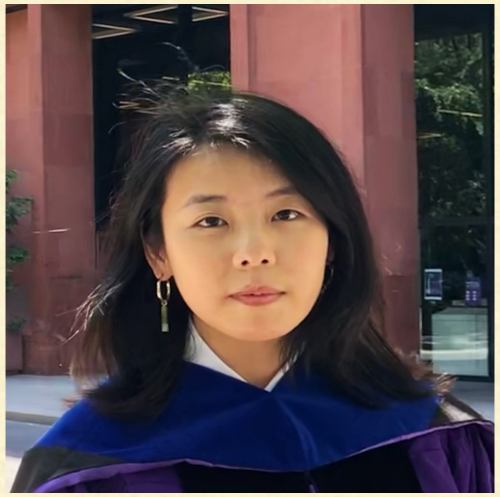
Data-driven methods for inference in dynamical systems

Björn Sandstede

Sam Maffa
(Broad Institute)



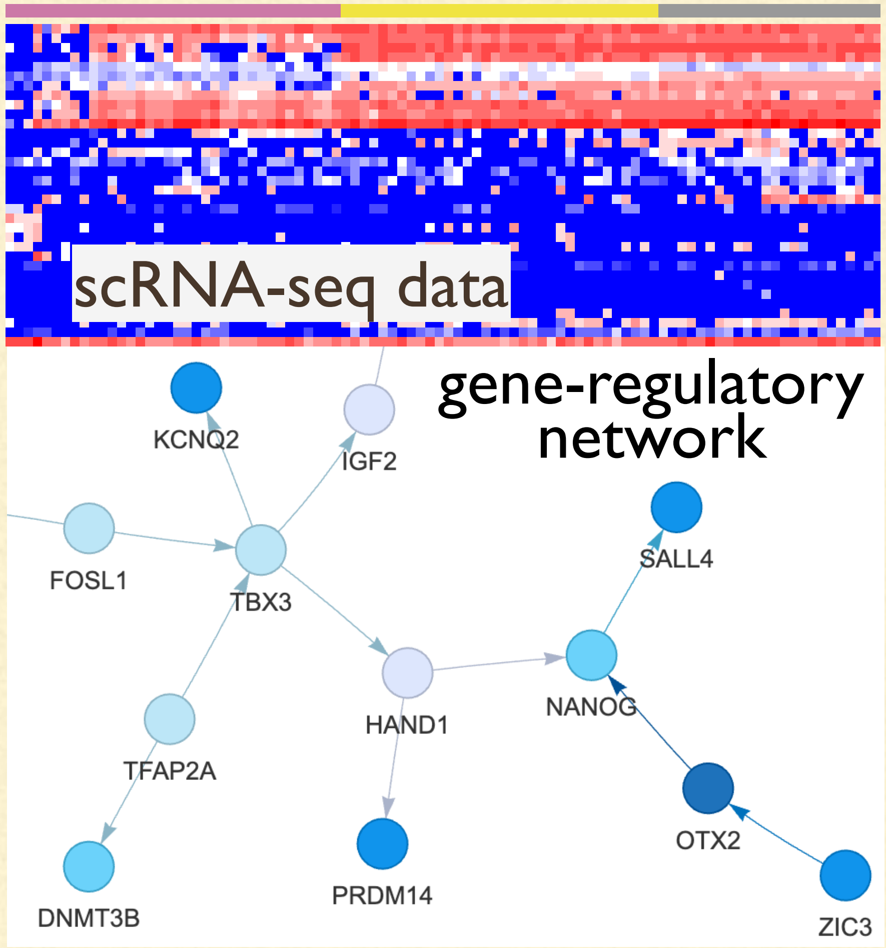
Wenjun Zhao
(Kantorovich Initiative
& Wake Forest U)

Erica Larschan
(Brown U)

Ritambhara Singh
(Brown U)

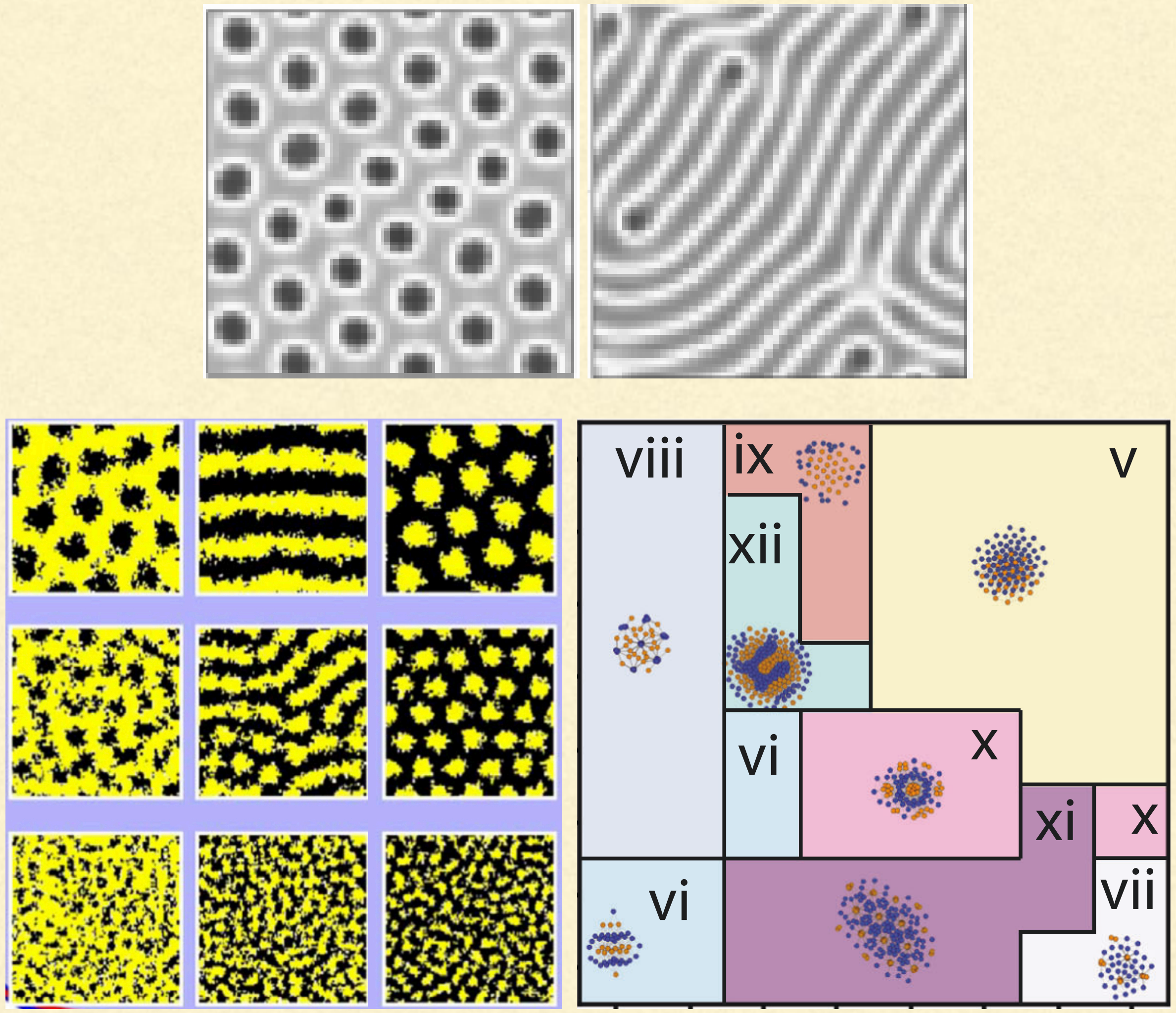
Two themes connected by optimal transport as the tool

Inferring gene-regulatory networks



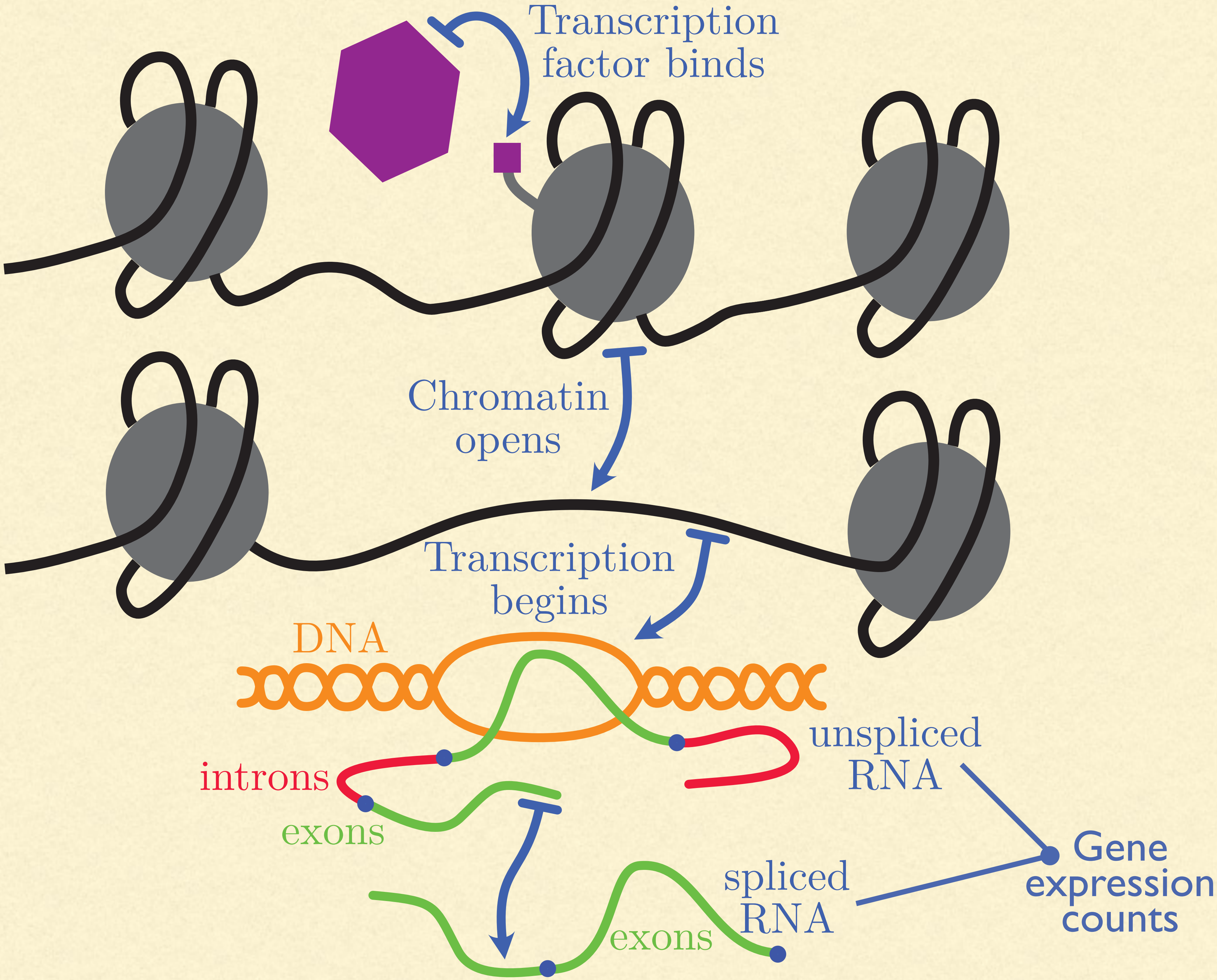
[Zhao, Larschan, S., Singh]

Quantifying patterns and tracing bifurcation curves

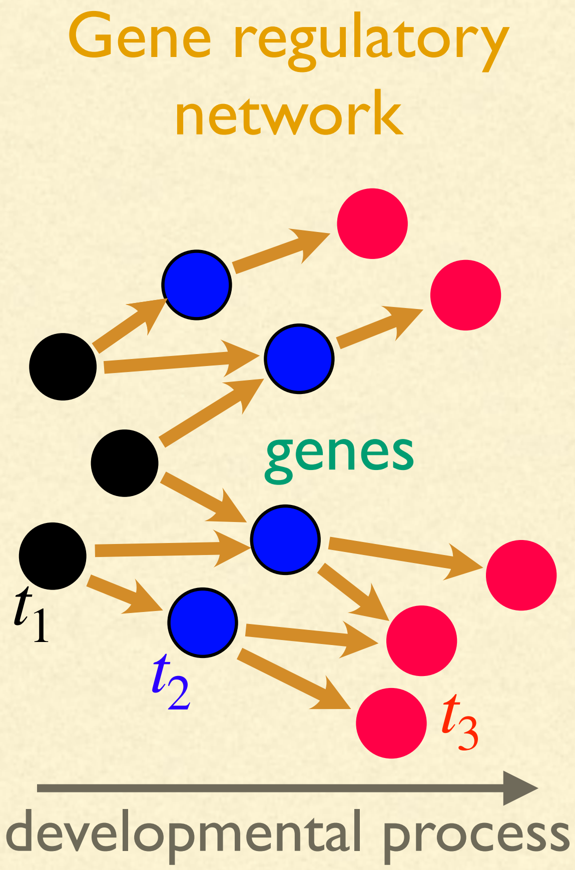


[Zhao, Maffa, S.]

Gene transcription in cells

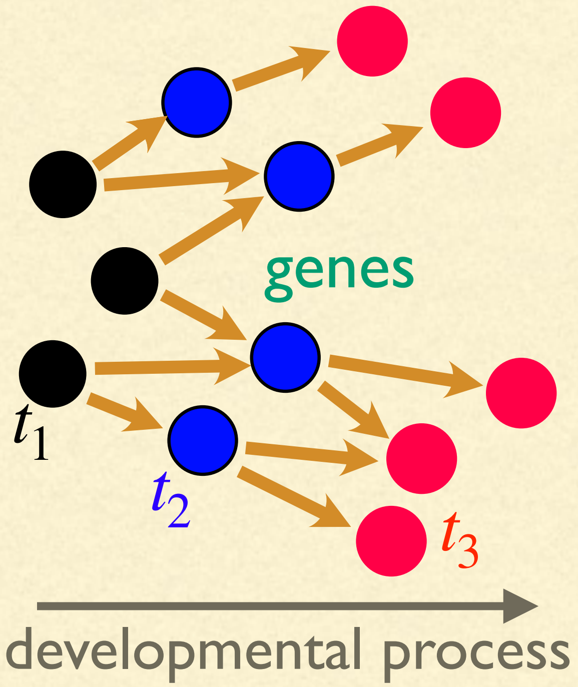


Infer gene-regulatory networks from single-cell data

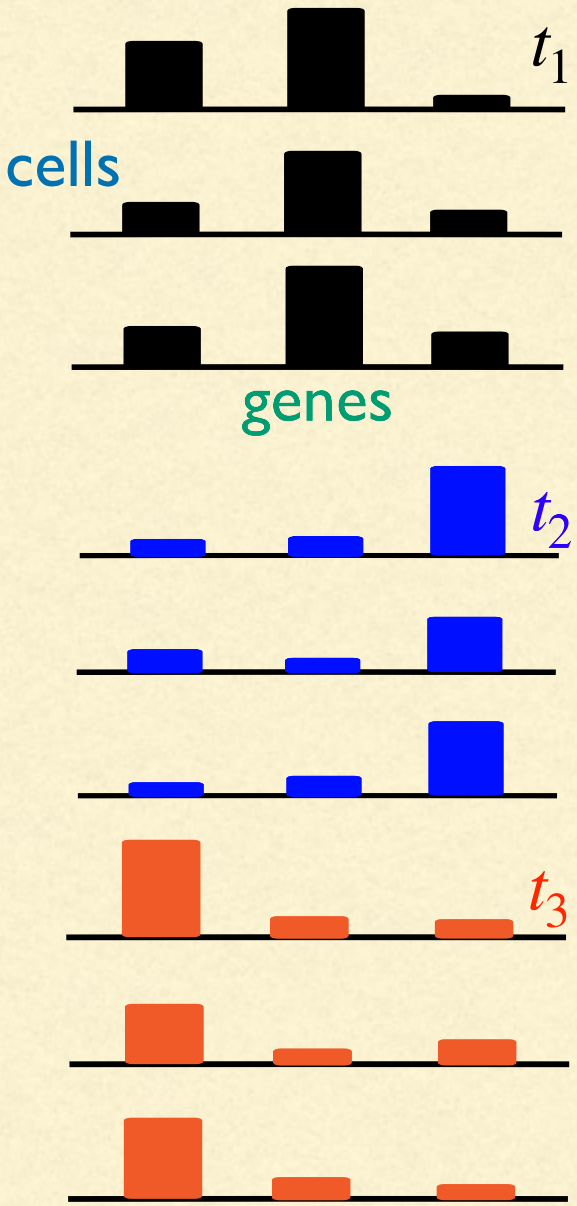


Infer gene-regulatory networks from single-cell data

Gene regulatory network

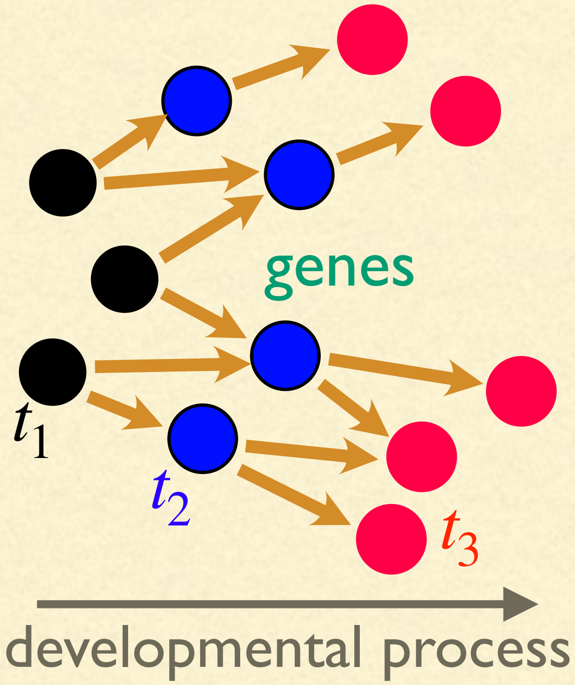


Gene expression levels

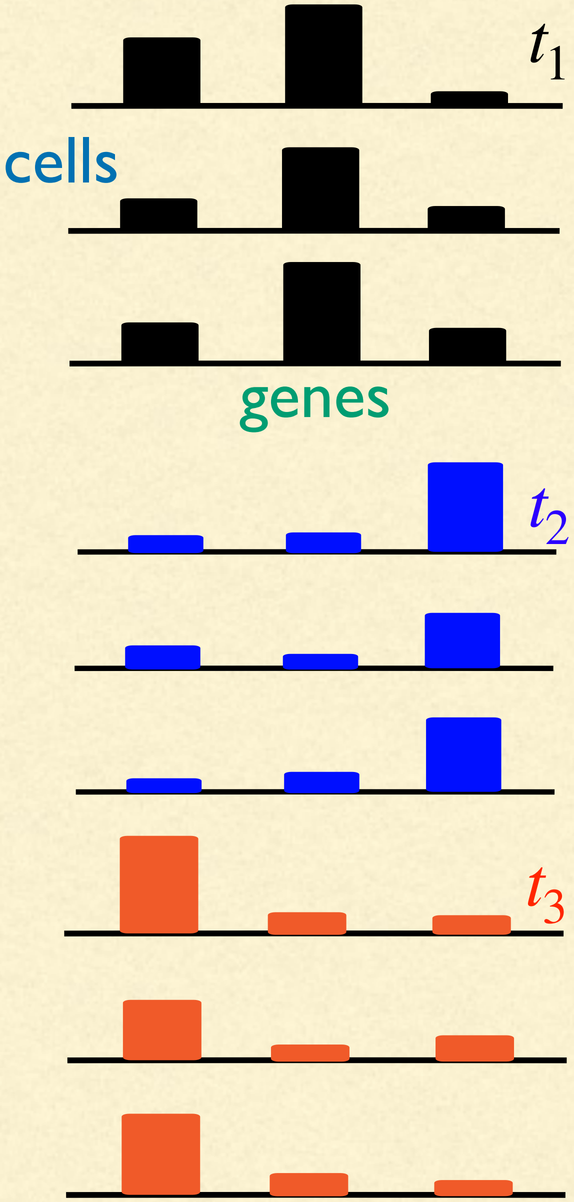


Infer gene-regulatory networks from single-cell data

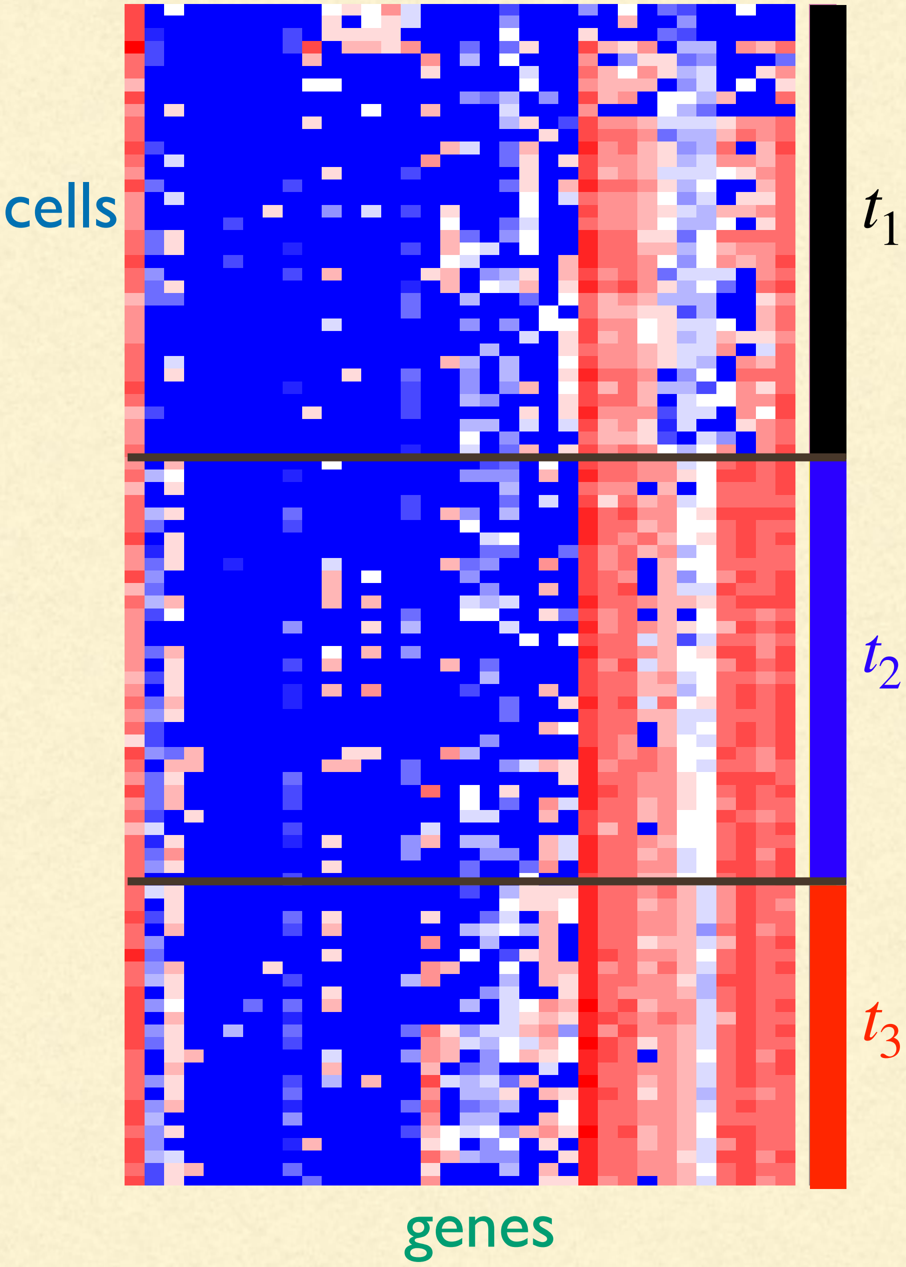
Gene regulatory network



Gene expression levels

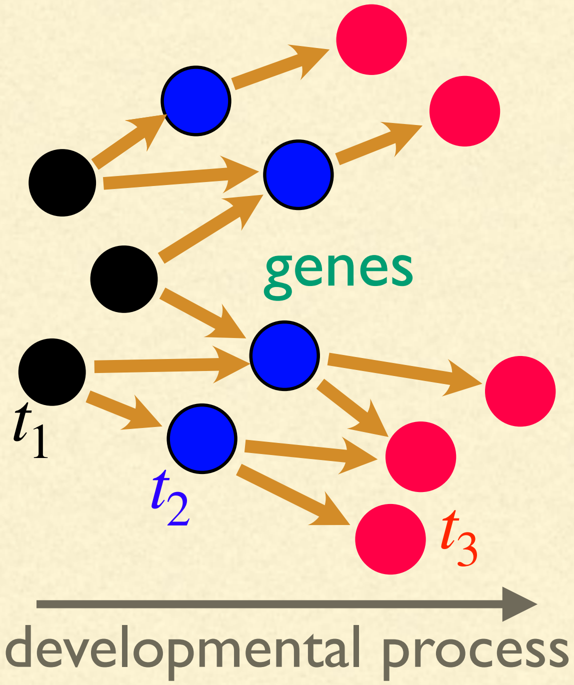


Single cell data

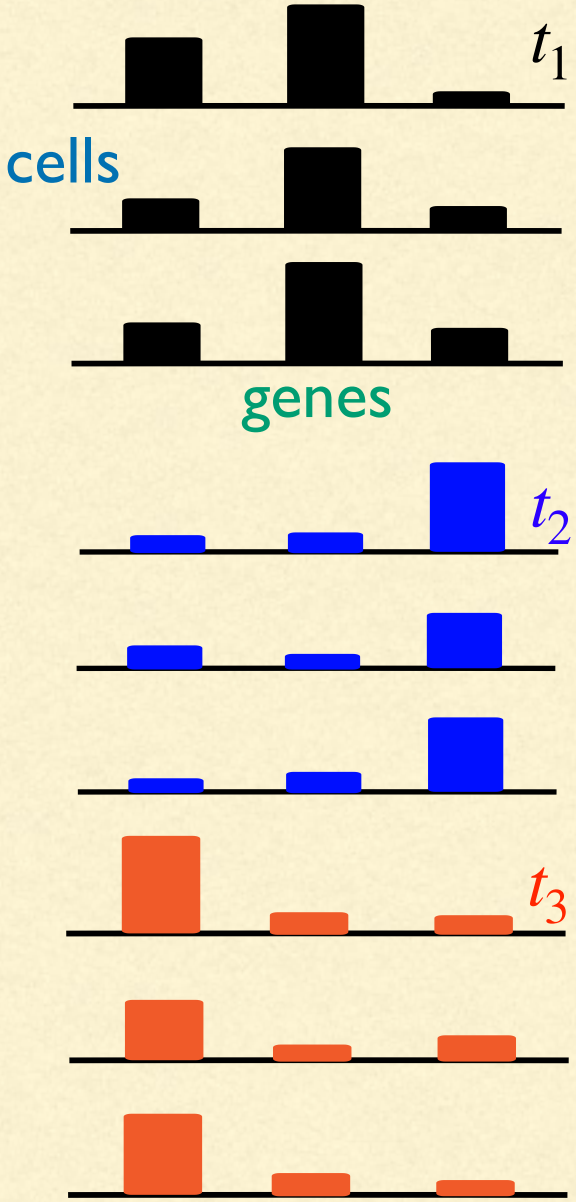


Infer gene-regulatory networks from single-cell data

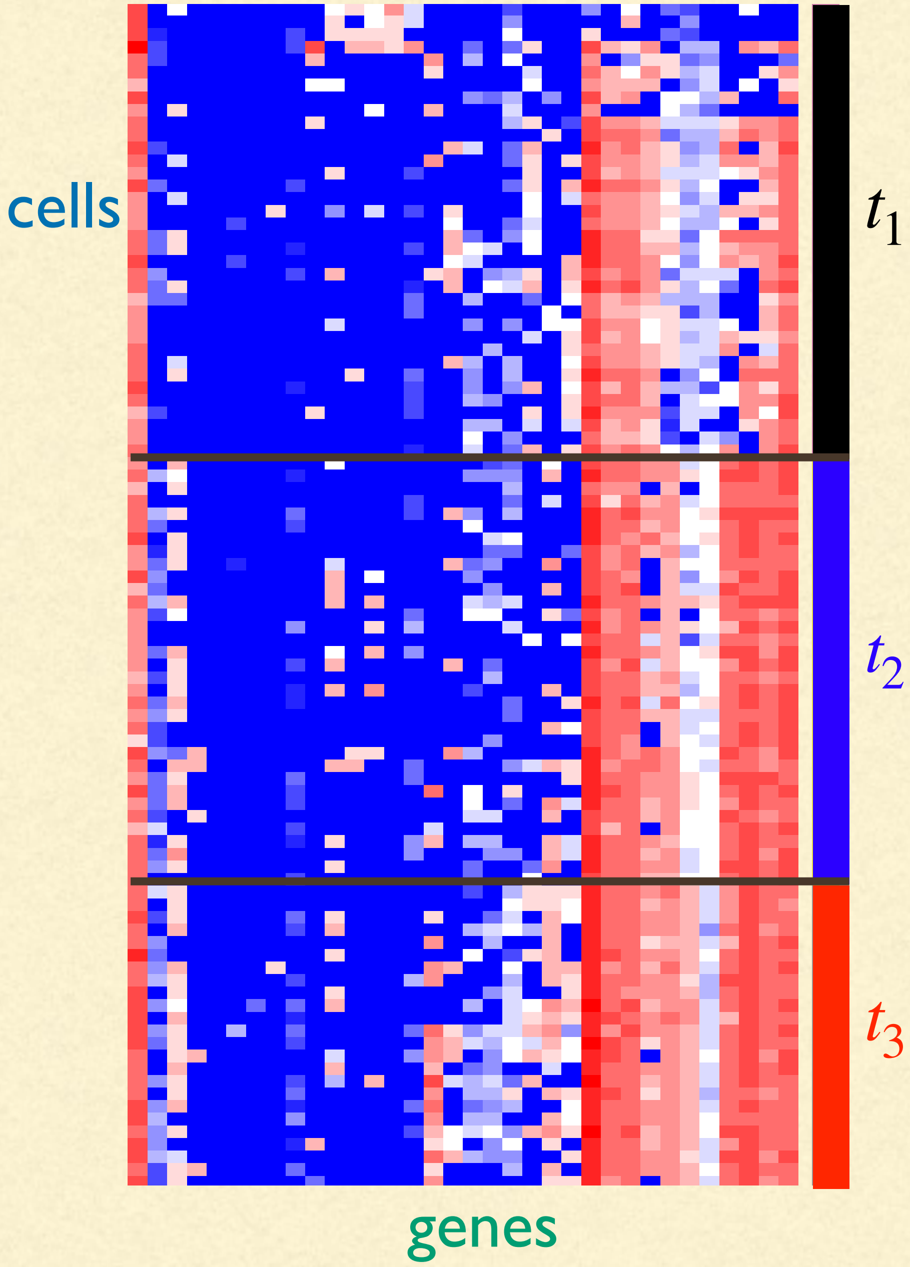
Gene regulatory network



Gene expression levels

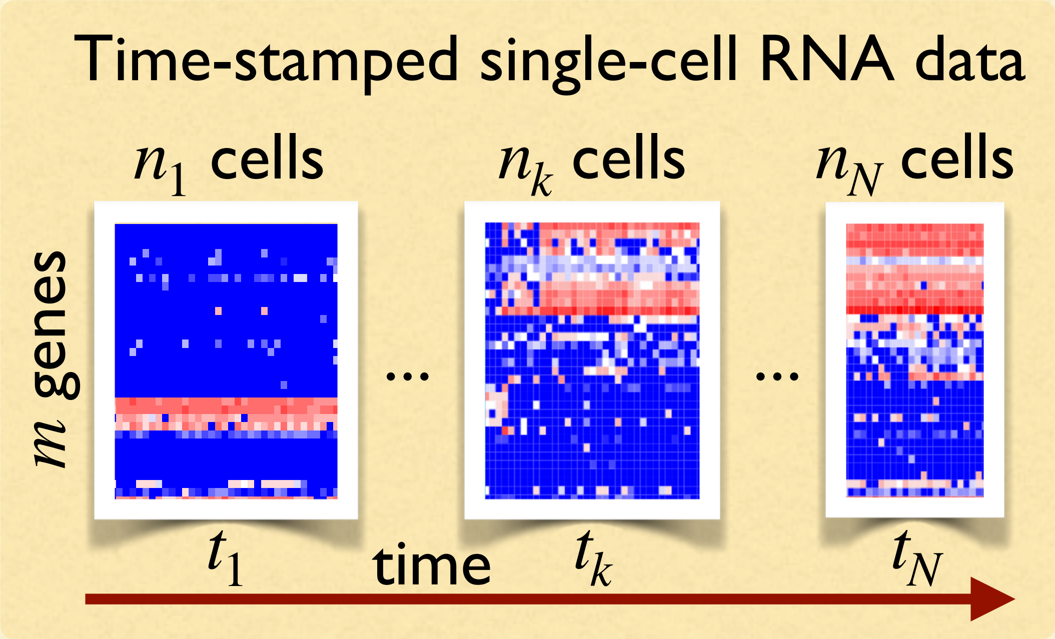


Single cell data



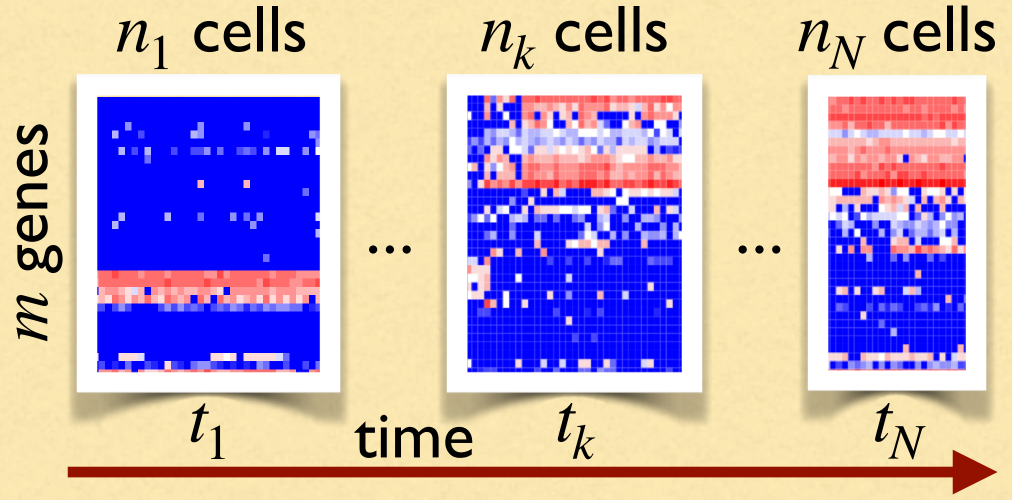
Inference: Infer gene-regulatory networks from time-stamped single-cell count matrices

OTVelo Pipeline

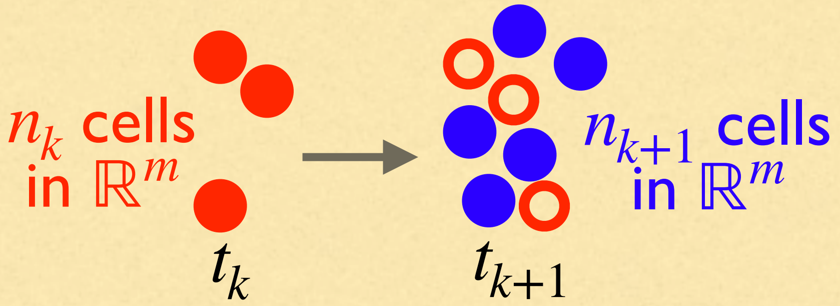


OTVelo Pipeline

Time-stamped single-cell RNA data



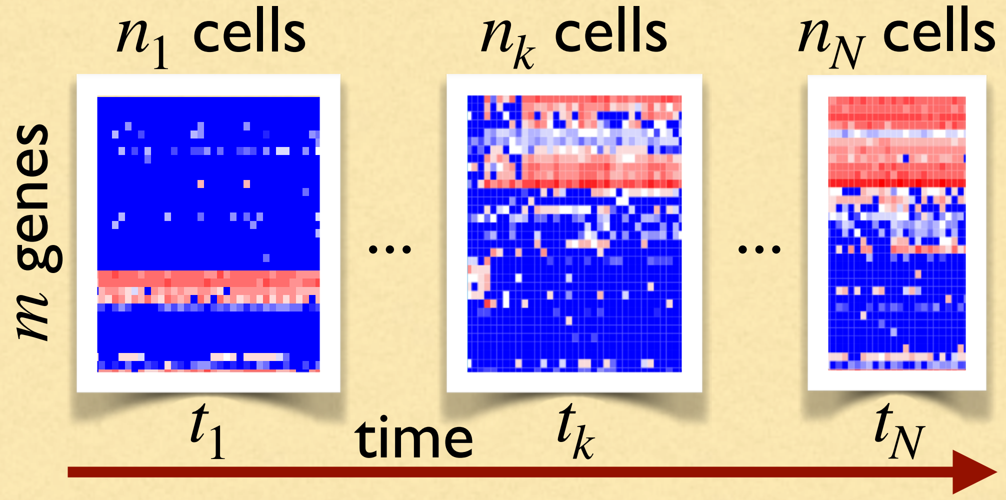
Predict cell trajectories


n_k cells in $\mathbb{R}^m$
t_k
n_{k+1} cells in $\mathbb{R}^m$
t_{k+1}

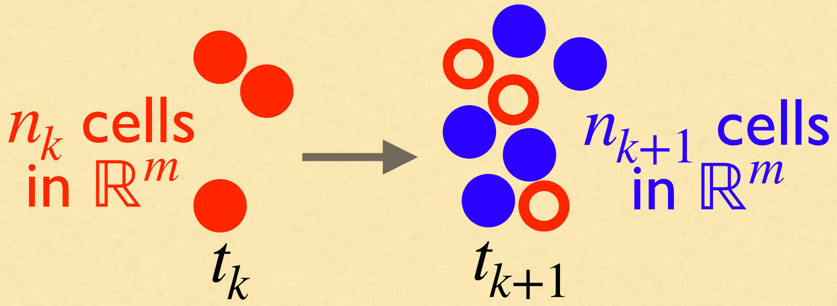
Fused Gromov–Wasserstein optimal transport

OTVelo Pipeline

Time-stamped single-cell RNA data

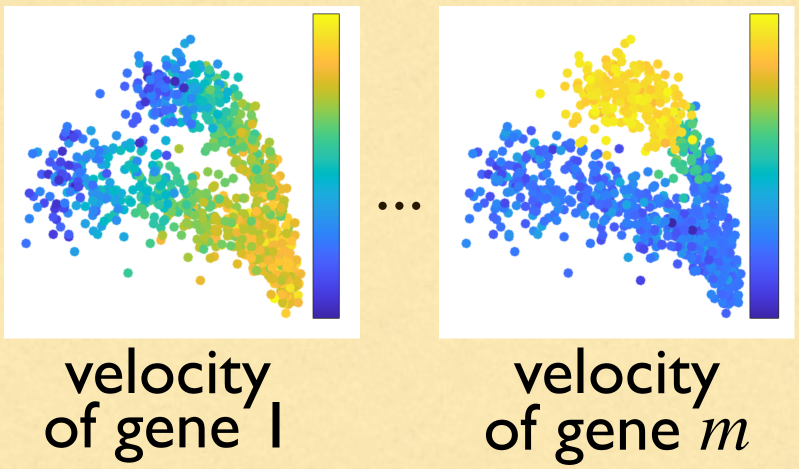


Predict cell trajectories


n_k cells in $\mathbb{R}^m$
t_k
n_{k+1} cells in $\mathbb{R}^m$
t_{k+1}

Fused Gromov–Wasserstein optimal transport

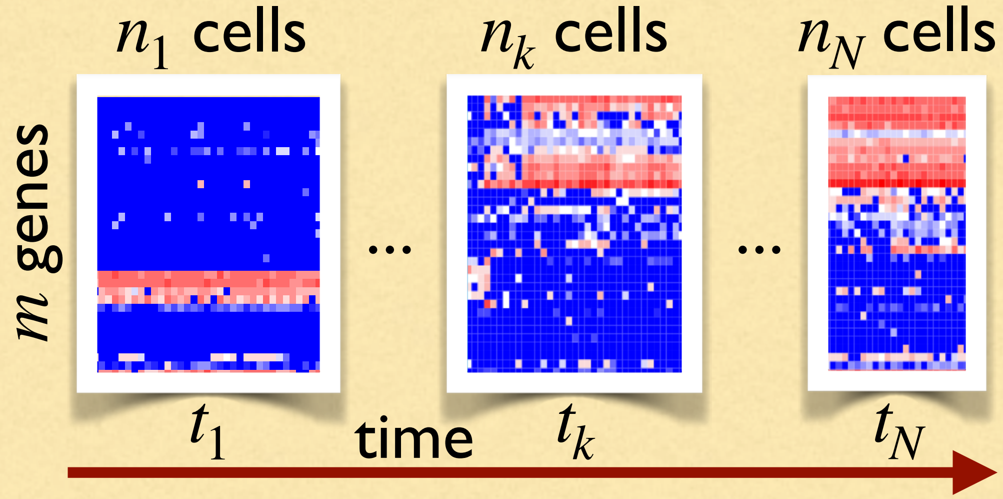
Infer gene velocity $v(x, t) \in \mathbb{R}^m$



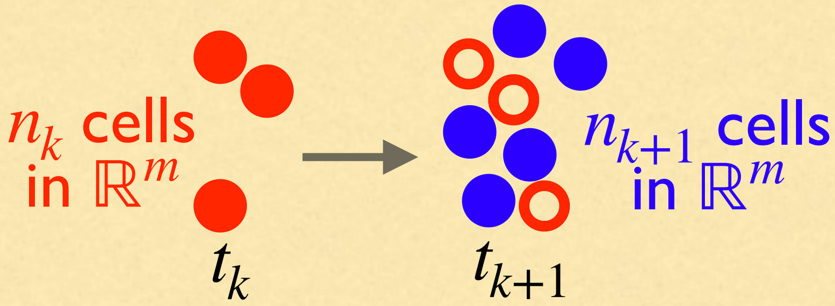
Finite differences on cell trajectories

OTVelo Pipeline

Time-stamped single-cell RNA data

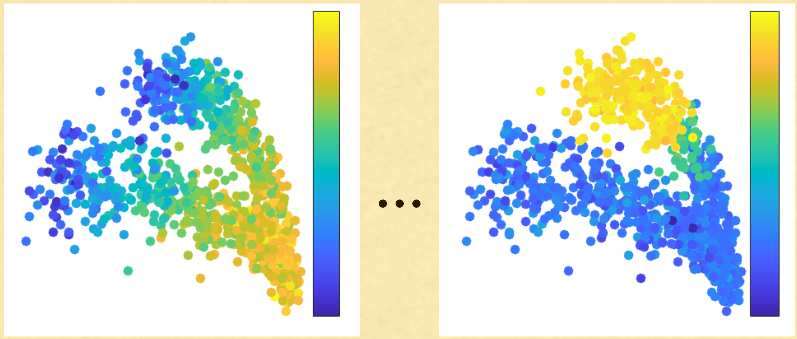


Predict cell trajectories


n_k cells in $\mathbb{R}^m$
t_k
n_{k+1} cells in $\mathbb{R}^m$
t_{k+1}

Fused Gromov–Wasserstein optimal transport

Infer gene velocity $v(x,t) \in \mathbb{R}^m$



velocity of gene 1

velocity of gene m

Finite differences on cell trajectories

Estimate gene-interaction matrix

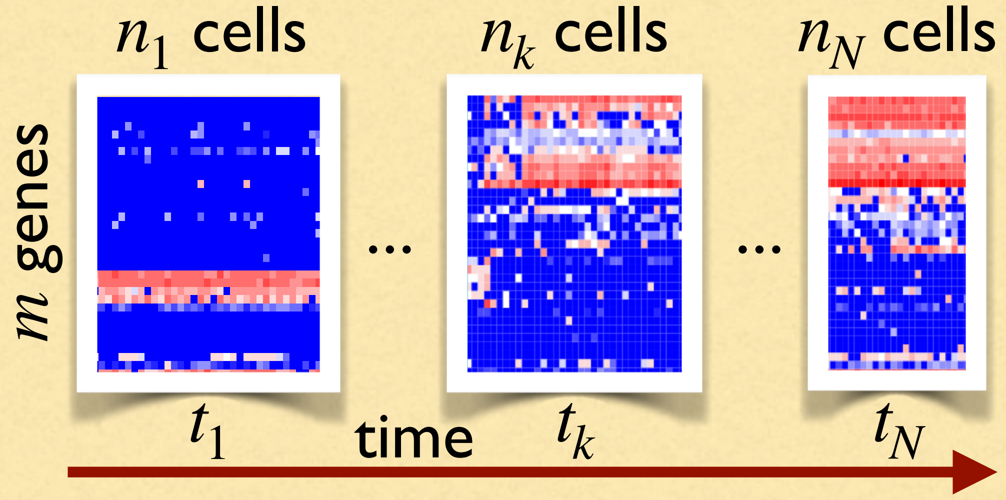
$$v(x, t_{k+1}) = A(t_k, t_{k+1}) v(x, t_k)$$

$$A(t_k, t_{k+1}) \in \mathbb{R}^{m \times m}$$

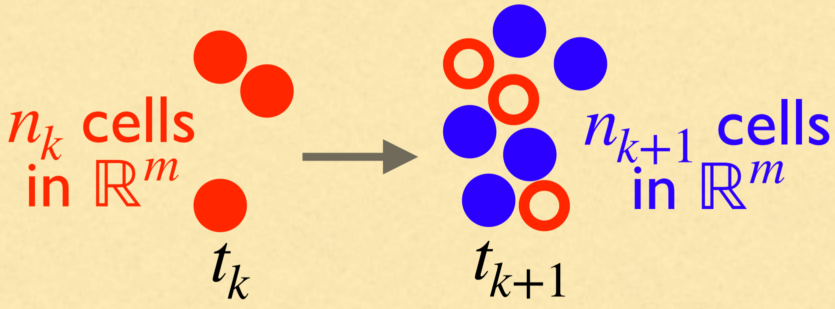
Regularized linear regression

OTVelo Pipeline

Time-stamped single-cell RNA data



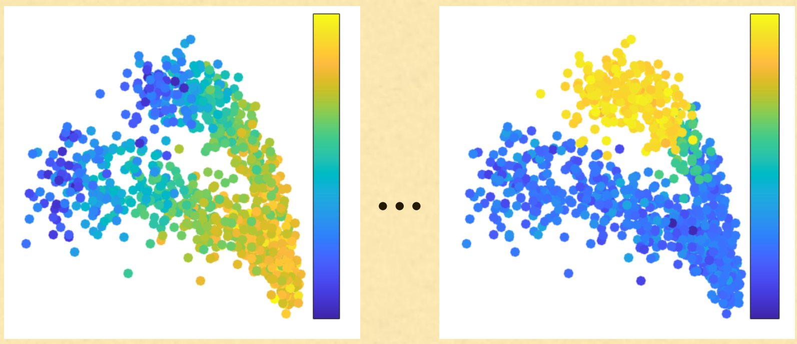
Predict cell trajectories



n_k cells in $\mathbb{R}^m$
t_k
n_{k+1} cells in $\mathbb{R}^m$
t_{k+1}

Fused Gromov–Wasserstein optimal transport

Infer gene velocity $v(x, t) \in \mathbb{R}^m$



velocity of gene 1

velocity of gene m

Finite differences on cell trajectories

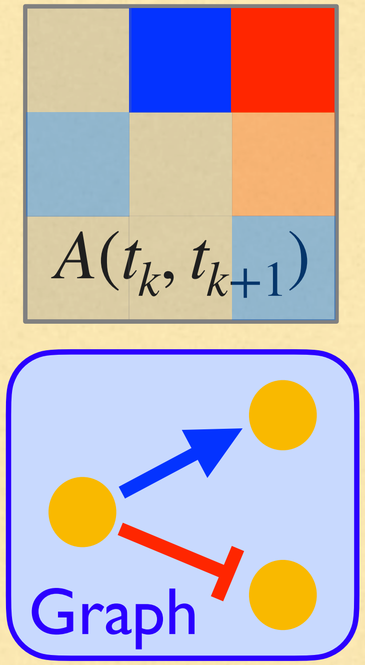
Estimate gene-interaction matrix

$$v(x, t_{k+1}) = A(t_k, t_{k+1}) v(x, t_k)$$

$$A(t_k, t_{k+1}) \in \mathbb{R}^{m \times m}$$

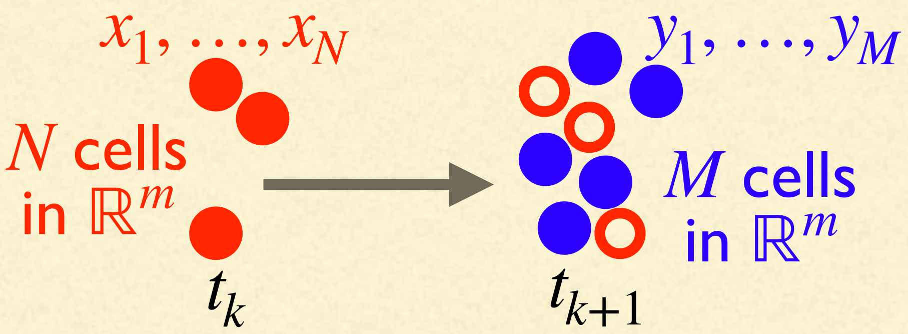
Regularized linear regression

Gene regulatory network



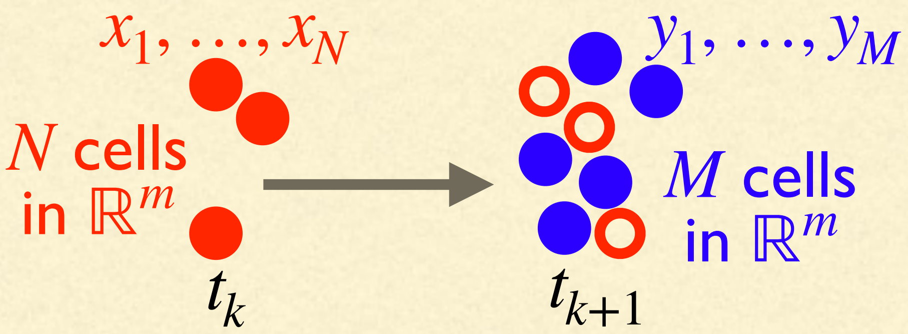
Optimal transport

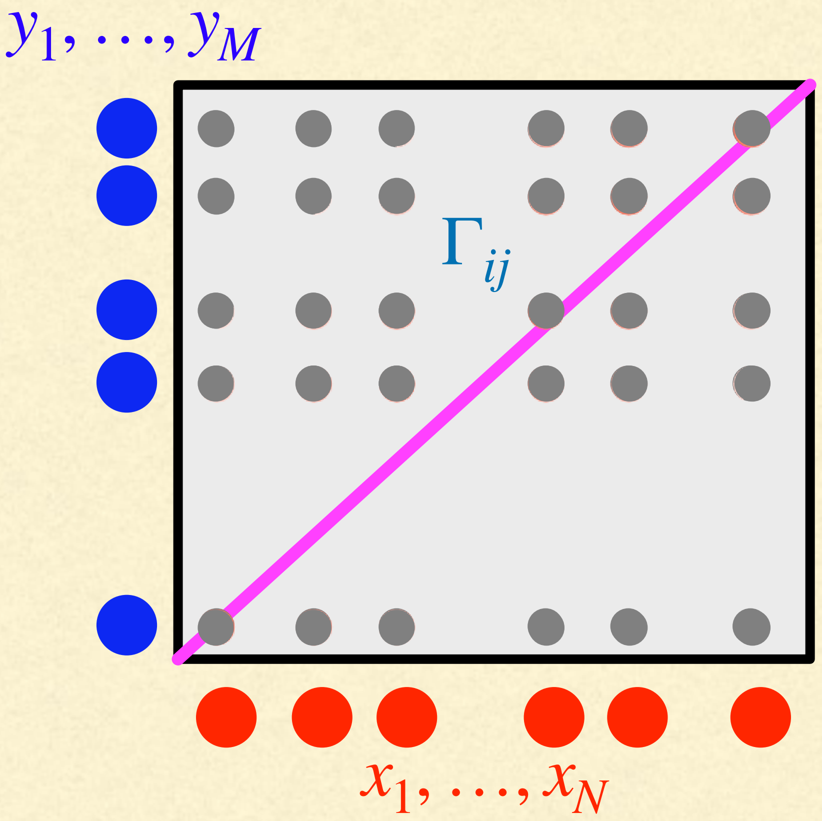
- Task: predict cell positions at later time points
- Assumption: cell samples at different time points represent full cell population



Optimal transport

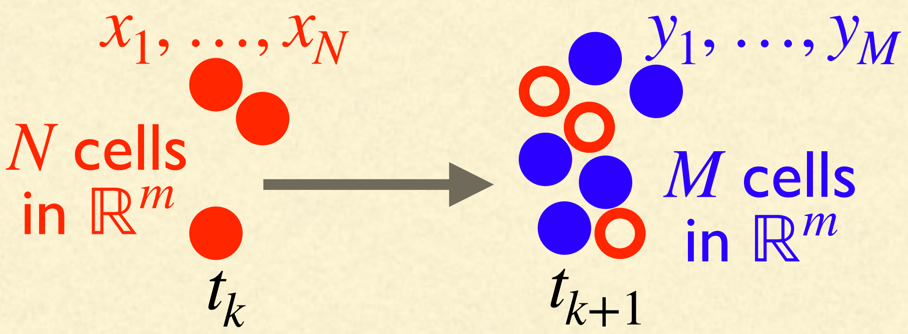
- Task: predict cell positions at later time points
- Assumption: cell samples at different time points represent full cell population





Optimal transport

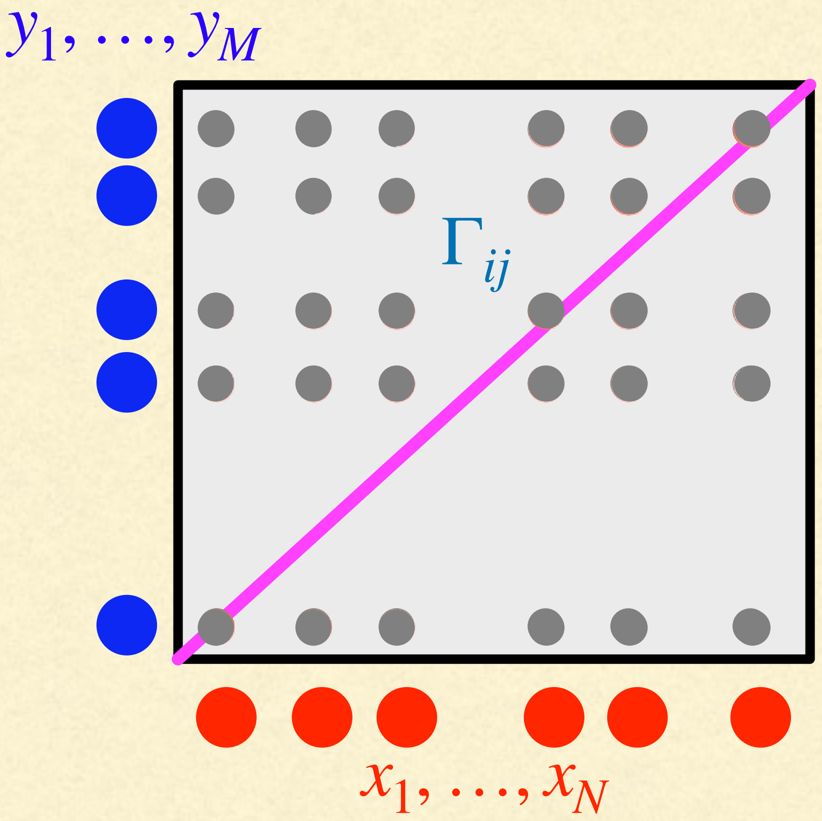
- Task: predict cell positions at later time points
- Assumption: cell samples at different time points represent full cell population



Find joint probability distribution Γ ($\Gamma_{ij} \geq 0$, $\sum_{i=1}^{N} \Gamma_{ij} = \frac{1}{M}$, $\sum_{j=1}^{M} \Gamma_{ij} = \frac{1}{N}$) as solution to

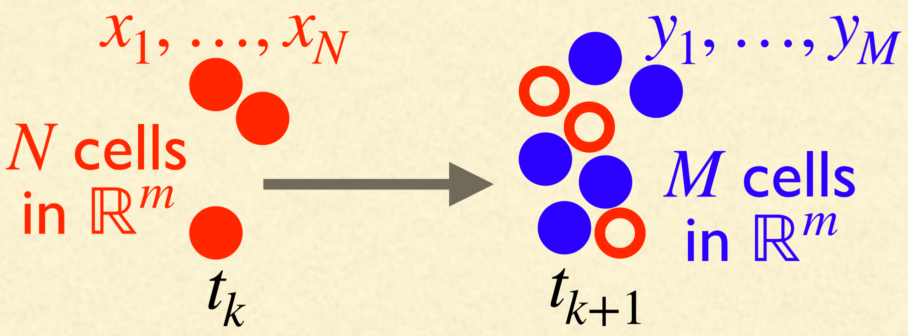
$$\Gamma = \arg\min_{\Gamma} \sum_{\substack{i=1,\ldots,N \\ j=1\ldots,M}} |x_i - y_j|^2_{\mathbb{R}^m} \, \Gamma_{ij}$$

We interpret Γ_{ij} as the probability that x_i is mapped to y_j



Optimal transport

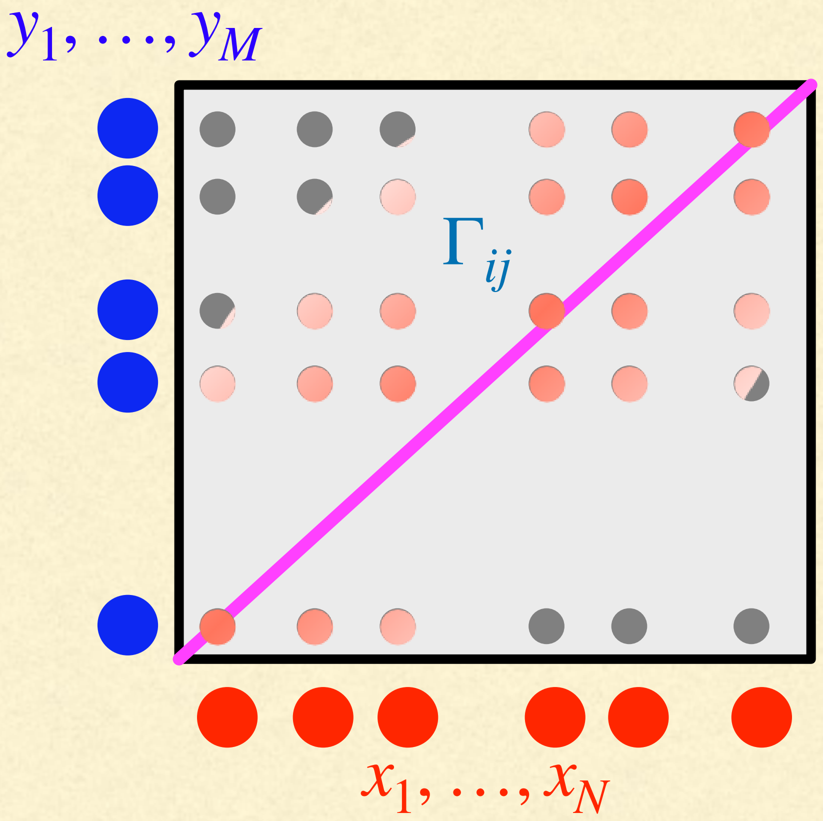
- Task: predict cell positions at later time points
- Assumption: cell samples at different time points represent full cell population



Find joint probability distribution Γ ($\Gamma_{ij} \geq 0$, $\sum_{i=1}^{N} \Gamma_{ij} = \frac{1}{M}$, $\sum_{j=1}^{M} \Gamma_{ij} = \frac{1}{N}$) as solution to

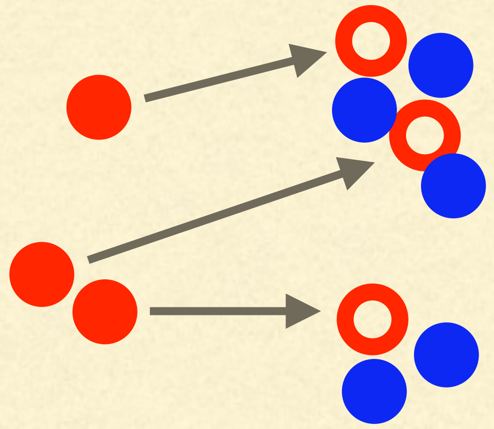
$$\Gamma = \arg\min_{\Gamma} \sum_{\substack{i=1,\ldots,N \\ j=1\ldots,M}} |x_i - y_j|^2_{\mathbb{R}^m} \, \Gamma_{ij}$$

We interpret Γ_{ij} as the probability that x_i is mapped to y_j



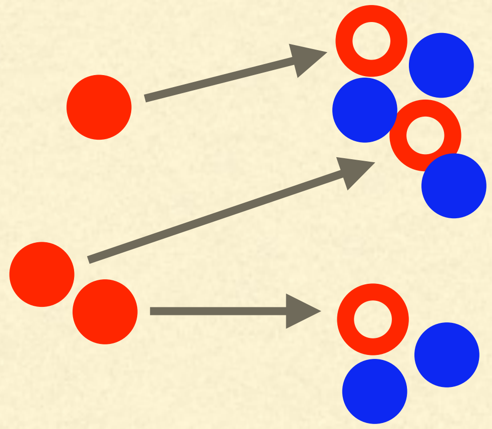
Gromov–Wasserstein optimal transport

Optimal transport
does not respect
local geometry

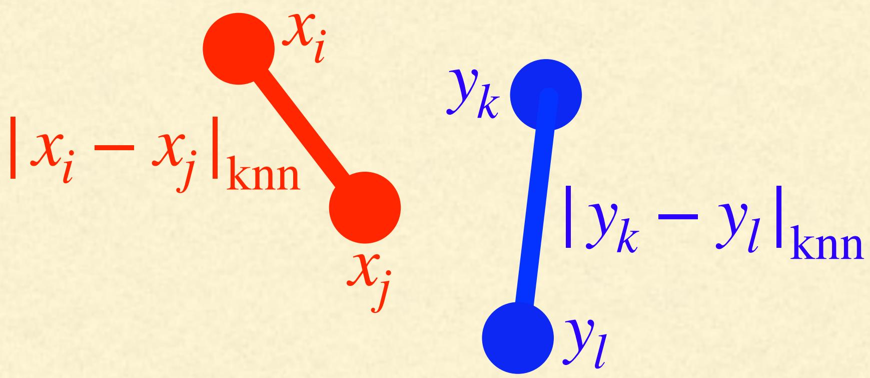


Gromov–Wasserstein optimal transport

Optimal transport
does not respect
local geometry



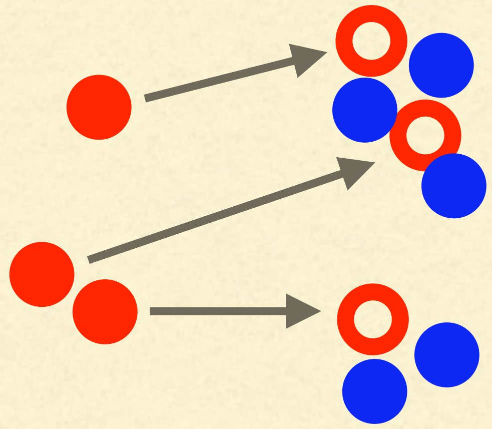
Include cost function
that aims to preserve
pairwise distances



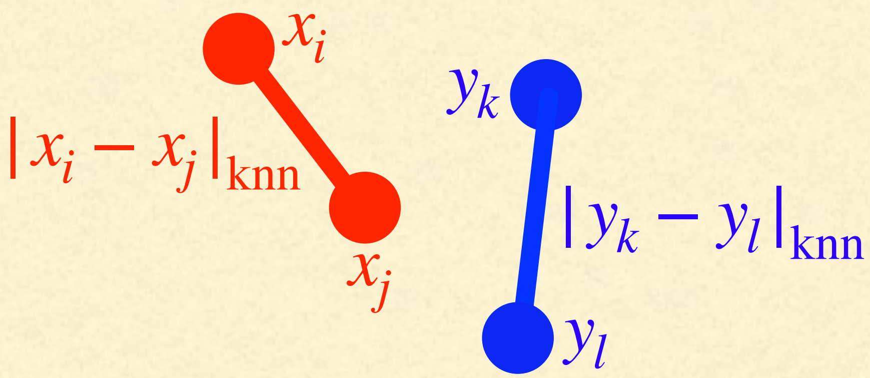
Distance based on
k-nearest-neighbor
graph

Gromov–Wasserstein optimal transport

Optimal transport does not respect local geometry



Include cost function that aims to preserve pairwise distances



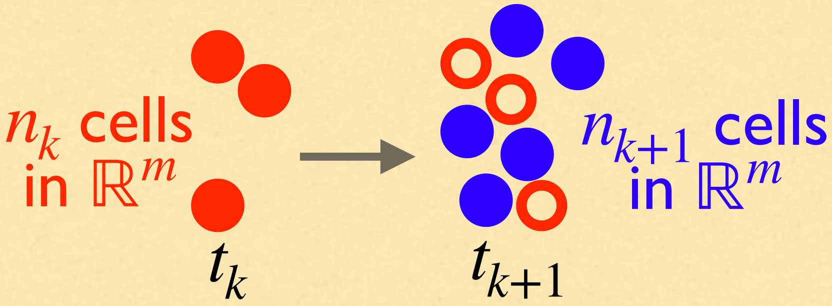
Distance based on k-nearest-neighbor graph

Find Γ as solution $$\arg\min_{\Gamma} \sum_{\substack{i,j=1,\dots,N \\ k,l=1\dots,M}} \Big| |x_i - x_j|_{\mathrm{knn}} - |y_k - y_l|_{\mathrm{knn}} \Big| \Gamma_{ik}\Gamma_{jl}$$

Cost function penalizes moving cells closer or farther apart but does not incorporate distance between the two data sets

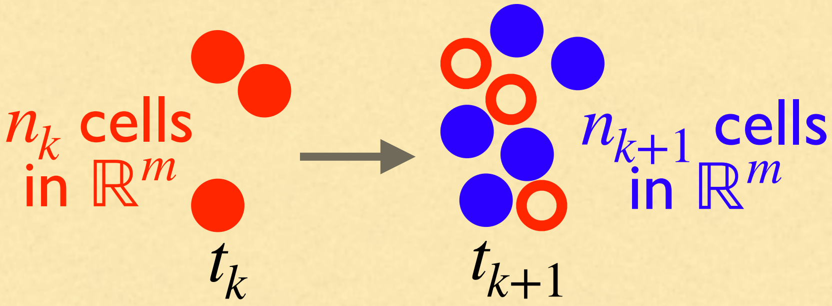
Fused Gromov–Wasserstein transport

Fused Gromov–Wasserstein optimal transport


n_k cells in $\mathbb{R}^m$
t_k
t_{k+1}
n_{k+1} cells in $\mathbb{R}^m$

Fused Gromov–Wasserstein transport

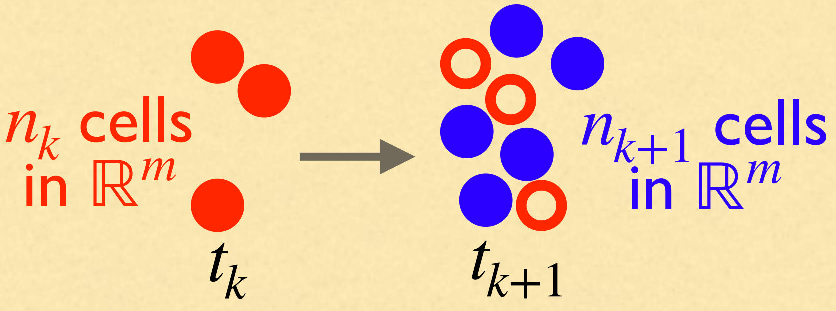
Fused Gromov–Wasserstein optimal transport


n_k cells in $\mathbb{R}^m$
t_k
t_{k+1}
n_{k+1} cells in $\mathbb{R}^m$

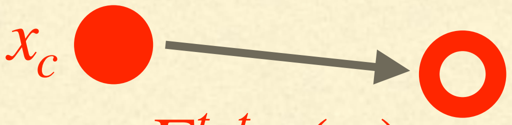
$$\underset{\Gamma}{\arg\min}\left[\alpha \sum_{\substack{1\le c\le n_k\\ 1\le d\le n_{k+1}}} \underset{\text{normal optimal transport}}{|x_c - y_d|^2_{\mathbb{R}^m}\Gamma_{cd}} + (1-\alpha)\sum_{\substack{1\le c,\tilde{c}\le n_k\\ 1\le d,\tilde{d}\le n_{k+1}}} \underset{\text{maps pairwise distances into each other: ensures geometry is preserved}}{\Big|\,|x_c - x_{\tilde{c}}|_{\mathrm{knn}} - |y_d - y_{\tilde{d}}|_{\mathrm{knn}}\Big|\,\Gamma_{cd}\Gamma_{\tilde{c}\tilde{d}}}\right]$$

Fused Gromov–Wasserstein transport

Fused Gromov–Wasserstein optimal transport


n_k cells in $\mathbb{R}^m$
t_k
t_{k+1}
n_{k+1} cells in $\mathbb{R}^m$

Barycentric projection is then used to map individual cells



$$\Gamma^{t_k,t_{k+1}}(x_c) := n_k \sum_{d=1,\ldots,n_{k+1}} \Gamma_{cd} y_d$$

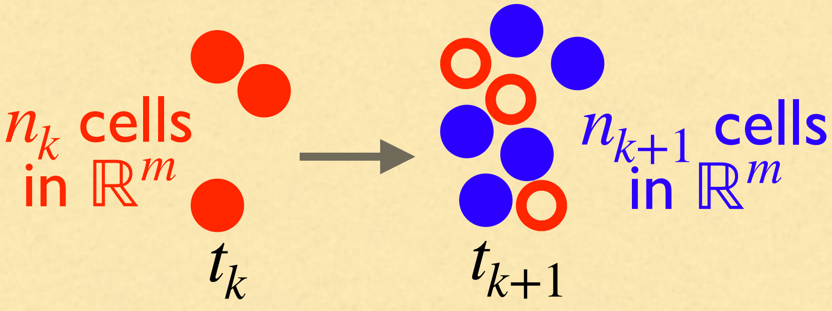
$$\underset{\Gamma}{\arg\min} \left[\alpha \sum_{\substack{1\leq c\leq n_k \\ 1\leq d\leq n_{k+1}}} |x_c - y_d|^2_{\mathbb{R}^m} \Gamma_{cd} + (1-\alpha) \sum_{\substack{1\leq c,\tilde{c}\leq n_k \\ 1\leq d,\tilde{d}\leq n_{k+1}}} \Big| |x_c - x_{\tilde{c}}|_{\mathrm{knn}} - |y_d - y_{\tilde{d}}|_{\mathrm{knn}} \Big| \Gamma_{cd}\Gamma_{\tilde{c}\tilde{d}} \right]$$

normal optimal transport

maps pairwise distances into each other: ensures geometry is preserved

Fused Gromov–Wasserstein transport

Fused Gromov–Wasserstein optimal transport


n_k cells in $\mathbb{R}^m$
t_k
t_{k+1}
n_{k+1} cells in $\mathbb{R}^m$

Barycentric projection is then used to map individual cells

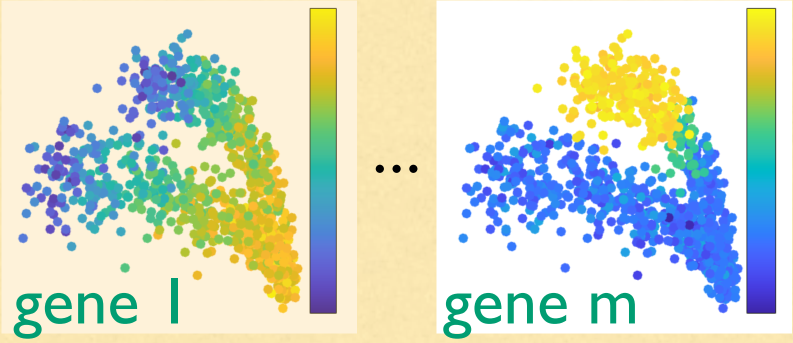
$$\Gamma^{t_k,t_{k+1}}(x_c) := n_k \sum_{d=1,\ldots,n_{k+1}} \Gamma_{cd} y_d$$

$$\arg\min_{\Gamma} \left[\alpha \sum_{\substack{1\le c\le n_k \\ 1\le d\le n_{k+1}}} |x_c - y_d|^2_{\mathbb{R}^m} \Gamma_{cd} + (1-\alpha) \sum_{\substack{1\le c,\tilde{c}\le n_k \\ 1\le d,\tilde{d}\le n_{k+1}}} \Big| |x_c - x_{\tilde{c}}|_{\mathrm{knn}} - |y_d - y_{\tilde{d}}|_{\mathrm{knn}} \Big| \Gamma_{cd}\Gamma_{\tilde{c}\tilde{d}} \right]$$

normal optimal transport

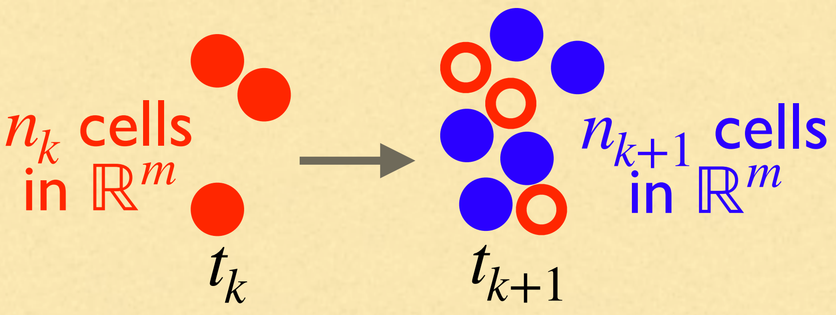
maps pairwise distances into each other: ensures geometry is preserved

Finite differences on cell trajectories



Fused Gromov–Wasserstein transport

Fused Gromov–Wasserstein optimal transport


n_k cells in $\mathbb{R}^m$
n_{k+1} cells in $\mathbb{R}^m$
t_k
t_{k+1}

Barycentric projection is then used to map individual cells

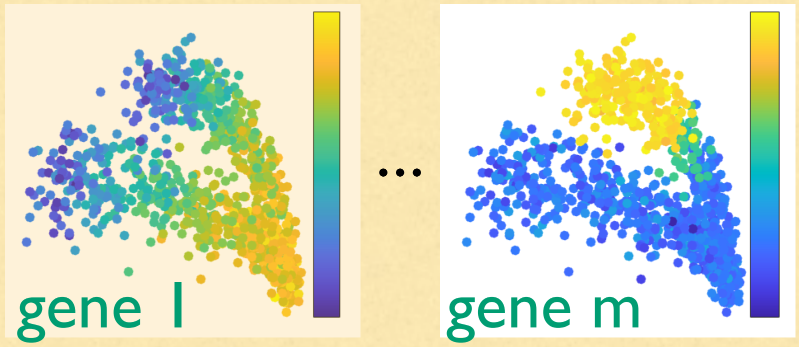
$$\Gamma^{t_k,t_{k+1}}(x_c) := n_k \sum_{d=1,\ldots,n_{k+1}} \Gamma_{cd}\, y_d$$

$$\arg\min_{\Gamma}\left[\alpha \sum_{\substack{1\le c\le n_k\\ 1\le d\le n_{k+1}}} |x_c - y_d|^2_{\mathbb{R}^m}\Gamma_{cd} + (1-\alpha)\sum_{\substack{1\le c,\tilde{c}\le n_k\\ 1\le d,\tilde{d}\le n_{k+1}}} \Big|\, |x_c - x_{\tilde{c}}|_{\text{knn}} - |y_d - y_{\tilde{d}}|_{\text{knn}} \Big| \Gamma_{cd}\Gamma_{\tilde{c}\tilde{d}}\right]$$

normal optimal transport

maps pairwise distances into each other: ensures geometry is preserved

Finite differences on cell trajectories

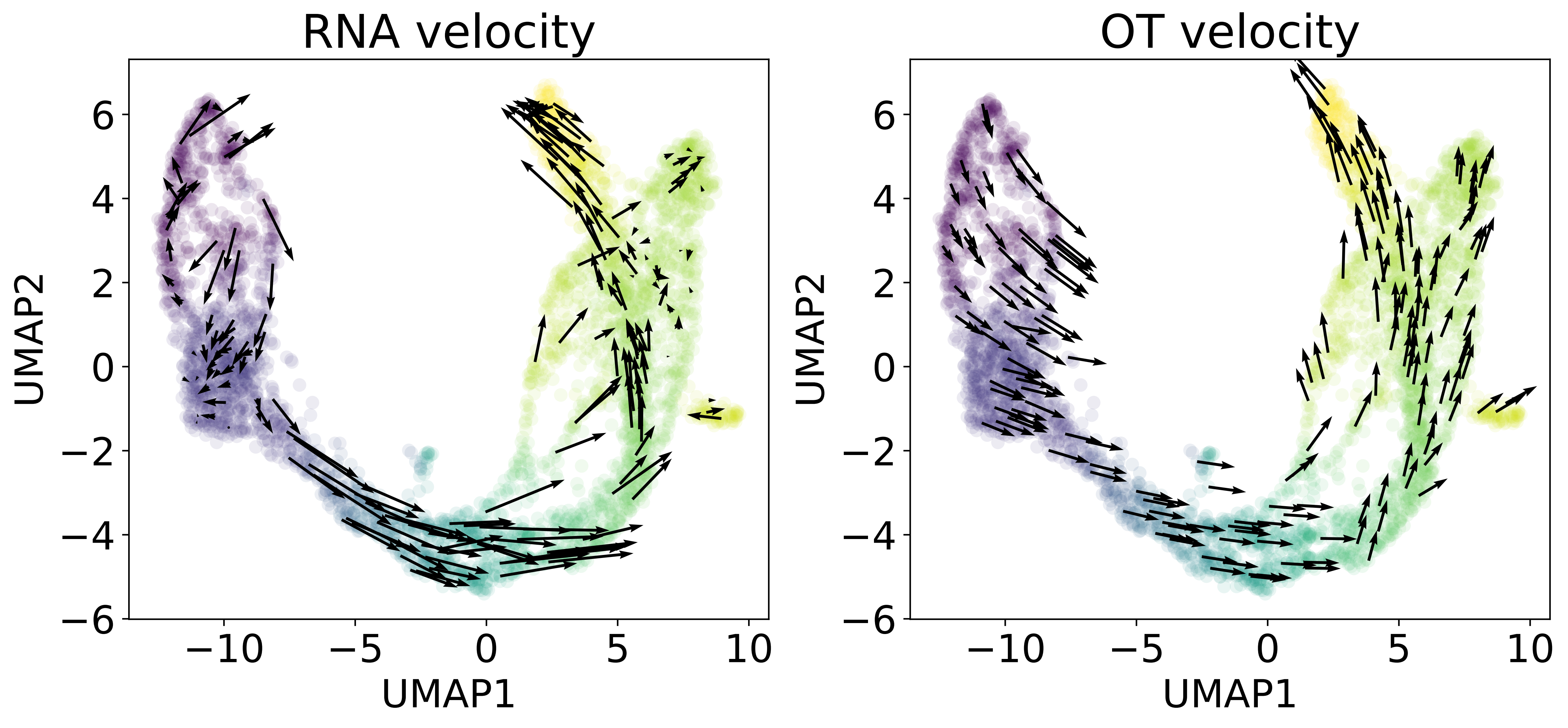


Gene velocities for each cell are defined by finite differences

$$v(x_c, t_k) := \frac{1}{t_{k+1} - t_k}\left(\Gamma^{t_k,t_{k+1}}(x_c) - x_c\right)$$

Comparison of RNA and OT cell velocities

scRNA data for pancreatic endocrinogenesis [Bastidas-Ponce et al.]



RNA velocity estimated in scVelo using reaction model for unspliced and spliced RNA counts [Bergen et al.]

Cell velocity estimated in OTVelo by finite differences

Inference of gene-to-gene interactions

Regularized linear regression

$$A = \arg \min_{A \in \mathbb{R}^{m \times m}} \Big[\| v(y, t_{k+1}) - Av(x, t_k) \| \\ + \lambda (r \|A\|_1 + (1 - r) \|A\|_2) \Big]$$

Predict velocities of data y at time t_{k+1} as linear function A of velocities of data x at time t_k and enforce sparsity through $\|A\|_1$

$\mathrm{sign}(A_{g1g2})$ indicates up- or down-regulation of gene g_2 by gene g_1

- Leads to sparse graphs (for $r \approx 1$)
- Computationally more expensive

Inference of gene-to-gene interactions

Regularized linear regression

$$A = \arg \min_{A \in \mathbb{R}^{m \times m}} \Big[\|v(y, t_{k+1}) - Av(x, t_k)\| \\ + \lambda(r\|A\|_1 + (1-r)\|A\|_2) \Big]$$

Predict velocities of data y at time t_{k+1} as linear function A of velocities of data x at time t_k and enforce sparsity through $\|A\|_1$

$\mathrm{sign}(A_{g1g2})$ indicates up- or down-regulation of gene g_2 by gene g_1

- Leads to sparse graphs (for $r \approx 1$)
- Computationally more expensive

Time-lagged correlation

$$C_{g_1 g_2} = \sum_{c=1}^{n_k} \sum_{d=1}^{n_{k+1}} v_{g_1}(x_c, t_k) v_{g_2}(y_d, t_{k+1}) \Gamma_{cd}^{t_k, t_{k+1}}$$

Correlation between velocities of cell c at time t_k and cell d at time t_{k+1} weighted by likelihood that cell d descended from cell c

$\mathrm{sign}(C_{g1g2})$ suggests up- or down-regulation of gene g_2 by gene g_1

- Leads to denser graphs
- Computationally efficient
  (scales well with number of genes)

Inference of gene-to-gene interactions

Regularized linear regression

$$A = \arg \min_{A \in \mathbb{R}^{m\times m}} \Big[\| v(y, t_{k+1}) - A v(x, t_k) \| + \lambda (r \|A\|_1 + (1-r)\|A\|_2) \Big]$$

Predict velocities of data y at time t_{k+1} as linear function A of velocities of data x at time t_k and enforce sparsity through $\|A\|_1$

$\mathrm{sign}(A_{g1g2})$ indicates up- or down-regulation of gene g_2 by gene g_1

- Leads to sparse graphs (for $r \approx 1$)
- Computationally more expensive

Time-lagged correlation

$$C_{g_1 g_2} = \sum_{c=1}^{n_k} \sum_{d=1}^{n_{k+1}} v_{g_1}(x_c, t_k) v_{g_2}(y_d, t_{k+1}) \Gamma_{cd}^{t_k, t_{k+1}}$$

Correlation between velocities of cell c at time t_k and cell d at time t_{k+1} weighted by likelihood that cell d descended from cell c

$\mathrm{sign}(C_{g1g2})$ suggests up- or down-regulation of gene g_2 by gene g_1

- Leads to denser graphs
- Computationally efficient
  (scales well with number of genes)

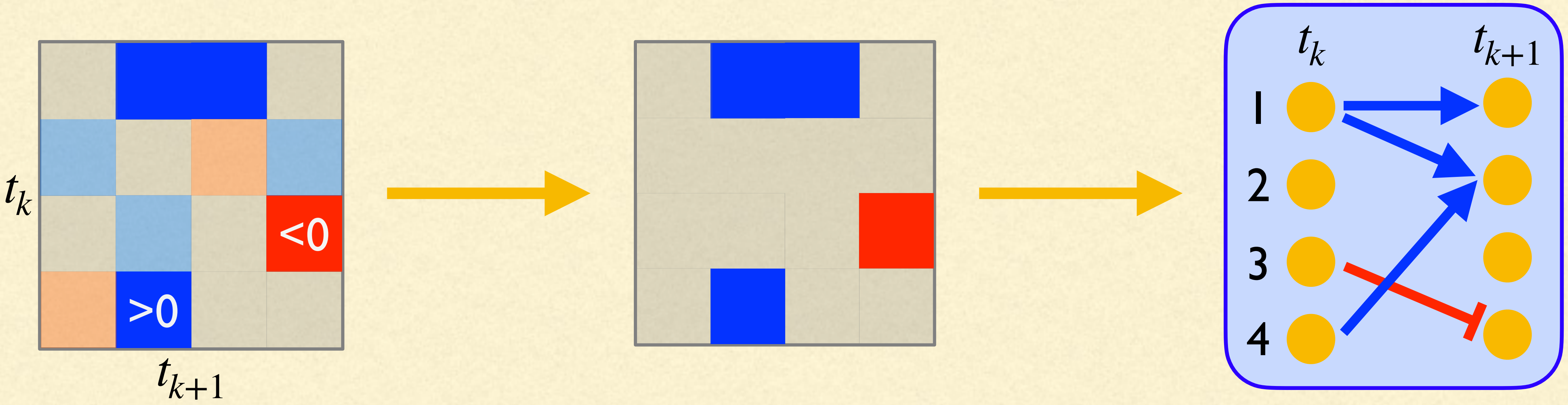
Summation over k leads to prediction of global gene-regulatory network rather than a dynamic network

Inference of gene-regulatory networks

Gene interaction matrix $A(t_k, t_{k+1})$

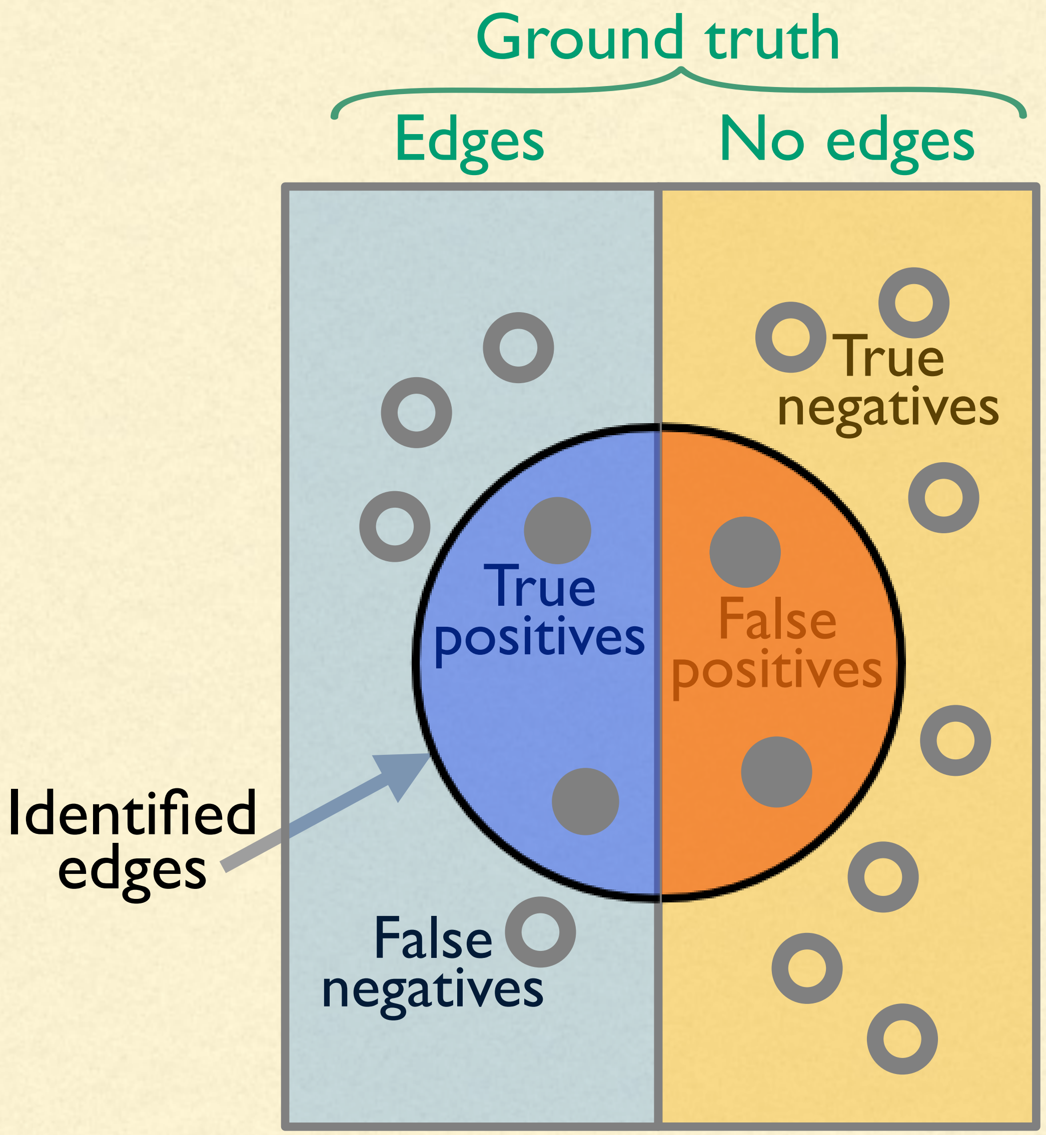
Apply threshold θ

Infer gene-regulatory network

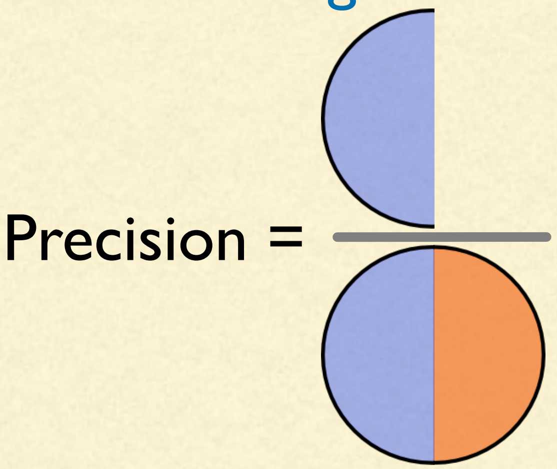


Threshold θ serves as a measure of confidence we have in the identified edge: this allows us to prioritize the predicted gene interactions

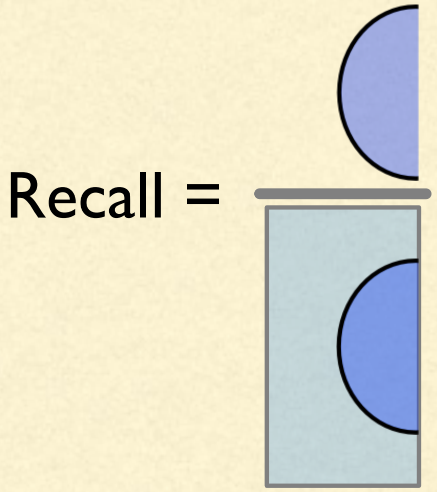
Quantifying success



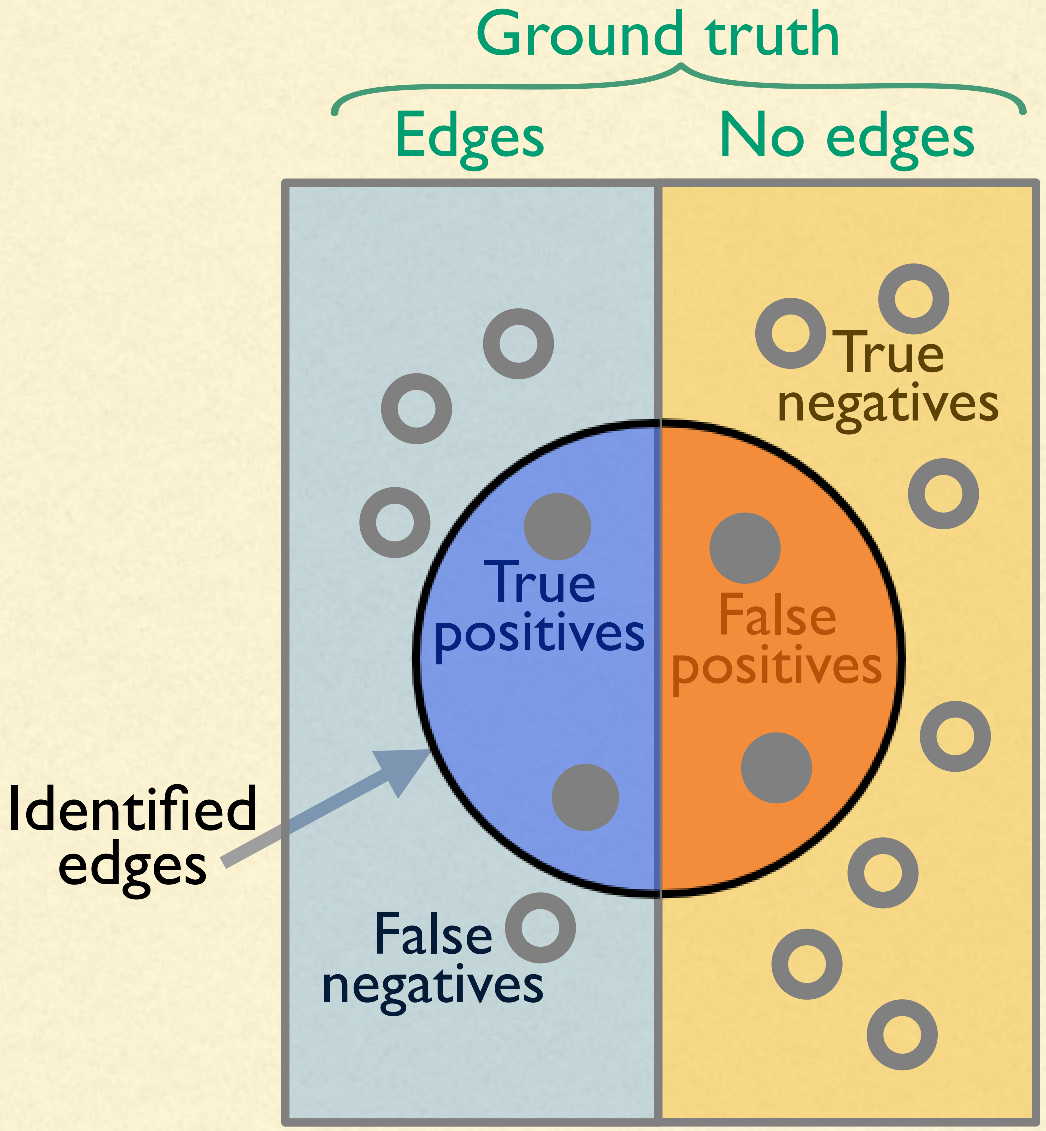
How many identified edges are ground-truth edges?



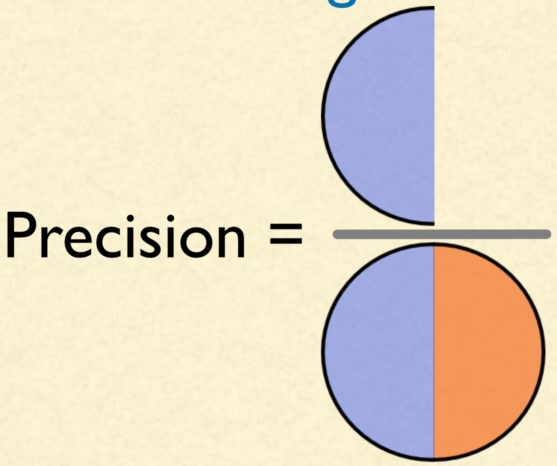
How many ground-truth edges are identified?



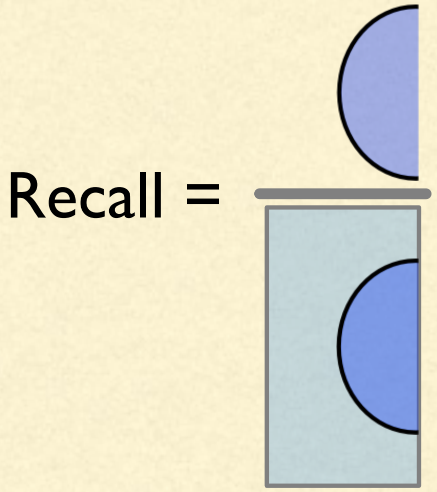
Quantifying success

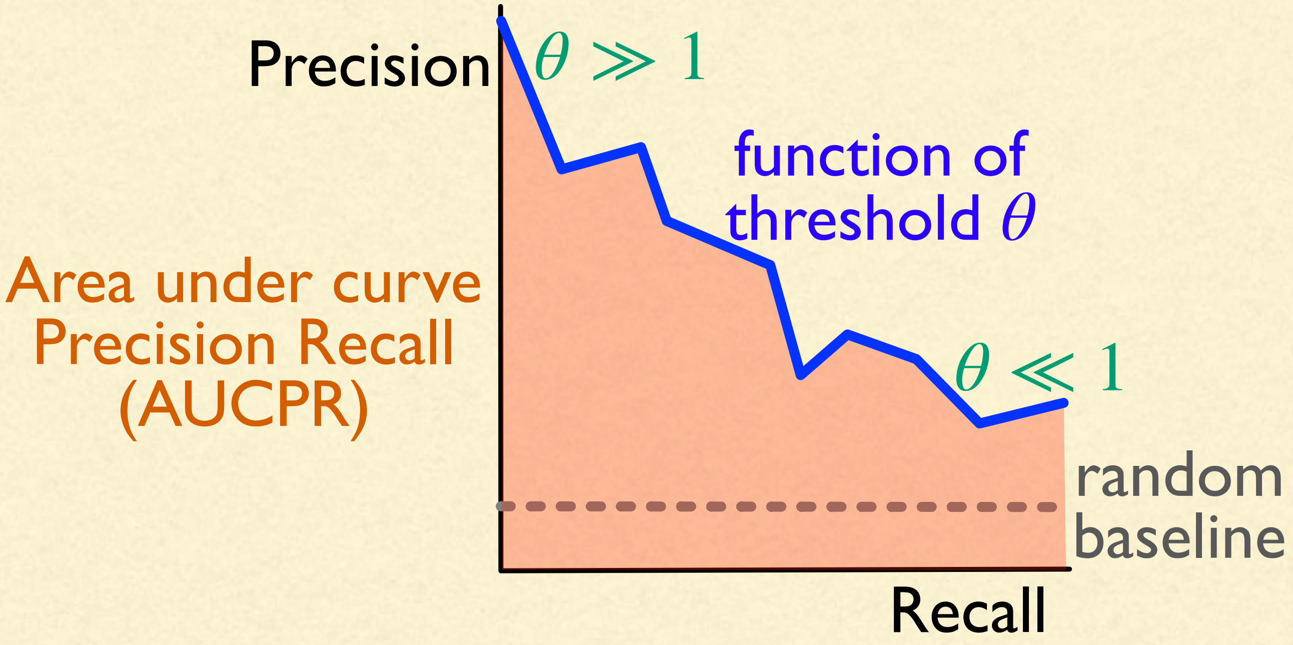


How many identified edges are ground-truth edges?



How many ground-truth edges are identified?





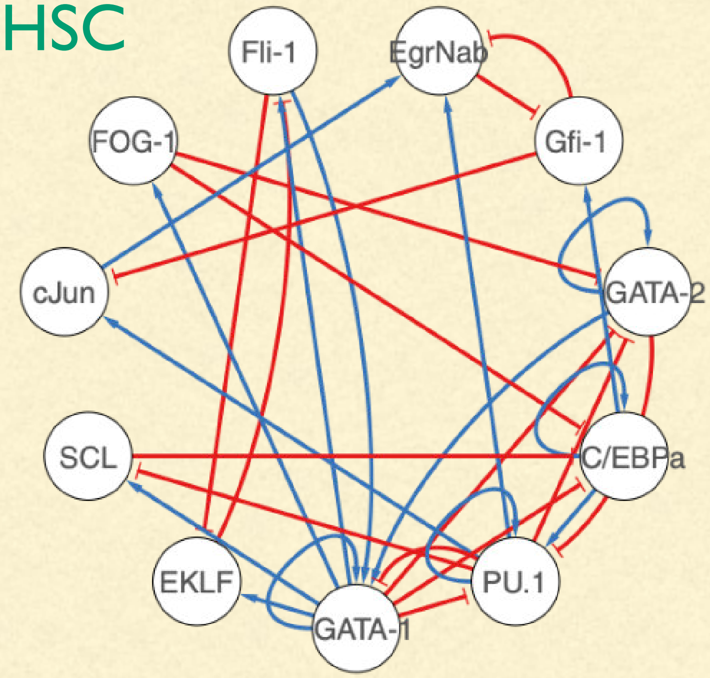
Results: Comparison with other algorithms

Beeline datasets [Pratapa et al.]

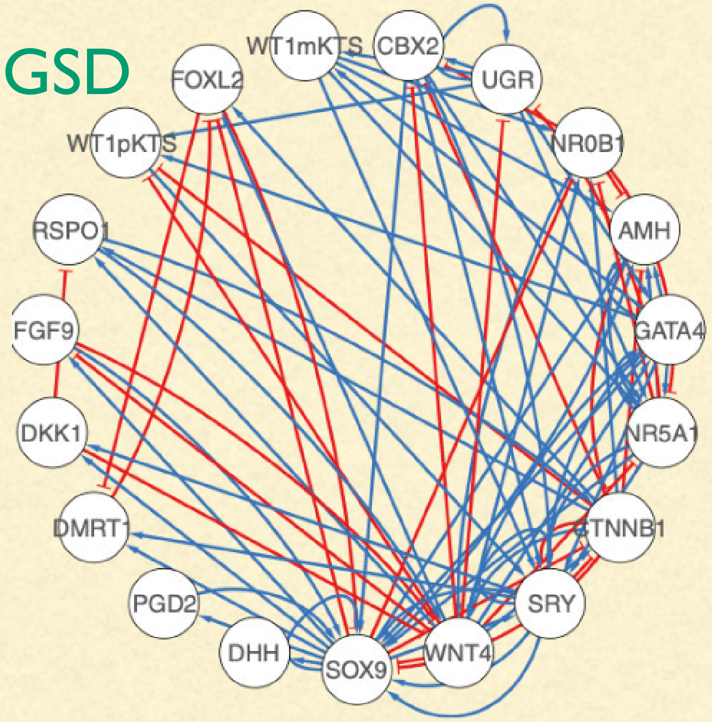
Hematopoietic Stem Cell Differentiation

Gonadal Sex Determination

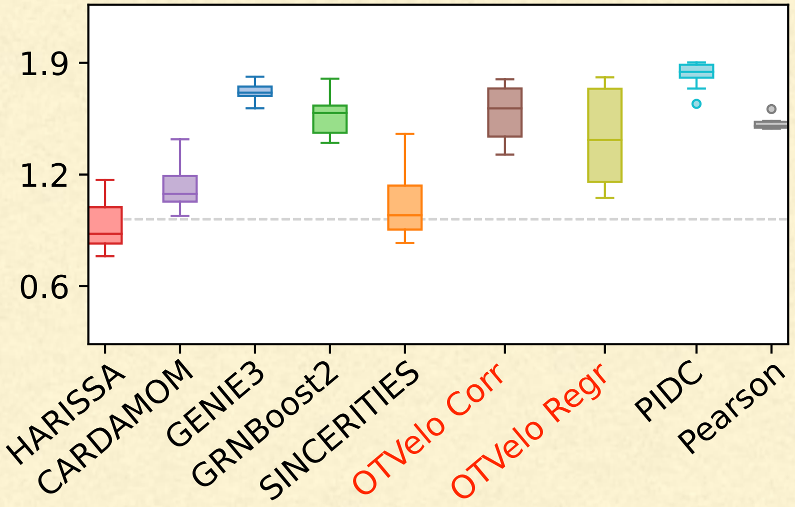
HSC

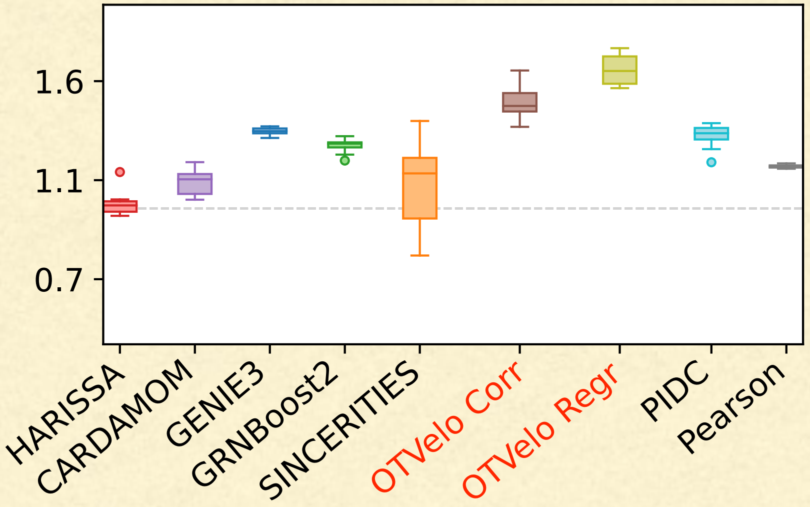


GSD

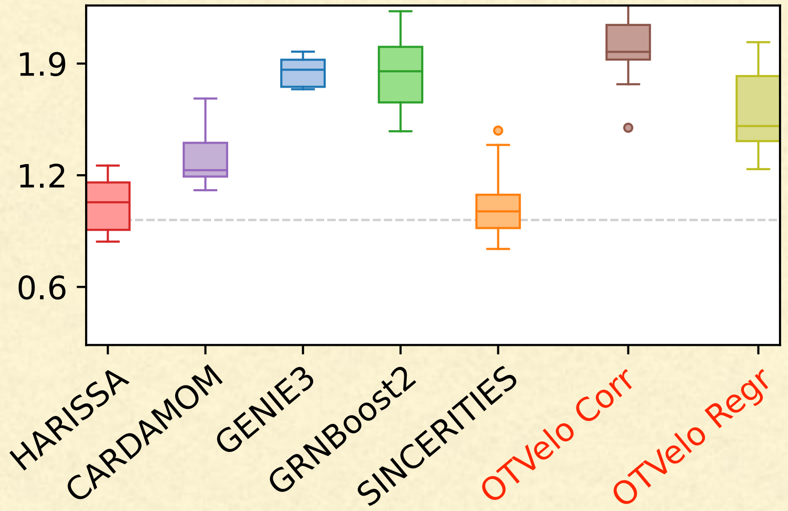


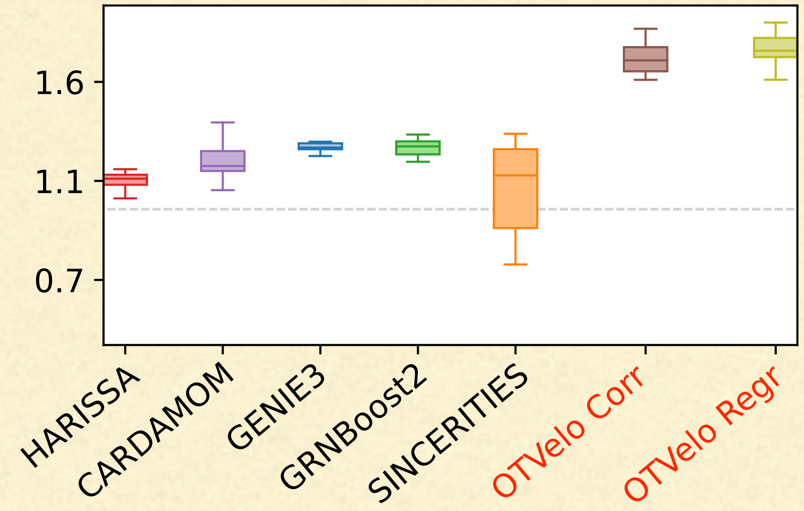
AUCPR ratio (undirected)





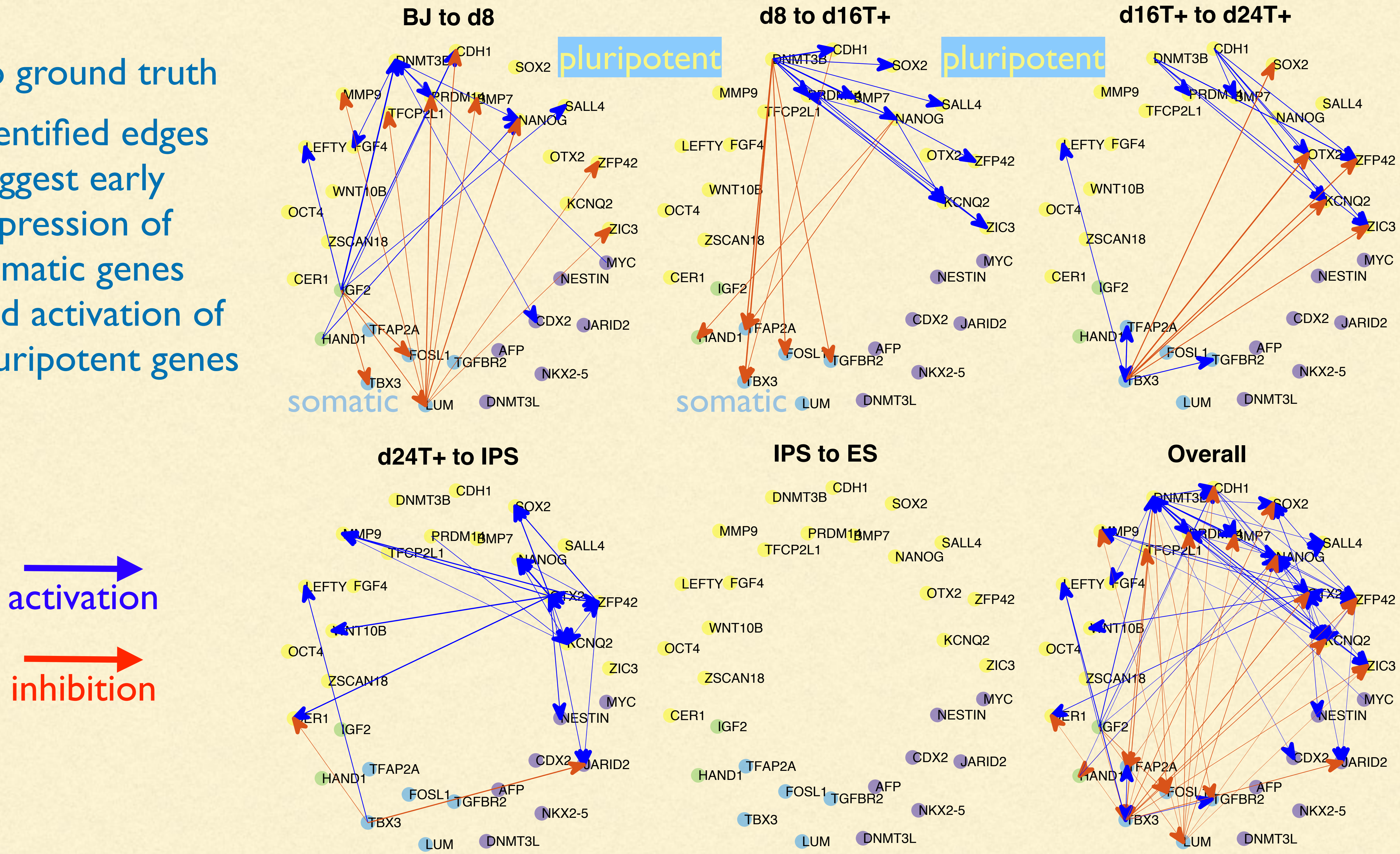
AUCPR ratio (directed)





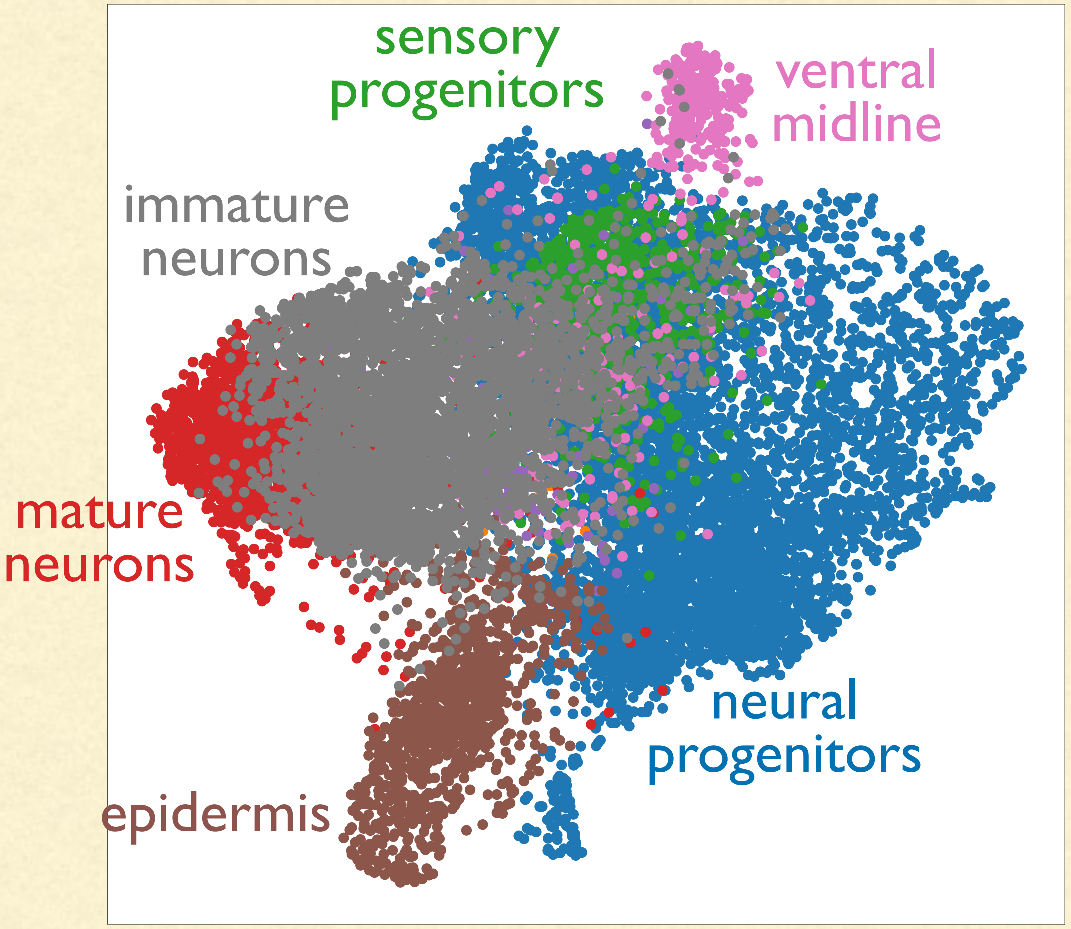
Results: scGEM

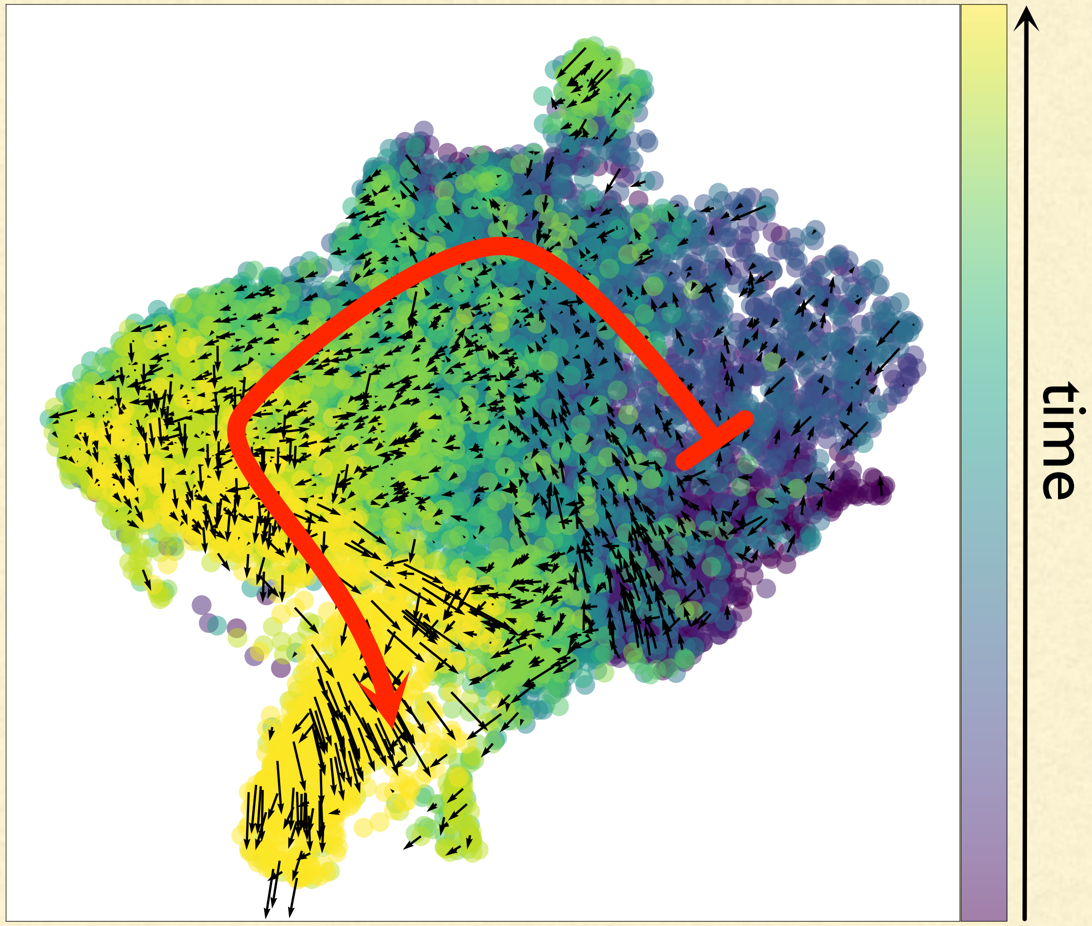
- no ground truth
- identified edges suggest early repression of somatic genes and activation of pluripotent genes



Reprogramming human somatic cells to pluripotent stem cells [Cheow et al. (2016)]

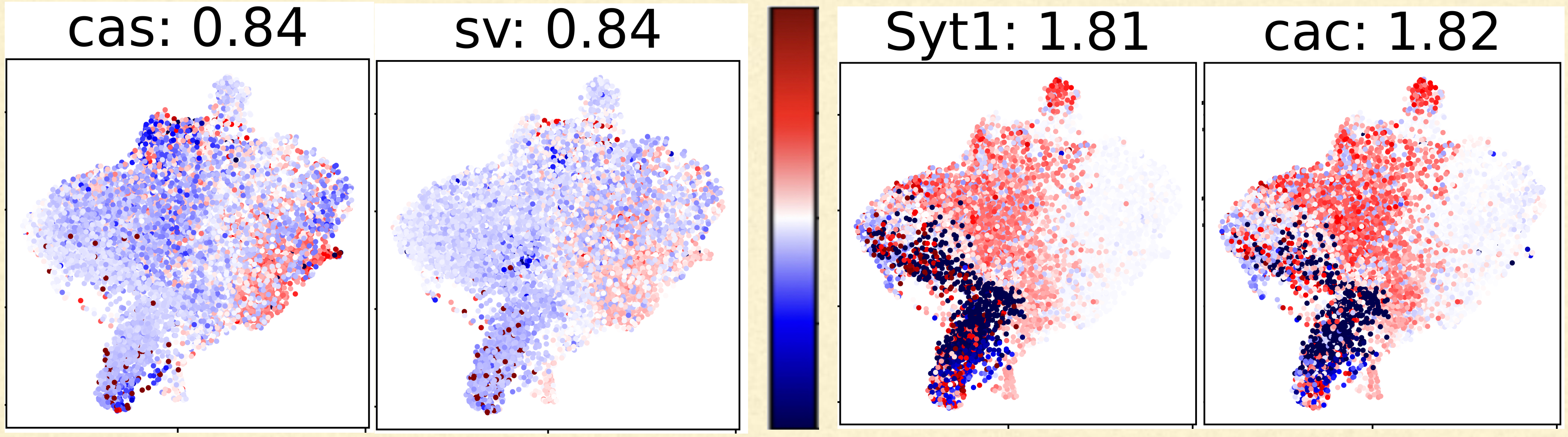
Results: *Drosophila*





In-out degree ratio

$$\deg(g) = \frac{\sum_{\tilde{g} \neq g} C_{\tilde{g},g}}{\sum_{\tilde{g} \neq g} C_{g,\tilde{g}}}$$



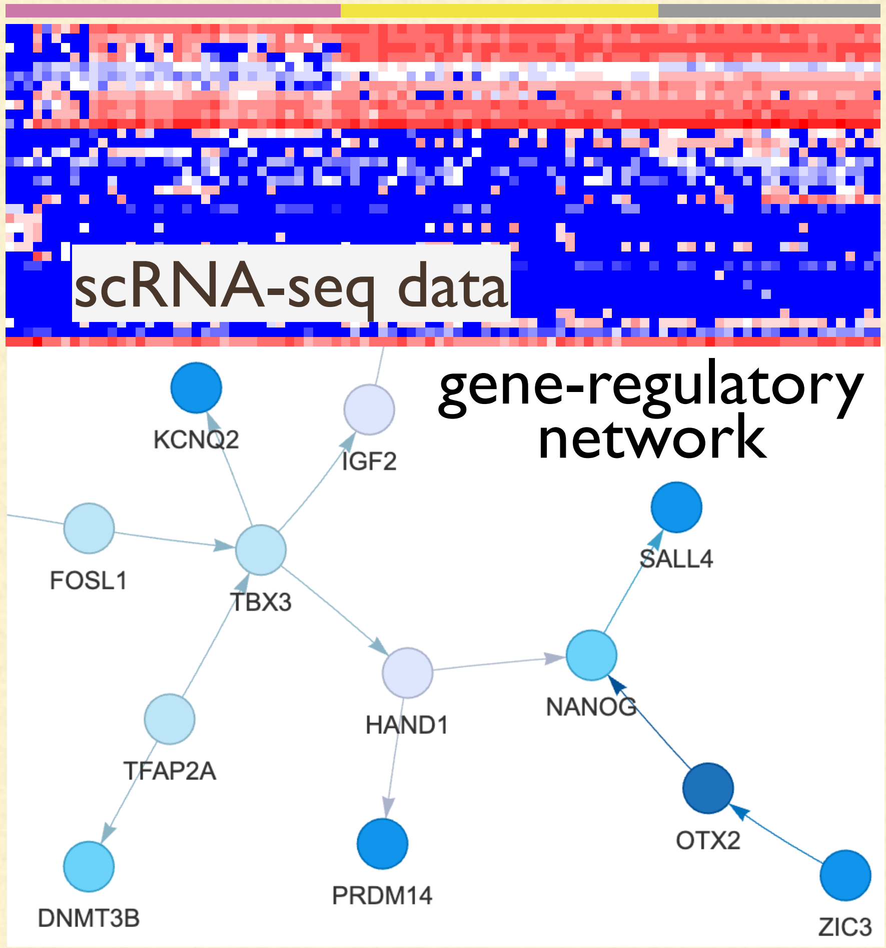
Upstream genes ($\deg(g) < 1$)

Downstream genes ($\deg(g) > 1$)

Drosophila embryonic development: neuroectoderm [Calderon et al. 2022]

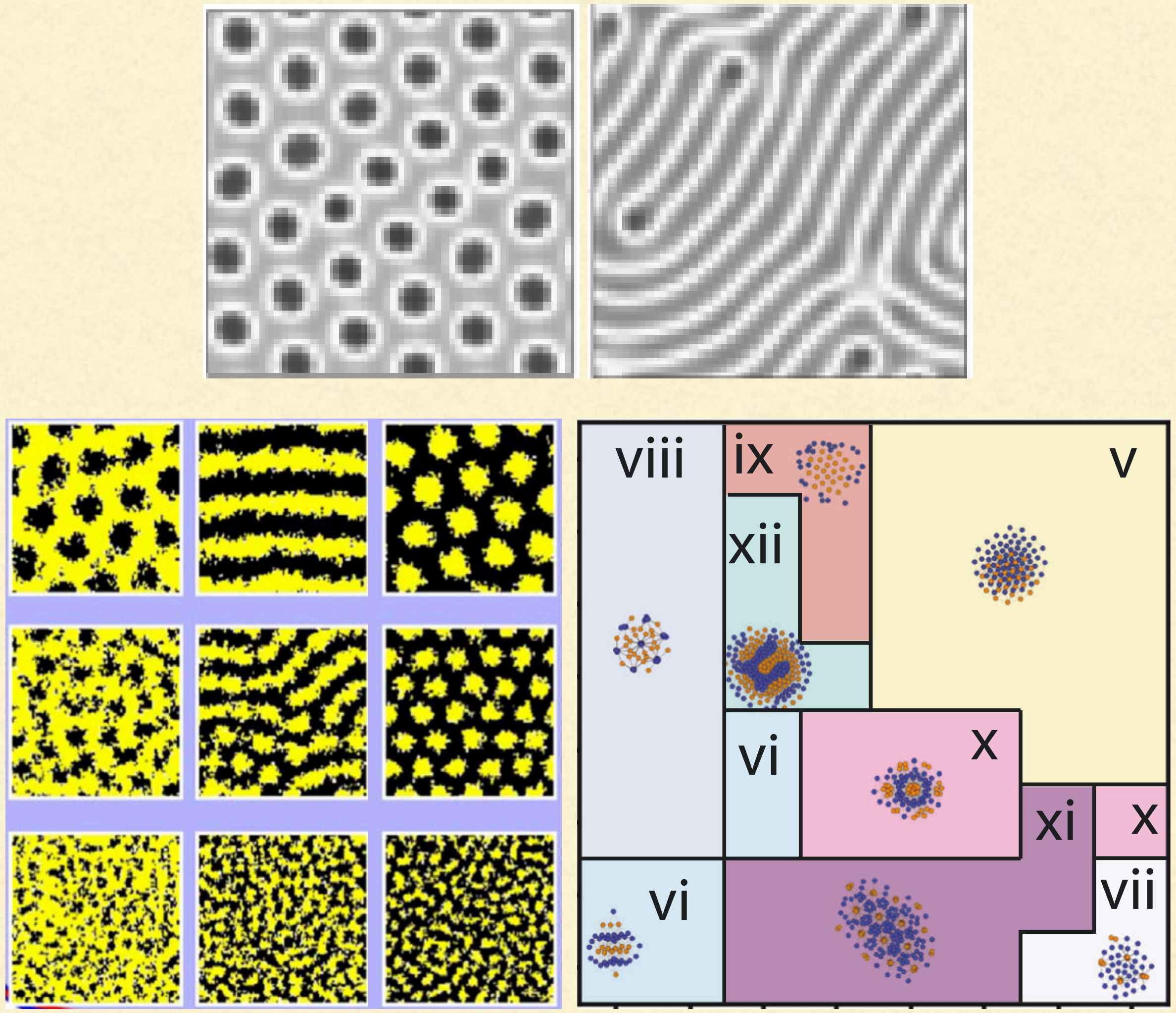
Two themes connected by optimal transport as the tool

Inferring gene-regulatory networks



[Zhao, Larschan, S., Singh]

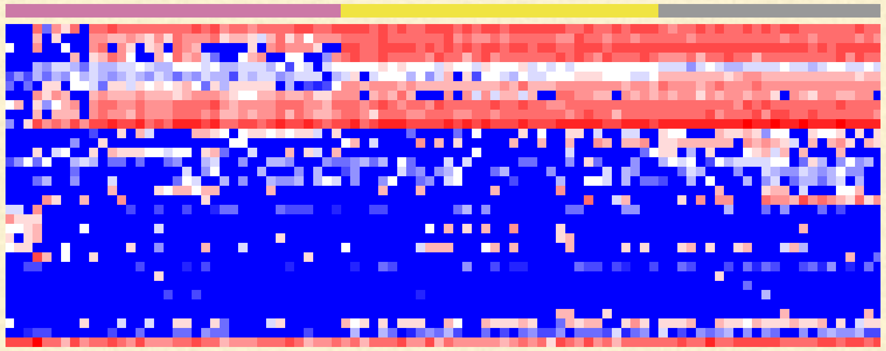
Quantifying patterns and tracing bifurcation curves



[Zhao, Maffa, S.]

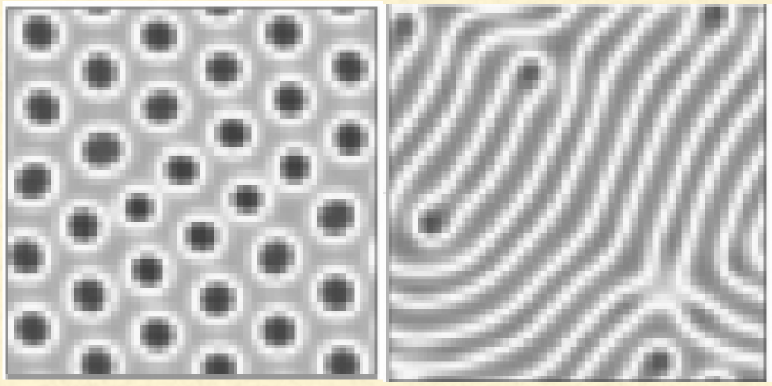
Next steps …

Inferring gene-regulatory networks



- Further validation:
  - ground-truth networks
  - identifiability of mathematical models
- Include other measurements: RNA velocity and Chromatin accessibility
- Identification of gene pathways conserved across fly & mice and neuron formation & learning/memory (with O'Connor–Giles, Fleischmann, Kaun, Larschan, Singh)

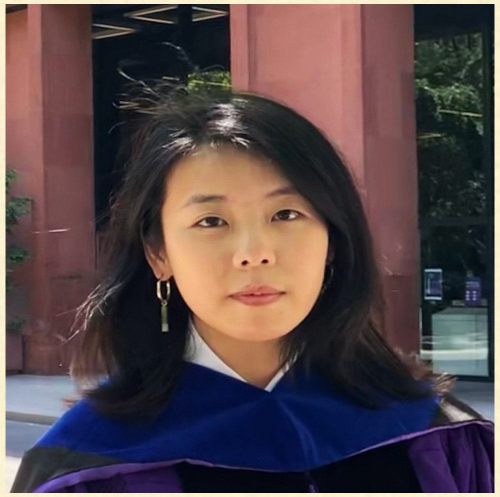
Quantifying patterns and tracing bifurcation curves



- Applications to:
  - stochastic agent-based models
  - contact and source defects
- Systematic convergence analysis and dependence on feature functions

Thanks to my fantastic collaborators!

Sam Maffa



Wenjun Zhao

Erica Larschan

Ritambhara Singh

And thank you for listening!