UNIVERSITY OF OXFORD

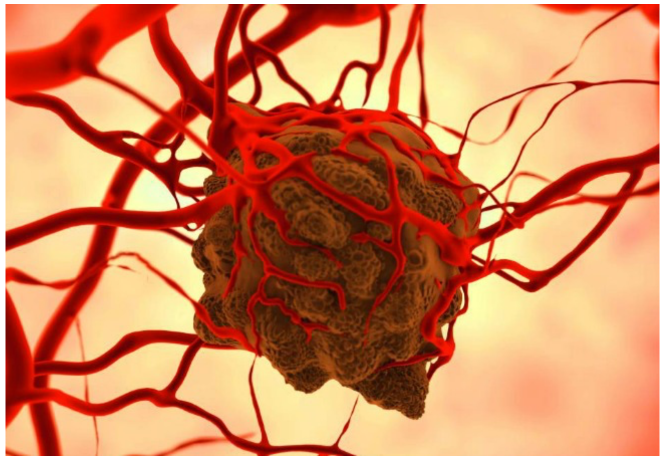
Leveraging concepts from stochastic simulation and machine learning for efficient Bayesian inference

Ruth Baker

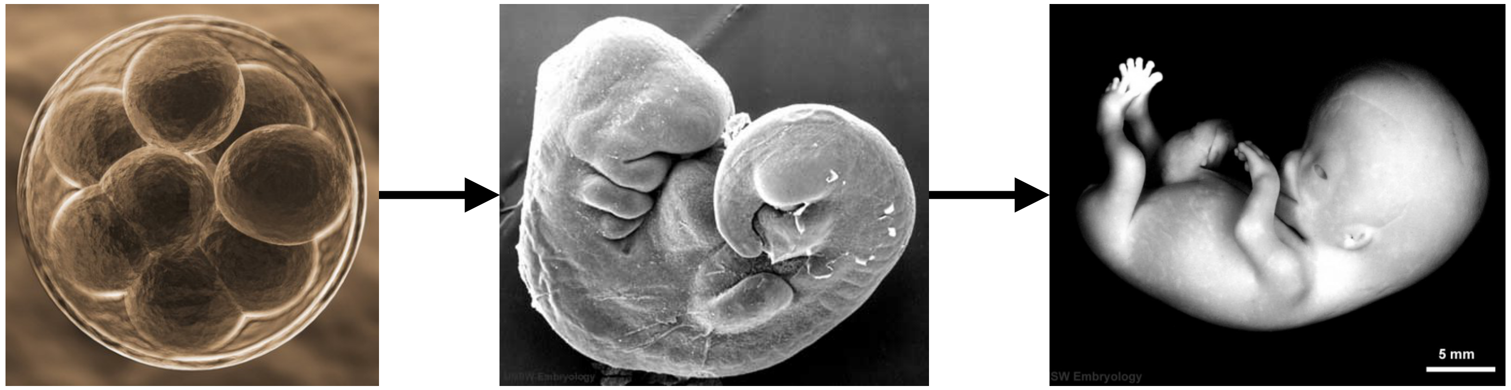
@ruth_baker

Applied research focus

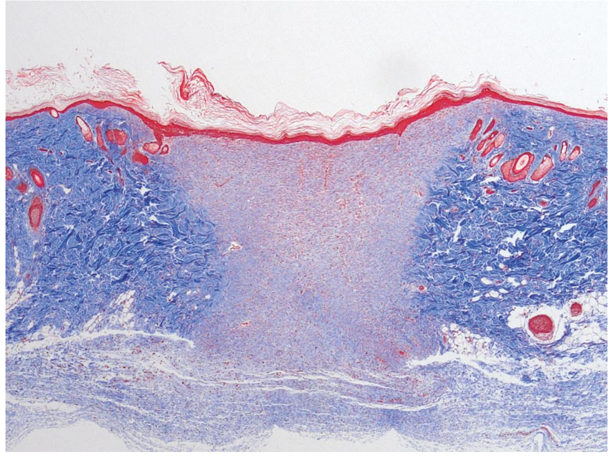
To understand the mechanisms driving collective cell motility, proliferation and death and their contributions to complex biological processes, such as those associated with development, disease and repair.



tumour growth



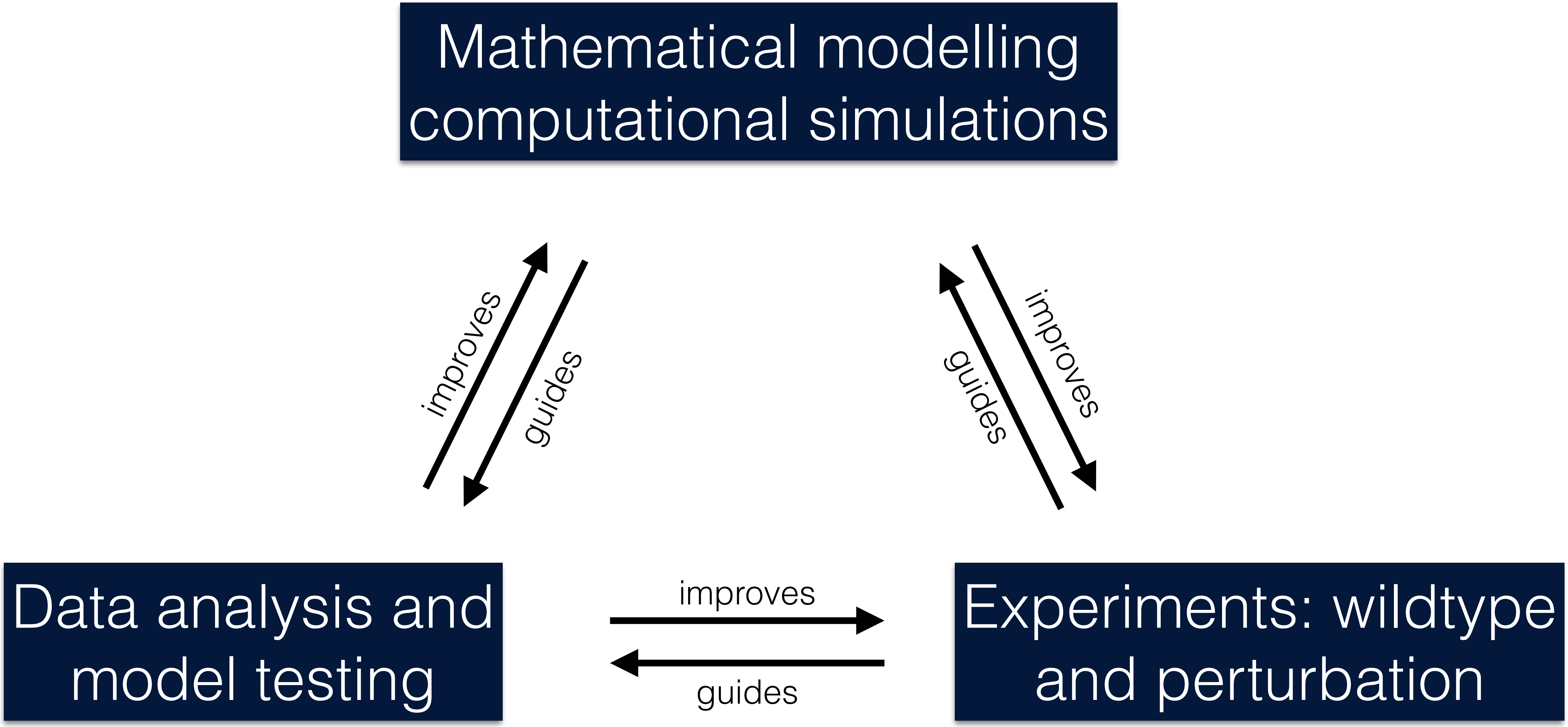
embryo development



wound healing

Combine mechanistic mathematical models and statistical / machine learning tools to provide new biological insights.

Interdisciplinary methodology



Mathematical modelling computational simulations

Data analysis and model testing

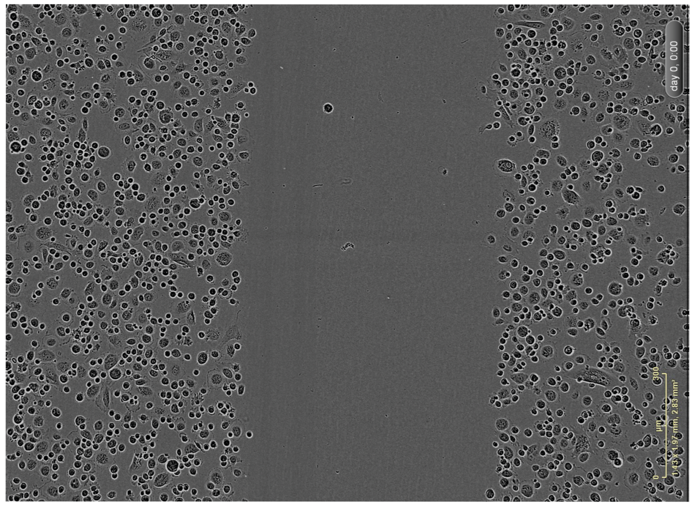
Experiments: wildtype and perturbation

Integral in the cycle of

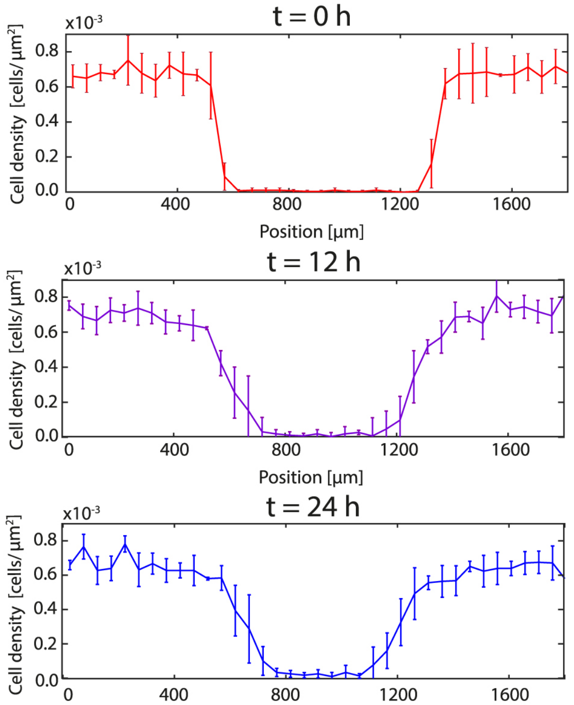
predict - test - refine - predict

The inverse problem

- Given quantitative data, can we estimate model parameters?



→



→

- Parameter inference using a Bayesian framework:

$$\mathbb{P}(\theta\,|\,\mathcal{D}) \propto \mathcal{L}(\mathcal{D}\,|\,\theta)\mathbb{P}(\theta)$$

posterior likelihood prior

Approximate Bayesian computation

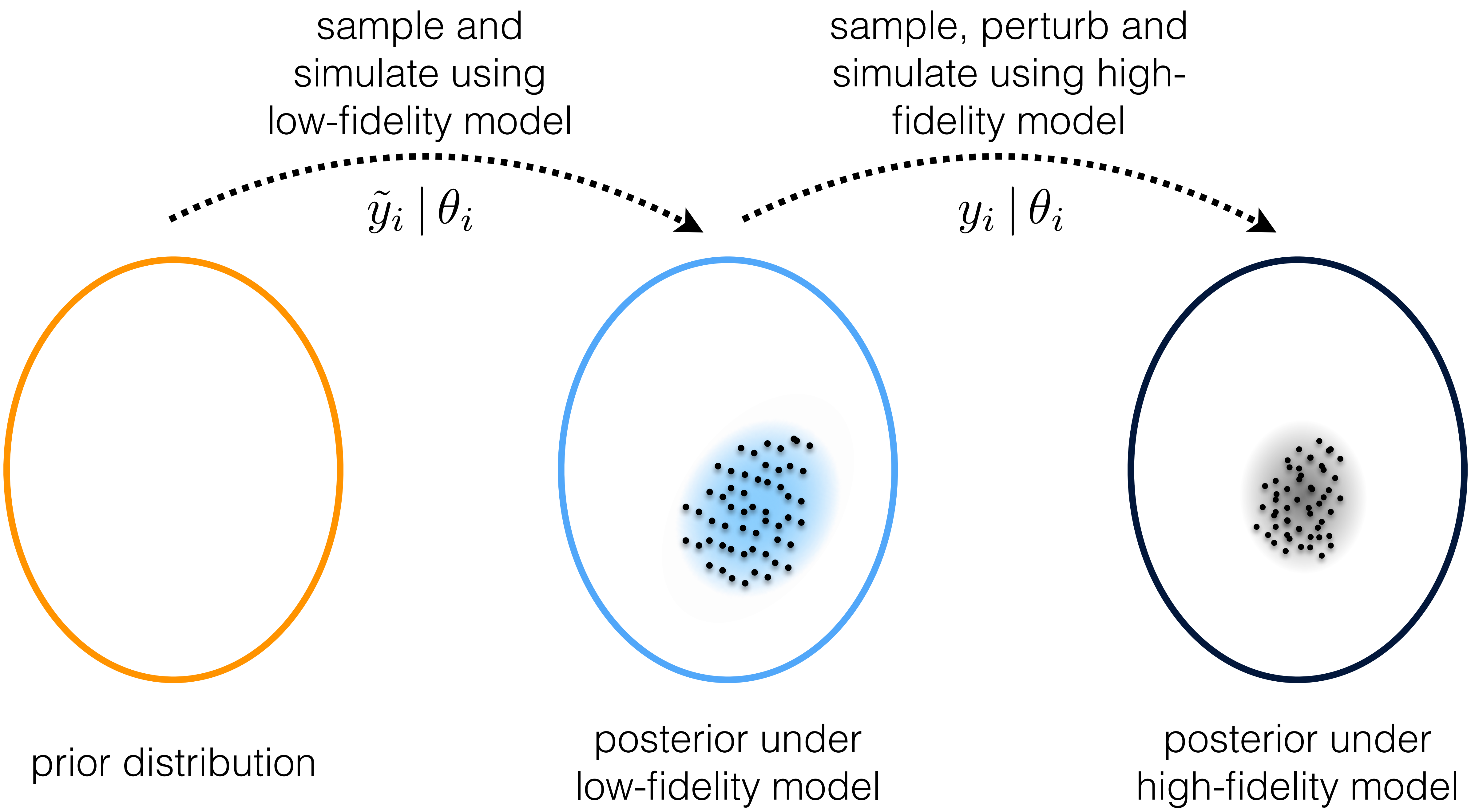
- Likelihood is generally intractable, so we use ABC:
    - A **very large number** of times:
        - sample parameter from some distribution;
        - evaluate how close model output is to data using a summary statistic and distance function;
        - assign a weight to the parameter - depends on this distance.

When a model takes minutes (or longer) to simulate, and / or the model has many parameters, ABC methods can be infeasible.

Increasing the efficiency of ABC

- Use a more intelligent exploration of parameter space.
  - Importance sampling, sequential Monte Carlo *etc.*
  - **Pre-conditioned ABC** - Use a simple model to help transition from prior to posterior.
- Make the weights less expensive to calculate.
  - **Moment-matching ABC** - Use a simple model to help estimate posterior.
  - **Multifidelity ABC** - Use hierarchies of models in the weight calculations.
  - **Minibatch ABC** - Use subsets of the data in evaluating the distance function.

Pre-conditioned ABC



sample and simulate using low-fidelity model

$\tilde{y}_i \mid \theta_i$

sample, perturb and simulate using high-fidelity model

$y_i \mid \theta_i$

prior distribution

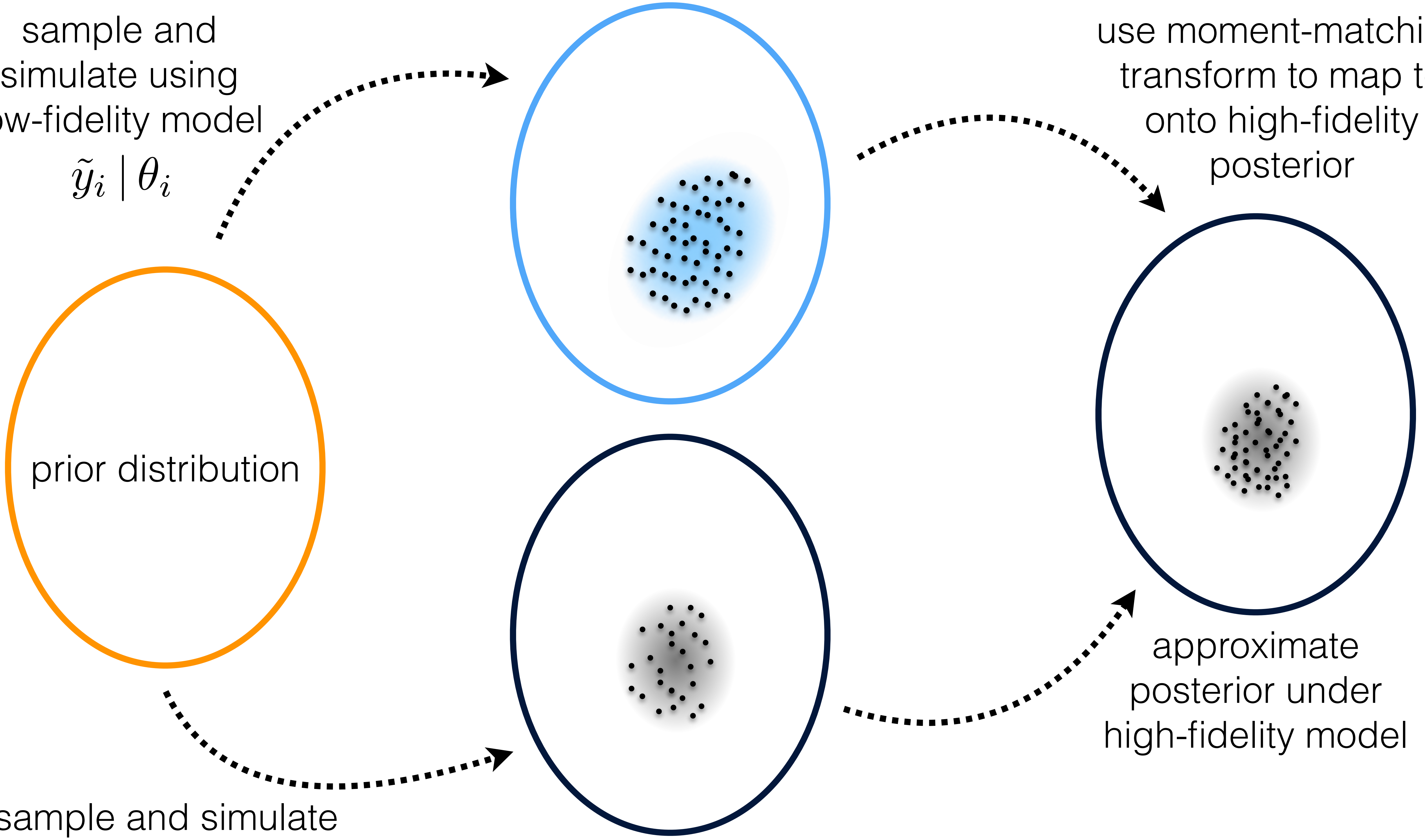
posterior under low-fidelity model

posterior under high-fidelity model

Weight accepted particles so that ABC posterior is returned.

Warne, Baker and Simpson, *J. Comp. Graphical Stat.* (2022).

Moment-matching ABC



sample and simulate using low-fidelity model

$\tilde{y}_i \mid \theta_i$

use moment-matching transform to map to onto high-fidelity posterior

prior distribution

approximate posterior under high-fidelity model

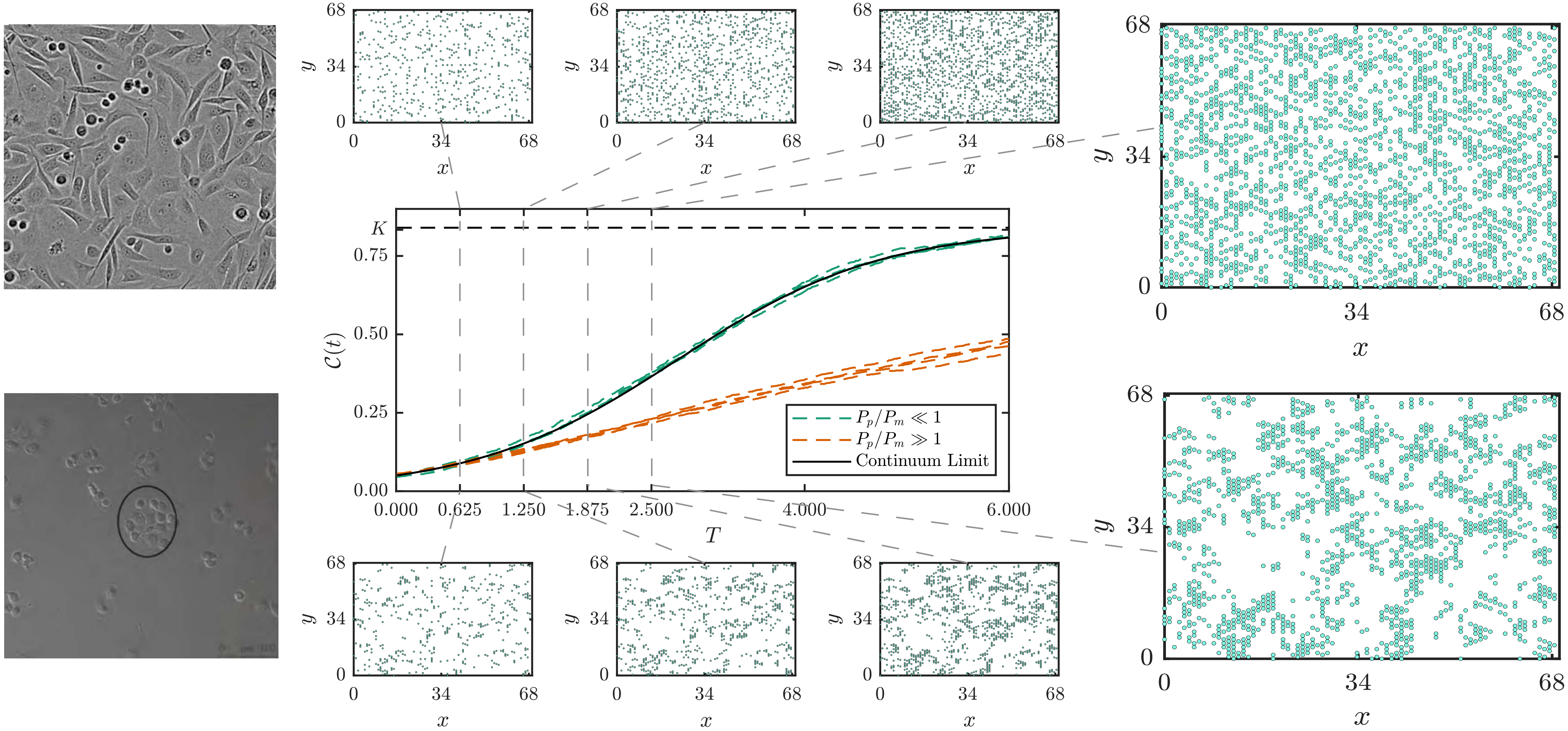
sample and simulate using high-fidelity model

$y_i \mid \theta_i$

Returns approximate posterior due to transform

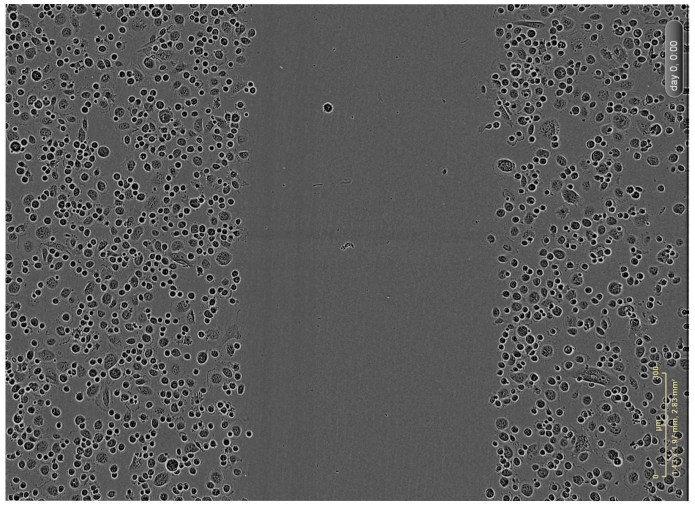
Pre-conditioned / MM ABC

- Discrete, on-lattice random walk model of cell migration and proliferation.



Warne, Baker and Simpson, *J. Comp. Graphical Stat.* (2022).

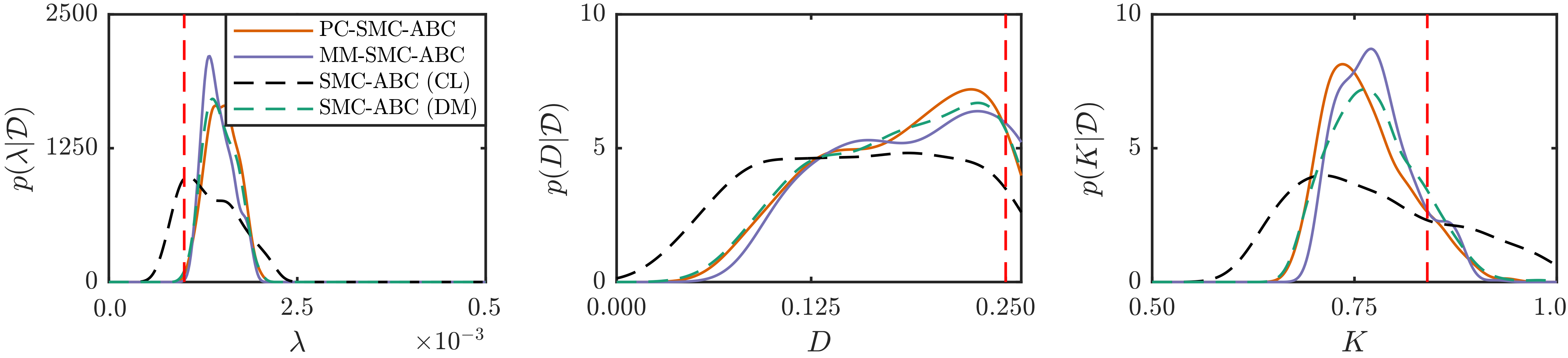
Pre-conditioned / MM ABC



Continuum limit - Fisher-KPP equation:

$$\frac{\partial C}{\partial t} = D\frac{\partial^2 C}{\partial x^2} + \lambda C\left(1 - \frac{C}{K}\right)$$

Marginal posterior parameter distributions



Warne, Baker and Simpson, *J. Comp. Graphical Stat.* (2022).

Pre-conditioned / MM ABC - summary

- Pre-conditioned and moment-matching ABC can provide significant time savings, through the combined use of high-fidelity and low-fidelity models.

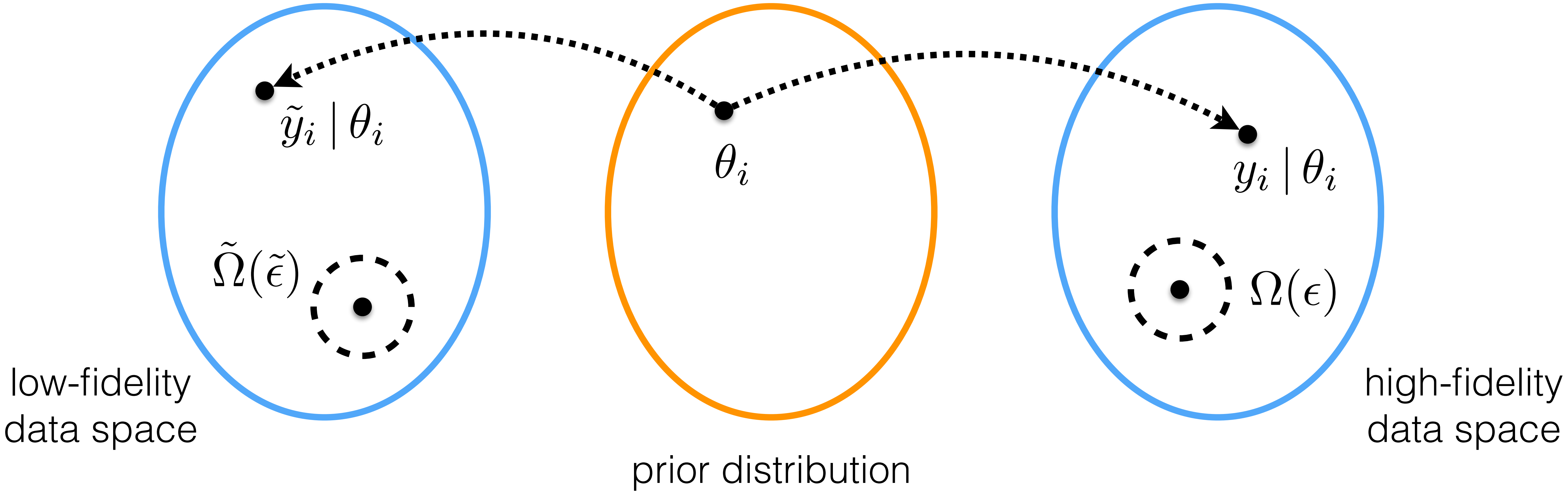
- Need to explore trade-off between fraction of high-fidelity and low-fidelity samples.

- Doesn't require the low-fidelity model to be particularly accurate, just that the model outputs depend in a qualitatively similar way on the input parameters.

Warne, Baker and Simpson, *J. Comp. Graphical Stat.* (2022).

Multifidelity ABC

- Can we use the low-fidelity model to allocate some weights (accept or reject parameters) and not bias the result?



Prescott and Baker, *SIAM / ASA J. UQ* (2020) and (2021).

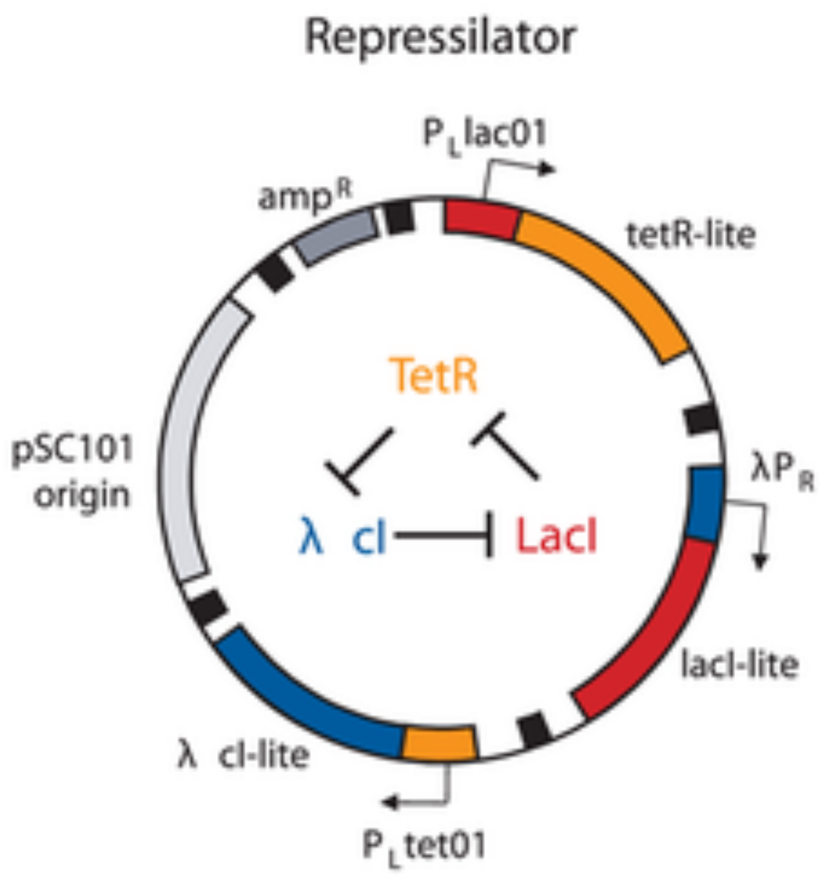
Multifidelity ABC - example

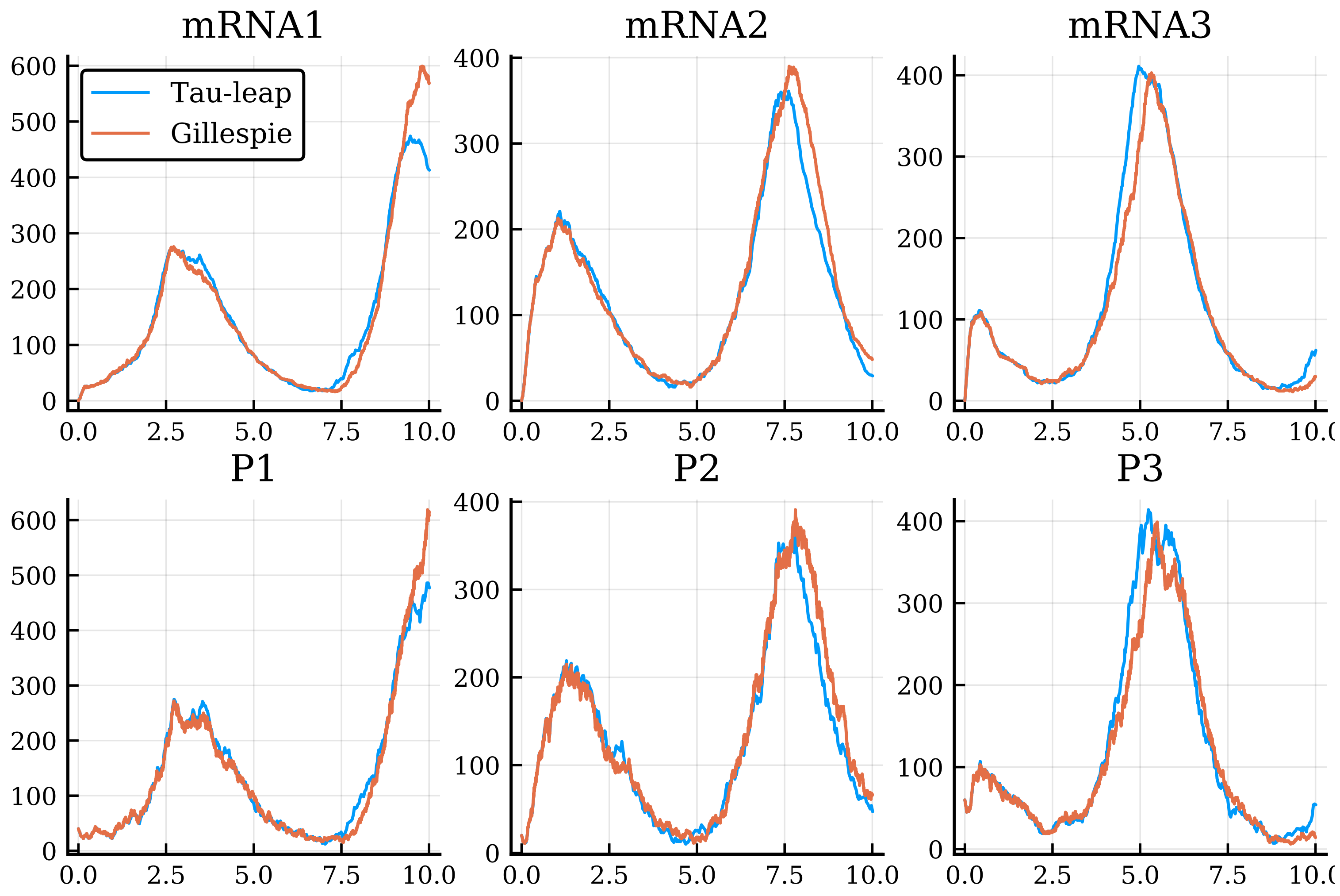
$$\xrightarrow{\alpha_0+\alpha f(p_j)} m_i \xrightarrow{1} \emptyset \qquad \text{for } (i,j)=(1,3),\ (2,1),\ \text{and } (3,2)$$

$$m_i \xrightarrow{\beta} m_i + p_i \qquad \text{for } i = 1,2,3$$

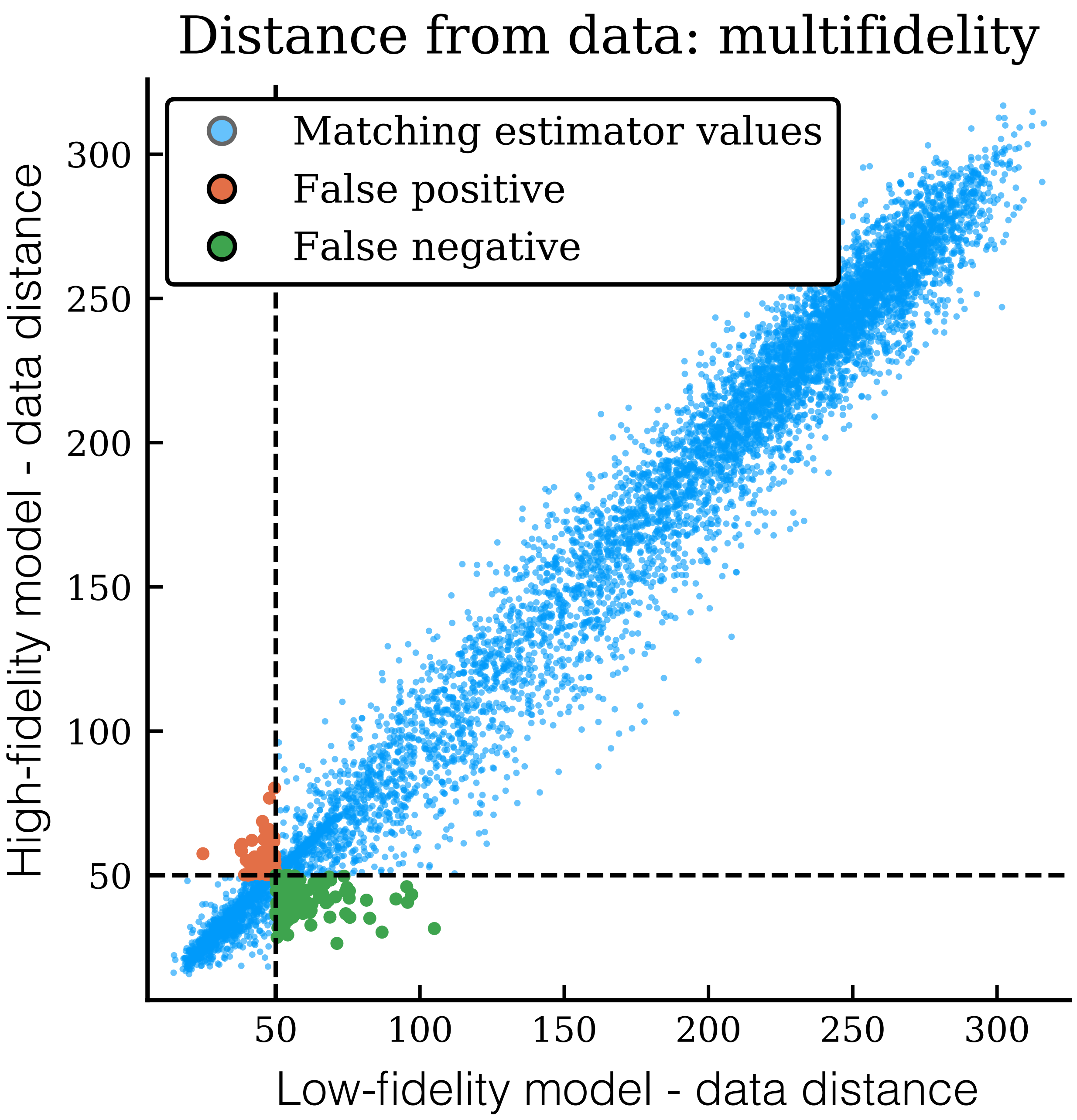
$$p_i \xrightarrow{\beta} \emptyset \qquad \text{for } i = 1,2,3$$

$$f(p) = \frac{K_h^n}{(K_h^n + p^n)}$$





Multifidelity ABC - the problem



- How to combine the outputs of the two models, so that the result is unbiased weights?

- How can we make this process efficient?

Prescott and Baker, *SIAM / ASA J. UQ* (2020) and (2021).

Multifidelity ABC - the algorithm

- Attempt to make an “early decision” using the low-fidelity model, and “sometimes” check that decision using the high-fidelity model.

- Decision to check is made uniformly at random, with probability $\alpha(\tilde{y}, \theta_n)$.

    - Here, w uming

$$\alpha(y, \theta_n) = \eta_1 \mathbb{1}(y \in \Omega(\epsilon)) + \eta_2 \mathbb{1}(y \notin \Omega(\epsilon))$$

Using a common noise input makes this a very efficient approach.

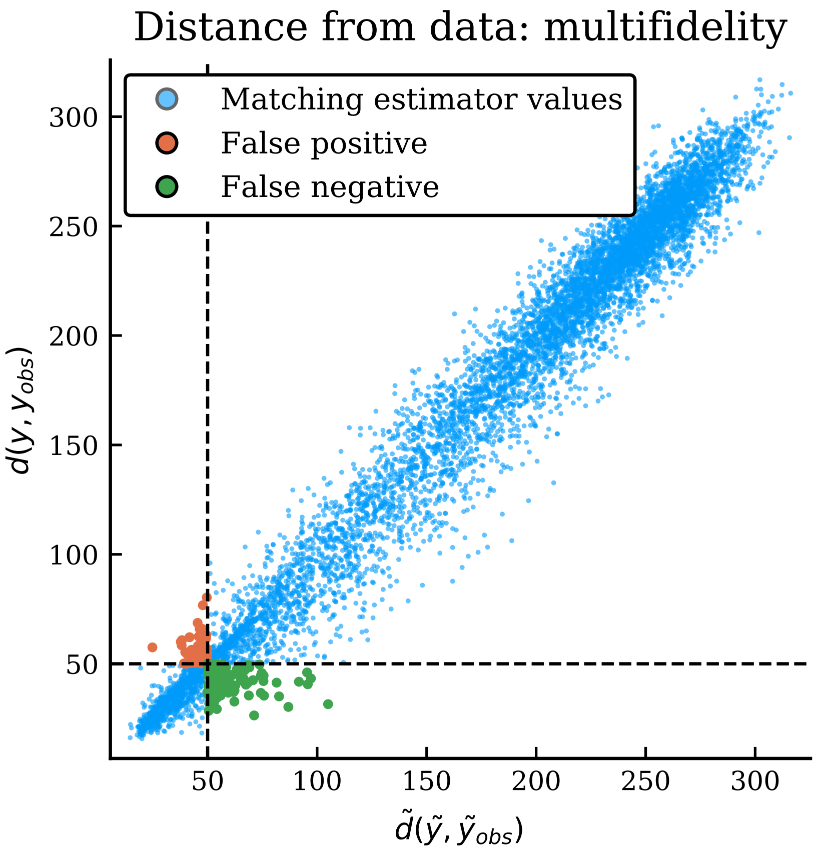
if low-fidelity model is **close to** data

if low-fidelity model is **far from** data

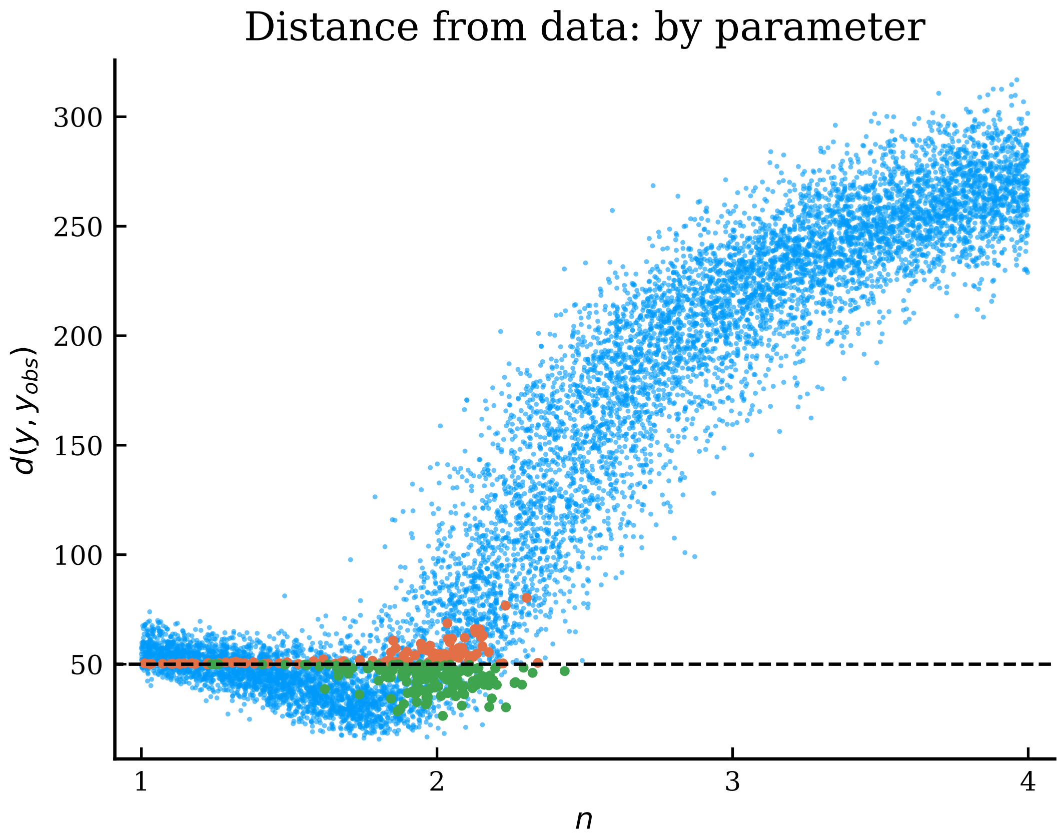
Prescott and Baker, *SIAM / ASA J. UQ* (2020) and (2021).

Multifidelity ABC - efficiency

- Derive analytical expressions for the optimal continuation probabilities, given estimates of various quantities.
    - In practice: adapt the continuation probabilities “on the fly”, as samples are generated…



Distance from data: multifidelity

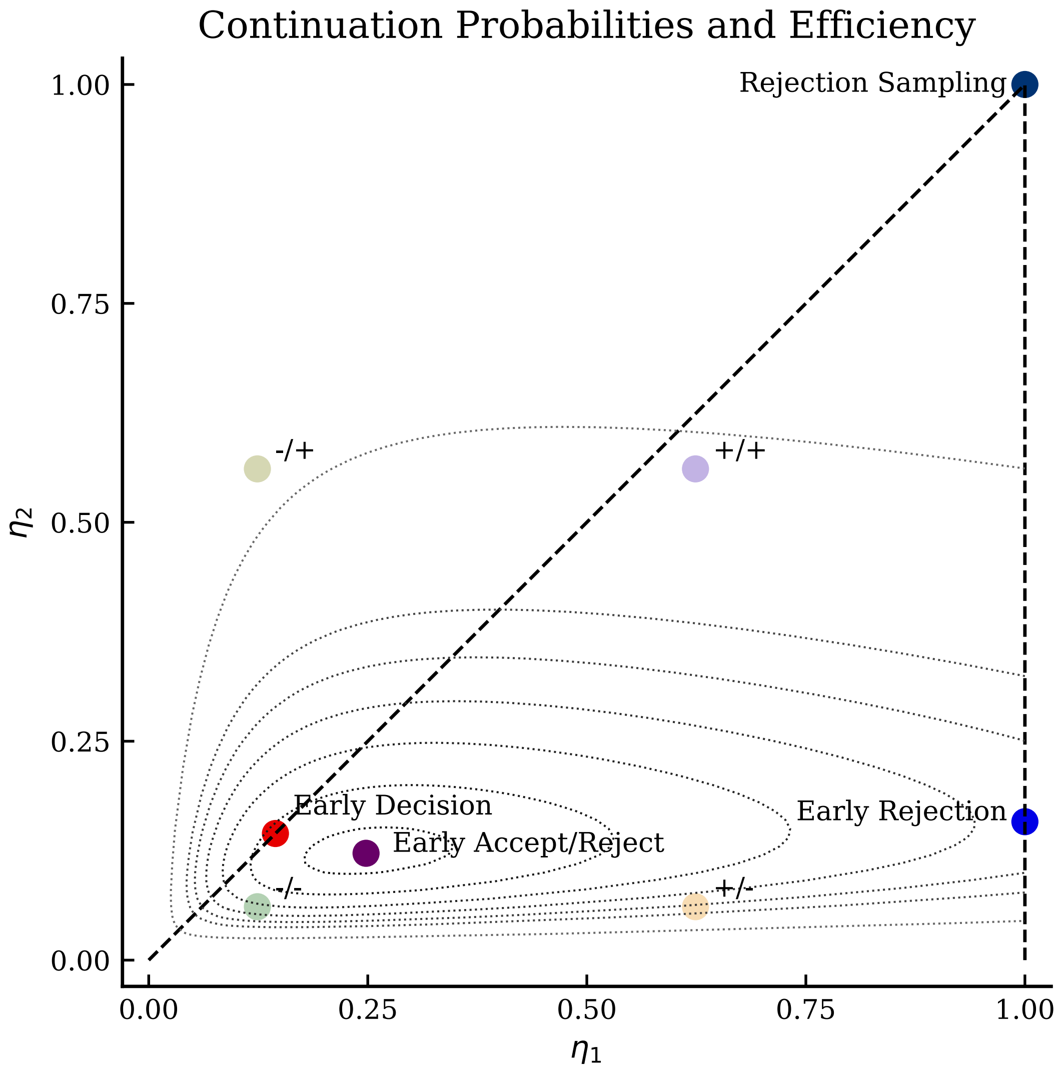


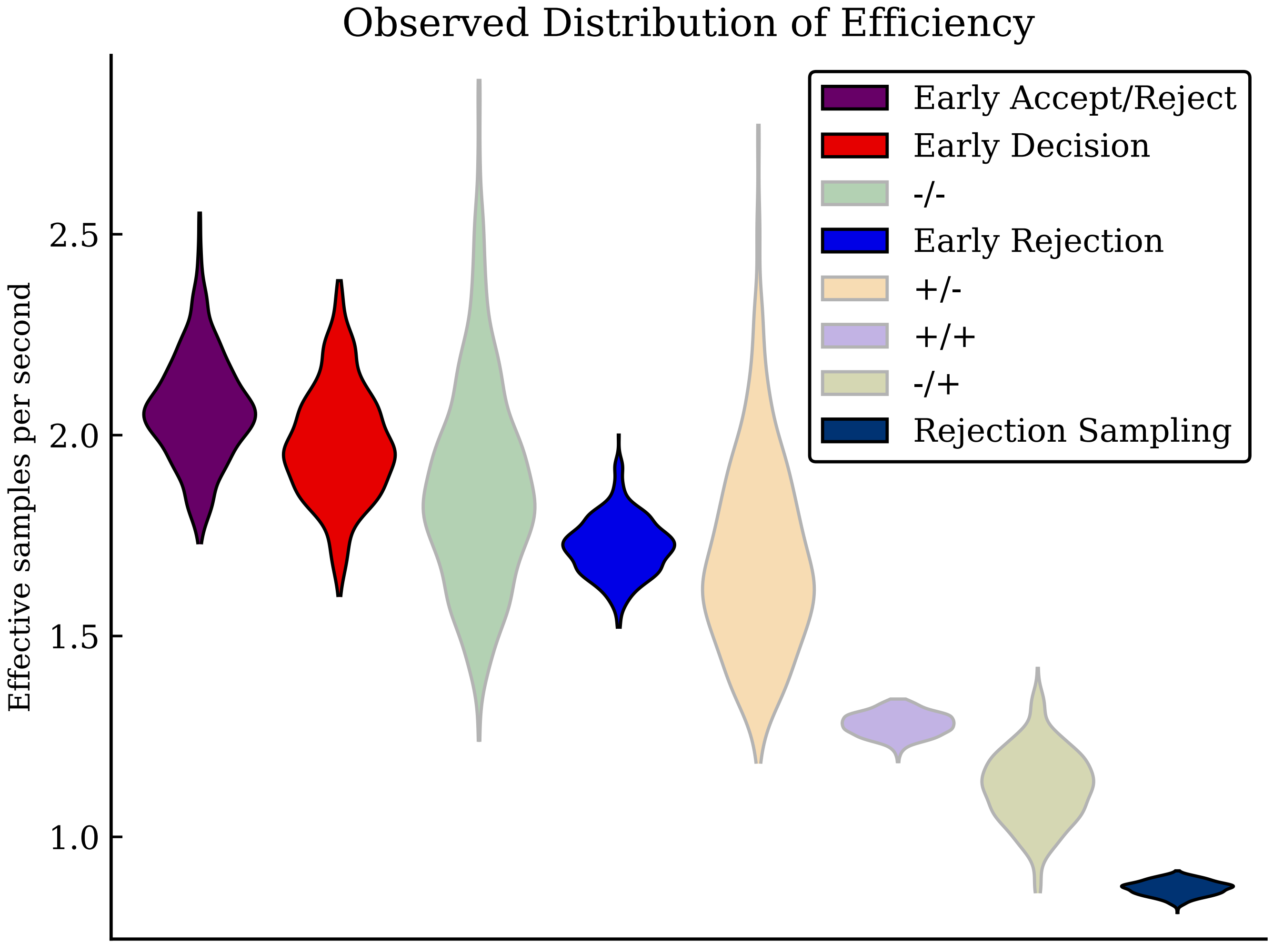
Distance from data: by parameter

Prescott and Baker, *SIAM / ASA J. UQ* (2020) and (2021).

Multifidelity ABC - example results

- Comparing results for a range of continuation probabilities:


Continuation Probabilities and Efficiency


Observed Distribution of Efficiency

Prescott and Baker, *SIAM / ASA J. UQ* (2020) and (2021).

Multifidelity ABC - conclusions

- Multifidelity ABC can provide time savings, through the combined use of high- and low-fidelity models.

- Can “learn” optimal continuation probabilities as the algorithm proceeds, separately controlling rates of checking early acceptance and early rejection.

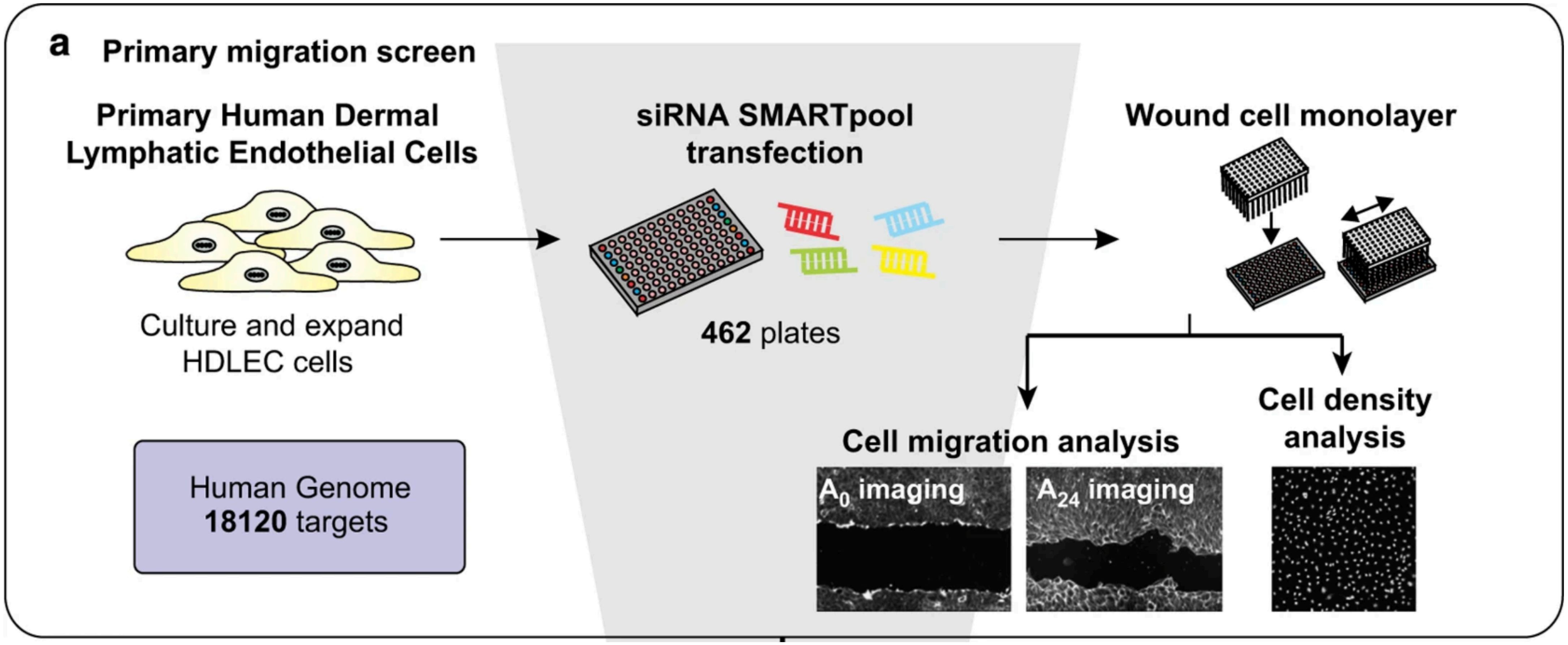
- Rates of false positives and negatives can be reduced by generating the high-fidelity model output conditional on the low-fidelity model output.

  - Enables smaller continuation probabilities and hence simulation cost.

Prescott and Baker, *SIAM / ASA J. UQ* (2020) and (2021).

Minibatch ABC

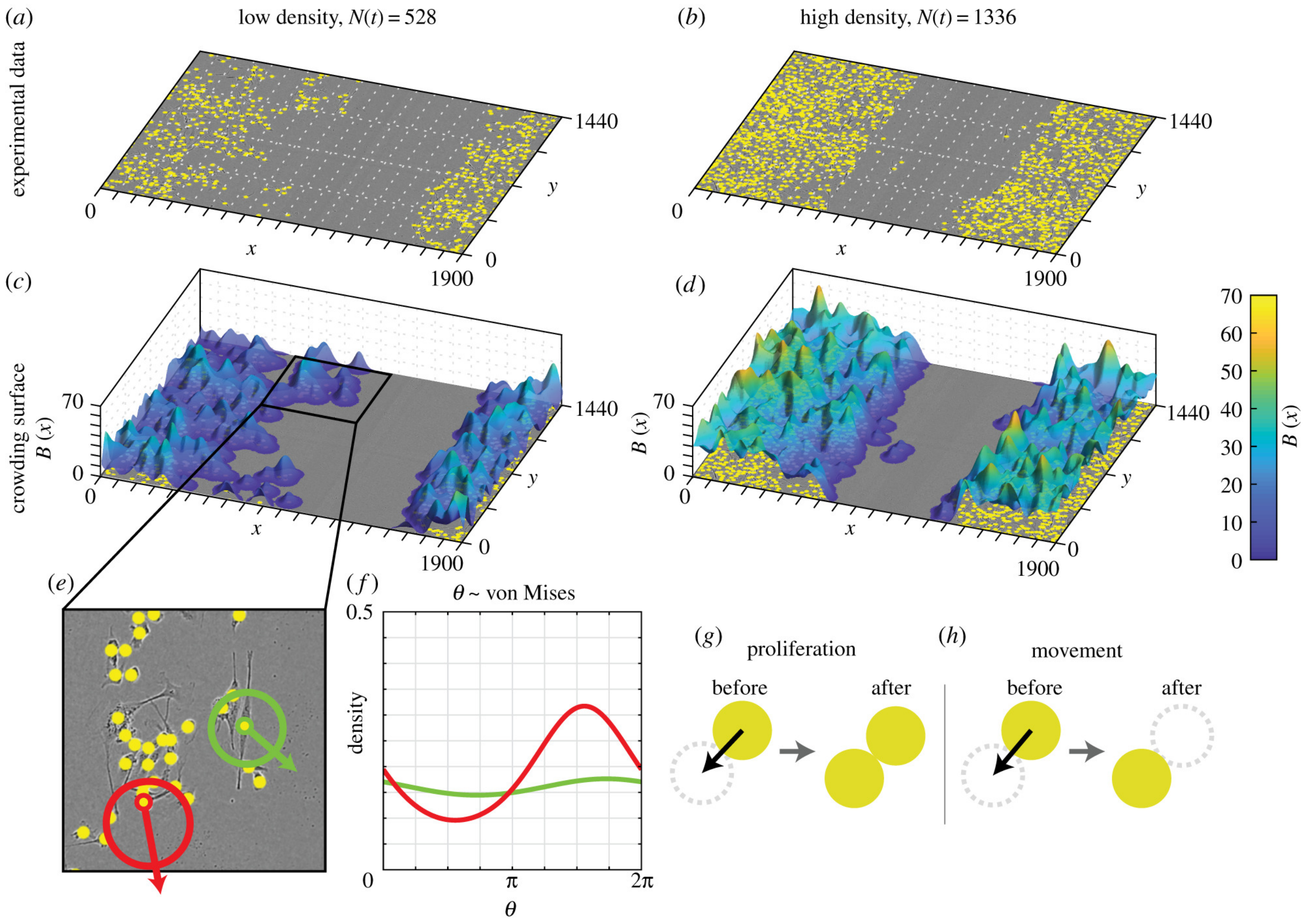
- Designed for use with high-throughput data i.e. large quantities of data.
- Motivated by genome-wide RNAi screen of endothelial cells.


a
Primary migration screen
Primary Human Dermal Lymphatic Endothelial Cells
Culture and expand HDLEC cells
Human Genome 18120 targets
siRNA SMARTpool transfection
462 plates
Wound cell monolayer
Cell migration analysis
A_0 imaging
A_{24} imaging
Cell density analysis

Williams *et al. Sci. Data* 4 (2017).

Individual-based model

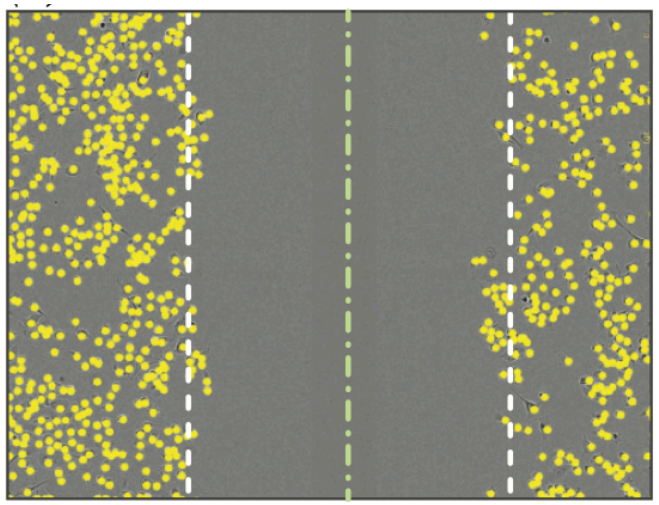
- Stochastic, off-lattice individual-based model of cell motility and proliferation - including density-dependent effects.



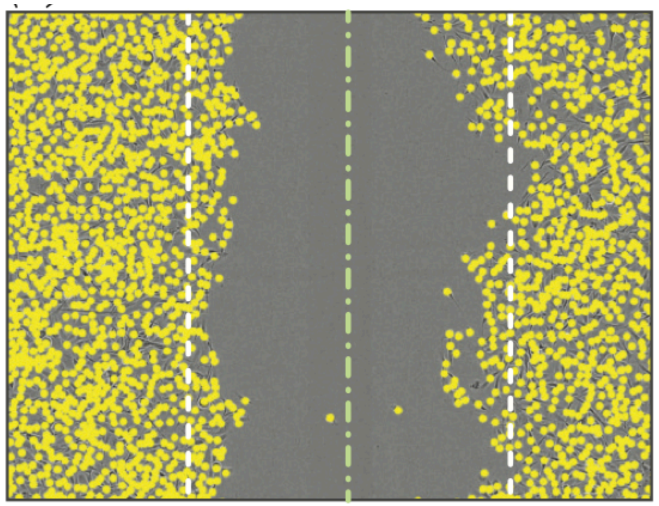
Browning *et al. J. Roy. Soc. Interface* 17 (2020).

Model - data comparison

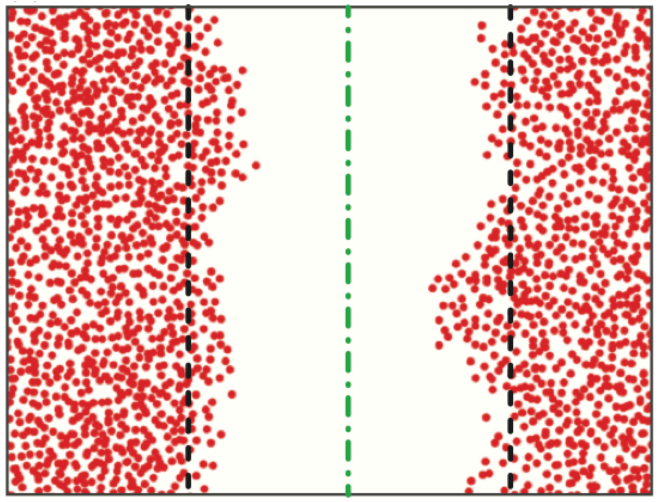
Data at t=0 hrs



Data at t=24 hrs



Model at t=24 hrs



Parameter inference using a Bayesian framework:

$$\mathbb{P}(\theta \,|\, \mathcal{D}) \propto \mathcal{L}(\mathcal{D} \,|\, \theta)\mathbb{P}(\theta)$$

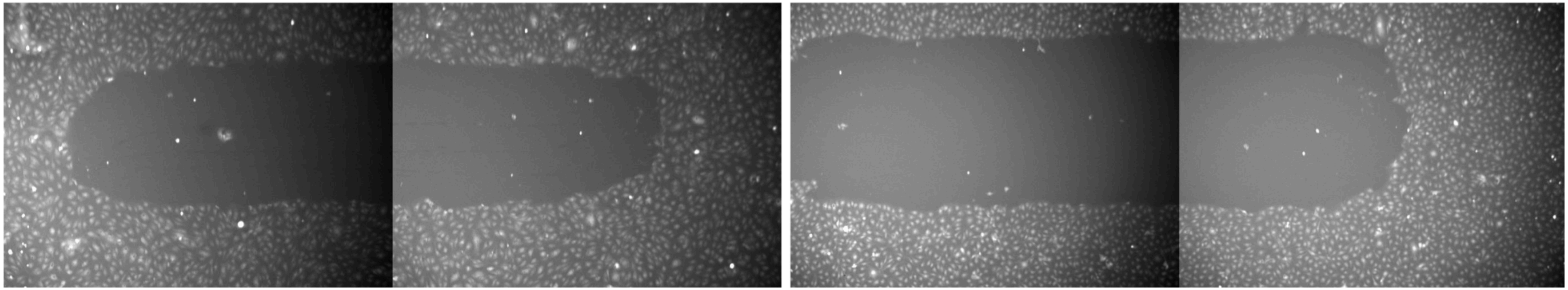
with

$$\theta = (m, p, d, \gamma_b, \gamma_m, \gamma_p)$$

Browning *et al. J. Roy. Soc. Interface* 17 (2020).

What’s the additional challenge?

- Large numbers of replicates (~100) for some knockdowns.
    - Huge variability in the initial wound size / shape (initial condition).



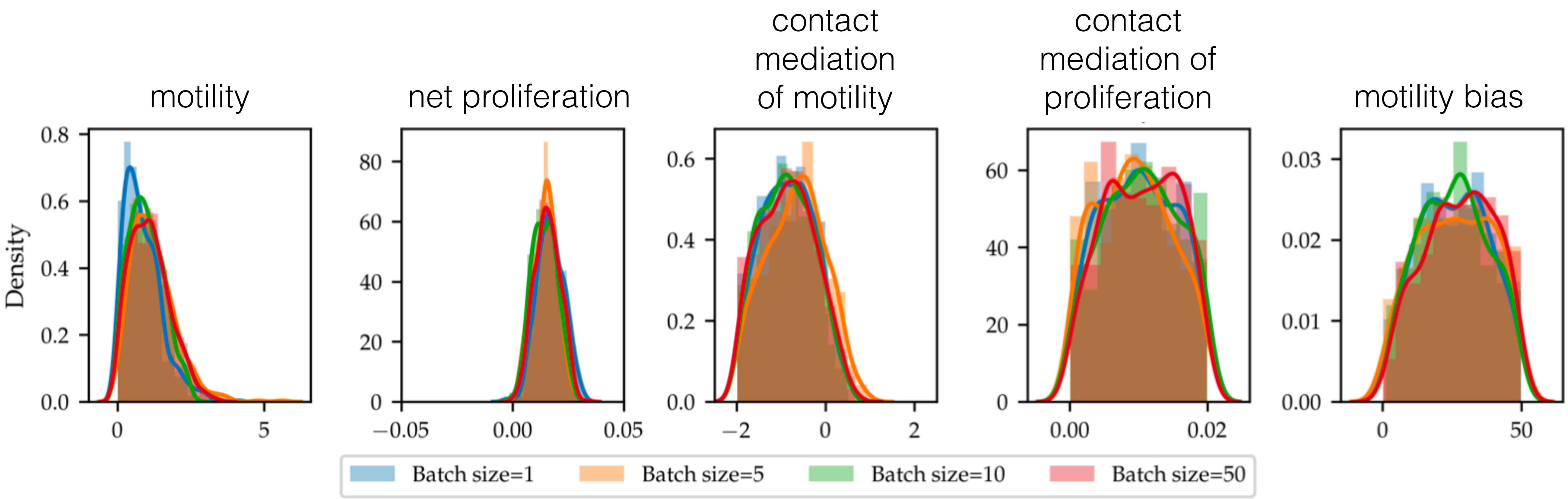
    - Need to generate individual simulations / make individual comparisons for each replicate.

Minibatch ABC

- A very large number of times:
  - sample parameter from prior distribution;
  - sample a **minibatch** of the data;
  - simulate model using this parameter - **do this for each sample from the minibatch individually**;
  - evaluate how close model output is to data using summary statistics and distance function - **do this for each sample from the minibatch individually**;
  - assign a weight to the parameter - depends on this distance.

Martina Perez, Sailem and Baker, *PLOS Comp. Biol.* (2022).

Minibatch ABC - test results


motility

net proliferation

contact mediation of motility

contact mediation of proliferation

motility bias

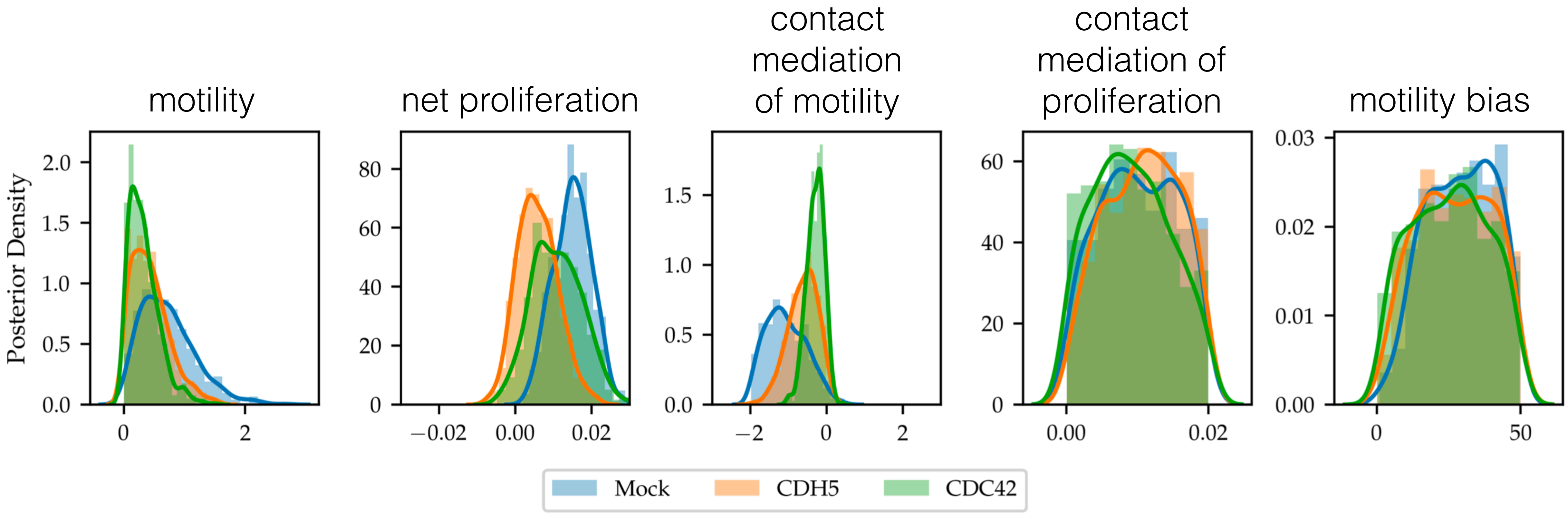
- Posterior distribution very similar over the different batch sizes.

- Confidence in estimates of motility, net proliferation, contact-mediation of motility parameters.

Martina Perez, Sailem and Baker, *PLOS Comp. Biol.* (2022).

Minibatch ABC - results

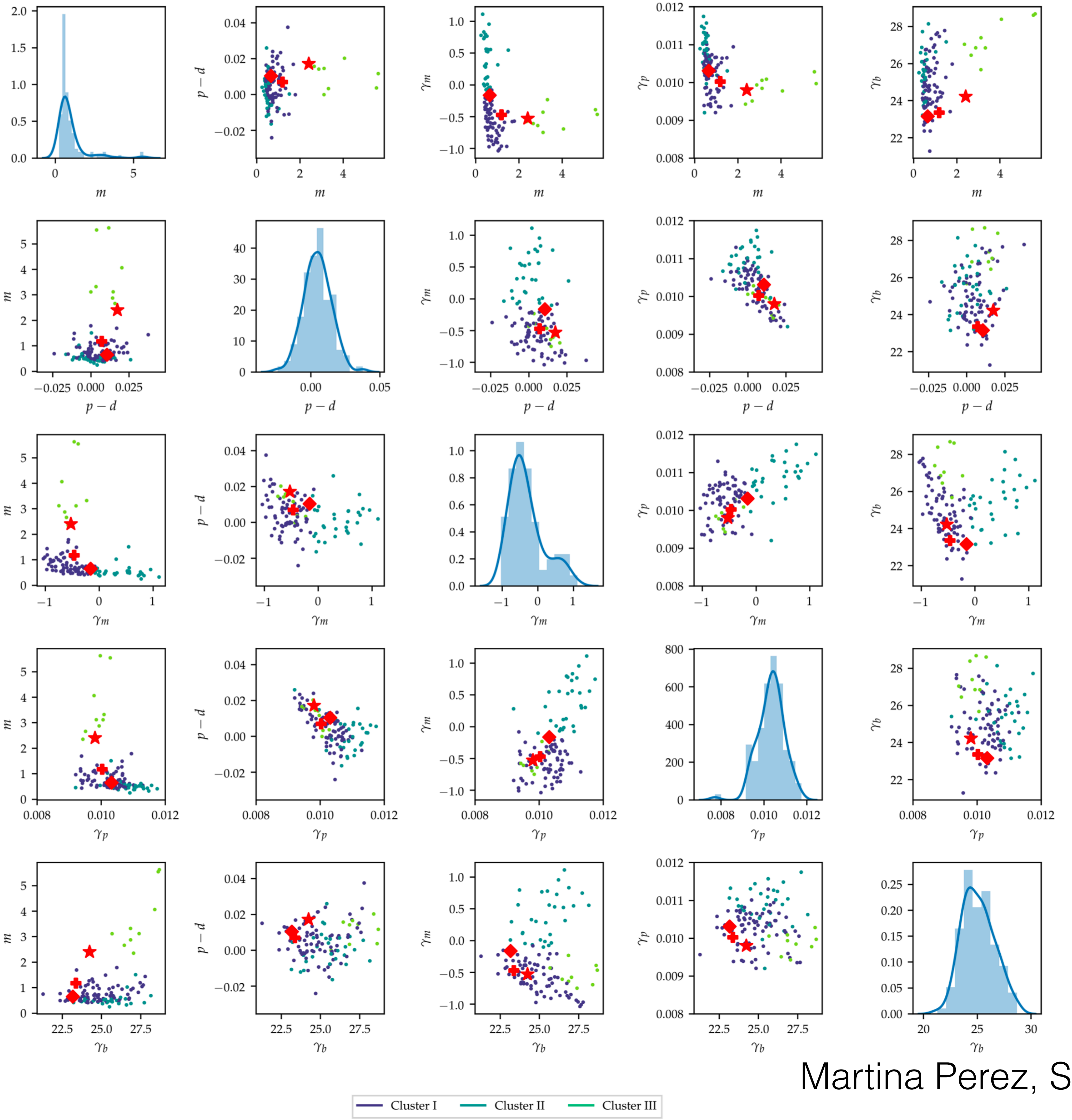
- Marginal posterior distributions



- Wildtype - motility strongly up-regulated in regions of high density compared to CDC42 and CDH5.
  - Consistent with current understanding of the roles of CDC42 and CDH5.

Martina Perez, Sailem and Baker, *PLOS Comp. Biol.* (2022).

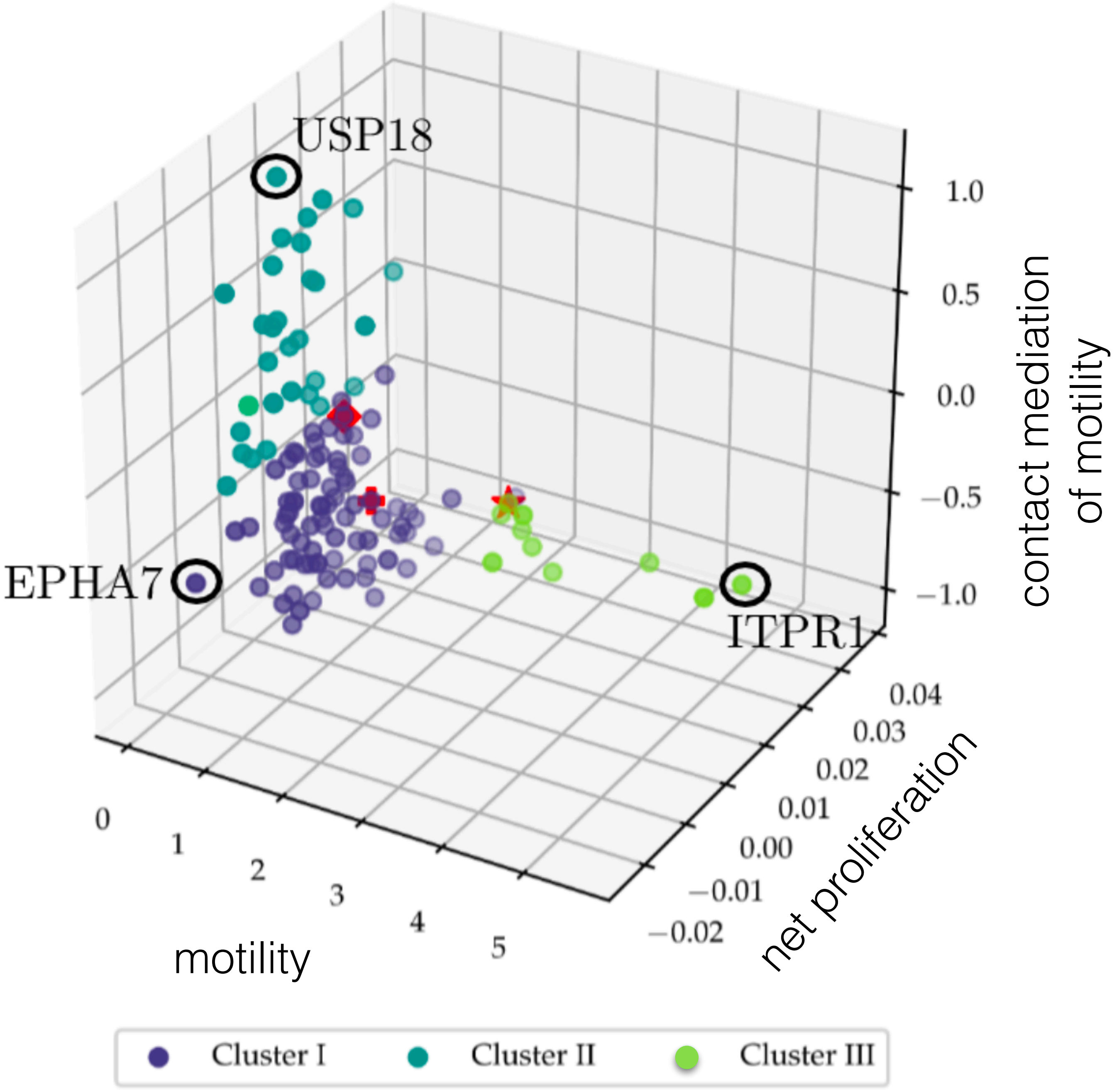
Interpreting results of RNAi screen

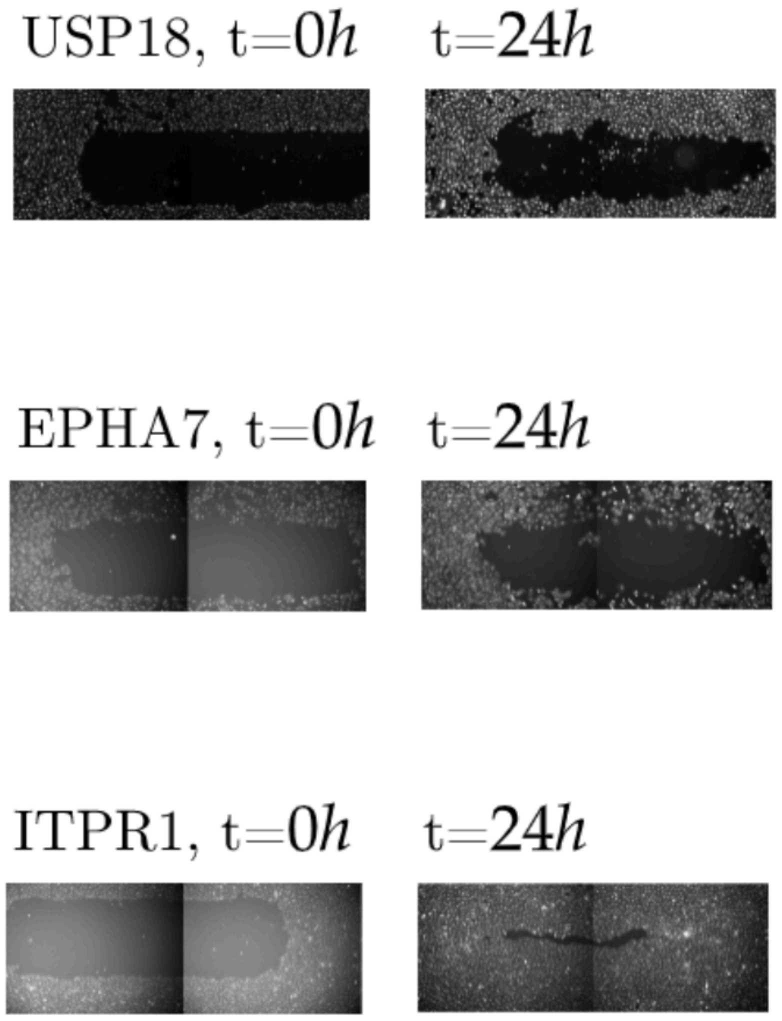


- Repeat for the other genes in the RNAi screen…
    - Plots of the posterior means for each knockdown.
    - Used K-means clustering to establish three main clusters in phenotype space.

Martina Perez, Sailem and Baker, *PLOS Comp. Biol.* (2022).

Interpreting results of RNAi screen





Martina Perez, Sailem and Baker, *PLOS Comp. Biol.* (2022).

Summary

- Possible to calibrate more complicated models to data, and use them to infer greater detail on the mechanisms driving wound closure in a range of gene knockdowns.

- Complex relationships between cell motility and proliferation drive wound closure.

- Mini-batch ABC provides a means to apply ABC approaches to high-throughput datasets, without huge computational costs.

Summary

- ABC is a fantastic tool for calibrating models to data since it relies only on forward simulation of the model.

- However, for modern mathematical biology studies the computational costs of naive forms of the method are prohibitive.

- Proposed three novel approaches to ABC - pre-conditioned, multi-fidelity and mini-batch.

- Importantly, ALL of these approaches can be combined with existing approaches e.g. ABC-SMC to provide further improvements.

- There's still much to do to optimise each approach!

Acknowledgements

UNIVERSITY OF OXFORD

Oxford
Tom Prescott
Simon Martina-Perez

Further afield
David Warne (QUT)
Mat Simpson (QUT)

THE ROYAL SOCIETY

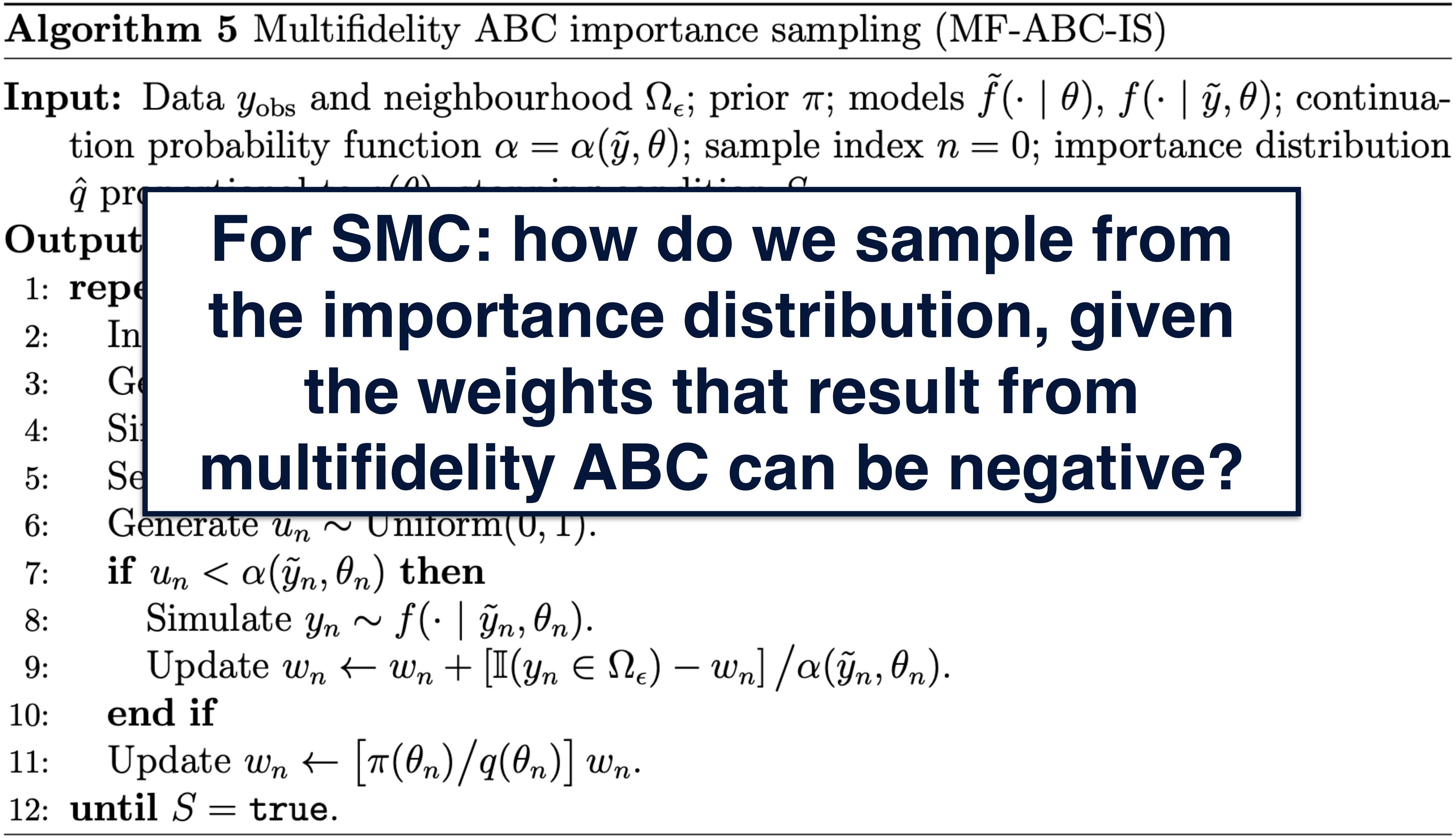
The Leverhulme Trust

UKRI Engineering and Physical Sciences Research Council

UKRI Biotechnology and Biological Sciences Research Council

Towards SMC: importance sampling

- First, need to integrate importance sampling into the multifidelity ABC framework:



Algorithm 5 Multifidelity ABC importance sampling (MF-ABC-IS)

Input: Data y_{obs} and neighbourhood Ω_ϵ; prior π; models $\tilde{f}(\cdot \mid \theta)$, $f(\cdot \mid \tilde{y}, \theta)$; continuation probability function $\alpha = \alpha(\tilde{y}, \theta)$; sample index $n = 0$; importance distribution $\hat{q}$ pr

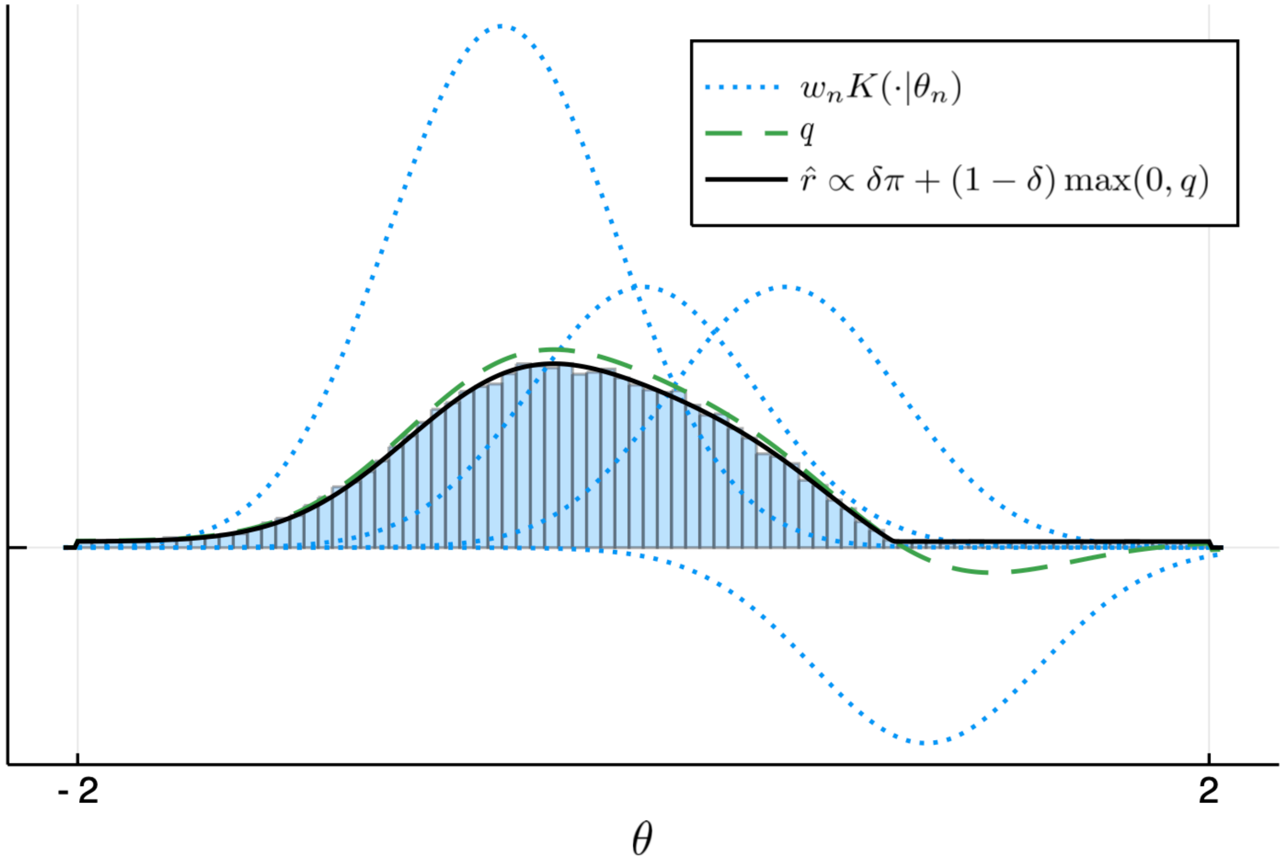
Output

```
repe
    In
    Ge
    Si
    Se
    Generate $u_n \sim$ Uniform(0, 1).
    if $u_n < \alpha(\tilde{y}_n, \theta_n)$ then
        Simulate $y_n \sim f(\cdot \mid \tilde{y}_n, \theta_n)$.
        Update $w_n \leftarrow w_n + [\mathbb{I}(y_n \in \Omega_\epsilon) - w_n] / \alpha(\tilde{y}_n, \theta_n)$.
    end if
    Update $w_n \leftarrow [\pi(\theta_n)/q(\theta_n)] \, w_n$.
until $S =$ true.
```

For SMC: how do we sample from the importance distribution, given the weights that result from multifidelity ABC can be negative?

Towards multifidelity SMC-ABC

- Use defensive importance sampling, first defining a new (non-negative) importance distribution.



- Estimate continuation probabilities for each generation "on the fly", using information from the previous generations.

MF ABC SMC in action

- Kuramoto oscillator network:

$$\dot{\phi}_i = \omega_i + \frac{K}{M}\sum_{j=1}^{M}\sin\left(\phi_j - \phi_i\right)$$

angular velocities drawn
from Cauchy distribution
median - ω_0
dispersion - γ

- Low-fidelity model - based on tracking Daido order parameters:

$$Z_n(t) = \frac{1}{M}\sum_{j=1}^{M}\exp(in\phi_j)$$

assume $Z_n(t) = Z_1(t)^n$

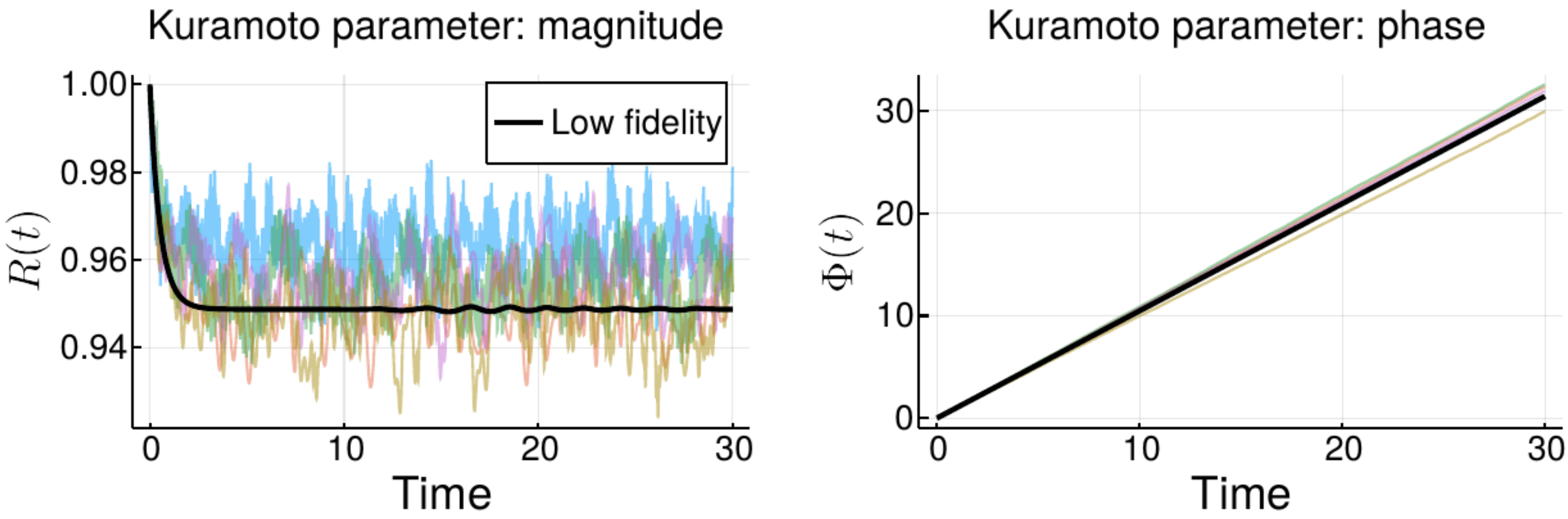
to get

$$\dot{\tilde{R}} = \left(\frac{K}{2} - \gamma\right)\tilde{R} - \frac{K}{2}\tilde{R}^3$$ (magnitude)

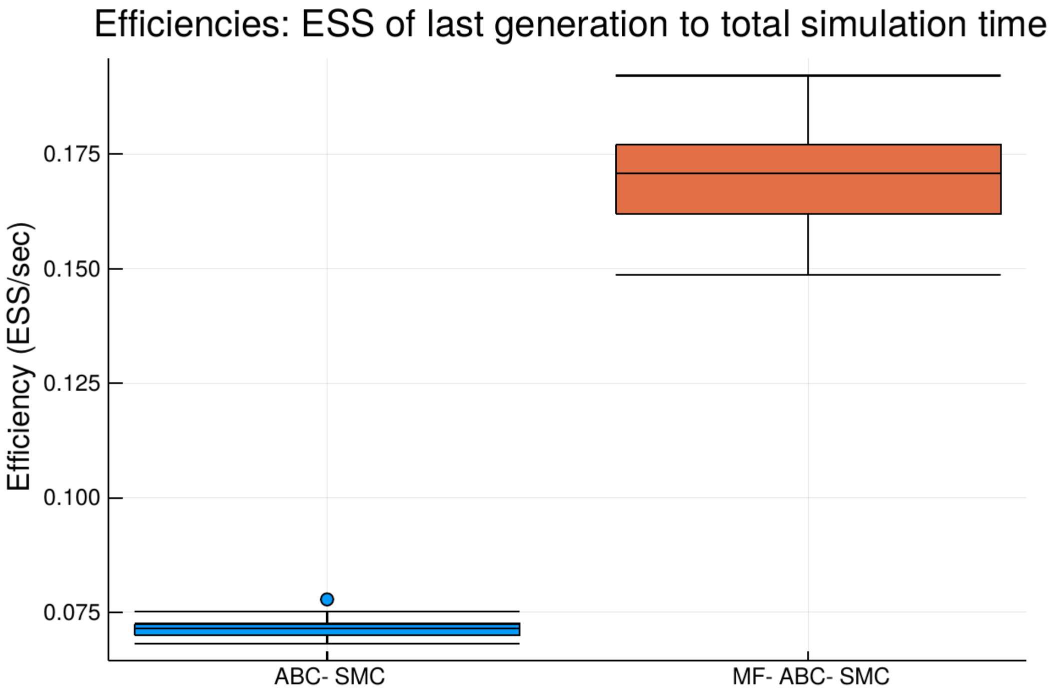
$$\dot{\tilde{\Phi}} = \omega_0$$ (phase)

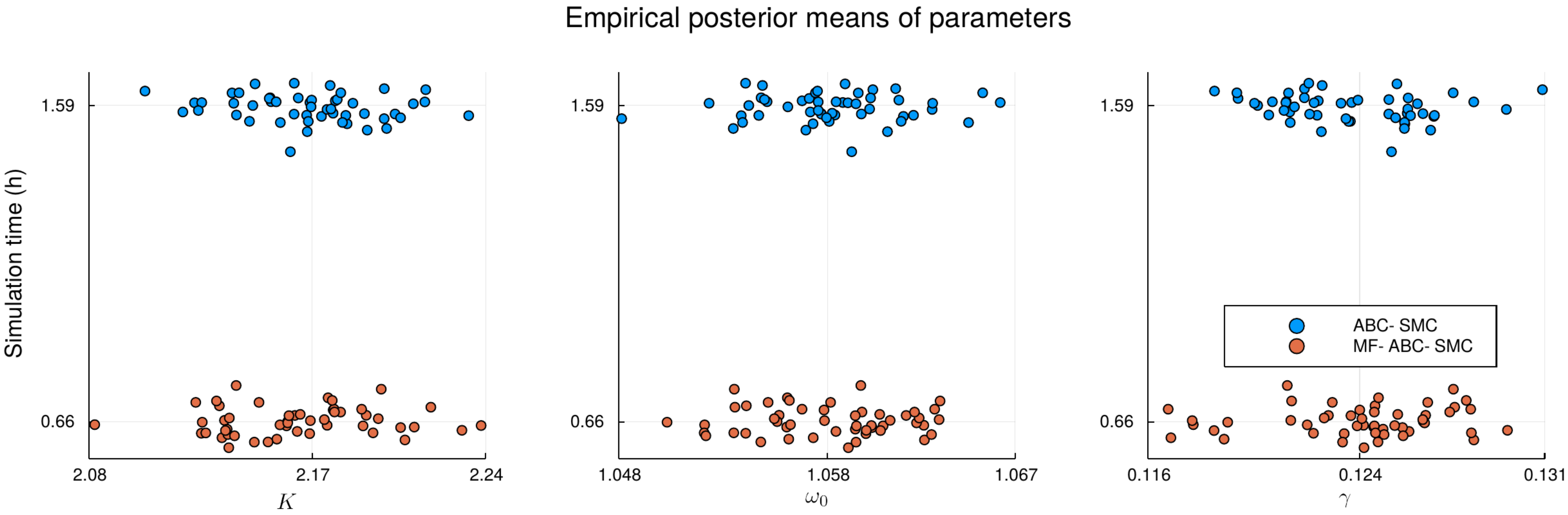
MF ABC SMC in action

- Typical simulation output:



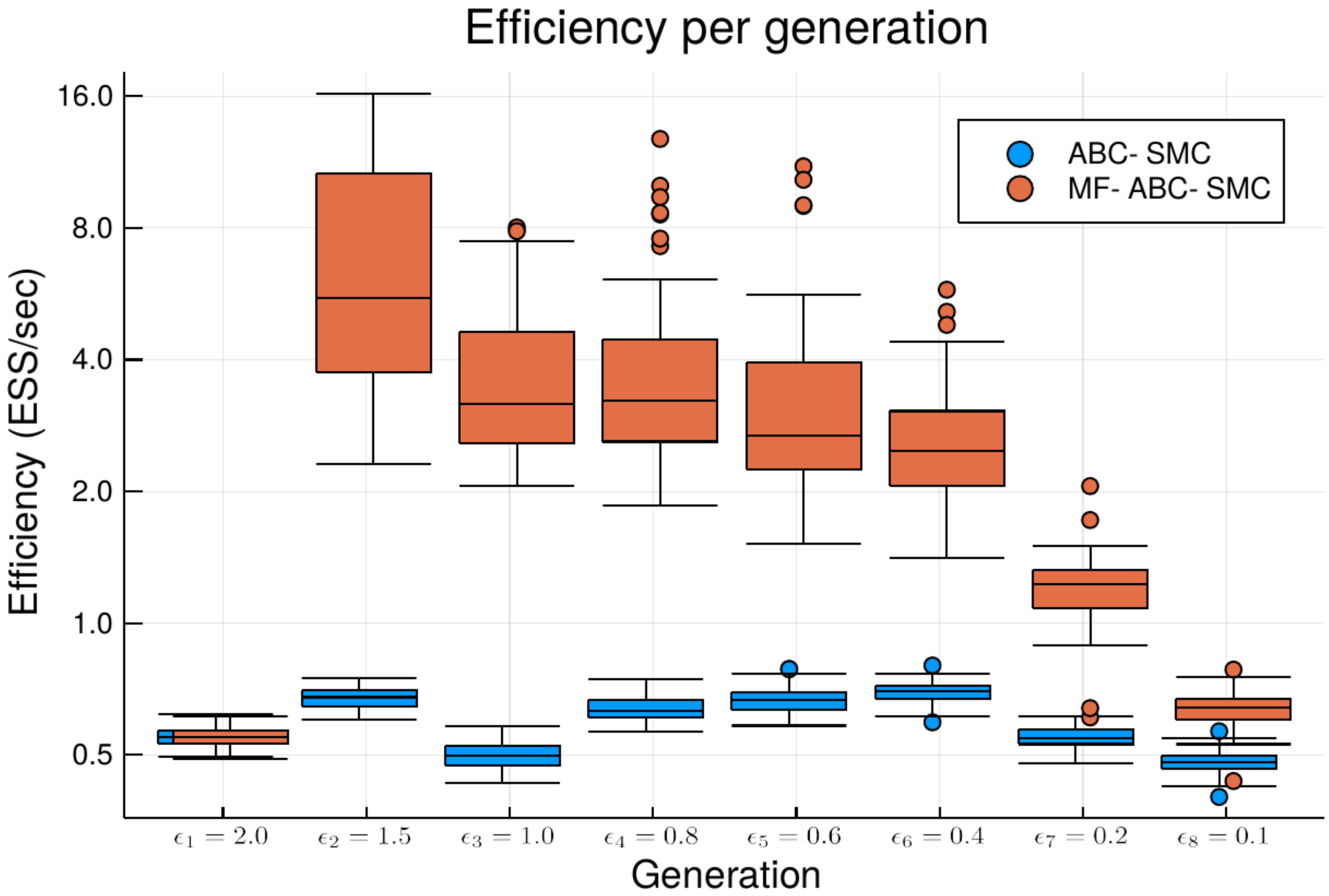
MF ABC SMC in action



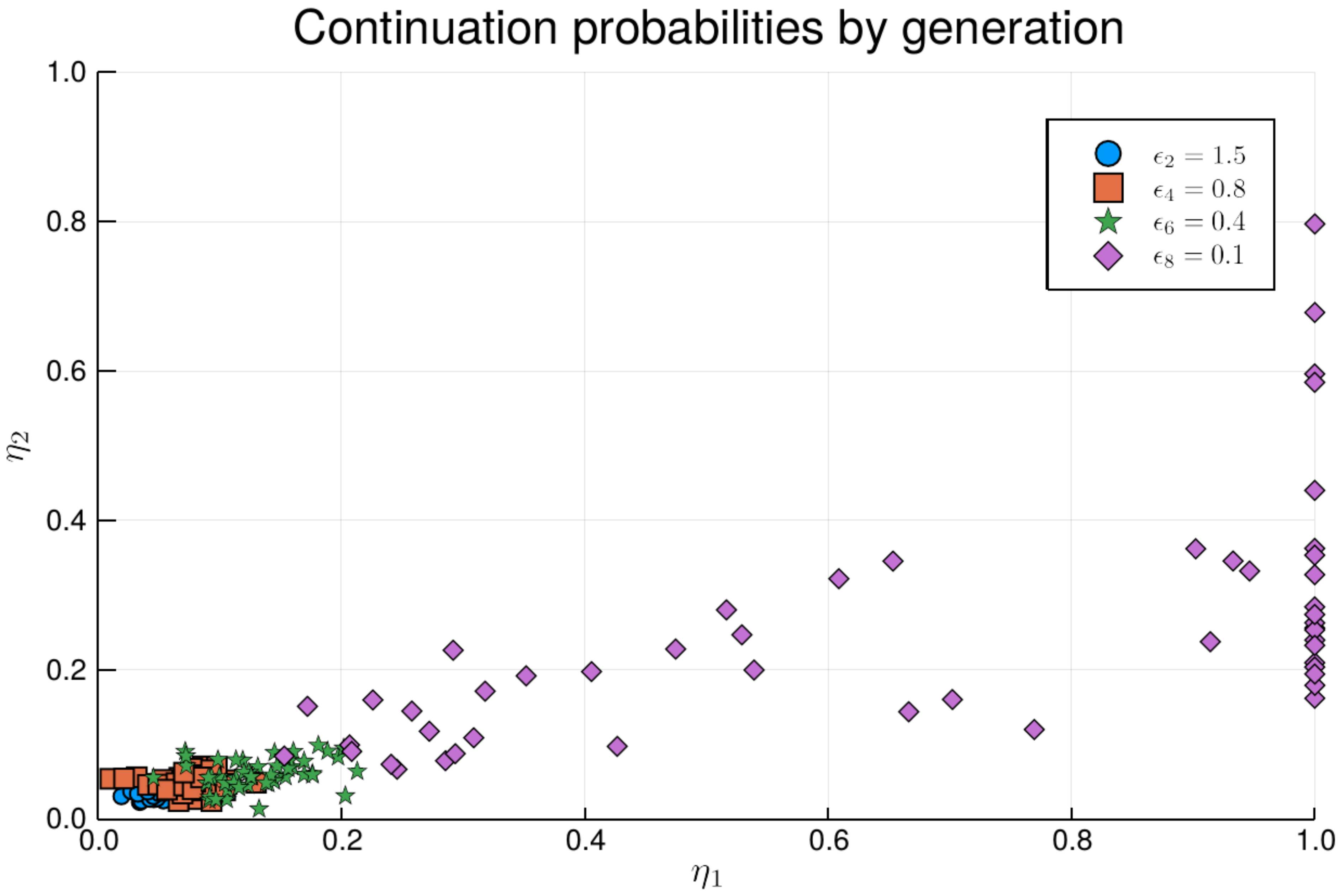


Towards multifidelity SMC-ABC

- Stopping criterion at each generation: $\mathrm{ESS} \geq 400$.



Towards multifidelity SMC-ABC



probability of requiring high-fidelity model simulation given low-fidelity model **far**

probability of requiring high-fidelity model simulation given low-fidelity model **close**