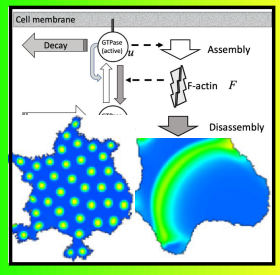
Inference and parameter identifiability for biological pattern formation

Yue Liu
Purdue University
Case Western Reserve Seminar
Jan.15.2026
liu4194@purdue.edu
https://liuyue002.github.io/

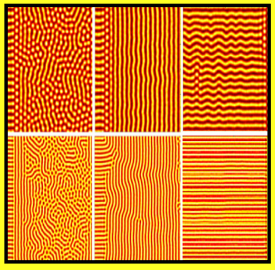
PURDUE UNIVERSITY

I: My journey in mathematical biology

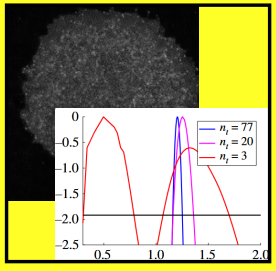
Pattern formation in models of cell polarisation



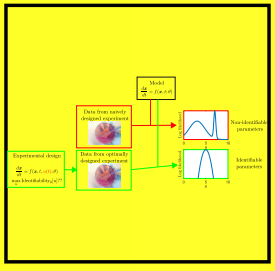
Pattern formation behind a wave of competency



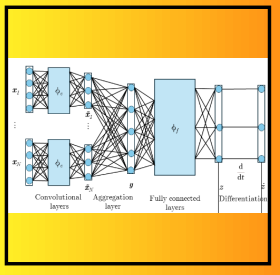
Parameter identifiability for PDE models of cell invasion



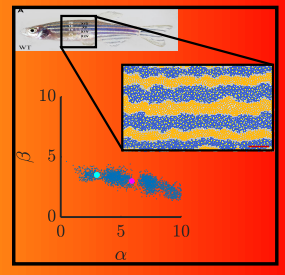
Optimal experiment design



Control of agent-based models with ML



Bayesian inference in ABM of zebrafish patterns via TDA



Conceptual PDE modelling

Data-driven PDE modelling

Data-driven approaches for agent-based models

Overarching theme:

- Biological patterns
- PDE models $\rightarrow$ Agent-based models (ABMs)
- Conceptual modelling $\rightarrow$ Data-driven approaches

I: My research program

Past:

- PDE models in signalling proteins regulating cell motility
- Diffusion-driven (Turing) instability in pattern formation
- Inference and identifiability of PDE models

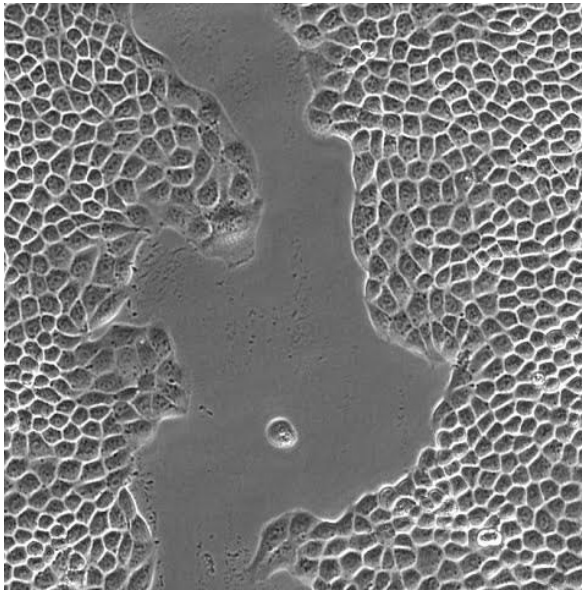
Current:

- Bayesian inference for ABM of cell-cell interaction and pattern formation with topological data analysis
- Machine learning-based dynamics discovery for cell movement during development
- Patterns in network models of opinion dynamics (undergrad mentorship)

Future:

- Combining PDE and ABM to build multi-scale model of immune response and wound healing
- Optimal therapy design with machine learning and control theory

I: Biological patterns are ubiquitous and fascinating



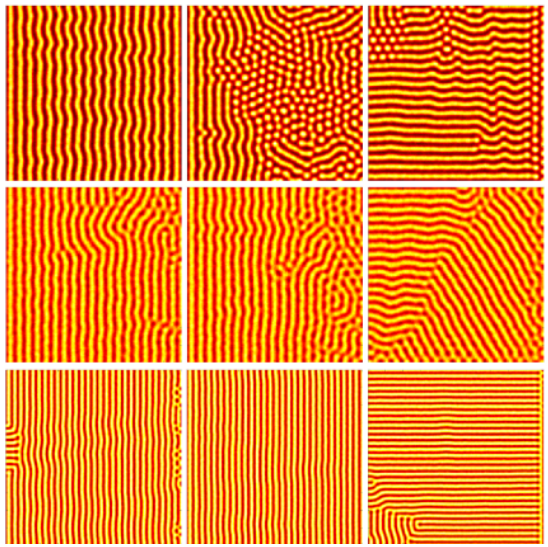
Epithelial Cells: Kozyrska et al, *Science*, 2022; Starlings: Baxter, *Wikipedia*, 2008;

Zebrafish: Azul, Wikipedia, 2005; Fern: Auer,1853.

Diffusion-driven instability was proposed by Turing (1952) as a possible driving mechanism behind many biological patterns

$$\frac{\partial u(x,t)}{\partial t} = \nabla \cdot (D_u \nabla u) + f(u,v),$$
$$\frac{\partial v(x,t)}{\partial t} = \varepsilon^2 \nabla \cdot (D_v \nabla u) + g(u,v),\ x \in \Omega \subseteq \mathbb{R}^n,\ t > 0$$
$$u(x,0) = u_0(x),\ v(x,0) = v_0(x,0),\ \partial_n u = \partial_n v = 0 \text{ on } \partial\Omega$$

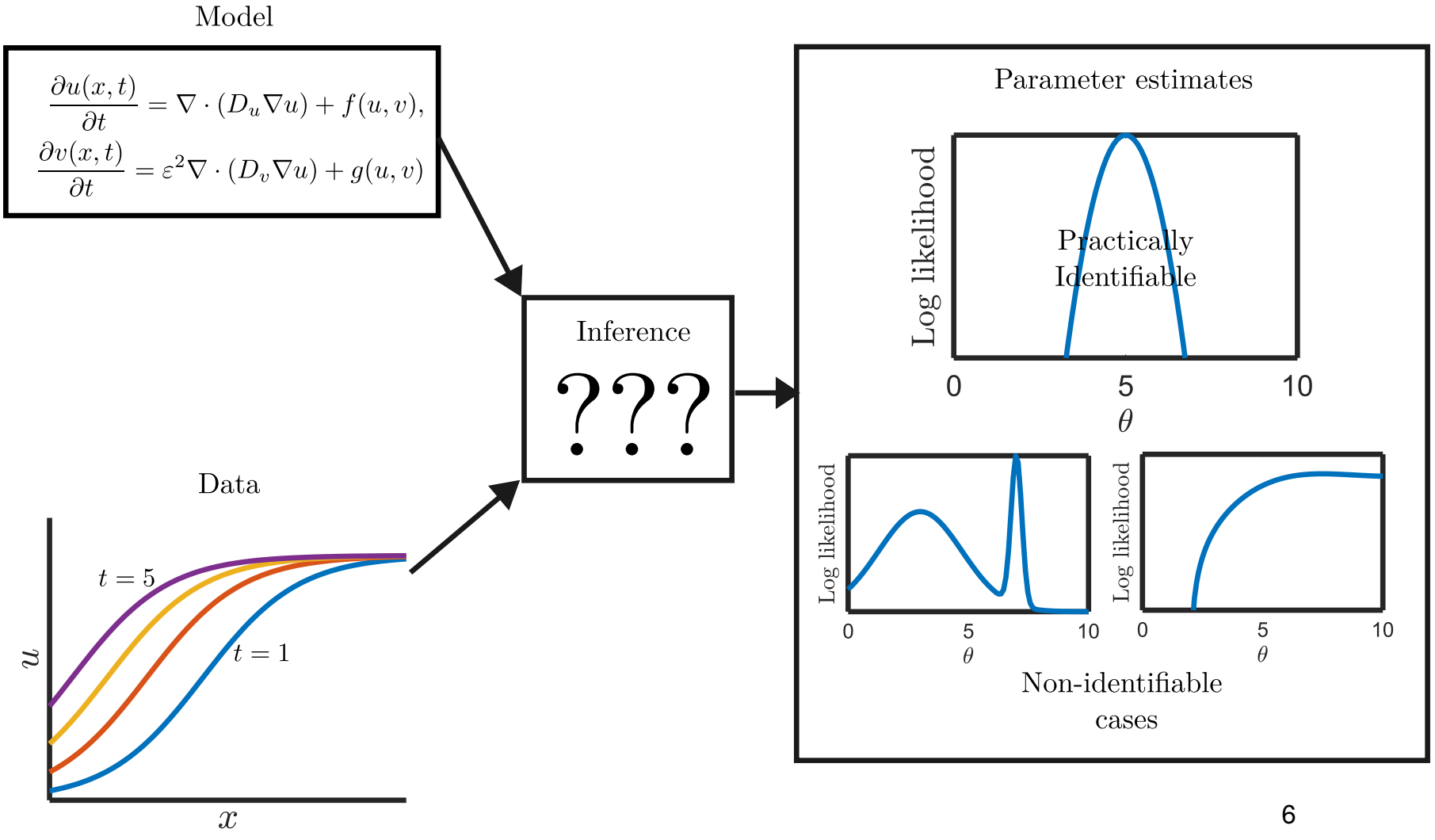
Many fascinating properties, but little evidence in biology



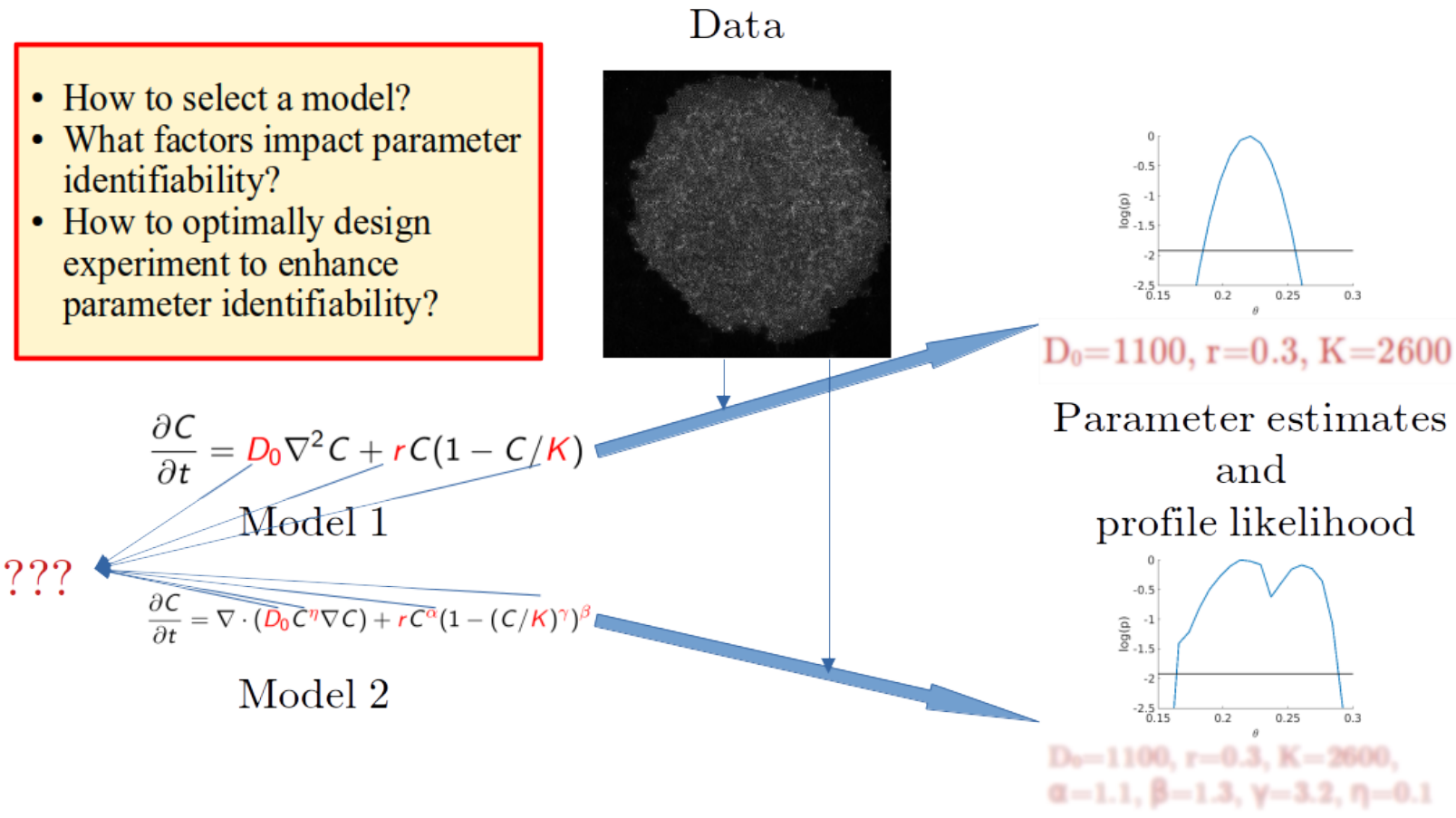
Liu, Maini, Baker, 2022. Control of diffusion-driven pattern formation behind a wave of competency. Physica D 438

I: Inference and Parameter identifiability

(Practical) Parameter identifiability: the ability to accurately infer values of model parameters with given data

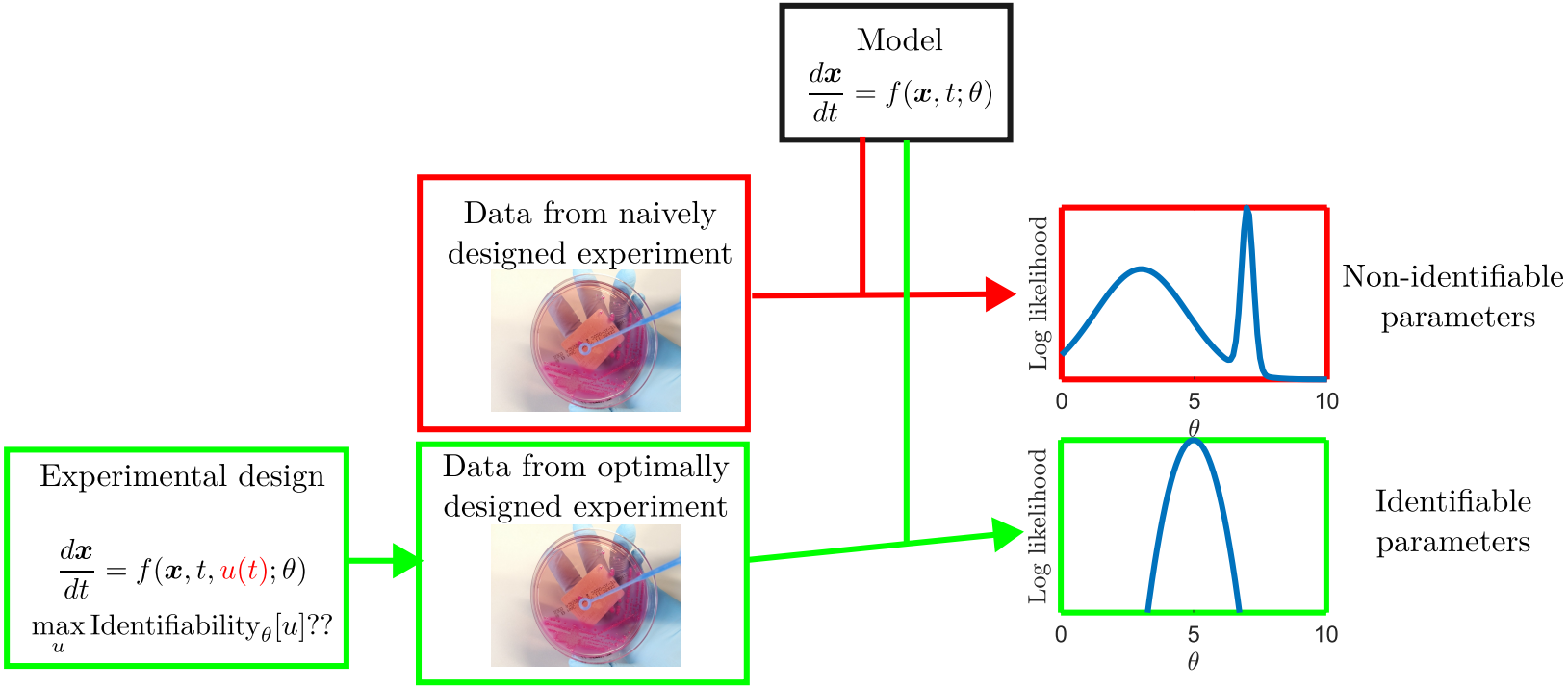


I: Data-driven PDE models of patterning



Liu, Suh, Maini, Cohen, Baker, 2024. Parameter identifiability and model selection for partial differential equation models of cell invasion. J R Soc Interface 21(212)

I: Experiment design



Liu, Maini, Baker, 2025. Optimal experiment design for practical parameter identifiability and model discrimination. arXiv:2506.11311

I: Pattern formation on zebrafish skin

zebrafish

About 1,310,000 results (0.07 sec)

SUBJECT REVIEWS | MAY 01 2008

Zebrafish as a Cancer Model FREE

Harma Feitsma; Edwin Cuppen

Check for updates

+ Author & Article Information

Mol Cancer Res (2008) 6 (5): 685–694.

The zebrafish: a new model organism for integrative physiology

Josephine P. Briggs

01 JAN 2002 // https://doi.org/10.1152/ajpregu.00589.2001

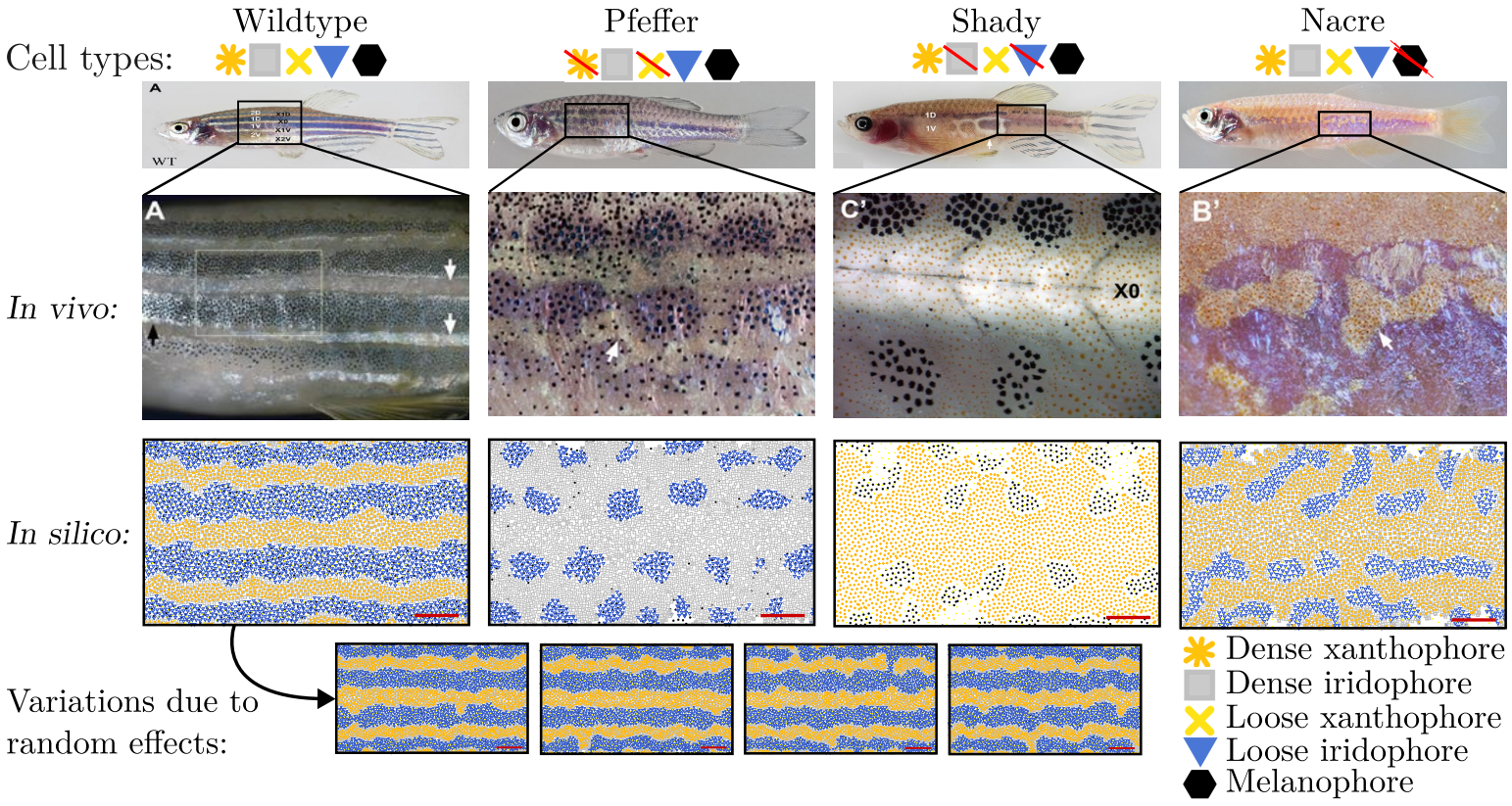
Studies of Turing pattern formation in zebrafish skin

Shigeru Kondo ✉, Masakatsu Watanabe and Seita Miyazawa

Published: 08 November 2021 https://doi.org/10.1098/rsta.2020.0274

Zebrafish photo: Azul, *Wikipedia*, 2005

I: The zebrafish ABM

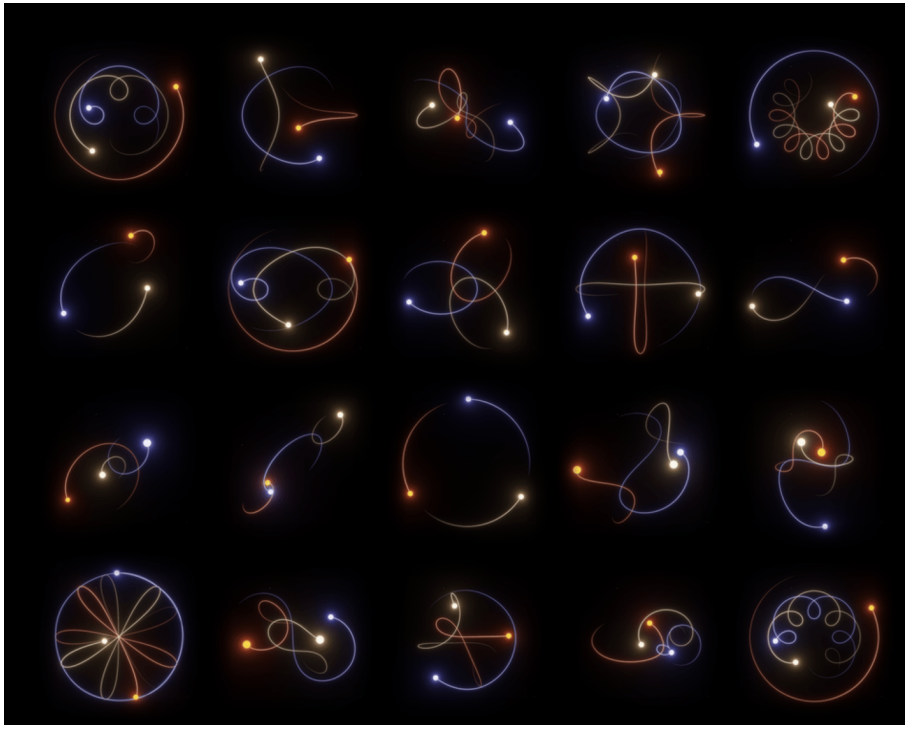


Fish and *in vivo* pictures: Frohnhöfer et al, *Development*, 2013. Model by Volkening & Sandstede, 2018

I: What are agent-based models (ABMs)?

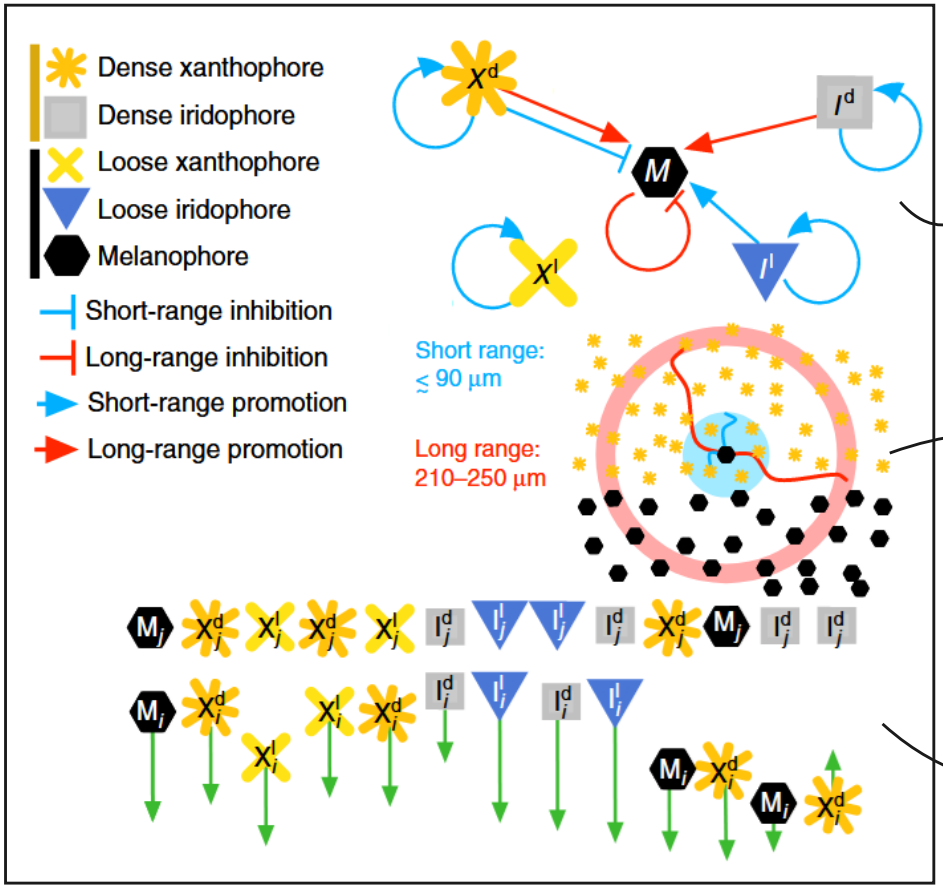
ABM: A collection of autonomous agents, and rules governing their interactions.

Example: Newtonian $n-$body problem



$$\frac{d^2x_i}{dt^2} = \sum_{j \neq i} -G \frac{m_i m_j}{|x_i - x_j|^2} \widehat{x_i - x_j}$$

I: The zebrafish model is more complex



Cell type transition rule example:

$I \longrightarrow I^{\mathrm{d}}$ if

$\#_{90} M < c$ and $[\#_\Omega X^d < d$ or $\#_{75} X^d > e]$

(poorly understood)

Cell differentiation rule example:

$\varnothing \longrightarrow M$ at z if

$\#_\Omega X^d + \#_\Omega I^{\mathrm{d}} > \alpha + \beta \#_\Omega M$

(better understood)

Movement rule example:

$$\frac{d\boldsymbol{X}_i}{dt} = -\sum_j \boldsymbol{F}_{ij},$$

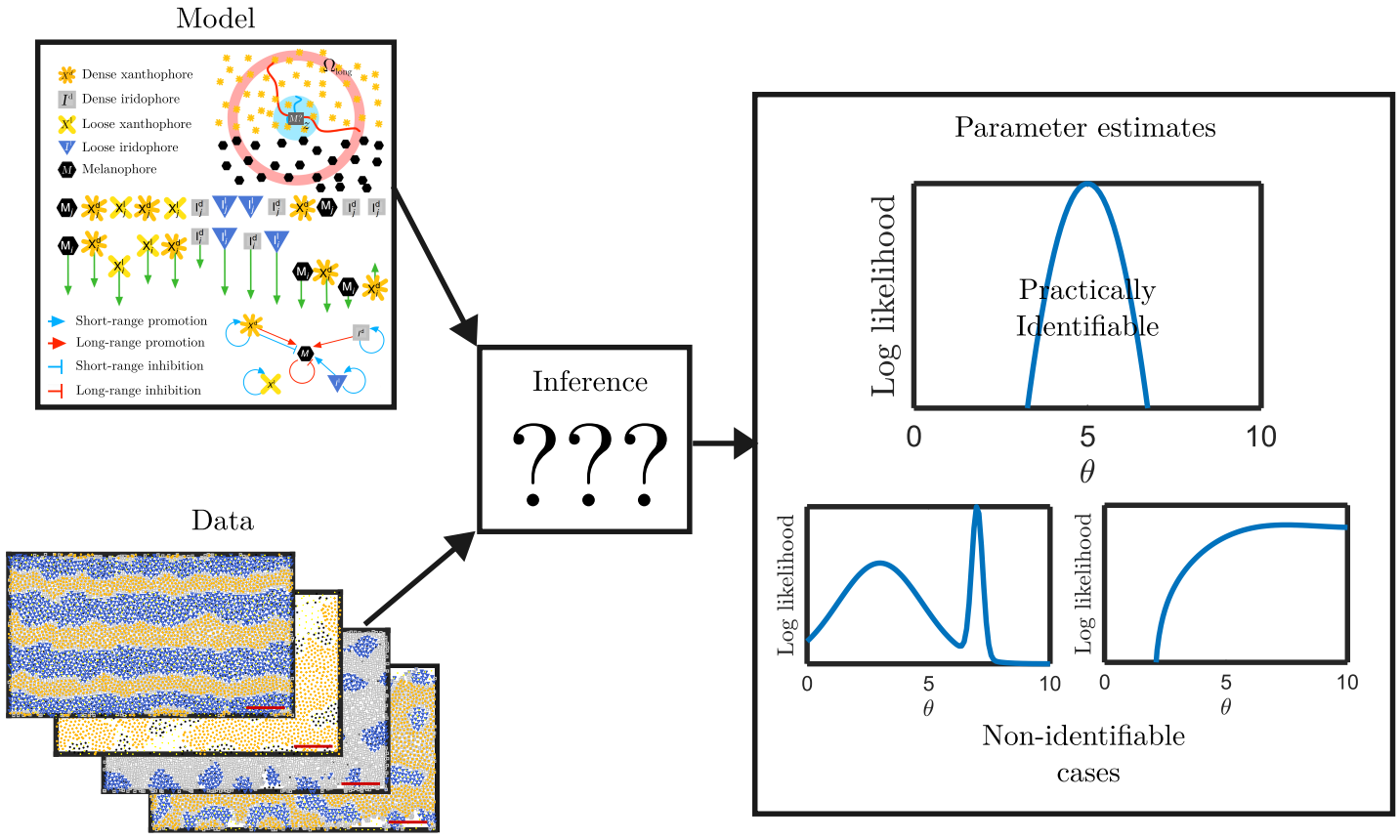
$$\boldsymbol{F}_{ij} = R_{I^d, I^d}\left[\frac{1}{2} + \frac{1}{2}\tanh\left(\frac{r_0 - ||\boldsymbol{r}||}{\gamma}\right)\frac{\boldsymbol{r}}{||\boldsymbol{r}||}\right]$$

(well understood)

Figure adapted from Volkening & Sandstede, *Nat. Comm.*, 2018

I: Parameter identifiability

Practical parameter identifiability: a quantification of uncertainty in parameter estimates with respect to data quality and quantity



I: The central aim

Biological question:
Does there exist alternative rulesets capable of reproducing the patterns?

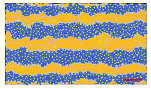
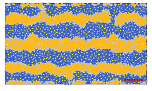
Spoilers: yes

Mathematical challenge:
How to perform parameter inference and identifiability analysis for such a complex, stochastic, model?

Challenge 1: Intractable likelihood, expensive simulations

$$L(\theta) = P(x_{\text{obs}}|\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x_{\text{obs}} - x(t_i))^2}{2\sigma^2}\right]$$

Challenge 2: No straight-forward way to compare patterns

$d($  ,  $) = ???$

AABC

TDA

Robust, computationally efficient inference pipeline

Talk outline

I Background and motivation

II Bayesian Inference

III Topological data analysis

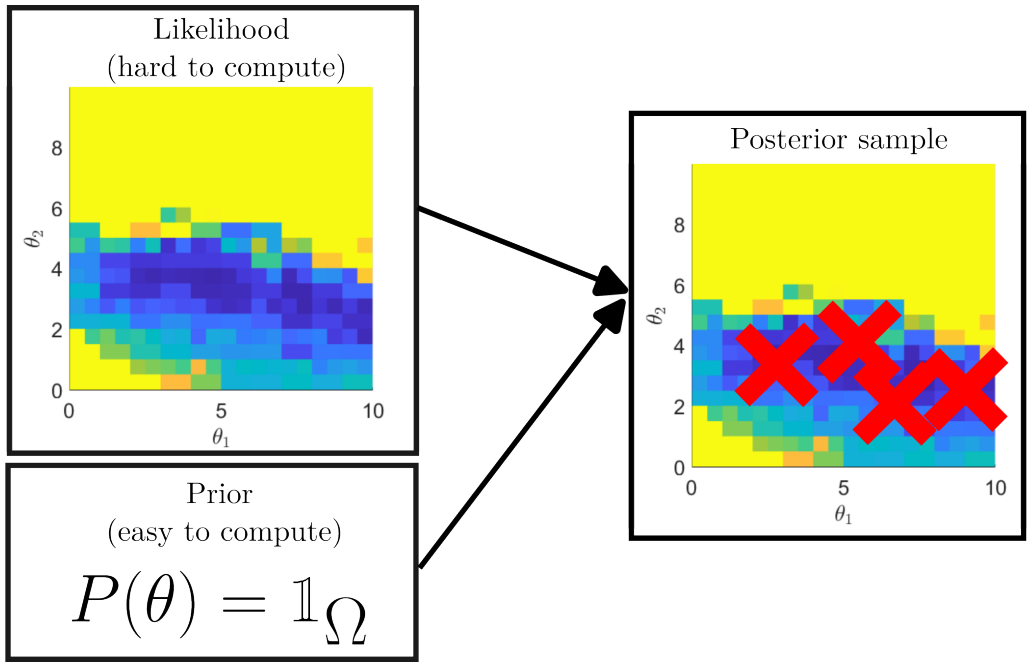
IV Inference result for the zebrafish model

V Future work

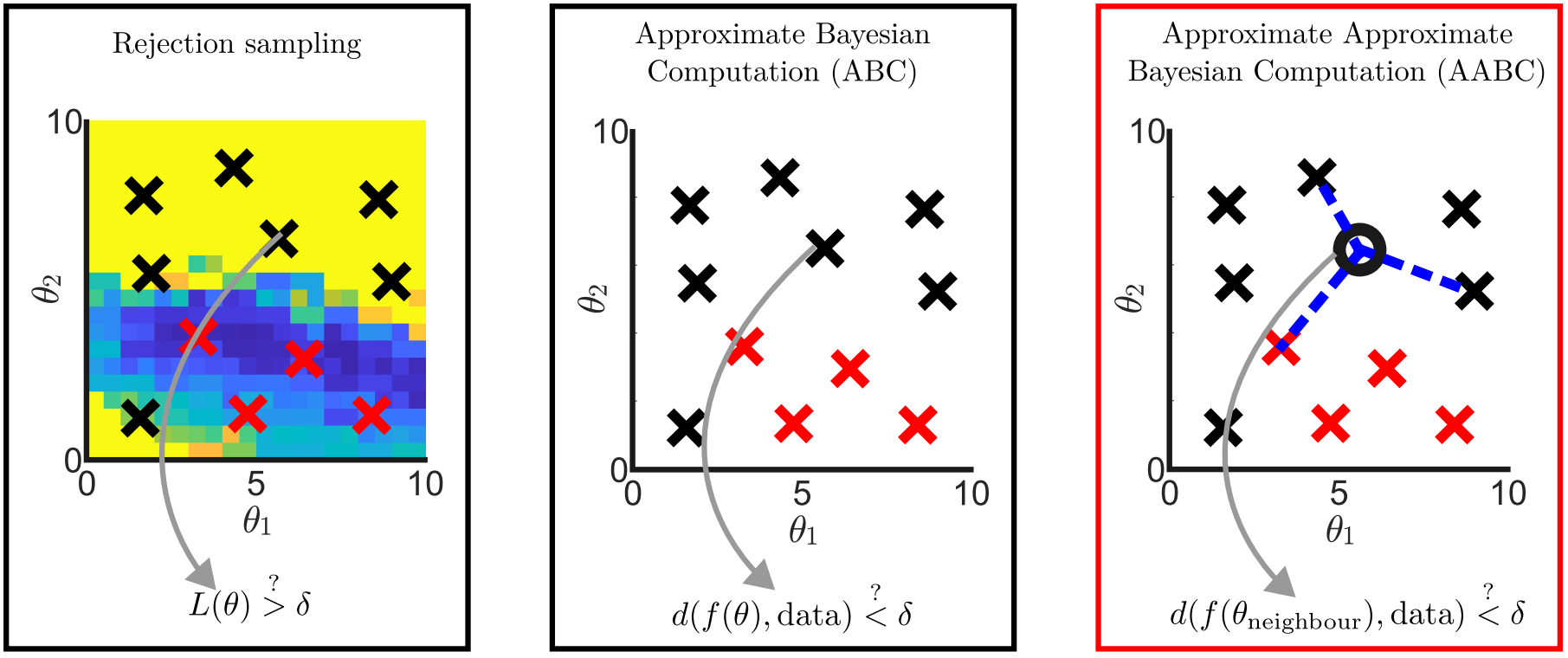
II: Bayesian parameter inference

$$\underbrace{P(\theta|x_{\text{data}})}_{\text{posterior}} \propto \underbrace{P(x_{\text{data}}|\theta)}_{\text{likelihood}} \underbrace{P(\theta)}_{\text{prior}}$$

Goal: Obtain the *posterior distribution* of parameter values by (approximately) sampling from it



II: From rejection sampling to AABC



Higher accuracy
More expensive

Lower accuracy
Computationally cheaper

Problem: how to define $d(\cdot, \cdot)$ and choose appropriate δ?

Talk outline

I Background and motivation

II Bayesian Inference

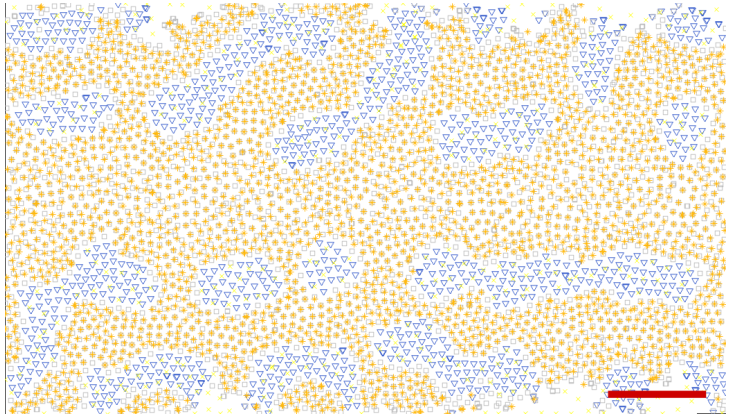
III Topological data analysis

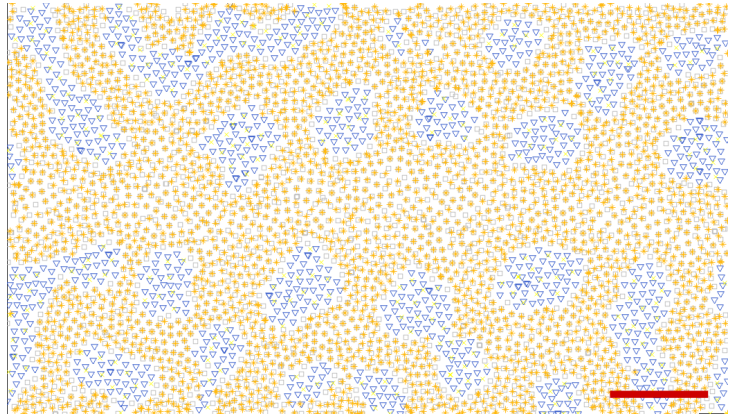
IV Inference result for the zebrafish model

V Future work

III: How to compare patterns?

Pixel-wise comparison does not respect the qualitative “essence" of patterns

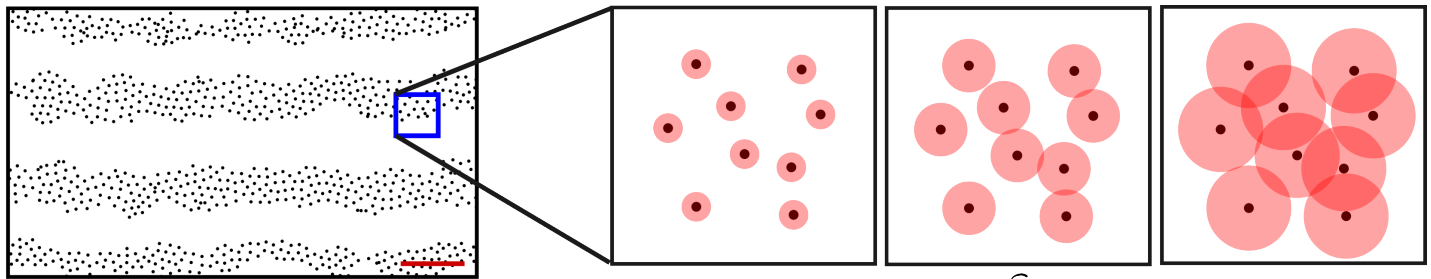




Common approaches for comparing spatial data:

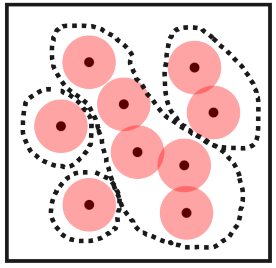
- Pair correlation functions
- Summary statistics
- Topological data analysis: good for summarising topological/geometric information

Step 1: Compute persistent homology
(we use the Vietoris–Rips filtration)

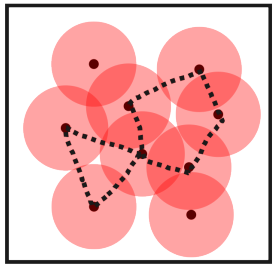


Point-cloud data

Persistent homology

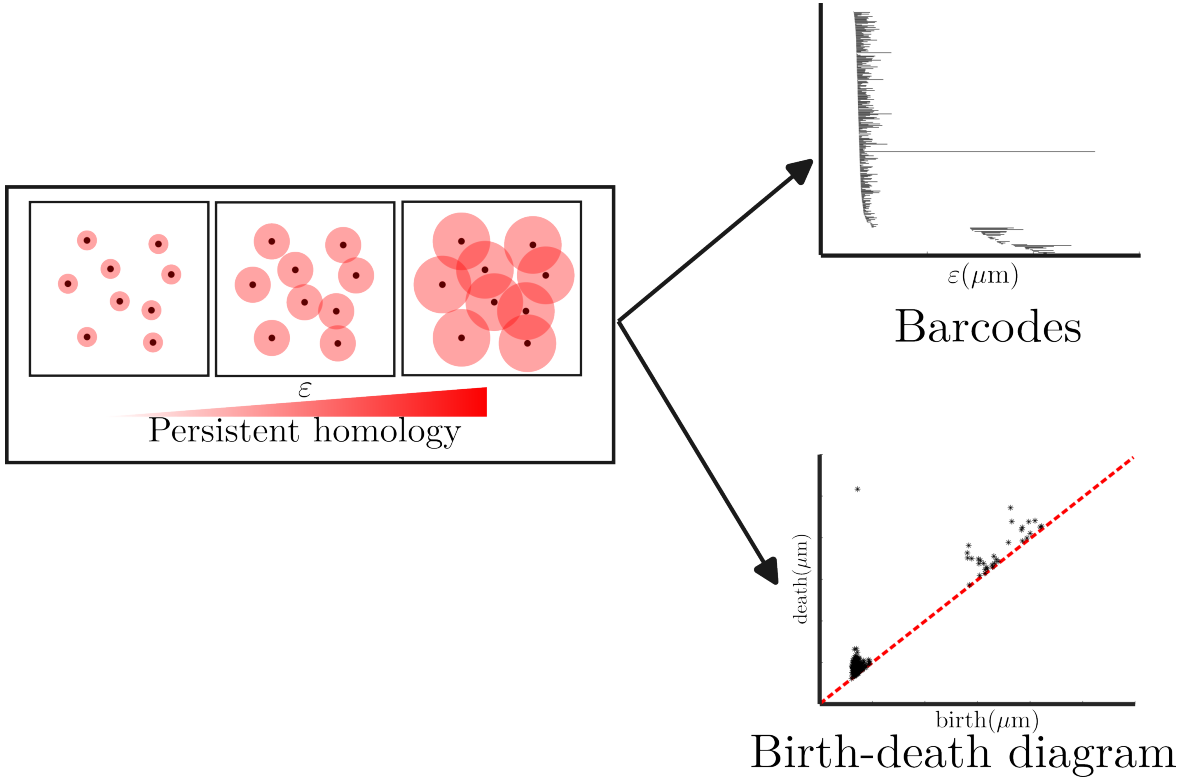


Dimension 0 features:
connected components



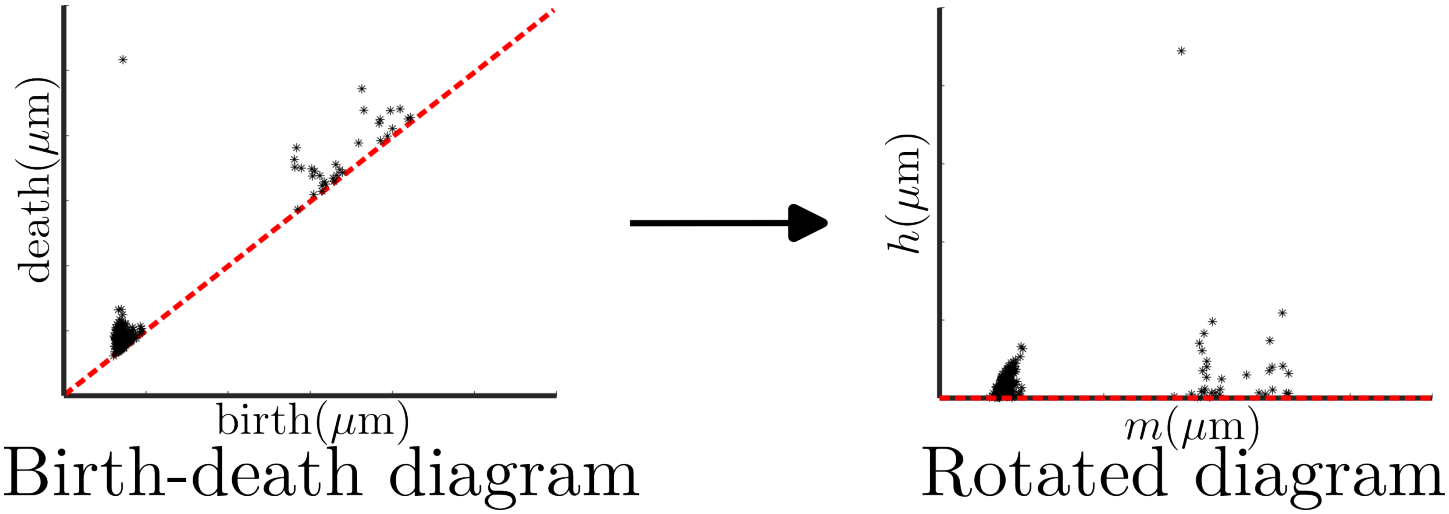
Dimension 1 features: loops

Step 2: Pick a dimension (usually 0 or 1), compute barcodes and birth-death diagrams

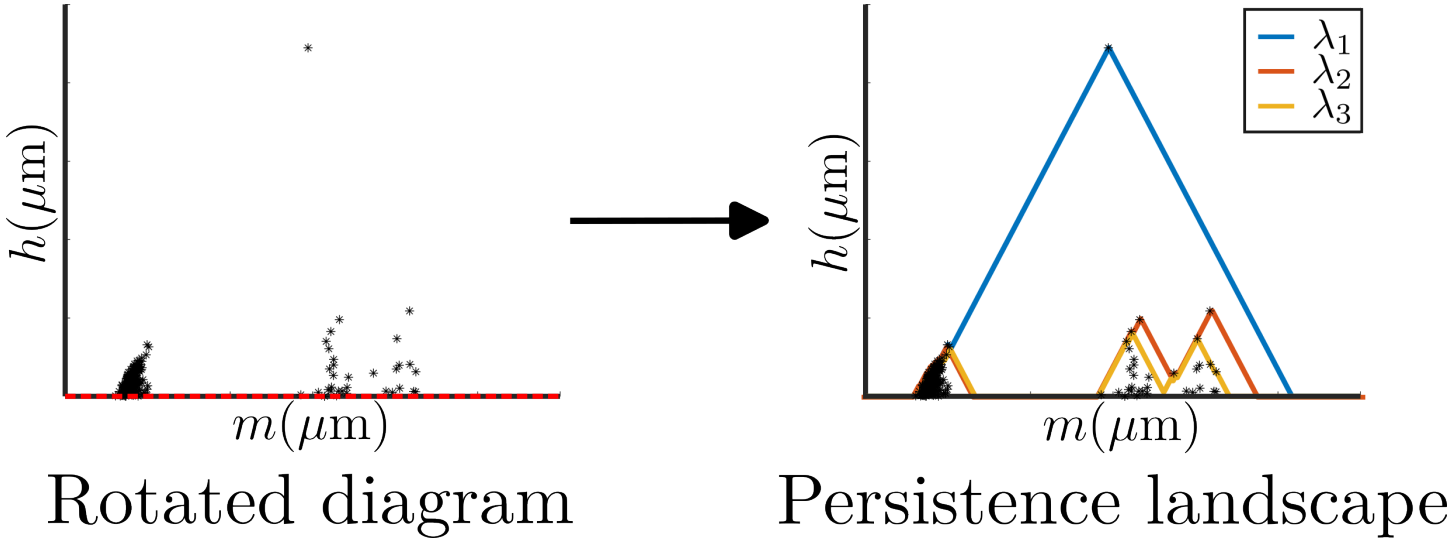


But directly comparing barcodes from different patterns is difficult
$\rightarrow$ persistence landscape

Step 3a: Rotate birth-death diagrams by $\pi/4$



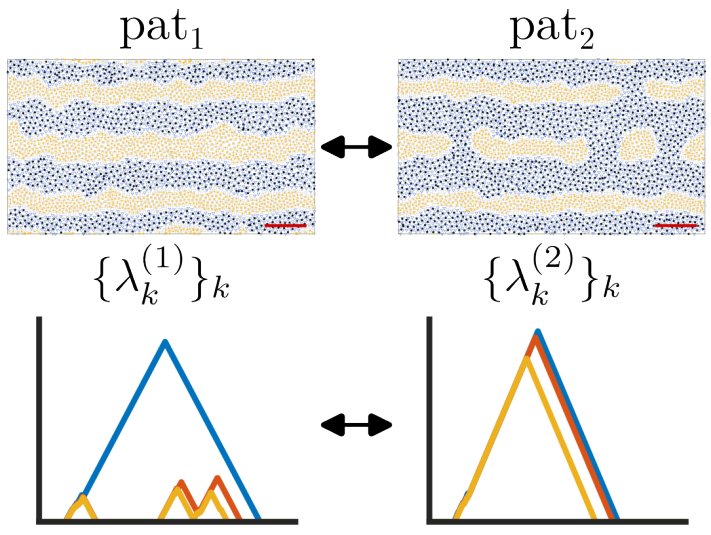
Step 3b: Draw triangular “mountains"



Persistence landscape is the collection of envelopes defined by the “mountains", $\{\lambda_k\}_{k=1}^{\infty}$

III: Metric for comparing patterns

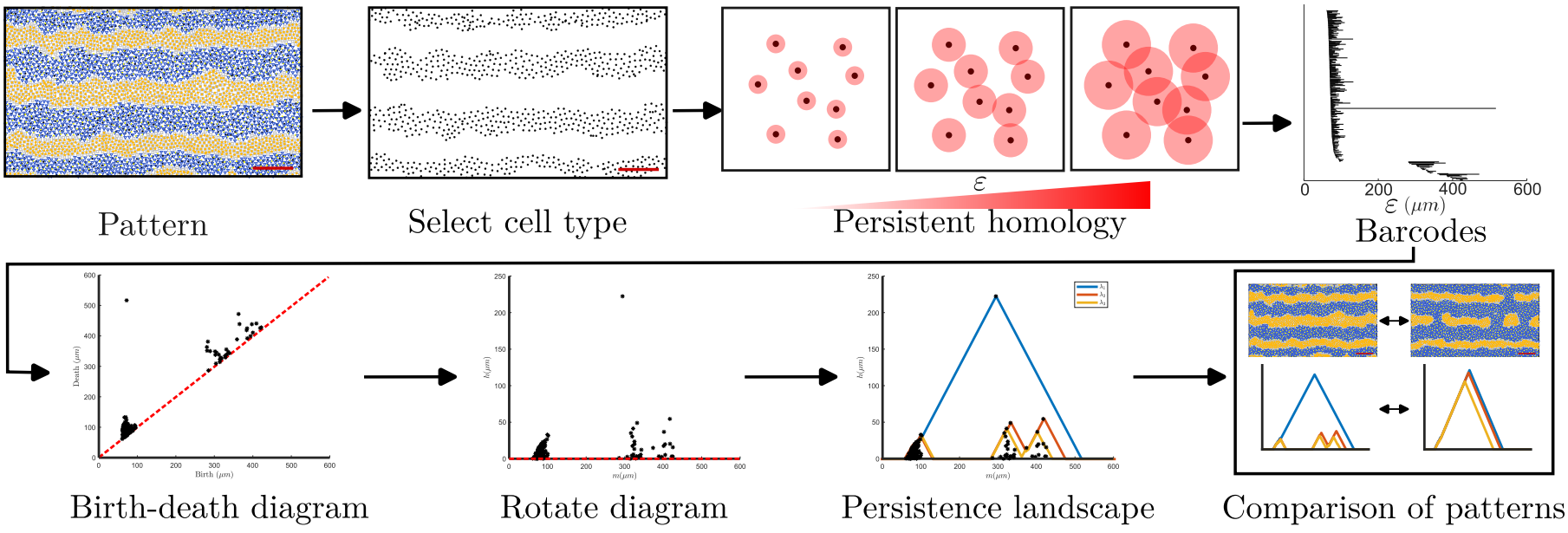
Persistence landscape offers a robust way to compare spatial data (Bubenik 2015, 2020), and consequently define score function for parameter values



$$d_{\mathrm{land}}(\mathrm{pat}_1, \mathrm{pat}_2)^2 = \sum_{k=1}^{\infty} d_{L^2}\left(\lambda_k^{(1)}, \lambda_k^{(2)}\right)^2$$

$$D(\theta) = d_{\mathrm{land}}(\mathrm{pat}(\theta), \mathrm{pat}_{\mathrm{data}}), \quad \mathrm{pat}_{\mathrm{data}} \text{ is synthetic}$$

III: The TDA pipeline

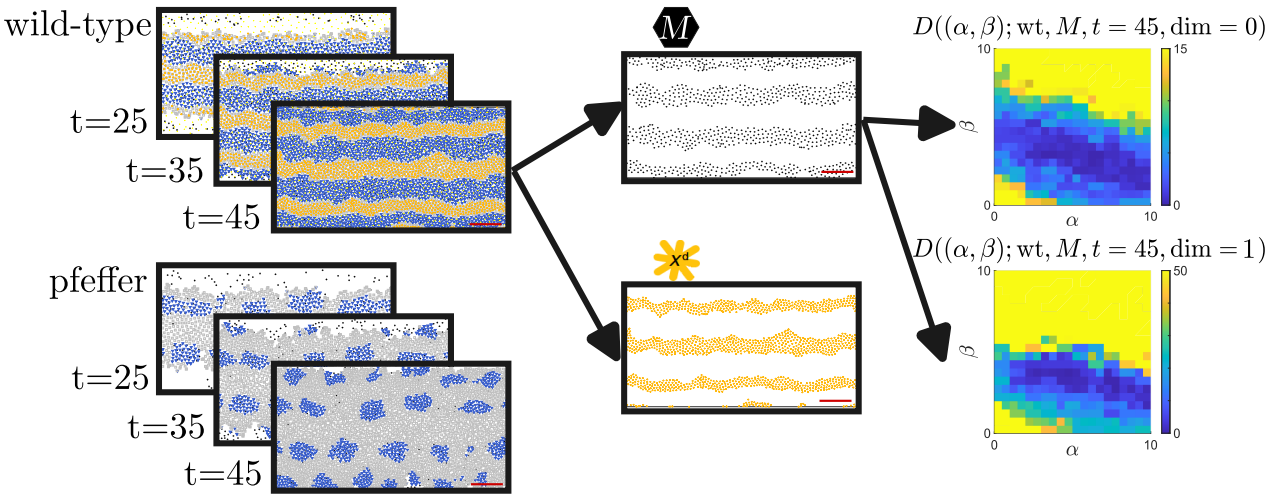


III: Choices in building a score function

Any choice of

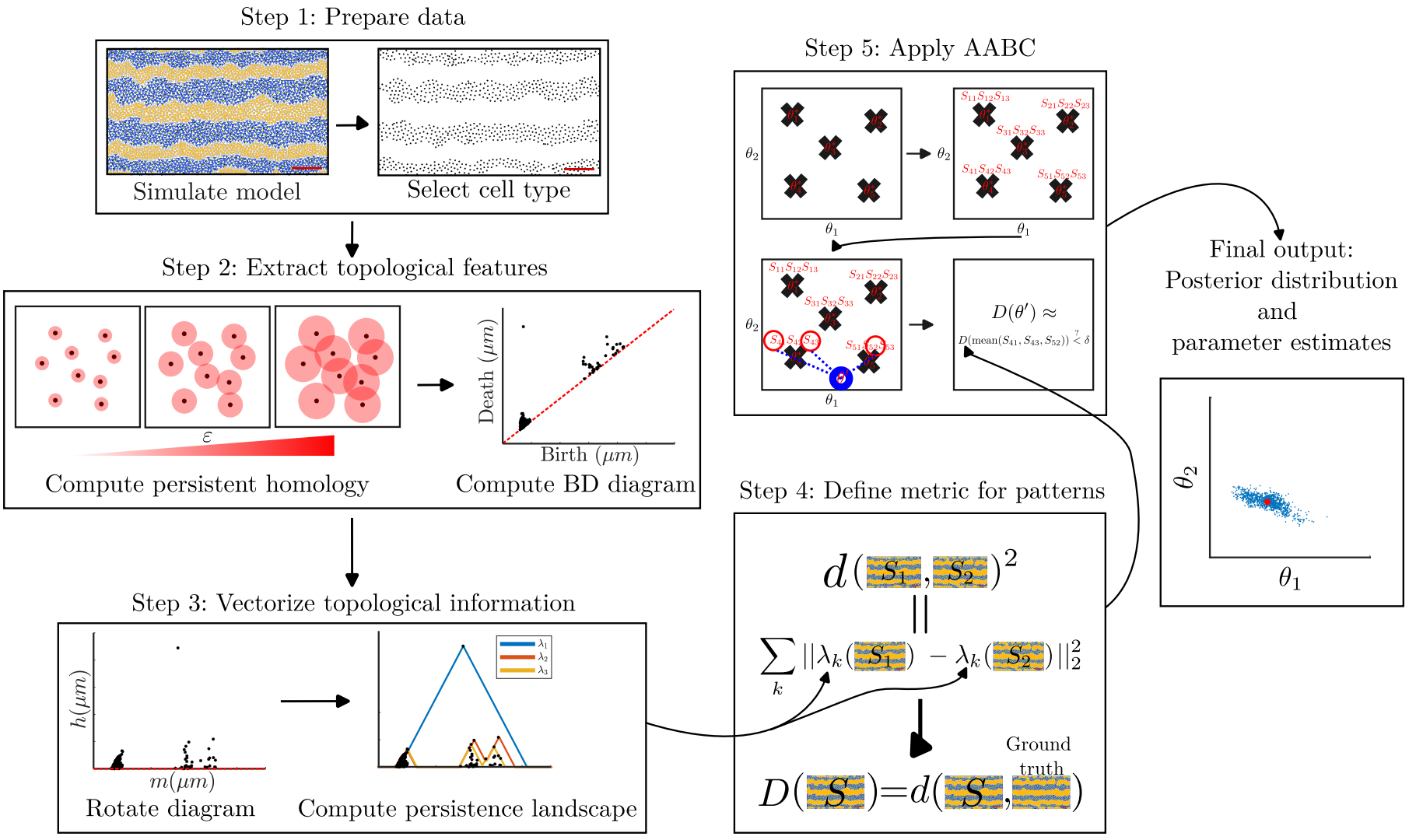
- Fish type (wild-type, *pfeffer, nacre, shady*),
- Cell type (M, X^d, X^l, I^d, I^l),
- time $t = 1, \ldots, 44, 45$,
- Topological dimension (0 or 1),

yields a distinct distance function $D((\alpha, \beta); \text{fish}, \text{cell}, t, \text{dim})$:



We can either use one such D individually, or combine them somehow (more on this later)

III: The entire pipeline



Talk outline

I Background and motivation

II Bayesian Inference

III Topological data analysis

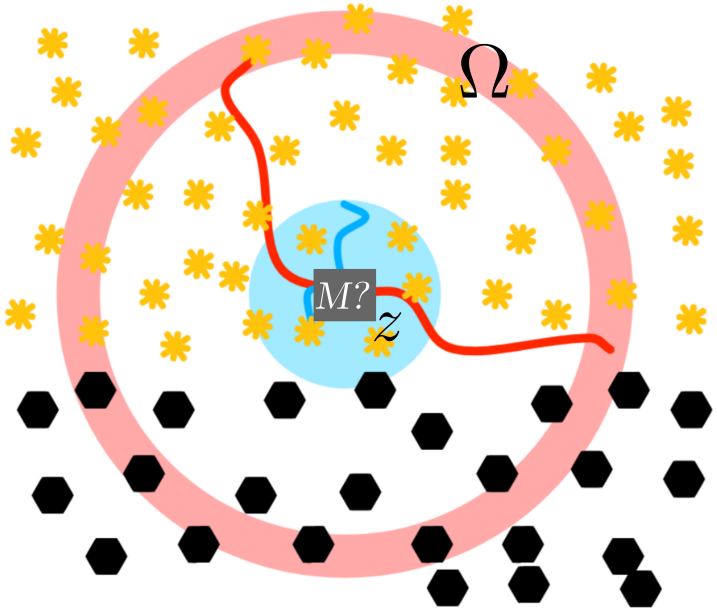
IV Inference result for the zebrafish model

V Future work

IV: Demonstration of methods on well-understood parameters

A new M at random location z may appear if

$$\#\Omega^{x^d} + \#\Omega^{I^d} > \alpha + \beta\, \#\Omega^{M}$$



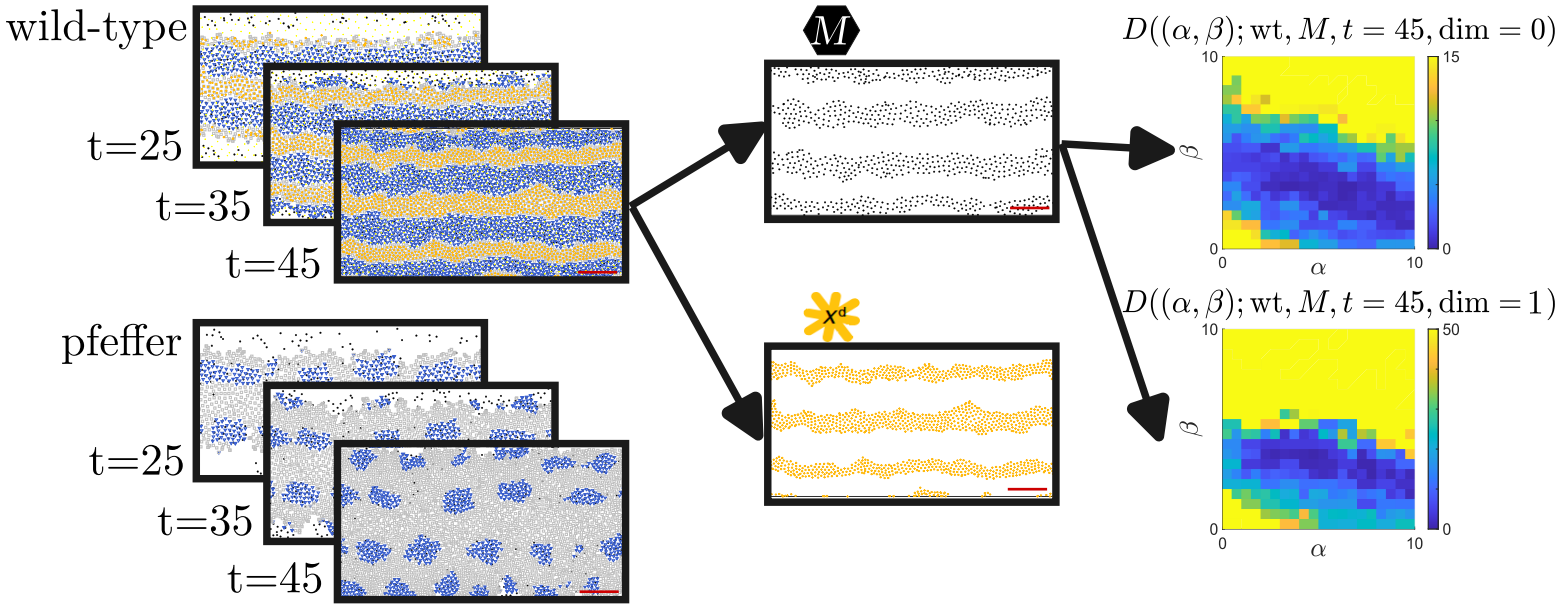
Parameters: α, β

IV: Recall the choices

Any choice of

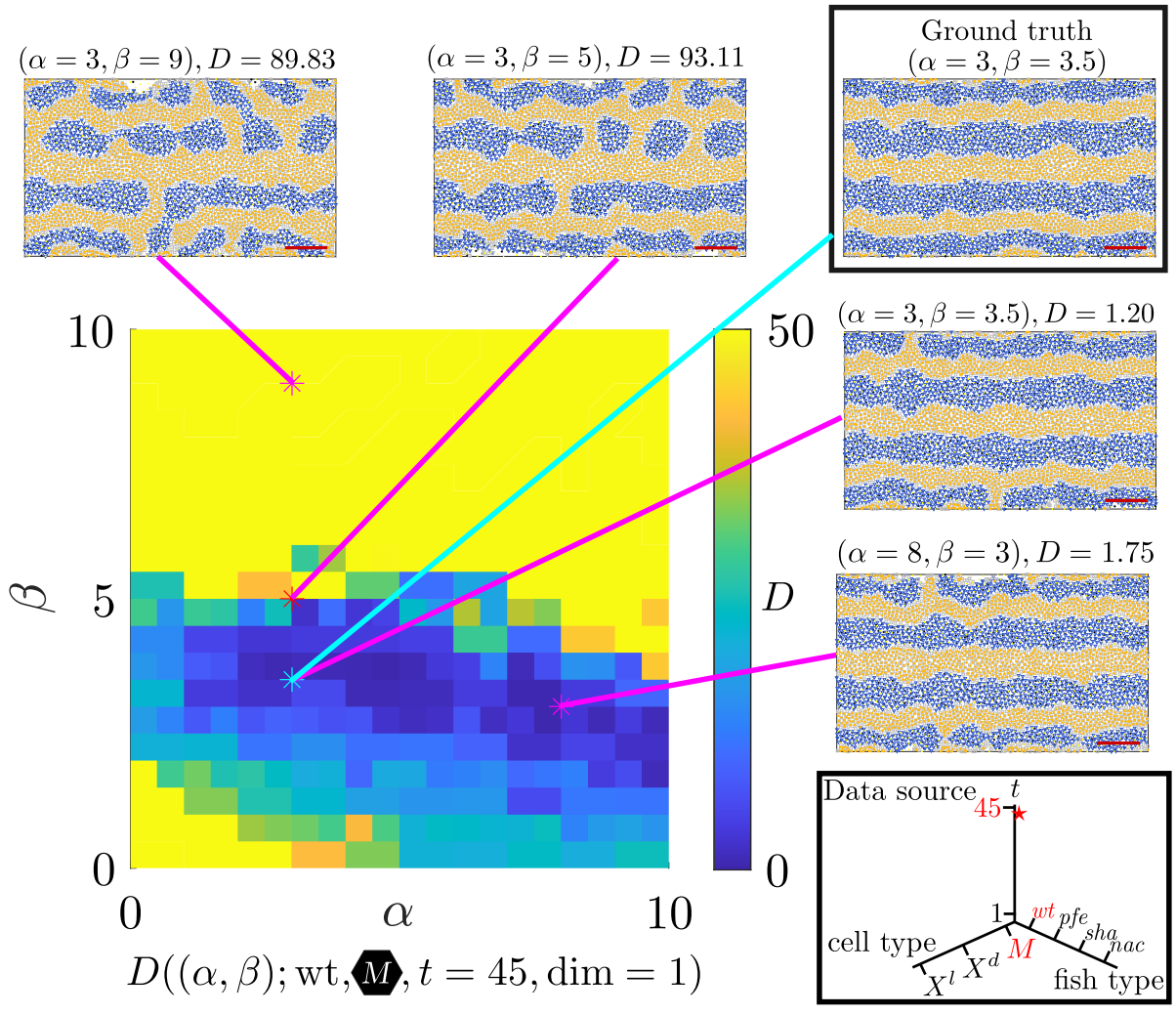
- Fish type (wild-type, *pfeffer, nacre, shady*),
- Cell type (M, X^d, X^l, I^d, I^l),
- time $t = 1, \ldots, 44, 45$,
- Topological dimension (0 or 1),

yields a distinct distance function $D((\alpha, \beta); \text{fish}, \text{cell}, t, \text{dim})$:



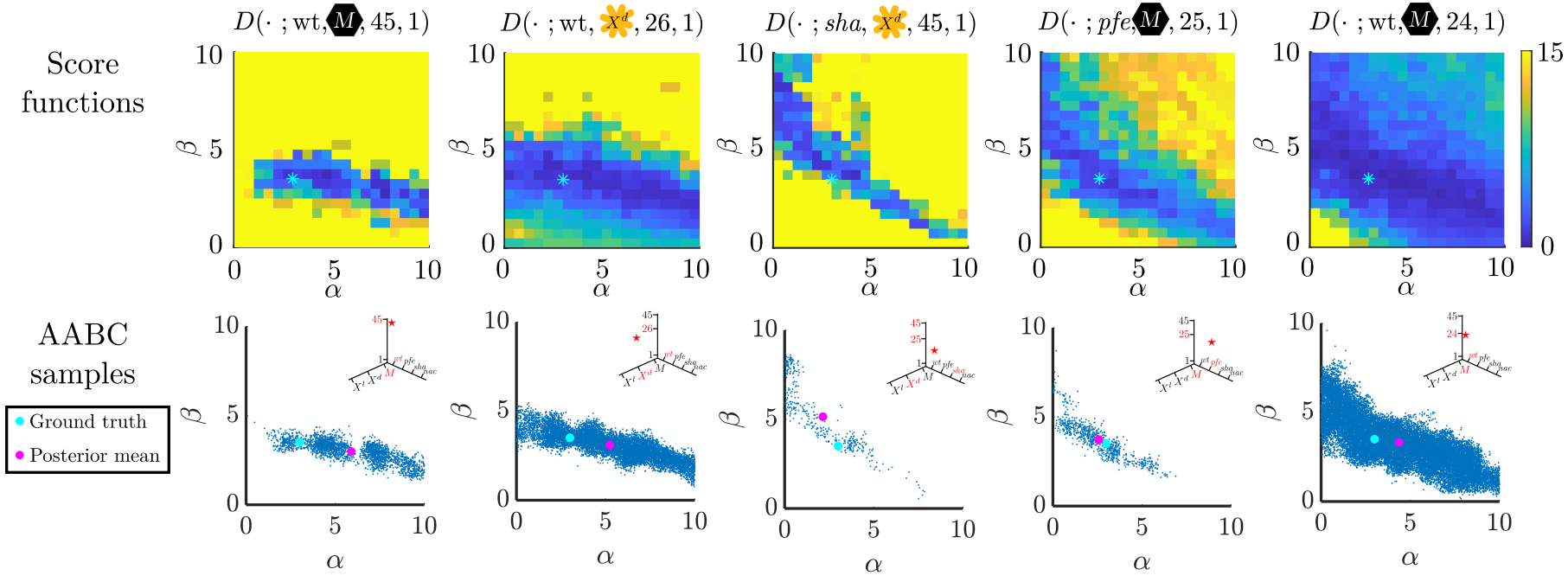
IV: Examining one particular score function

The score function captures qualitative characteristics



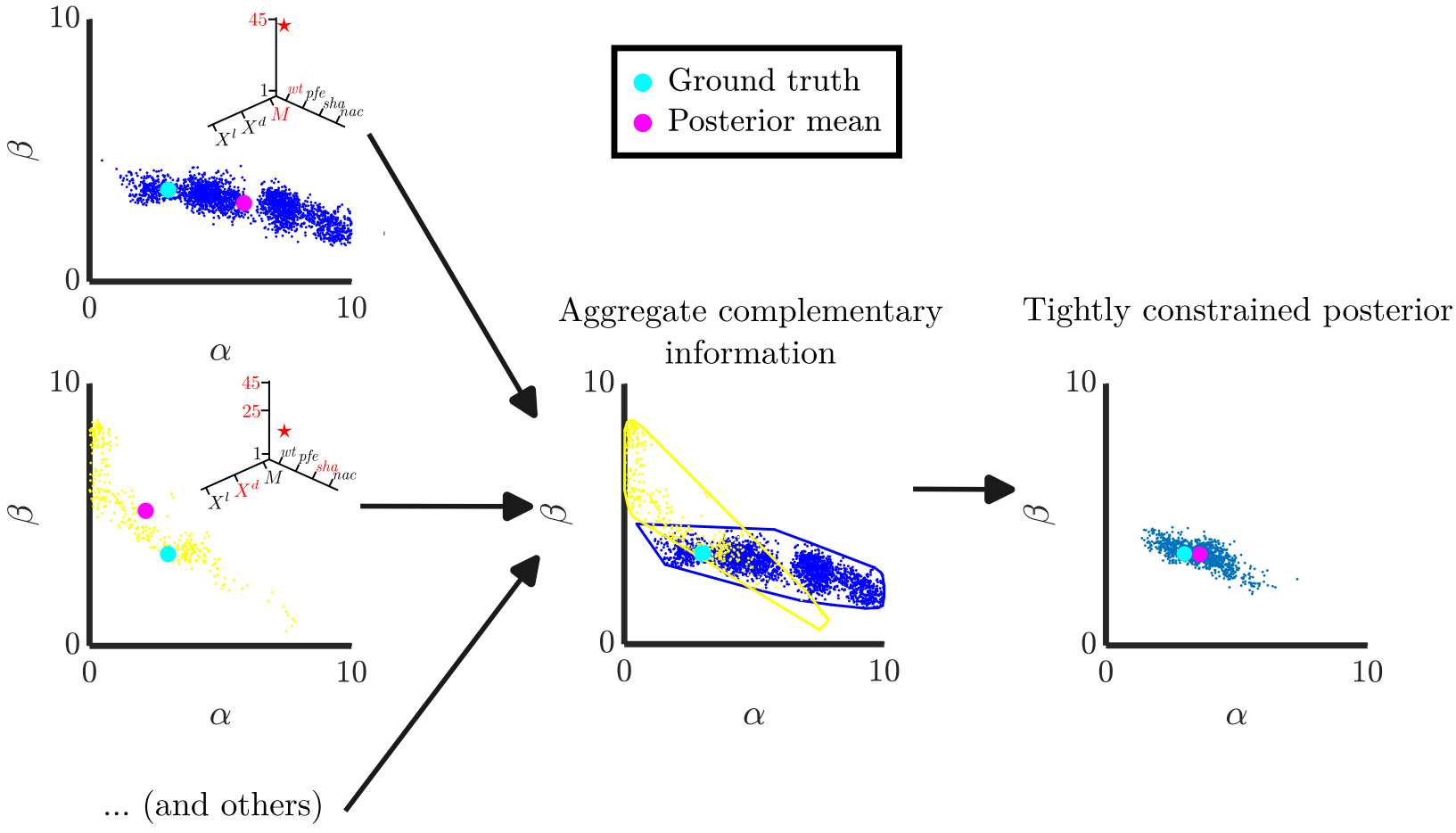
IV: Difference between score functions

No single score function is sufficiently informative to provide practical identifiability



IV: Combining score functions

Combination of score functions provides practical identifiability

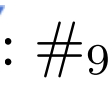
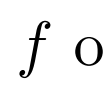


Many ways to combining multiple slices of information, but in this case the result is similar

V: Inference results for poorly-understood parameters

 and can transition to each other if:

 $\rightarrow$  : $\#_{90}$  $< c$ and $[\#_{\Omega}$  $< d$ or $\#_{75}$ $> e]$

 $\rightarrow$  : $\#_{90}$  $> f$ or $[\#_{\Omega}$  $> g$ and $\#_{75}$  $< h]$

Parameters: c, d, e, f, g, h

Biological question:

Are all six of these interactions necessary to produce observed patterns?

V: Approaches for combining information

How to combine the information encoded in each individual score function $D^{(i)}(\theta)$?
Approach 1: Weighted sum: accept θ if

$$\sum_i w_i D^{(i)}(\theta) < \delta$$

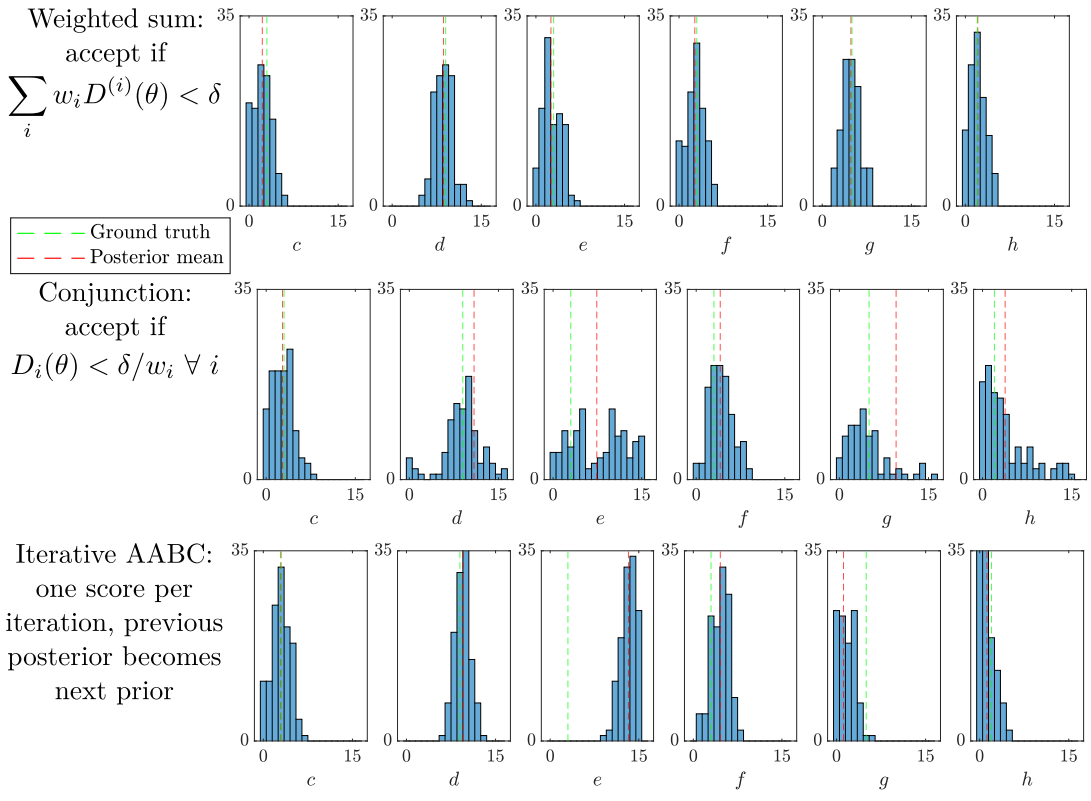
Approach 2: Conjunction: accept θ if

$$D^{(i)}(\theta) < \delta / w_i \ \forall \, i$$

Approach 3: Iterative AABC: Multiple rounds of inference with updated prior

V: Combining information

Different approaches for combining information lead to different outcome



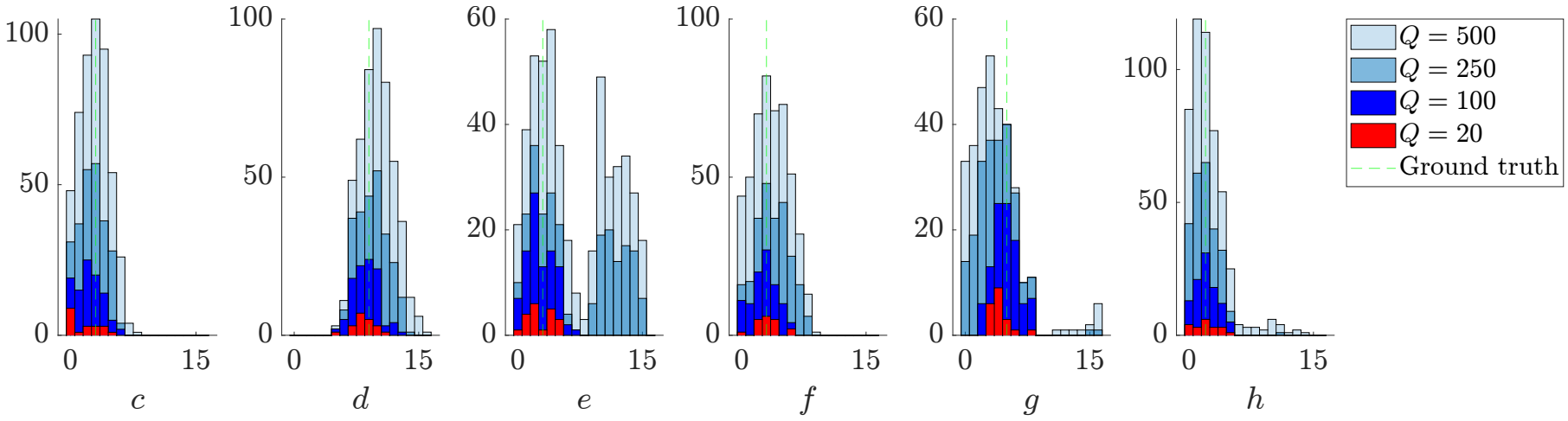
Apparently identifiable: c, d, f, h; Apparently non-identifiable: e, g

So which one is right?

Answer: posterior predictive check

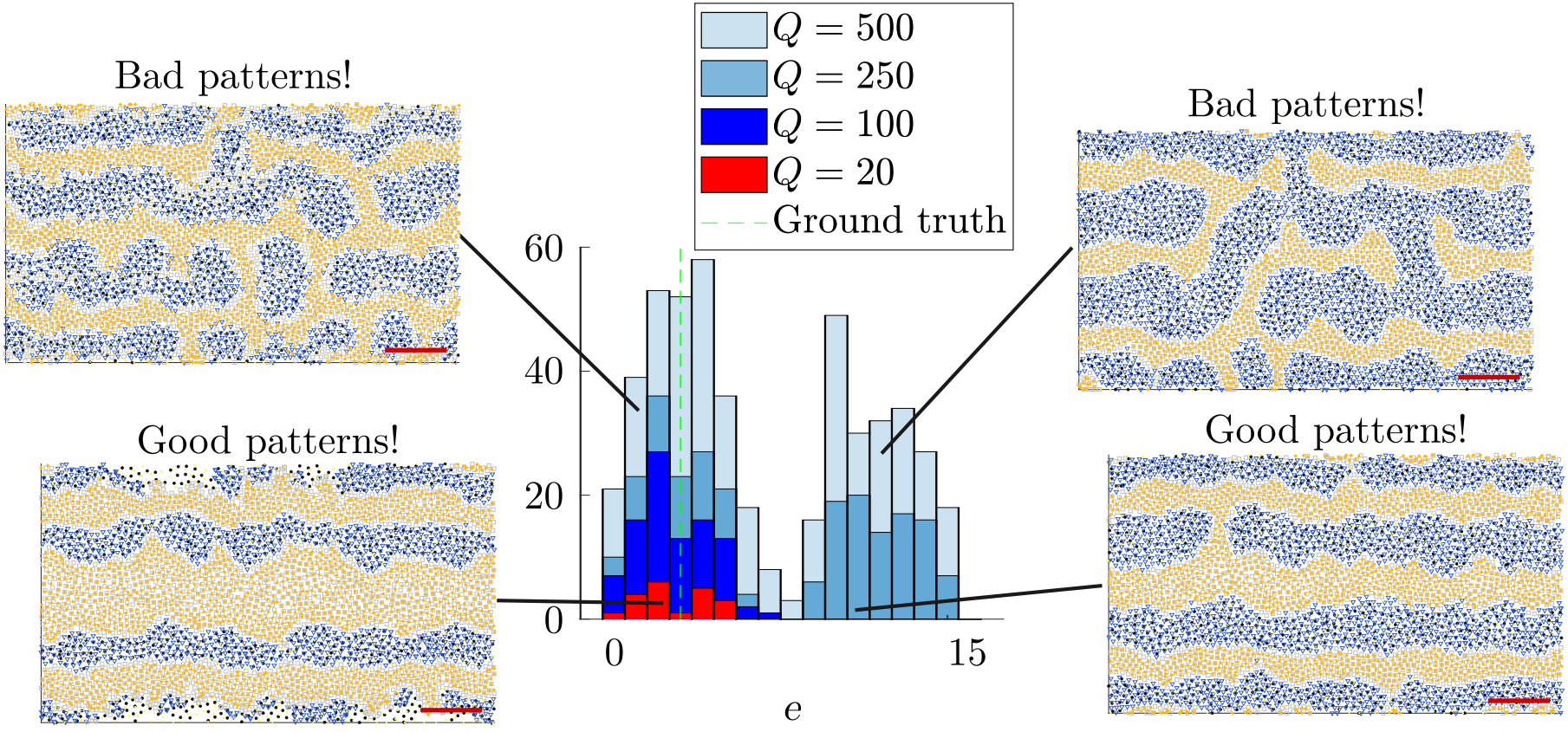
V: Hyperparameter tuning for δ

Posterior obtained using weighted sum, varying δ:



$Q =$ quota for number of samples to accept (out of 10^6).
Larger $Q \Leftrightarrow$ larger δ
Substantiate difference in posterior as δ changes!
Difference in posterior outcomes can be explained by hyperparameter tuning

V: Posterior predictive check



Q : number of parameter samples to accept

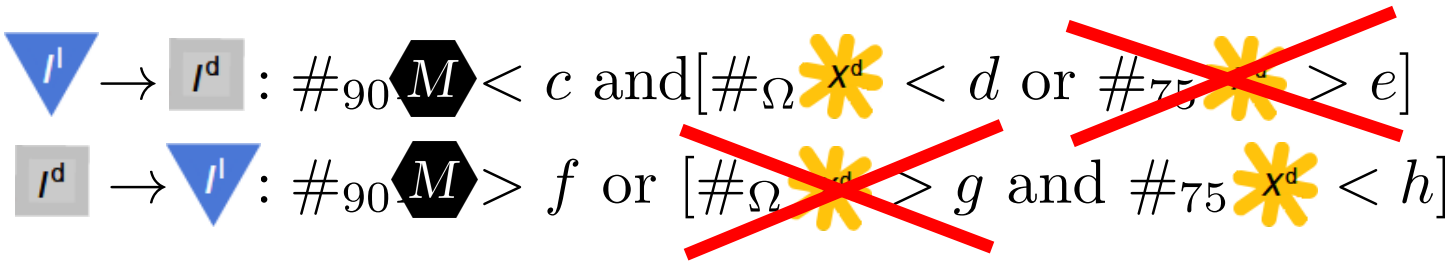
$Q = 20, 100$: rejected too many good parameters

$Q = 500$: accepted too many bad parameters

$Q = 250$: just right

V: Biological insight

- Bimodal posterior distribution⇒ non-identifiable parameters
- There exist parameter sets with very high values of *e* and *g*, effectively turning off the corresponding rules, but still capable of reproducing data ⇒ alternative mechanistic hypothesis
- Consequence of redundancy in cell interaction rules
- Robustness of pattern formation
- We inferred not only parameter values, but also the interaction rules themselves

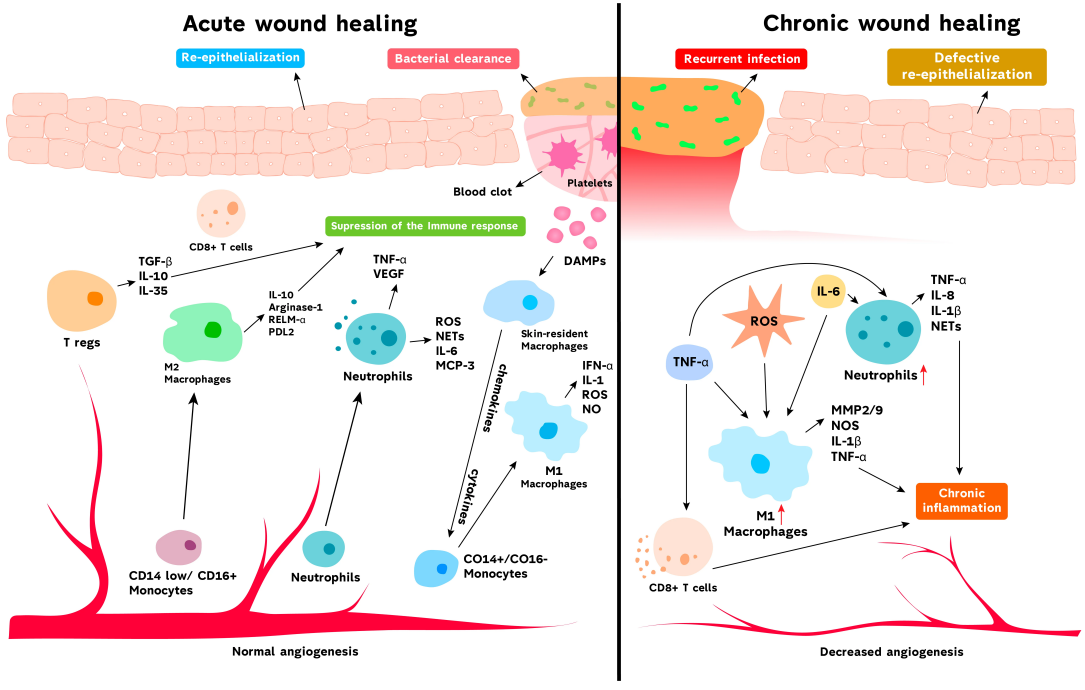


VI: Future directions

- Parameter inference for ABMs is important, but challenging
- Developed analysis pipeline with TDA+AABC, computationally tractable, readily generalisable to wide class of spatial ABMs
- Many choices and hyperparameter in the pipeline, effects to be studied in future work:
  - Choice of filtration for TDA
  - Methods for vectorising persistence homology
  - Sampling approaches for AABC
  - Methods for combining score functions
  - Hyperparameter tuning for weights and δ
- Sweeping Plane Filtration or Pair Correlation Functions as alternatives to Vietoris-Rips

VI: Further application: data-driven models of wound healing

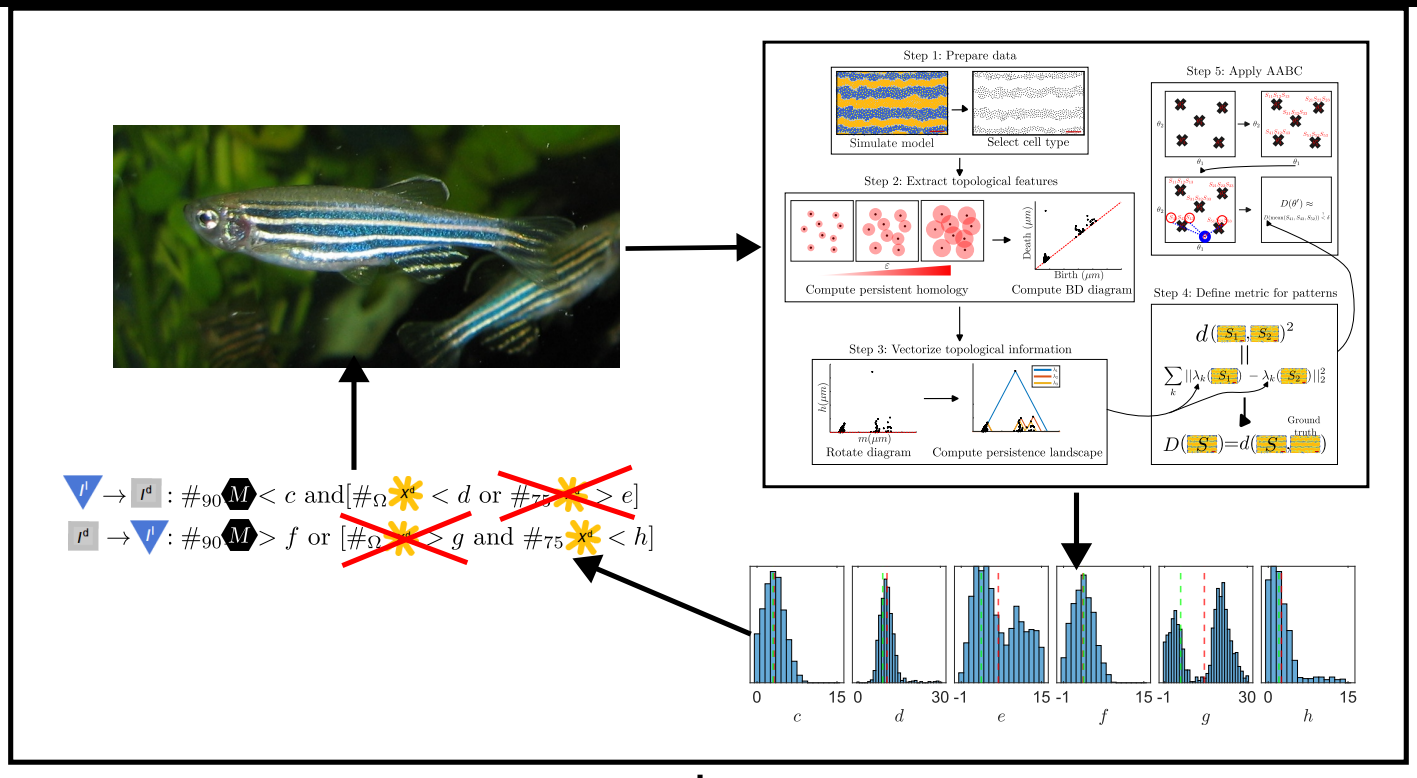
Modelling wound healing and associated immune response requires a multi-scaled approach combining PDEs and ABMs.

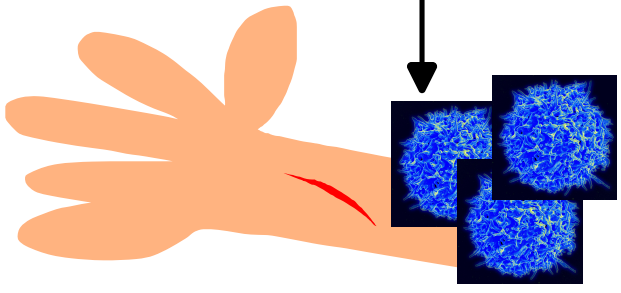


Raziyeva et al, 2021. Immunology of Acute and Chronic Wound Healing. Biomolecules 11(5)

Studying such model requires inference and and identifiability analysis, enabling optimal therapy design via control theory and machine learning.

VI: Summary and outlook

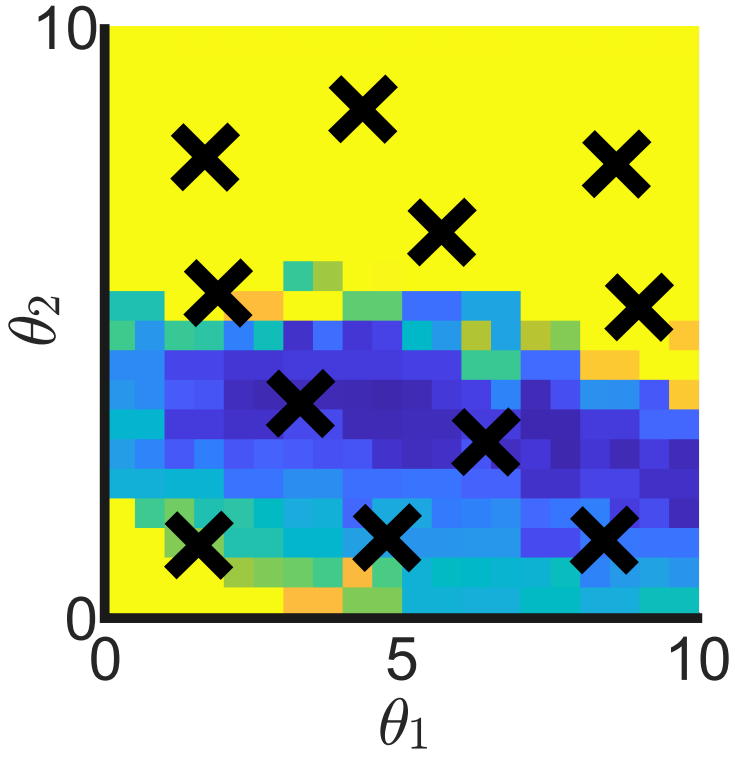




Thank you!

II: Rejection sampling

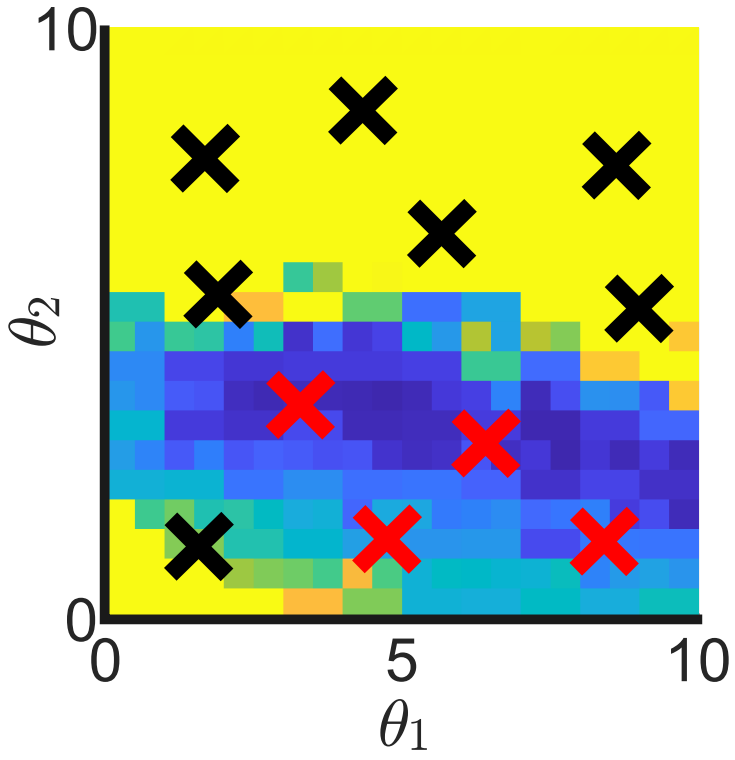
Suppose we have the likelihood function, rejection sampling (von Neumann 1940s) is a classic method for inference.
First, generate *proposals* (**X**) by sampling from the prior



Darker colour $\sim$ higher likelihood

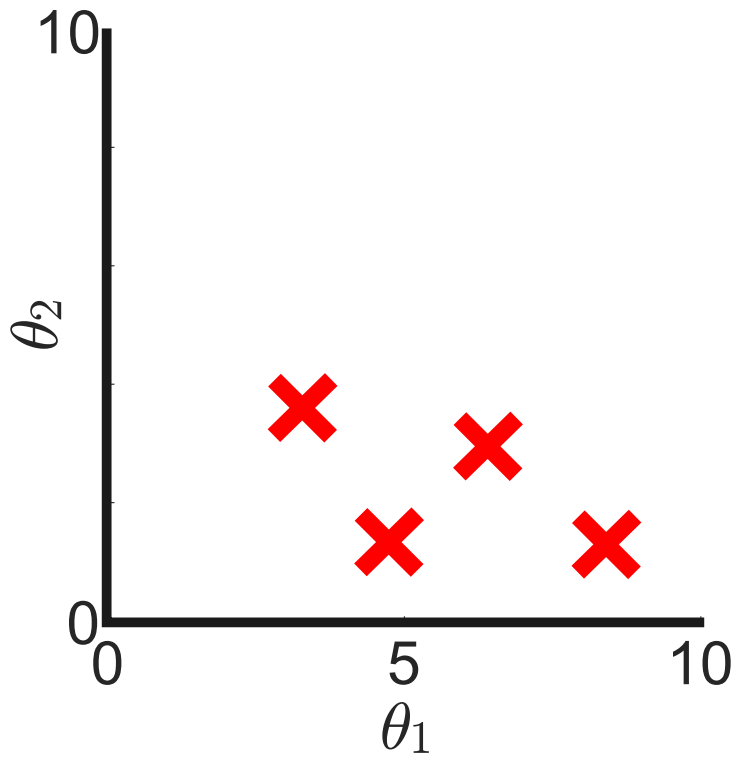
II: Rejection sampling

Next, evaluate the likelihood at each proposal, and accept if it is sufficiently high, reject otherwise



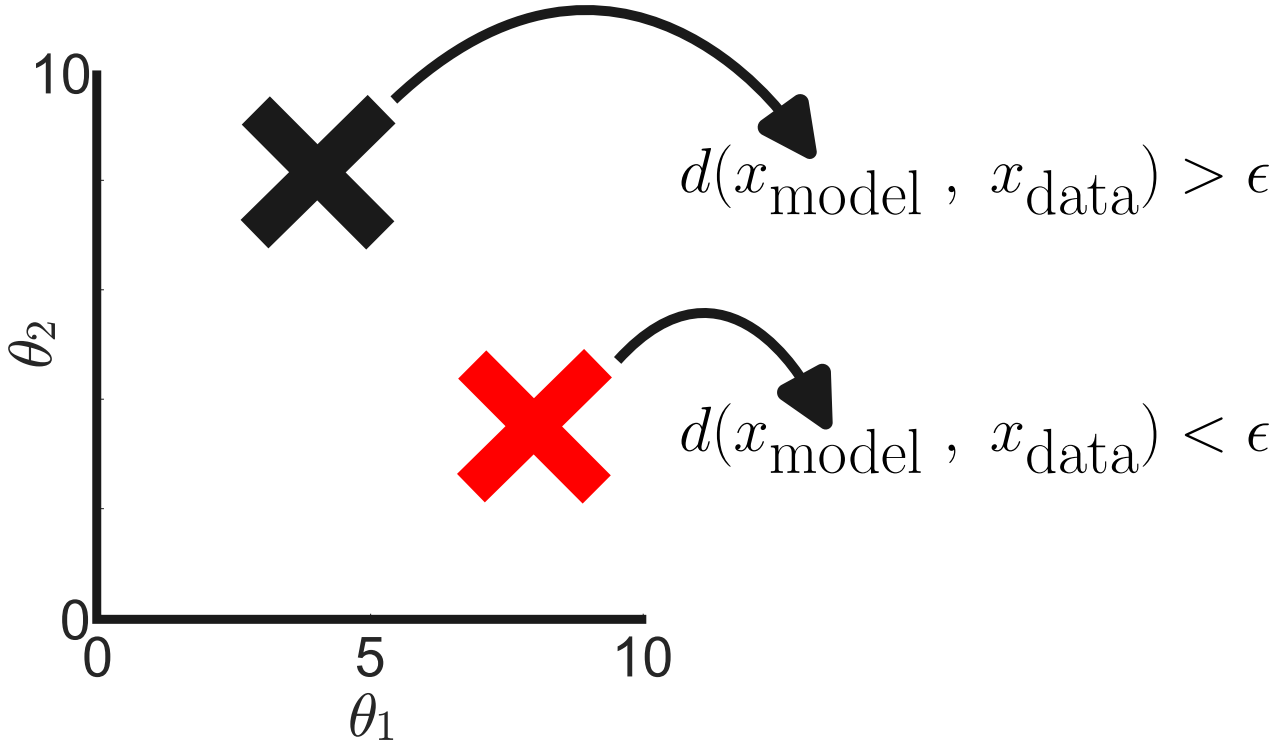
II: Rejection sampling

Repeat many times $\rightarrow$ *voila!* we obtained a sample from the posterior



II: ABC + rejection sampling

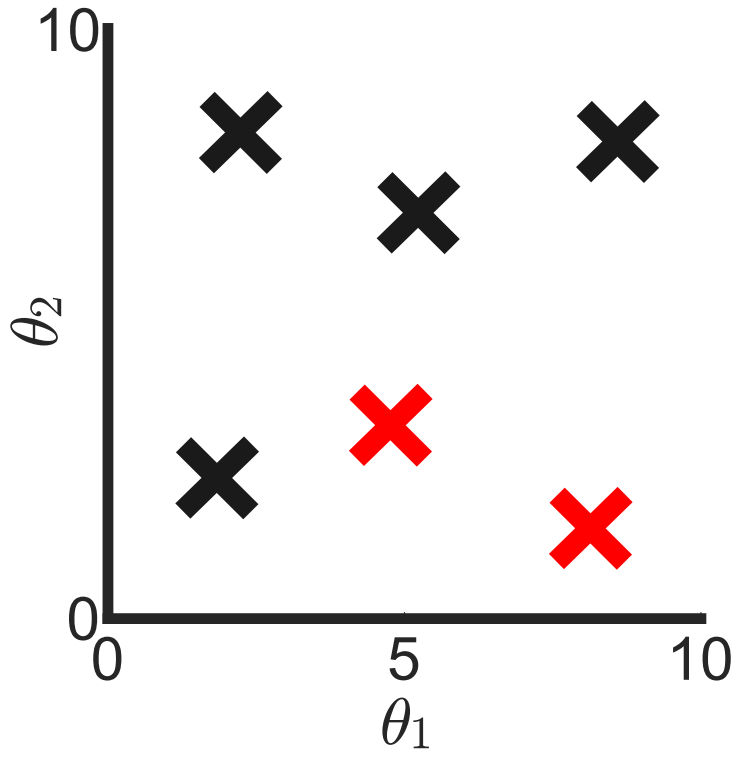
Approximate Bayesian Computation (Rubin 1984, Pritchard et al 1999): evaluating the likelihood is expensive, so use a *likelihood-free* acceptance criterion:



Defining this $d(\cdot\,,\,\cdot)$ is another challenge

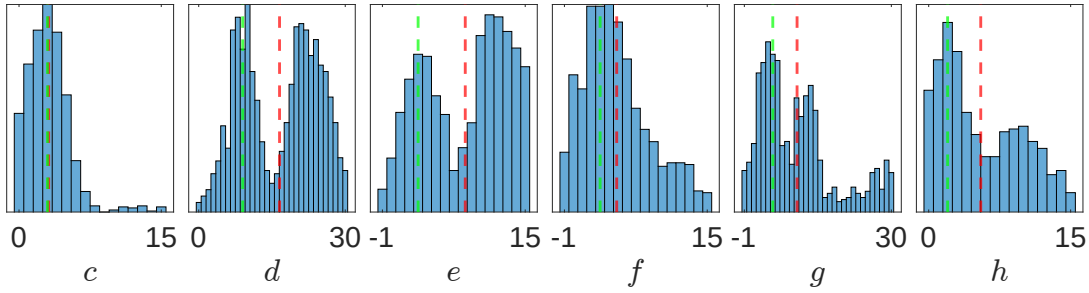
II: AABC + rejection sampling

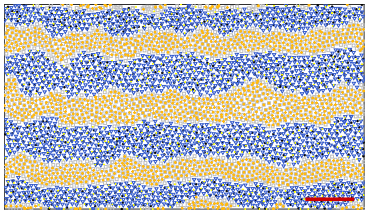
ABC is still expensive from running too many simulations $\rightarrow$ approximate again!
We first simulate the model for a small number of proposals (**X**)



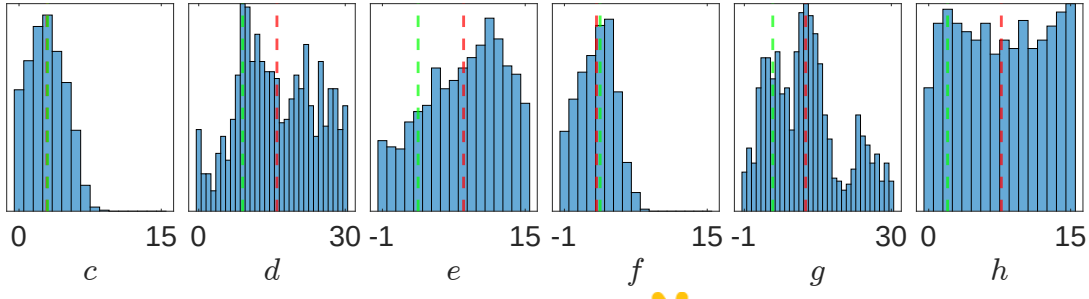
V: Examining score functions individually

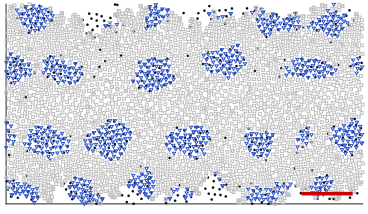
$D(\,\cdot\,;\mathrm{wt},\mathcal{M},t=45,\mathrm{dim}=1)$



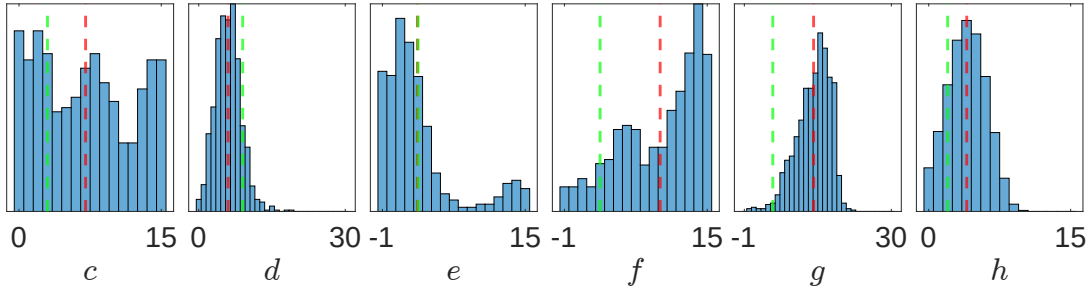


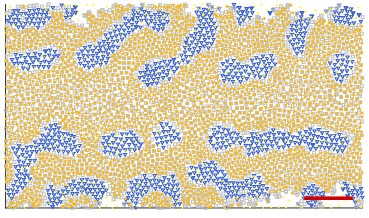
$D(\,\cdot\,;\mathrm{pfe},\mathcal{M},t=40,\mathrm{dim}=0)$





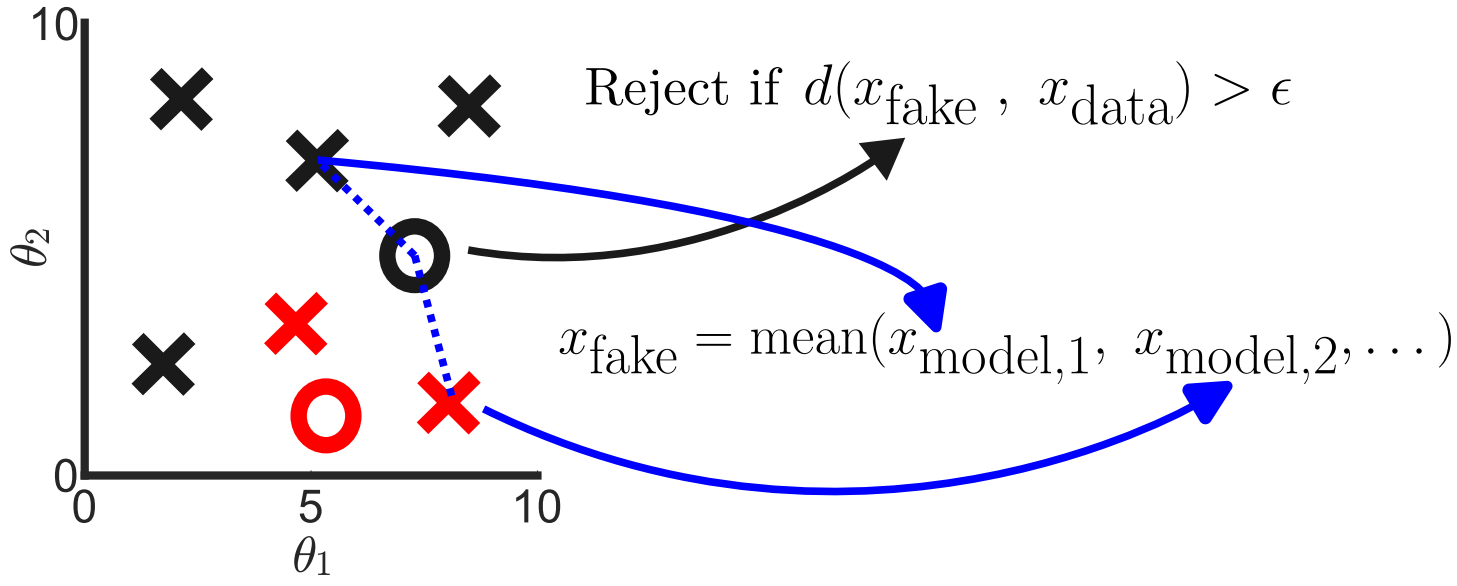
$D(\,\cdot\,;\mathrm{nac},\chi^{d},t=45,\mathrm{dim}=1)$





II: AABC + rejection sampling

Then, for a much larger number of proposals, (**O**), we aggregate the model output from neighbouring **X**s to stand-in as its output



The neighbours are chosen with some degree of randomness.

Buzbas & Rosenberg (2015) proved AABC converges to ABC as

#**X** $\rightarrow \infty$

Appendix: AABC algorithm

Algorithm AABC for the zebrafish model with persistence landscape-based metrics

1. Obtain the pre-proposals independently from the prior: $\theta_j^* \sim P(\theta),\ j = 1, \ldots, m$.

2. For each θ_j^*, simulate the model s times with different random seed: $\boldsymbol{x}_{j,k}^* \sim f(\theta^*), j = 1, \ldots, m,\ k = 1, \ldots, s$, and compute their corresponding persistence landscapes

3. For each further proposal θ':

3.1 Compute the distances $d_p(\theta', \theta_j^*)$, and denote $\theta^{*(j)}$ as the pre-sample with j^{th} lowest distance from θ.

3.2 Compute weights ω_j for selecting neighbouring parameter sets according to the Epanechnikov kernel:

$$\omega_j = \frac{3}{4} \frac{1}{d_p(\theta', \theta^{*(s+1)})} \left[1 - \left(\frac{d_p(\theta', \theta_j^*)}{d_p(\theta', \theta^{*(s+1)})} \right)^2 \right], j = 1, \ldots, m\ .$$

3.3 Sample ϕ according to the Dirichlet distribution with weight $\boldsymbol{\omega}$:

$$P(\boldsymbol{\phi}|\boldsymbol{\omega}) \propto \prod_{j=1}^{m} \phi_j^{\omega_j - 1}.$$

Here ϕ is a vector of non-negative weights, where ϕ_j is the weight of selecting the j^{th} pre-proposal as the neighbour

3.4 Sample s indices $\{i_1, \ldots, i_s\}$ from $\{1, \ldots, m\}$ with weights ϕ, and construct a set of s surrogate model output, $\{\boldsymbol{x}_k | k = 1, \ldots, s\}$:

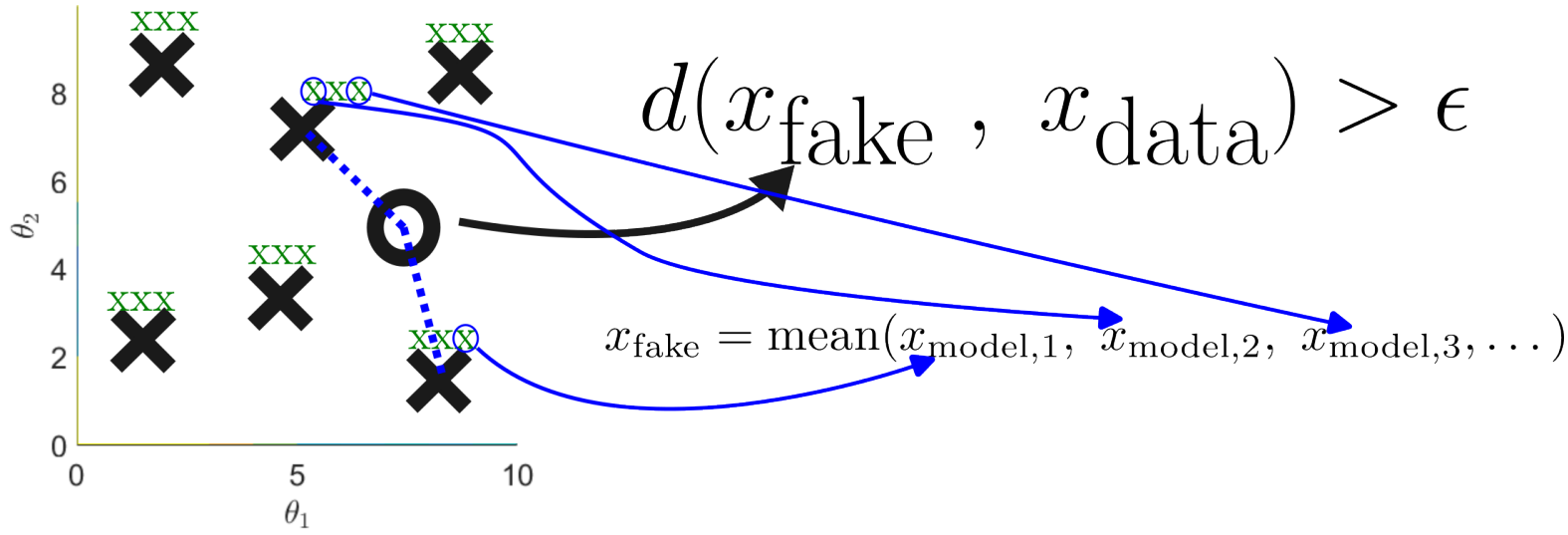
$$\boldsymbol{x}_k = \boldsymbol{x}_{i_k, \eta_k}^*, \text{ where } \eta_k \sim \mathrm{Unif}\{1, \ldots, s\},$$

and aggregate the corresponding persistence landscapes or surfaces

3.5 Compute the chosen score function $D(\theta')$ using the aggregated persistence landscapes or surfaces, and accept θ' if $D(\theta') < \delta$

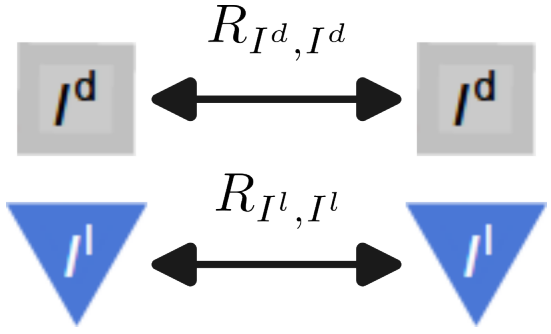
AABC + rejection sampling

Technicality: each sample actually has multiple simulations to take model stochasticity into account



Parameters for movement

I^d and I^l repel cells of the same kind with strength R_{I^d,I^d}, R_{I^l,I^l}, respectively



The story is the same