LORENZO TARRICONE - 14.03.2024

OPTIMALLY DESIGNED MODEL SELECTION

FOR SYNTETIC BIOLOGY

ETH zürich

OVERVIEW

1. How to use OED in Synthetic Biology
2. The Model(s)
3. Frequentist vs Bayesian approach
4. Inferring parameters and first model selection
   I. Frequentist approach: Evolutionary Algorithm
   II. Bayesian approach: NUTS Algorithm (HMC)
5. Optimal Experimental Design
6. Stability properties of the solution (if time)
7. Conclusion
8. Recap

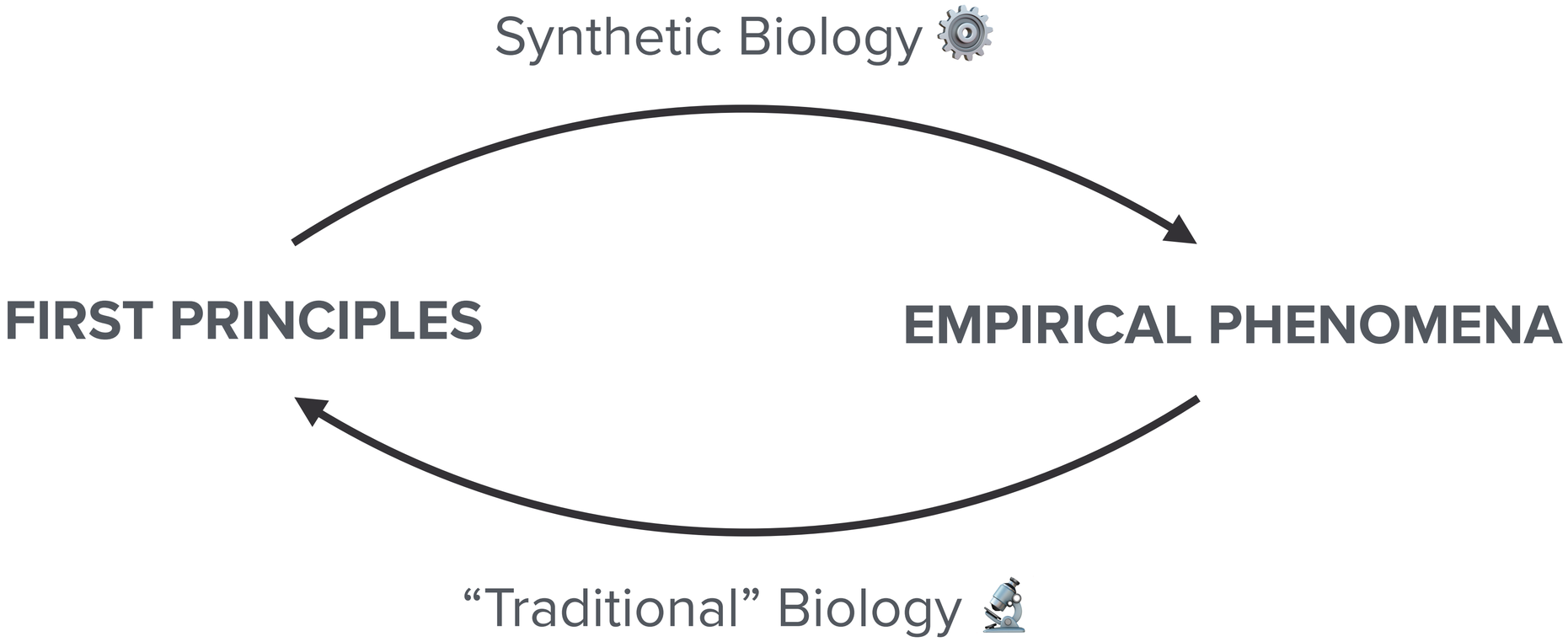
OVERVIEW

1. How to use OED in Synthetic Biology
2. The Model(s)
3. Frequentist vs Bayesian approach
4. Inferring parameters and first model selection
   I. Frequentist approach: Evolutionary Algorithm
   II. Bayesian approach: NUTS Algorithm (HMC)
5. Optimal Experimental Design
6. Stability properties of the solution (if time)
7. Conclusion
8. Recap

PAY ATTENTION TO THIS LINE!

- easy and intuitive
- a bit more involved
- sloppy and mathsy

HOW TO USE OED IN SYNTHETIC BIOLOGY



HOW TO USE OED IN SYNTHETIC BIOLOGY

HOW TO USE OED IN SYNTHETIC BIOLOGY

1. Models are essential for studying synthetic circuits
    I. Biology to include?

HOW TO USE OED IN SYNTHETIC BIOLOGY

1. Models are essential for studying synthetic circuits
   I. Biology to include?
2. We need experimental data to assess parameters of the model

HOW TO USE OED IN SYNTHETIC BIOLOGY

1. Models are essential for studying synthetic circuits
   I. Biology to include?
2. We need experimental data to assess parameters of the model
   I. Time ⏱️

HOW TO USE OED IN SYNTHETIC BIOLOGY

1. Models are essential for studying synthetic circuits
   I. Biology to include?
2. We need experimental data to assess parameters of the model
   I. Time ⏱️
   II. Cost 💰

HOW TO USE OED IN SYNTHETIC BIOLOGY

1. Models are essential for studying synthetic circuits
   I. Biology to include?
2. We need experimental data to assess parameters of the model
   I. Time ⏱️
   II. Cost 💰
   III. Nonlinearity 📈

HOW TO USE OED IN SYNTHETIC BIOLOGY

1. Models are essential for studying synthetic circuits
   I. Biology to include?
2. We need experimental data to assess parameters of the model
   I. Time ⏱️
   II. Cost 💰 → **(partial) SOLUTION**
   III. Nonlinearity 📈

HOW TO USE OED IN SYNTHETIC BIOLOGY

1. Models are essential for studying synthetic circuits
   I. Biology to include?
2. We need experimental data to assess parameters of the model
   I. Time ⏱️
   II. Cost 💰
   III.Nonlinearity 📈

→ **(partial) SOLUTION**

↓

Optimal Experimental Design!

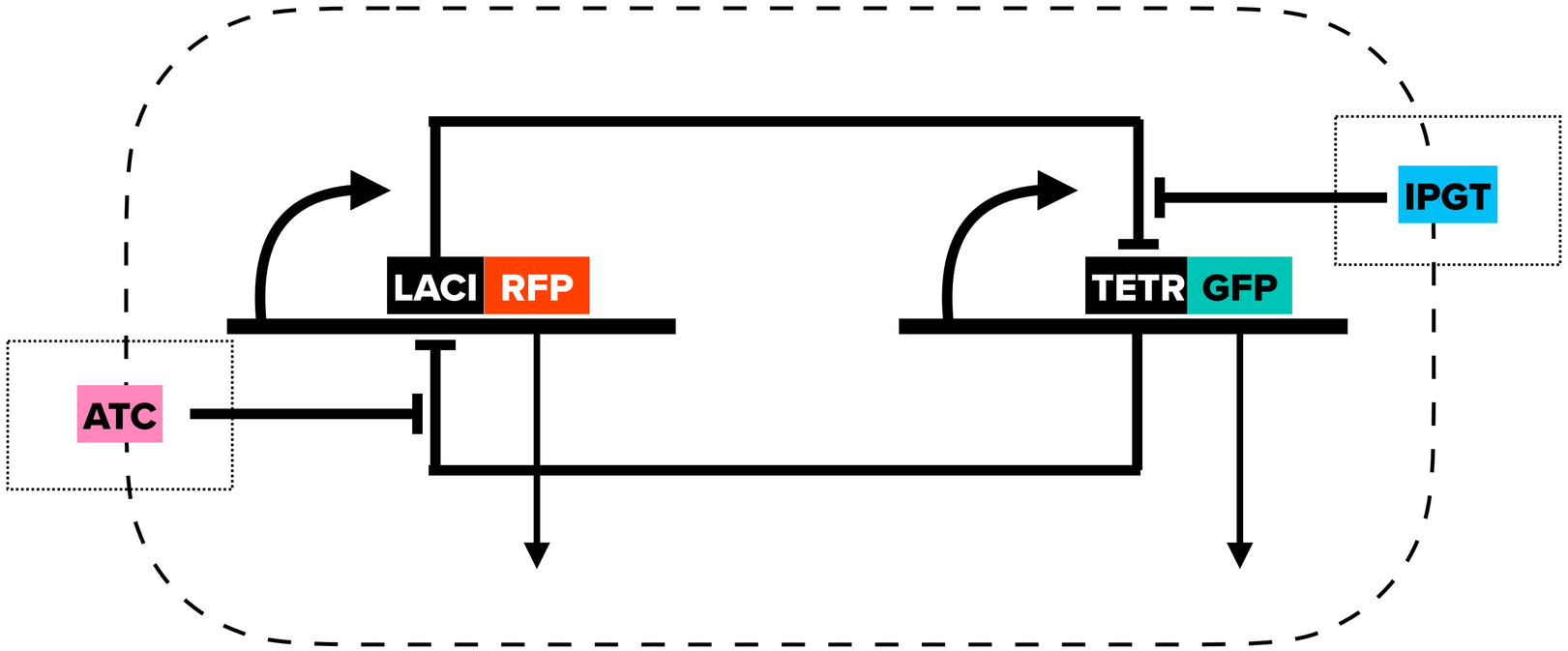
Try to balance tradeoff between information gained and experimental effort

Can we use OED also to solve problem I?

(i.e. can we use OED for model selection?)

THE MODEL(S)

- **Toggle Switch:** two transcriptional inhibitors (**LacI** and **TetR**) and two inducers (**aTc** and **IPTG**) in *E.Coli*

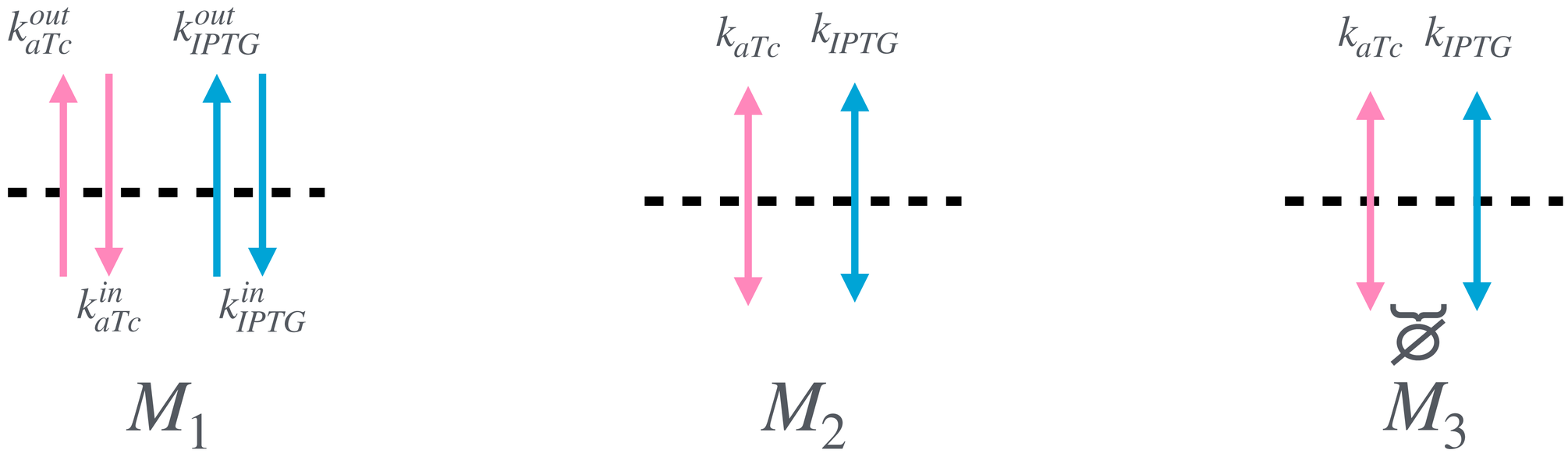


Fixed

Changeable

THE MODEL(S)

- M_1 Non symmetrical influx/outflux of aTc/IPTG through the membrane
- M_2 Same influx/outflux rate (simple diffusion)
- M_3 Same influx/outflux rate (simple diffusion) + Dilution due to cell growth



FREQUENTIST VS BAYESIAN

MODELS

↓

PARAMETERS
of the model

↓

first
MODEL SELECTION

↓

OED
final choice

FREQUENTIST
approach

BAYESIAN
approach

INFERRING PARAMETERS: FREQUENTIST

- For a given set of parameters you can simulate the system and compare to experimental data
  - Loss function: weighted Sum of Squared Error (SSE) [$w_k = \frac{1}{\bar{\sigma}_k}$, k exp.]
- For a each parameter i we have an initial estimate μ_i (other publication)
- **Enhanced Scatter Search** Algorithm (Evolutionary Algorithm) is used

INFERRING PARAMETERS: FREQUENTIST

ESS Algorithm initialisation (for each model):

Select a model

M_1
or
M_2
or
M_3

x 100 times

$$\mu_1 \longrightarrow \mathcal{U}([\frac{1}{10}\mu_1, 10\mu_1])$$

$$\mu_2 \longrightarrow \mathcal{U}([\frac{1}{10}\mu_2, 10\mu_2])$$

$$\vdots$$

$$\mu_I \longrightarrow \mathcal{U}([\frac{1}{10}\mu_I, 10\mu_I])$$

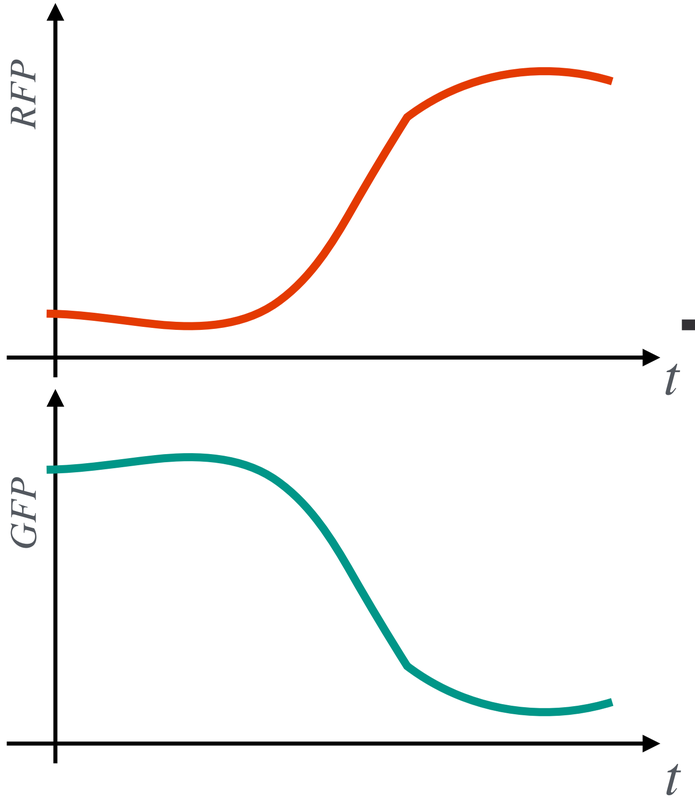
Set of parameters
$[1,2,\ldots,I]$

e.g. the 9th sample is

$$\begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_I \end{bmatrix} = \begin{bmatrix} 3 \\ 0.2 \\ \vdots \\ 54 \end{bmatrix}$$

is the 9th **individual** of the population

Simulate forward

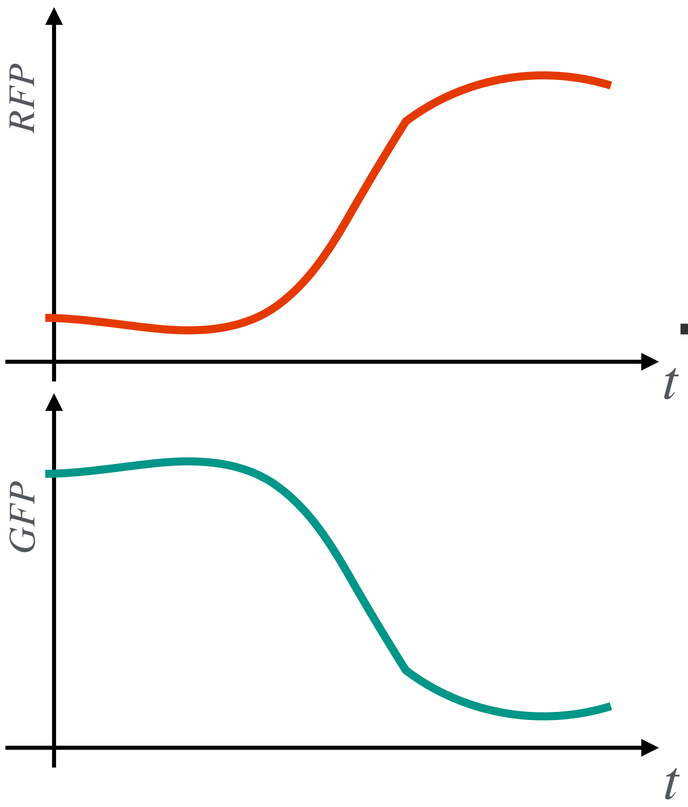


With those parameters

INFERRING PARAMETERS: FREQUENTIST

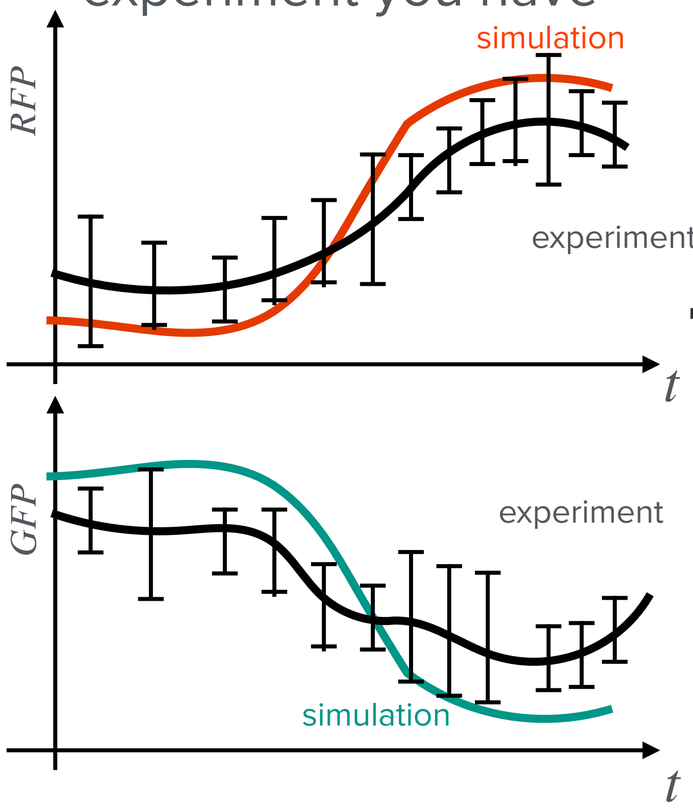
ESS Algorithm evaluation (for each individual):

Simulate forward



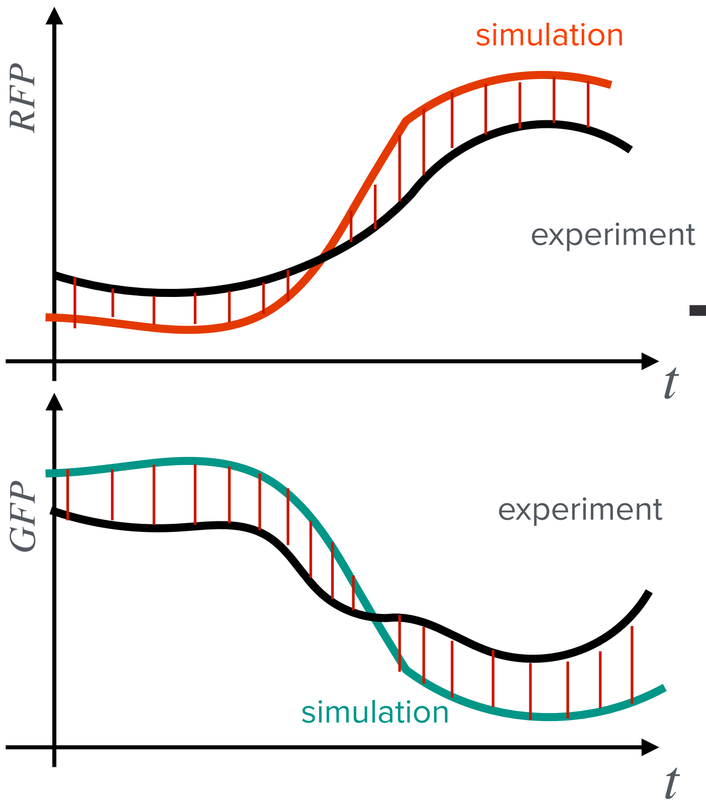
With those parameters

Compare to the experiment you have



and calculate residuals

And calculate square residuals

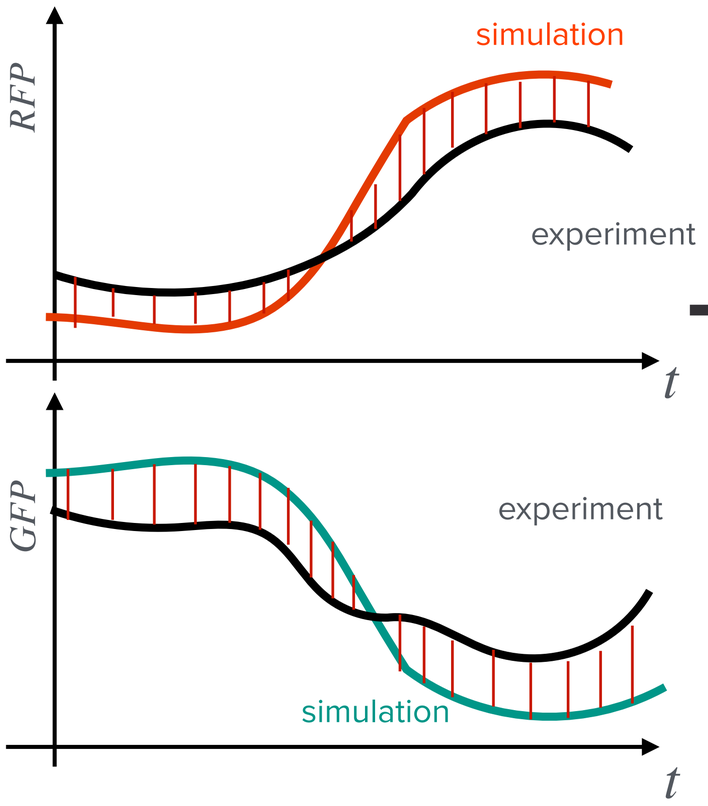


and calculate residuals

INFERRING PARAMETERS: FREQUENTIST

ESS Algorithm evaluation (for each individual):

And calculate square residuals



and calculate residuals

Do it for all 10 experiments and sum the residuals with weight

$$w_k = \frac{1}{\bar{\sigma}_k} \text{ for } k \in \{1,2,...10\}$$

This is how you obtain the (un)**fitness value** for your individual!

INFERRING PARAMETERS: FREQUENTIST

ESS Algorithm in action (for each model):

Initialise 100 individuals and calculate their fitness value

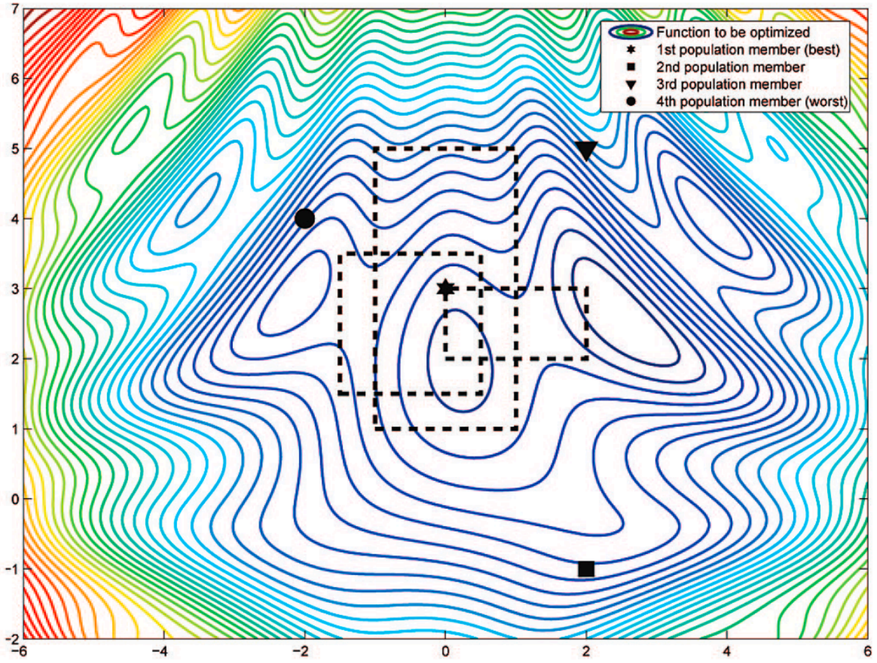
COMBINATION RULE

Select best $b/2$ individuals

Sample the other $b/2$ from the rest of the population

COMBINATION METHOD

For all couple of points create biased hyper rectangle to sample from



POPULATION UPDATE

For each parent: if one of the $b-1$ solutions is fitter than it, substitute the parent

(ENHANCING)

If parent substituted: continue the search in the direction parent/ fittest child, possibly iterating three times

INFERRING PARAMETERS: FREQUENTIST

ESS Algorithm in action (for each model):

We now have

$$\hat{\theta} = \begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \\ \vdots \\ \hat{\theta}_I \end{bmatrix}$$

→ How can we get a **distribution**? → we use the covariance matrix and the Cramer Rao lower bound!

↓

This will create confidence intervals and a multivariate Gaussian

INFERRING PARAMETERS: FREQUENTIST

MODEL SELECTION:

$$\hat{\theta} = \begin{bmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \\ \vdots \\ \hat{\theta}_I \end{bmatrix}$$

→ Simulate forward → Use AIC_i to rank the models

$$AIC_i = 2d_i + \sum_{k \in \{g,r\}} \underbrace{\sum_t \frac{(\hat{y}_{k,t}(\theta_i, u) - y_{k,t}(\theta_i, u))^2}{\sigma^2_{k,t}(u)}}_{\text{Fitness of our Evolutionary Algorithm}}$$

INFERRING PARAMETERS: BAYESIAN

We want $P(\theta_i \mid \{E\}) \;\; i \in \{1,2,3\}$

INFERRING PARAMETERS: BAYESIAN

We want $P(\theta_i \mid \{E\}) \;\; i \in \{1,2,3\}$

From Bayes: $P(\theta_i \mid \{E\}) \propto P(\{E\} \mid \theta_i, M_i) P(\theta_i, M_i)$

INFERRING PARAMETERS: BAYESIAN

We want $P(\theta_i \mid \{E\}) \;\; i \in \{1,2,3\}$

From Bayes: $P(\theta_i \mid \{E\}) \propto P(\{E\} \mid \theta_i, M_i) P(\theta_i, M_i)$

That we cannot use directly 🥲

INFERRING PARAMETERS: BAYESIAN

We want $P(\theta_i \mid \{E\}) \;\; i \in \{1,2,3\}$

From Bayes: $P(\theta_i \mid \{E\}) \propto P(\{E\} \mid \theta_i, M_i) P(\theta_i, M_i)$

That we cannot use directly 🥲

↓

INFERRING PARAMETERS: BAYESIAN

We want $P(\theta_i \mid \{E\}) \;\; i \in \{1,2,3\}$

From Bayes: $P(\theta_i \mid \{E\}) \propto P(\{E\} \mid \theta_i, M_i) P(\theta_i, M_i)$

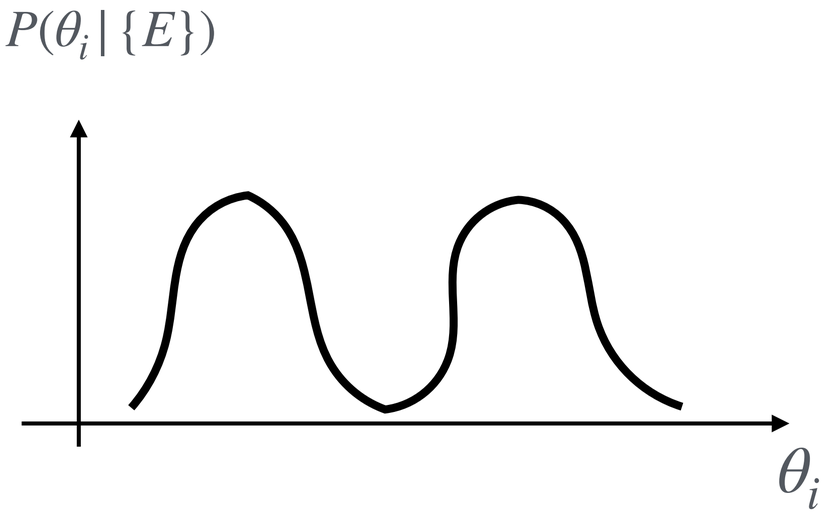
That we cannot use directly 🥲

↓

Refresher: **Metropoli-Hastings** (for each model):

INFERRING PARAMETERS: BAYESIAN

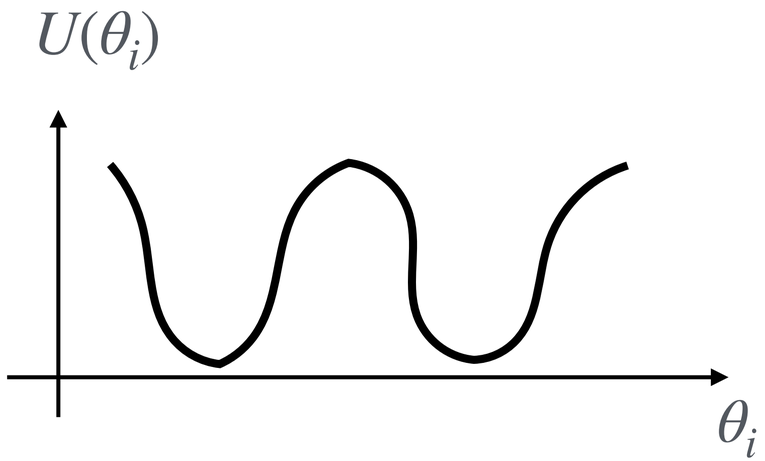
Refresher: **Metropoli-Hastings**:



"Probability view"

Boltzmann Distribution

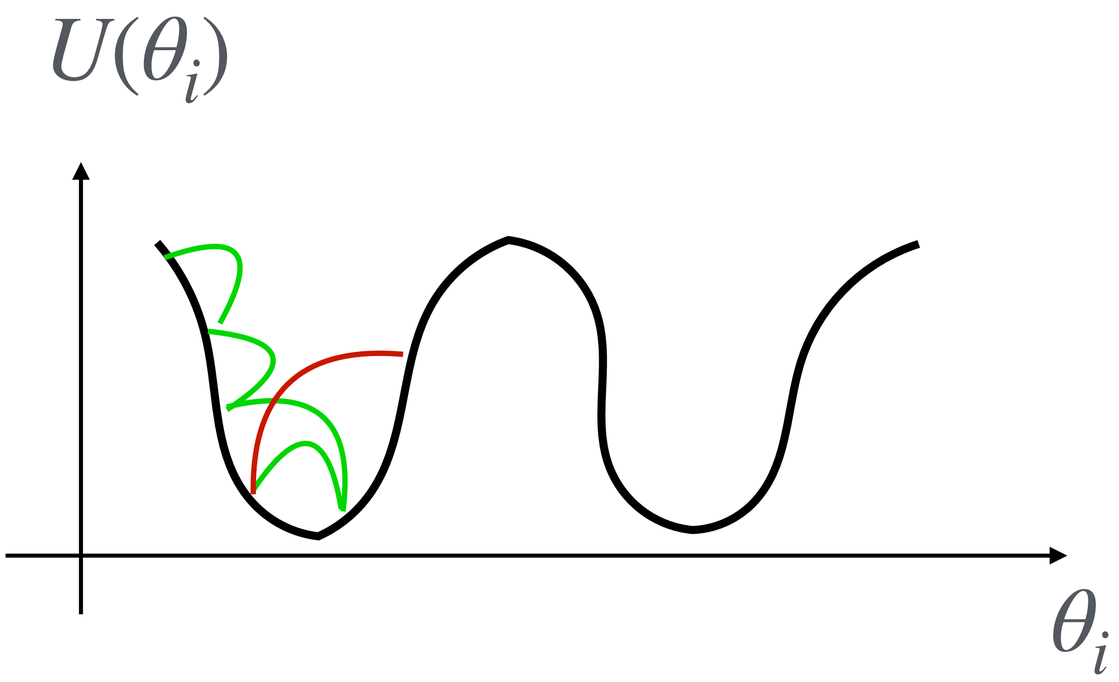
$$P(H) \propto e^{-\frac{H}{k_B T}}$$



"**Potential** energy view"

INFERRING PARAMETERS: BAYESIAN

Refresher: **Metropoli-Hastings**:



MH not really good at going around high dimensional/highly correlated probability distributions

INFERRING PARAMETERS: BAYESIAN

Hamiltonian Monte Carlo: Using both Potential and Kinetic energy

$$H(\theta, m) = U(\theta) + K(m) \text{ where } K(m) = \sum_{j=1}^{I} \frac{m_j^2}{2mass} \text{ and } U(\theta) = -\,log(P(E\,|\,\theta)P(\theta))$$

$$\downarrow$$

$$P(\theta, m) \propto P(E\,|\,\theta)P(\theta)e^{-\frac{m^2}{2}} = P(E\,|\,\theta)P(\theta)\mathcal{N}(m; 0,1)$$

$$\downarrow$$

Meaning that we can just sample from this joint and then discard the m values to sample from the posterior

INFERRING PARAMETERS: BAYESIAN

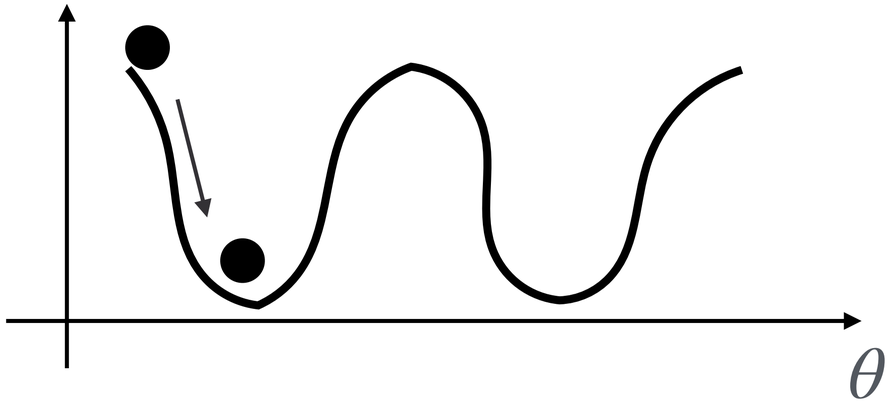
Hamiltonian Monte Carlo: Intuition

(Here there are some tricks to maintain detail balance)

Mechanical energy is conserved!

We (almost) always accept the next move!

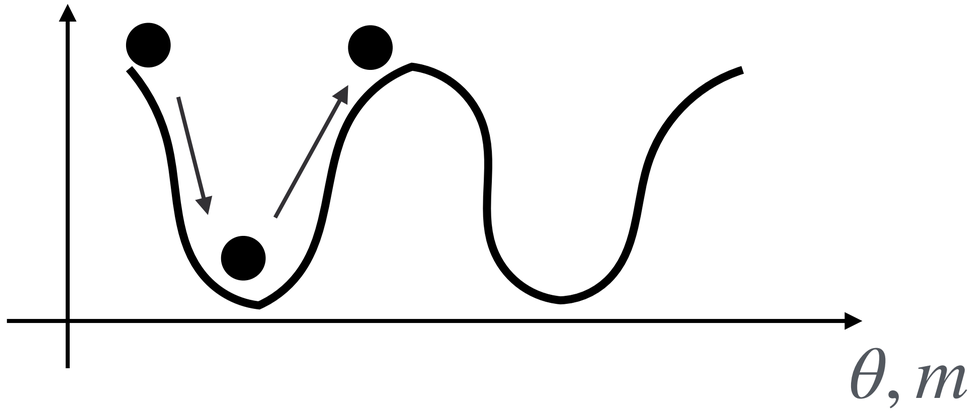
$U(\theta)$



Metropolis-Hastings

Random walk to move around

$H(\theta, m) = U(\theta) + K(m)$



Hamiltonian Monte Carlo

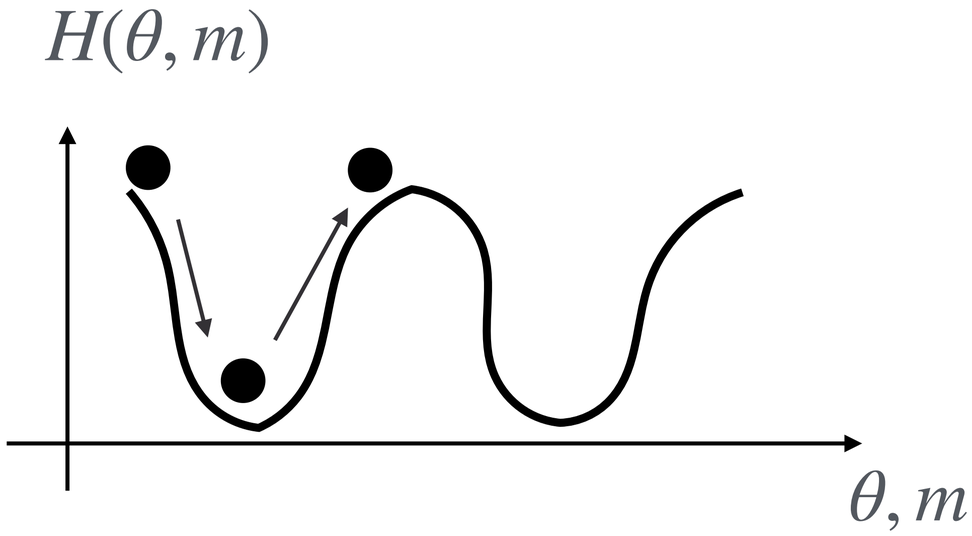
sample m and then Newton to move around

INFERRING PARAMETERS: BAYESIAN

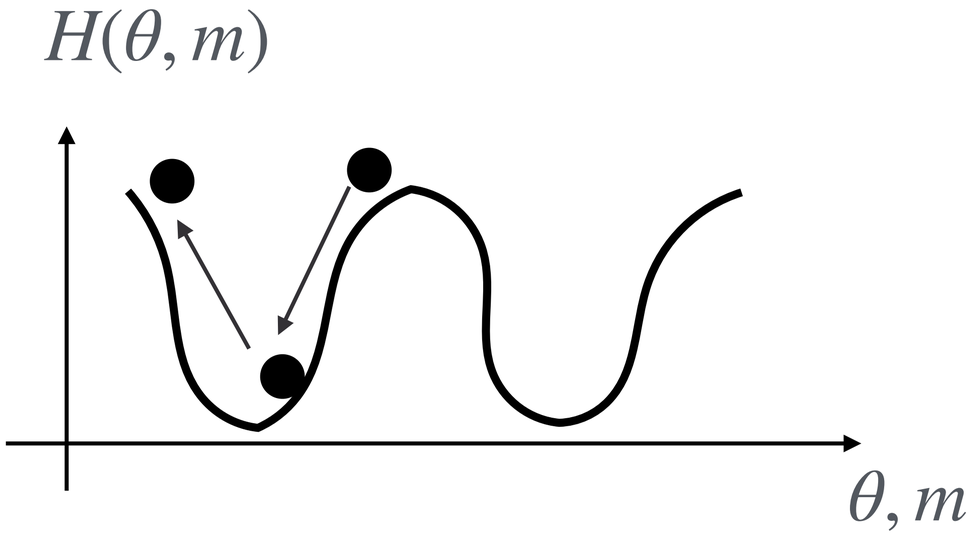
No-U-Turns Algorithm: Improving HMC

PROBLEMS:

- Inaccuracy in integrating the equation of motion
- Trajectory might go back and forth many times



Step 1



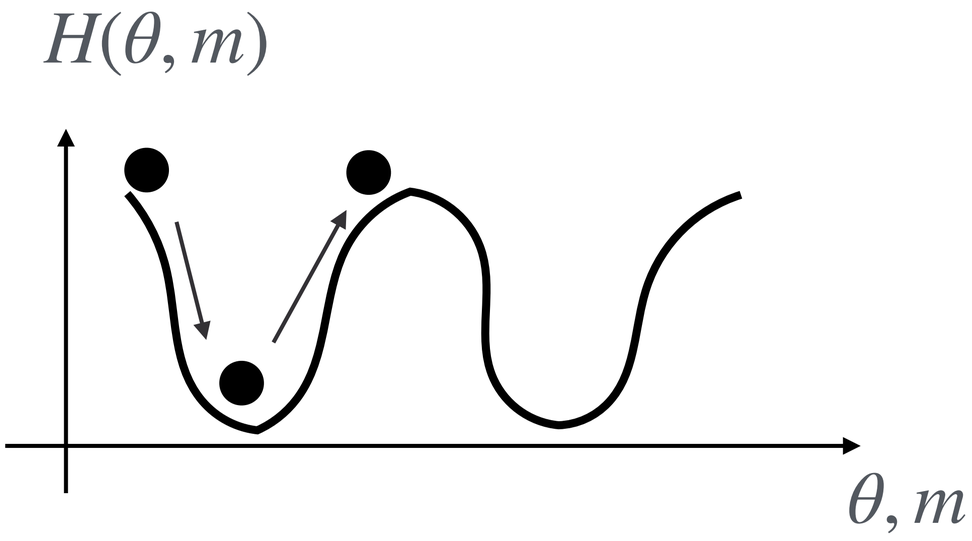
Step 2

INFERRING PARAMETERS: BAYESIAN

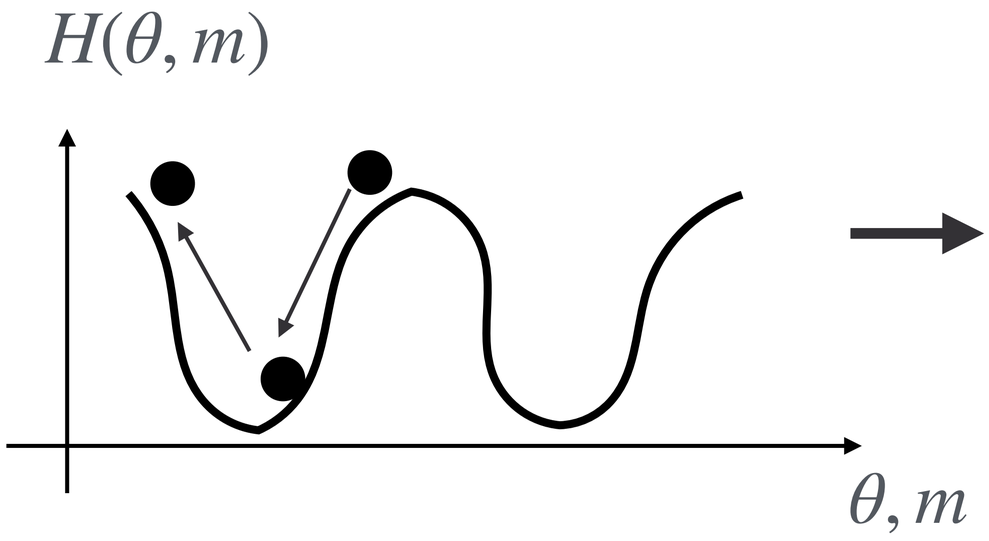
No-U-Turns Algorithm: Improving HMC

PROBLEMS:

- Inaccuracy in integrating the equation of motion
- Trajectory might go back and forth many times



Step 1



Step 2

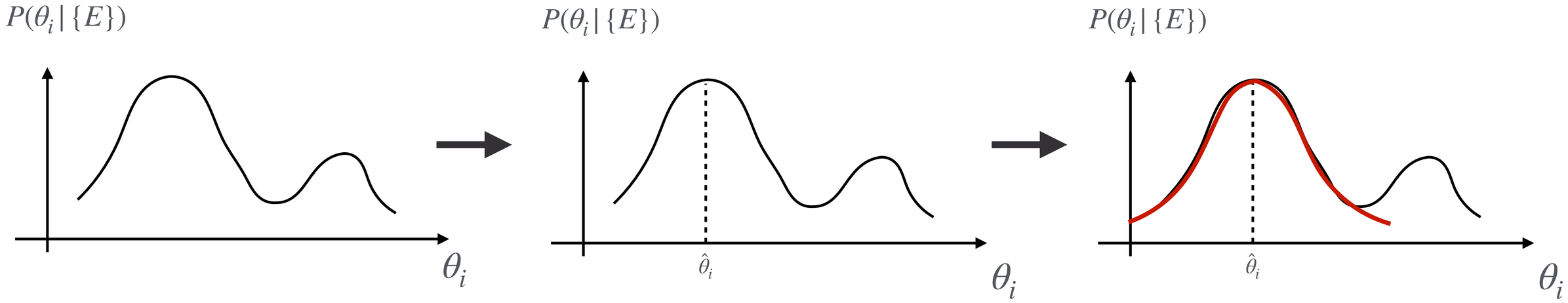
Similar m means similar points sampled, computationally useless!

STOP SAMPLING

INFERRING PARAMETERS: BAYESIAN VS FREQUENTIST

MODEL SELECTION

We use **Laplace approximation**

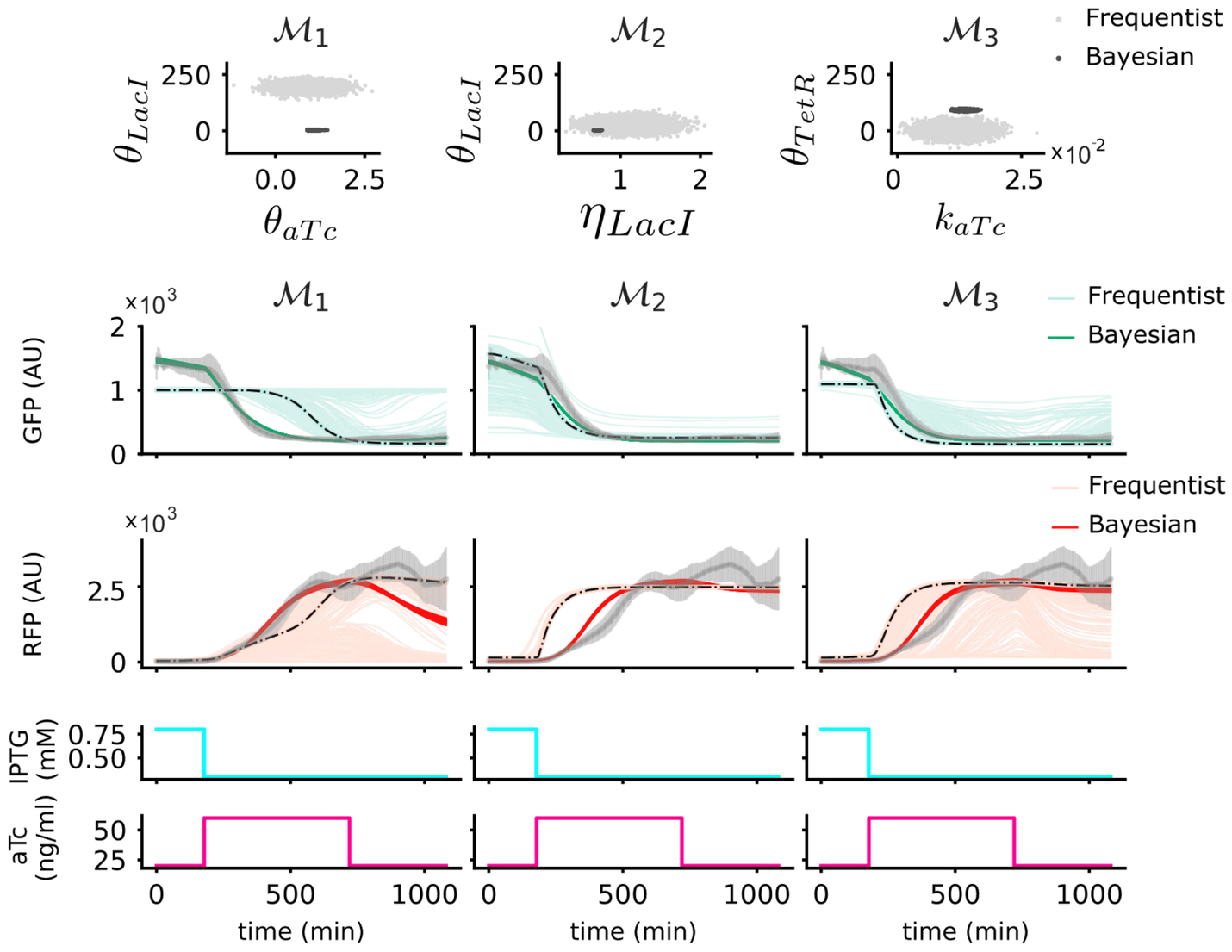


Posterior

Find argmax with an MCMC method

Fit a (multivariate) Gaussian and get the Gaussian likelihood and the evidence for each model

INFERRING PARAMETERS: BAYESIAN VS FREQUENTIST



Bayesian approach more “sure” than Frequentist approach

Bayesian approach better at predicting real data overall

(8k samples from each model’s distribution)

OPTIMAL EXPERIMENTAL DESIGN

MODEL SELECTION RESULTS

FREQUENTIST
Prefers models M_1 and M_3

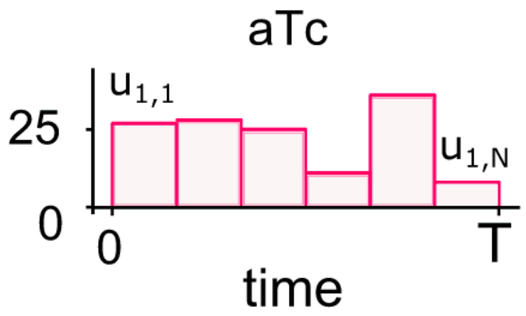
BAYESIAN
Prefers models M_1 and M_2

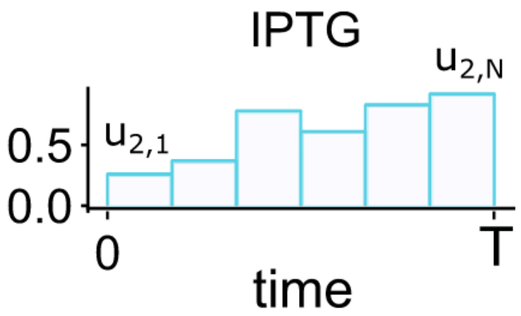
OPTIMAL ESPERIMENTAL DESIGN (OED)
to take the final decision!

Finding the best next experiment to discriminate between the two candidate models

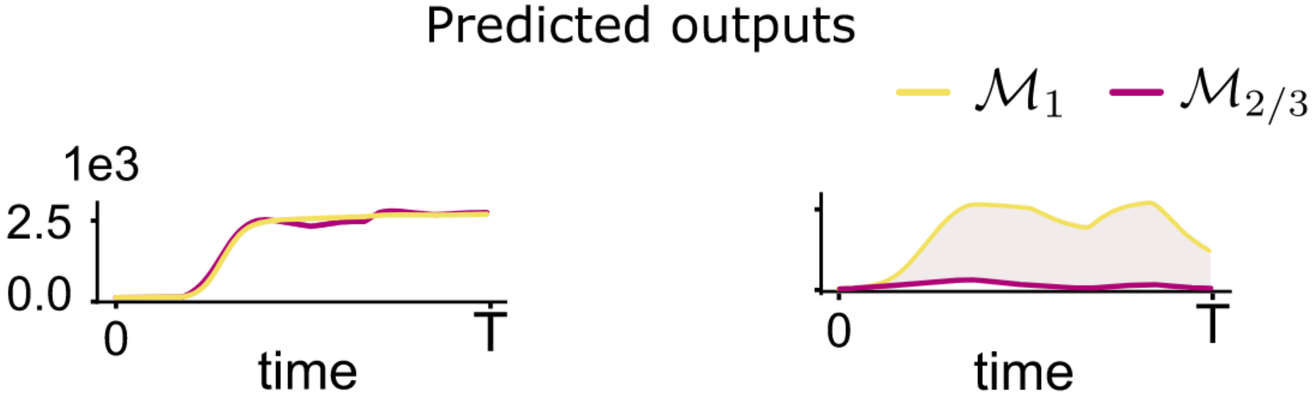
OPTIMAL EXPERIMENTAL DESIGN

INTUITION





We can choose a step-wise input



BAD INPUT ❌

GOOD INPUT ✅

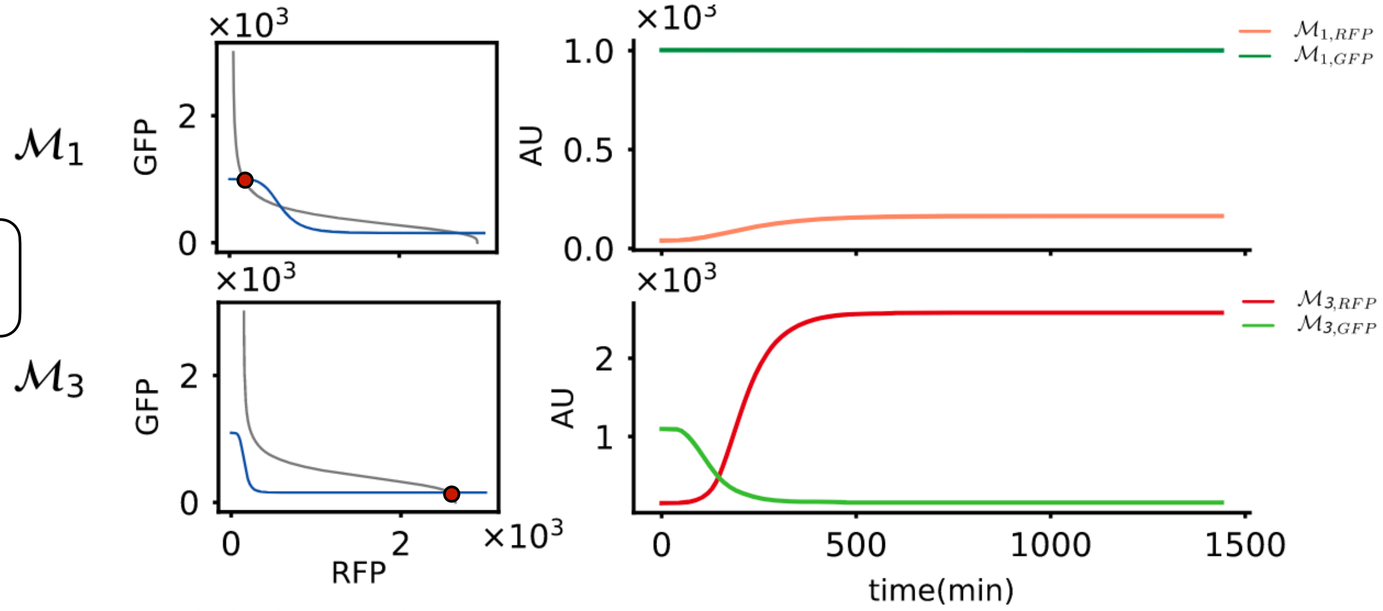
To maximize the divergence between the simulated values (we fix the best parameters)

OPTIMAL EXPERIMENTAL DESIGN

INTUITION

IPTG = 0.25mM
aTc = 23 ng/ml

for this input



We have this dynamics with final steady states

THE TWO MODELS DISAGREE!

According to:

M_1 GFP dominates

M_3 RFP dominates

↓

Good experiment

OPTIMAL EXPERIMENTAL DESIGN

FREQUENTIST

Enhanced Scatter Sampling Algorithm with new fitness function!

$$d(M_\alpha, M_\beta) = \sqrt{\sum_{t=1}^{T} (\hat{y}_{t,\alpha}(\theta_\alpha, u_j) - \hat{y}_{t,\beta}(\theta_\beta, u_j))^2}$$

800 loops of the algorithm

OPTIMAL EXPERIMENTAL DESIGN

BAYESIAN

"We want to maximise the distance between the predictive posteriors with Bhattacharyya distance using Bayesian Optimisation"

OPTIMAL EXPERIMENTAL DESIGN

BAYESIAN

"We want to maximise the distance between the predictive posteriors with Bhattacharyya distance using Bayesian Optimisation"

↓

Probably you right now: 😕

OPTIMAL EXPERIMENTAL DESIGN

BAYESIAN

"We want to maximise the distance between the predictive posteriors with Bhattacharyya distance using Bayesian Optimisation"

↓

Probably you right now: 😕

↓

Let's try to build some intuition:

1. Predictive posterior distribution
2. Bhattacharyya distance
3. Bayesian Optimisation

OPTIMAL EXPERIMENTAL DESIGN

BAYESIAN

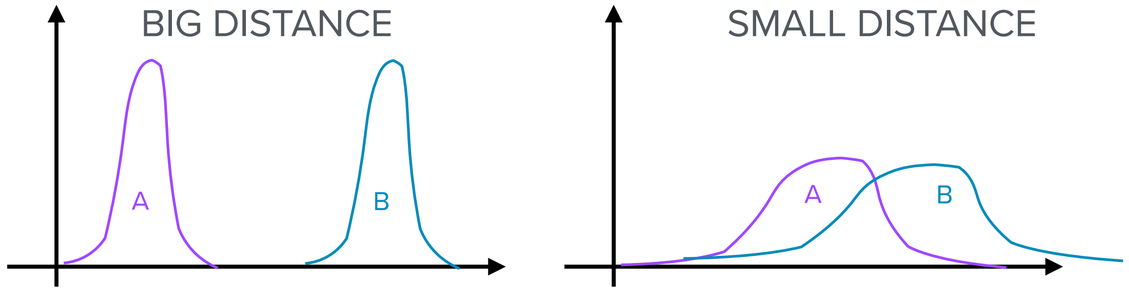
Predictive posterior: probability of new data marginalising over the posterior

$$p(\tilde{E} \mid \{E\}, M_i) = \int p(\tilde{E} \mid \theta_i) p(\theta_i \mid \{E\}) d\theta = \mathbb{E}_{post}[p(\tilde{E})]$$

Bhattacharyya distance: how distance are the distributions (amount of overlap)

$$d_B(p, q) = -\log\left(\int_X \sqrt{q(x)p(x)}dx\right)$$

[For Gaussians (VI on posterior) we want distant means and small variance]



OPTIMAL EXPERIMENTAL DESIGN

BAYESIAN OPTIMISATION: Gaussian process and Upper Confidence Bound (GP + UCB)

u

Input aTc and IPTG step-wise function

$E'(u)$

Make new experiment with this input

$p(\theta_i \mid \{E\} \cup E'(u))$
$p(\theta_j \mid \{E\} \cup E'(u))$

"Calculate" new posteriors

$p = p(\tilde{E} \mid \{E\} \cup E'(u), M_i)$
$q = p(\tilde{E} \mid \{E\} \cup E'(u), M_j)$

Calculate new predictive posteriors

$f(u) = d_B(p, q)$

Calculate Bhattacharyya distance

OPTIMAL EXPERIMENTAL DESIGN

BAYESIAN OPTIMISATION: Gaussian process and Upper Confidence Bound (GP + UCB)

u

Input aTc and IPTG step-wise function

$E'(u)$

Make new experiment with this input

$p(\theta_i \mid \{E\} \cup E'(u))$

$p(\theta_j \mid \{E\} \cup E'(u))$

"Calculate" new posteriors

BLACK BOX

$p = p(\tilde{E} \mid \{E\} \cup E'(u), M_i)$

$q = p(\tilde{E} \mid \{E\} \cup E'(u), M_j)$

Calculate new predictive posteriors

$f(u) = d_B(p, q)$

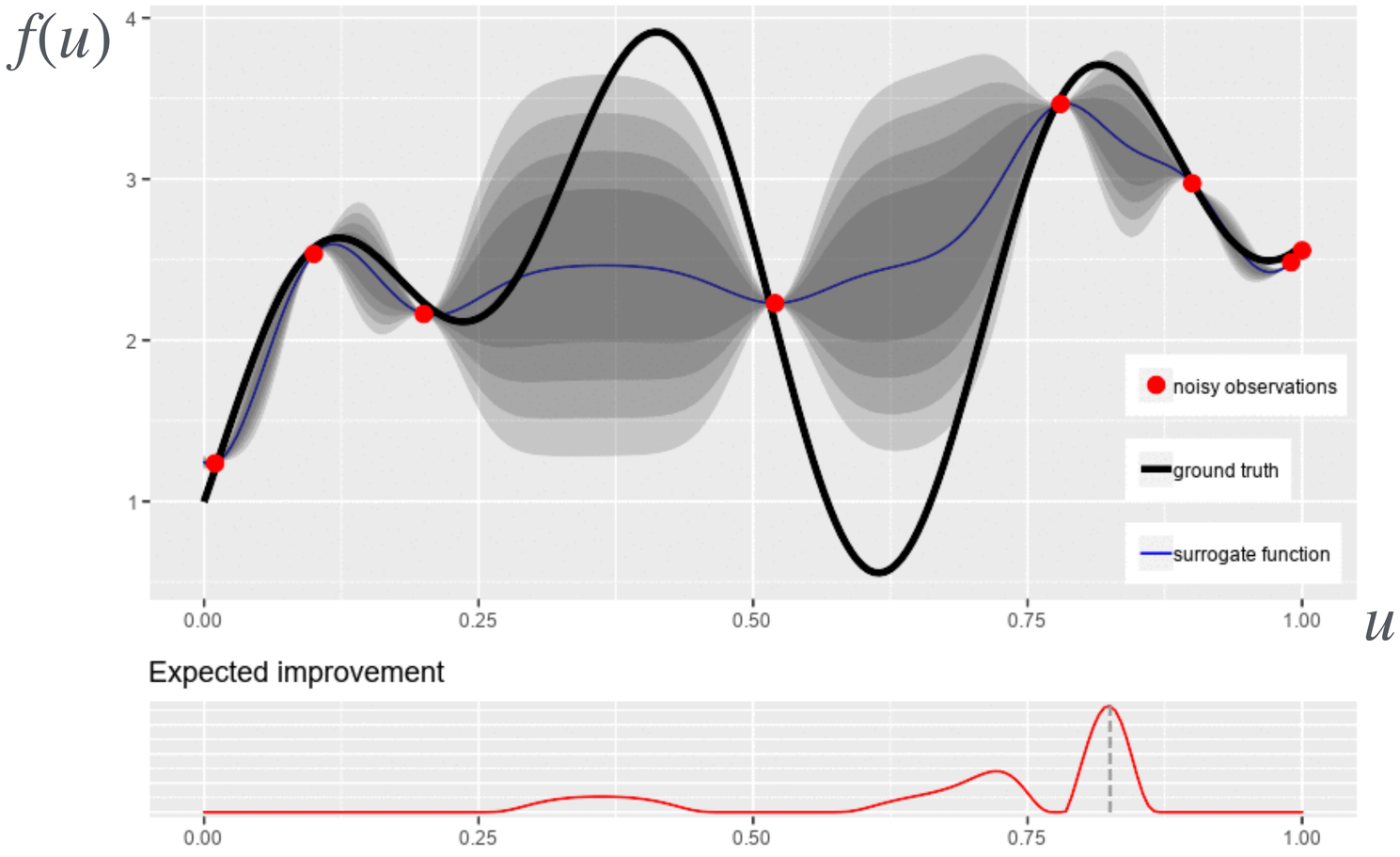
Calculate Bhattacharyya distance

CANNOT TAKE DERIVATIVES OF POSTERIOR

Cannot directly maximise! Consider it BlackBox

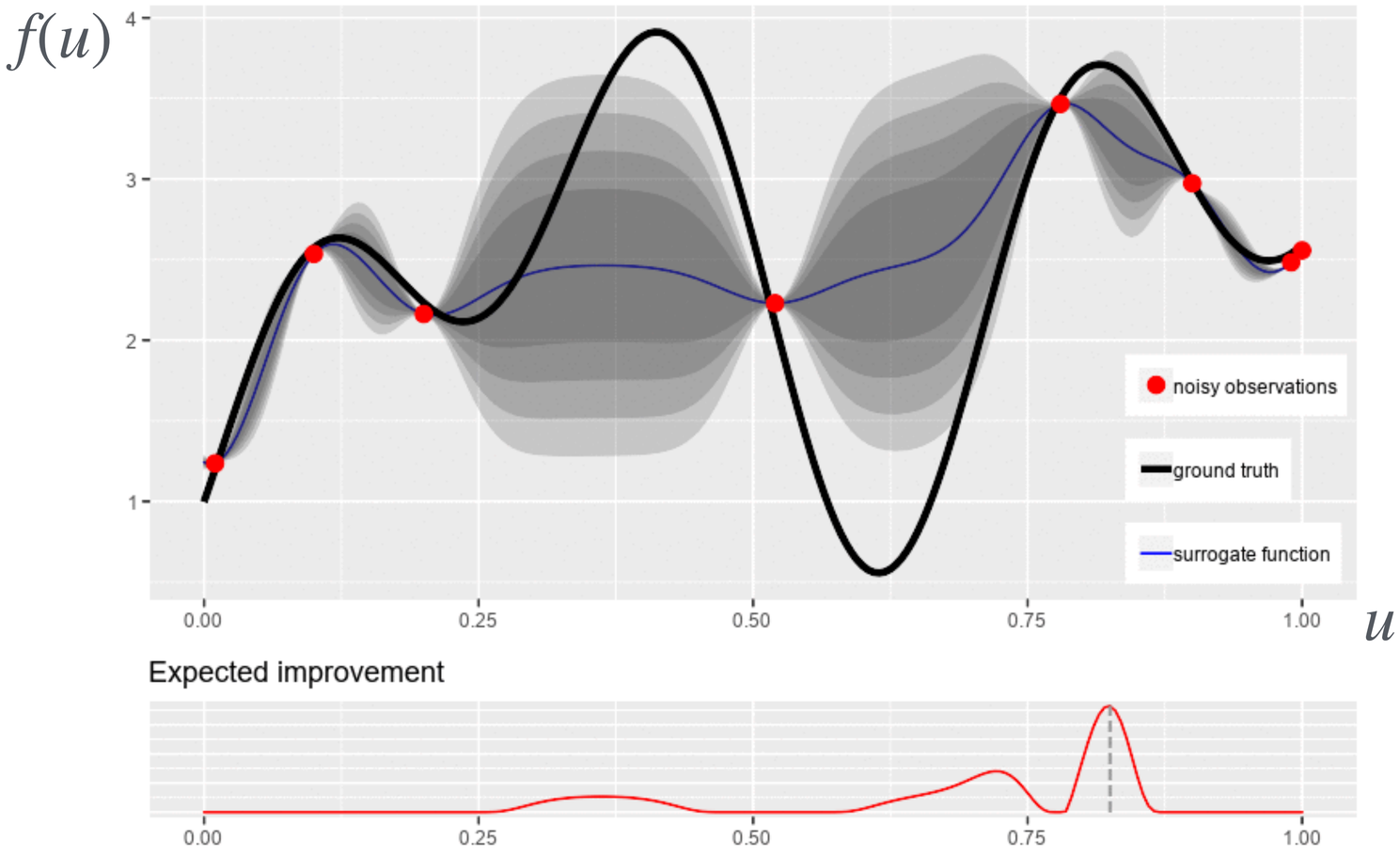
OPTIMAL EXPERIMENTAL DESIGN

BAYESIAN OPTIMISATION: Gaussian process and Upper Confidence Bound (GP + UCB)



OPTIMAL EXPERIMENTAL DESIGN

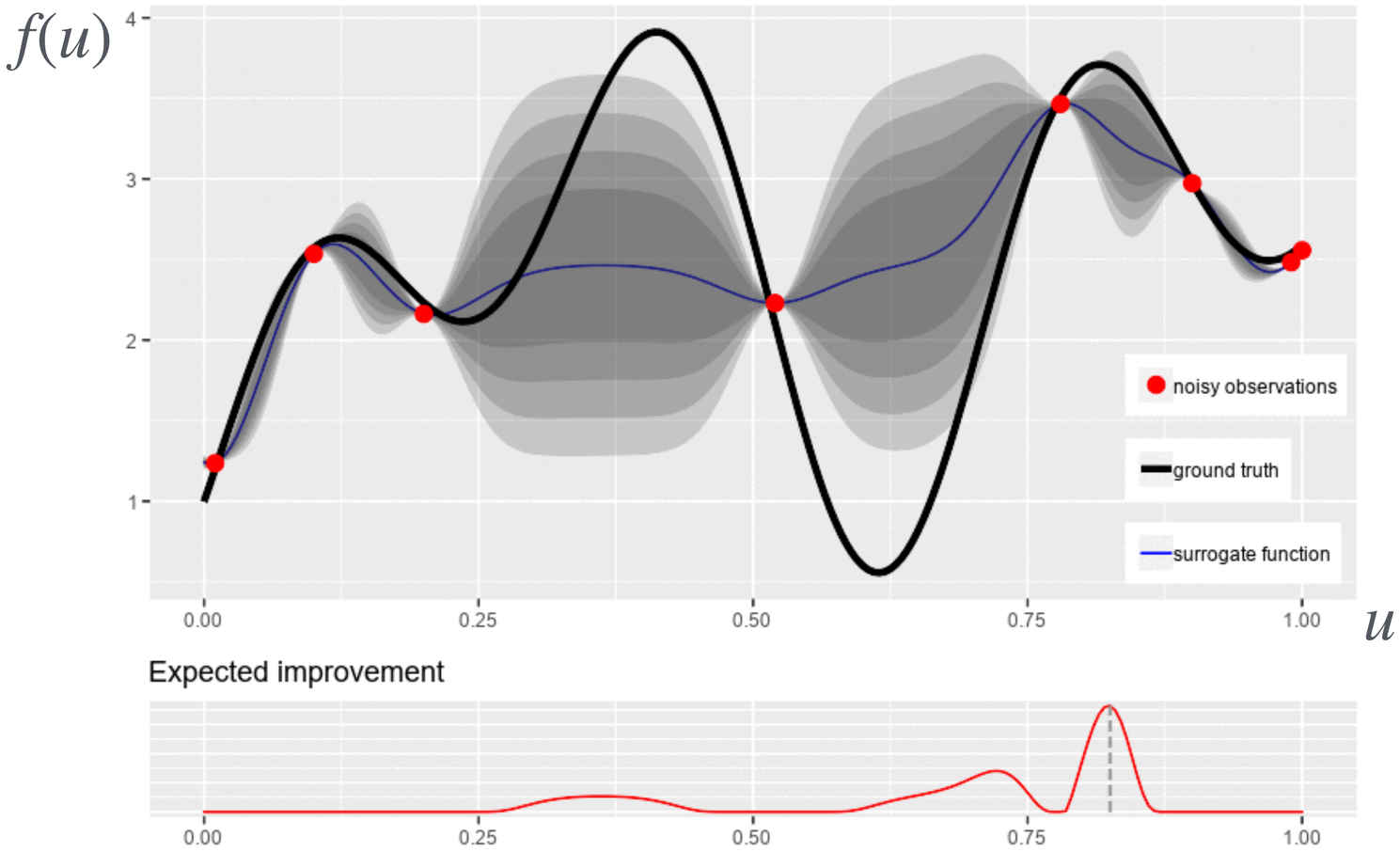
BAYESIAN OPTIMISATION: Gaussian process and Upper Confidence Bound (GP + UCB)



OPTIMAL EXPERIMENTAL DESIGN

BAYESIAN OPTIMISATION: Gaussian process and Upper Confidence Bound (GP + UCB)

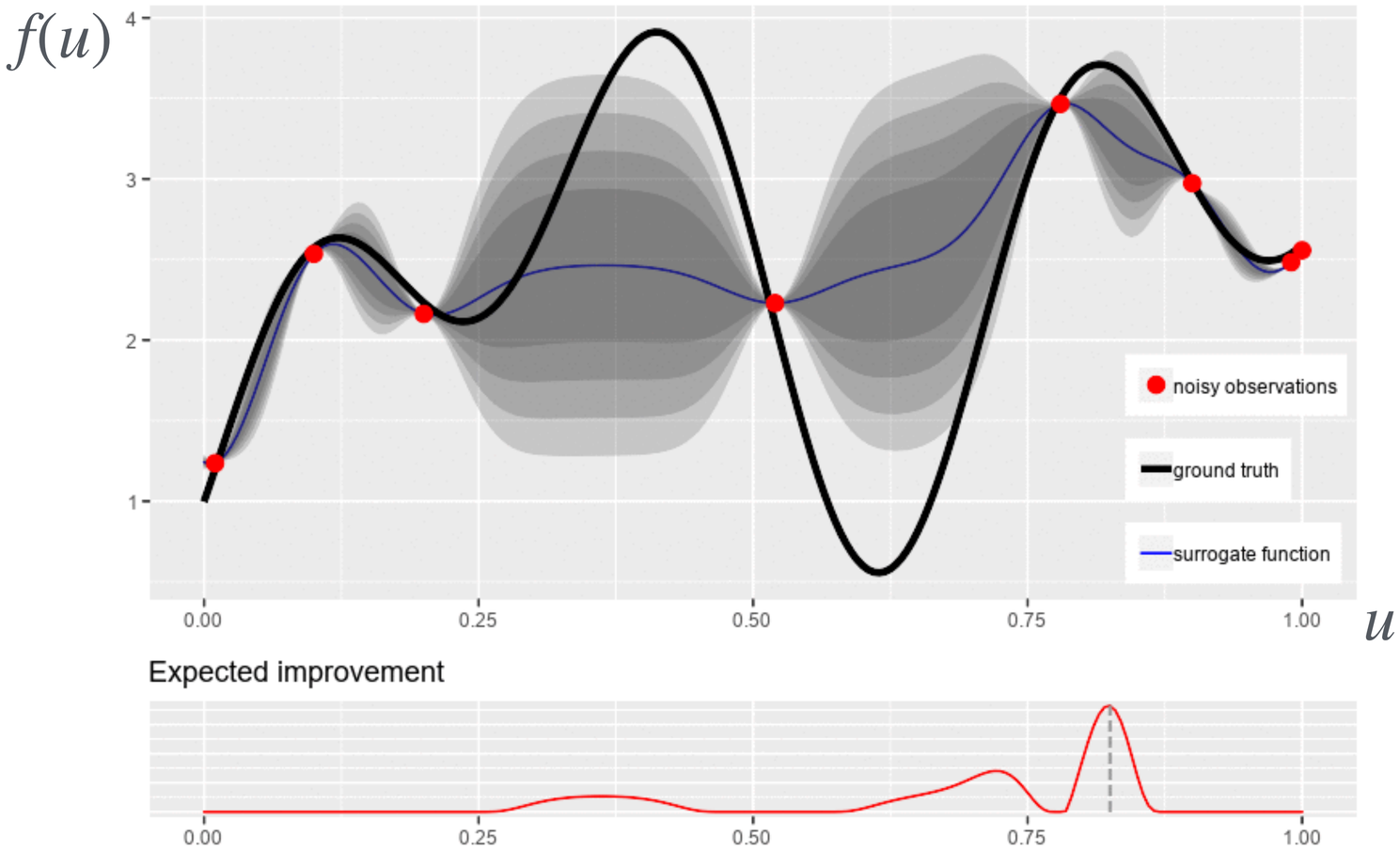
Sorry this is very sloppy 😬



OPTIMAL EXPERIMENTAL DESIGN

BAYESIAN OPTIMISATION: Gaussian process and Upper Confidence Bound (GP + UCB)

Sorry this is very sloppy 😬



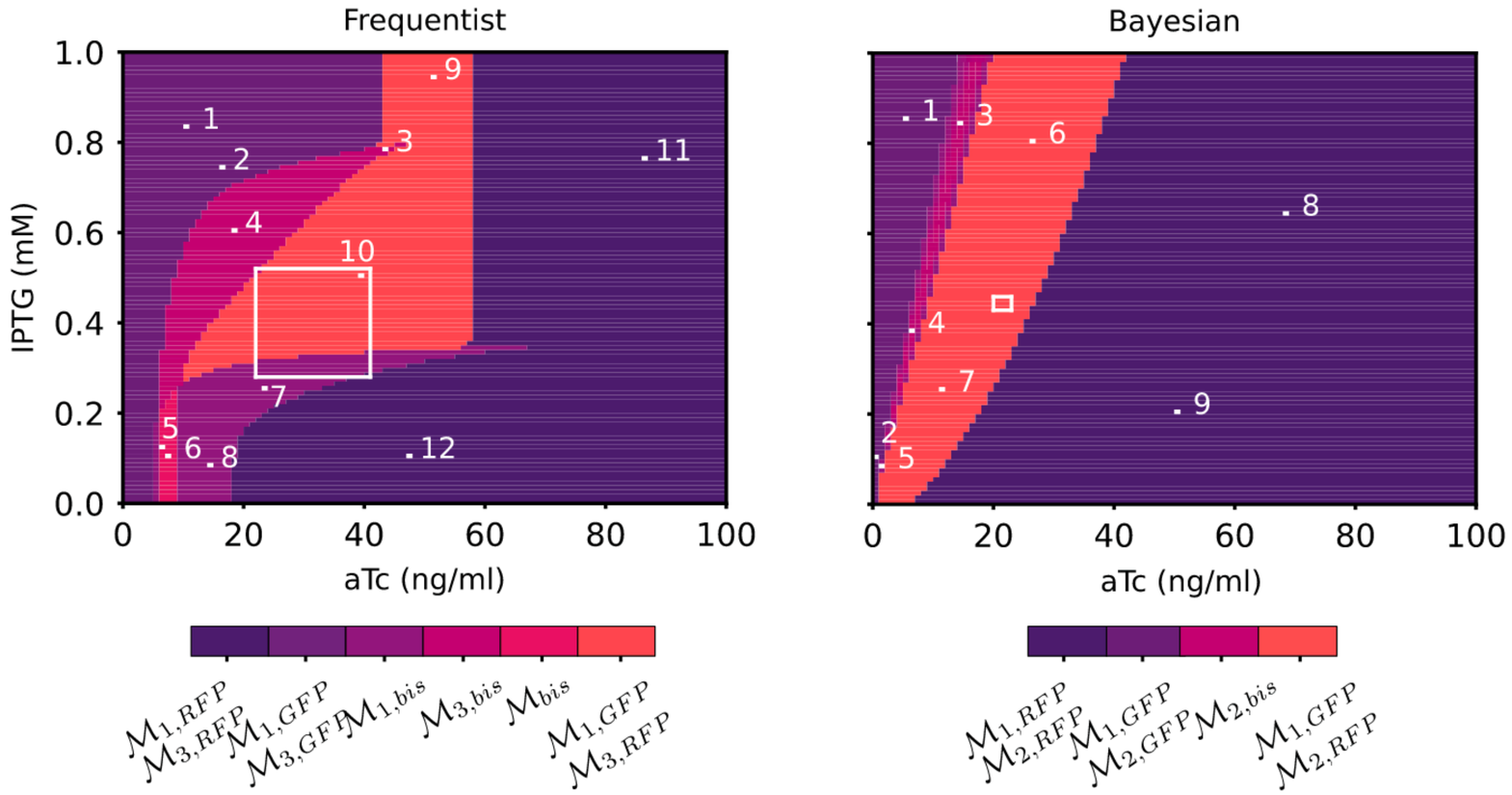
Follow PAI for more! 😉

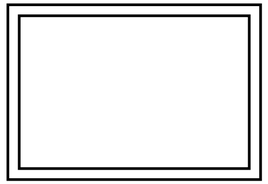
STABILITY PROPERTIES OF THE SOLUTION

We have optimal inputs and different steady states, but **how sure are we** of the result?

↓

We want to study the stability of these steady states depending on the input





Part of the input space returned by the optimisers

↓

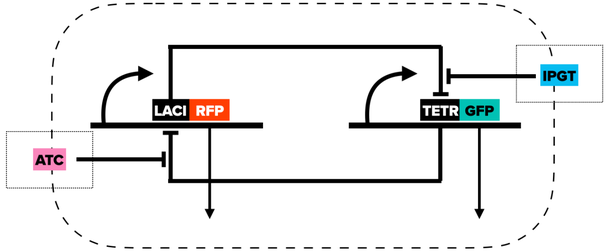
They are the best at discriminating!

CONCLUSION

IN THIS FRAMEWORK

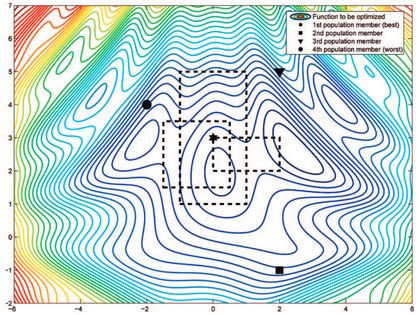
1. Create candidate model for your biological problem
2. We proceed with traditional model selection
   I. Find best set of parameters for each candidate model
      a) Frequentist: Evolutionary Algorithm
      b) Bayesian: MCMC method
   II. Eliminate some candidate models
      a) Frequentist: AIC
      b) Bayesian: Laplace Approximation
3. Augment you dataset with optimal experiment
   I. Define you metric
      a) Frequentist: (simil) RMSE
      b) Bayesian: Bhattacharyya distance
   II. Optimise for the metric
      a) Frequentist Evolutionary Algorithm
      b) Bayesian: GP + UCB

RECAP



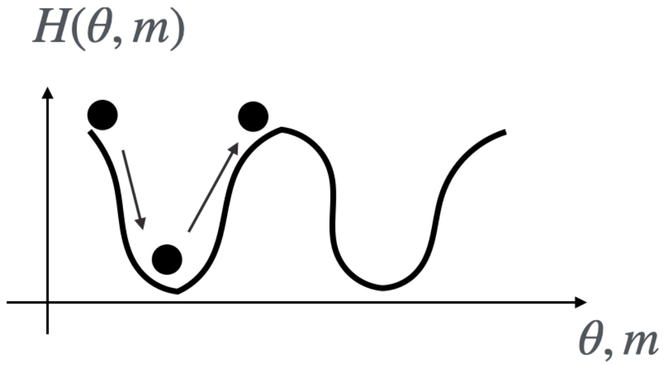
Toggle Switch Model

(CSB)



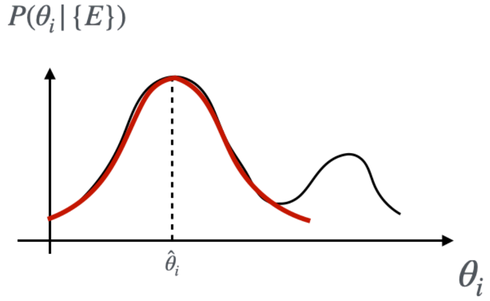
Enhanced Scatter Search

(“CSB”)



Hamiltonian Monte Carlo

(“CSBMS”)



Laplace Approximation

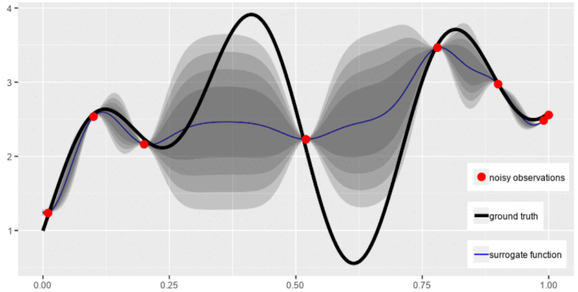
(PAI)

$$p(\tilde{E} \mid \{E\}, M_i) = \int p(\tilde{E} \mid \theta_i) p(\theta_i \mid \{E\}) d\theta = \mathbb{E}_{post}[p(\tilde{E})]$$

$$d_B(p, q) = -\log\left(\int_X \sqrt{q(x)p(x)}dx\right)$$

Predictive posterior and Bhattacharyya distance

(PAI)



Bayesian Opaimization

(PAI)

THANK YOU FOR YOUR ATTENTION

ANY QUESTIONS?