Applied Mathematics

Two Inference Problems in Dynamical Systems from Mathematical and Computational Biology

Wenjun Zhao
PIMS Postdoc Seminar
03.26.2025

Two inference problems

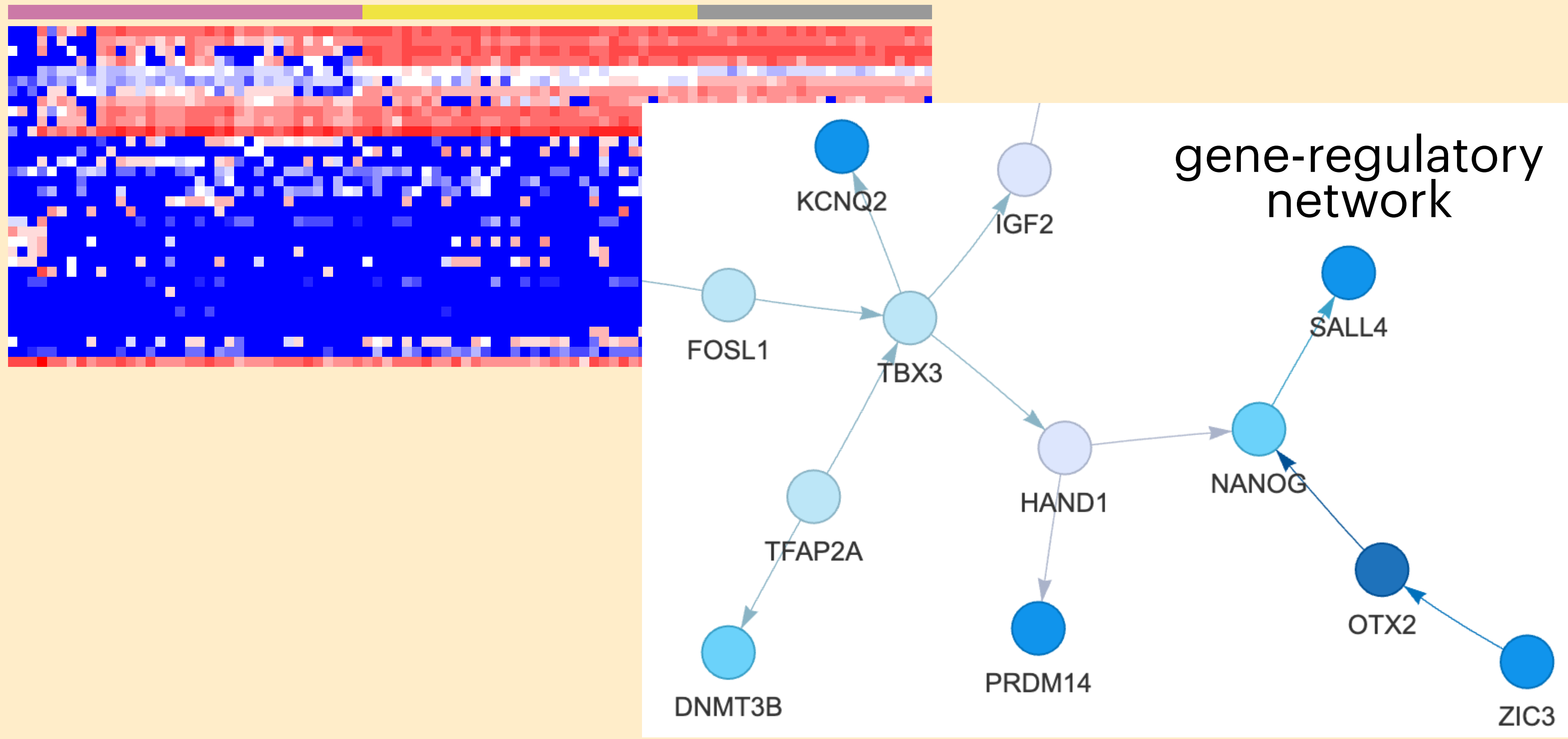
Inferring mechanisms from single-cell sequencing data

Joint work with

Bjorn Sandstede

Ritambhara Singh

Erica Larschan

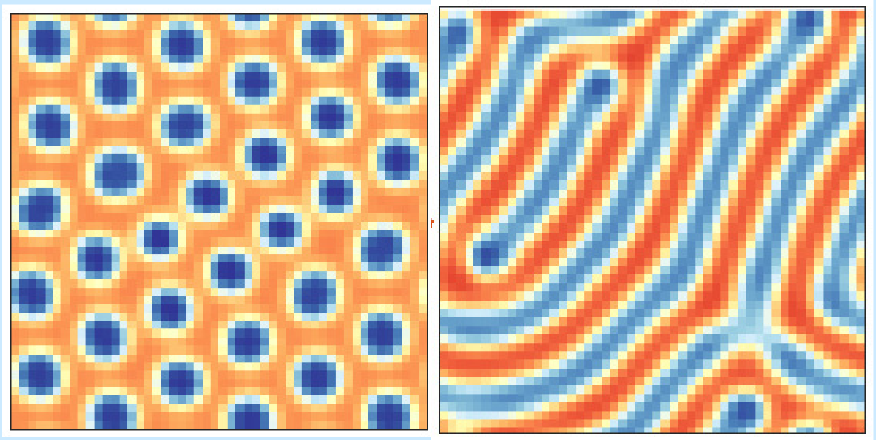


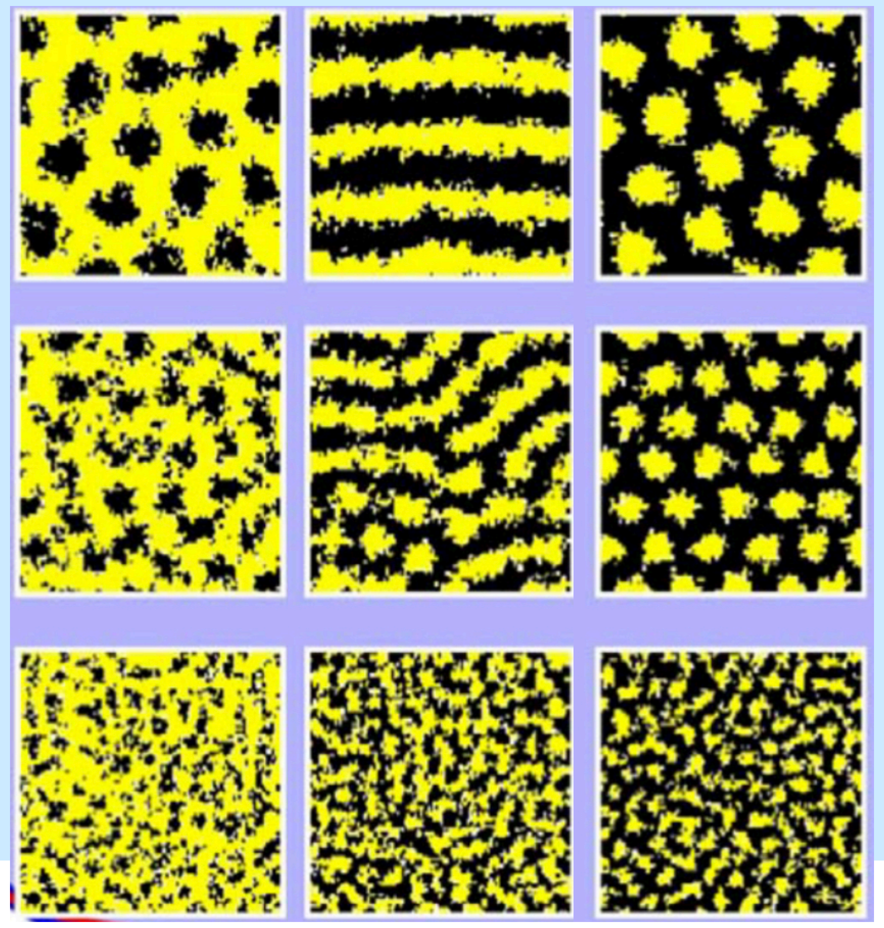
Quantifying Turing patterns and their bifurcation

Joint work with

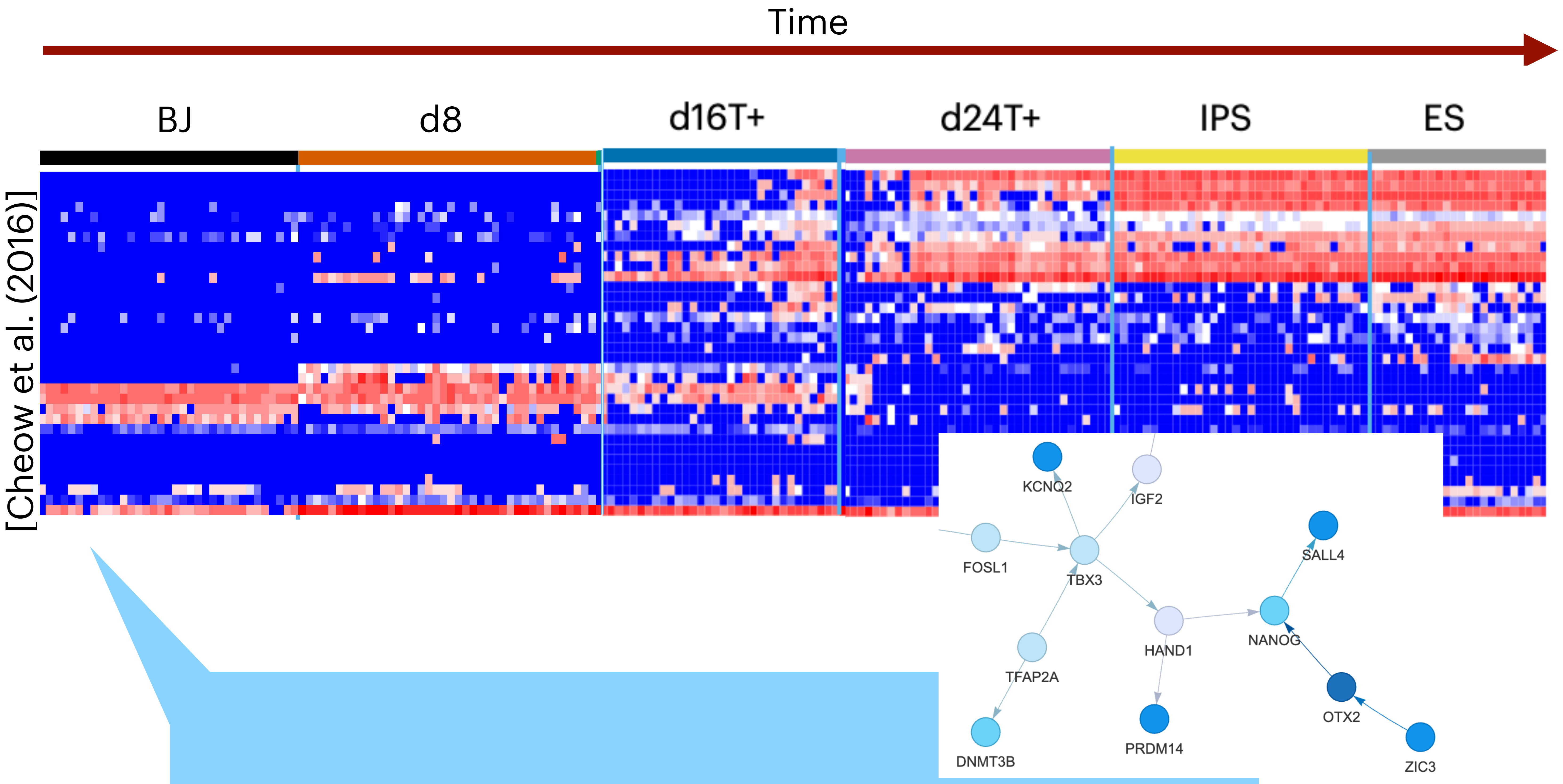
Bjorn Sandstede

Sam Maffa



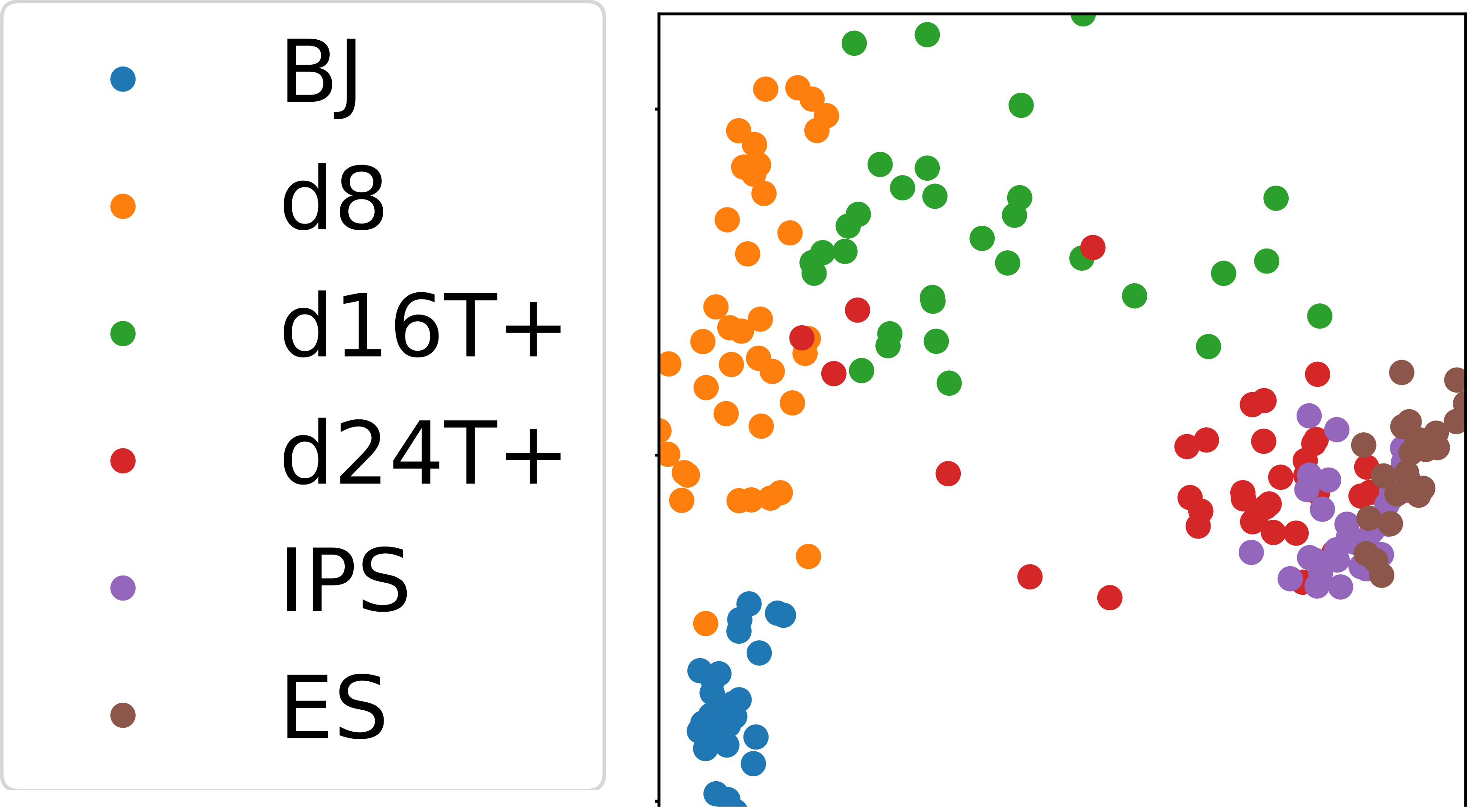


Problem statement



[Cheow et al. (2016)]

Which gene (row) activated/ inhibited which gene (row)?



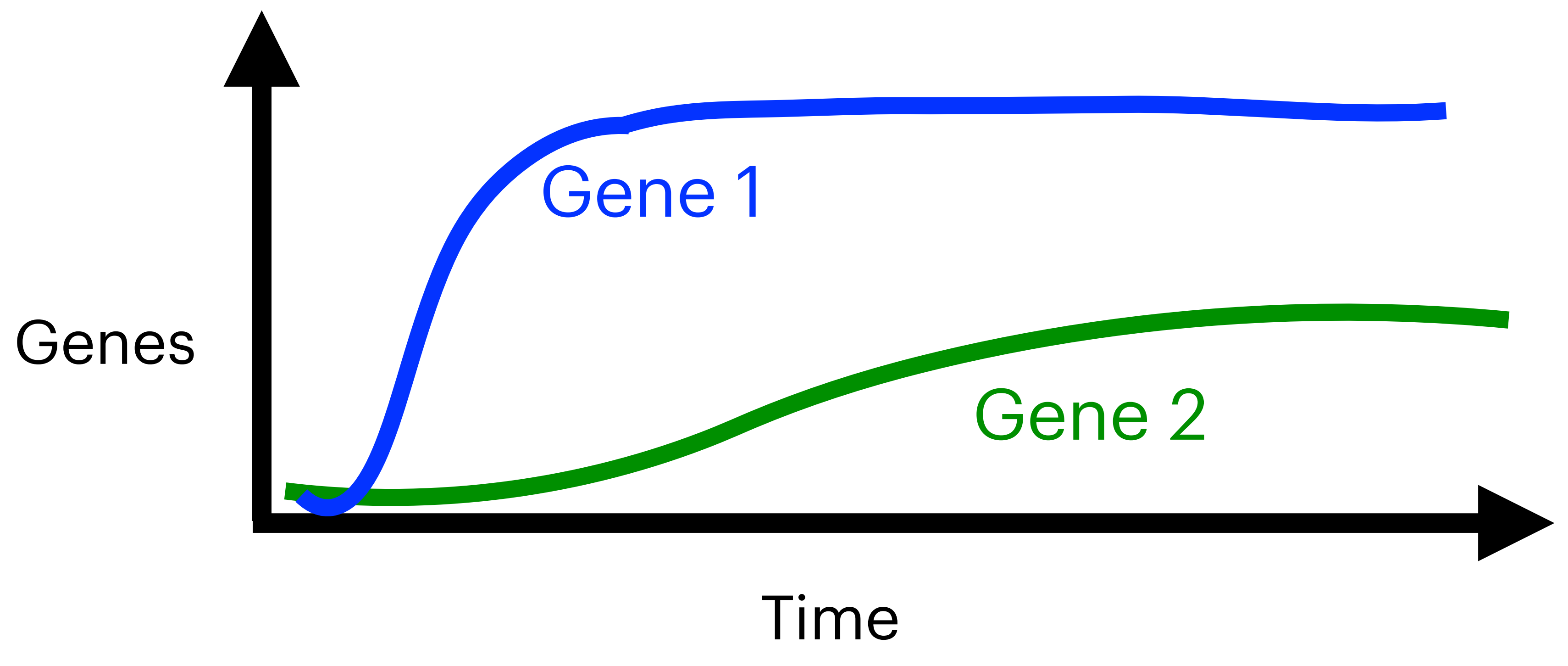
Cell types

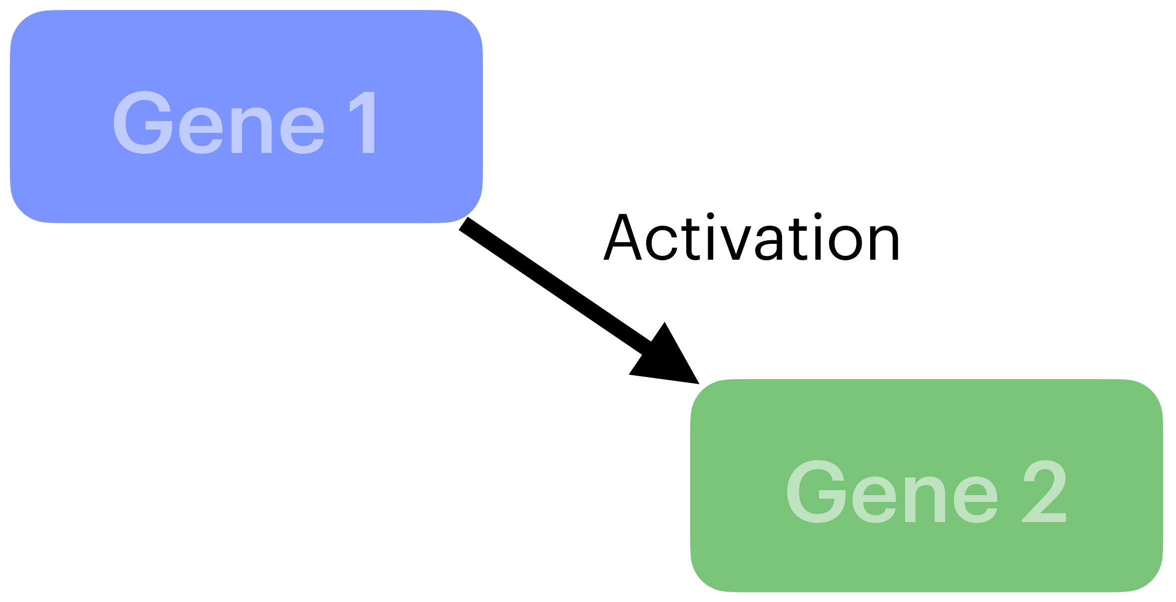
First 2 PCs

Causation between time series

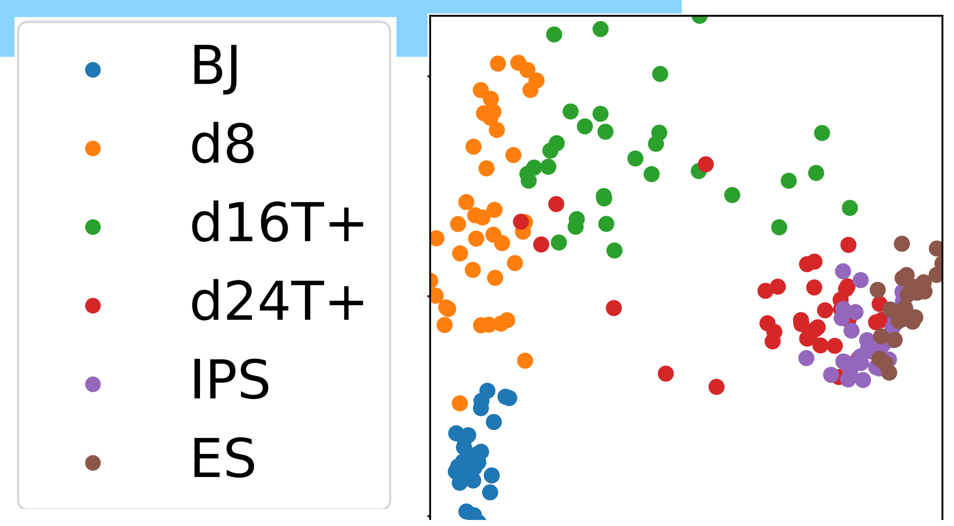
Granger causality:

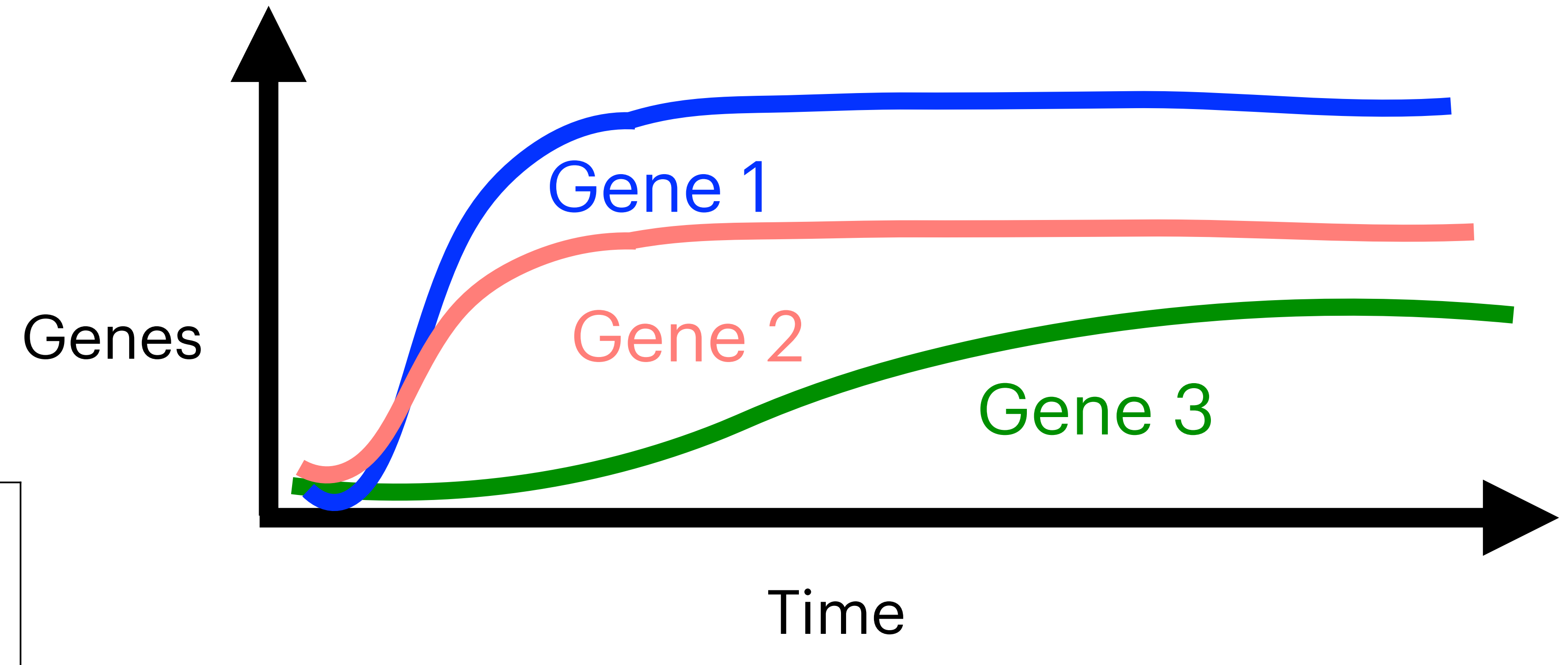
- (1) The cause happens prior to its effect
- (2) The cause has unique information about the future values of its effect

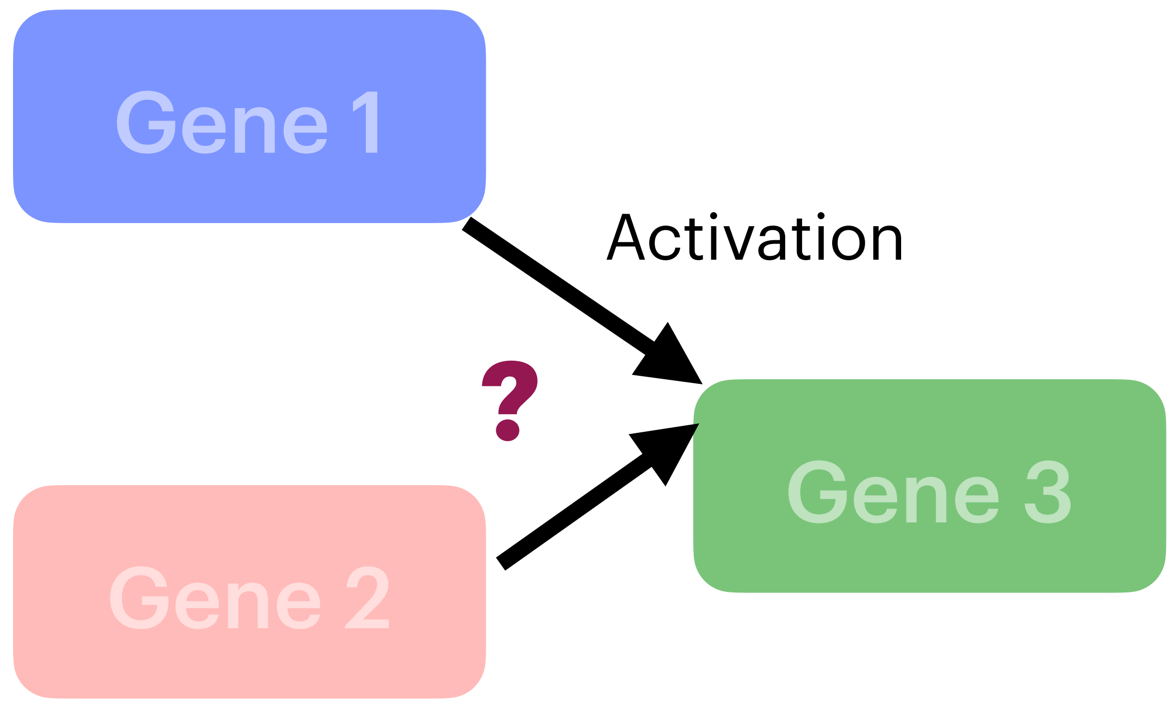




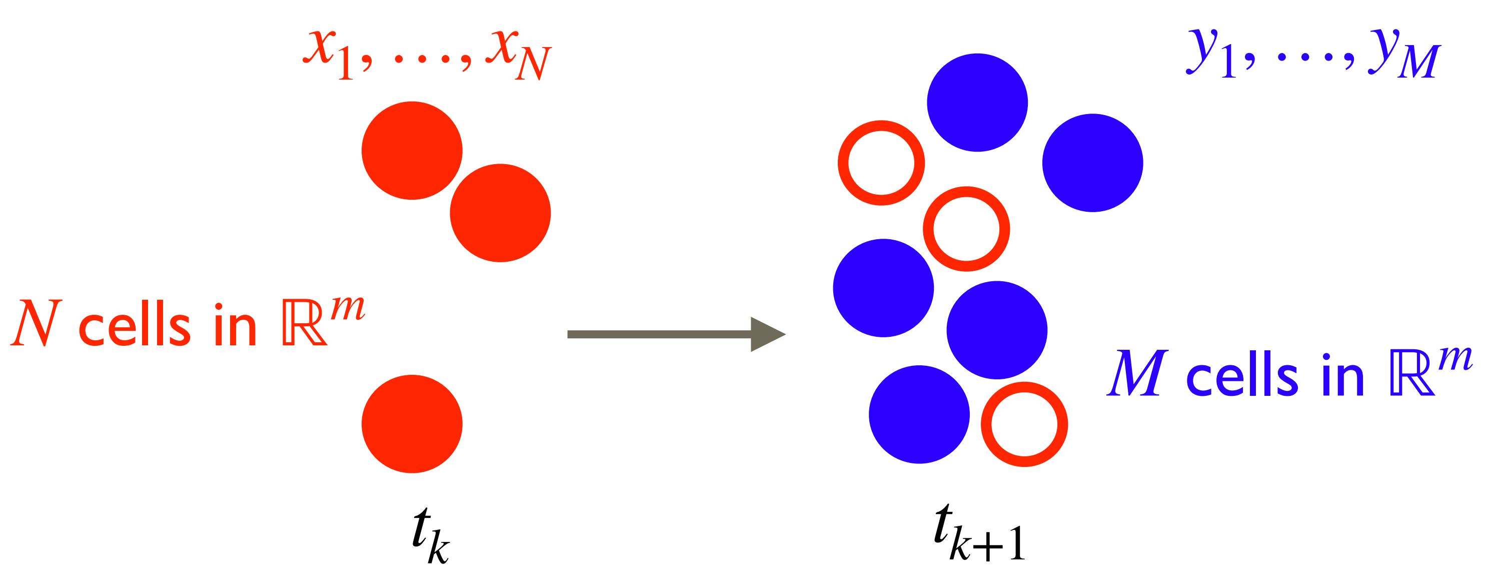
Problem:
No actual time series:
Cells are destroyed during measurement







Trajectory inference through optimal transport



Find joint probability distribution Γ

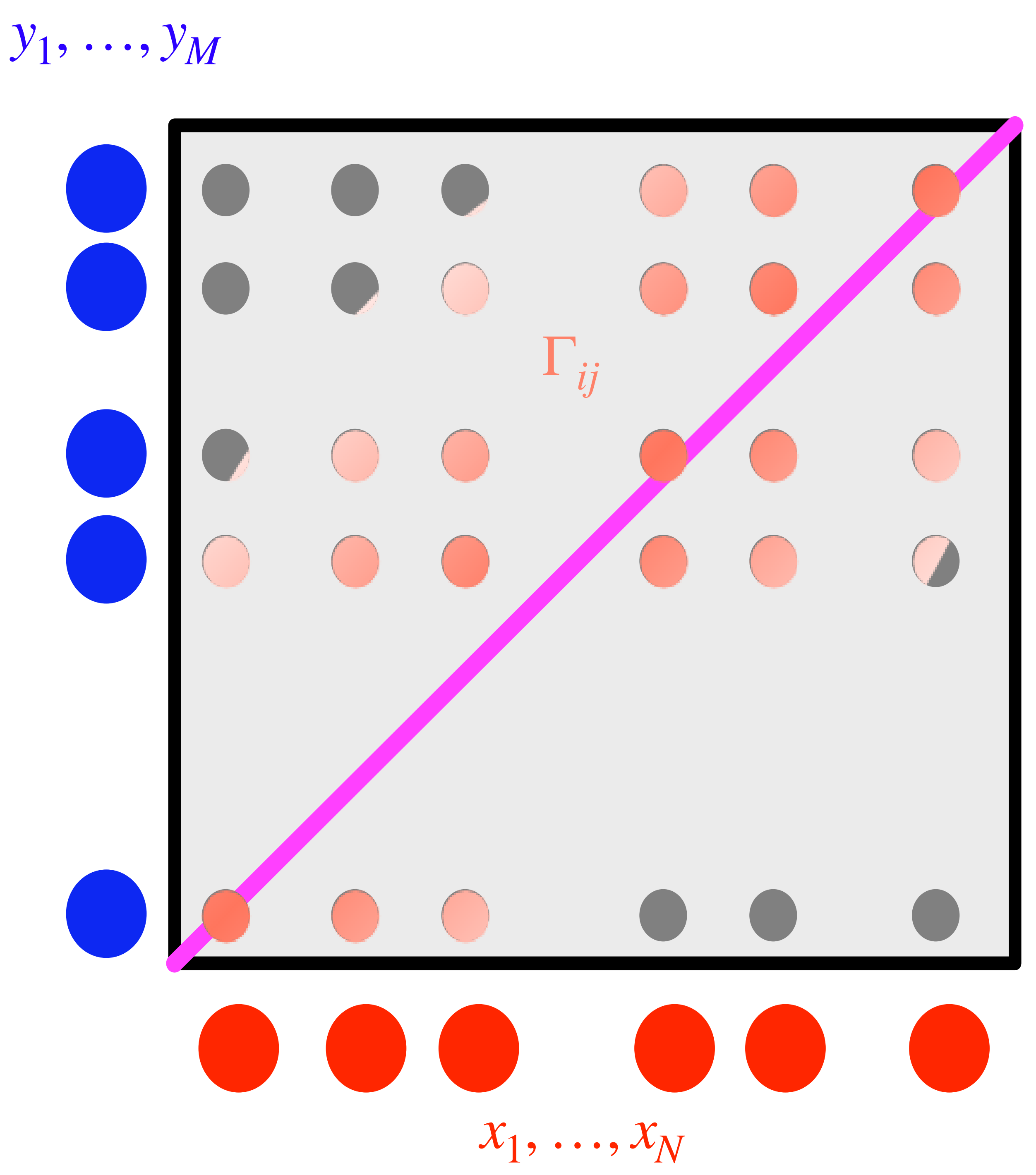
($\Gamma_{ij} \geq 0,\ \sum_{i=1}^{N} \Gamma_{ij} = \frac{1}{M},\ \sum_{j=1}^{M} \Gamma_{ij} = \frac{1}{N}$) as solution to

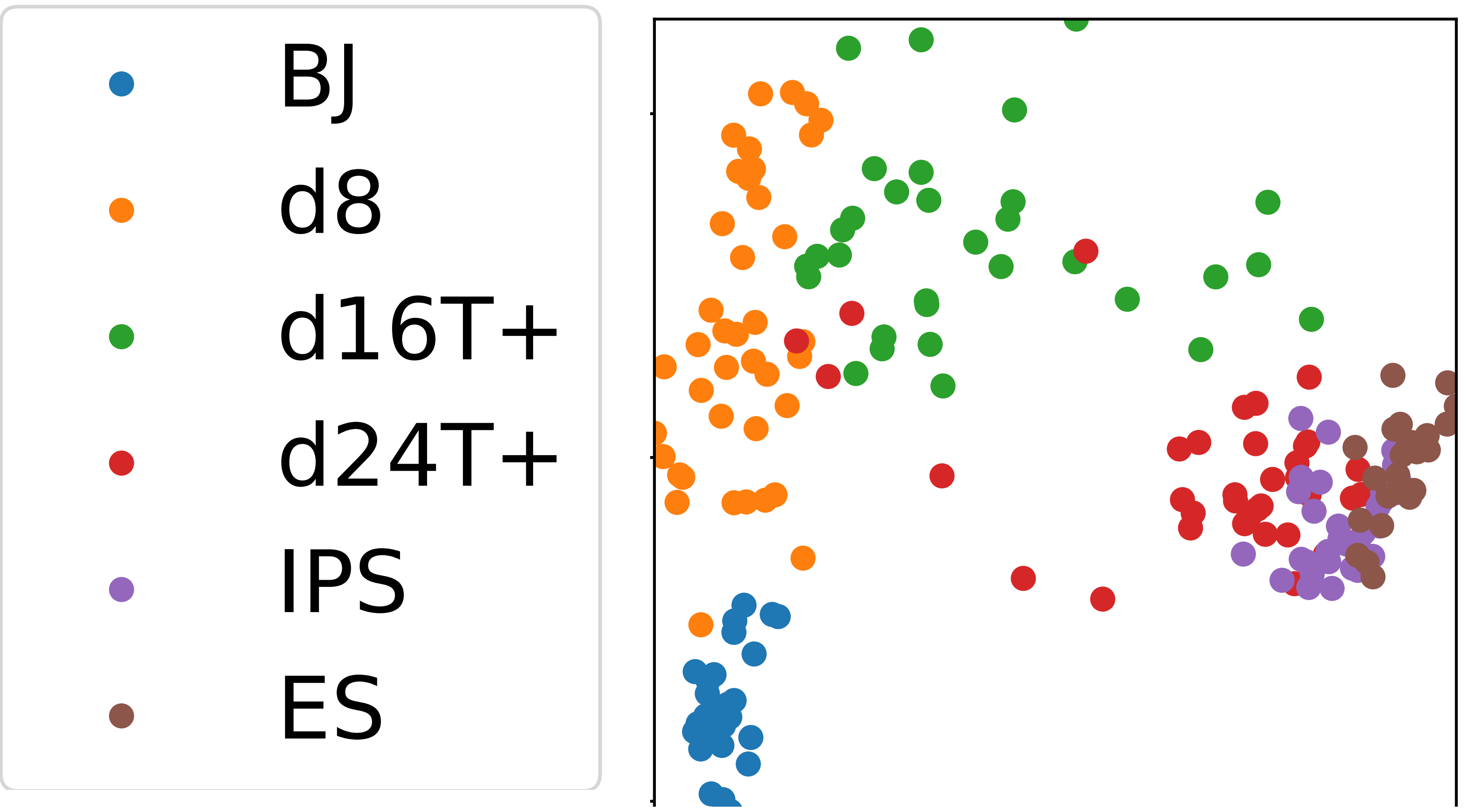
$$\Gamma = \arg\min_{\Gamma} \sum_{\substack{i=1,\ldots,N \\ j=1\ldots,M}} |x_i - y_j|^2_{\mathbb{R}^m}\, \Gamma_{ij}$$

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

Barycentric projection:

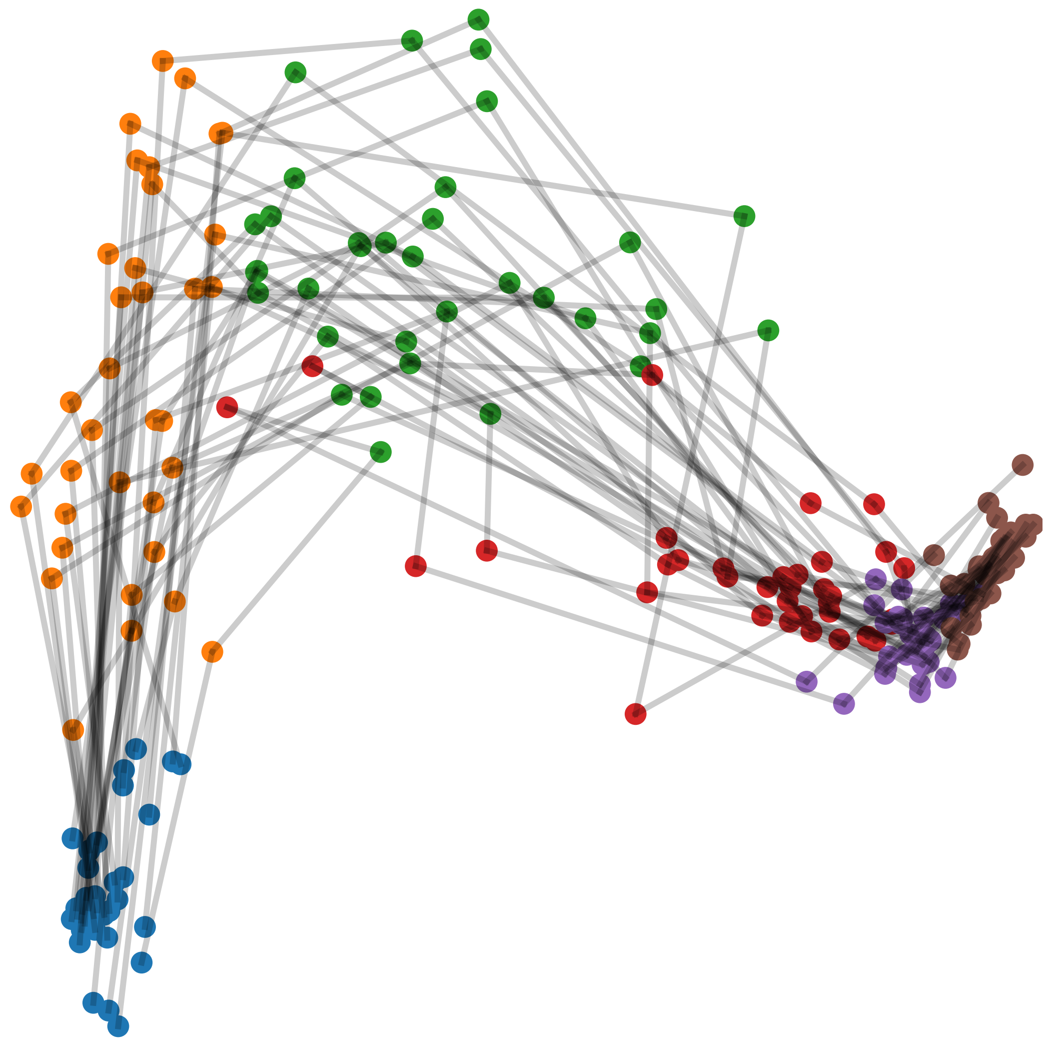
$$x_{i,t-1} \to \sum_j \frac{\Gamma_{i,j}}{\sum_k \Gamma_{i,k}} x_{j,t}$$





Cell types

First 2 PCs



Recovered cell trajectories

OT velocity

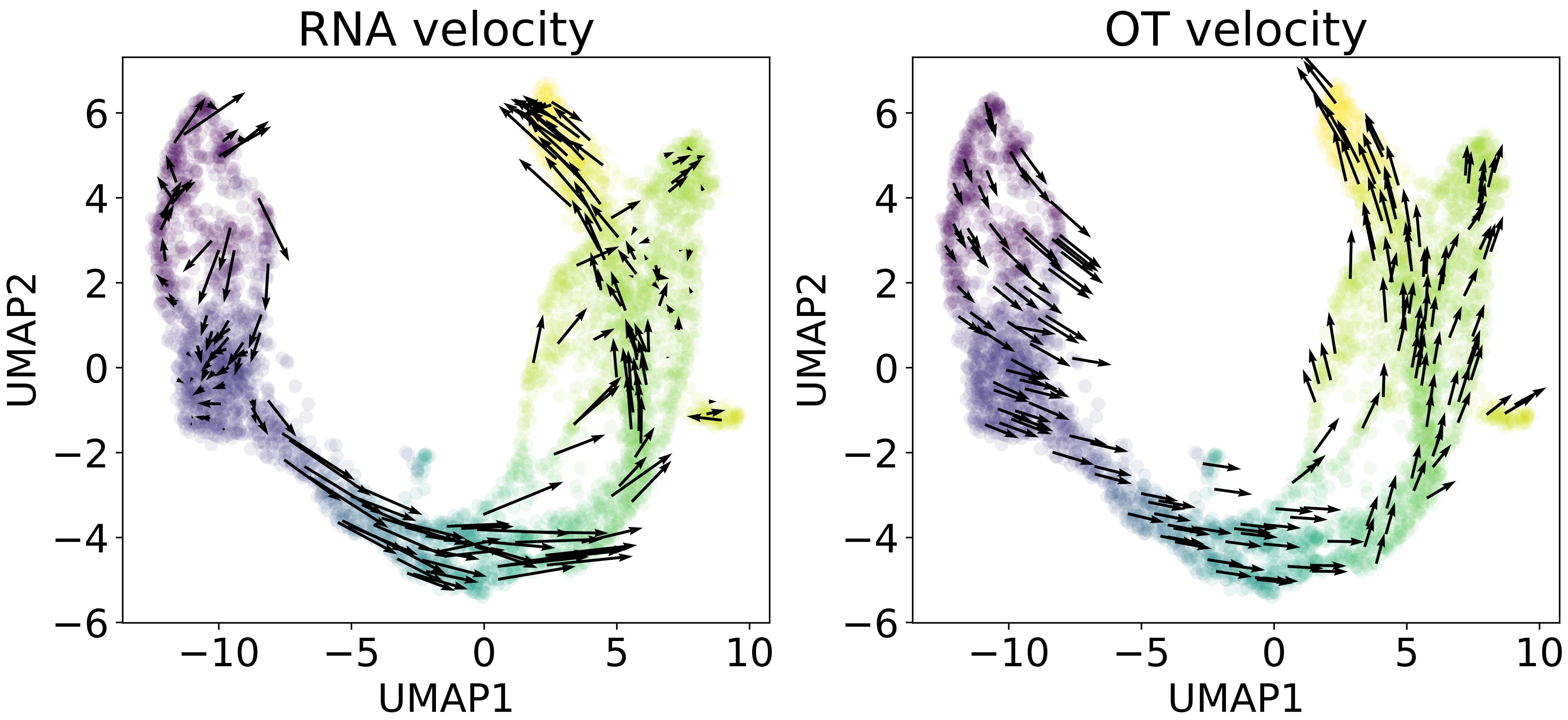
- Solve the trajectory problem via OT as mentioned

For cell $x_{i,t} \in \mathbb{R}^m$, infer its past $\hat{x}_{i,t-1}$ and future $\hat{x}_{i,t+1}$ via barycentric projection

- Estimate the ‘OT velocity’ for each cell via finite difference

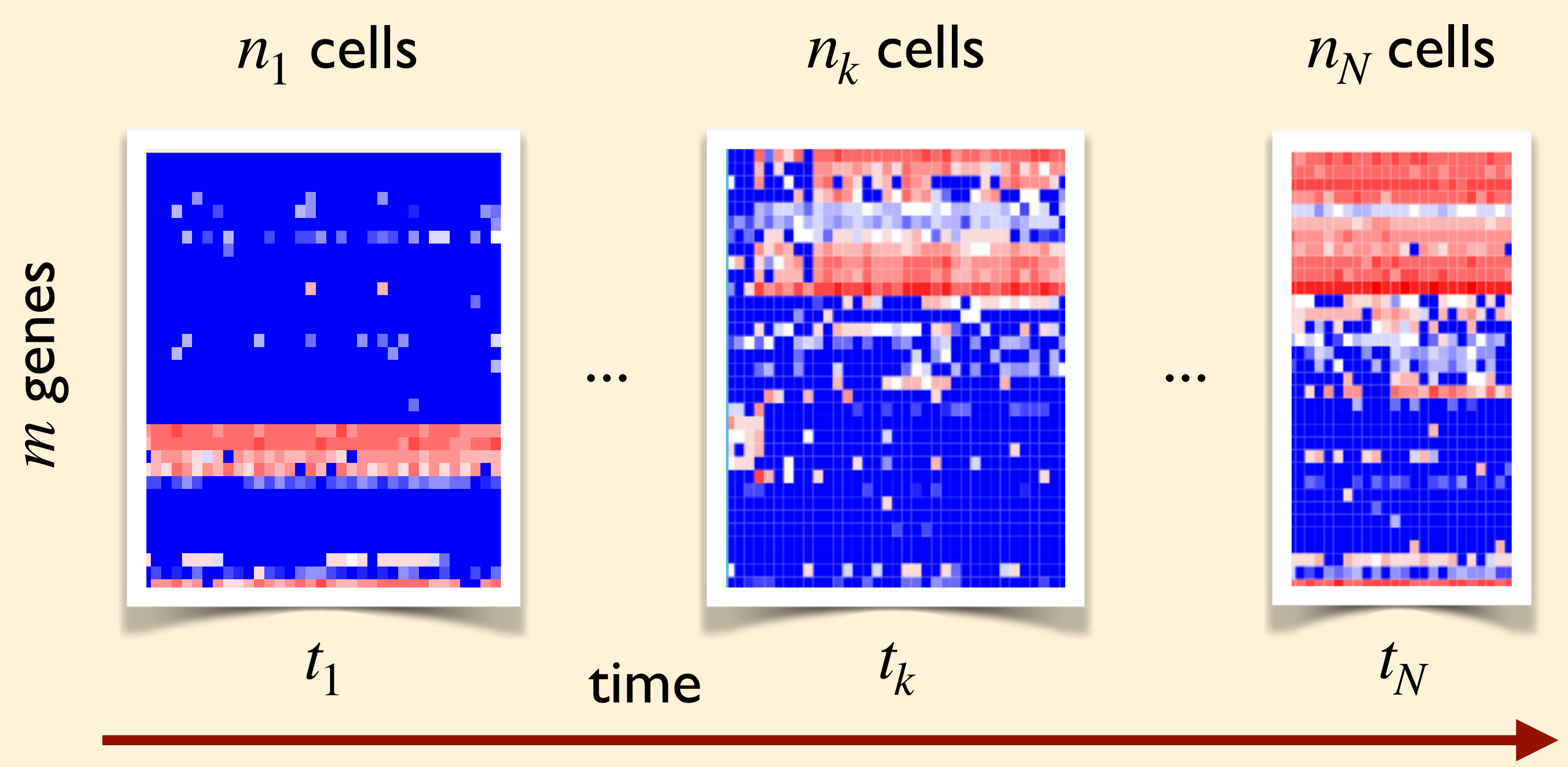
Velocity for gene g in each cell: $v_{i,t}^{g} = \dfrac{dx_{i,t}^{g}}{dt} \approx \dfrac{\hat{x}_{i,t+1}^{g} - \hat{x}_{i,t-1}^{g}}{2}$

Validation:
(1) OT velocity is similar to RNA velocity (a method using reaction model for unspliced/spliced RNA counts);
(2) RNA velocity is only available in deeply sequenced datasets.

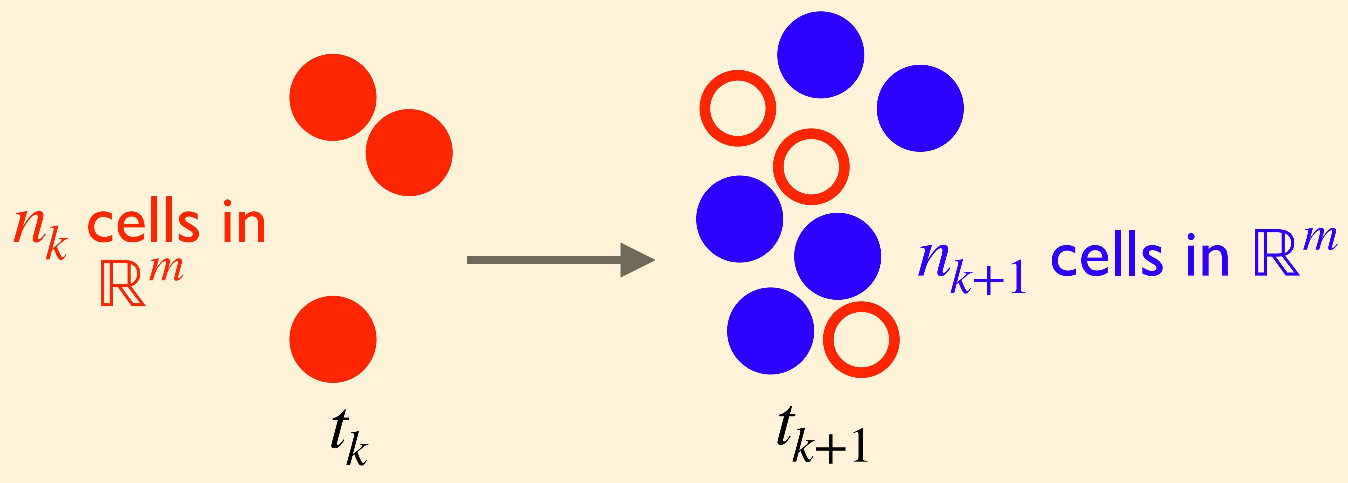


Pipeline...

Time-stamped single-cell RNA data

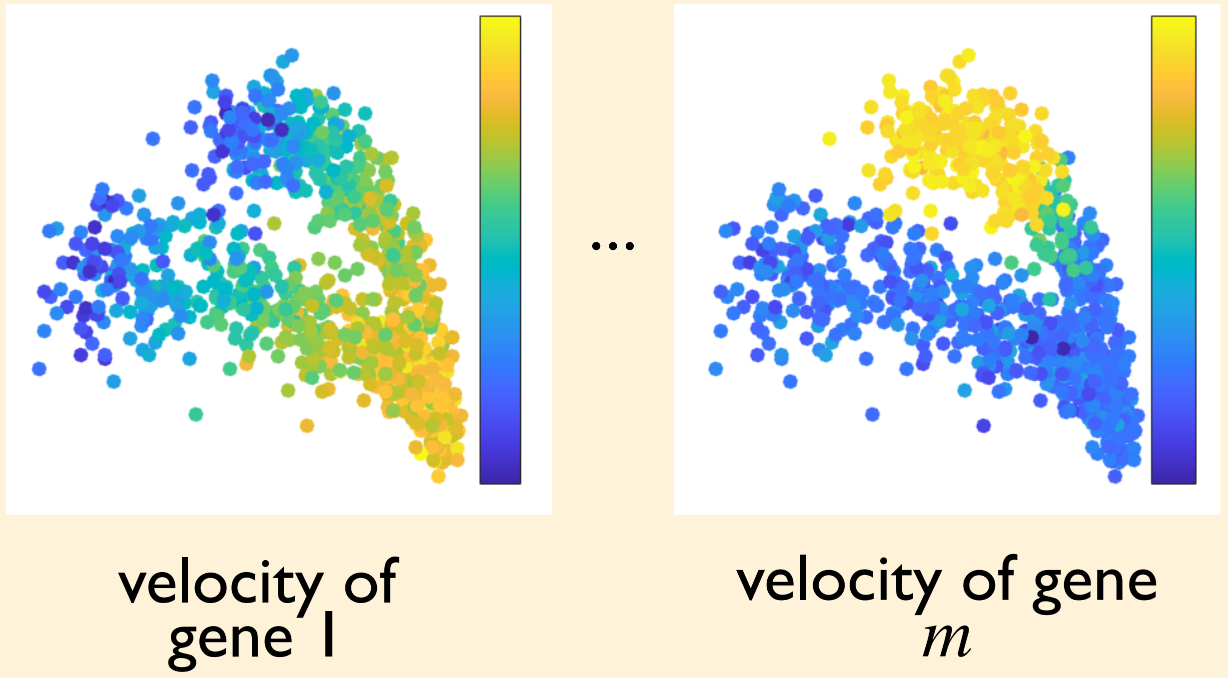


Optimal transport



predict cell trajectories

Finite differences on cell trajectories



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

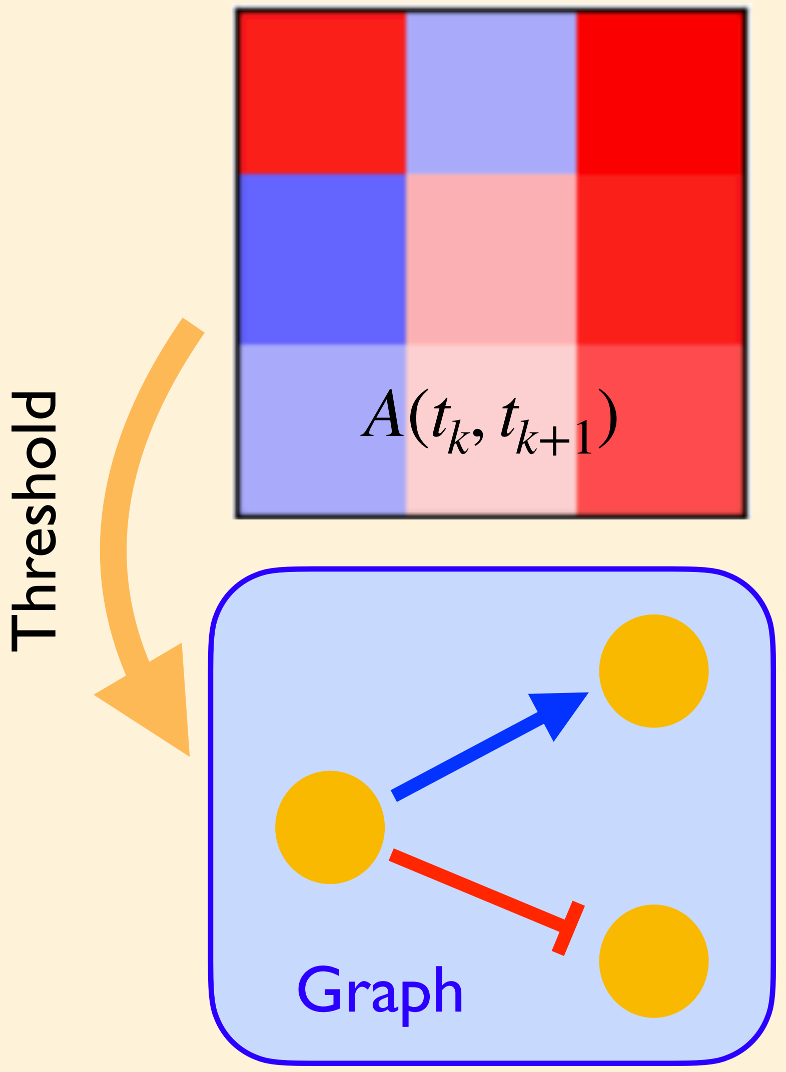
Granger causality with linear reduction

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

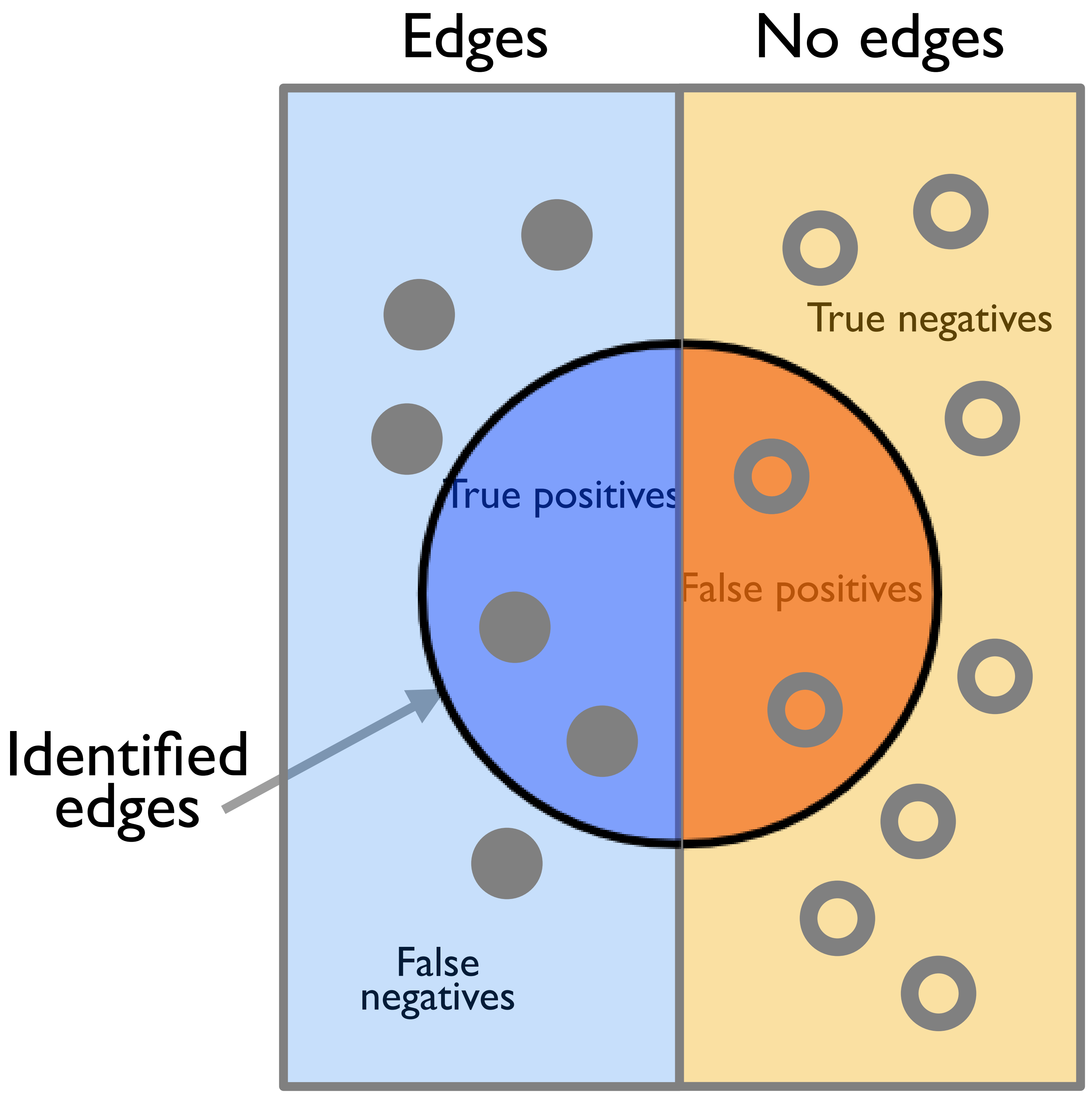
$$A(t_k,t_{k+1}) \in \mathbb{R}^{m \times m}$$

estimate gene-interaction matrix

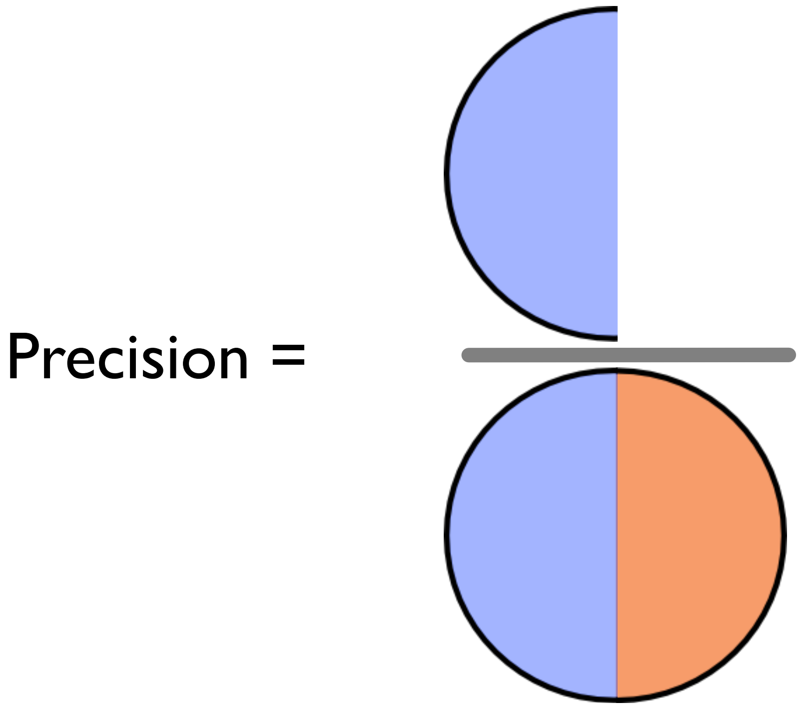
Gene regulatory network



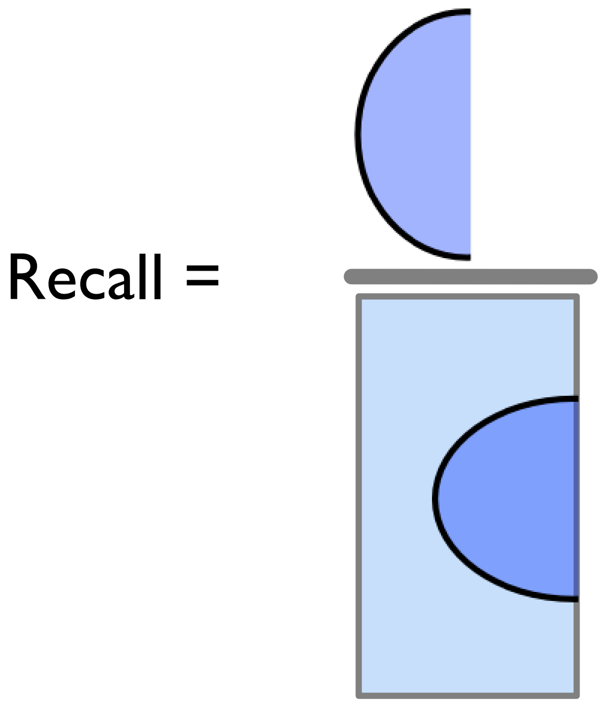
Performance quantification

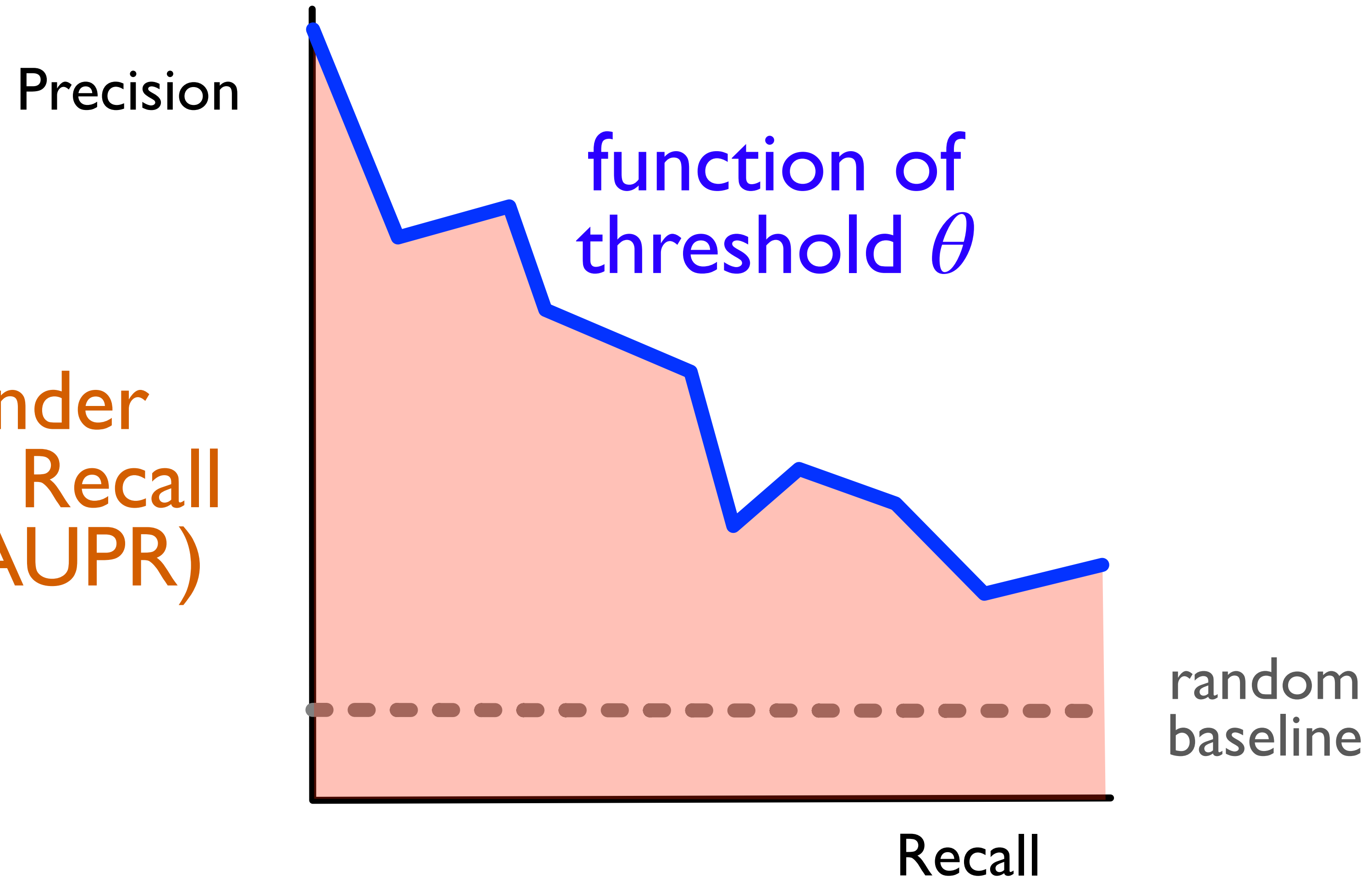


How many identified edges are ground-truth edges?



How many ground-truth edges are identified?





Area under Precision Recall Curve (AUPR)

Features for AUPR:

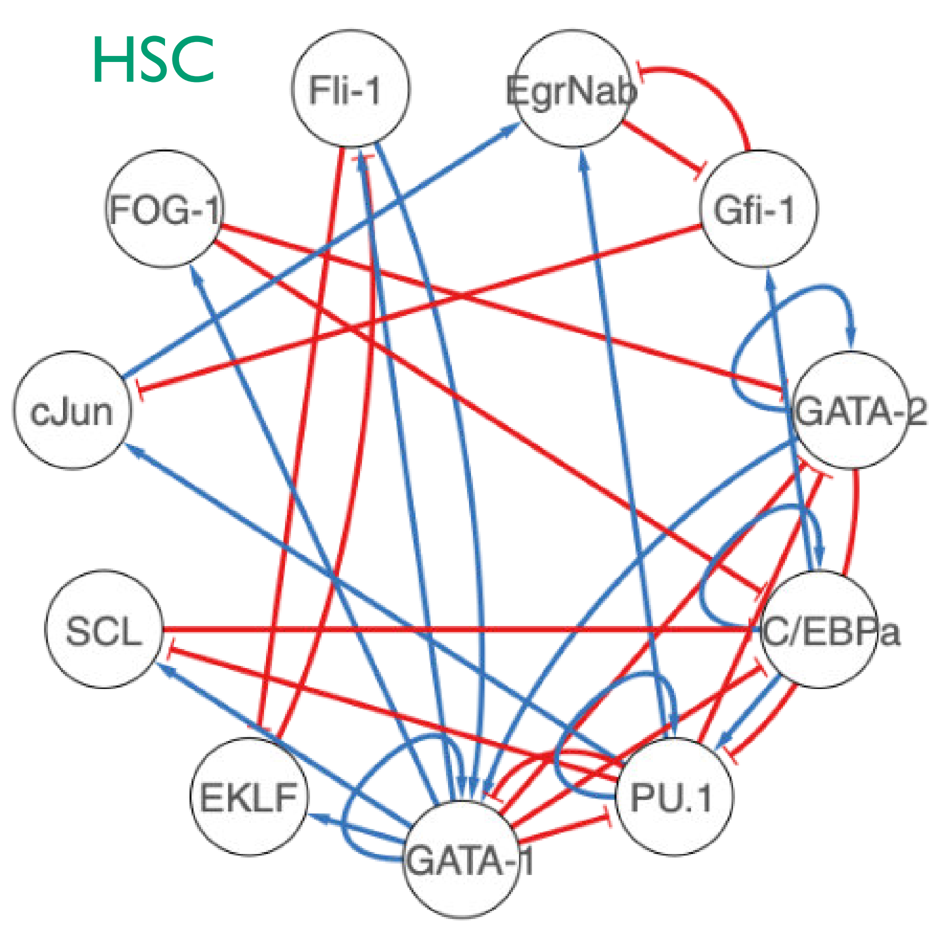
(1) Requires ground truth;
(2) Bounded between 0 and 1;
(3) Higher is better;
(4) Random baseline = density of truth.

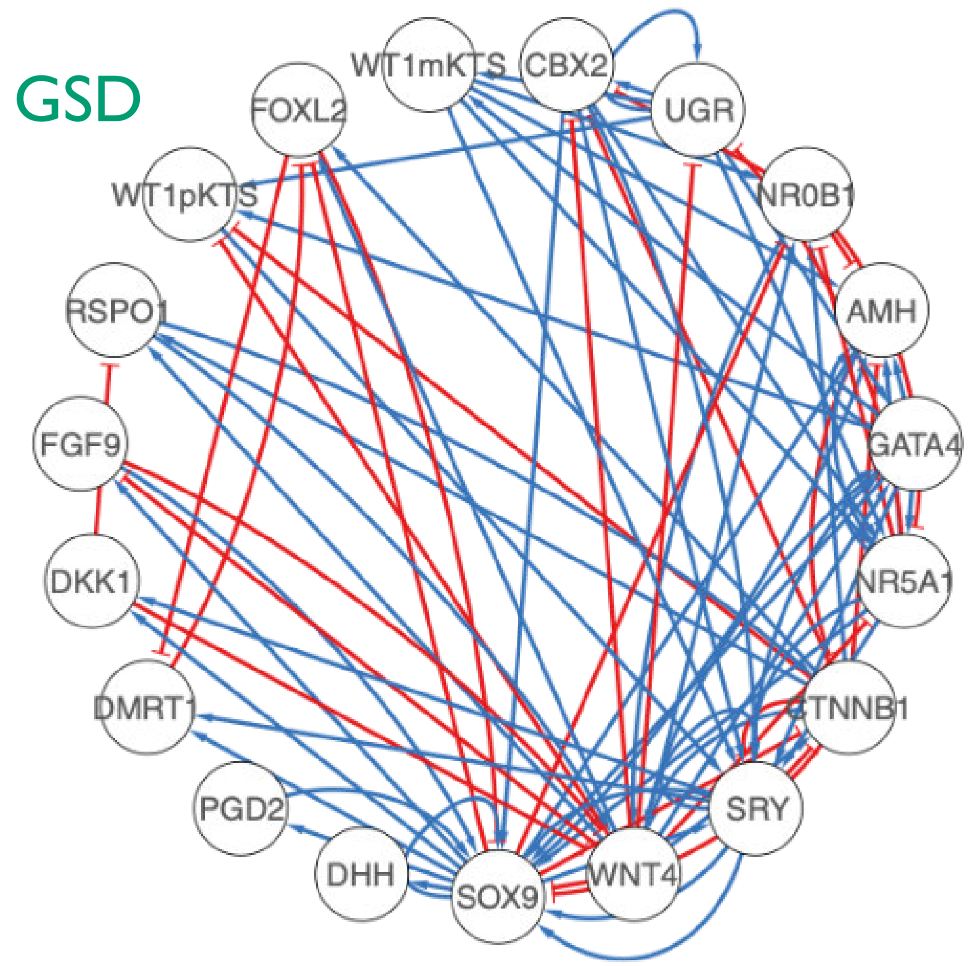
Result on simulated data

Beeline datasets

Real networks: (1) Hematopoietic Stem Cell Differentiation (HSC) (2) Gonadal Sex Determination (GSD)

Count matrices simulated from SDEs



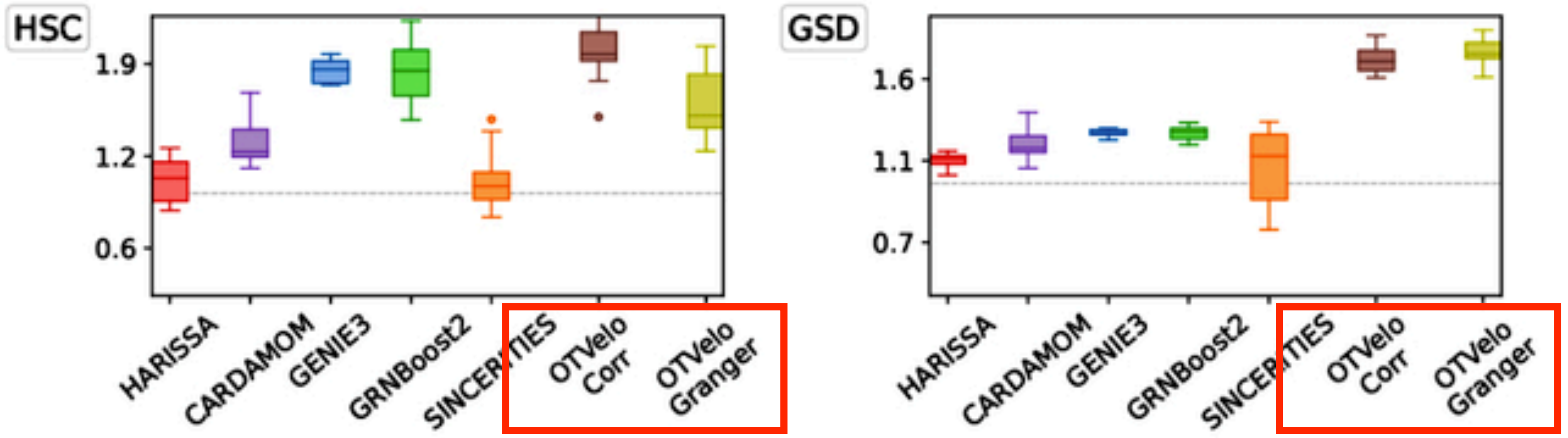


Metrics:

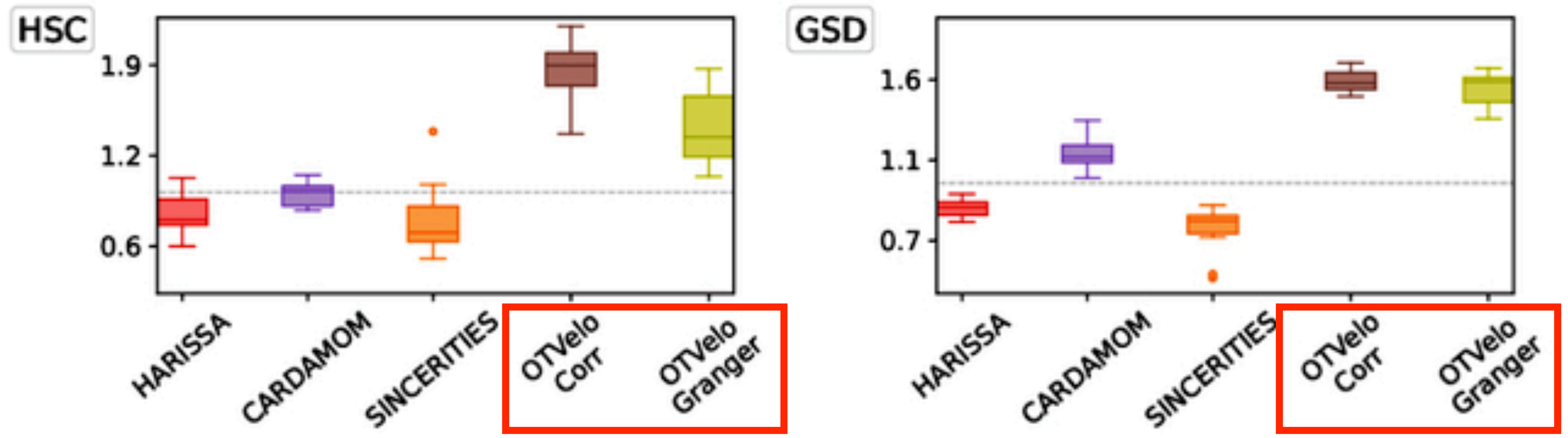
(1) AUPR ratio = AUPR over random baseline;

(2) For signed prediction, 'true positive' requires correct sign. Some algorithms don't have signed prediction.

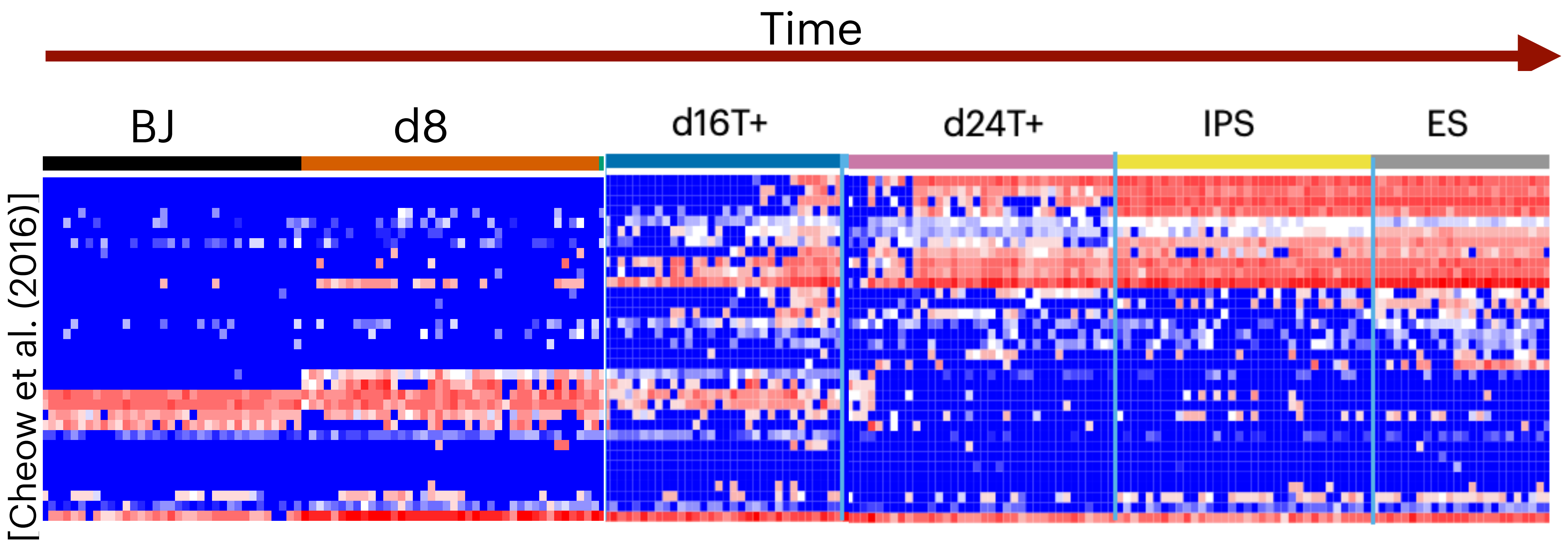
AUPR ratio (unsigned)



AUPR ratio (signed)



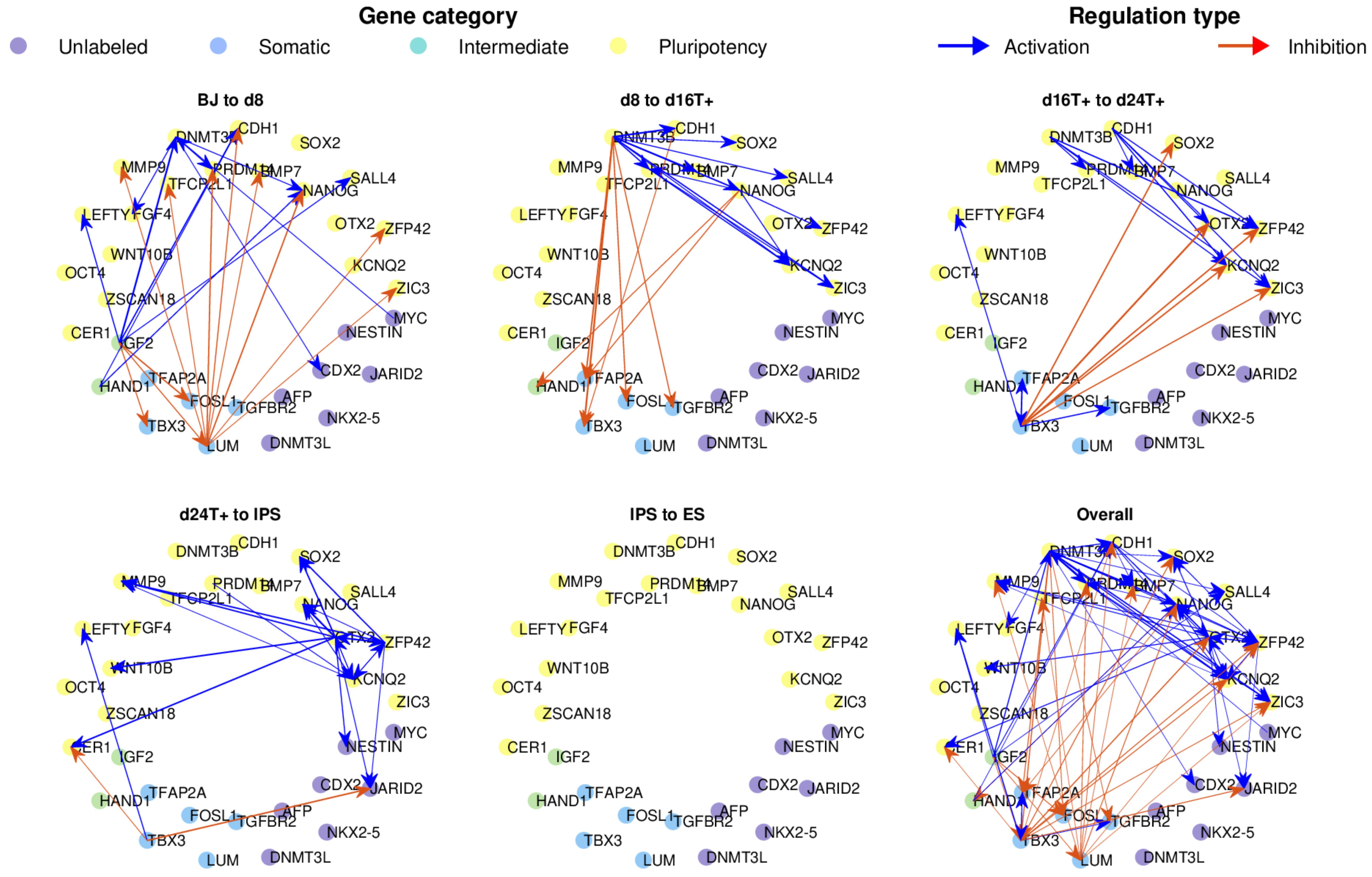
Result on real data



[Cheow et al. (2016)]

	experimental network	
	AUPRC	AUROC
GENIE3	0.027	0.592
GRNBoost2	**0.033**	0.616
SINCERITIES	0.018	0.474
HARISSA	0.028	0.586
CARDAMOM	0.024	0.566
OTVelo-Corr	0.031	**0.667**
OTVelo-Granger	0.020	0.480

Table: AUC values across different methods compared against experimental network
Random baselines: 0.02 and 0.5



Two inference problems

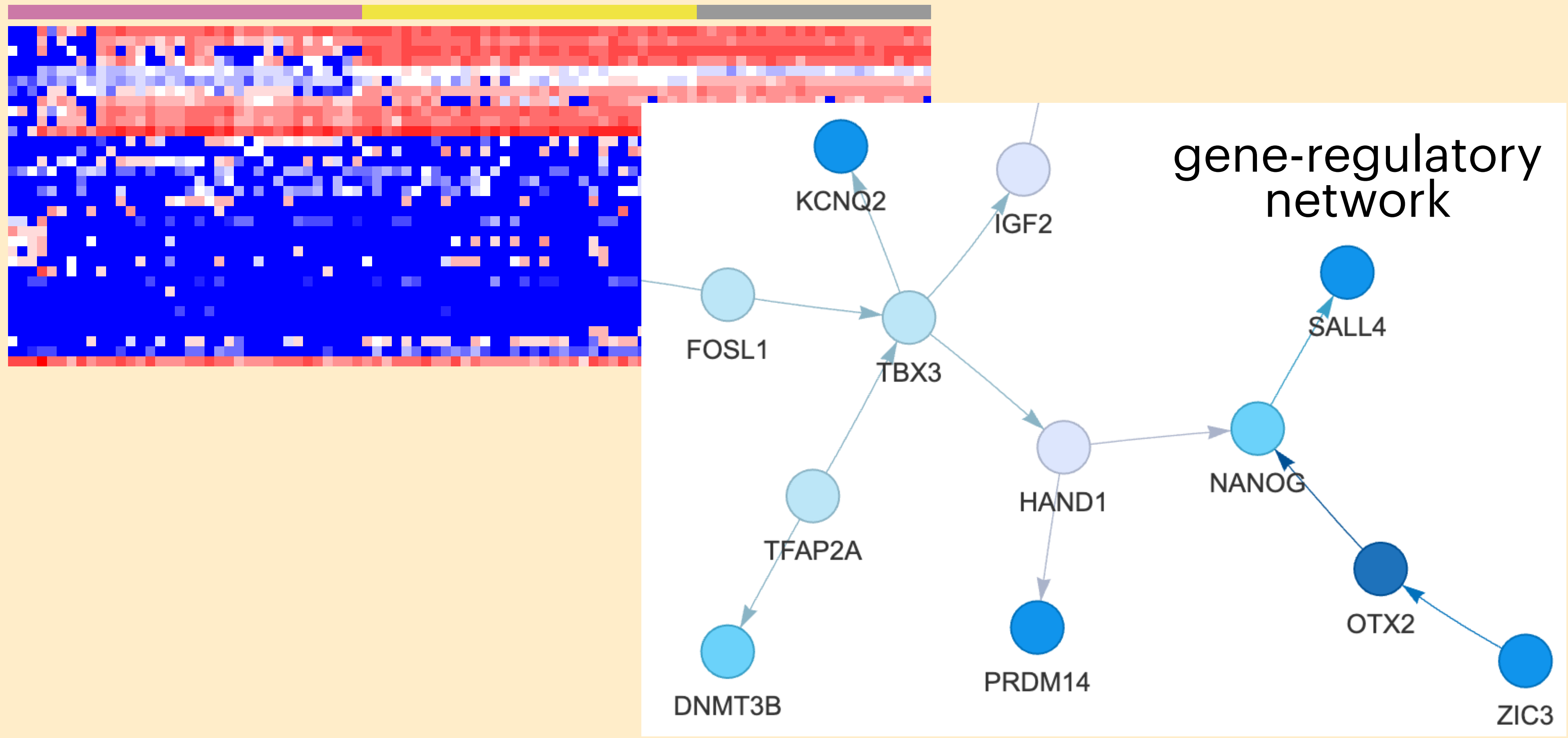
Inferring mechanisms from single-cell sequencing data

Joint work with

Bjorn Sandstede

Ritambhara Singh

Erica Larschan

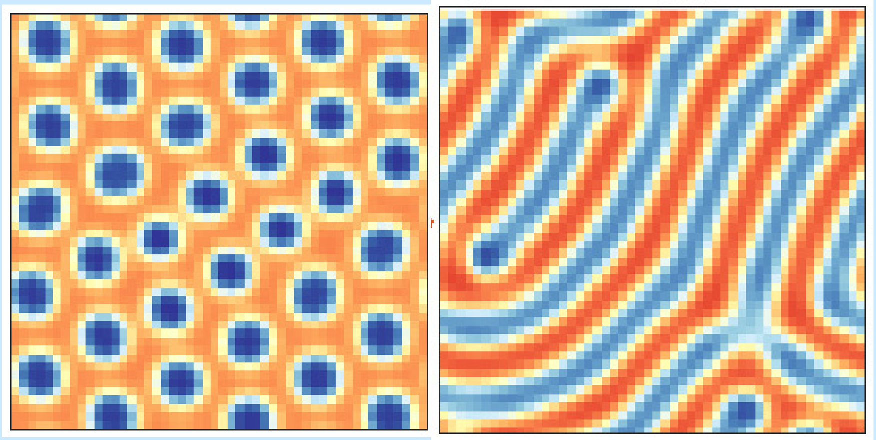


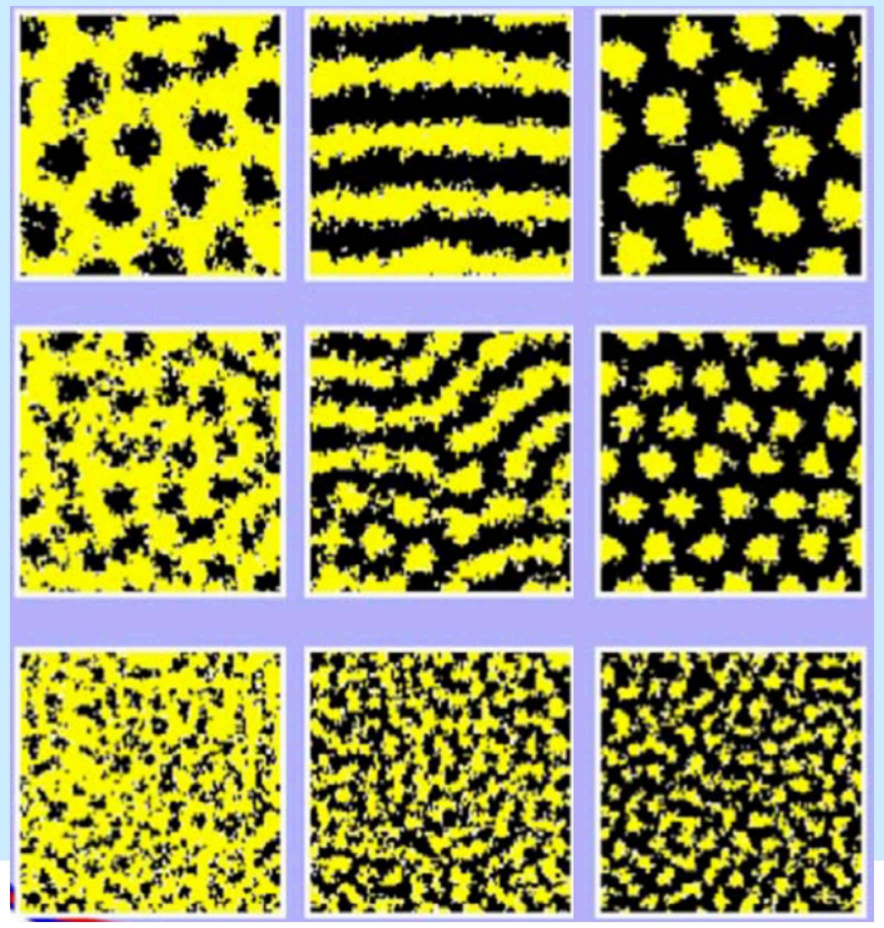
Quantifying Turing patterns and their bifurcation

Joint work with

Biorn Sandstede

Sam Maffa

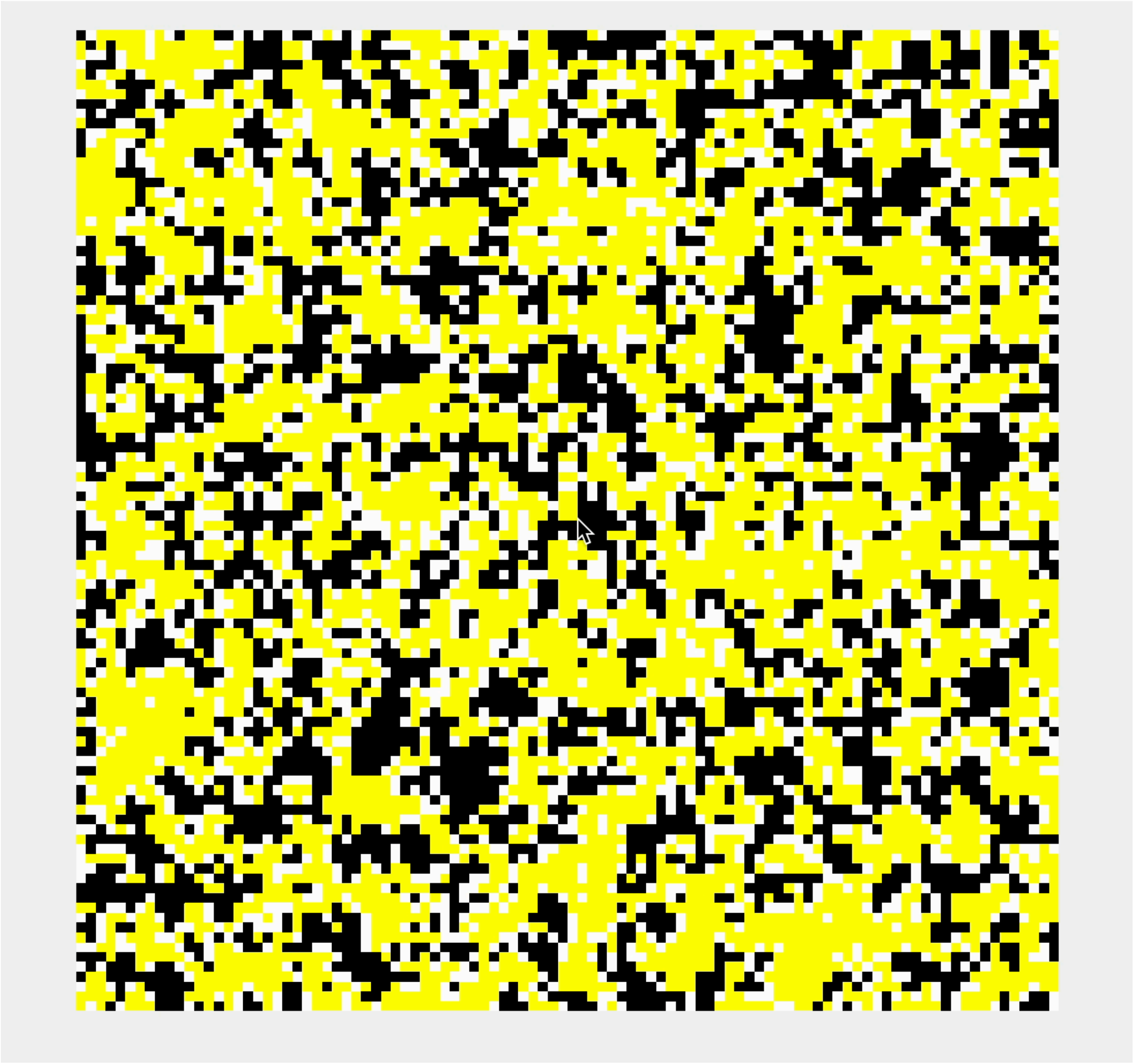


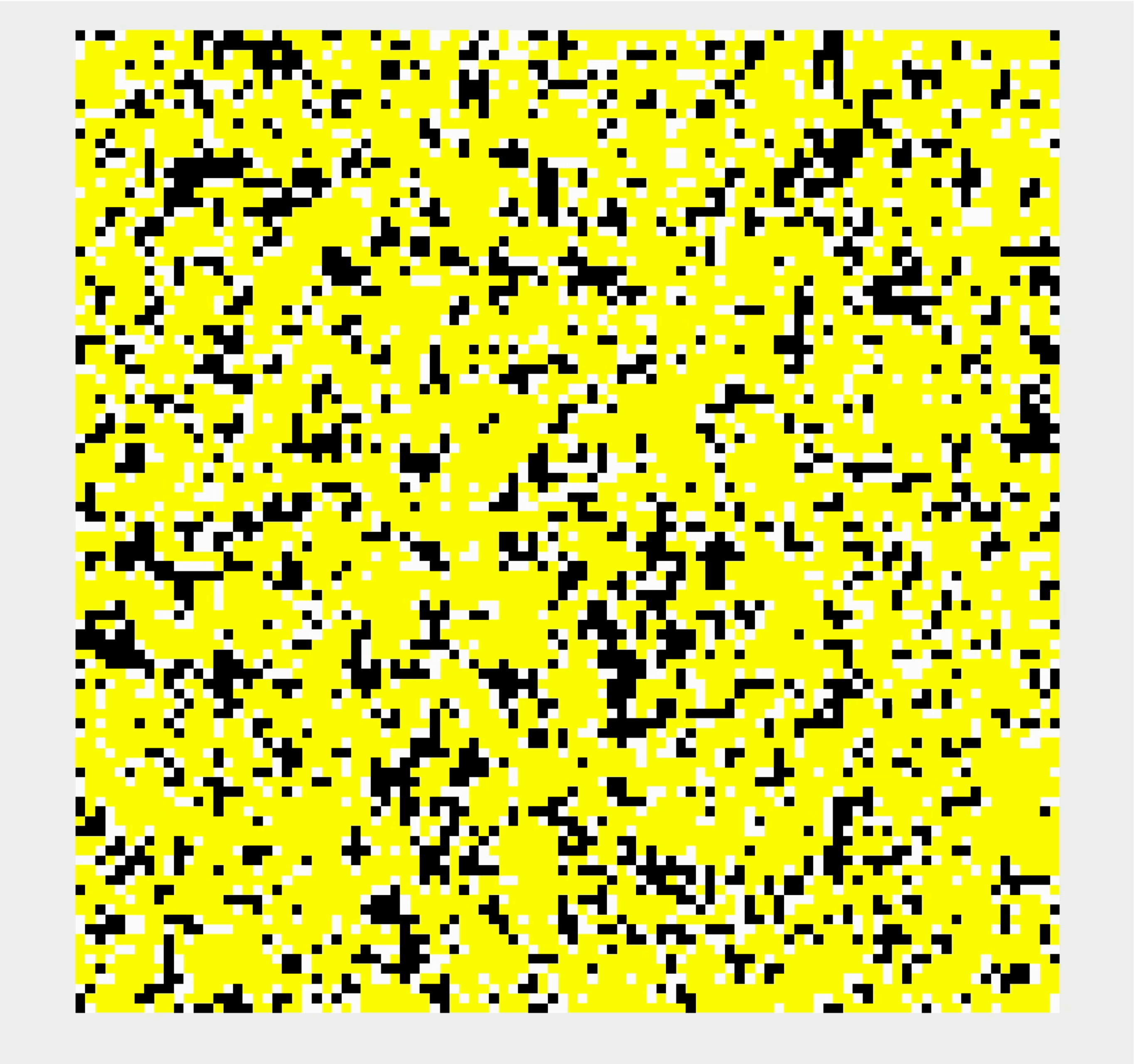


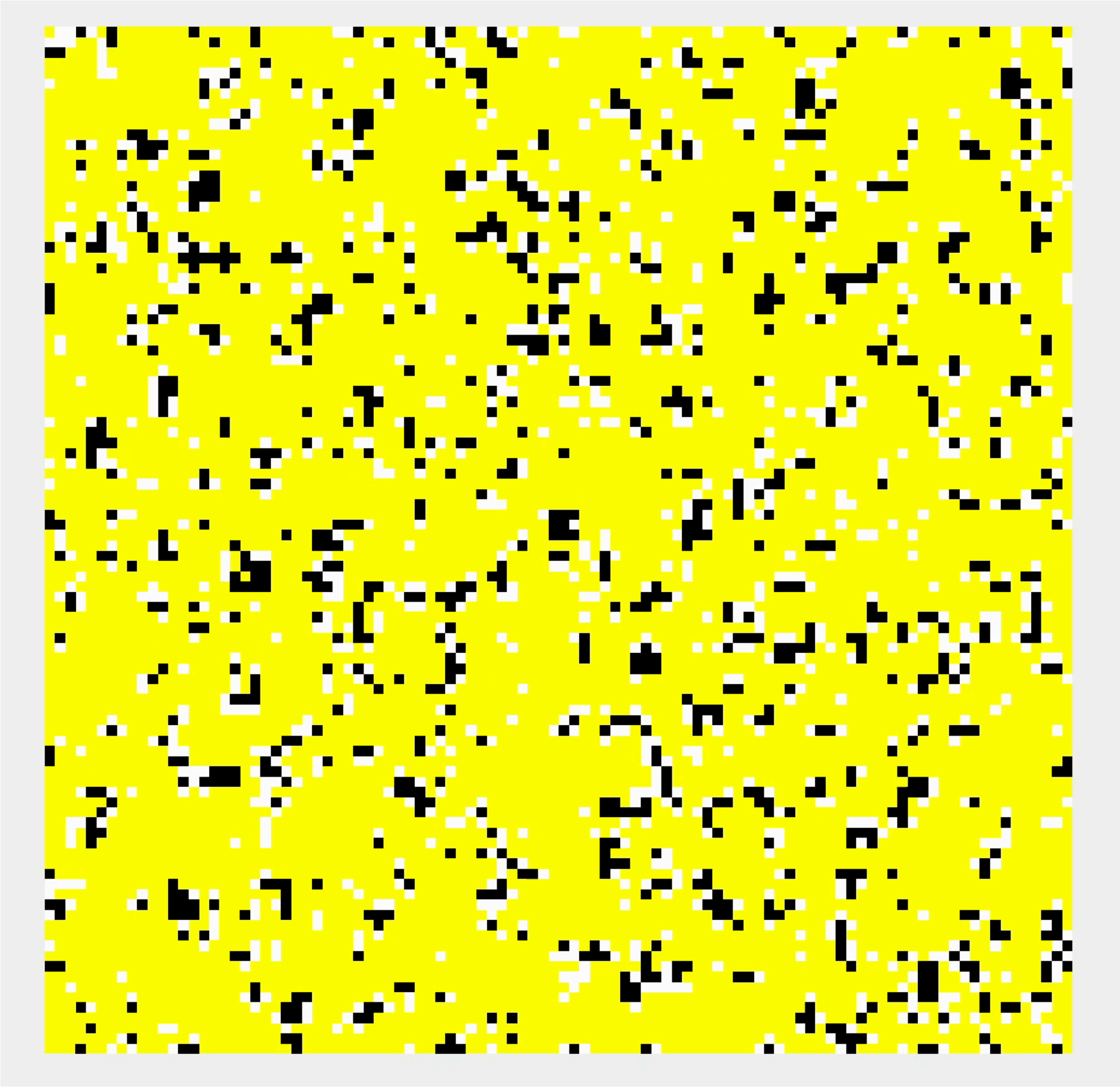
How to distinguish different patterns?

Example 1: fish skin model

- How does fish skin form their patterns?
- Example [Bullara & Decker] on-lattice model
- Stripe/Spots/Nothing?
- Depends on spatial extent and strength of interaction between cells







How to distinguish different patterns?

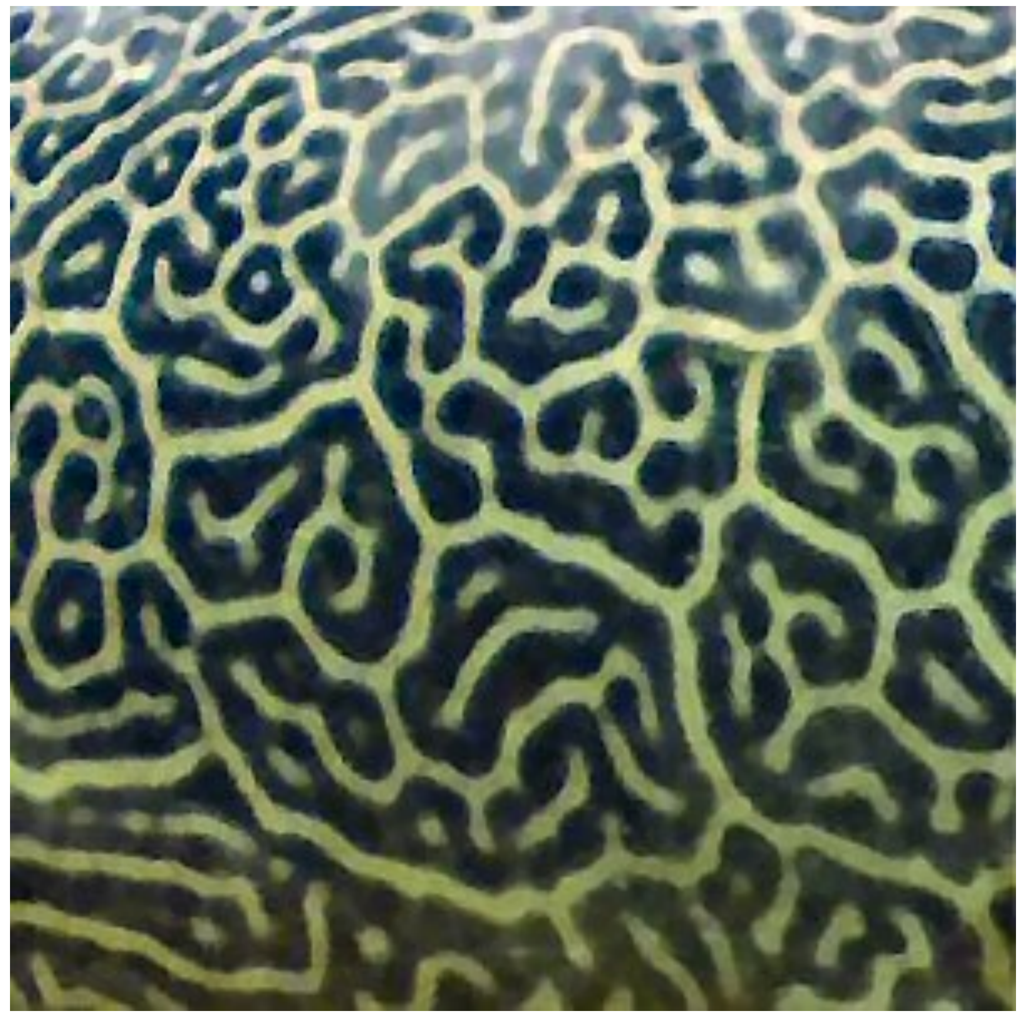
Example 2: reaction-diffusion equation

- Turing patterns: introduced by Alan Turing in 1952 to describe patterns in nature;
- Spatially extended reaction-diffusion systems are known to generate such patterns, such as Brusselator (we focus on the case with $x = (x_1, x_2) \in \mathbb{R}^2$):

$$\begin{aligned} \partial_t u &= \boxed{D_0 \nabla^2 u} + \boxed{a - (1+b)u + vu^2} \\ \partial_t v &= \boxed{D_1 \nabla^2 v} + \boxed{bu - vu^2} \end{aligned}$$

Diffusion

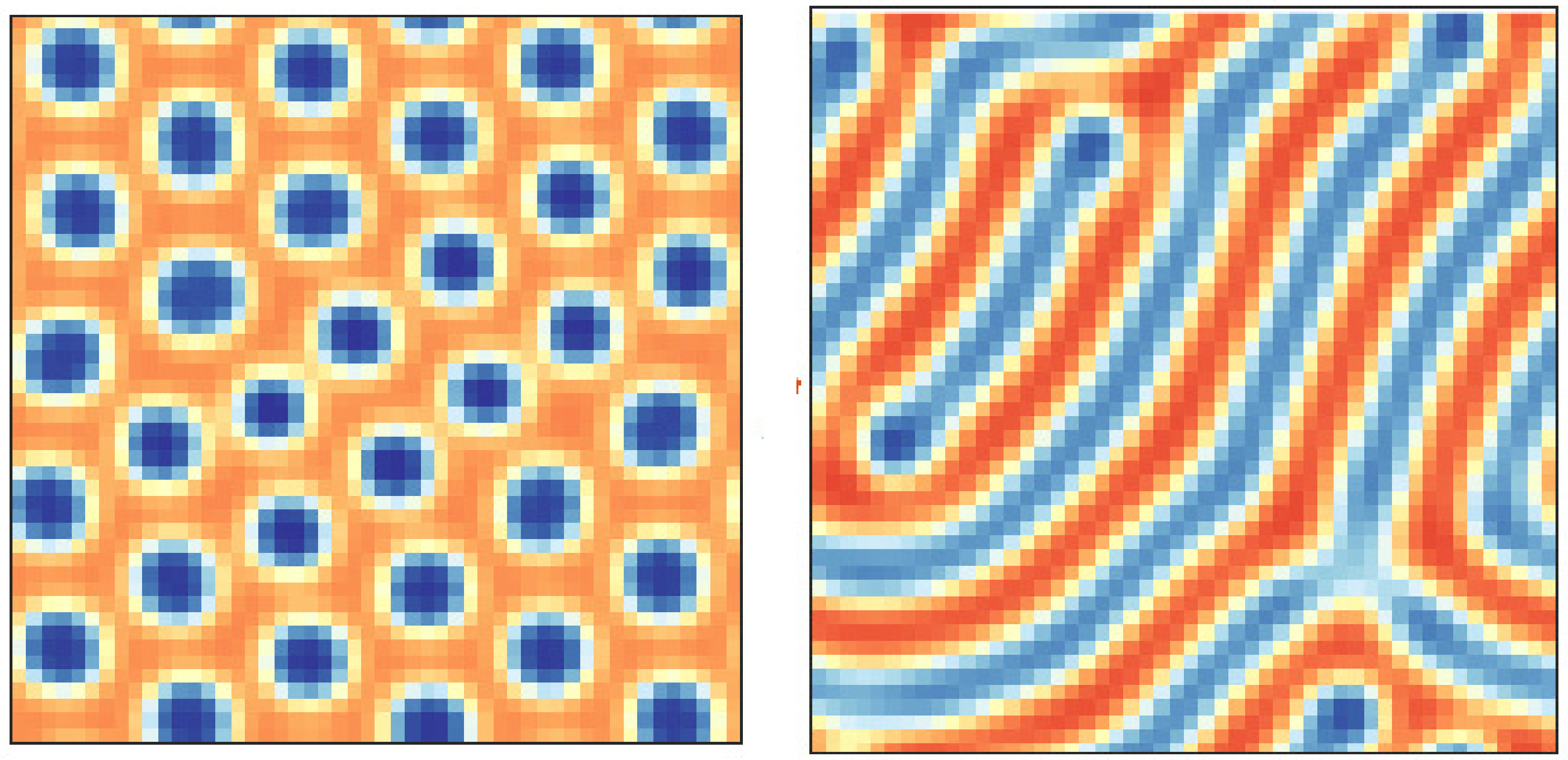
Reaction



Wikipedia (Pufferfish skin)

Wikipedia (Patterned vegetation)

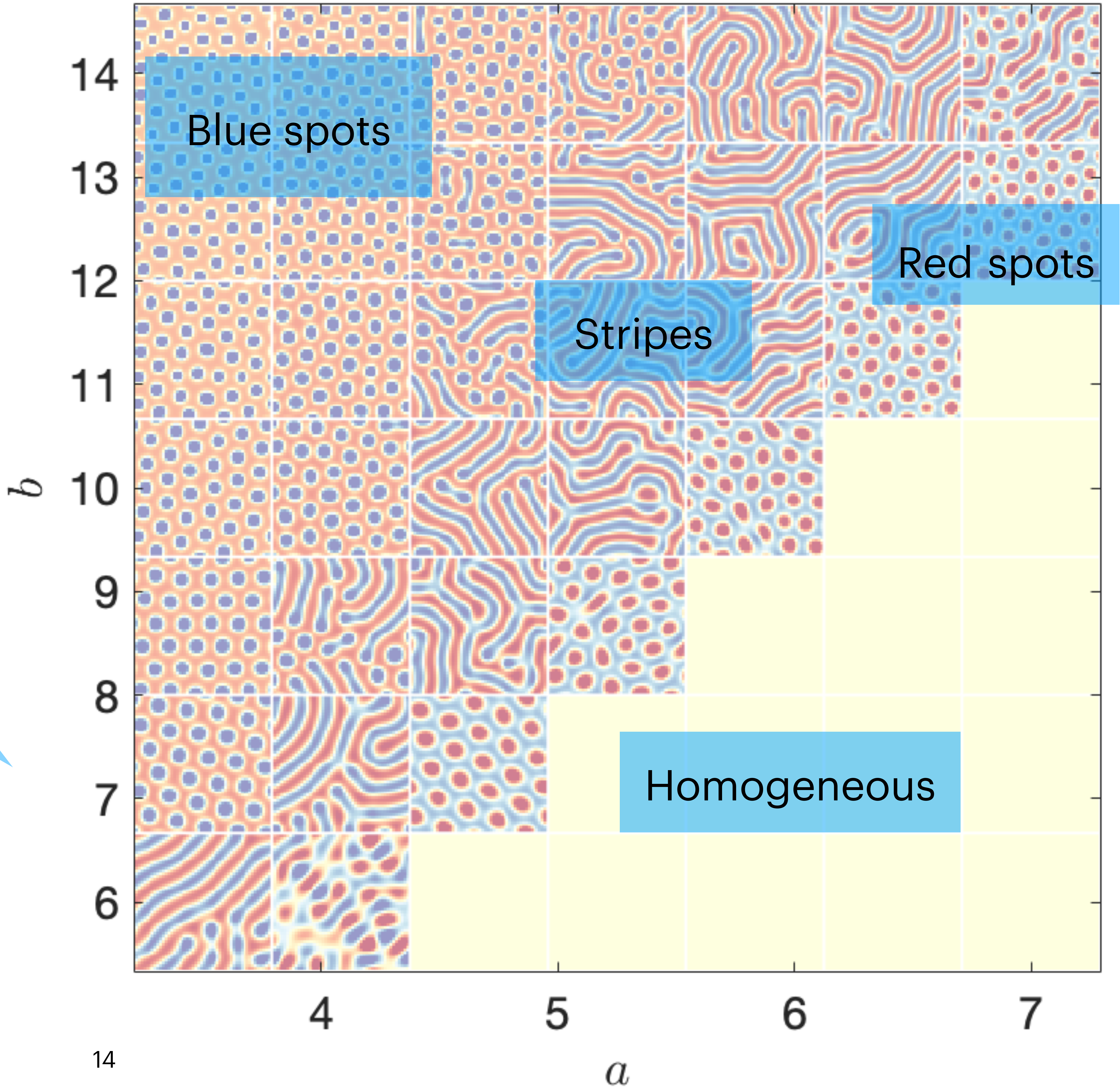
$u(x_1, x_2, T)$ for $T \gg 1$



Bifurca

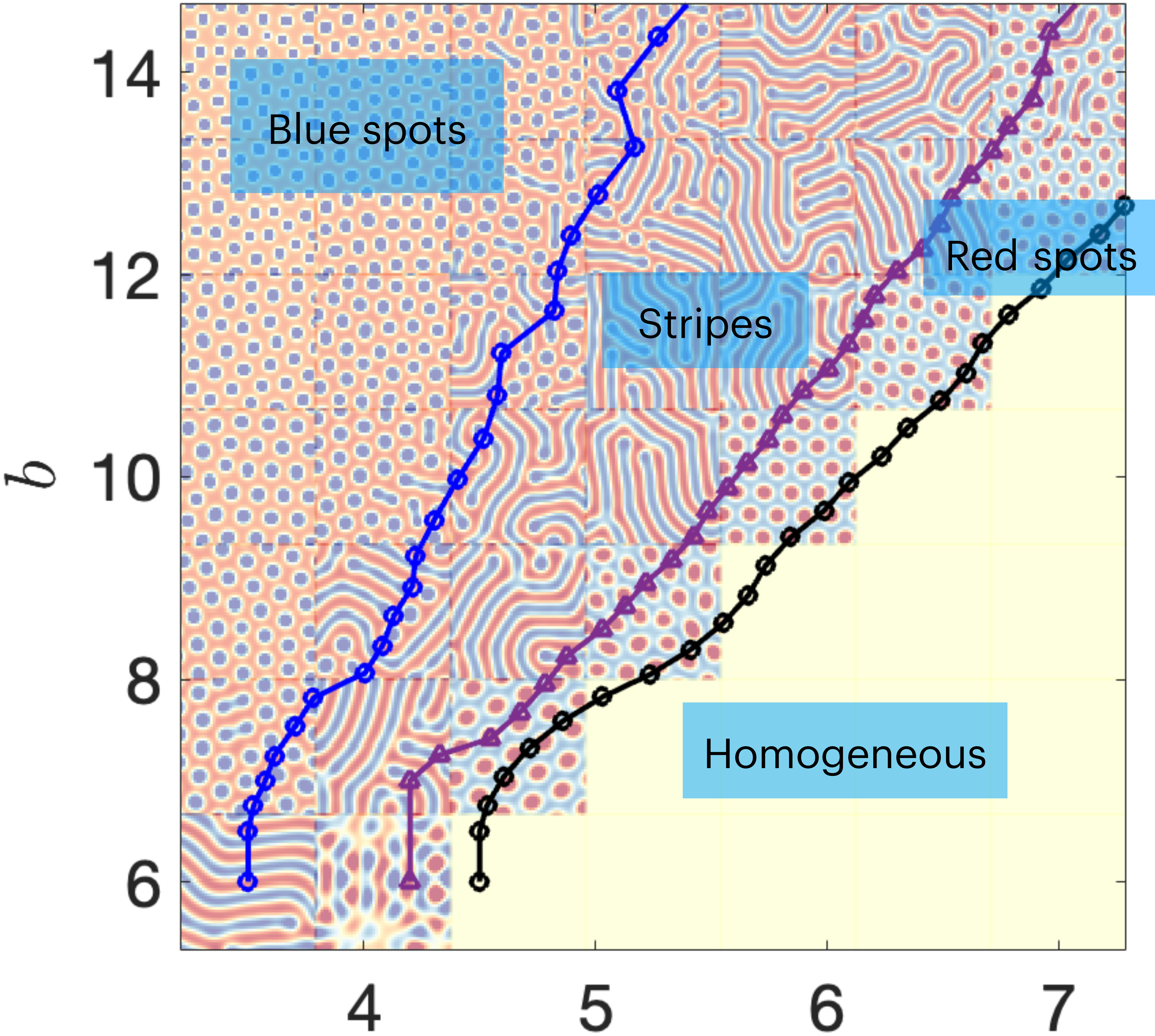
$$\partial_t u = D_0 \nabla^2 u + a - (1 + b) u +$$
$$\partial_t v = D_1 \nabla^2 v + bu - vu^2$$

Where does the pattern transition happen in the parameter space?



Spoiler (what we found)

Where does the pattern transition happen in the parameter space?



Challenges

Traditional approaches:

- Visual inspection for direct simulation: **cheap, but** **less accurate**
- Existing software (continuation): given initial value problems, compute each pattern and their stability and find their interfaces between different patterns.

  accurate, but **expensive and specialized** (wave number selection? bistability?)

  They don’t work for agent-based models like the fish skin!

Solve a classification problem via neural network?

- Requires a training set with pre-labeled solutions — lots of human effort…
- Interpretability?
- Active learning (collect data strategically, only near bifurcation curve)?

Data-driven algorithm via cheap, direct simulation?

Sure, but how should we *quantify the patterns*?

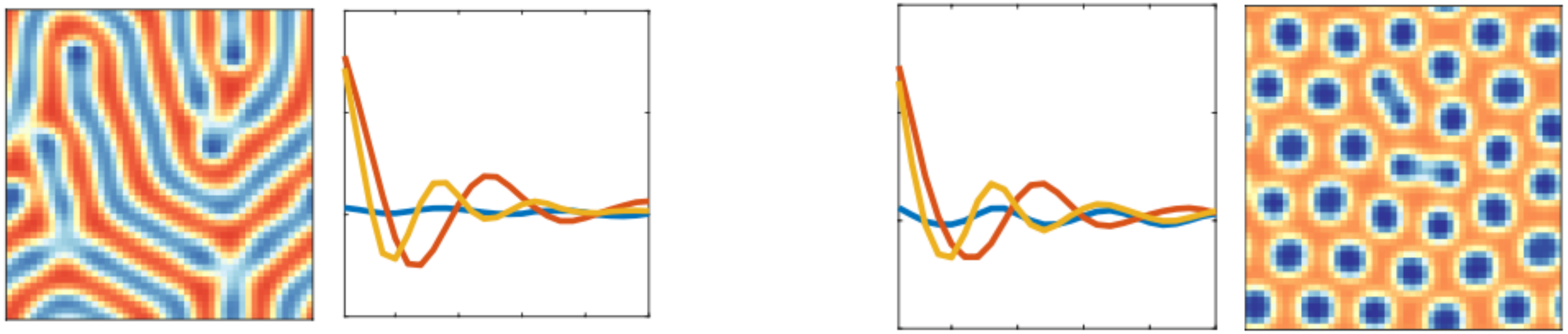
Furthermore, how do we *compare the patterns*?

Some options for pattern quantification

- Option 1: 2D Fourier coefficients
- Option 2: pattern correlation function [Gavagnin et al 2018]:

$$f(m) = \frac{c(m)}{\mathbb{E}[c(m)]}, \quad c(m) = \left|\{(x, y) \in \text{sublevel set} \mid \|x - y\| = m\}\right|$$

- They are not good enough at distinguishing spots/stripes with random orientations!



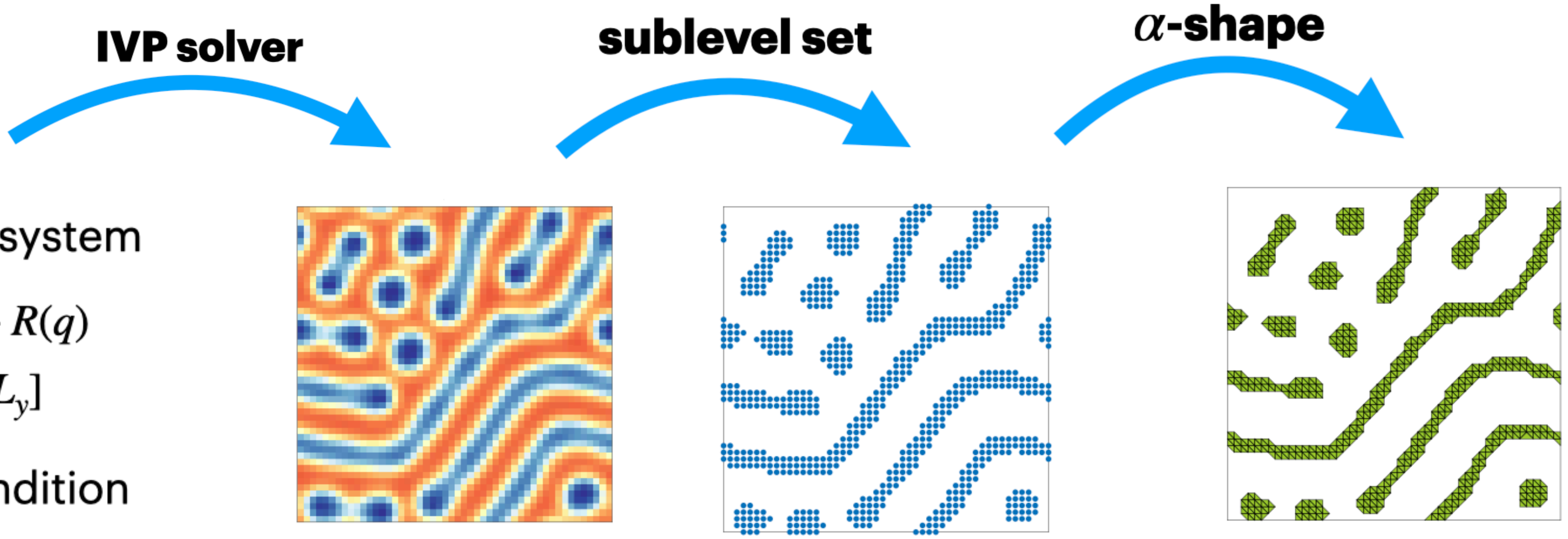
Our approach: from PDE solution to patterns

Reaction-diffusion system

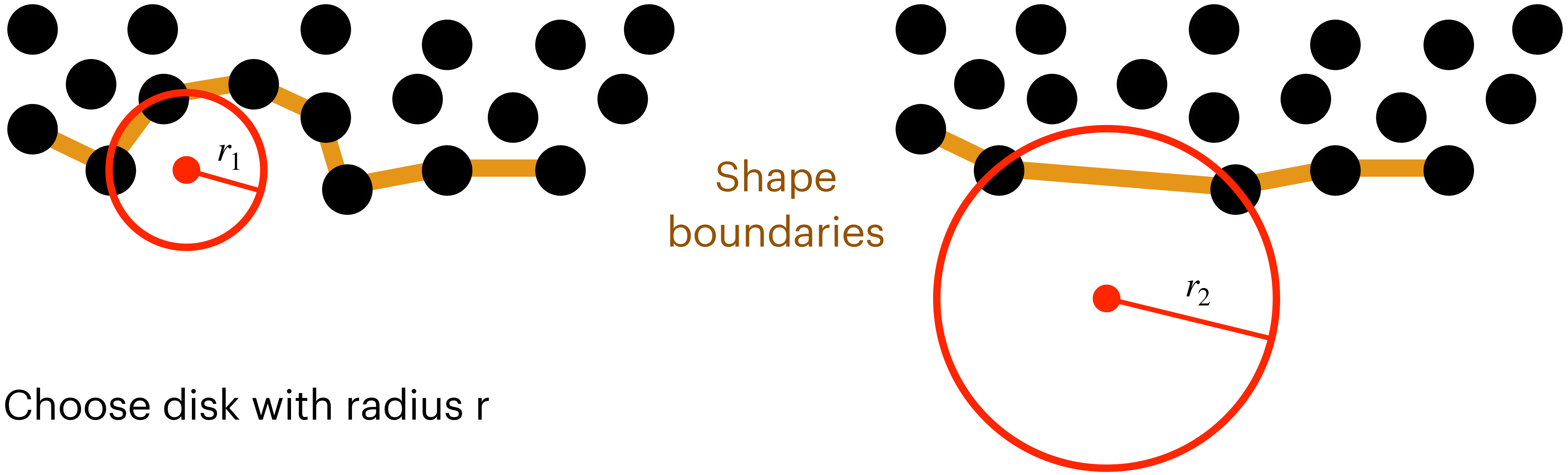
$$\partial_t q(x,t) = D\nabla^2 q + R(q)$$

$$x \in [0,L_x] \times [0,L_y]$$

Random initial condition



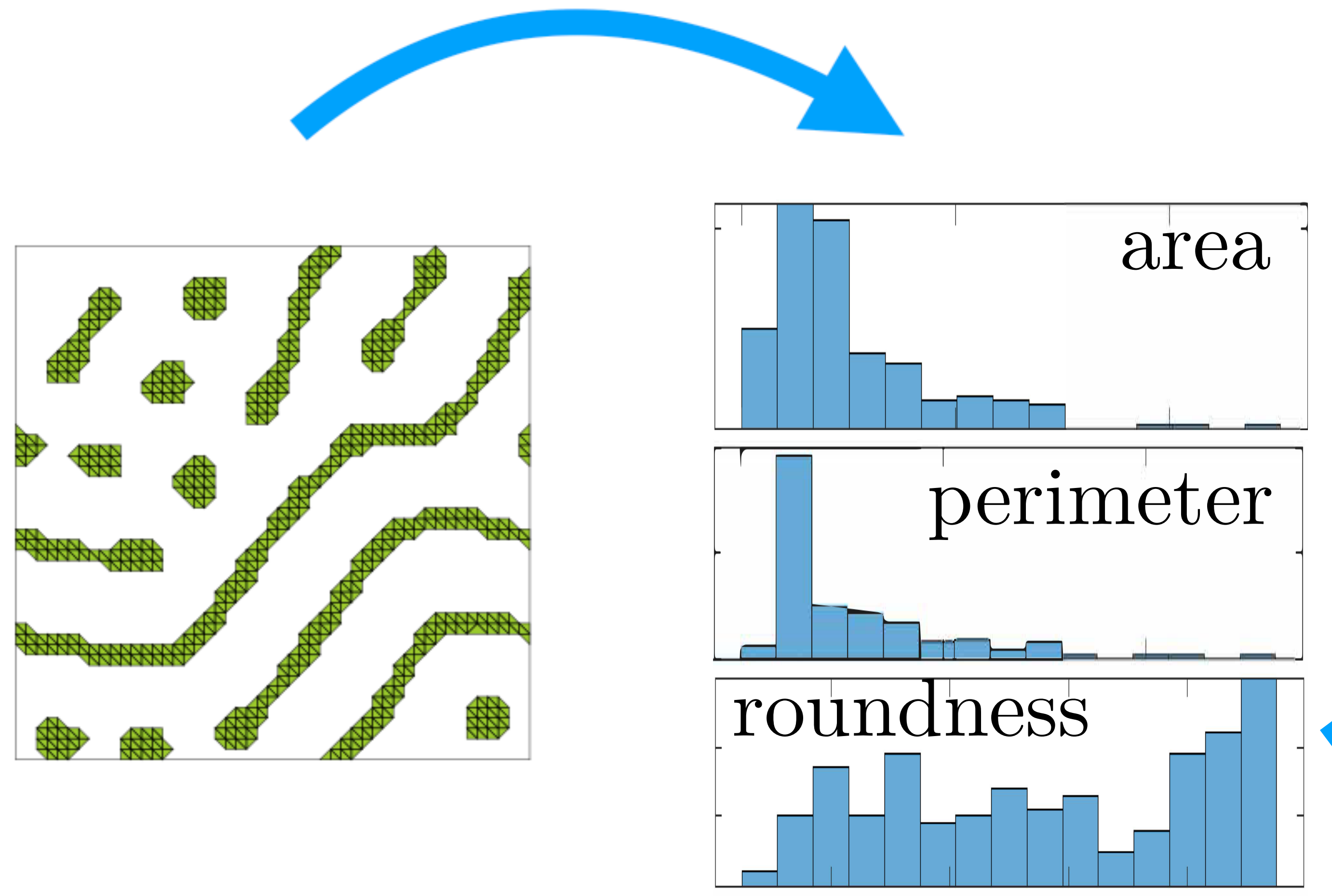
α-shapes



- Choose disk with radius r
- Connect two data points precisely when we can position the disk so that it contains only these two data points

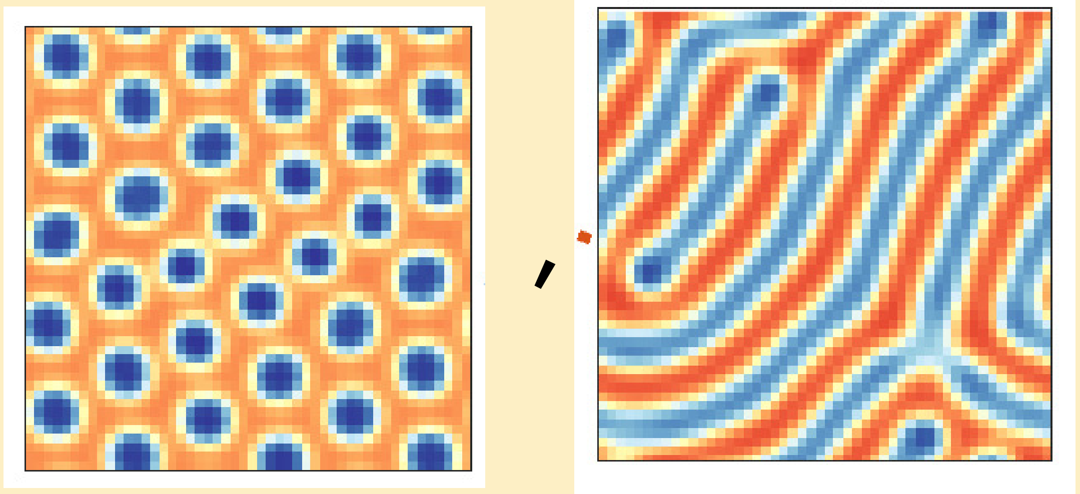
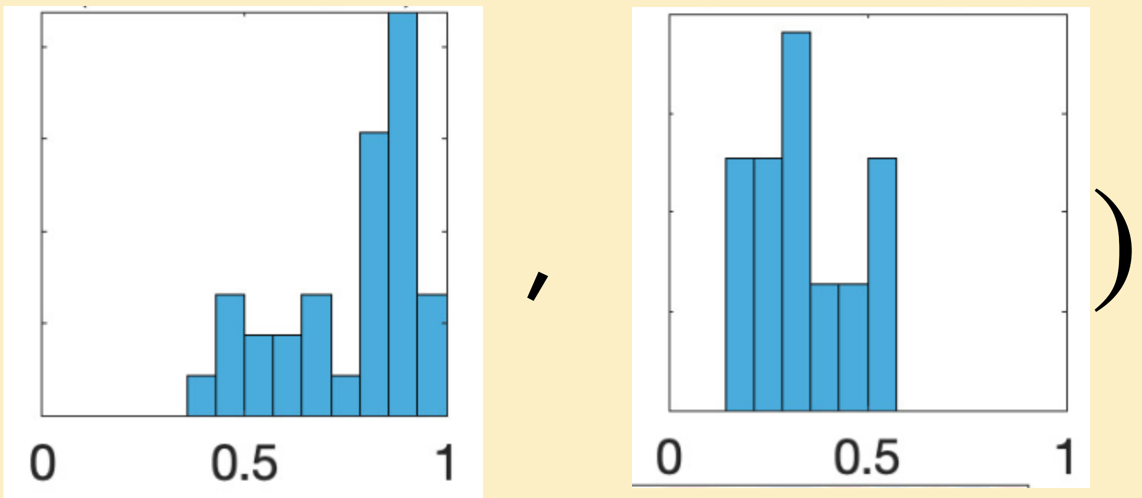
Our approach: from patterns to metric

cluster-wise features



$roundness = \frac{4\pi \times area}{perimeter^2}$, bounded between 0 (stripes) and 1 (spots)

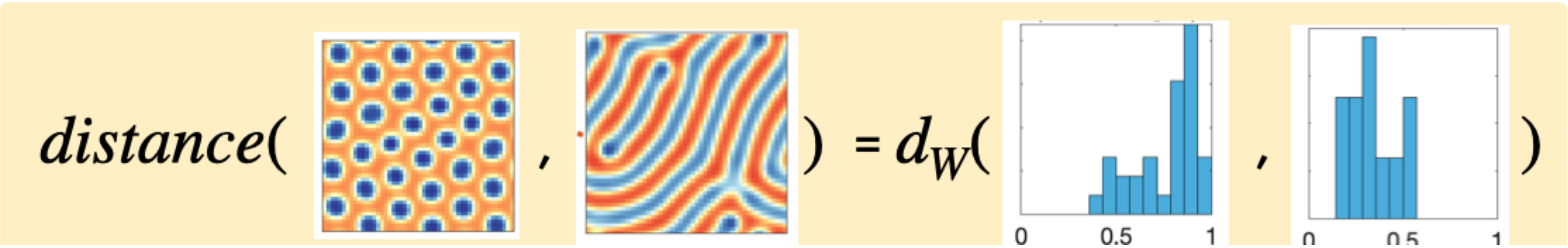
2-Wasserstein distance

$distance($  $) = d_W($  $)$

Theorem (under some assumptions): The pattern statistics $\mu_f : \mathbb{R}^2 \longrightarrow \mathrm{Prob}(\mathcal{Z}),\ p \longmapsto \mu_f(p)$

is continuous in the Wasserstein metric on $\mathrm{Prob}(\mathcal{Z})$

Empirical evaluation of Wasserstein distance in feature space

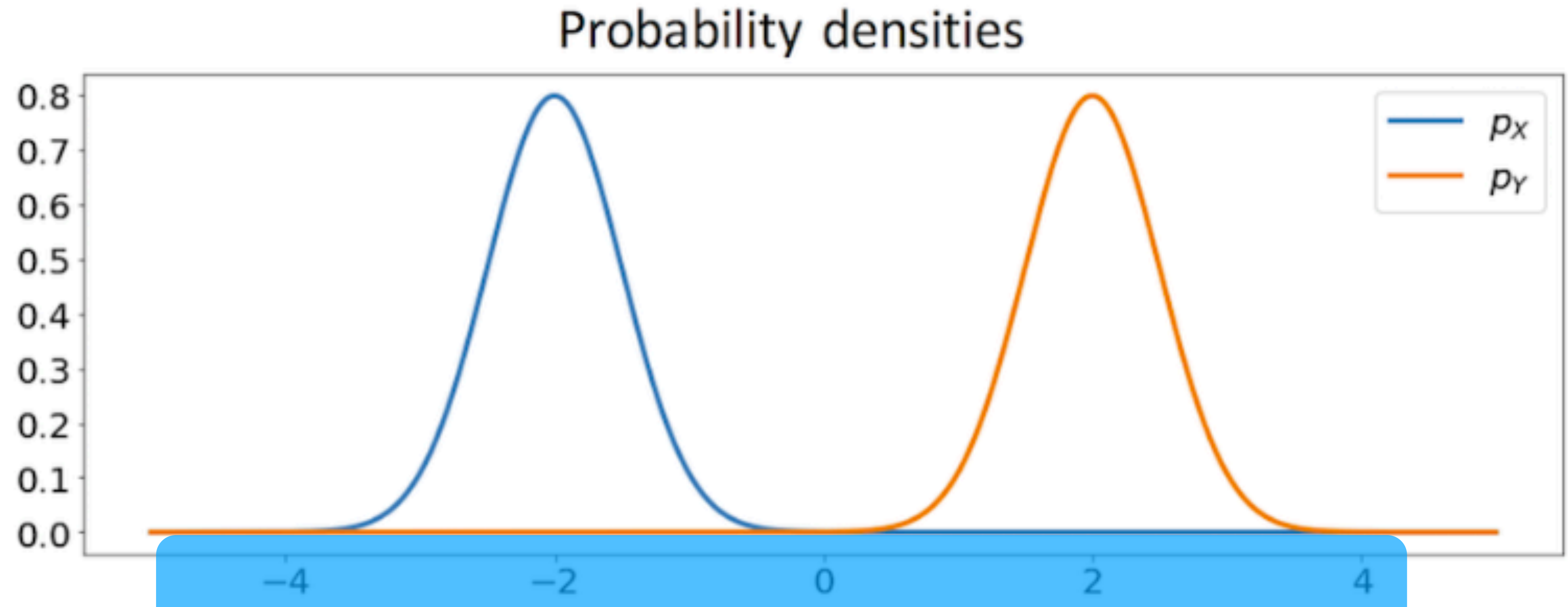


- For 1D histograms (our case!), with Euclidean cost, it reduces to:

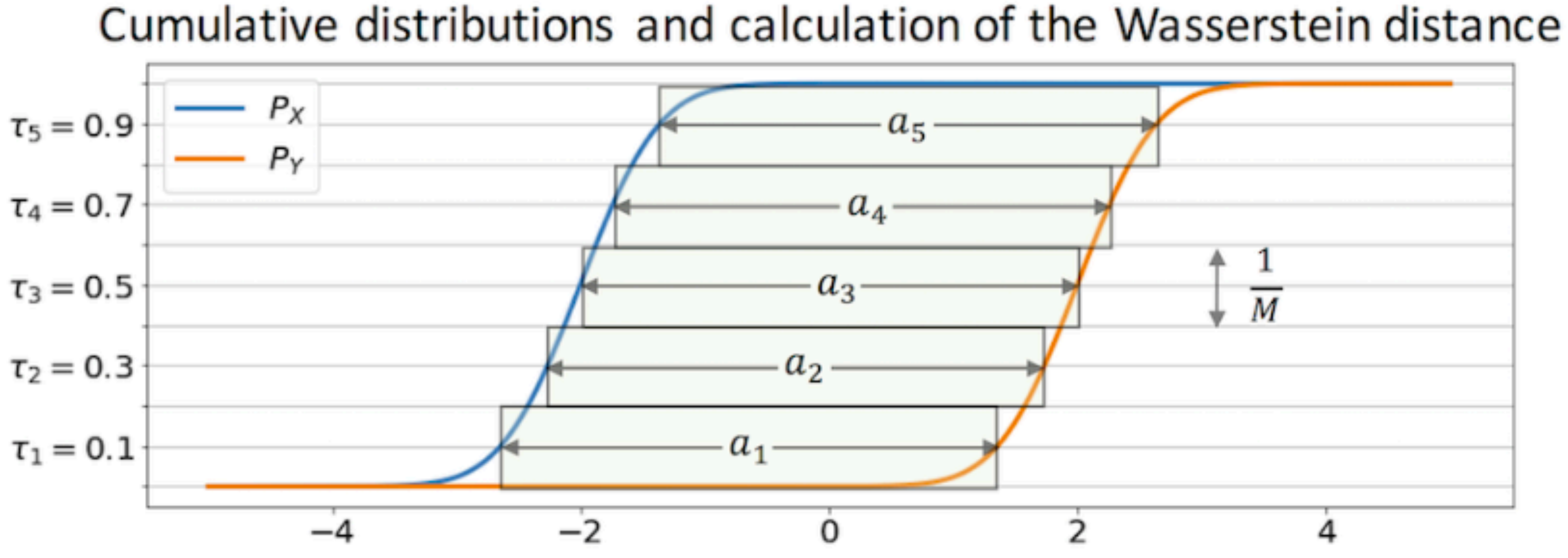
$$d_W^2(\rho,\mu) = \int_0^1 (F_\rho^{-1}(p) - F_\mu^{-1}(p))^2 \, dp,$$ [Kolouri and Martin 2018]

where $F_\rho^{-1}(p)$ is the inverse cumulative density function of ρ.

- Its sample-based approximation can be computed efficiently via a sorting algorithm. Existing implementation: *https://github.com/nklb/wasserstein-distance*

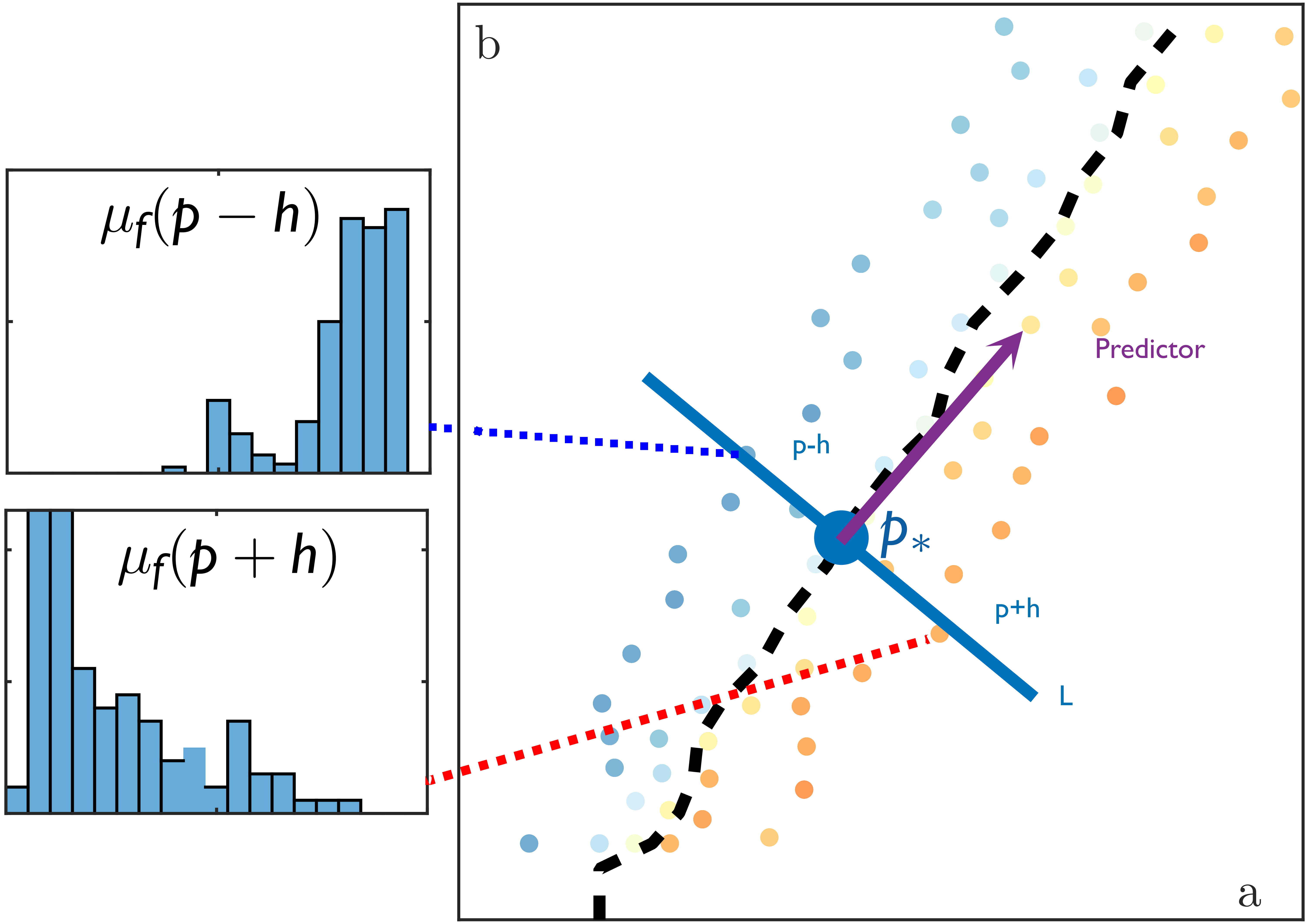


Intuitively, the metric measures the gap between all quantiles



Figures reproduced from [Kolouri and Martin 2018]

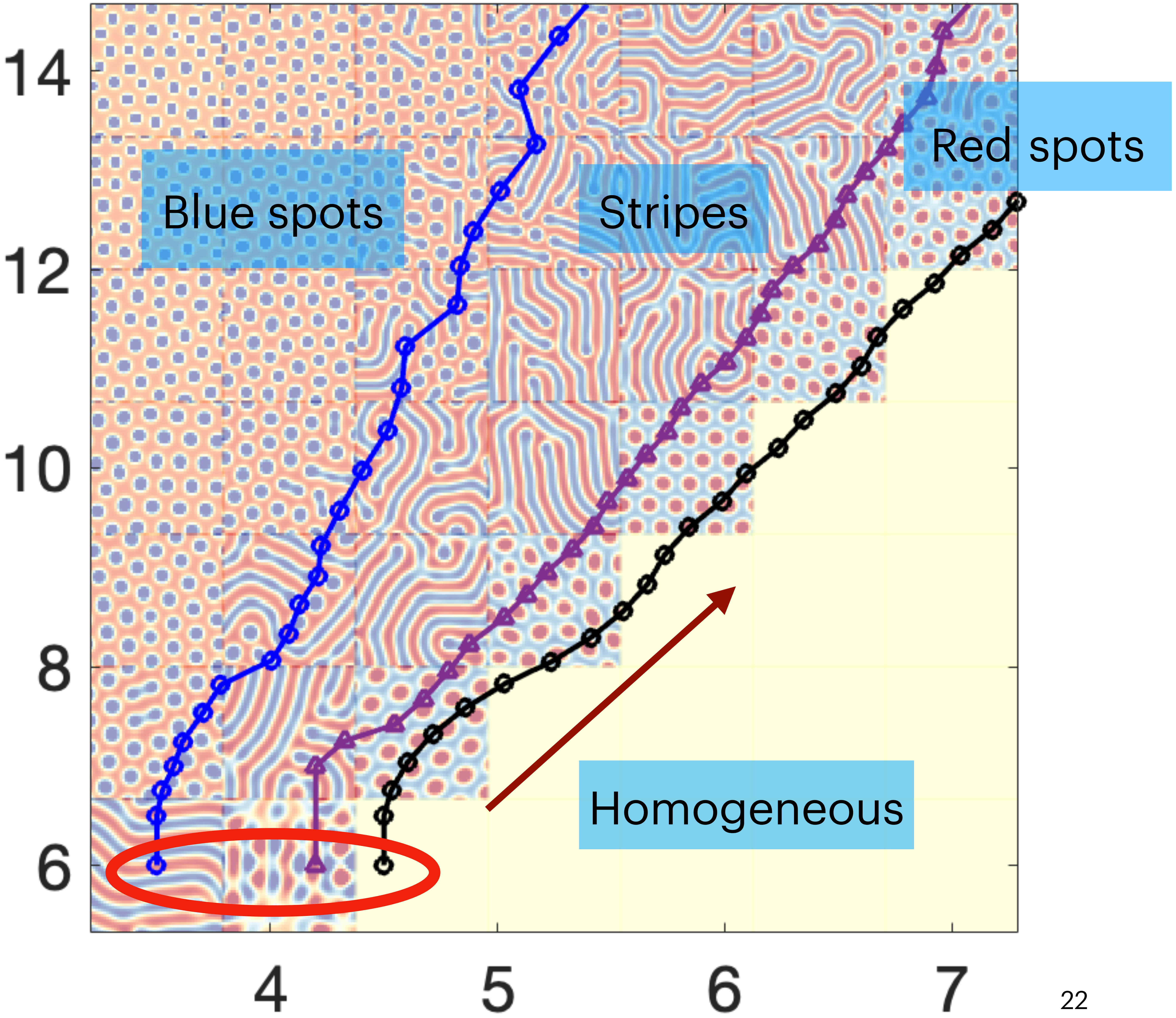
Numerical: continuation framework



Bifurcation curves are determined by maximizing Wasserstein distance of pattern statistics across curve

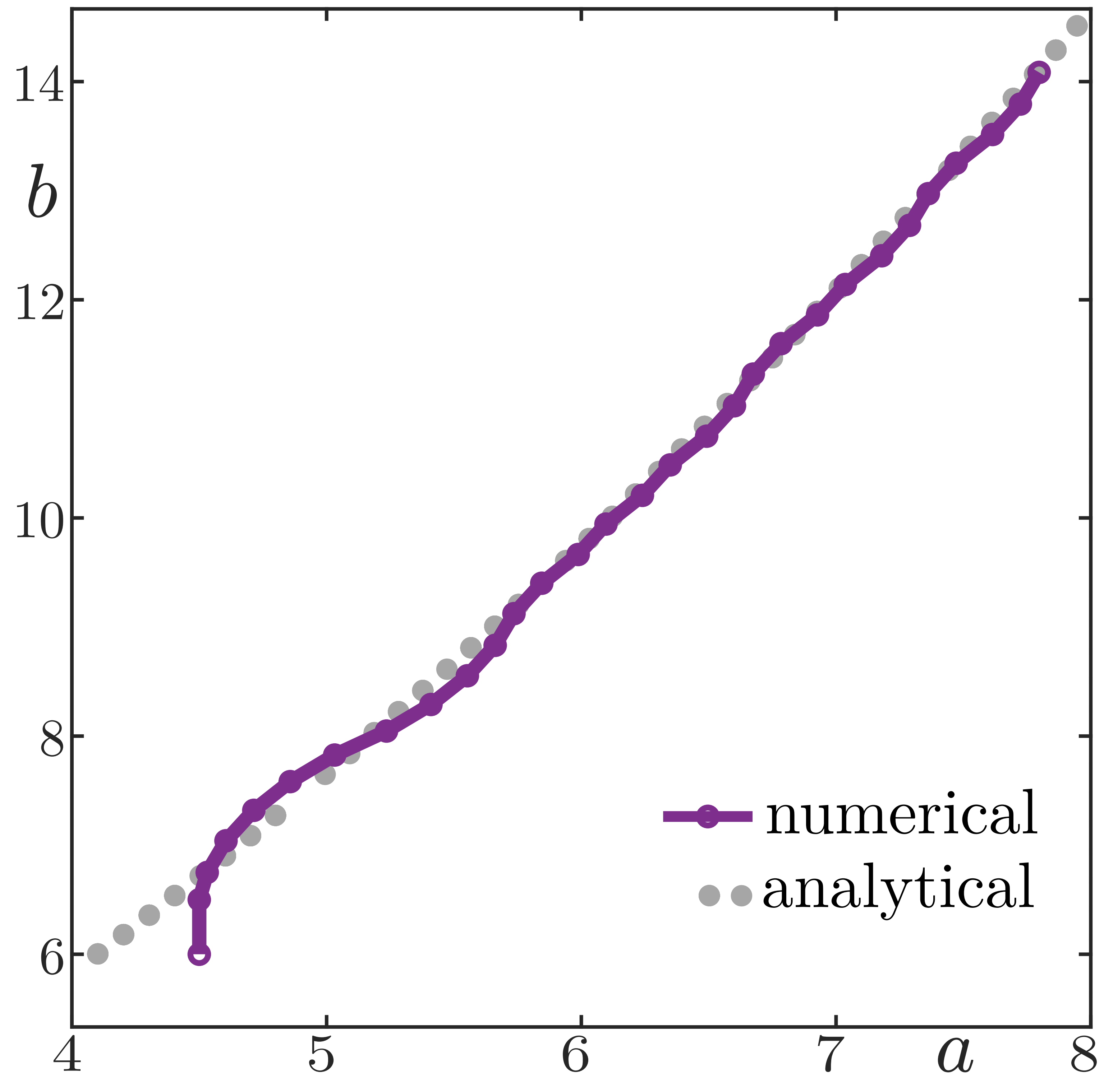
$$p_* := \arg\max_{p\in L} d_W(\mu_f(p-h), \mu_f(p+h))$$

Application: Brusselator model



Validation:

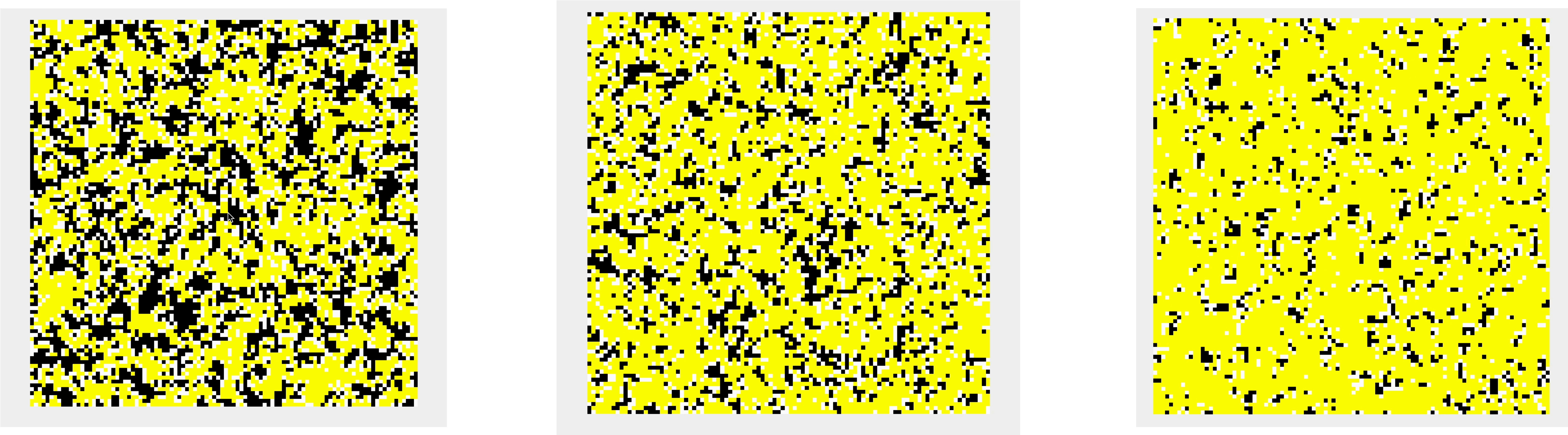
- Stripes/Spots: visual inspection…
- Homogeneous state: analytical solution



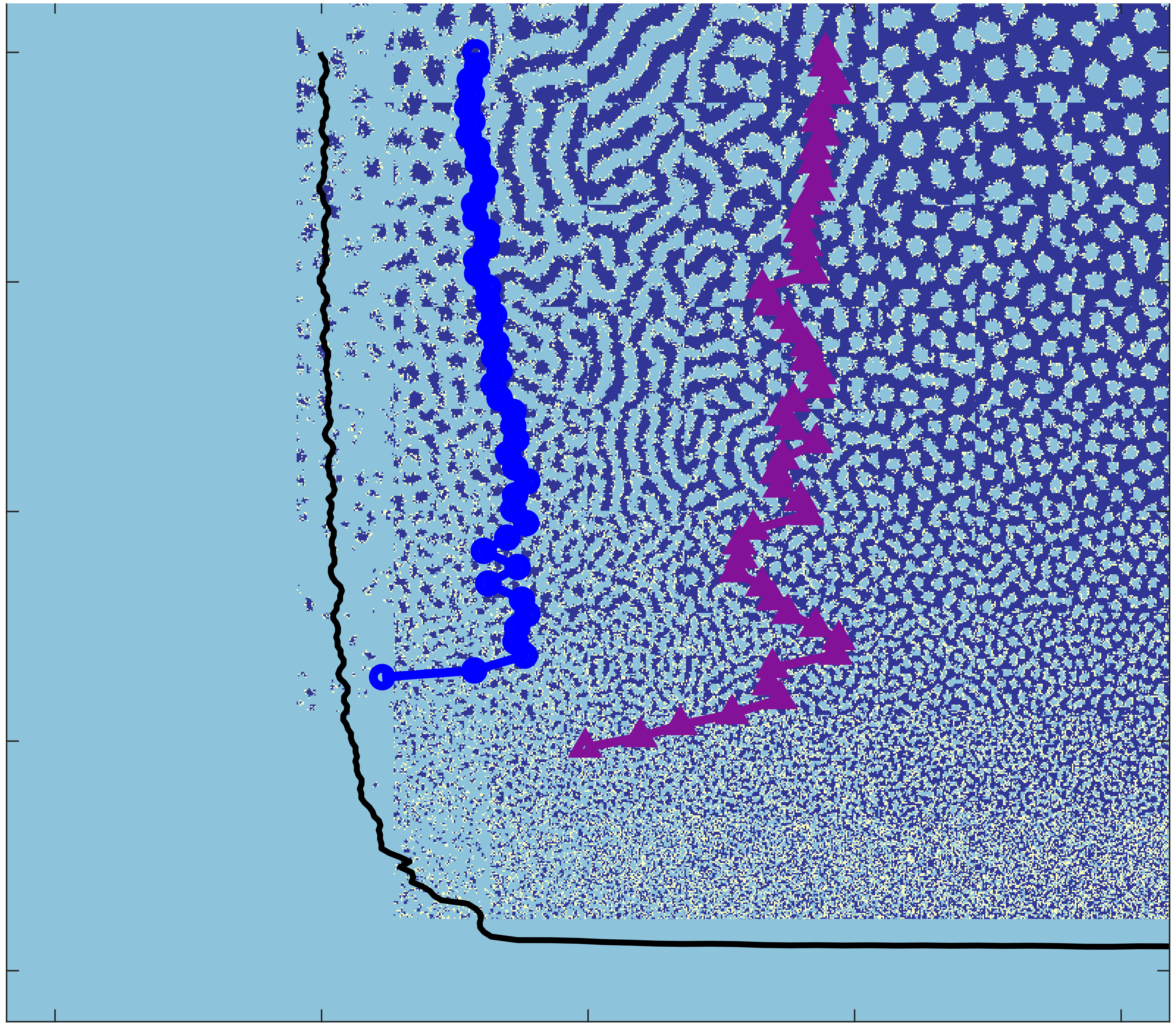
Moving back to the original motivation...

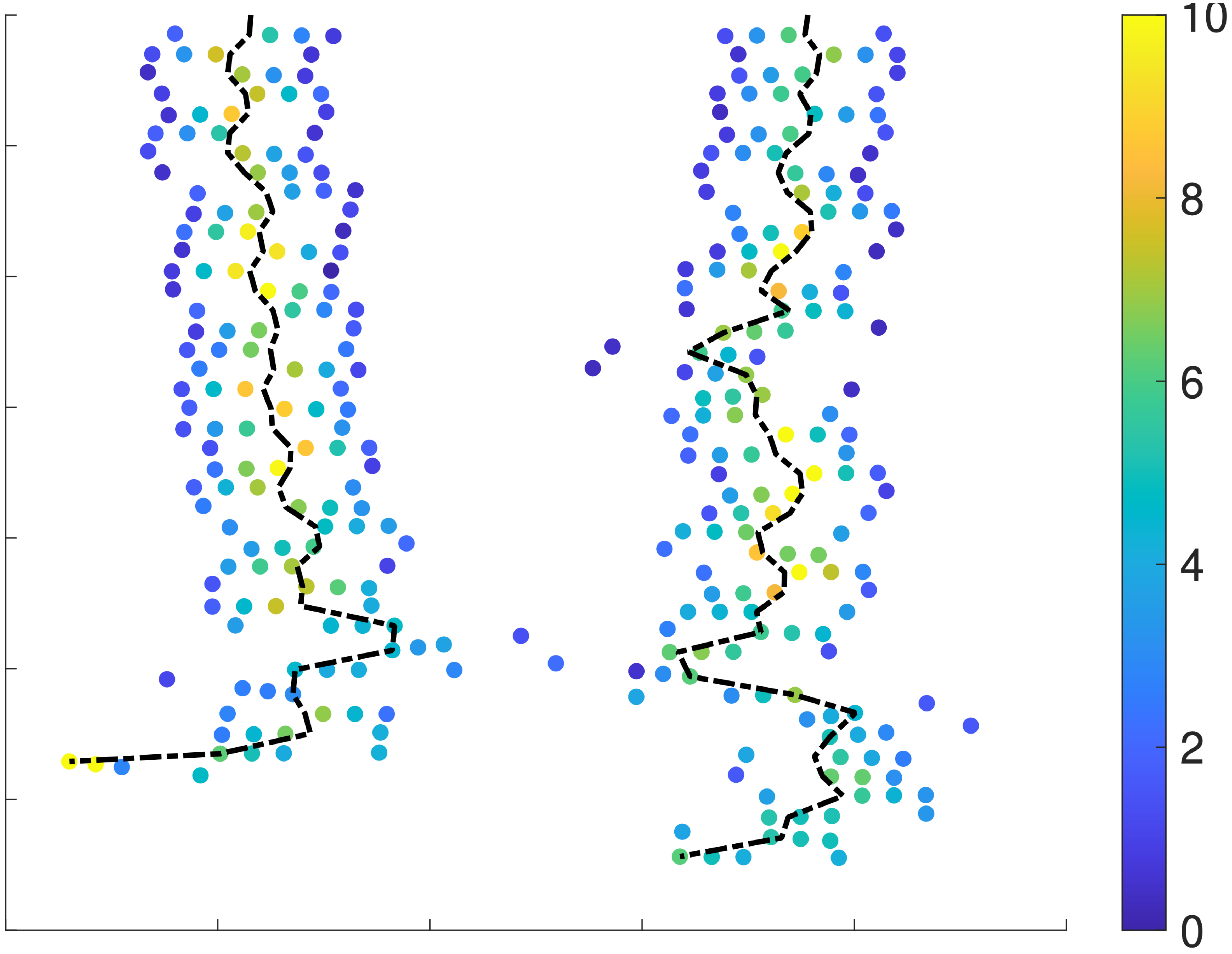
Example 1: fish skin model

- How does fish skin form their patterns?
- Example [Bullara & Decker] on-lattice model
- Stripe/Spots/Nothing?
- Depends on spatial extent and strength of interaction between cells



Fish skin model?





Horizontal: strength of interaction

Vertical: spatial extent of interaction

Wasserstein distance decreases when patterns are destroyed

Thanks! Questions?