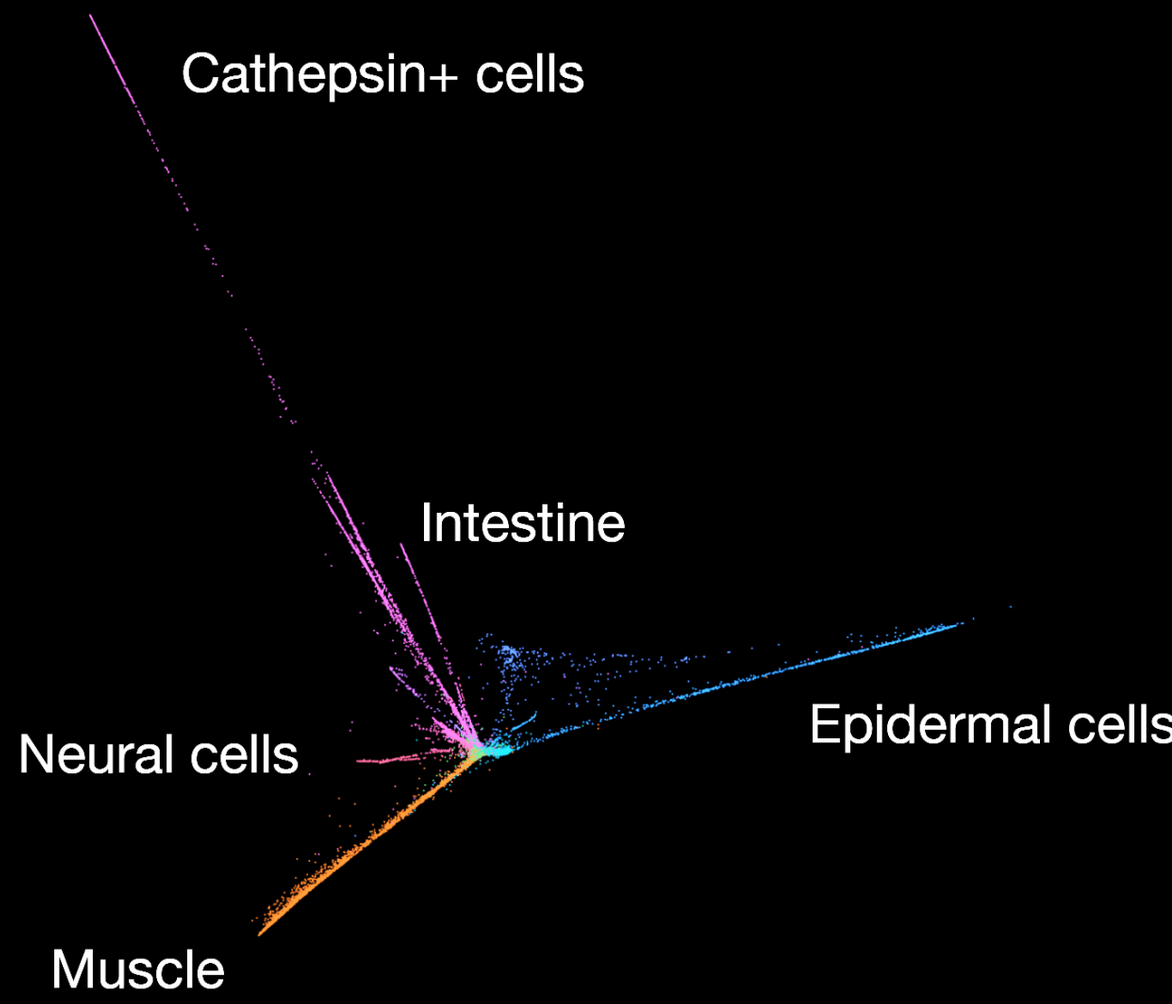


Planarian whole animal (Plass et. al.)

Improved comparison cross datasets with graph-based alignment

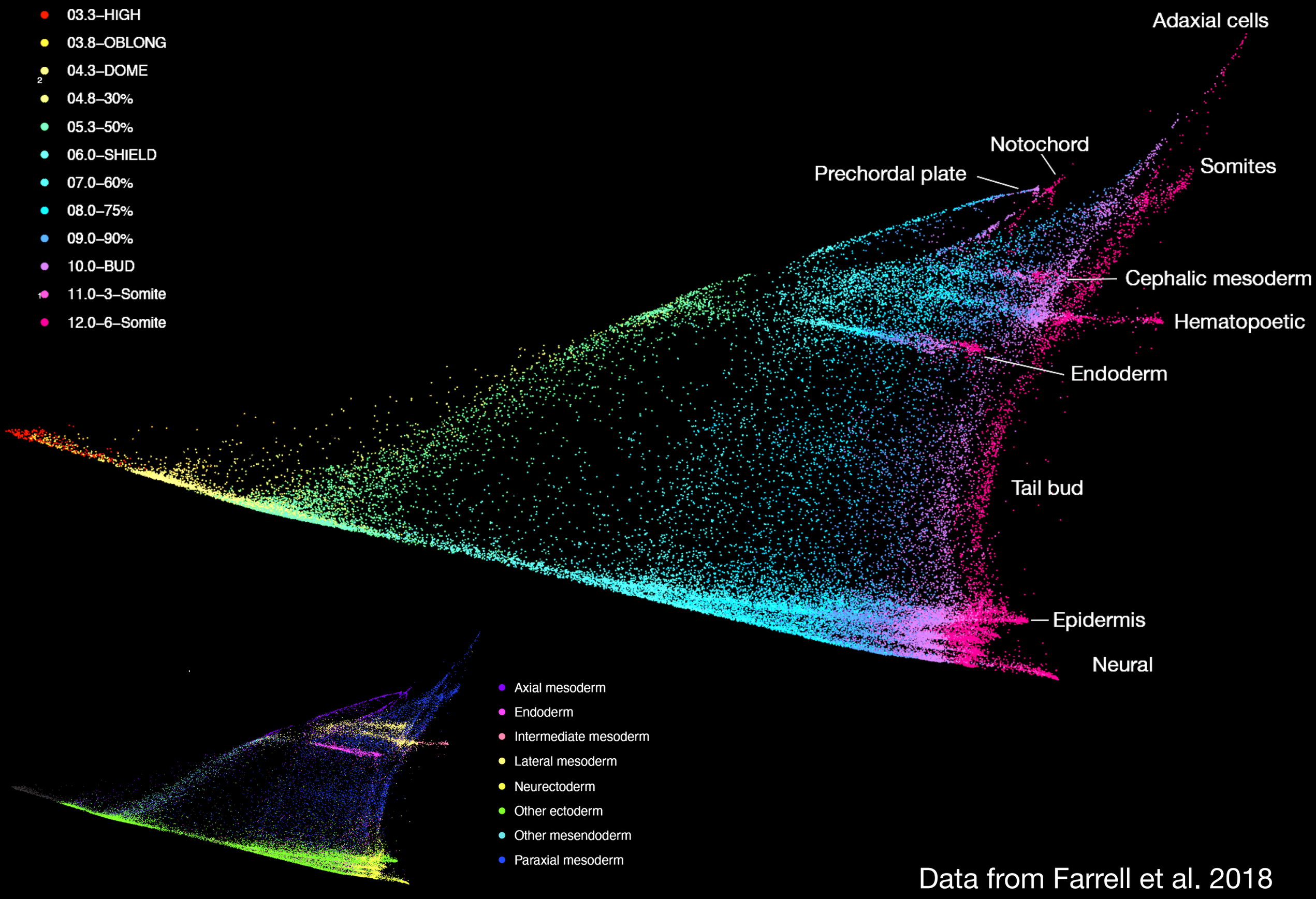


Fincher et. al.



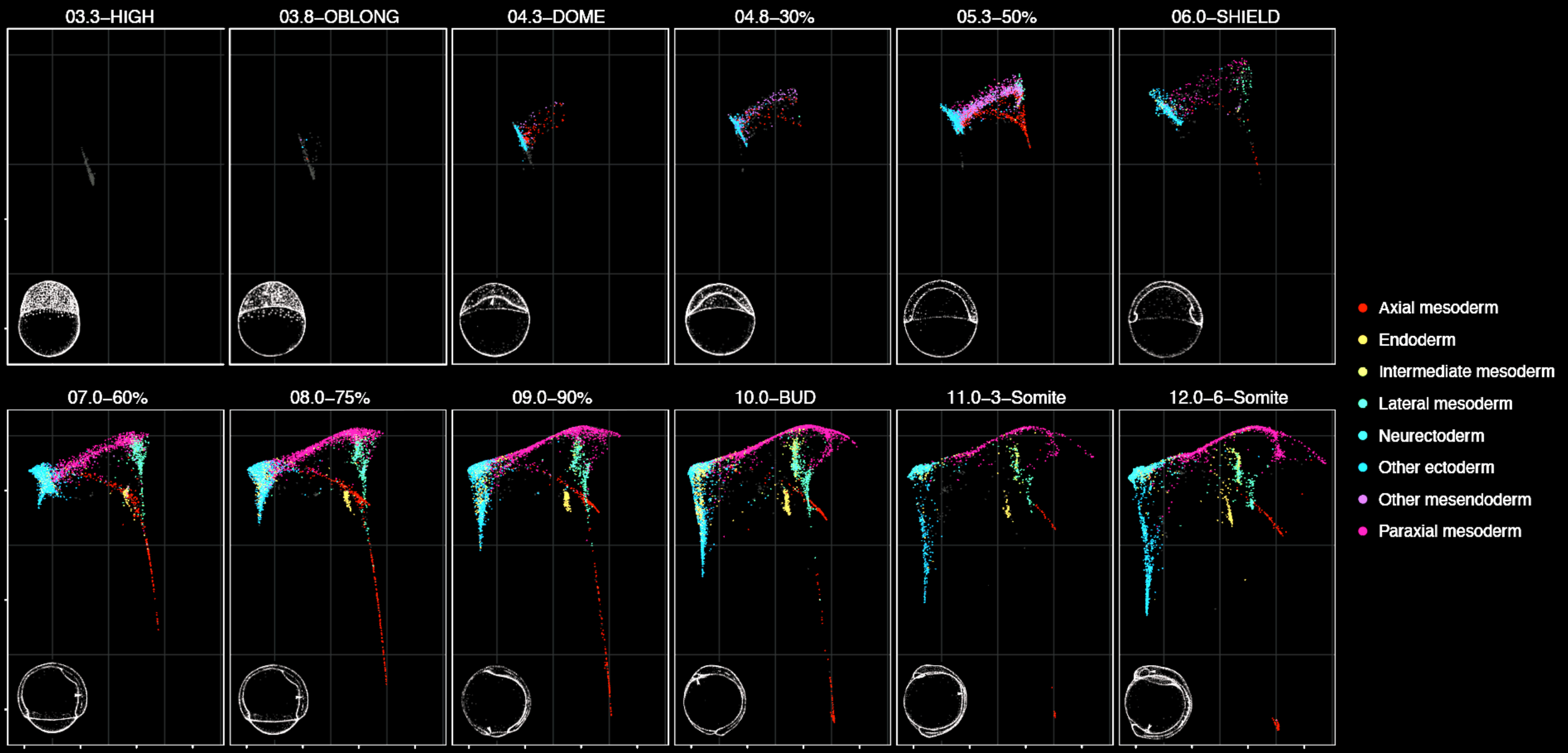
Plass et. al.

GraphDR visualization of zebrafish embryonic development

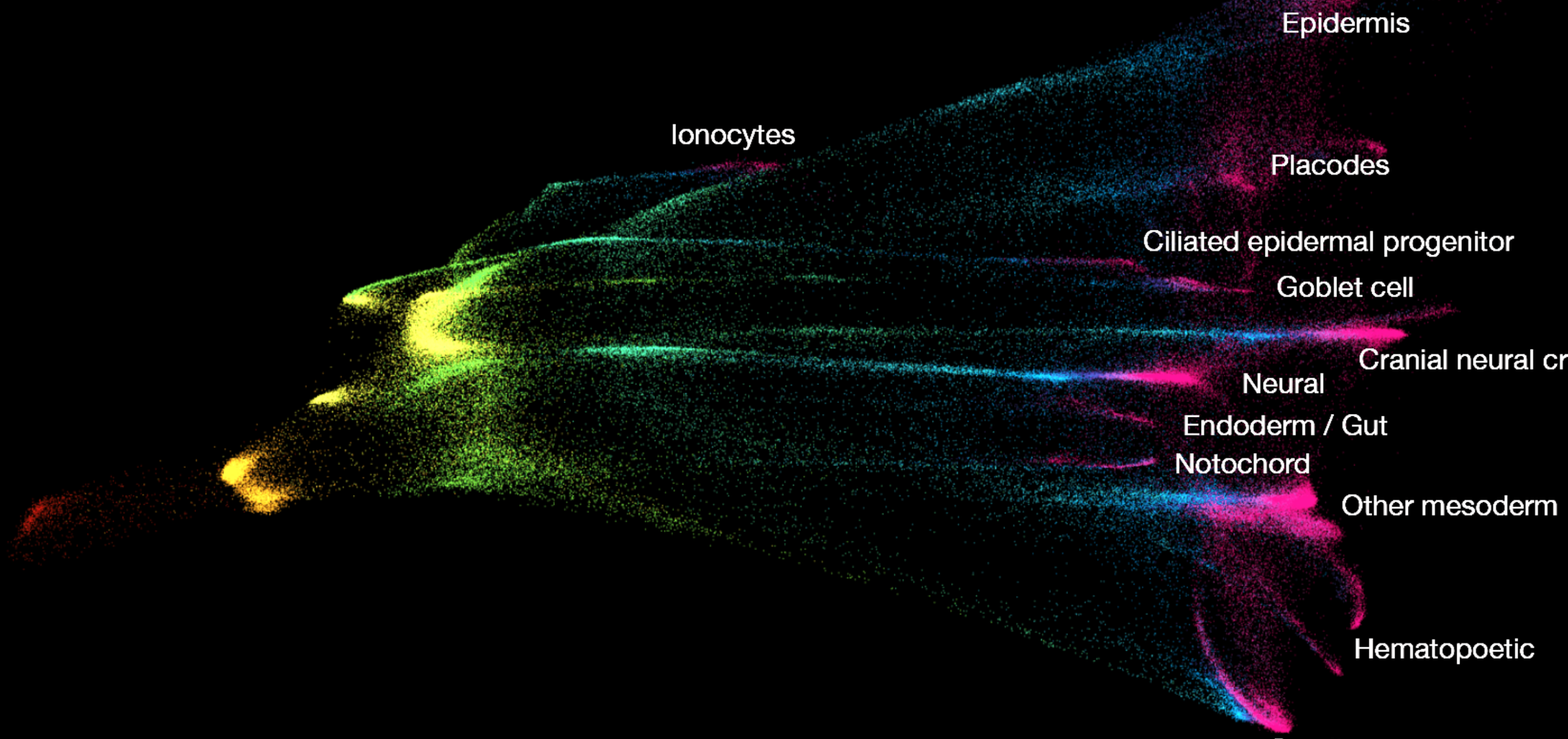


Data from Farrell et al. 2018

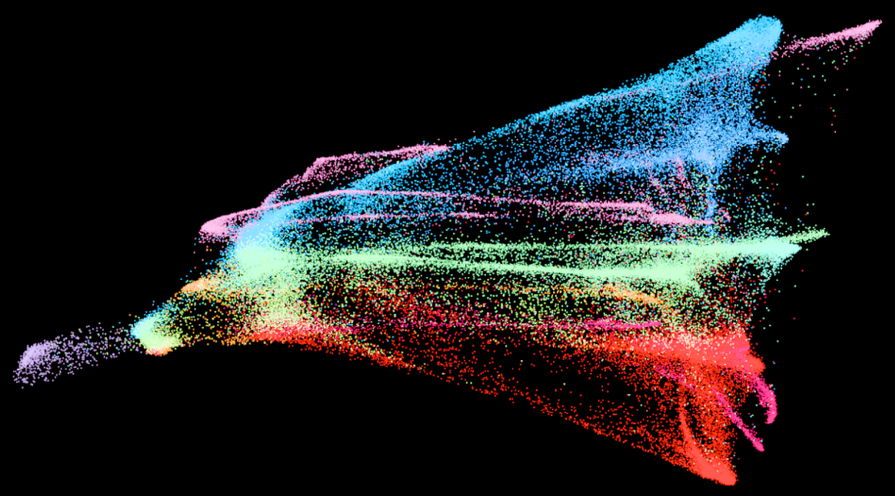
Uniform interpretability allows direct comparison across temporal “slices”



GraphDR visualization of Xenopus embryonic development

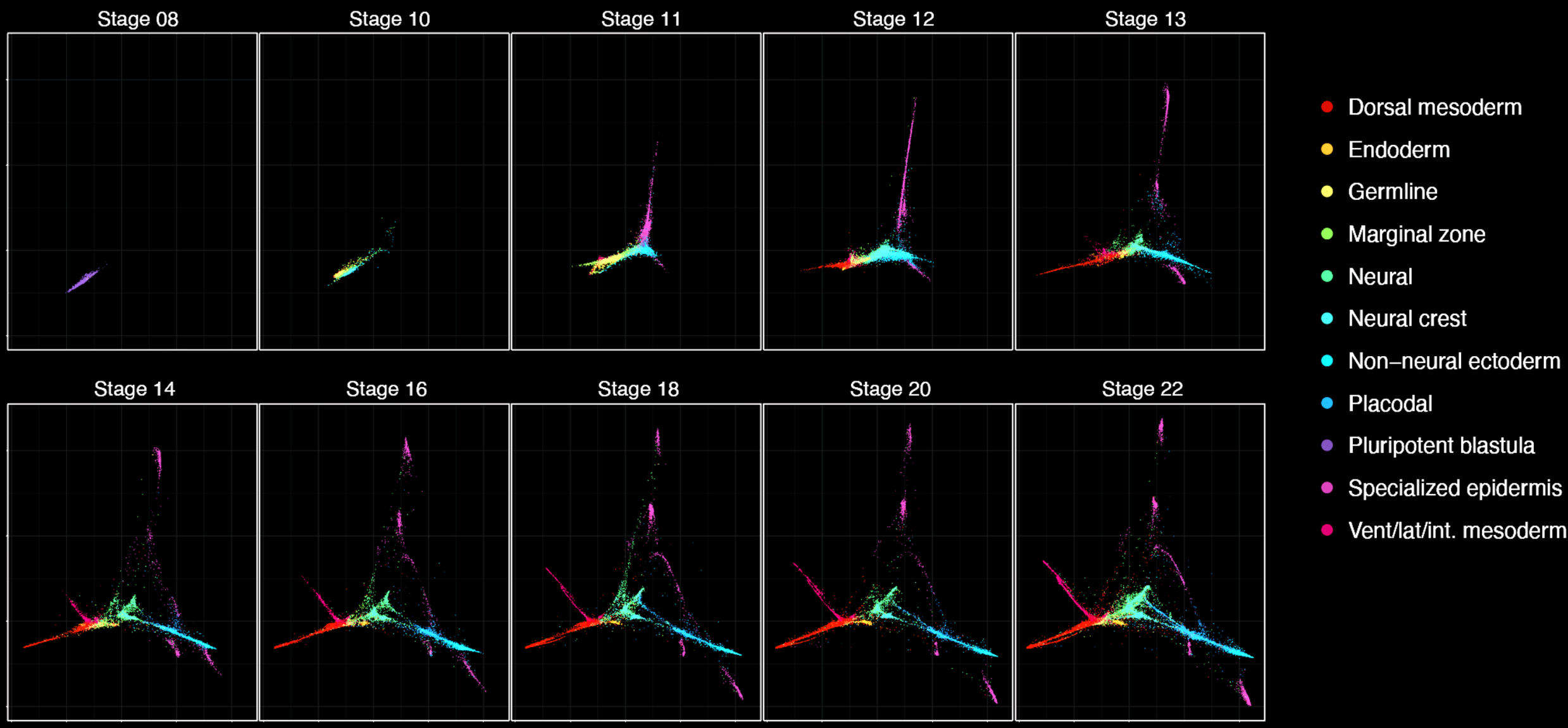


- Stage 08
- Stage 10
- Stage 11
- Stage 12
- Stage 13
- Stage 14
- Stage 16
- Stage 18
- Stage 20
- Stage 22



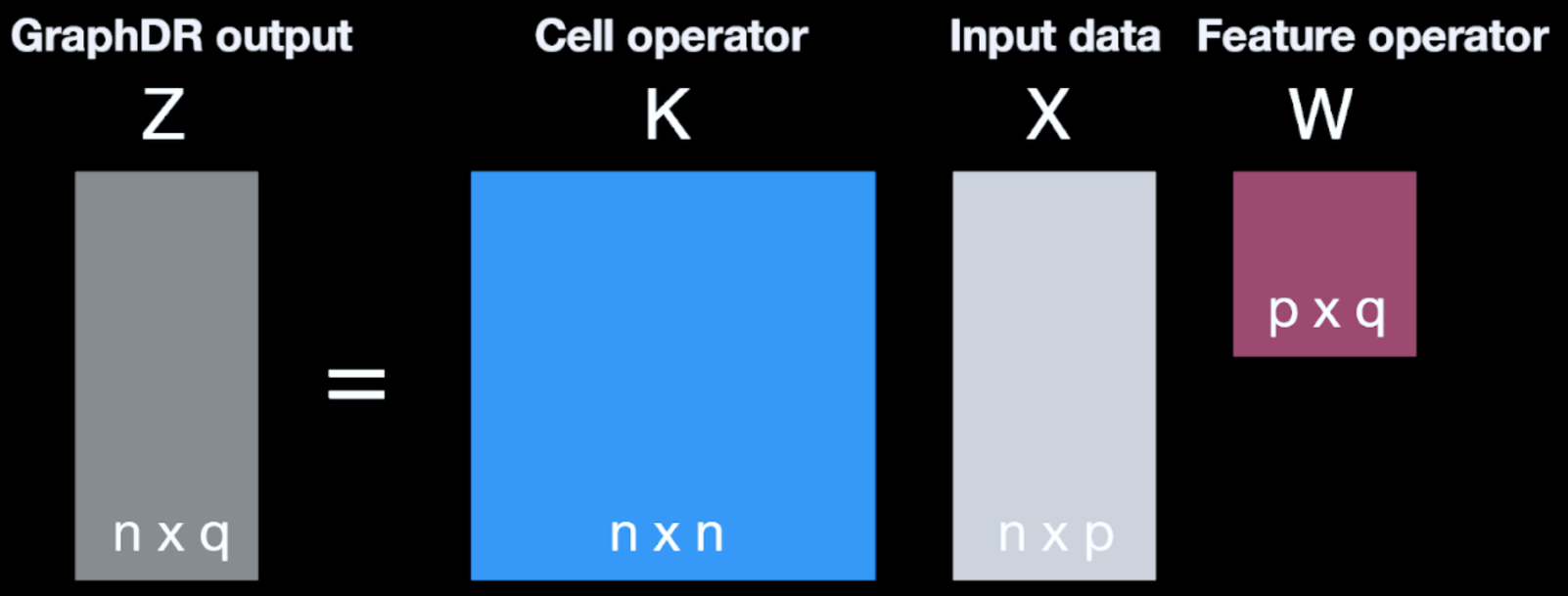
- Dorsal mesoderm
- Endoderm
- Germline
- Marginal zone
- Neural
- Neural crest
- Non–neural ectoderm
- Placodal
- Pluripotent blastula
- Specialized epidermis
- Vent/lat/int. mesoderm

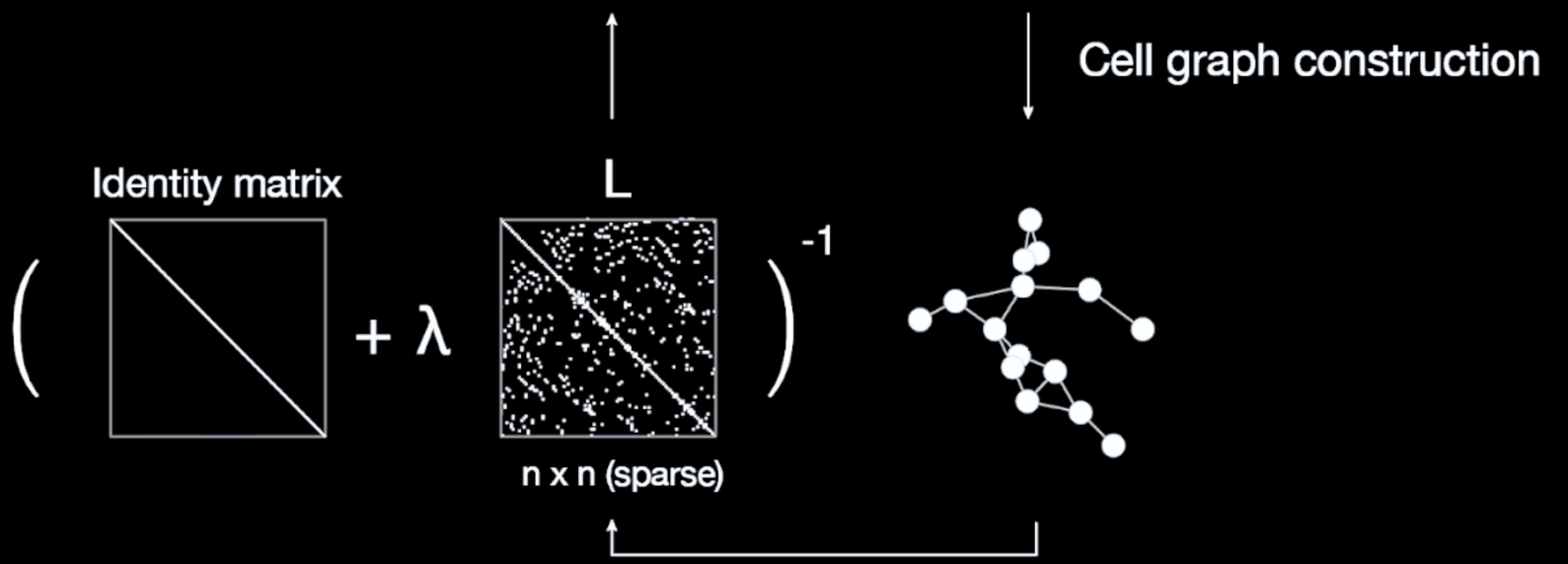
Data from Briggs et al. 2018


Stage 08
Stage 10
Stage 11
Stage 12
Stage 13
Stage 14
Stage 16
Stage 18
Stage 20
Stage 22
Dorsal mesoderm
Endoderm
Germline
Marginal zone
Neural
Neural crest
Non−neural ectoderm
Placodal
Pluripotent blastula
Specialized epidermis
Vent/lat/int. mesoderm

“Quasilinear” approaches:

GraphDR - visualization and general-purpose representation

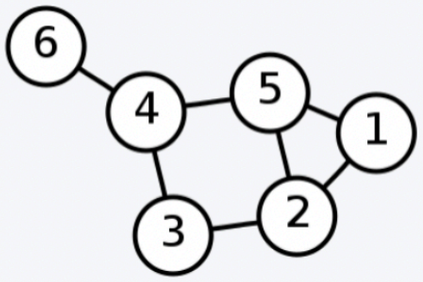




GraphDR objective: $\underset{W,Z}{minimize} \quad \|X - Z\mathrm{W}^{\mathrm{T}}\|_2^2 + \lambda \sum_{\{i,j\}\in G} G_{ij}\|Z_i - Z_j\|_2^2, \qquad s.t.\ W^TW = I$

New release! github.com/jzthree/quasildr

Graph Laplacian

Labelled graph	Degree matrix	Adjacency matrix	Laplacian matrix
	$\begin{pmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 2 & -1 & 0 & 0 & -1 & 0 \\ -1 & 3 & -1 & 0 & -1 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ -1 & -1 & 0 & -1 & 3 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix}$

From https: // en. wikipedia. org/ wiki/ Laplacian_ matrix .

Why does it appear?

It automatically emerge when you compute sum of pairwise L2 distances

$$\text{trace}(X^T L X) = \sum_{(i,j)\in G} \|x_i - x_j\|^2$$

"Quasilinear" approaches:

Implement GraphDR (you can do it in ~10 lines of code!)

GraphDR objective: $\underset{W,Z}{minimize} \quad \|X - Z\mathrm{W}^{\mathrm{T}}\|_2^2 + \lambda \sum_{\{i,j\}\in G} G_{ij}\|Z_i - Z_j\|_2^2, \qquad s.t.\ W^TW = I$

Analytical solution: $Z = (I + \lambda L)^{-1}XW$

W represents top d eigenvectors of $X^T(I + \lambda L)^{-1}X$

1. You may use: sklearn.neighbors.kneighbors_graph, scipy.sparse.csgraph.laplacian, scipy.sparse.eye, numpy.linalg.inv

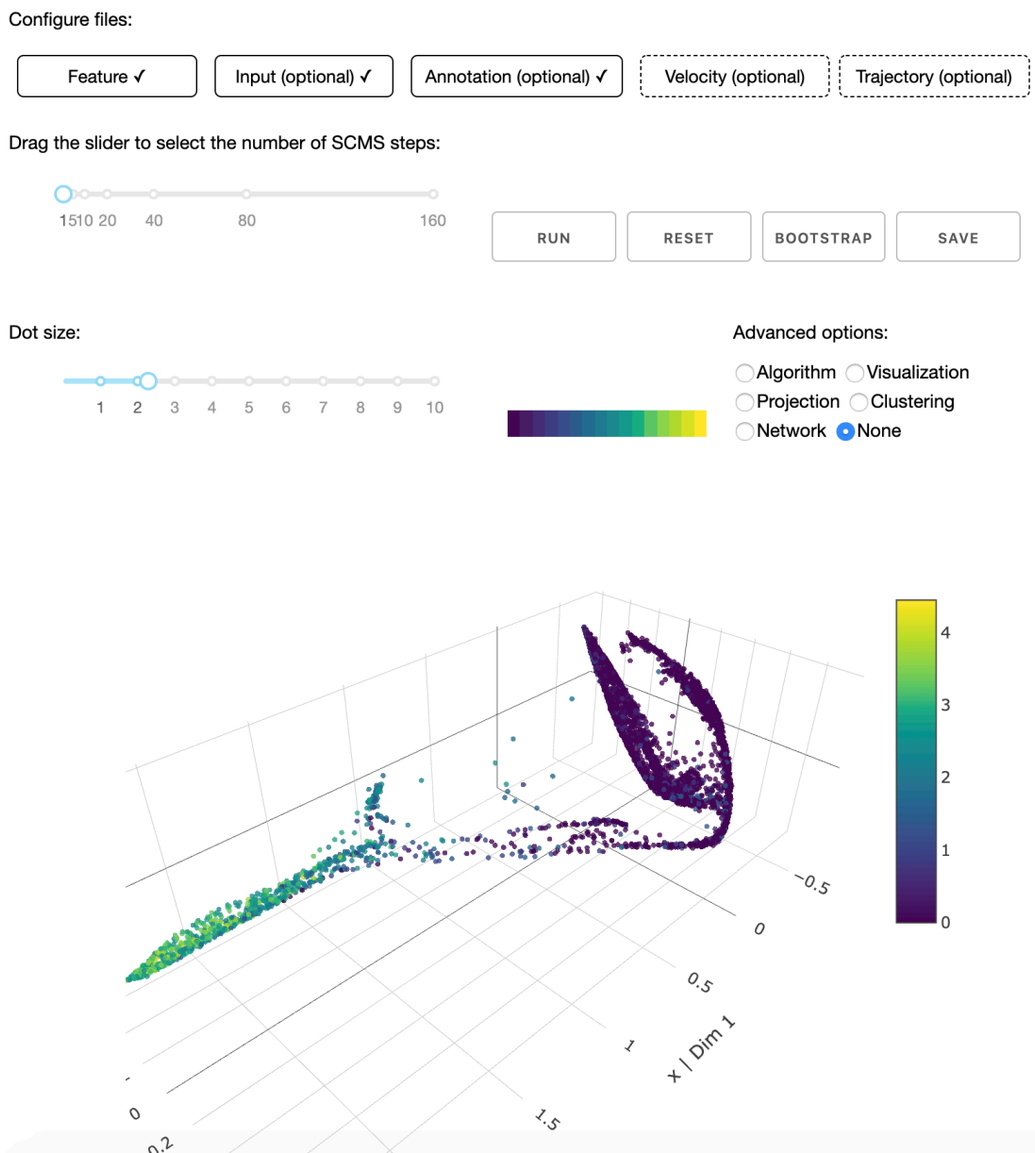
2. Make skipping computing W an option (`no_rotation=True`):
   $Z' = (I + \lambda L)^{-1}X$ is the solution to the objective when we add the constraint $W = I$

3. Z does not have to reduce the number of dimensions in GraphDR

4. (Optional) make it scalable to > 1 million cells

Task for Day 3:

1. **Code review for Day 2 (Docstring for t-SNE)**
2. **Implement GraphDR (Code + Docstring)**
3. **Optional: 3D visualization (with plot.ly) or / and interactive visualization interface that allows adjusting regularization parameter (e.g. with Dash)**



GraphDR: scaling to >10 million cells

$$Z = (I + \lambda L)^{-1} X W$$

Approximate nearest neighbor (ANN)

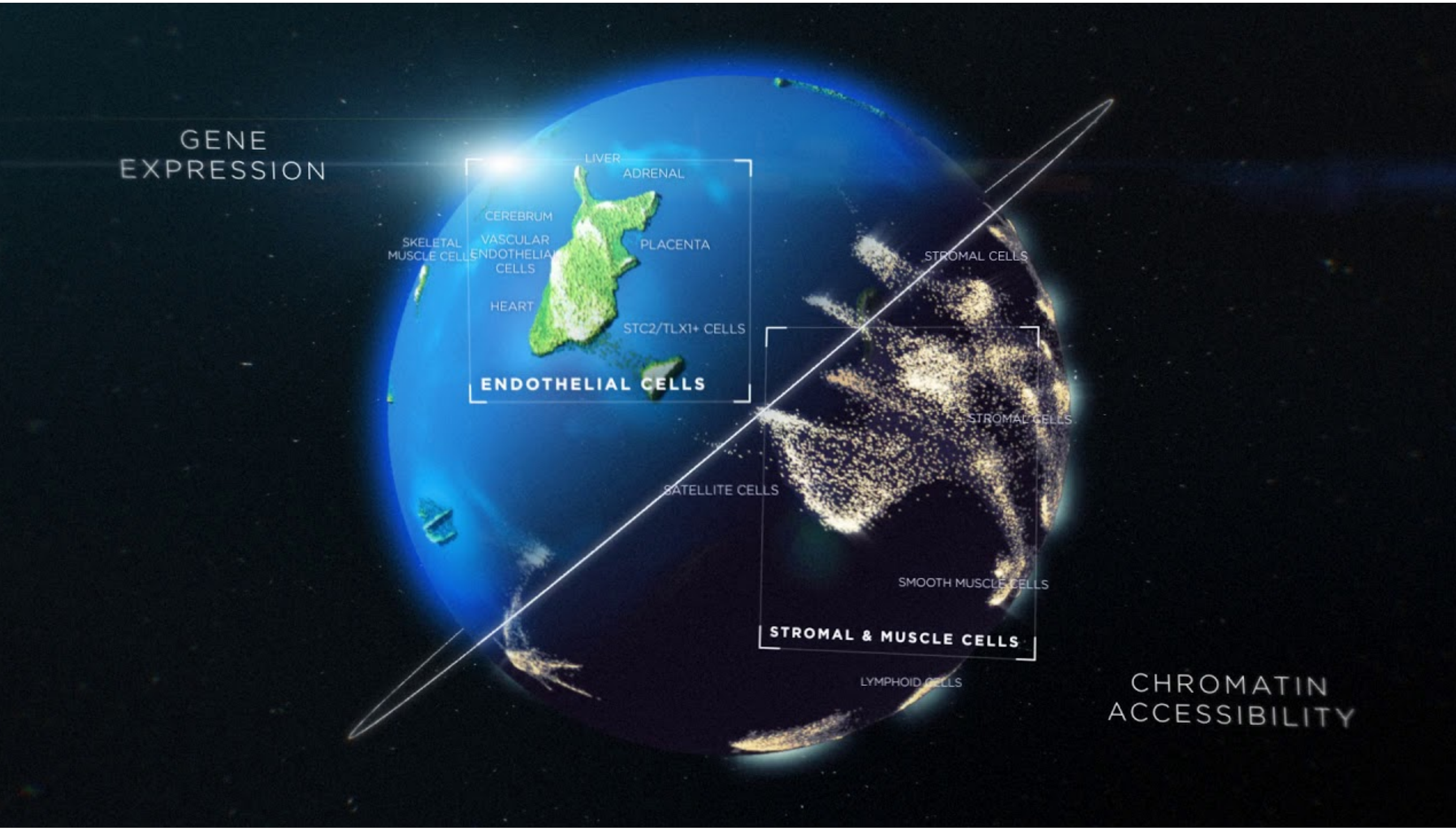
Brute force	n^2
KD-Tree / Ball-tree	n log n
ANN (**HNSW**, NN descent)	Close to O(n)

Don’t do matrix inversion

$I + \lambda L$ is sparse, the inverse of it is not sparse

Solve the linear equation $(I + \lambda L)Z = XW$ **with a sparse solver**

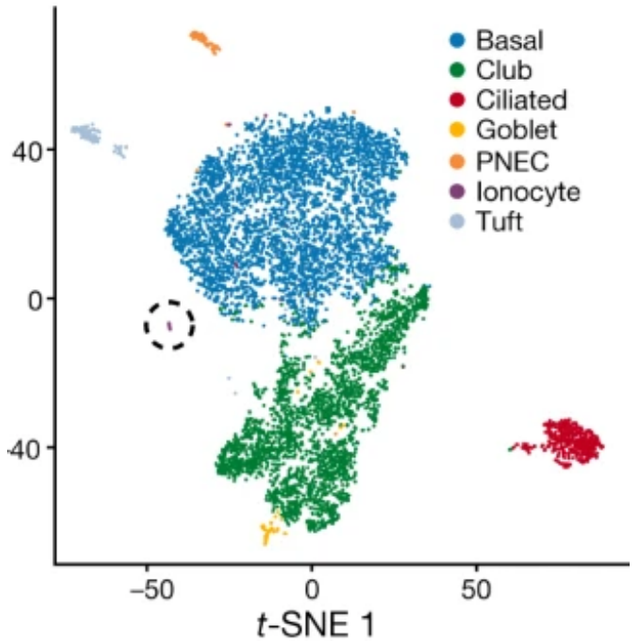
For further speed up, use GPU

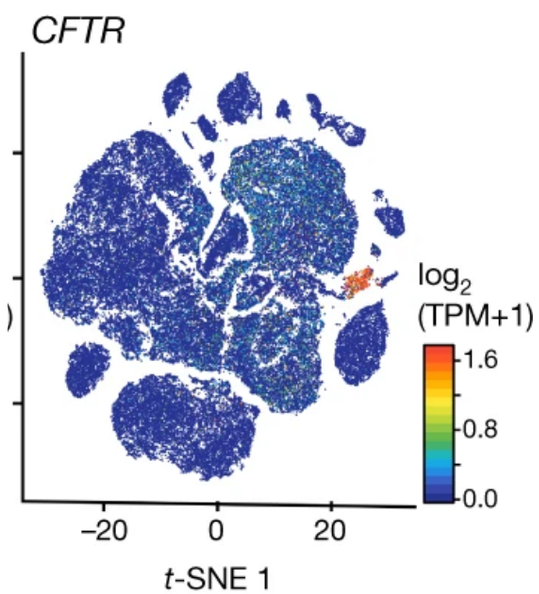

GENE EXPRESSION
LIVER
ADRENAL
CEREBRUM
SKELETAL MUSCLE CELLS
VASCULAR ENDOTHELIAL CELLS
PLACENTA
HEART
STC2/TLX1+ CELLS
ENDOTHELIAL CELLS
STROMAL CELLS
STROMAL CELLS
SATELLITE CELLS
SMOOTH MUSCLE CELLS
STROMAL & MUSCLE CELLS
LYMPHOID CELLS
CHROMATIN ACCESSIBILITY

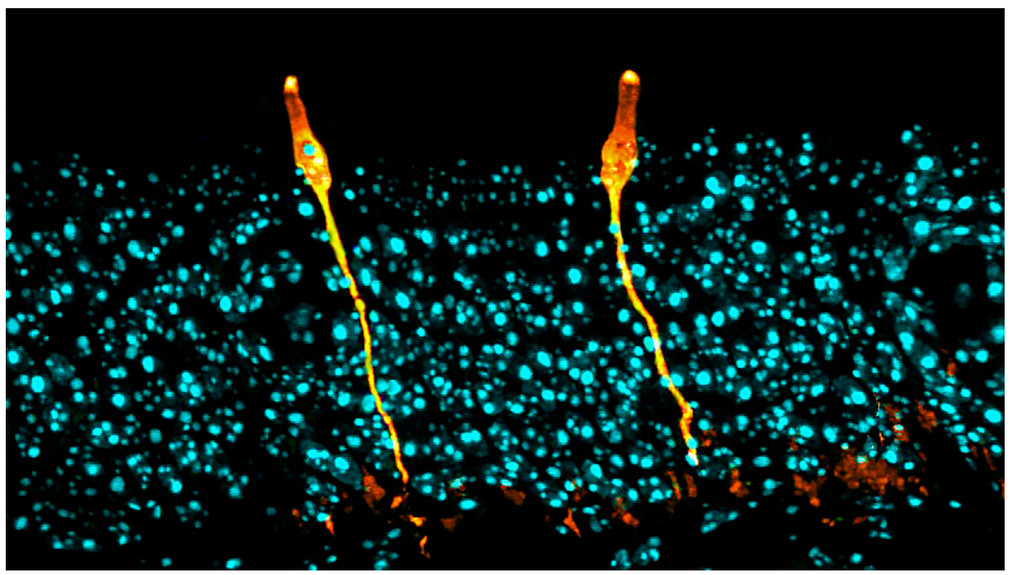
Single-cell RNA-seq identified rare CFTR-expression cell types that is key to cystic fibrosis

Cystic fibrosis is caused by mutations from the CFTR gene, discovered in the 1980s

What cell type does CFTR function in?

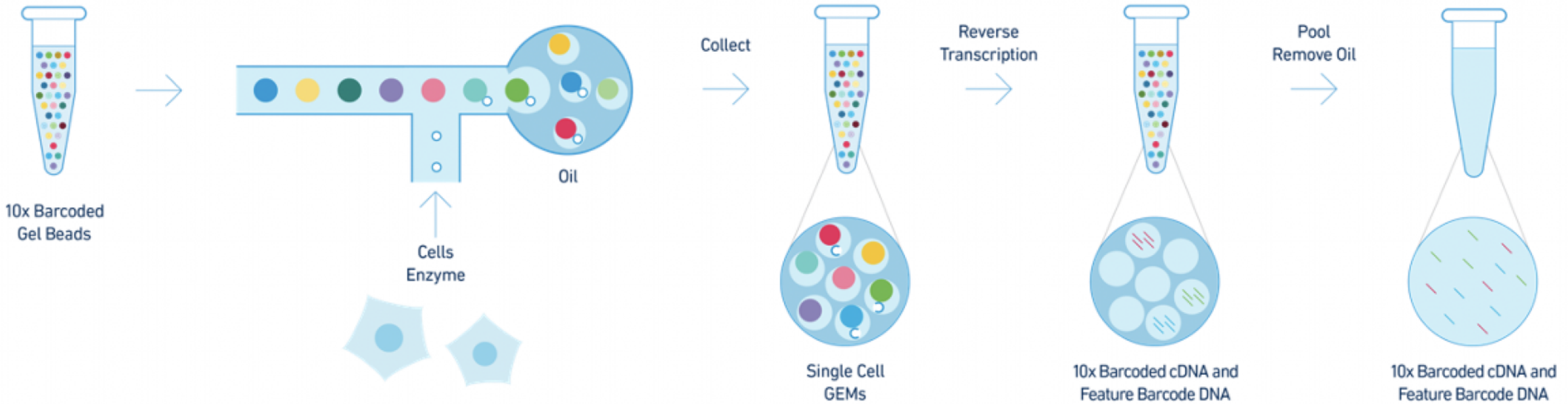




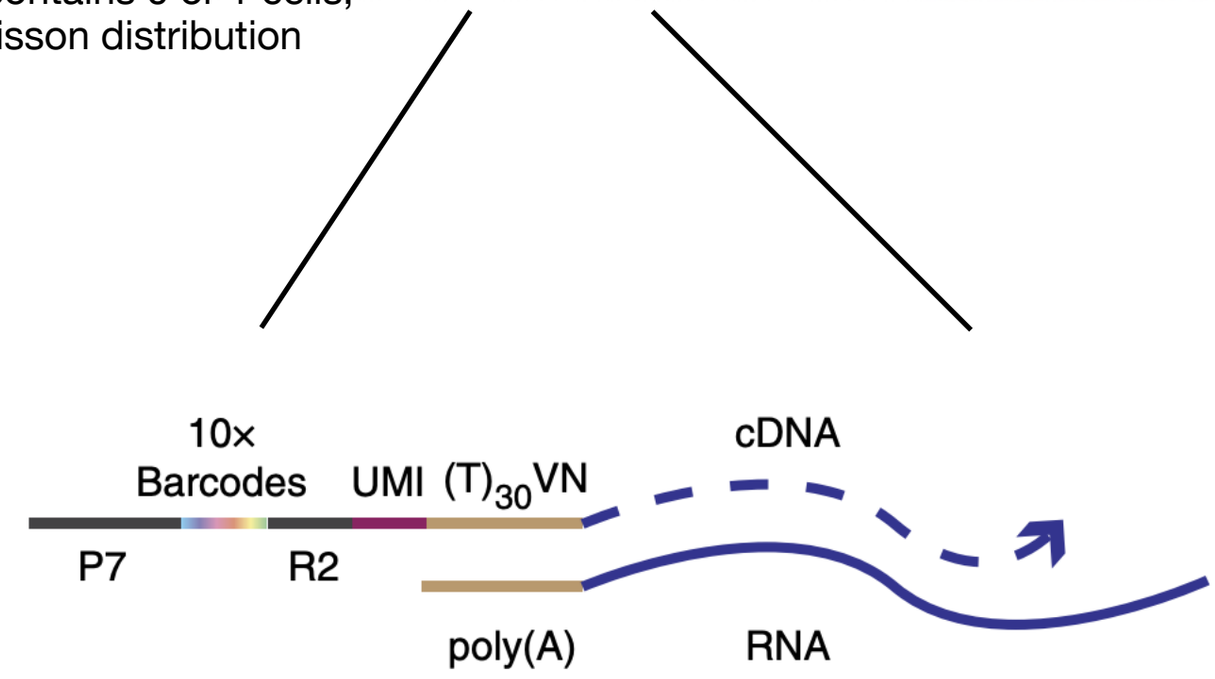


Plasschaert et al. 2018

Understand the data: single cell RNA-seq

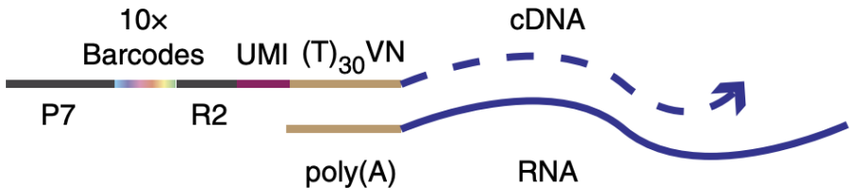


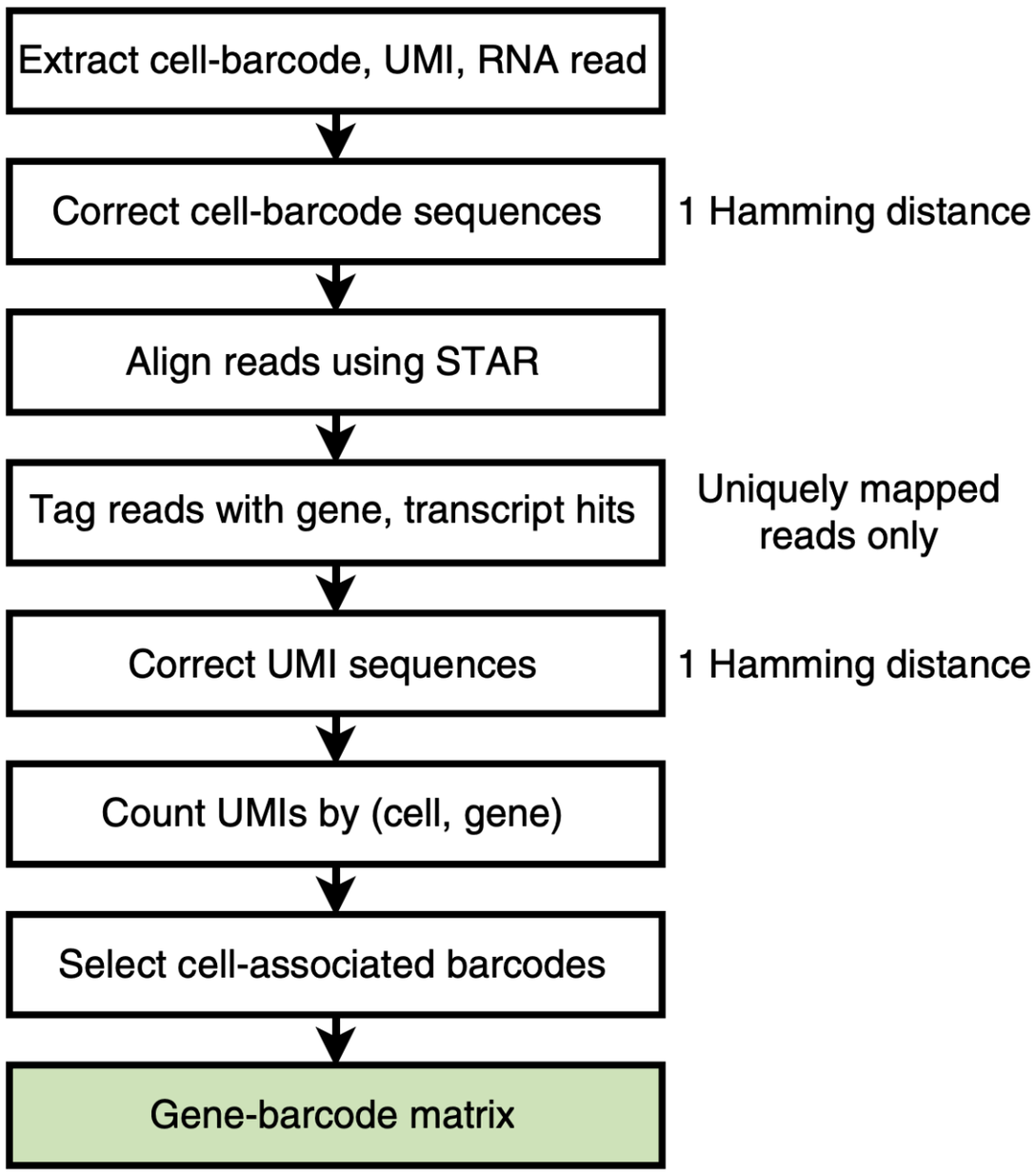
Most droplets contains 0 or 1 cells,
following Poisson distribution



Zheng et al, 2016, Massively parallel digital transcriptional profiling of single cells

Understand the data: single cell RNA-seq





Standard processing pipeline

Filter out genes and cells with mostly zero counts

Filter out cells with high mitochondria gene content

Normalize by total UMI count per cell (library size)

Log-transformation with pseudo count log(x+1)

Optional: Standard scaling (zero mean and unit variance for each gene)

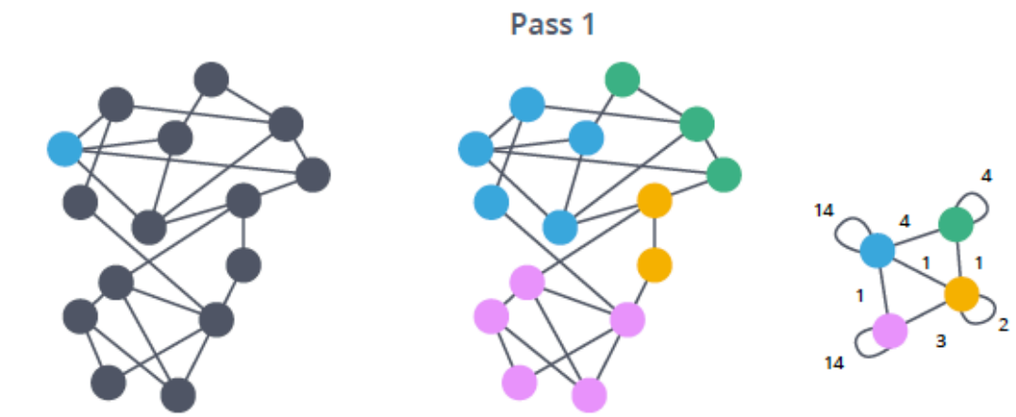
Zheng et al, 2016, Massively parallel digital transcriptional profiling of single cells

Identify cell types: Graph-based clustering

Use the topological structure of the data:
Nearest neighbor graph / shared nearest neighbor graph is often good representation for identify clusters

https://github.com/vtraag/**louvain-igraph**

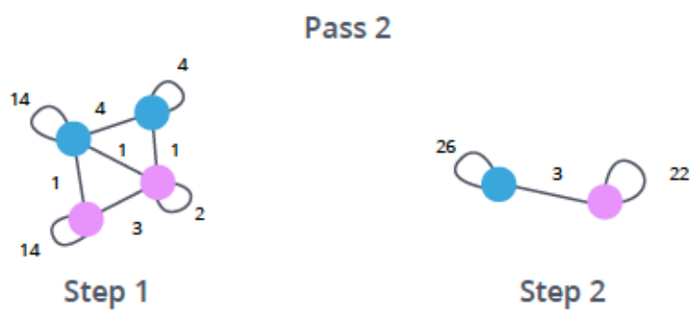
Modularity:
# of within-cluster edges -
expectation (based on product of node degrees)



Choose a start node and calculate the change in modularity that would occur if that node joins and forms a community with each of its immediate neighbors.

The start node joins the node with the highest modularity change. The process is repeated for each node with the above communities formed.

Communities are aggregated to create super communities and the relationships between these super nodes are weighted as a sum of previous links. (Self-loops represent the previous relationships now hidden in the super node.)

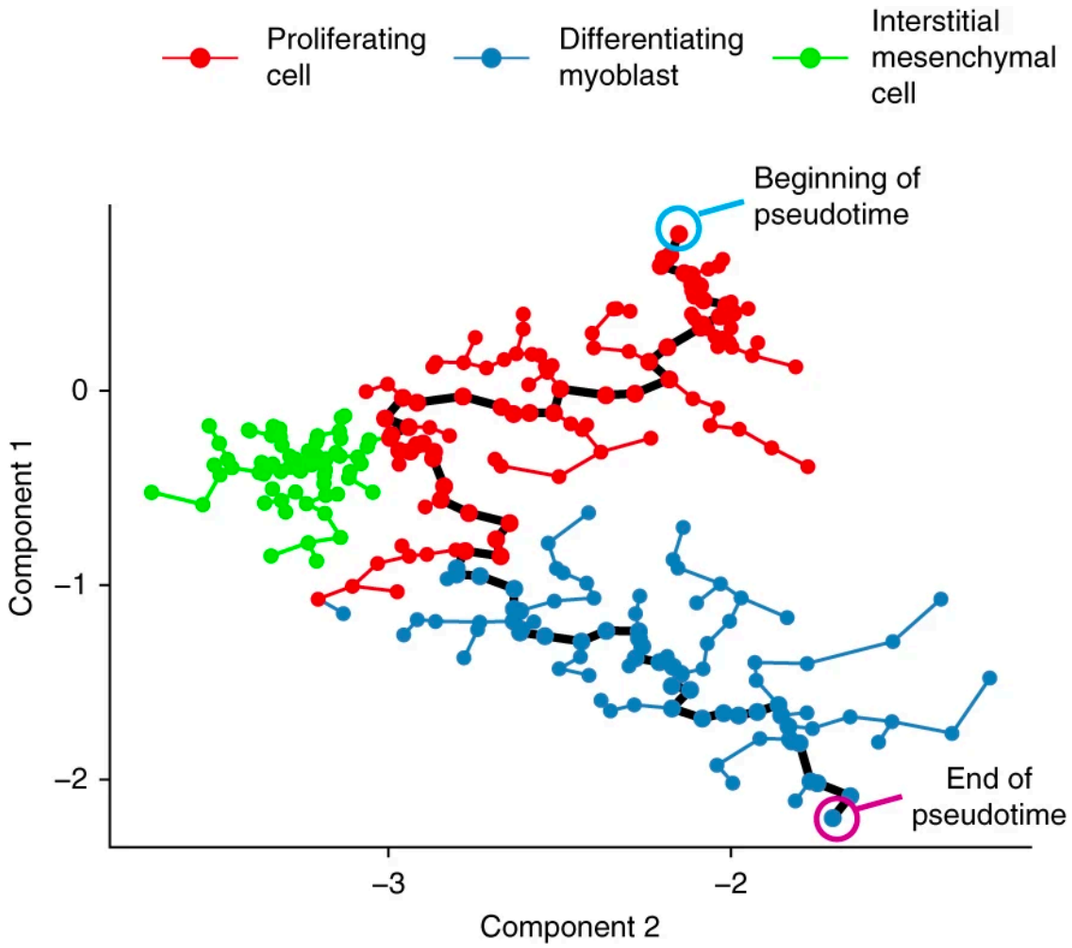


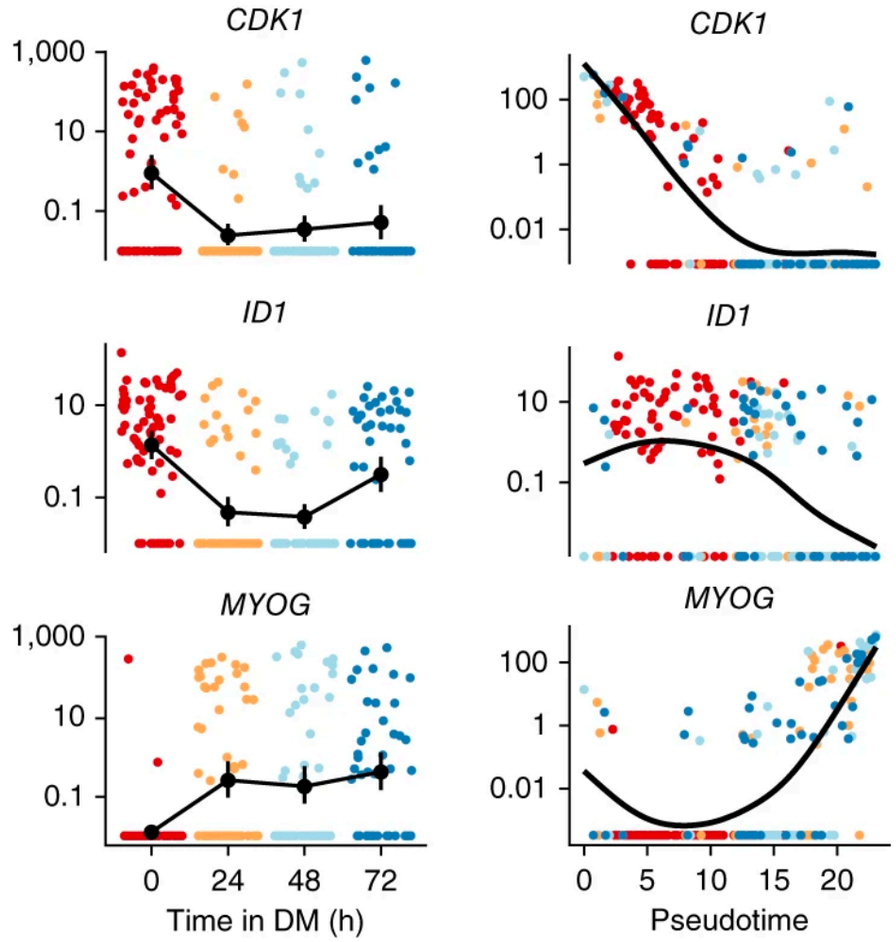
Steps 1 and 2 repeat in passes until there is no further increase in modularity or a set number of iterations have occurred.

Louvain Modularity Algorithm

https://neo4j.com/blog/graph-algorithms-neo4j-louvain-modularity/

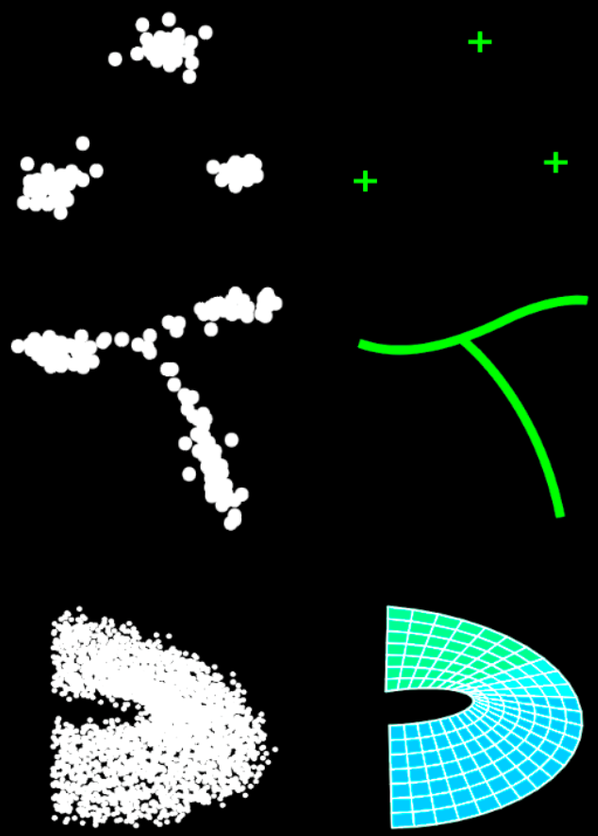
Order cells by developmental state: pseudotime / trajectory analysis





Trapnell, C. et al. (2014)

StructDR - unified structure extraction and inference of confidence set



Cluster

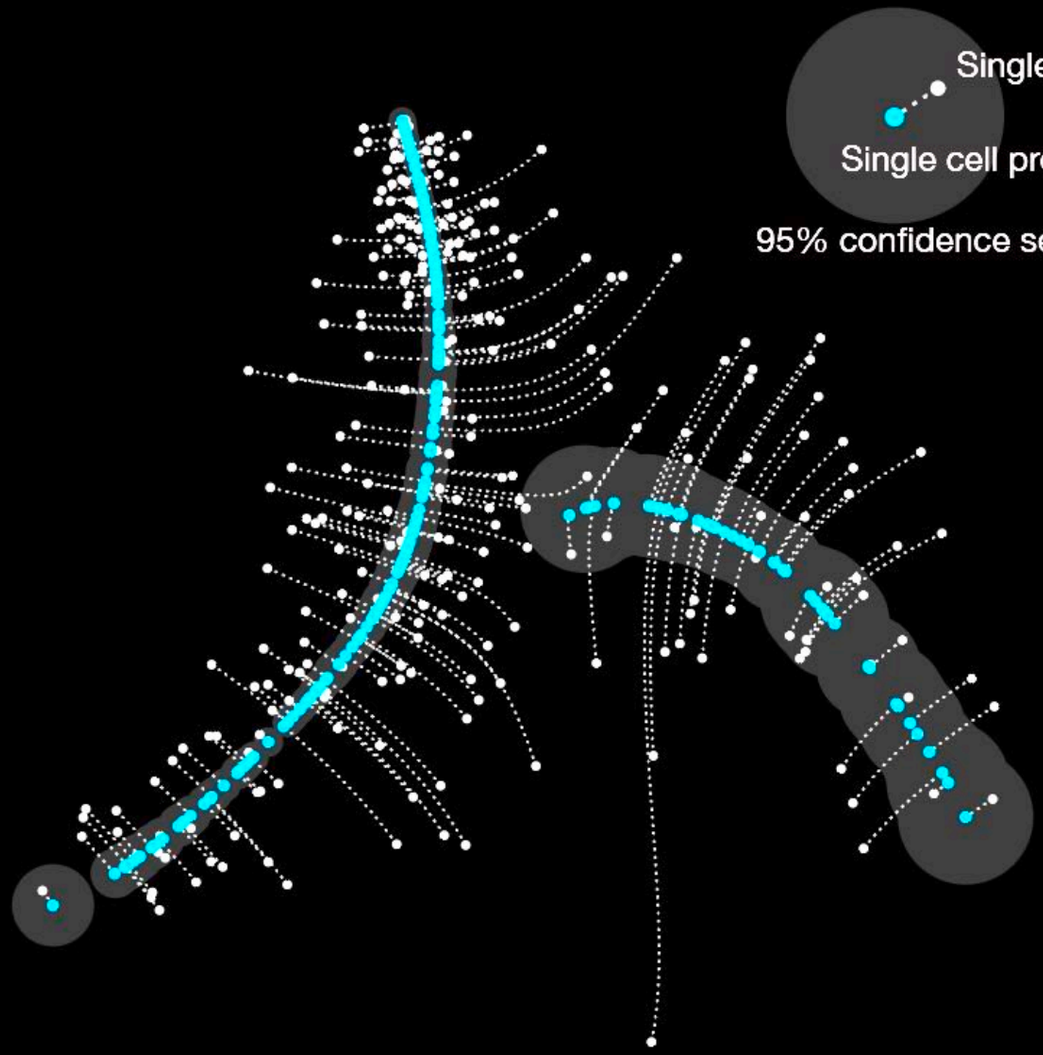
Cell type / state

Trajectory

Developmental trajectories

Surface

Spatial maps,
complex heterogeneous
population



StructDR - unified structure extraction and inference of confidence set