UNIVERSITY OF CAMBRIDGE

Lancaster University

To Bayesian Optimisation and Beyond

Gaussian Processes as Decision Makers

Henry Moss

What is Active Learning?

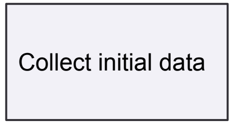
Bayesian search for learning functions

Sequential data collection

Let's make use of uncertainty estimates to make better models

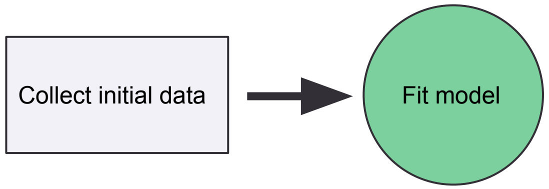
Sequential data collection

Let's make use of uncertainty estimates to make better models



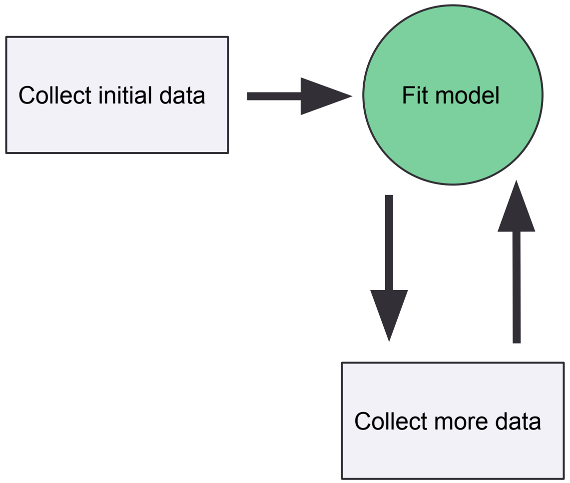
Sequential data collection

Let's make use of uncertainty estimates to make better models



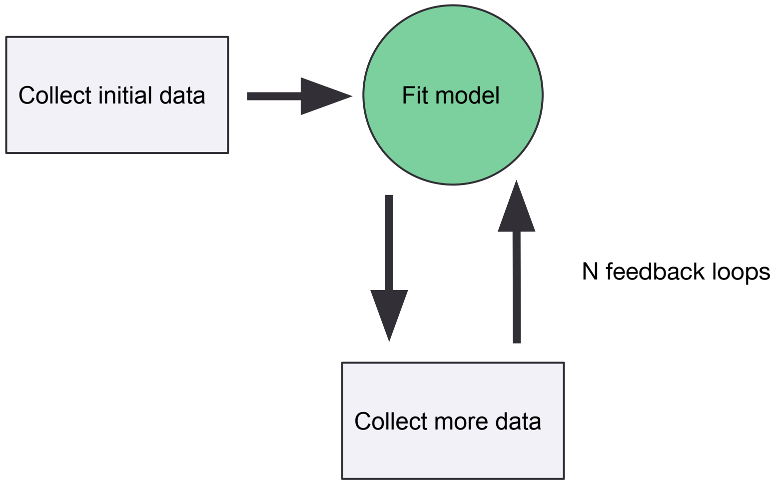
Sequential data collection

Let's make use of uncertainty estimates to make better models



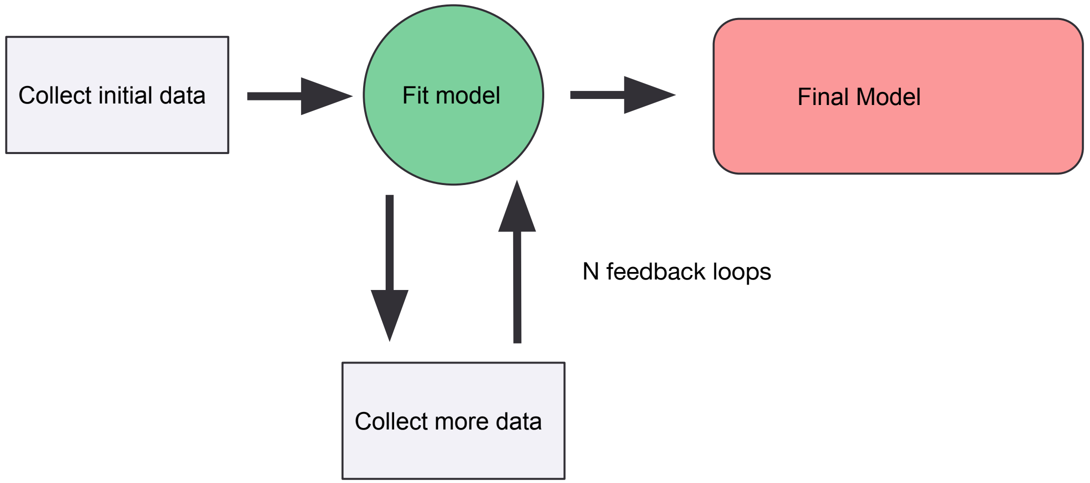
Sequential data collection

Let's make use of uncertainty estimates to make better models



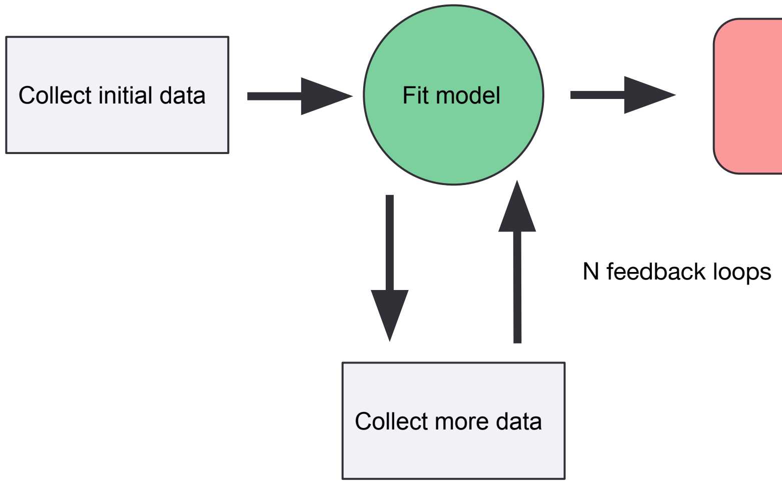
Sequential data collection

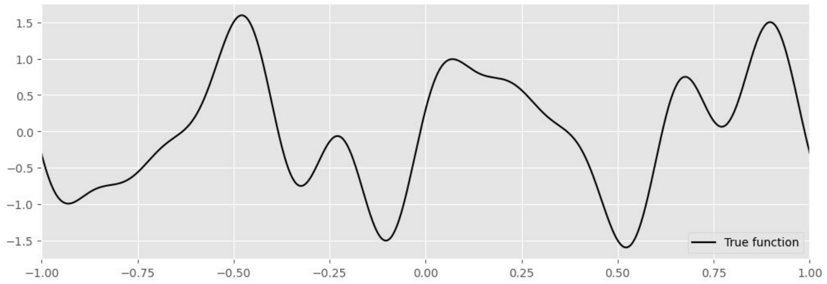
Let's make use of uncertainty estimates to make better models

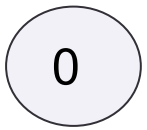


Sequential data collection

Let's make use of uncertainty estimates to make better models

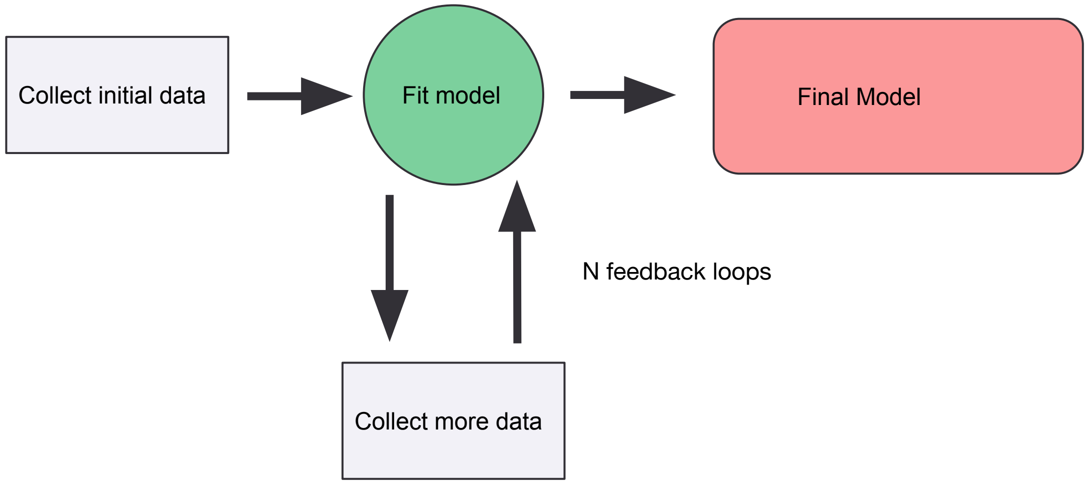


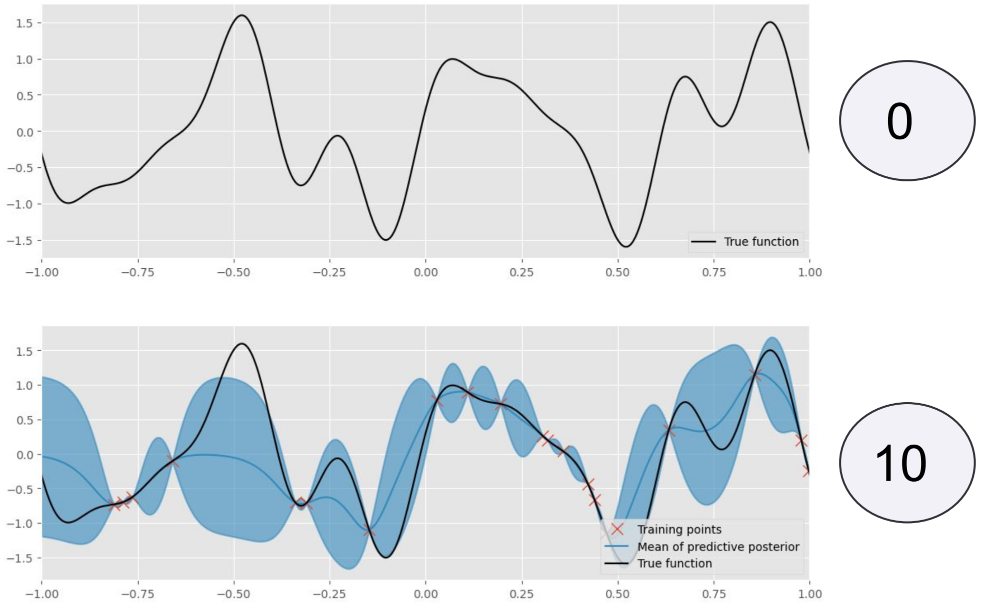




Sequential data collection

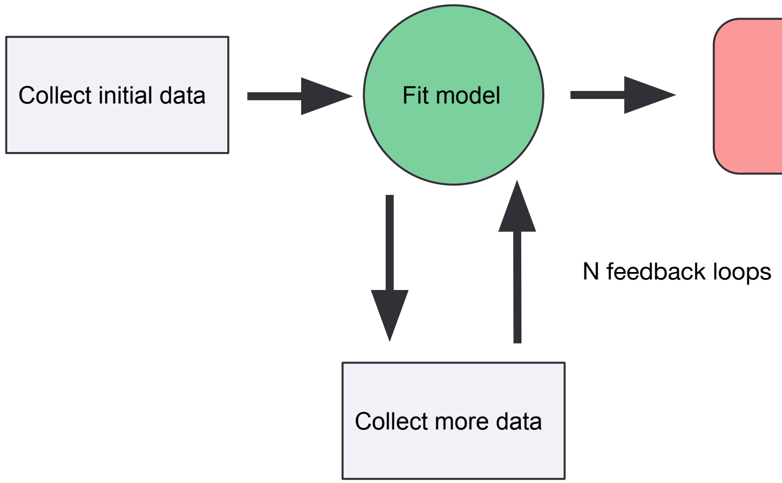
Let's make use of uncertainty estimates to make better models

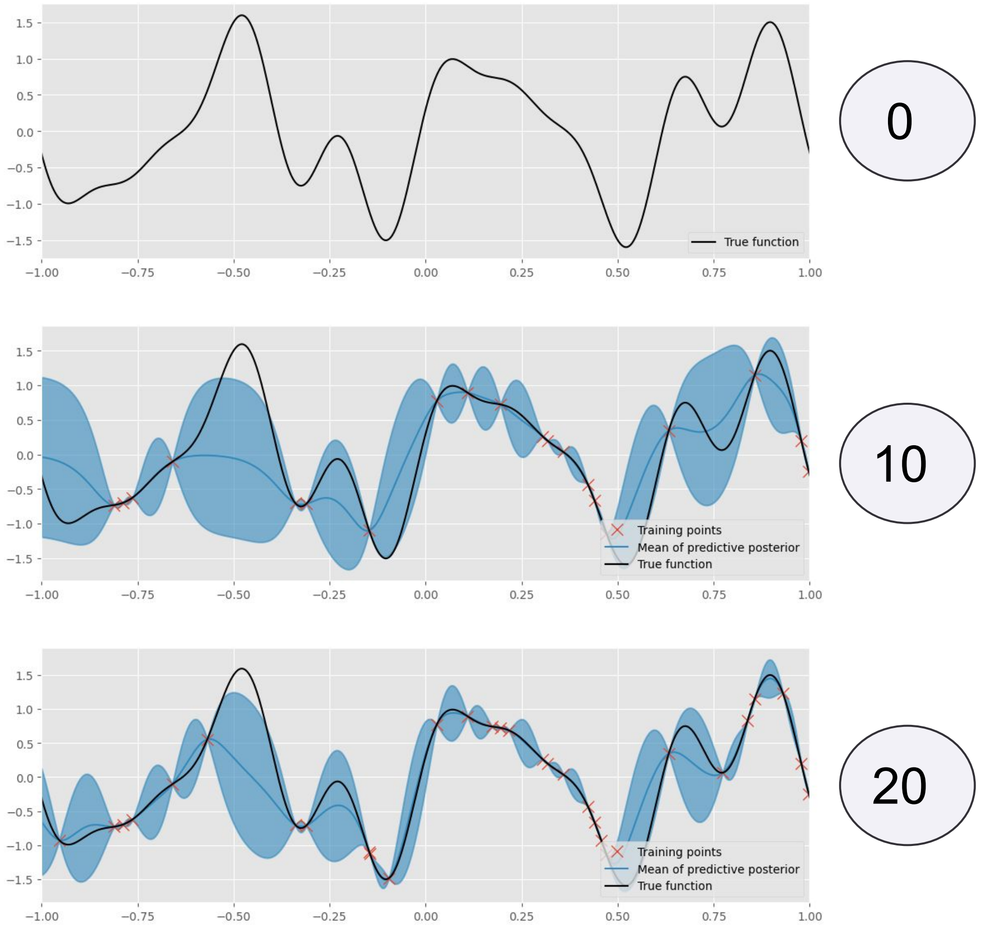




Sequential data collection

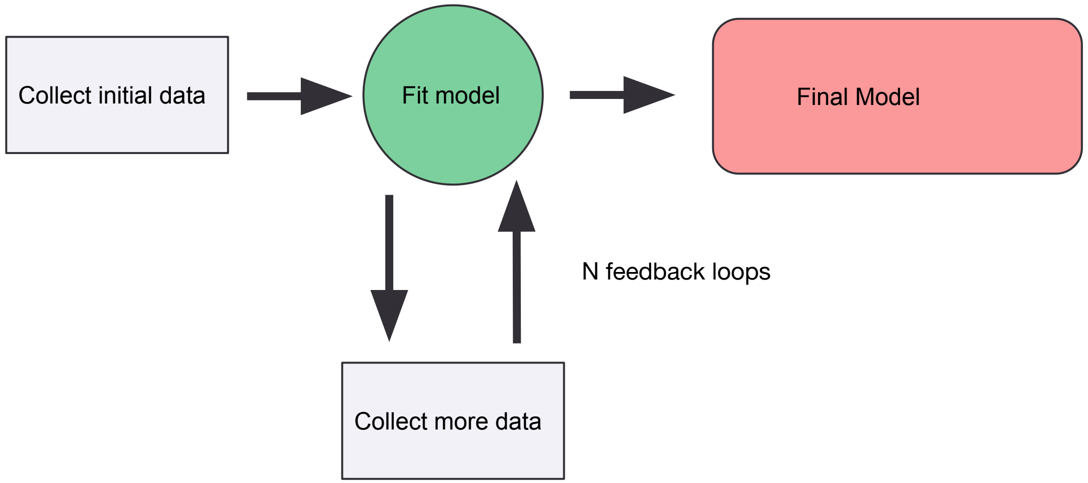
Let's make use of uncertainty estimates to make better models

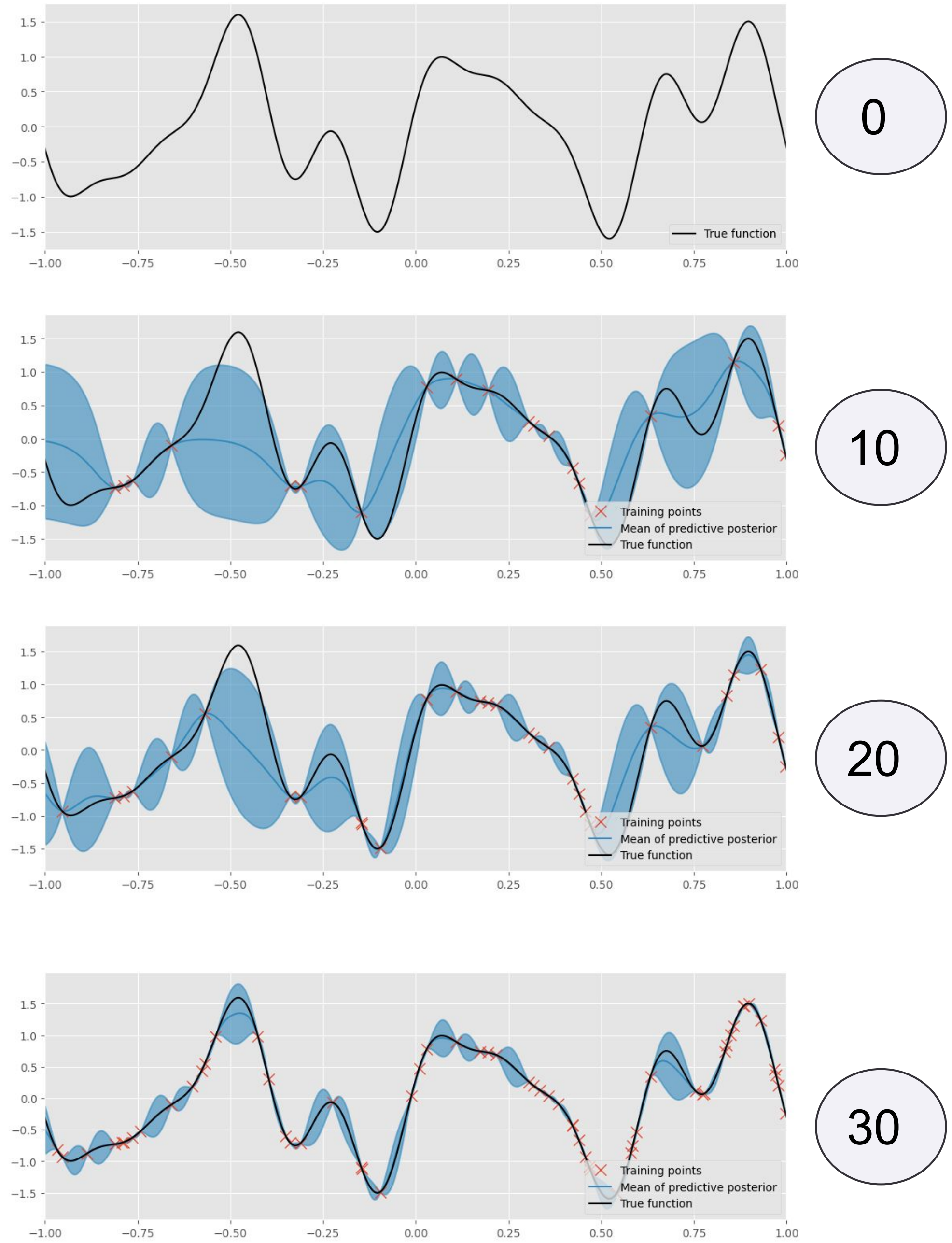




Sequential data collection

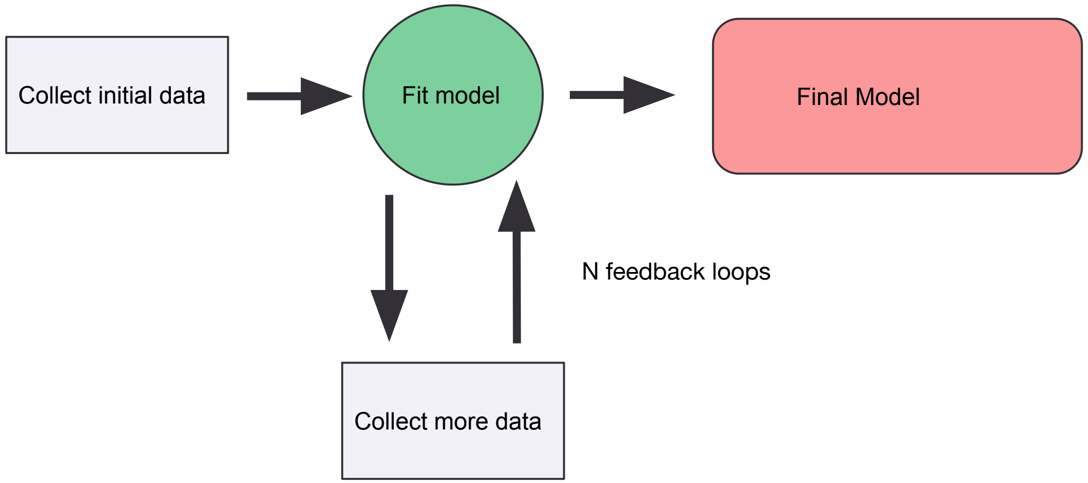
Let's make use of uncertainty estimates to make better models

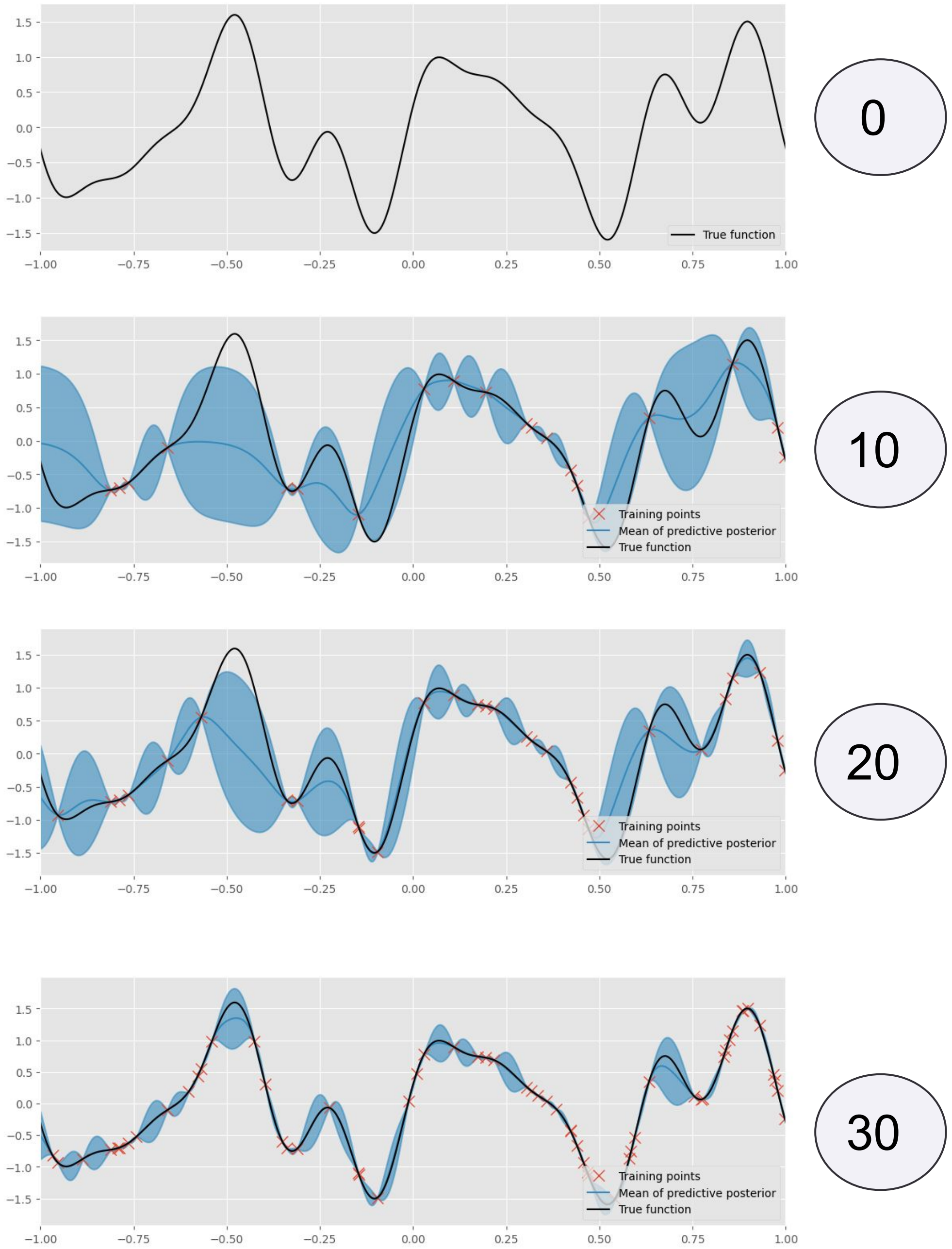




Sequential data collection

Let's make use of uncertainty estimates to make better models





But can we do better than **random**???

Active learning

Sequentially collecting more data to improve your model for the task at hand

Active learning

Sequentially collecting more data to improve your model for the task at hand

- I care about **regression** —> collect data to improve global model accuracy

Active learning

Sequentially collecting more data to improve your model for the task at hand

- I care about **regression** —> collect data to improve global model accuracy
- I care about the **maximum** value of my process —> collect data in promising regions (Bayesian Optimisation)

Active learning

Sequentially collecting more data to improve your model for the task at hand

- I care about **regression** —> collect data to improve global model accuracy
- I care about the **maximum** value of my process —> collect data in promising regions (Bayesian Optimisation)
- I'm interested in **multiple objectives** -> populate the Pareto front (Multi-objective Bayesian Optimisation)

Active learning

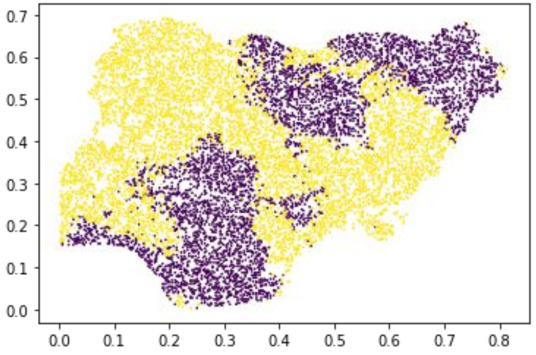
Sequentially collecting more data to improve your model for the task at hand

- I care about **regression** —> collect data to improve global model accuracy
- I care about the **maximum** value of my process —> collect data in promising regions (Bayesian Optimisation)
- I'm interested in **multiple objectives** -> populate the Pareto front (Multi-objective Bayesian Optimisation)
- I care about predicting a **threshold** -> choose data close to threshold (level-set design)

Active learning

Sequentially collecting more data to improve your model for the task at hand

- I care about **regression** —> collect data to improve global model accuracy
- I care about the **maximum** value of my process —> collect data in promising regions (Bayesian Optimisation)
- I'm interested in **multiple objectives** -> populate the Pareto front (Multi-objective Bayesian Optimisation)
- I care about predicting a **threshold** -> choose data close to threshold (level-set design)

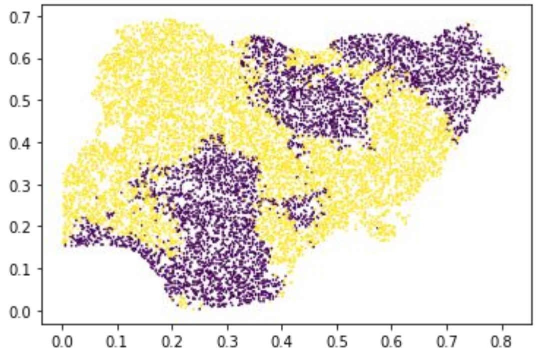


Malaria incidence
in Nigeria

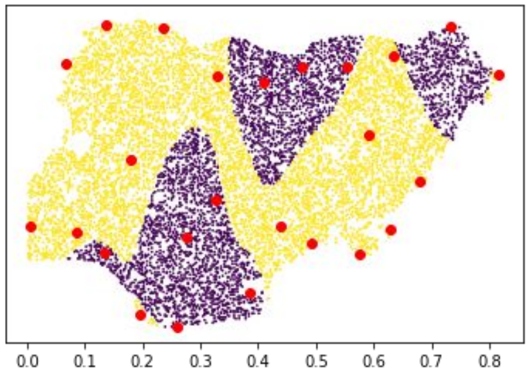
Active learning

Sequentially collecting more data to improve your model for the task at hand

- I care about **regression** —> collect data to improve global model accuracy
- I care about the **maximum** value of my process —> collect data in promising regions (Bayesian Optimisation)
- I'm interested in **multiple objectives** -> populate the Pareto front (Multi-objective Bayesian Optimisation)
- I care about predicting a **threshold** -> choose data close to threshold (level-set design)



Malaria incidence in Nigeria

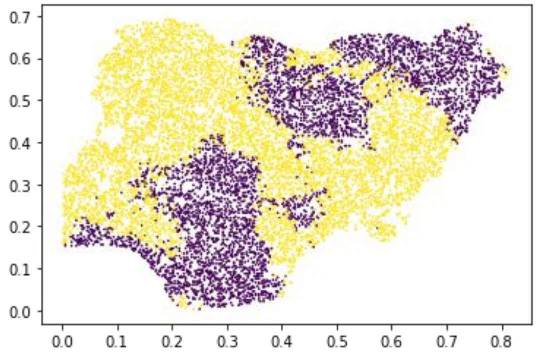


Model on Random data

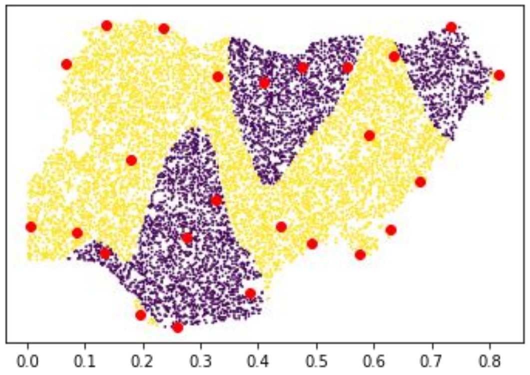
Active learning

Sequentially collecting more data to improve your model for the task at hand

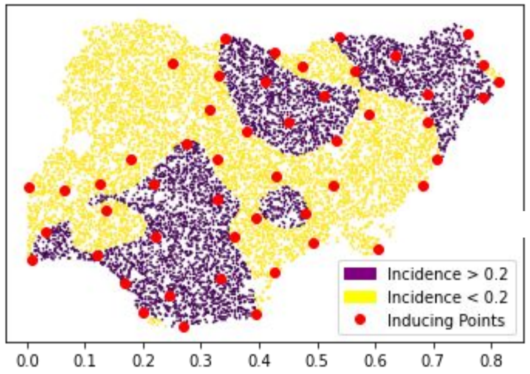
- I care about **regression** —> collect data to improve global model accuracy
- I care about the **maximum** value of my process —> collect data in promising regions (Bayesian Optimisation)
- I'm interested in **multiple objectives** -> populate the Pareto front (Multi-objective Bayesian Optimisation)
- I care about predicting a **threshold** -> choose data close to threshold (level-set design)



Malaria incidence in Nigeria



Model on Random data



Model from data chosen by Active learning

So, Bayesian Optimisation?

i.e. Active learning for optimisation

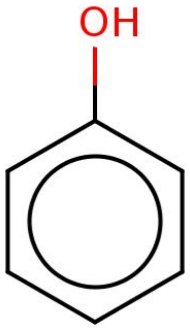
A molecular design pipeline

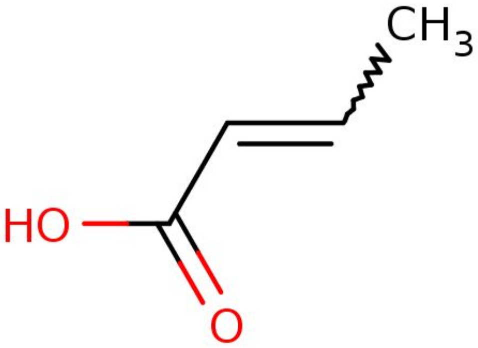
Efficiently explore molecule space

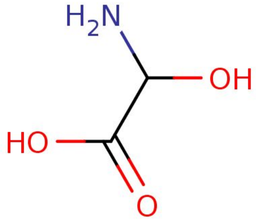
A molecular design pipeline

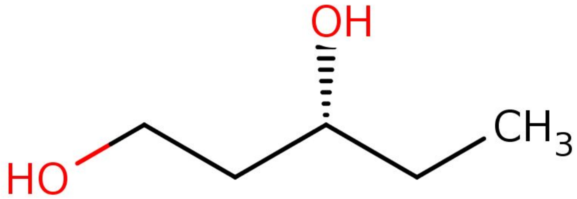
Efficiently explore molecule space

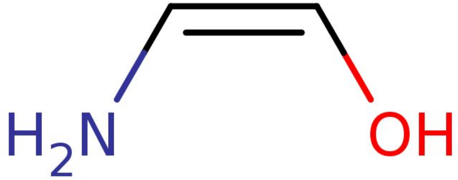
- **Large** library of candidates









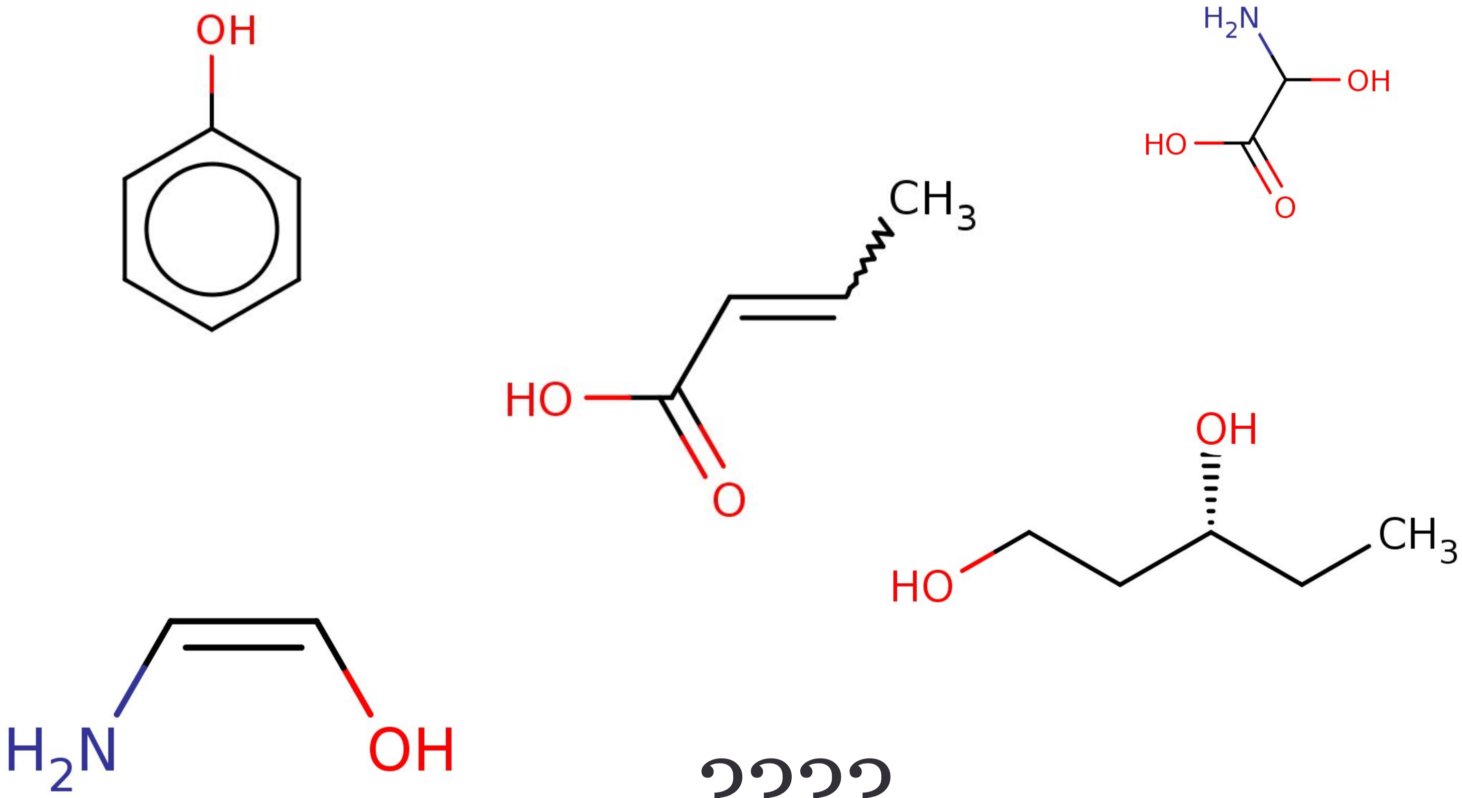


????

A molecular design pipeline

Efficiently explore molecule space

- **Large** library of candidates
- **Expensive** experiments (<10)



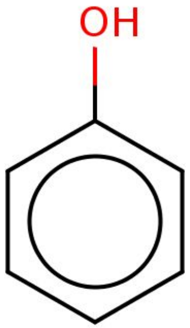
????

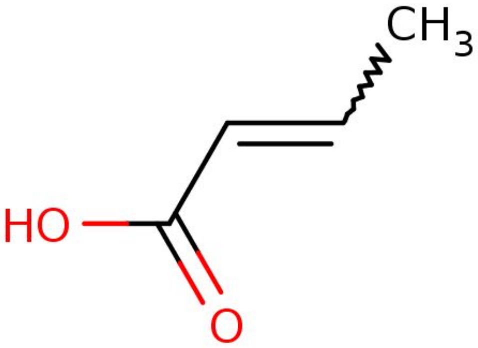
UNIVERSITY OF CAMBRIDGE   Lancaster University

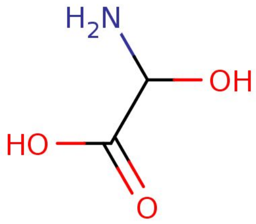
A molecular design pipeline

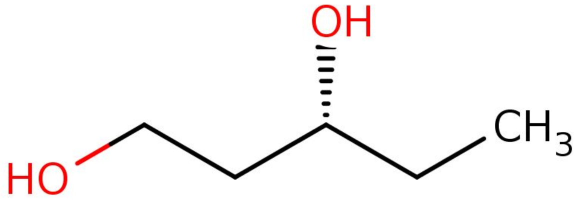
Efficiently explore molecule space

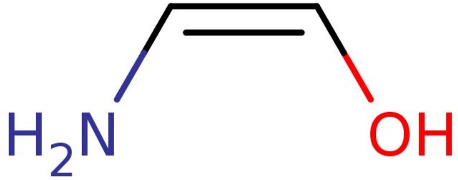
- **Large** library of candidates
- **Expensive** experiments (<10) (**IN A LAB !!!**)









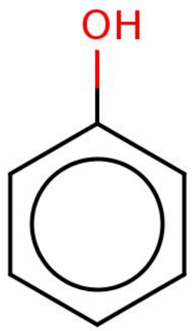


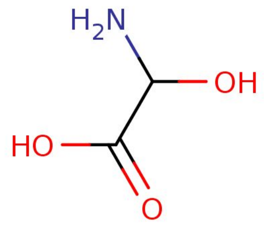
????

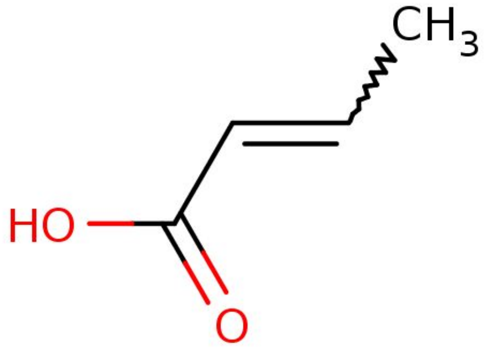
A molecular design pipeline

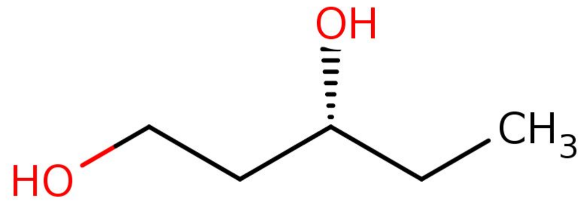
Efficiently explore molecule space

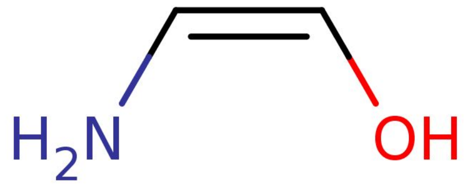
- **Large** library of candidates
- **Expensive** experiments (<10)
- High degree of **parallelism**









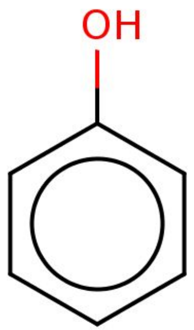


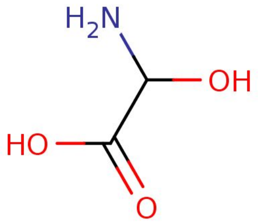
????

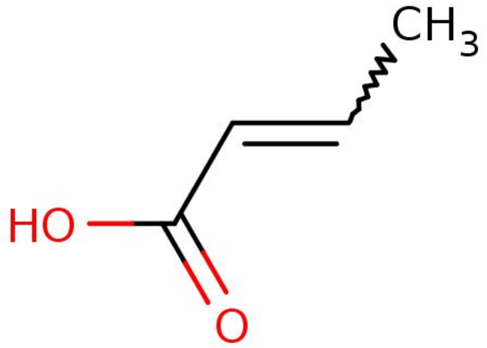
A molecular design pipeline

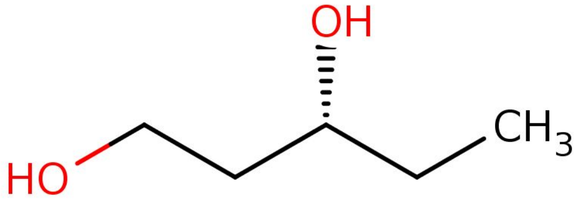
Efficiently explore molecule space

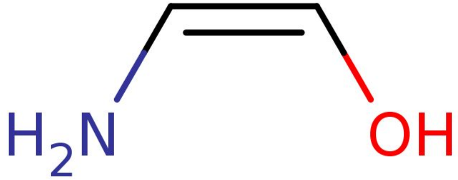
- **Large** library of candidates
- **Expensive** experiments (<10)
- High degree of **parallelism**
- Want molecules with high **affinity**









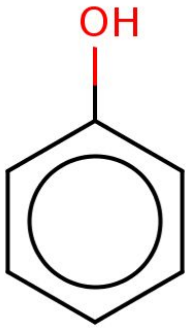


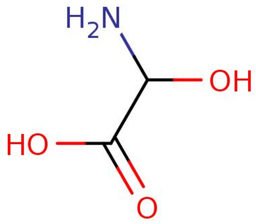
????

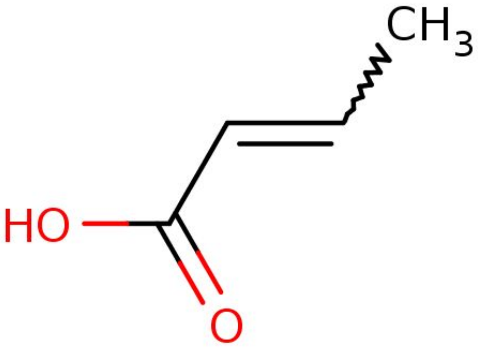
A molecular design pipeline

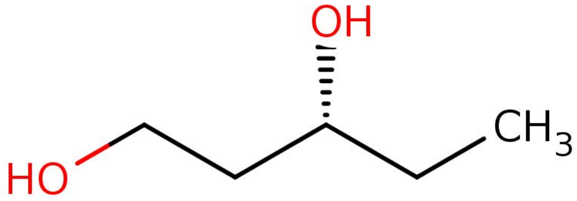
Efficiently explore molecule space

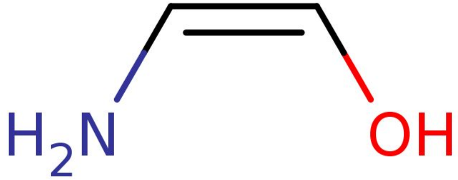
- **Large** library of candidates
- **Expensive** experiments (<10)
- High degree of **parallelism**
- Want molecules with high **affinity**
    - Also easy to make









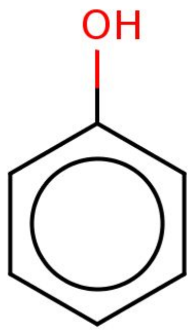


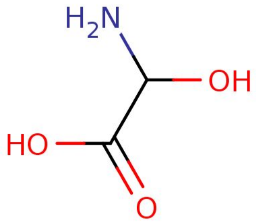
????

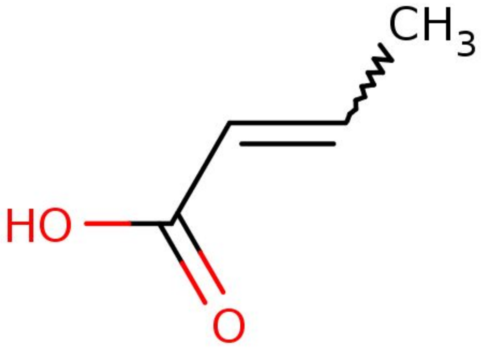
A molecular design pipeline

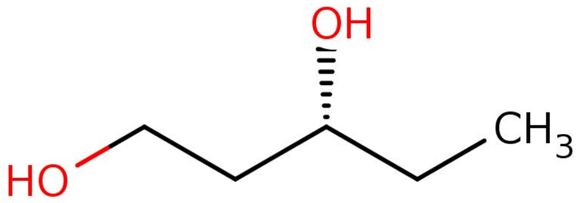
Efficiently explore molecule space

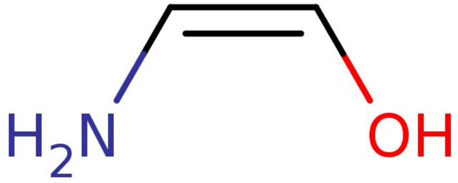
- **Large** library of candidates
- **Expensive** experiments (<10)
- High degree of **parallelism**
- Want molecules with high **affinity**
  - Also easy to make
  - Don't stick to themselves









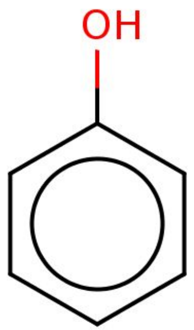


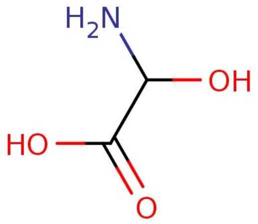
????

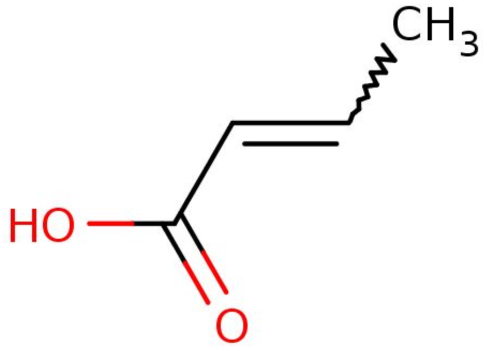
A molecular design pipeline

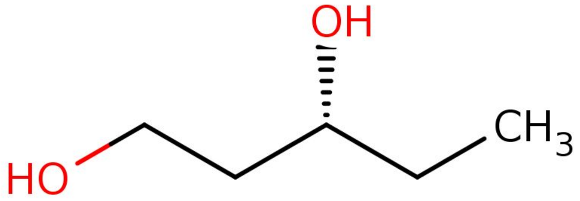
Efficiently explore molecule space

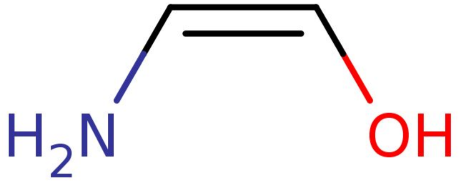
- **Large** library of candidates
- **Expensive** experiments (<10)
- High degree of **parallelism**
- Want molecules with high **affinity**
    - Also easy to make
    - Don’t stick to themselves
    - Stable









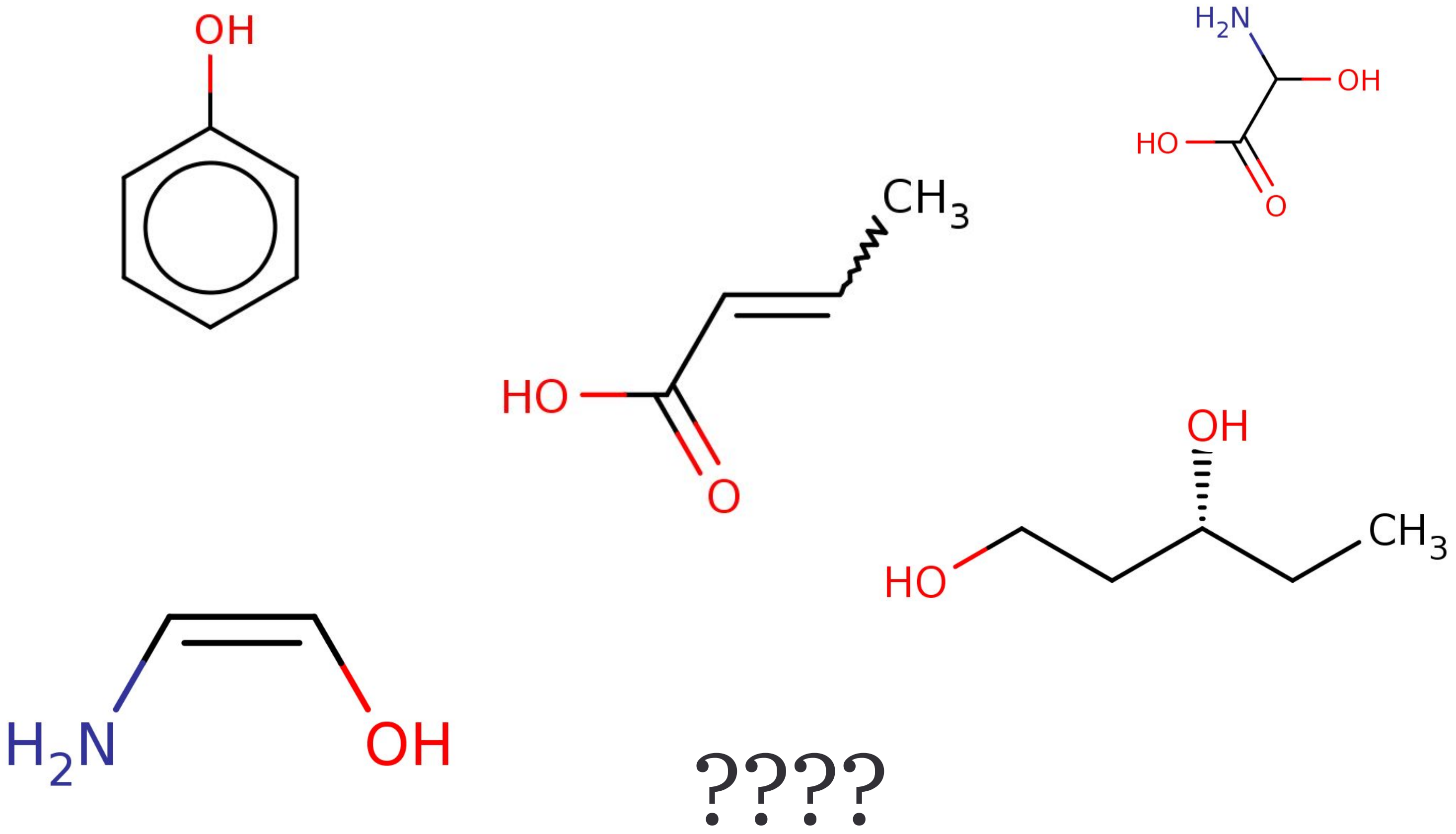


????

A molecular design pipeline

Efficiently explore molecule space

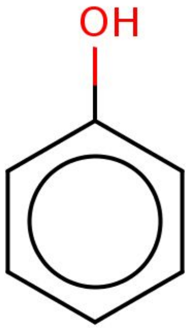
- **Large** library of candidates
- **Expensive** experiments (<10)
- High degree of **parallelism**
- Want molecules with high **affinity**
  - Also easy to make
  - Don’t stick to themselves
  - Stable
  - In a new area of “patent space”

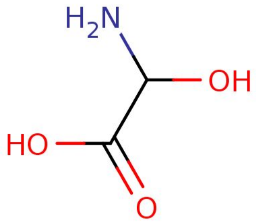


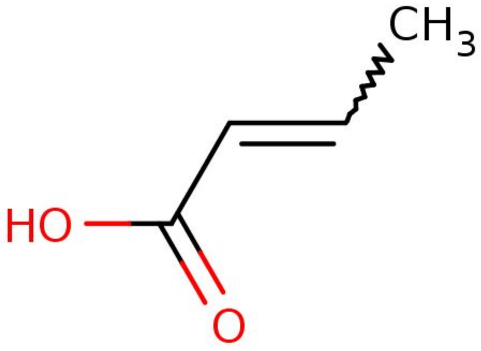
A molecular design pipeline

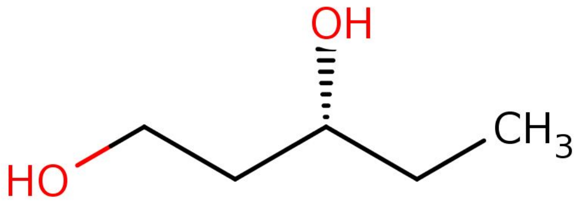
Efficiently explore molecule space

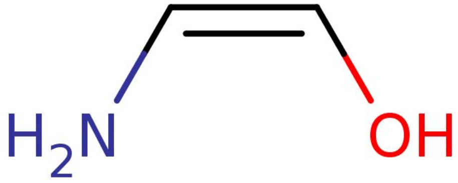
- **Large** library of candidates
- **Expensive** experiments (<10)
- High degree of **parallelism**
- Want molecules with high **affinity**
    - Also easy to make
    - Don't stick to themselves
    - Stable
    - In a new area of "patent space"









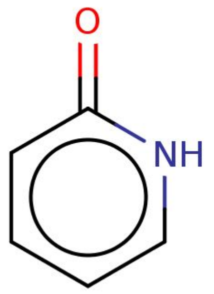


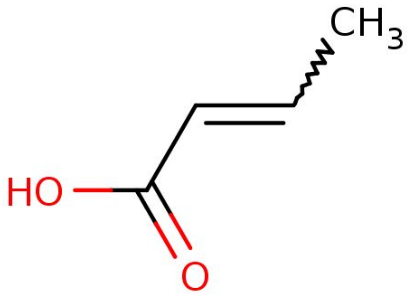
????

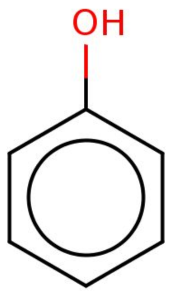
Any ideas?

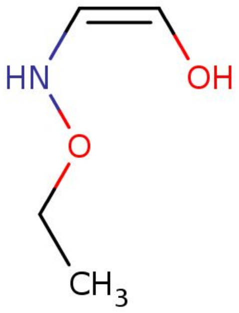
A Simpler Example

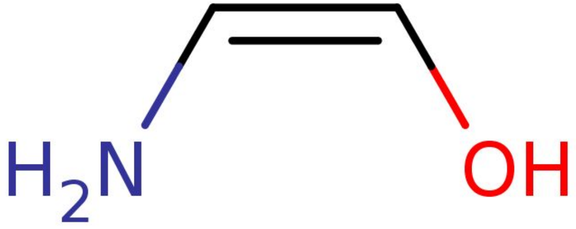
Can evaluate **at most** 4

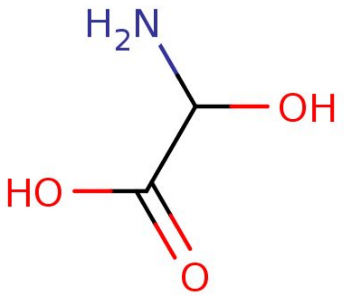






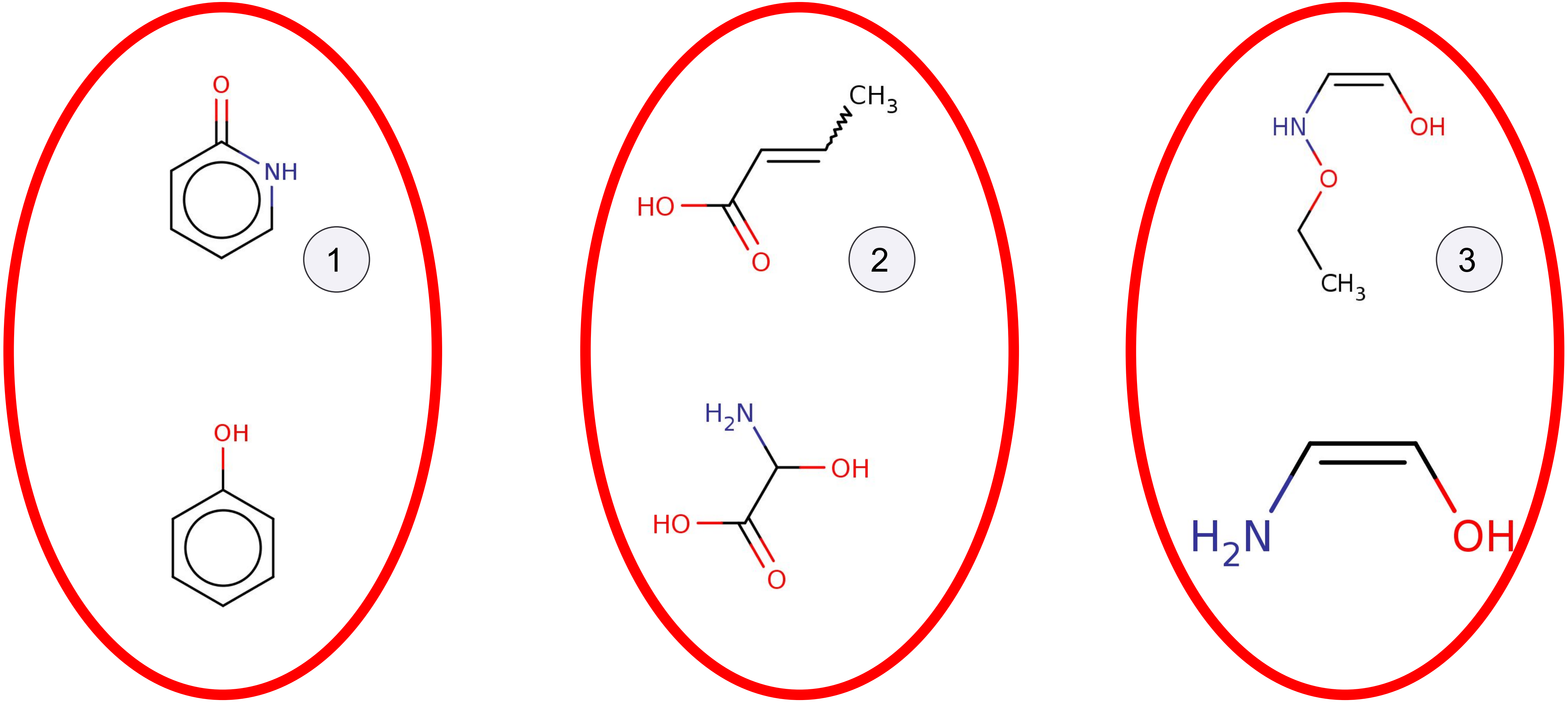






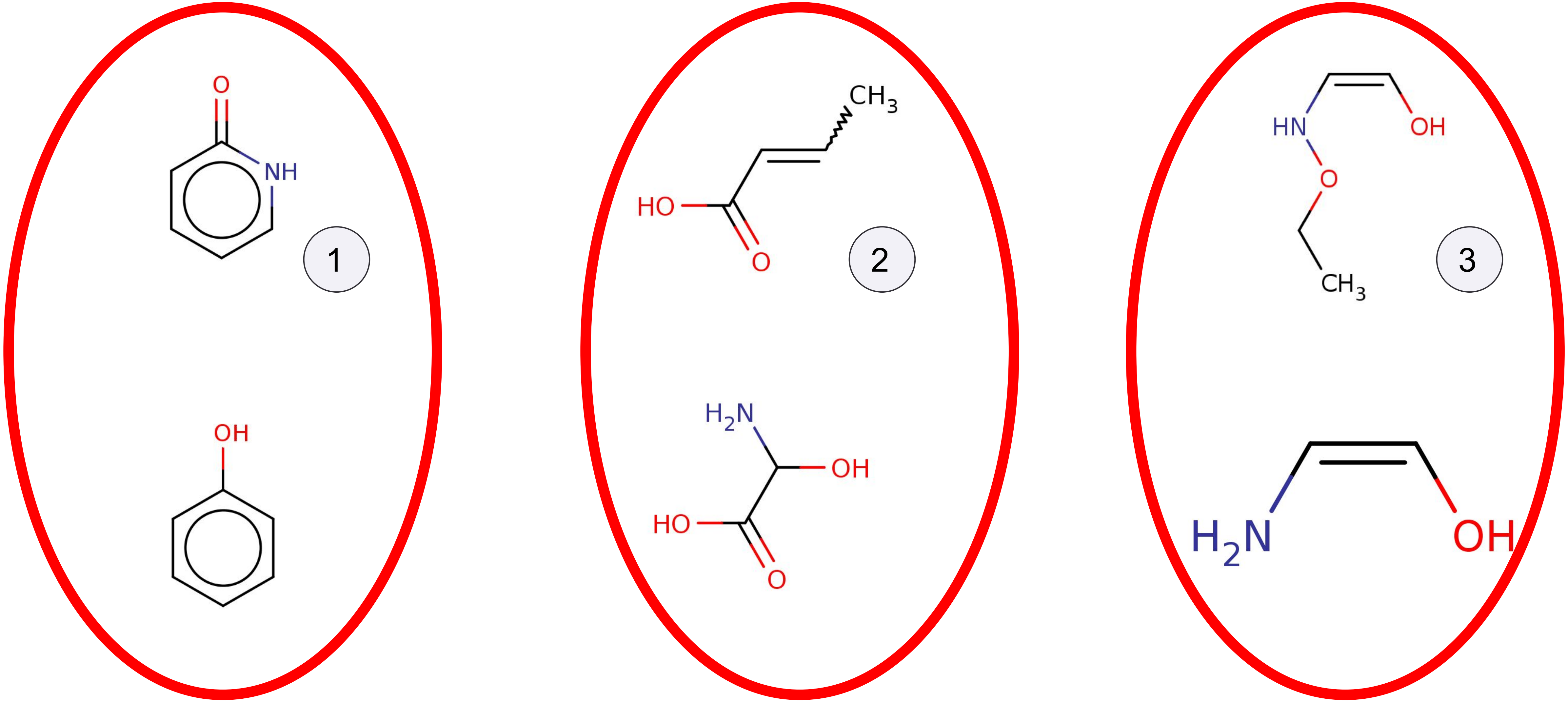
A Simpler Example (grouped)

Can evaluate **at most** 4



A Simpler Example (grouped)

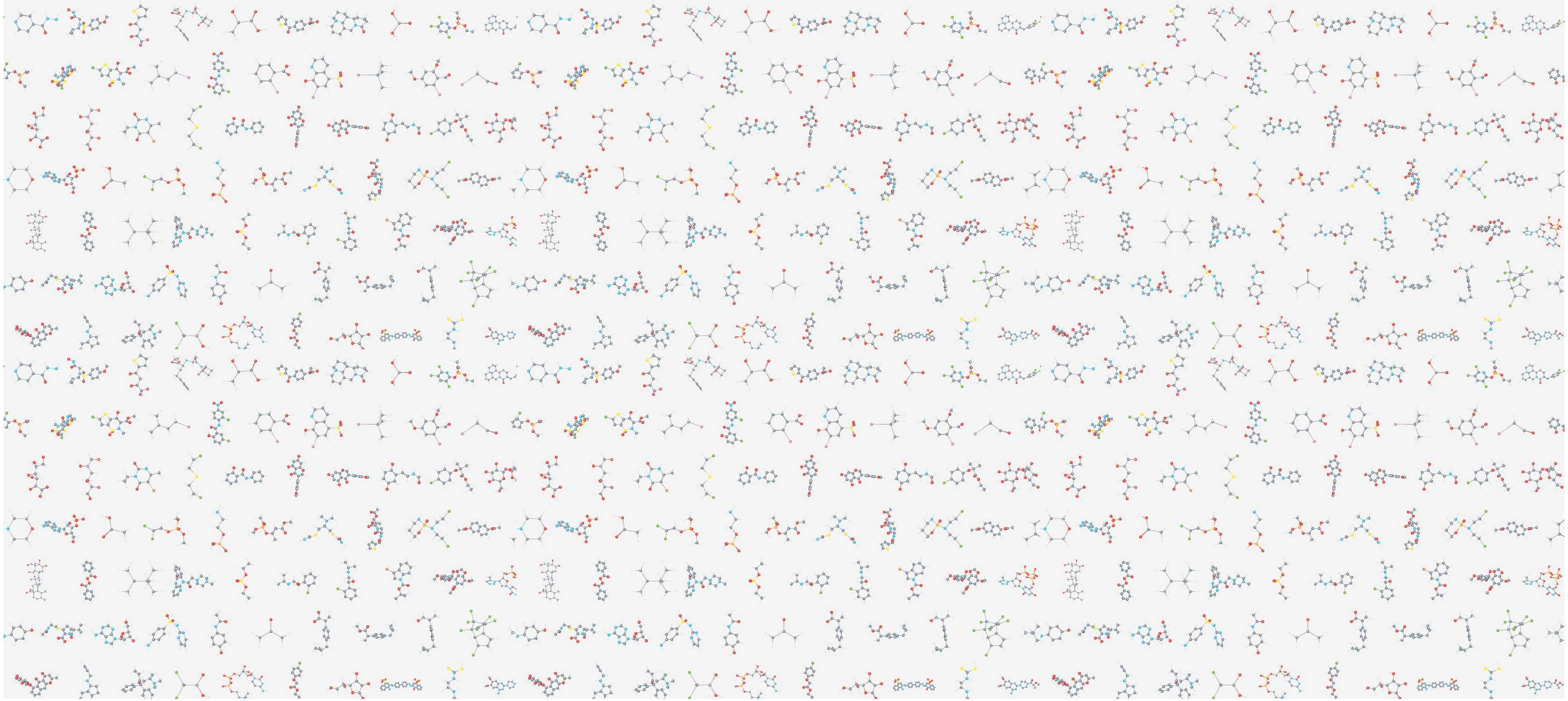
Can evaluate **at most** 4



Explore v.s. **exploit**?

What about at scale?

eek



https://www.horizons-mag.ch/2018/09/06/the-universe-of-molecules/

What about at scale?

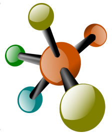
eek

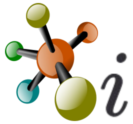
Use a GP!

https://www.horizons-mag.ch/2018/09/06/the-universe-of-molecules/

An Aside: GPs for Molecules

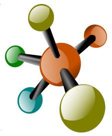
Structured Input Spaces

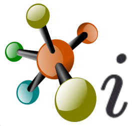
$y_i = f($$_i) + \epsilon_i$

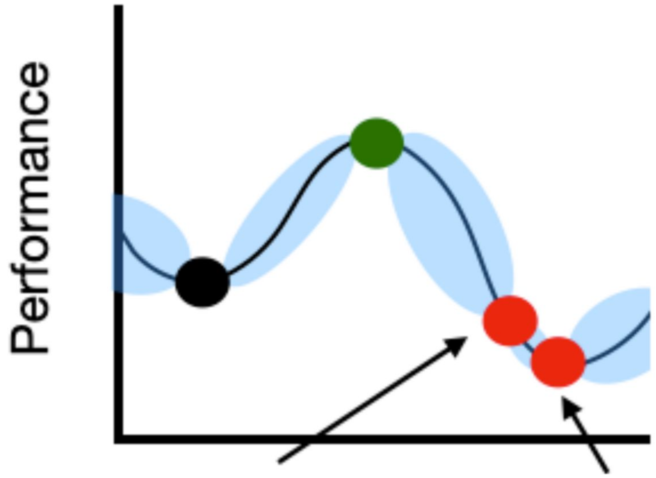
$D_N = \{($$_i, y_i)\}_i^N$

An Aside: GPs for Molecules

Structured Input Spaces

$y_i = f($$_i) + \epsilon_i$

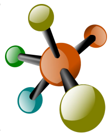
$D_N = \{($$_i, y_i)\}_i^N$

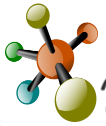


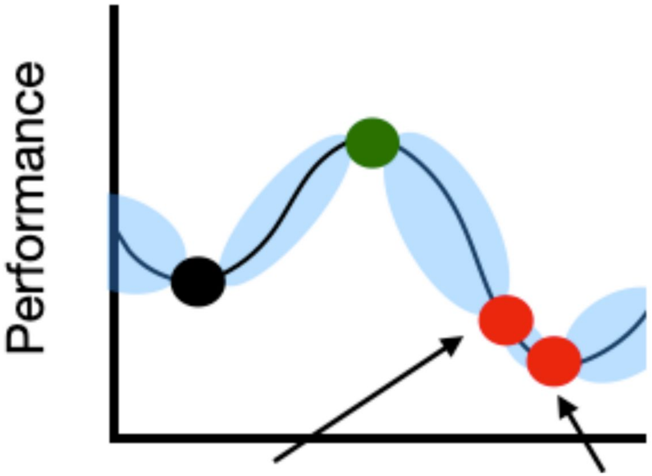
What do we require to define a GP?

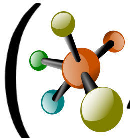
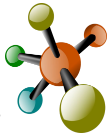
An Aside: GPs for Molecules

Structured Input Spaces

$y_i = f($  $_i) + \epsilon_i$

$D_N = \{($  $_i, y_i)\}_i^N$



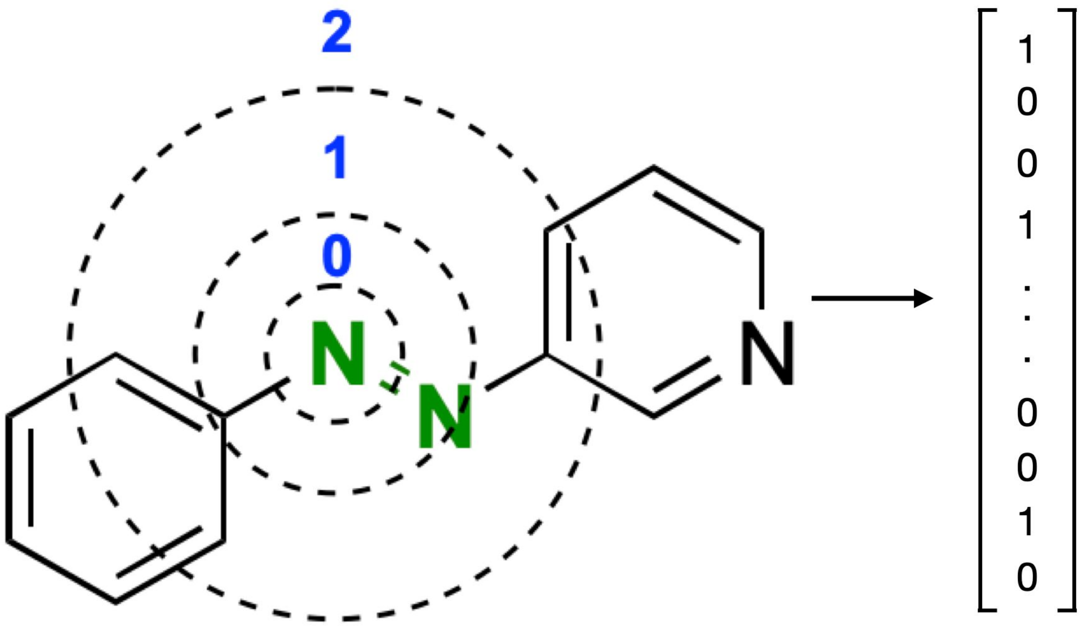
$k($  $_i,$  $_j) = ?$

What do we require to define a GP?

An Aside: GPs for Molecules

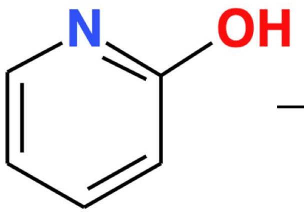
Fingerprint Kernels

$$k(\;_i,\;_j) = k_{\text{linear}}(\Phi(\;_i), \Phi(\;_j))$$

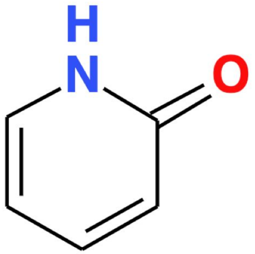


An Aside: GPs for Molecules

String kernels between SMILES strings



$\longrightarrow$ $Oc1ncccc1$



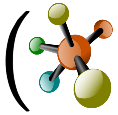
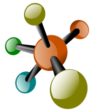
$\longrightarrow$ $O = c1[nH]cccc1$

Automatically choosing next molecules

Using GP posteriors and utility functions

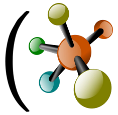
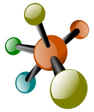
Automatically choosing next molecules

Using GP posteriors and utility functions

- $U_f($  $)$: what is the utility of evaluating  (if it will return f)

Automatically choosing next molecules

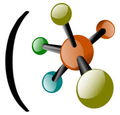
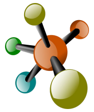
Using GP posteriors and utility functions

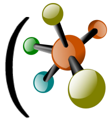
- $U_f($  $)$: what is the utility of evaluating  (if it will return f)

    - $f^\star$ Is best so far

Automatically choosing next molecules

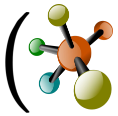
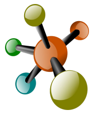
Using GP posteriors and utility functions

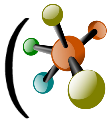
- $U_f($  $)$: what is the utility of evaluating  (if it will return f)

  - $f^\star$ Is best so far
  - Has there been an improvement? $U_f($  $) = \mathbb{1}_{(f > f^\star)}$

Automatically choosing next molecules

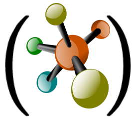
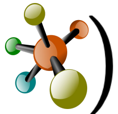
Using GP posteriors and utility functions

- $U_f($  $)$: what is the utility of evaluating  (if it will return f)

  - $f^\star$ Is best so far
  - Has there been an improvement? $U_f($  $) = \mathbb{1}_{(f > f^\star)}$
  - How big was the improvement? $U_f($  $) = \max(f - f^\star, 0)$

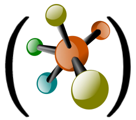
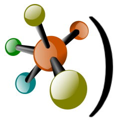
Automatically choosing next molecules

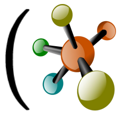
Using GP posteriors and utility functions

- $\alpha($  $) = \mathbb{E}_f[U_f($  $)]$: what utility is predicted by my model of f

Automatically choosing next molecules

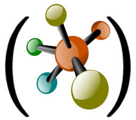
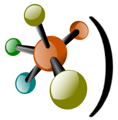
Using GP posteriors and utility functions

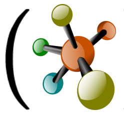
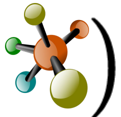
- $\alpha($  $) = \mathbb{E}_f[U_f($  $)]$: what utility is predicted by my model of f

- What the probability of improvement? $\alpha_{\mathrm{PI}}($  $) = \mathbb{E}_f\left[\mathbb{1}_{(f > f^\star)}\right]$

Automatically choosing next molecules

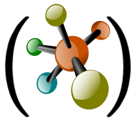
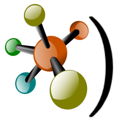
Using GP posteriors and utility functions

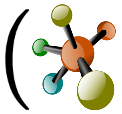
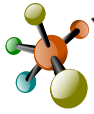
- $\alpha($$) = \mathbb{E}_f[U_f($$)]$: what utility is predicted by my model of f

  - What the probability of improvement? $\alpha_{\mathrm{PI}}($$) = \mathbb{E}_f\left[\mathbb{1}_{(f>f^\star)}\right]$
  - How much improvement do we expect? $\alpha_{\mathrm{EI}}($$) = \mathbb{E}_f[\max(f - f^\star, 0)]$

Automatically choosing next molecules

Using GP posteriors and utility functions

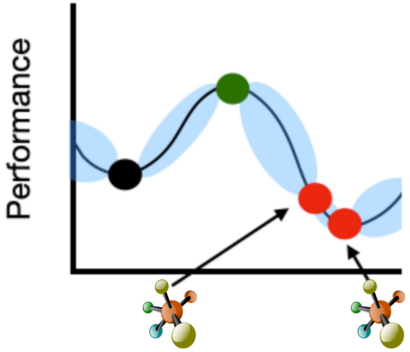
- $\alpha($$) = \mathbb{E}_f[U_f($$)]$: what utility is predicted by my model of f

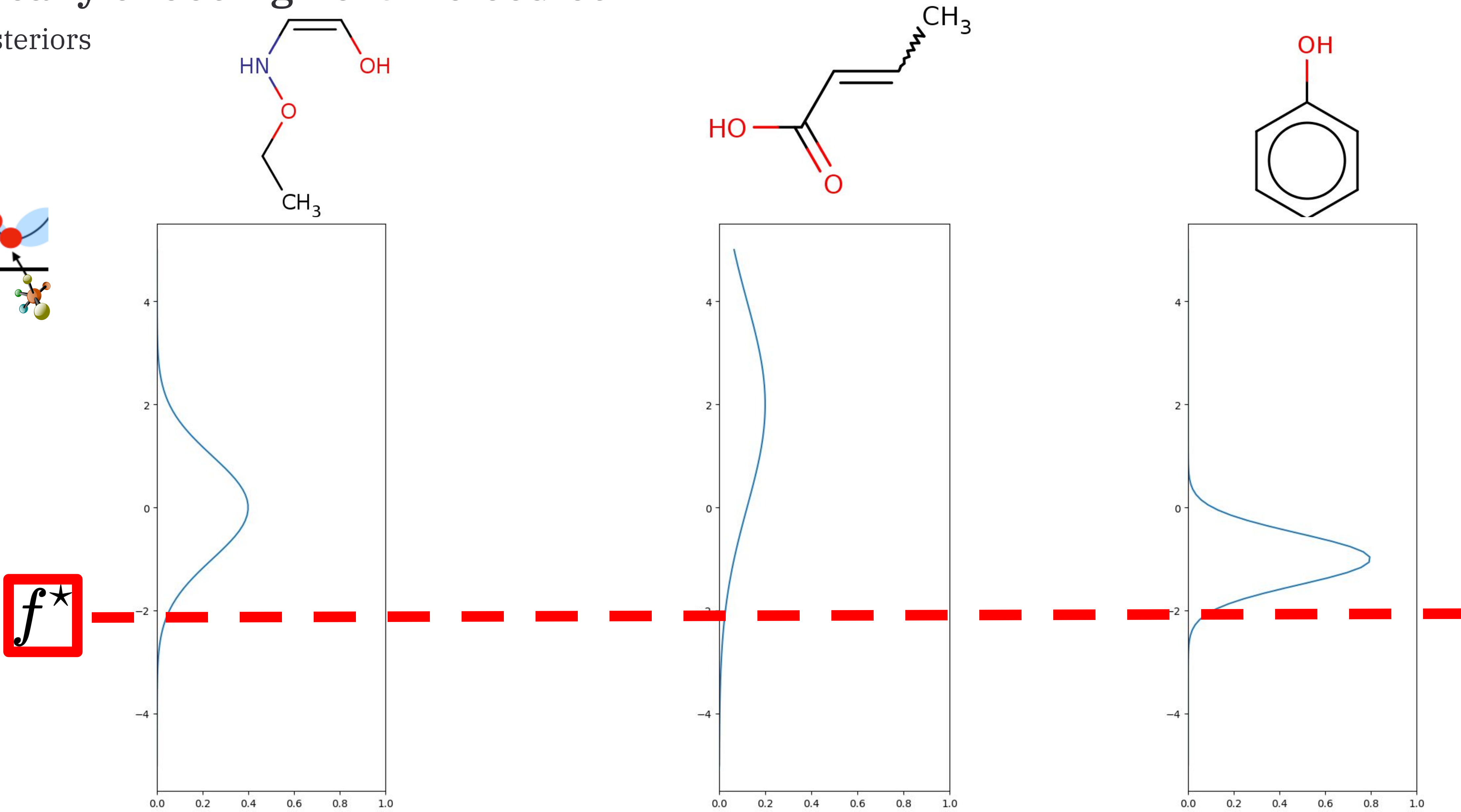
  - What the probability of improvement? $\alpha_{\mathrm{PI}}($$) = \mathbb{E}_f\left[\mathbb{1}_{(f > f^\star)}\right]$
  - How much improvement do we expect? $\alpha_{\mathrm{EI}}($$) = \mathbb{E}_f[\max(f - f^\star, 0)]$

$$f \sim \mathcal{N}(\mu, \sigma^2)$$

Automatically choosing next molecules

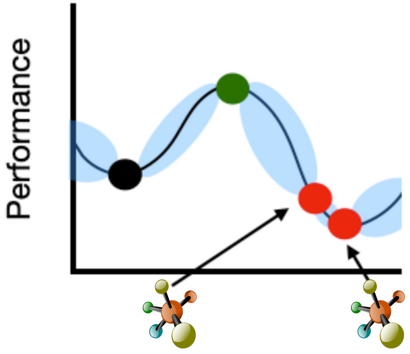
Using GP posteriors

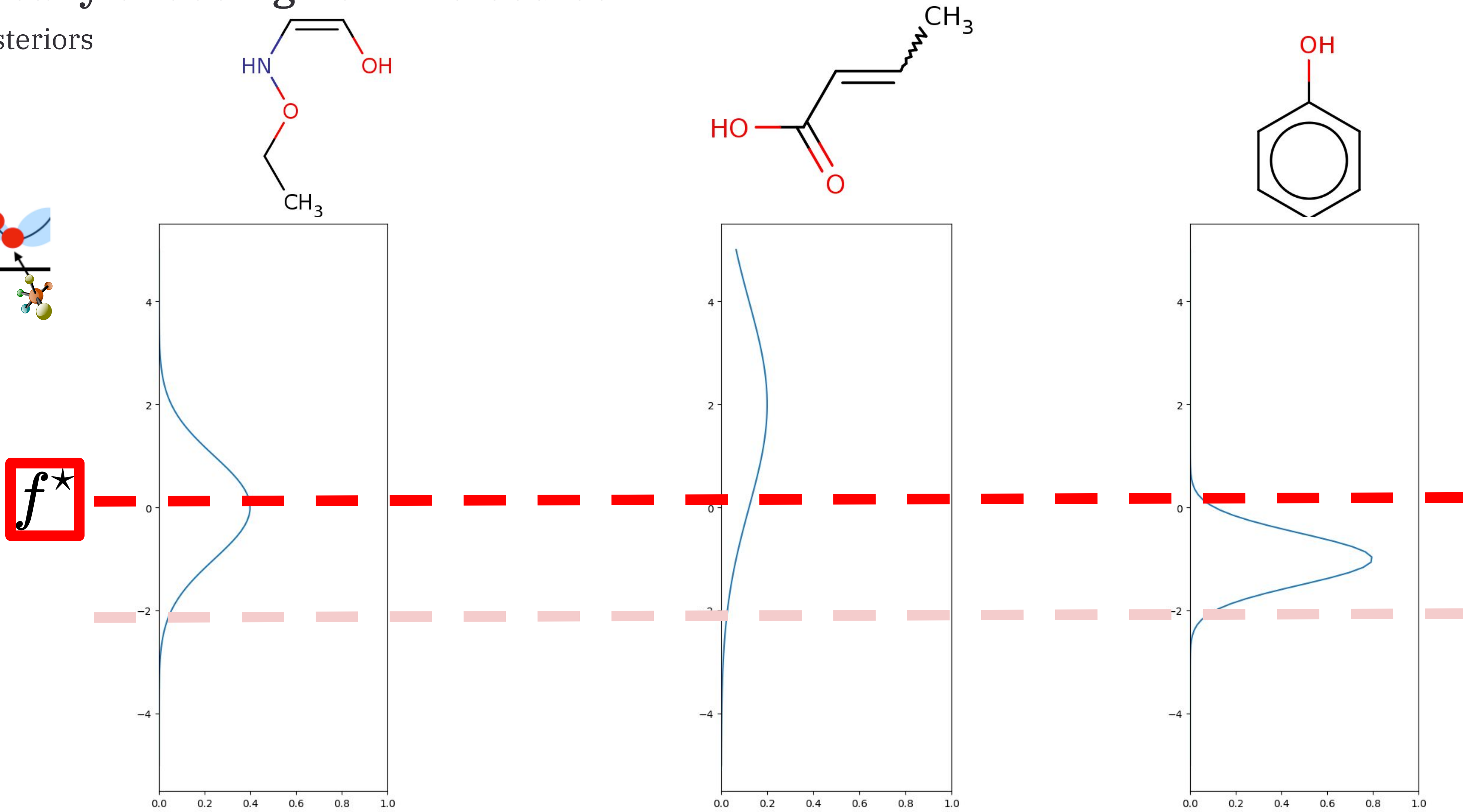




Automatically choosing next molecules

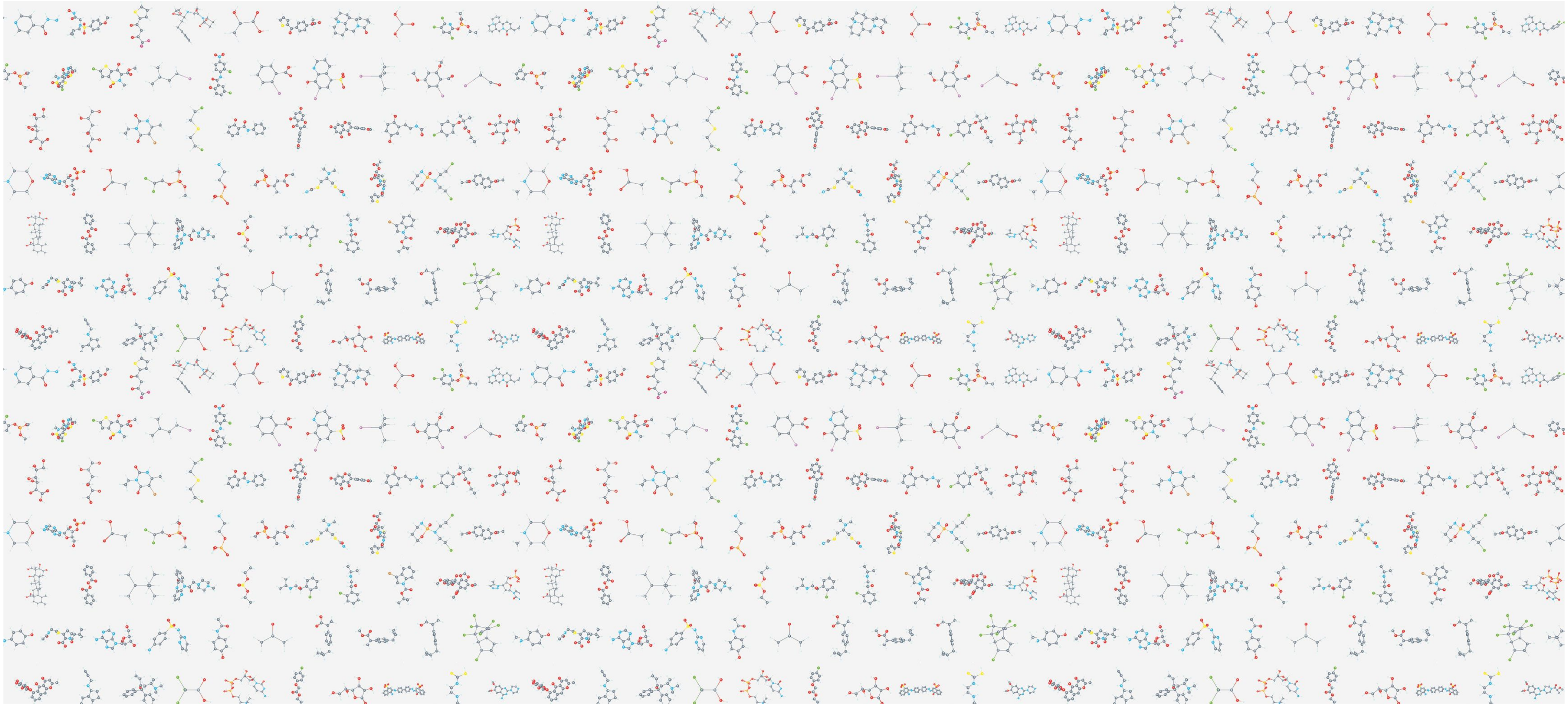
Using GP posteriors





Automatically choosing next molecules

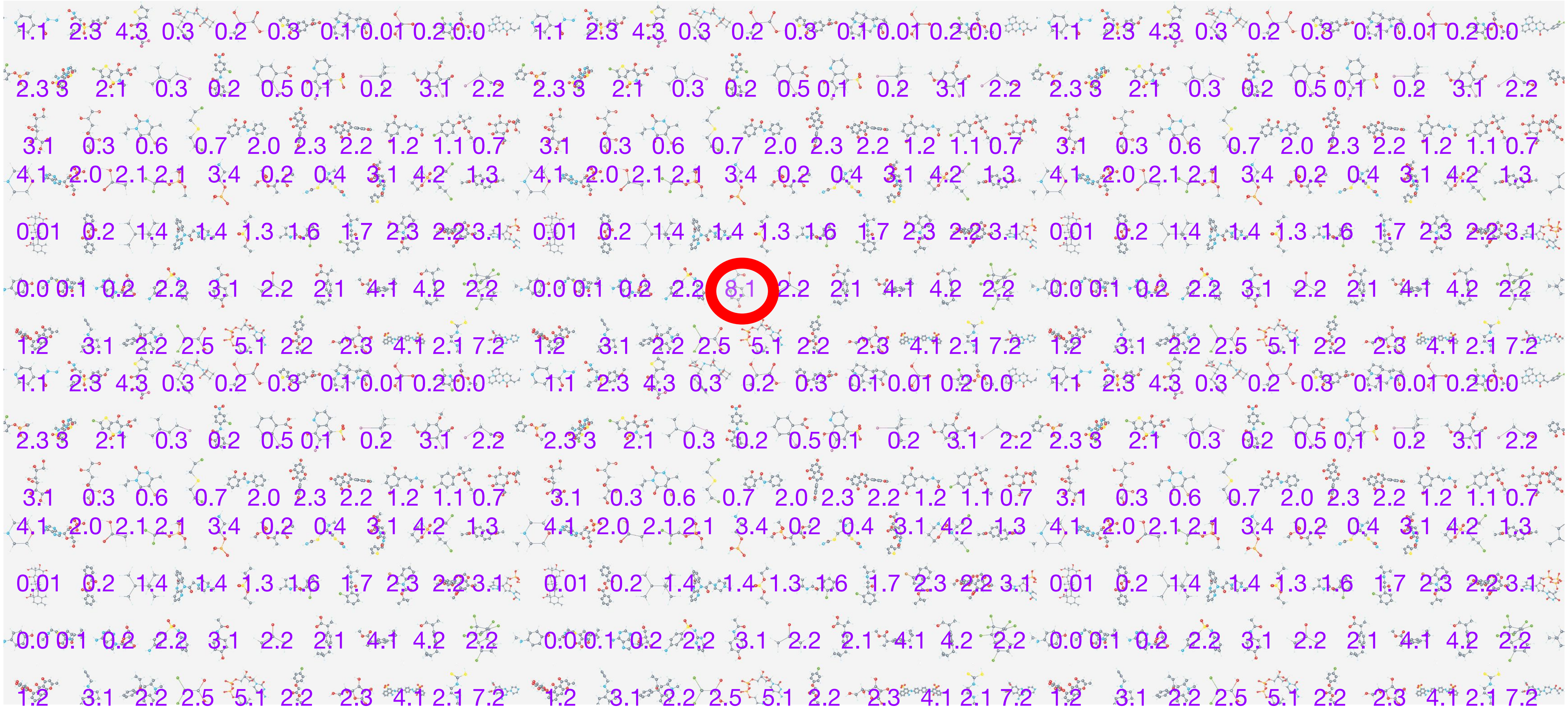
Calc acquisition function and pick best



https://www.horizons-mag.ch/2018/09/06/the-universe-of-molecules/

Automatically choosing next molecules

Calc acquisition function and pick best



https://www.horizons-mag.ch/2018/09/06/the-universe-of-molecules/

Automatically choosing next molecules

Full Bayesian optimisation loop

1. Evaluate 2 random molecules

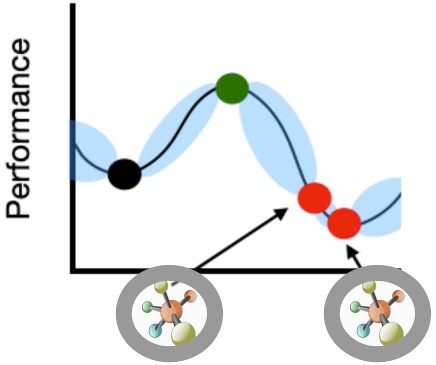
https://www.horizons-mag.ch/2018/09/06/the-universe-of-molecules/

Automatically choosing next molecules

Full Bayesian optimisation loop

1. Evaluate 2 random molecules

2. Fit GP model to measurements



https://www.horizons-mag.ch/2018/09/06/the-universe-of-molecules/

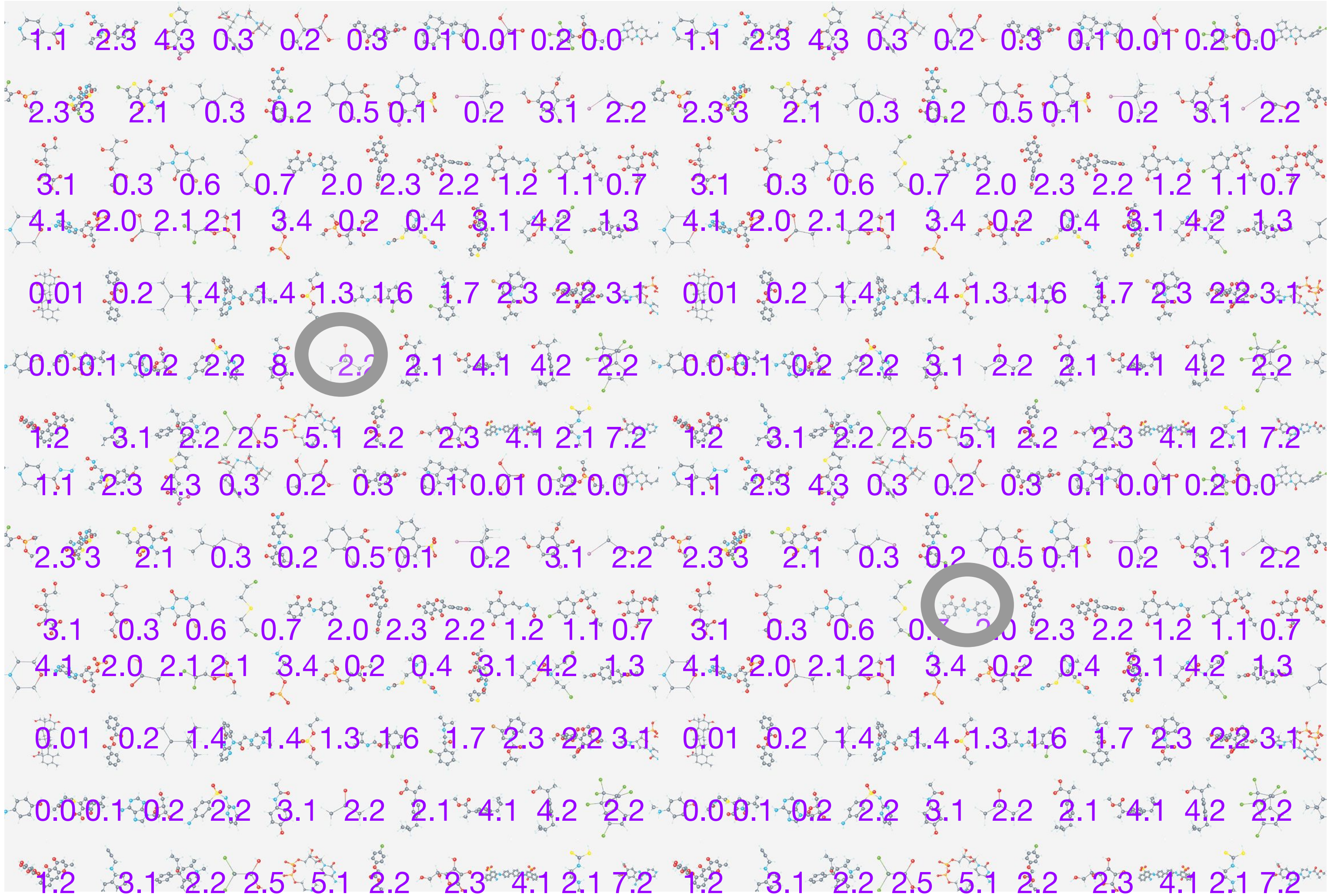
UNIVERSITY OF CAMBRIDGE

Lancaster University

Automatically choosing next molecules

Full Bayesian optimisation loop

1. Evaluate 2 random molecules
2. Fit GP model to measurements
3. Calc acquisition function

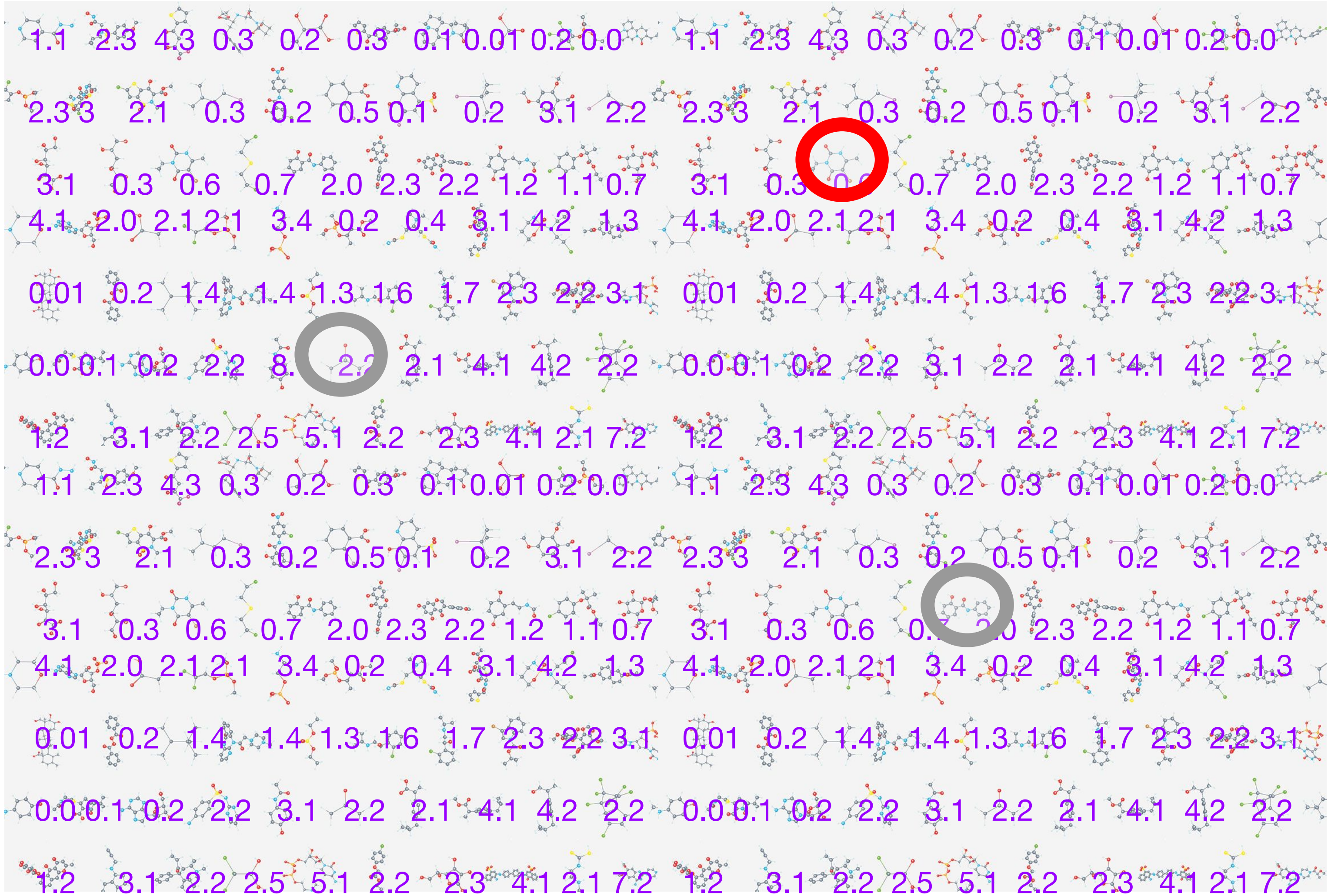


https://www.horizons-mag.ch/2018/09/06/the-universe-of-molecules/

Automatically choosing next molecules

Full Bayesian optimisation loop

1. Evaluate 2 random molecules
2. Fit GP model to measurements
3. Calc acquisition function
4. Choose new molecule



https://www.horizons-mag.ch/2018/09/06/the-universe-of-molecules/

Automatically choosing next molecules

Full Bayesian optimisation loop

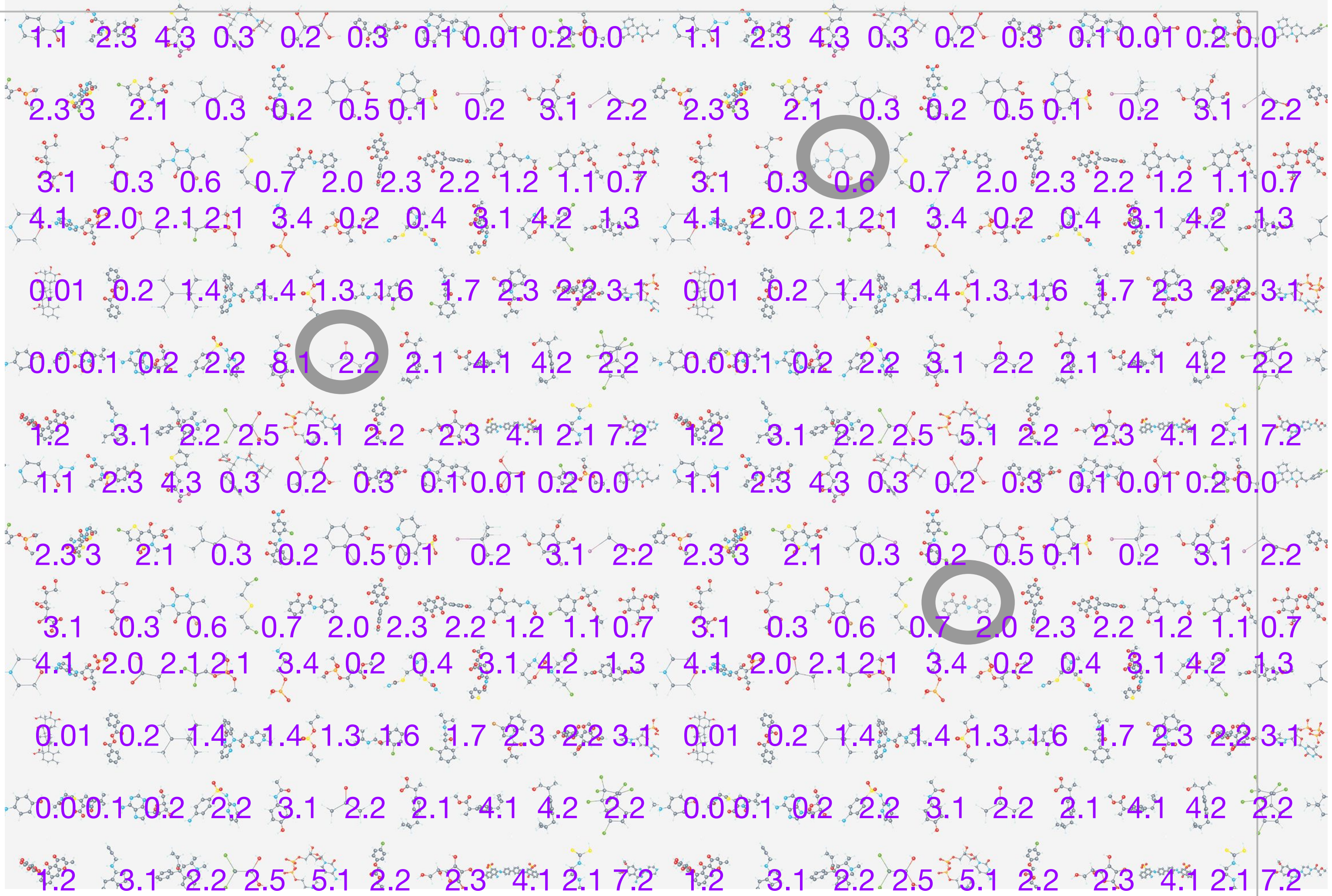
1. Evaluate 2 random molecules
2. Fit GP model to measurements
3. Calc acquisition function
4. Choose new molecule
5. Go to step 2.

https://www.horizons-mag.ch/2018/09/06/the-universe-of-molecules/

Automatically choosing next molecules

Full Bayesian optimisation loop

1. Evaluate 2 random molecules
2. Fit GP model to measurements
3. Calc new acquisition function
4. Choose new molecule
5. Go to step 2.

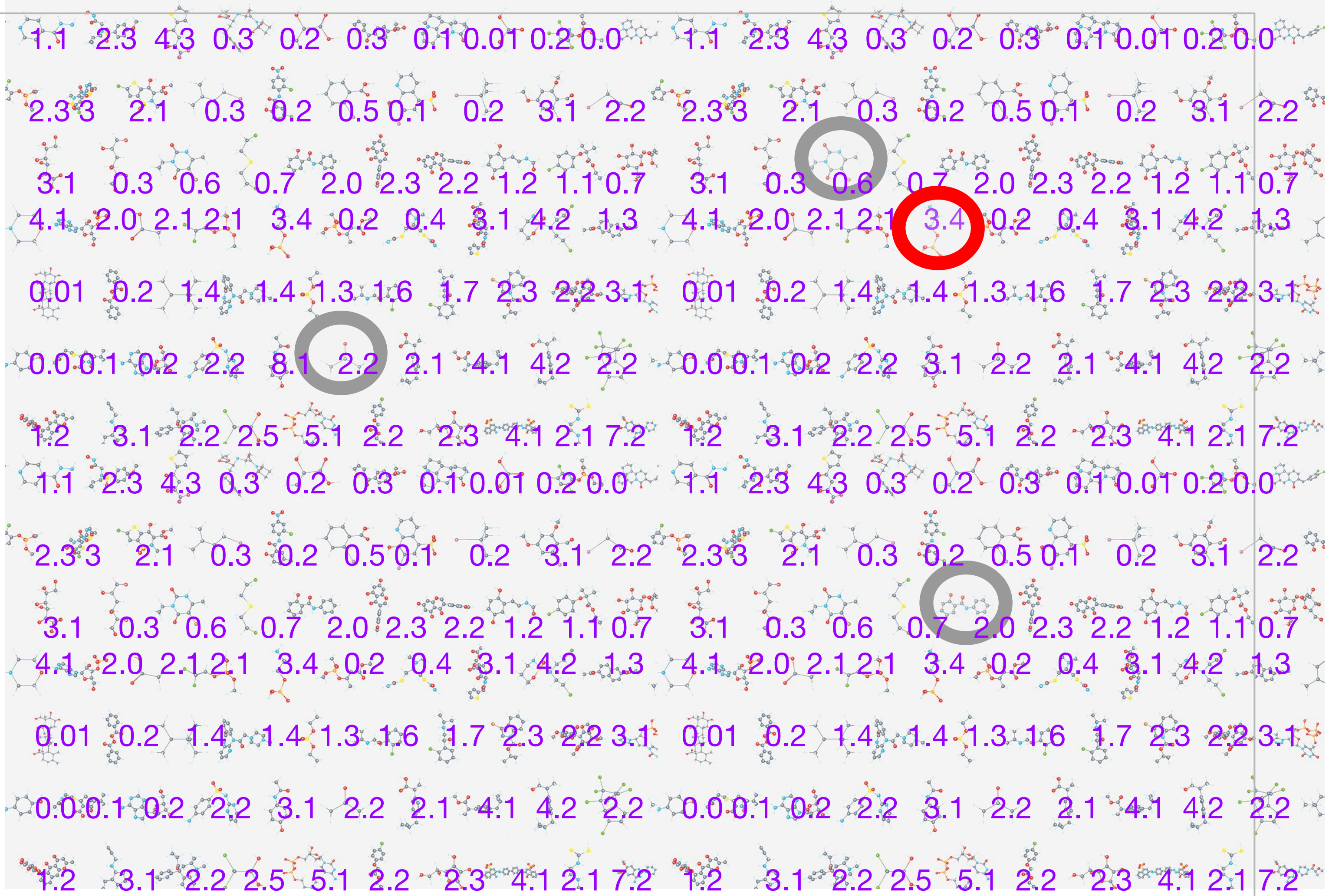


https://www.horizons-mag.ch/2018/09/06/the-universe-of-molecules/

Automatically choosing next molecules

Full Bayesian optimisation loop

1. Evaluate 2 random molecules
2. Fit GP model to measurements
3. Calc new acquisition function
4. Choose new molecule
5. Go to step 2.

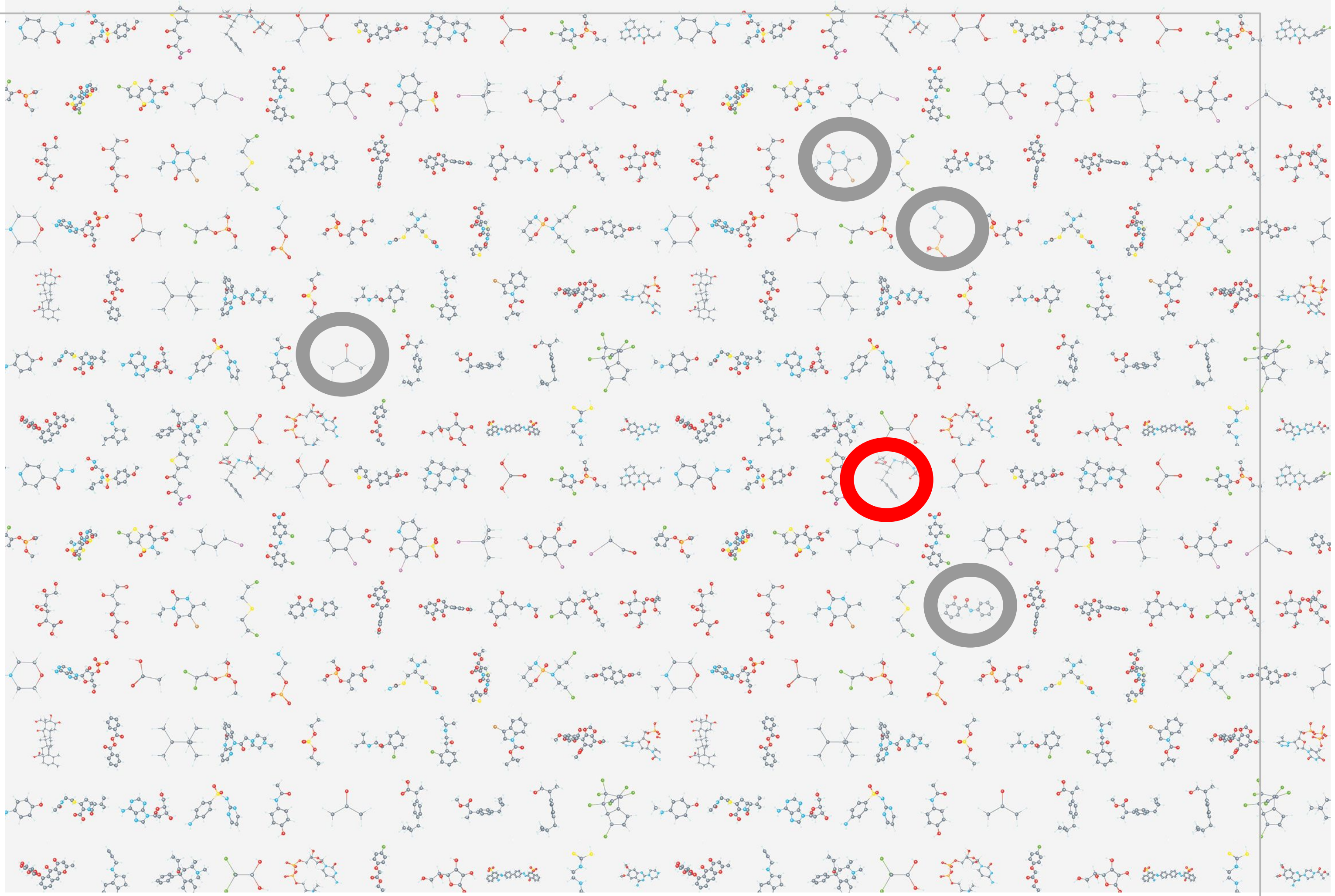


https://www.horizons-mag.ch/2018/09/06/the-universe-of-molecules/

Automatically choosing next molecules

Full Bayesian optimisation loop

1. Evaluate 2 random molecules
2. Fit GP model to measurements
3. Calc new acquisition function
4. Choose new molecule
5. Go to step 2.



https://www.horizons-mag.ch/2018/09/06/the-universe-of-molecules/

UNIVERSITY OF CAMBRIDGE

Lancaster University

Automatically choosing next molecules

Full Bayesian optimisation loop

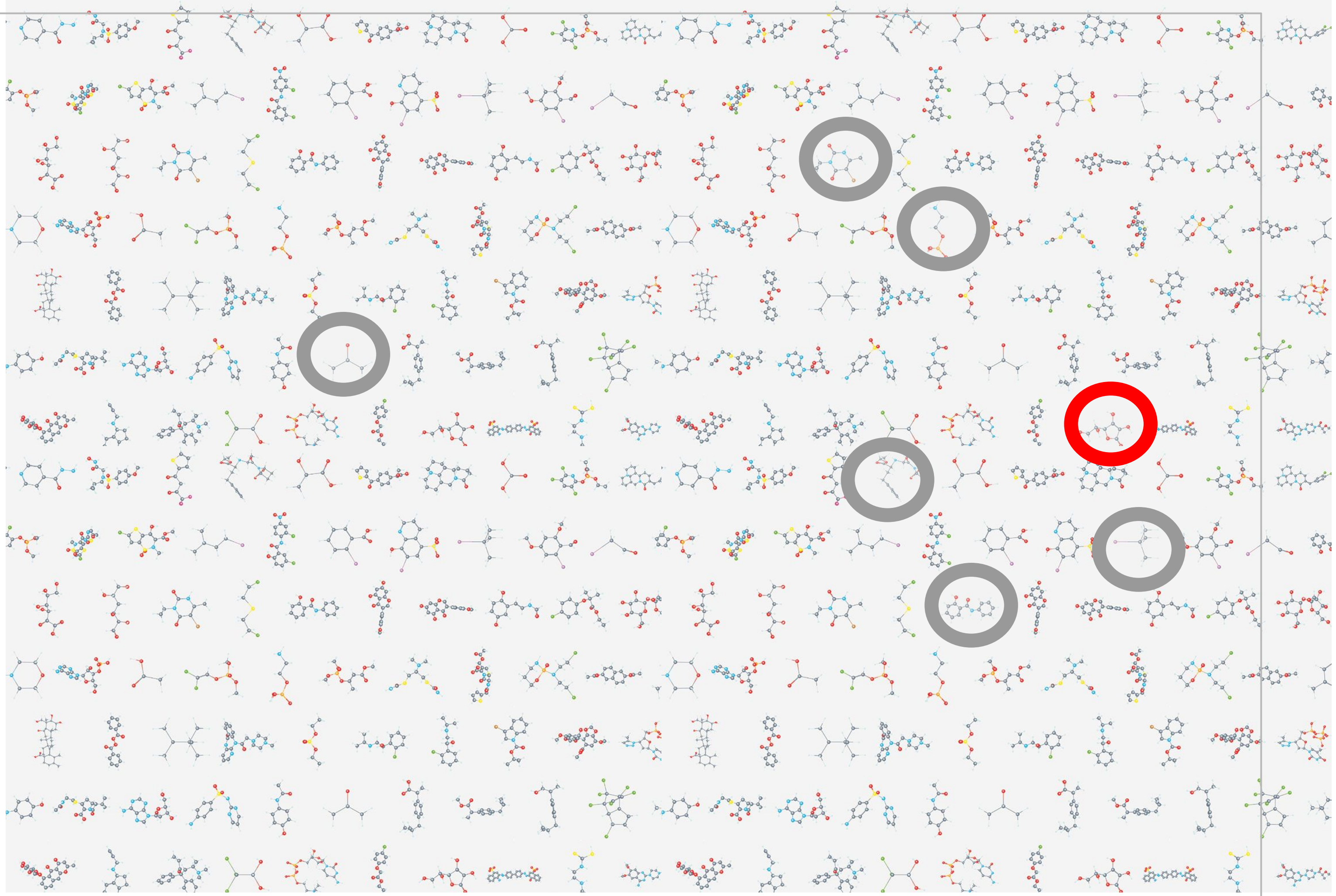
1. Evaluate 2 random molecules
2. Fit GP model to measurements
3. Calc new acquisition function
4. Choose new molecule
5. Go to step 2.

https://www.horizons-mag.ch/2018/09/06/the-universe-of-molecules/

Automatically choosing next molecules

Full Bayesian optimisation loop

1. Evaluate 2 random molecules
2. Fit GP model to measurements
3. Calc new acquisition function
4. Choose new molecule
5. Go to step 2.



https://www.horizons-mag.ch/2018/09/06/the-universe-of-molecules/

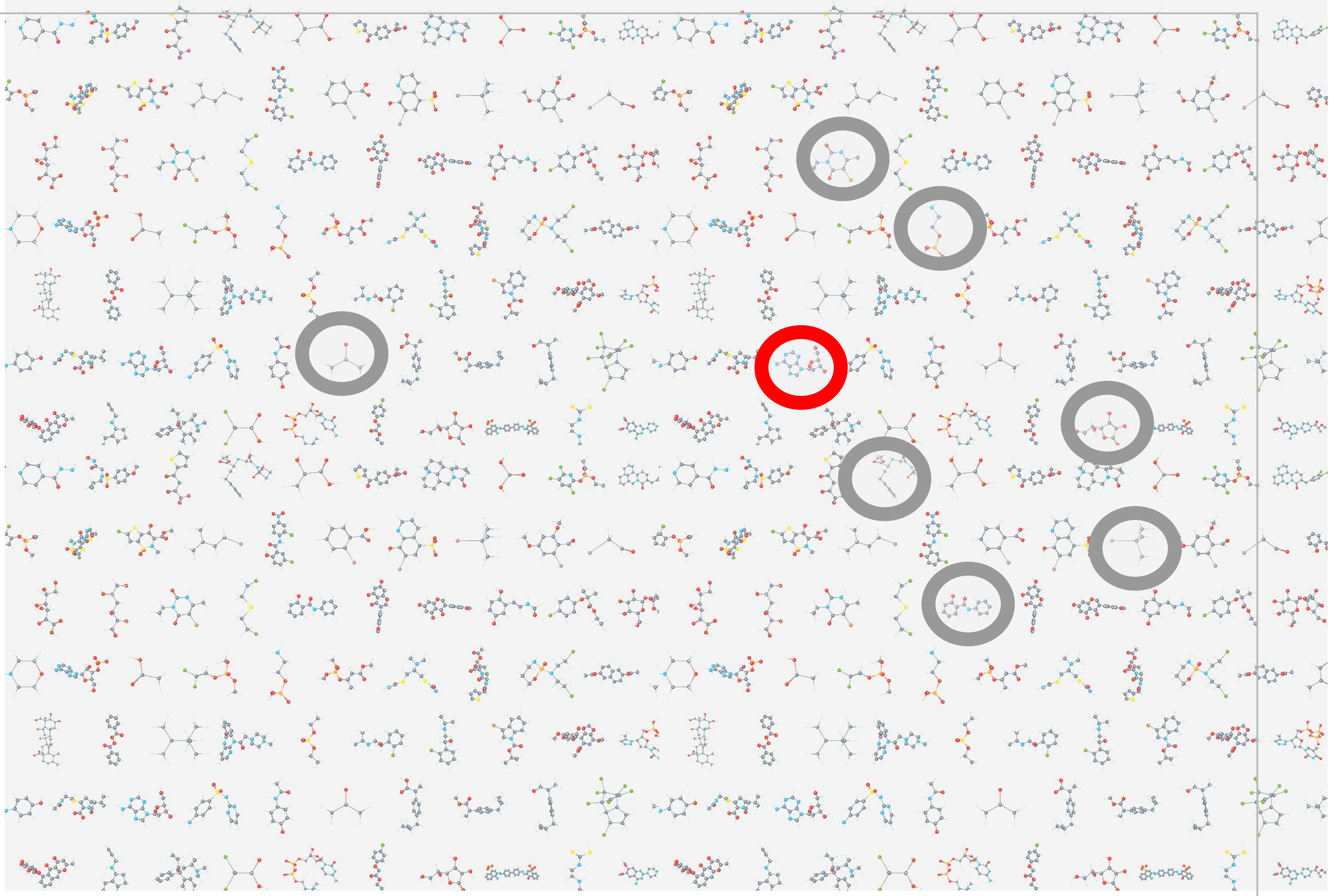
UNIVERSITY OF CAMBRIDGE

Lancaster University

Automatically choosing next molecules

Full Bayesian optimisation loop

1. Evaluate 2 random molecules
2. Fit GP model to measurements
3. Calc new acquisition function
4. Choose new molecule
5. Go to step 2.



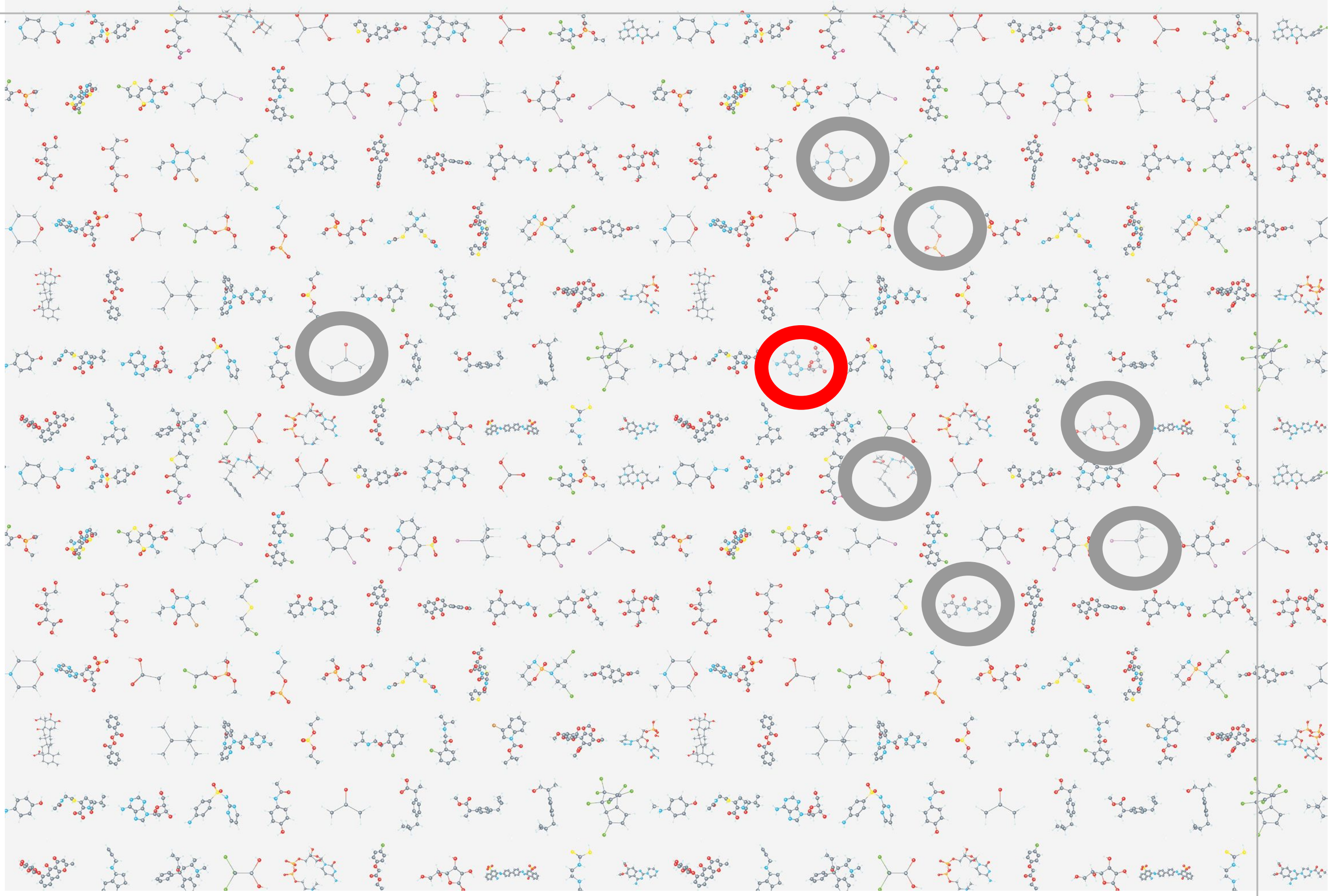
https://www.horizons-mag.ch/2018/09/06/the-universe-of-molecules/

Automatically choosing next molecules

Full Bayesian optimisation loop

1. Evaluate 2 random molecules
2. Fit GP model to measurements
3. Calc new acquisition function
4. Choose new molecule
5. Go to step 2.

And so on ....



https://www.horizons-mag.ch/2018/09/06/the-universe-of-molecules/

What about standard optimisation problems?

i.e. infinite candidates

BO Demo

Let's find the maximum of a 1D function:

BO Demo

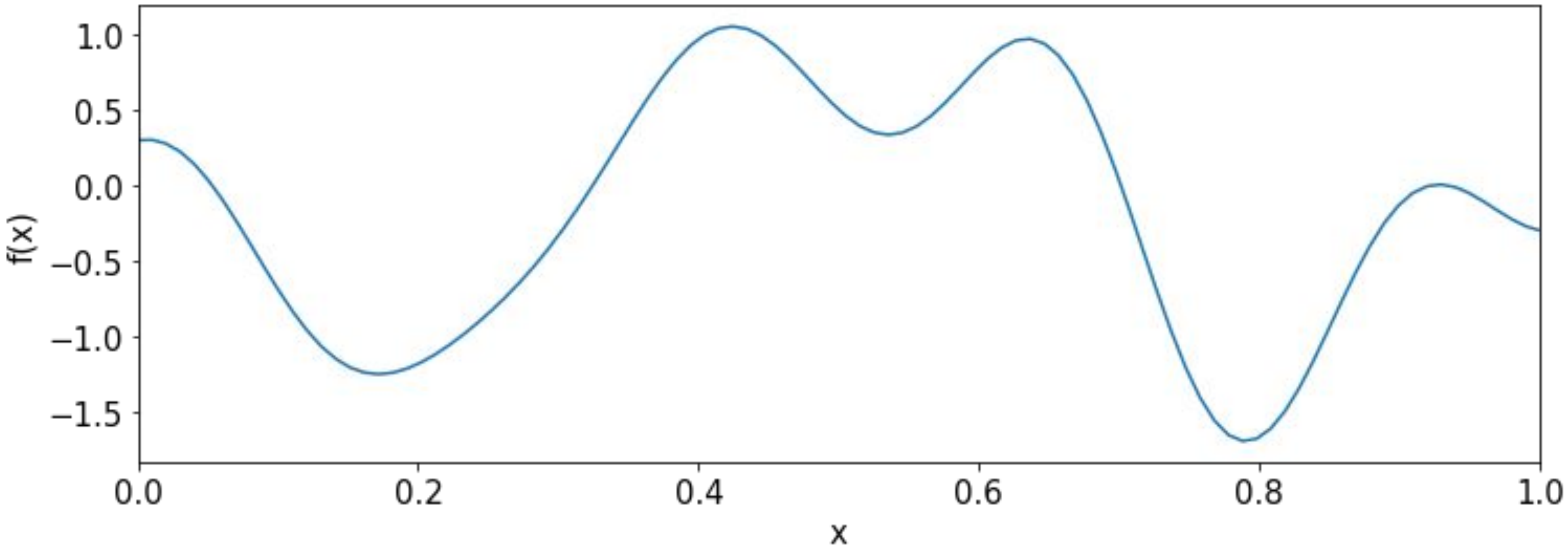
Let's find the maximum of a 1D function:

Using as **few** function evaluations as possible!

BO Demo

Let's find the maximum of a 1D function:

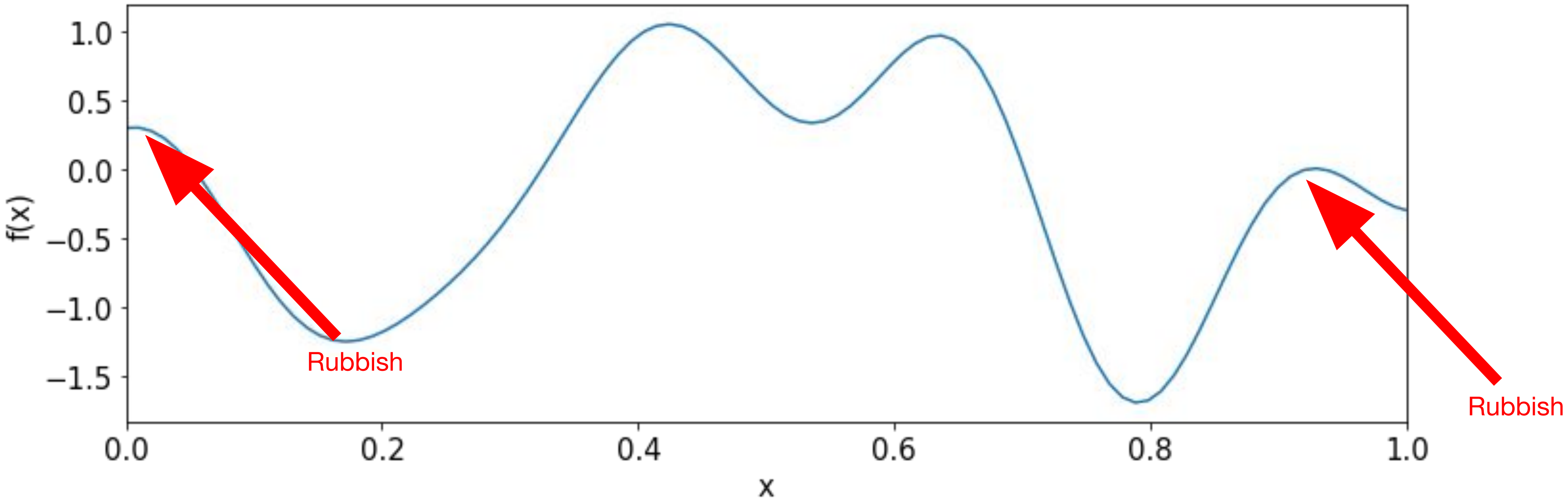
Using as **few** function evaluations as possible!



BO Demo

Let's find the maximum of a 1D function:

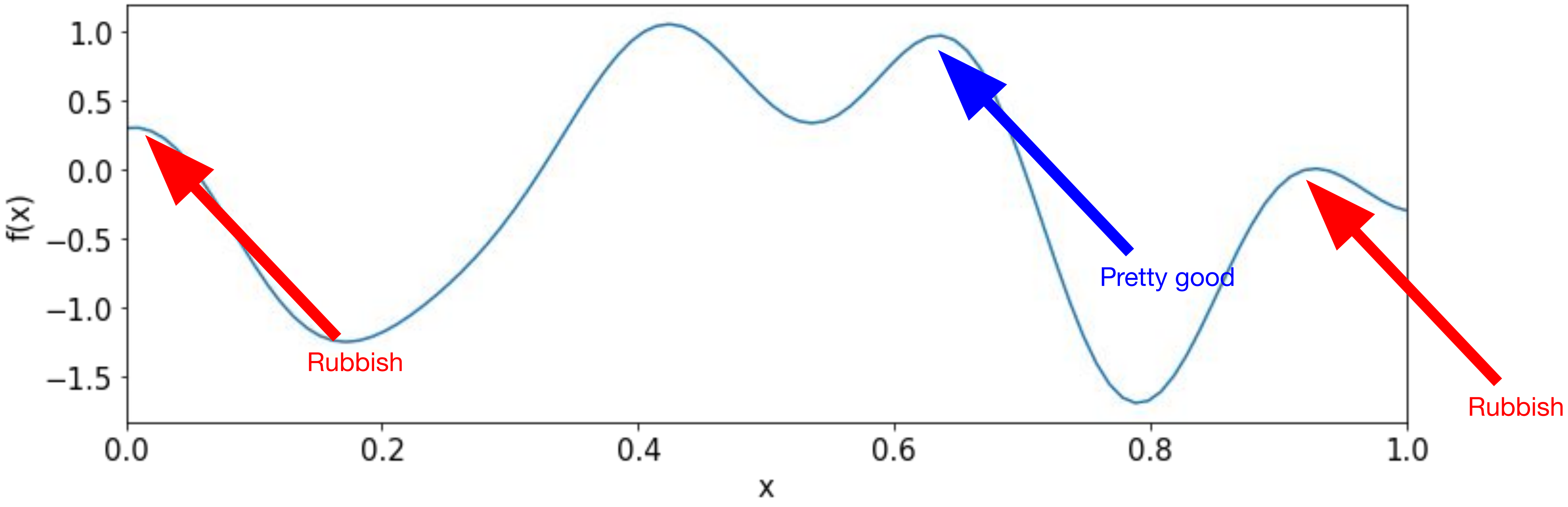
Using as **few** function evaluations as possible!



BO Demo

Let's find the maximum of a 1D function:

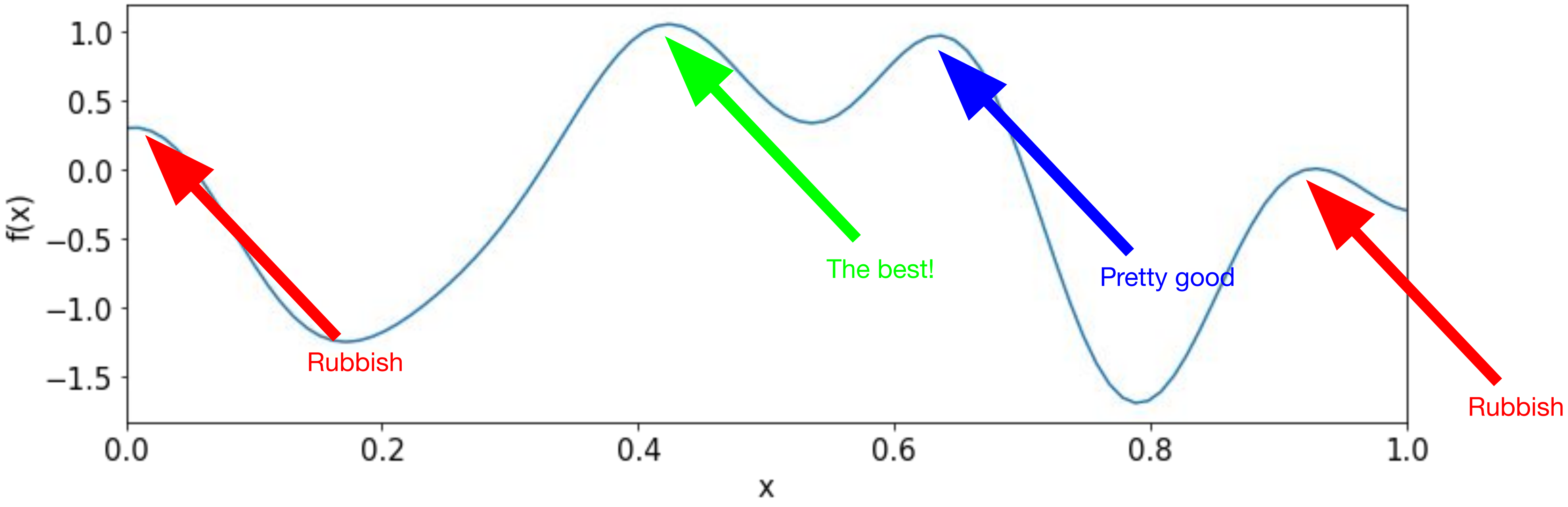
Using as **few** function evaluations as possible!



BO Demo

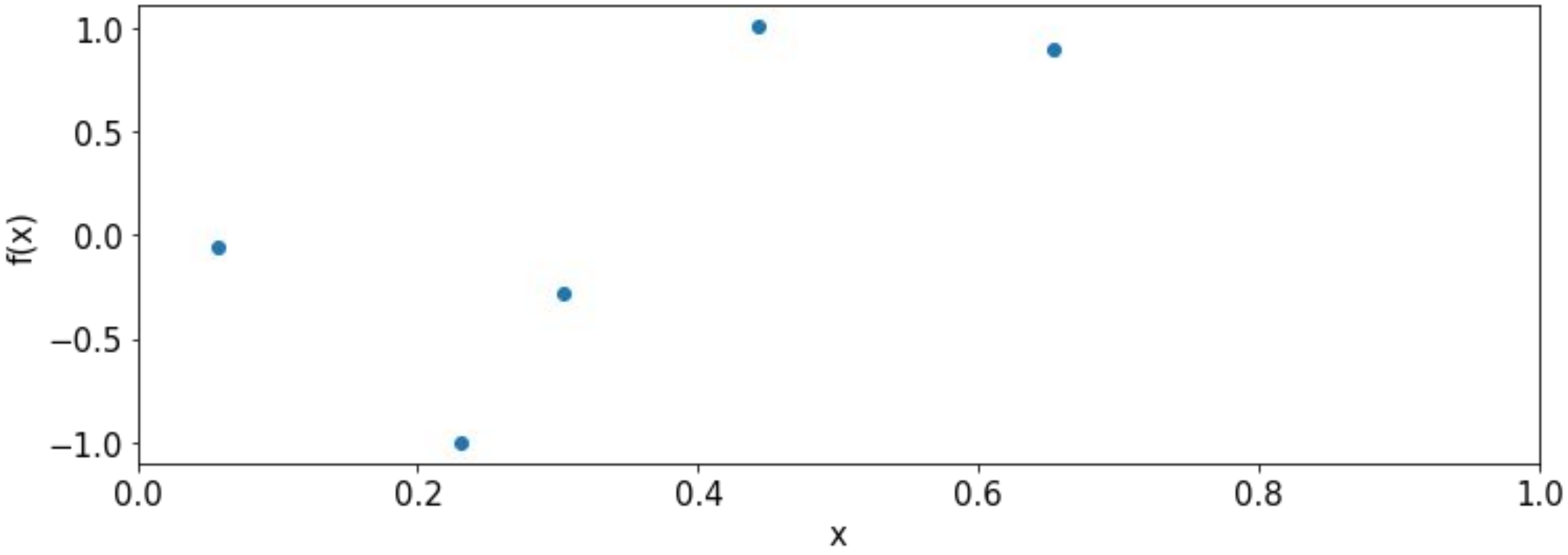
Let's find the maximum of a 1D function:

Using as **few** function evaluations as possible!



BO Demo

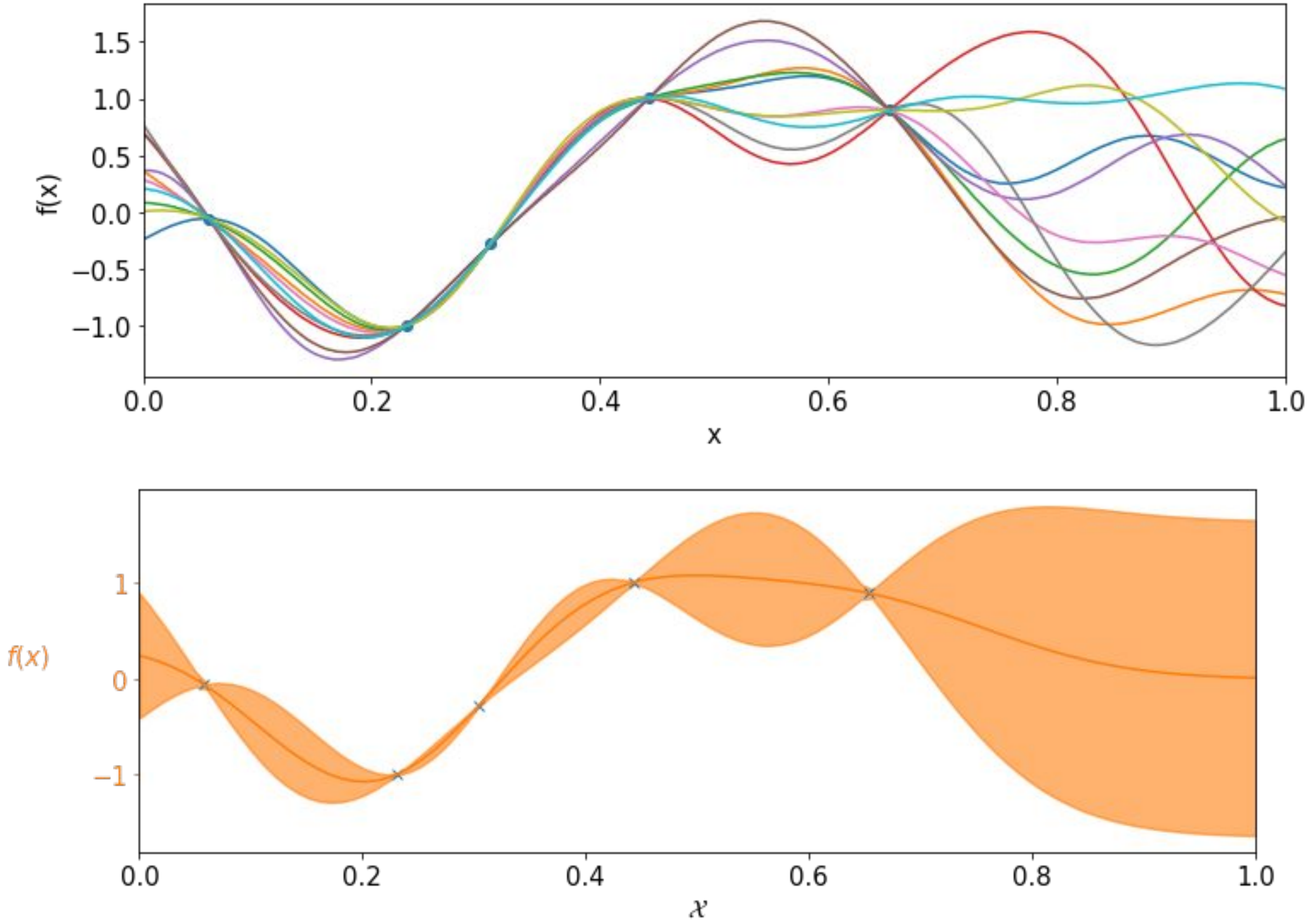
Suppose we make 5 evaluations



Where should we next evaluate? Explore/Exploit?

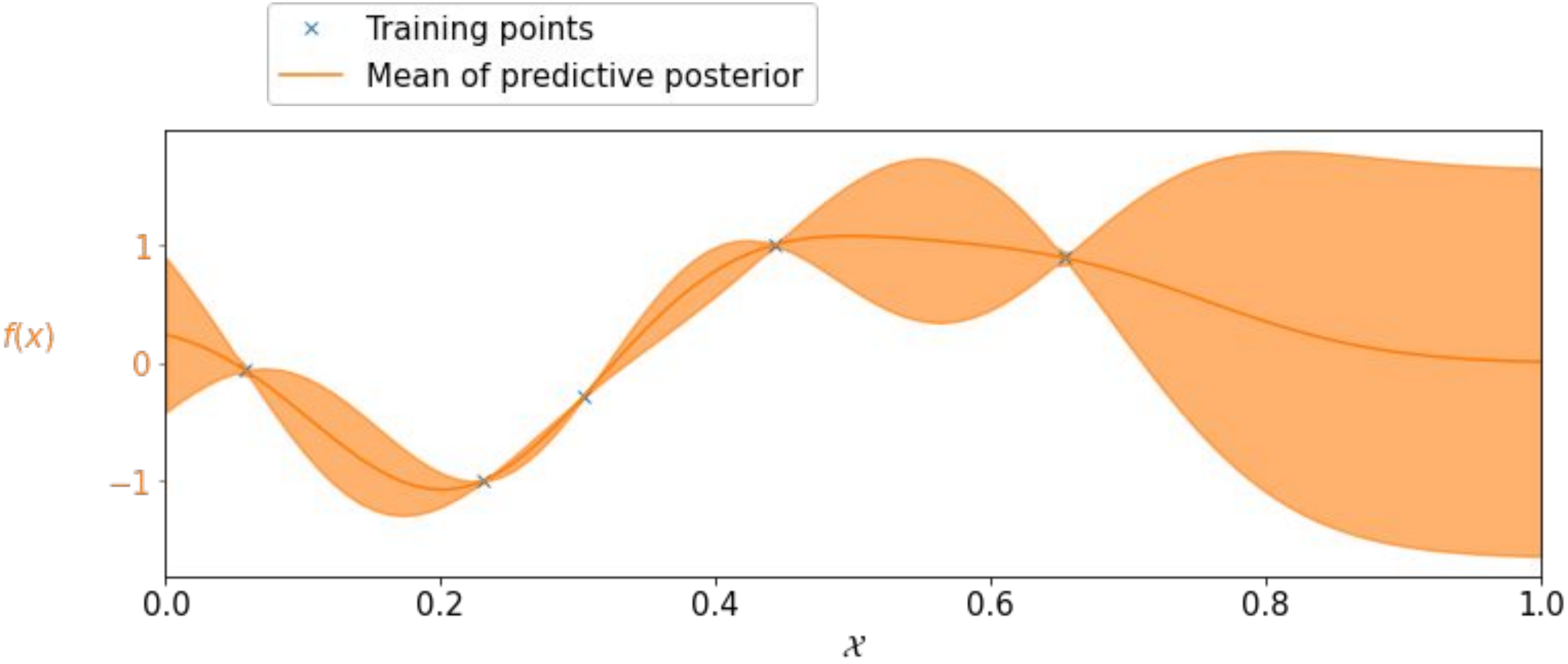
How to automate BO: step 1

Use a statistical model like a Gaussian process



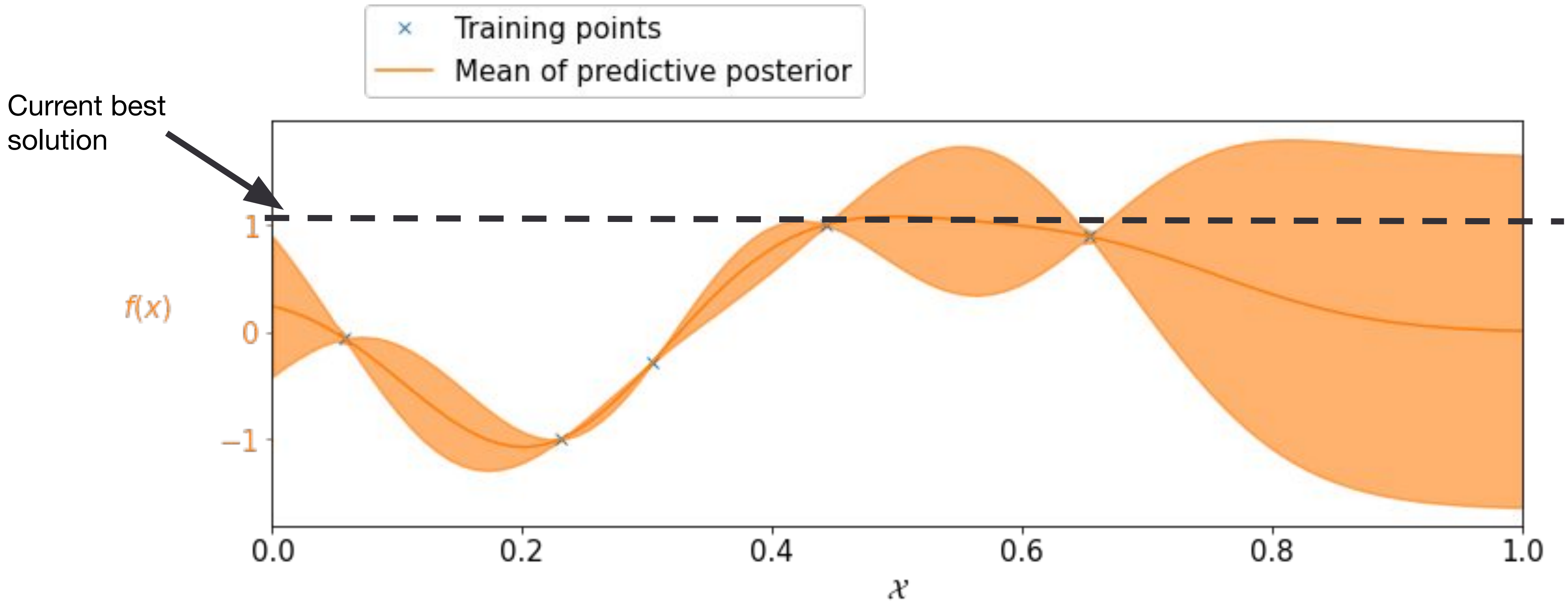
How to automate BO: step 2

Automated decision making via an acquisition function like expected improvement



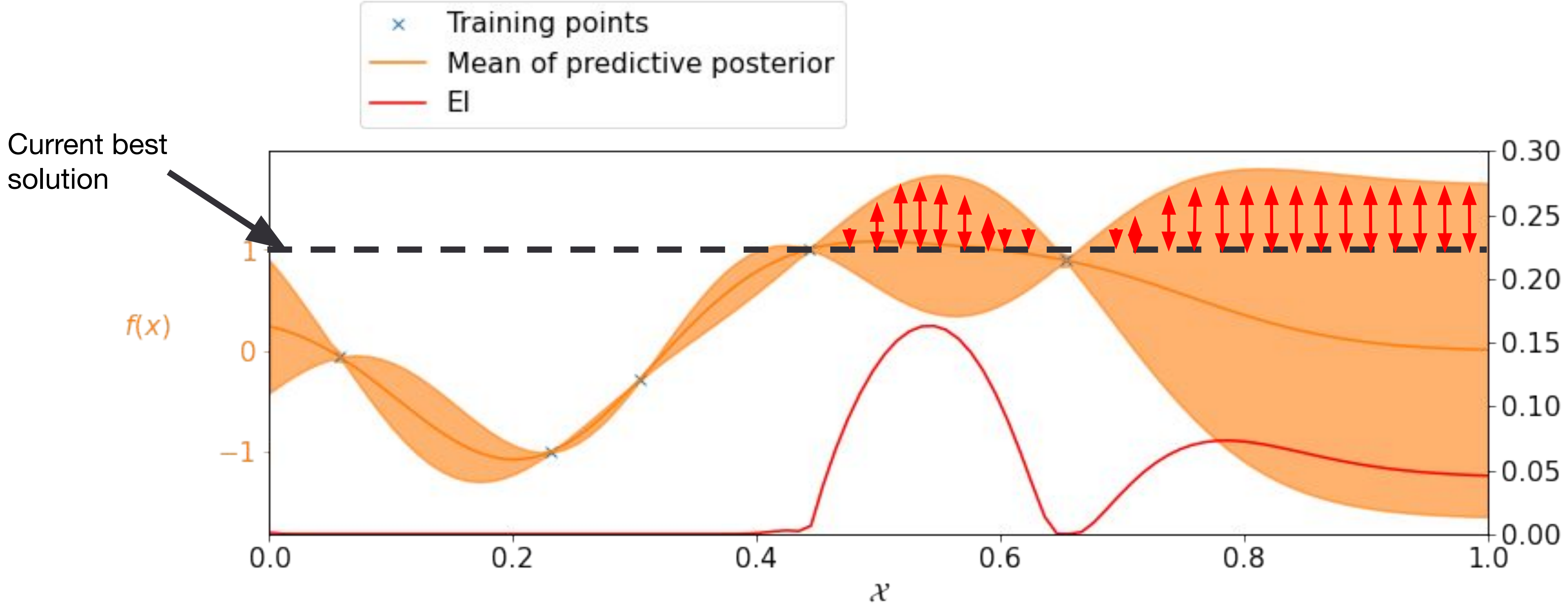
How to automate BO: step 2

Automated decision making via an acquisition function like expected improvement



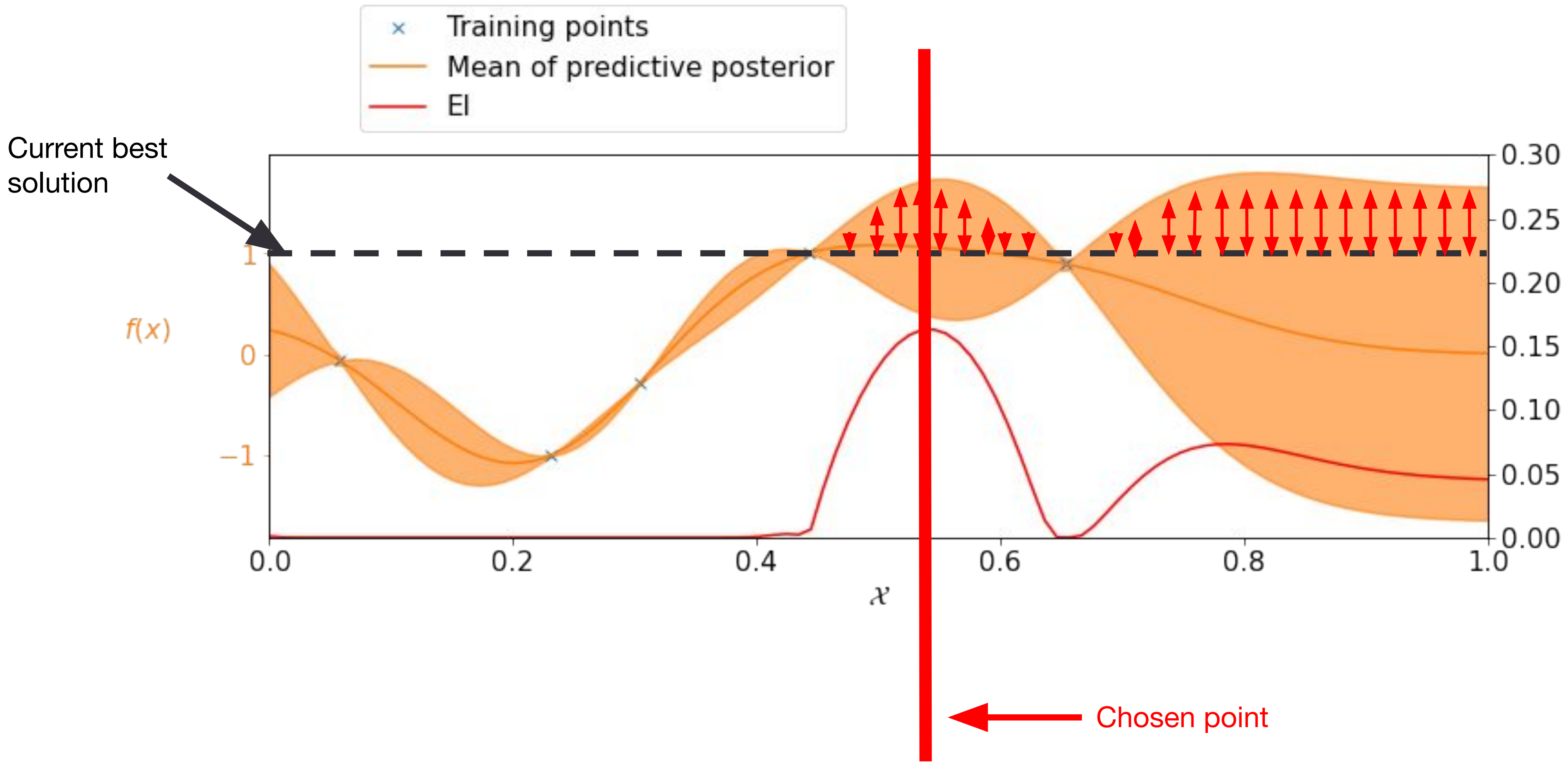
How to automate BO: step 2

Automated decision making via an acquisition function like expected improvement



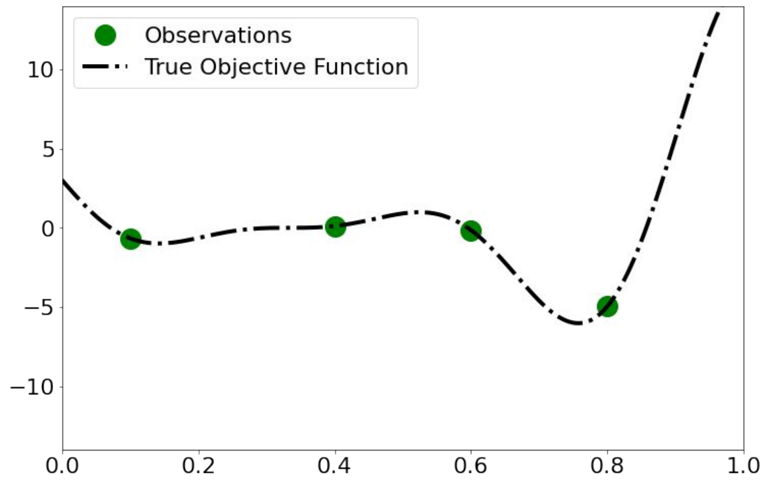
How to automate BO: step 2

Automated decision making via an acquisition function like expected improvement



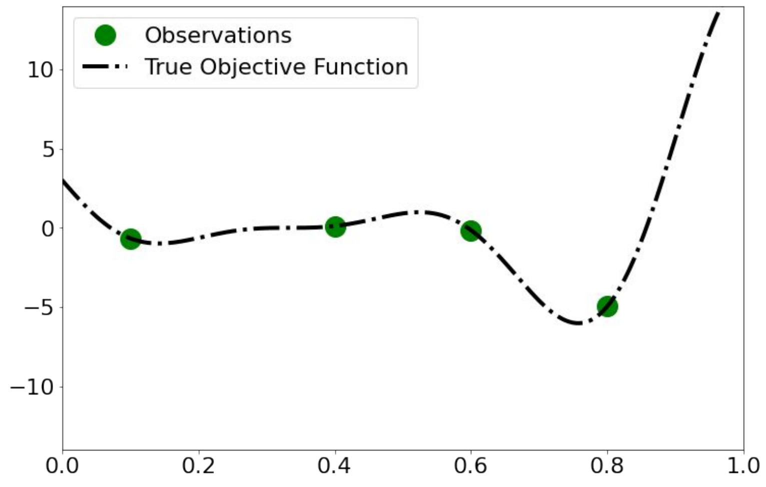
Expected Improvement

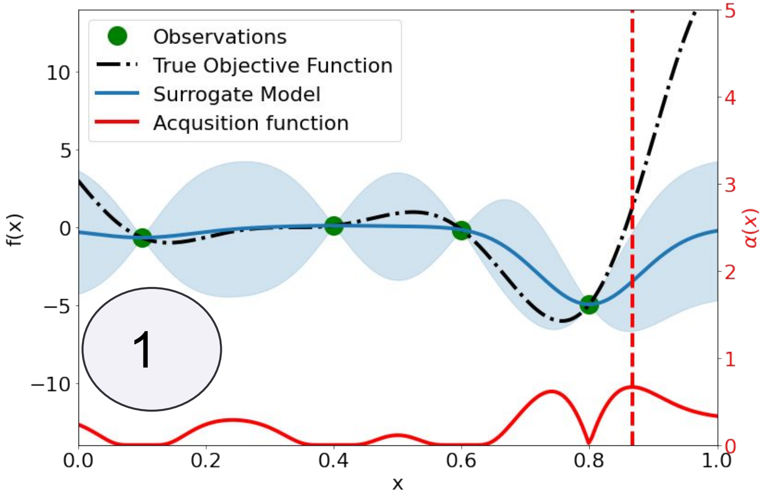
Demo BO loop



Expected Improvement

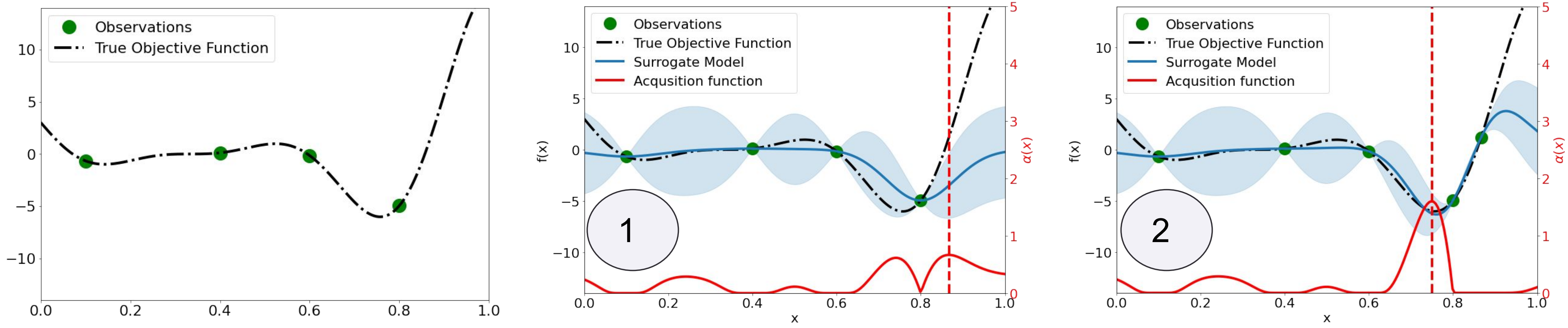
Demo BO loop





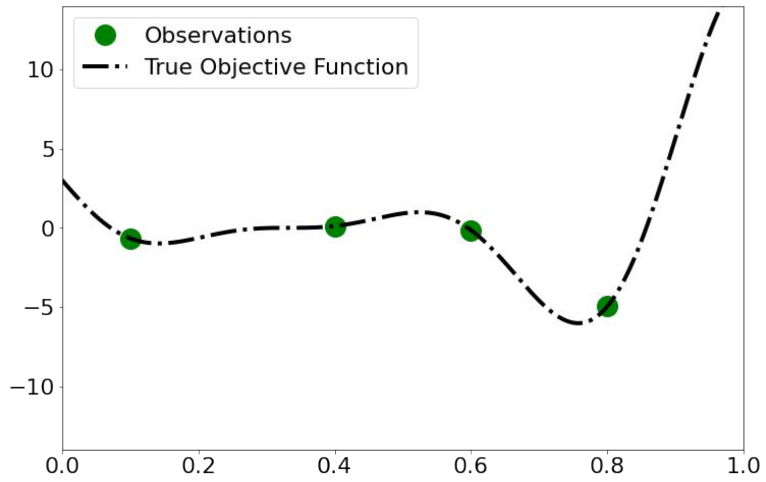
Expected Improvement

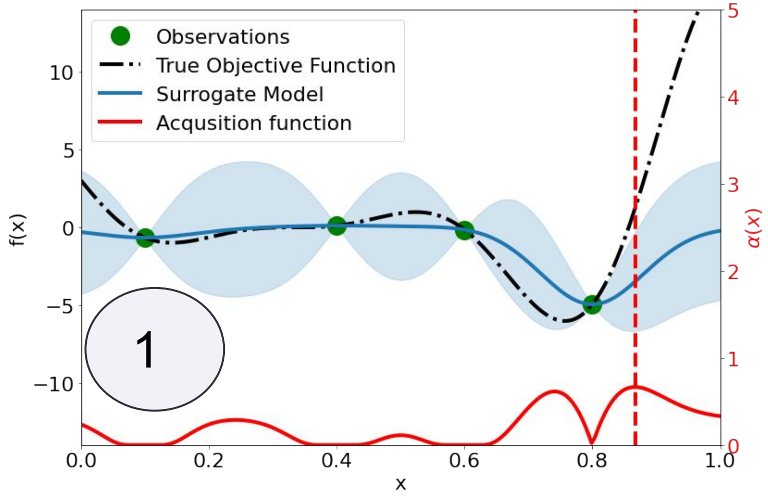
Demo BO loop

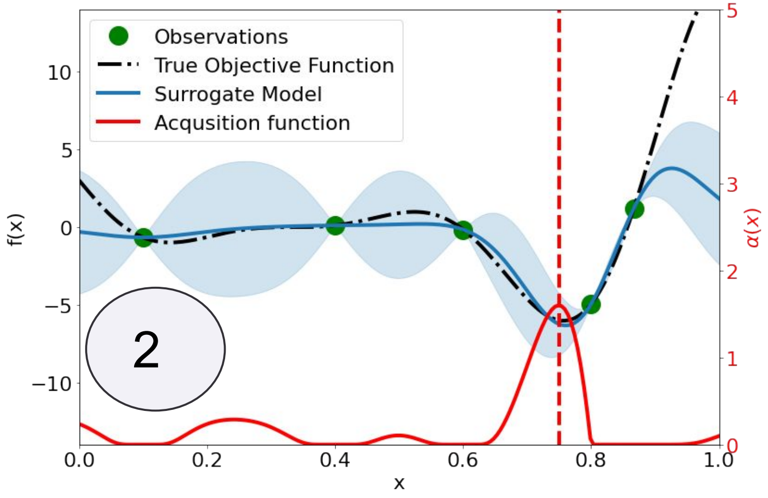


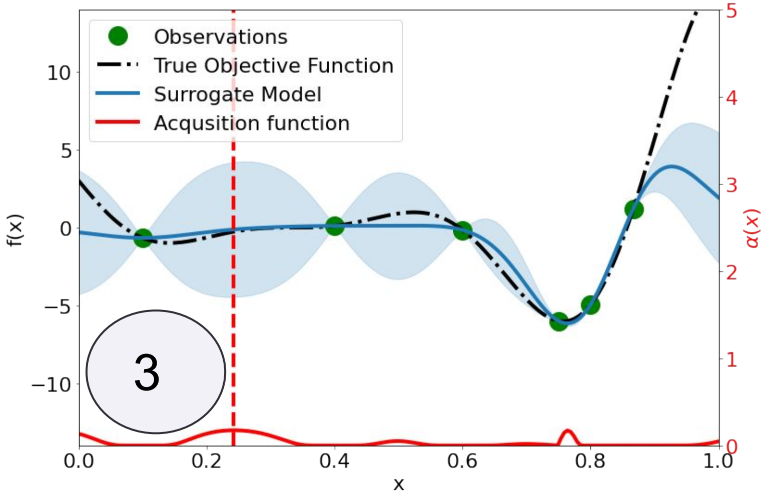
Expected Improvement

Demo BO loop



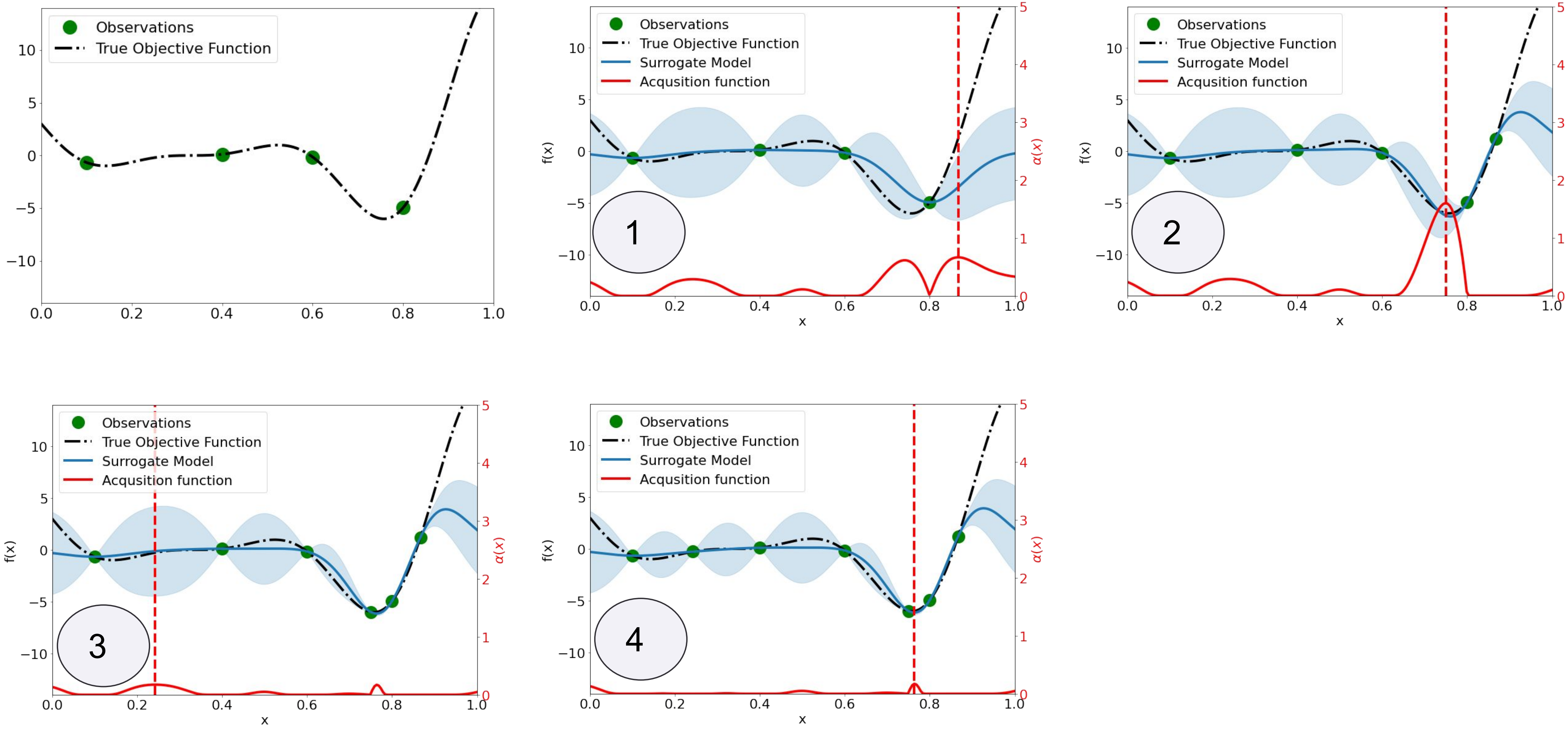






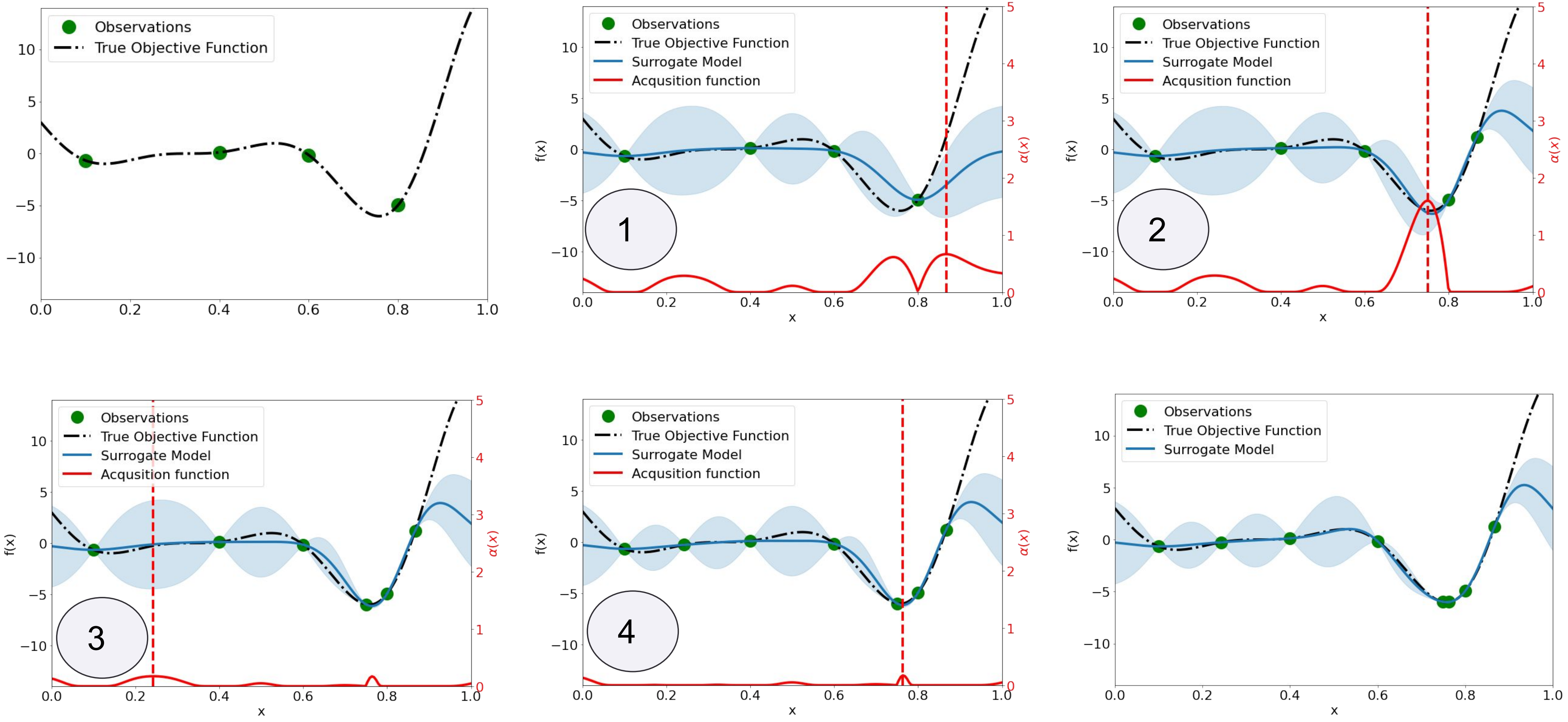
Expected Improvement

Demo BO loop



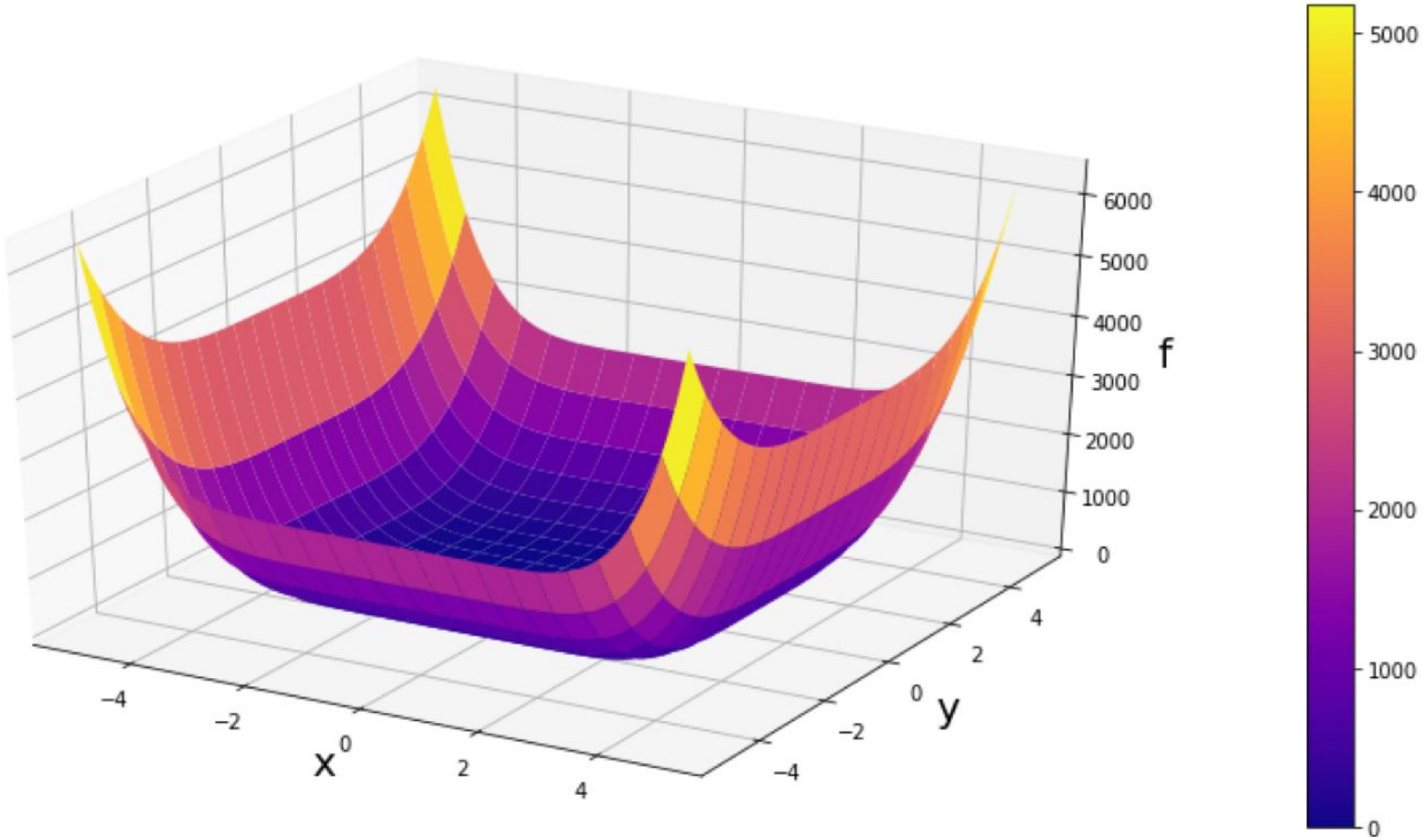
Expected Improvement

Demo BO loop



BO Demo 2

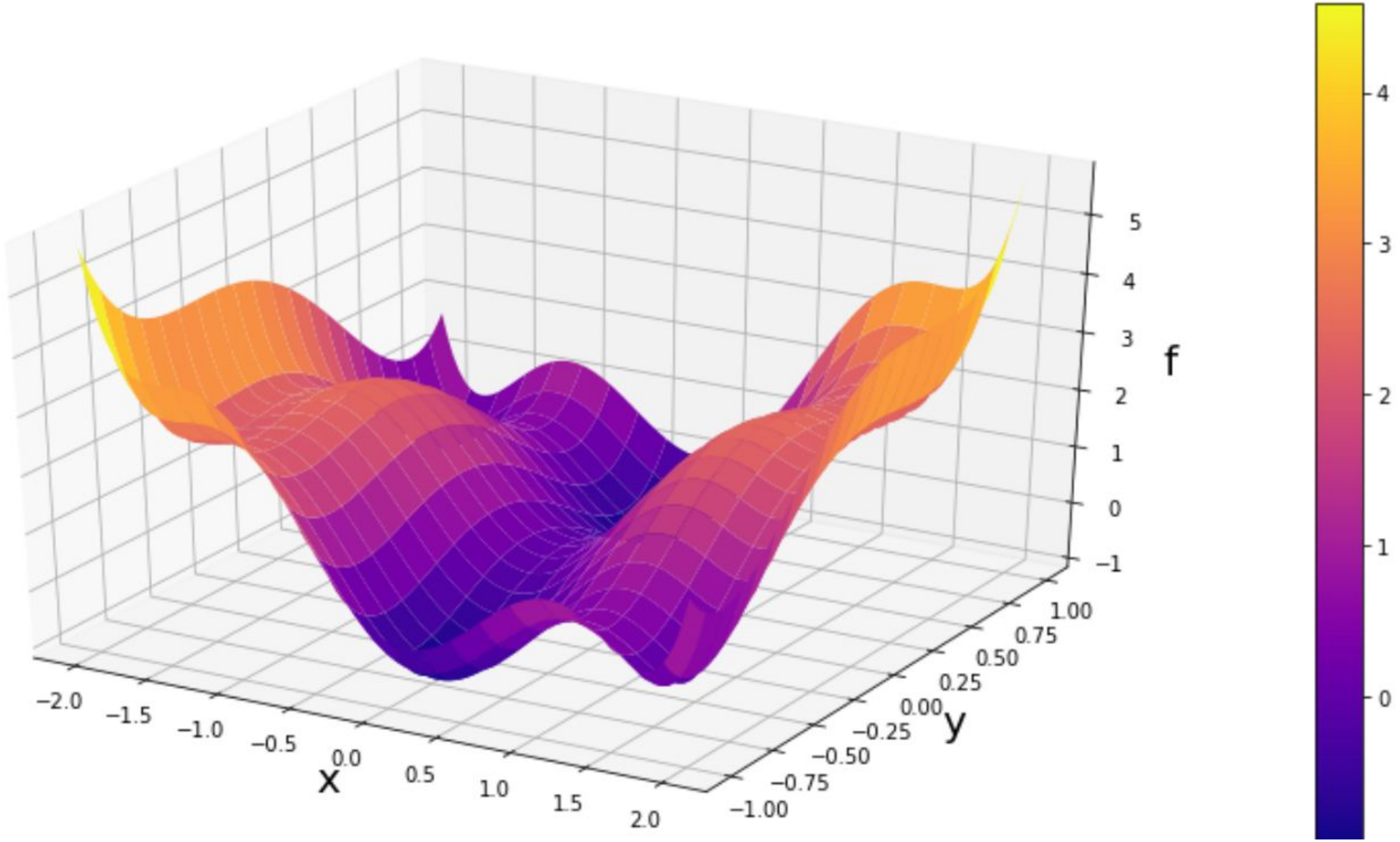
Let minimize the 6 Hump Camel function



Looks like we **can** use a local optimizer!

BO Demo 2

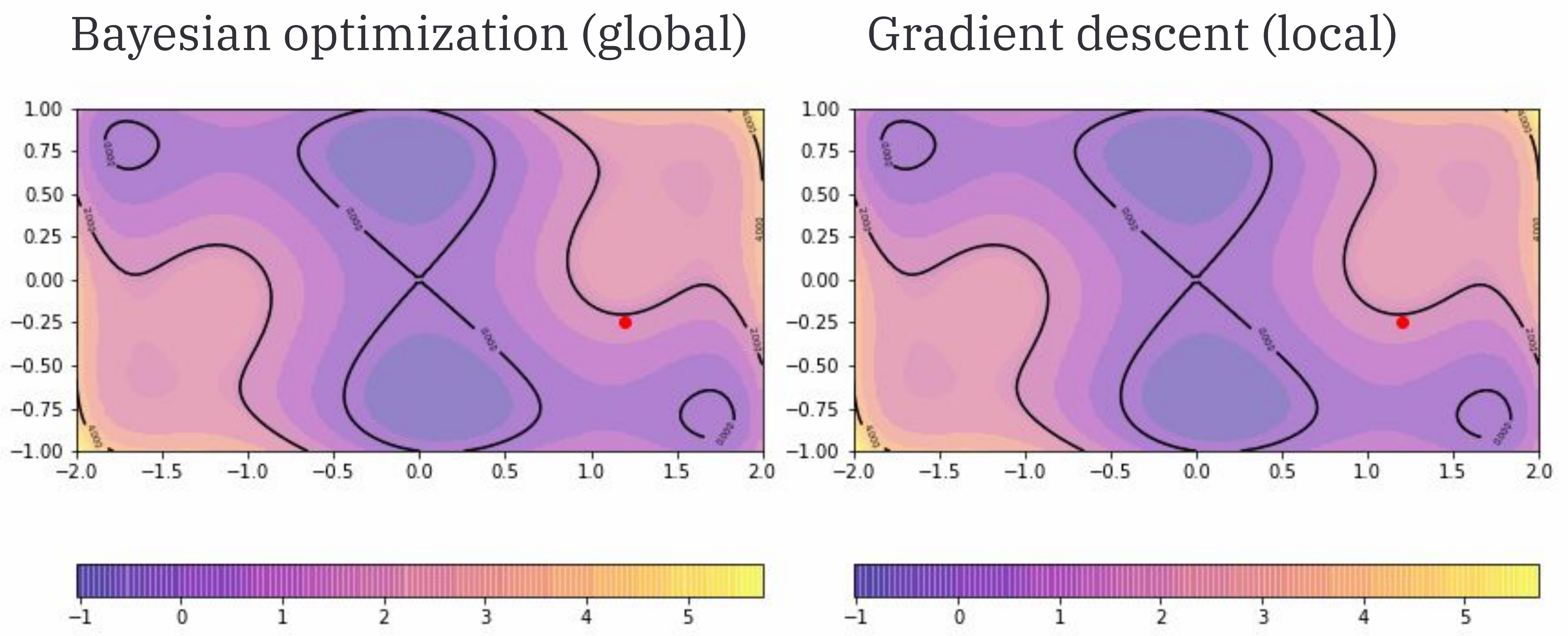
Zoom in: Perhaps not quite as easy?



Looks like we **cannot** use a local optimizer!

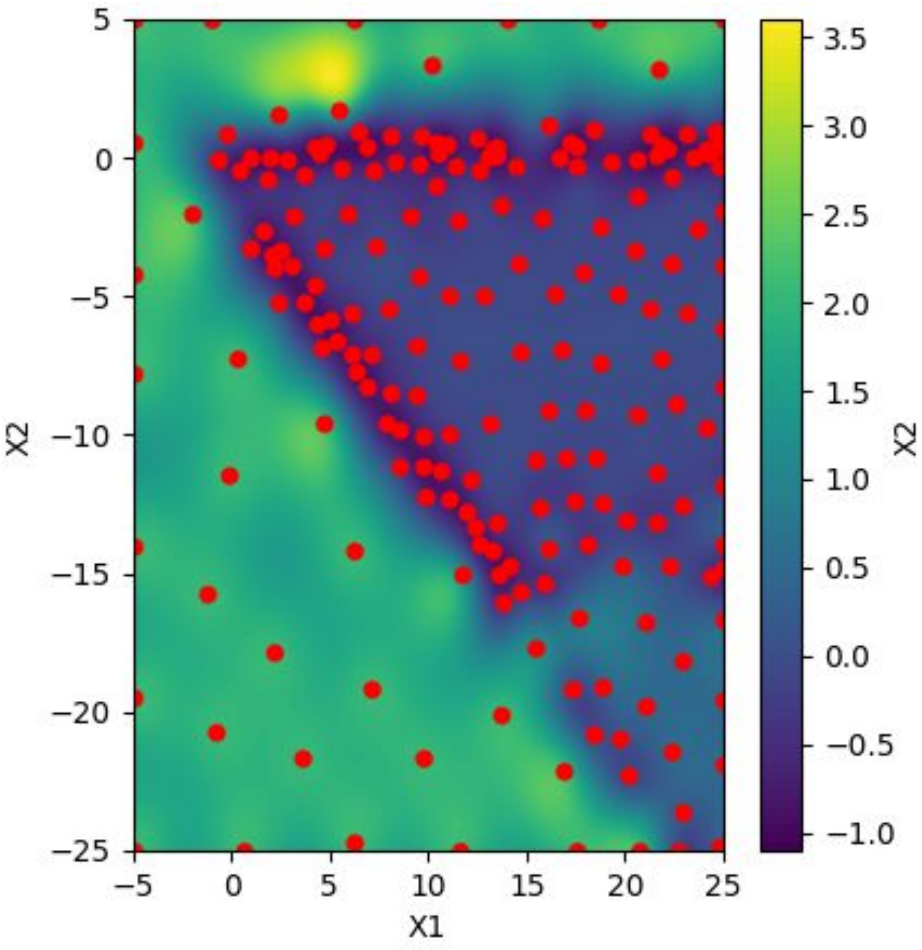
BO Demo 2

Bayesian optimization is a global optimizer



BO Demo 3

Efficient coverage of the search space



So why do we care about Bayesian Optimization?

So why do we care about Bayesian Optimization?

- BO performs **global** optimization (good for multi-modal functions)

So why do we care about Bayesian Optimization?

- BO performs **global** optimization (good for multi-modal functions)
- BO can optimize under a **limited evaluation budget** (great for problems with high evaluation costs)

So why do we care about Bayesian Optimization?

- BO performs **global** optimization (good for multi-modal functions)
- BO can optimize under a **limited evaluation budget** (great for problems with high evaluation costs)
    - Simulating performance of a car engine (mins)
    - Training a large ML model (hours)
    - Synthesising a new molecule (weeks)
    - Testing performance of a wind turbine in real world (months

Increasing cost

So why do we care about Bayesian Optimization?

- BO performs **global** optimization (good for multi-modal functions)
- BO can optimize under a **limited evaluation budget** (great for problems with high evaluation costs)
  - Simulating performance of a car engine (mins)
  - Training a large ML model (hours)
  - Synthesising a new molecule (weeks)
  - Testing performance of a wind turbine in real world (months

Increasing cost

- We do not need gradients or noiseless observations (i.e. **black-box** optimization)

So why do we care about Bayesian Optimization?

- BO performs **global** optimization (good for multi-modal functions)
- BO can optimize under a **limited evaluation budget** (great for problems with high evaluation costs)
  - Simulating performance of a car engine (mins)
  - Training a large ML model (hours)
  - Synthesising a new molecule (weeks)
  - Testing performance of a wind turbine in real world (months

Increasing cost

- We do not need gradients or noiseless observations (i.e. **black-box** optimization)

BO: clever modelling rather than brute force!

Cool things that you can do with BO

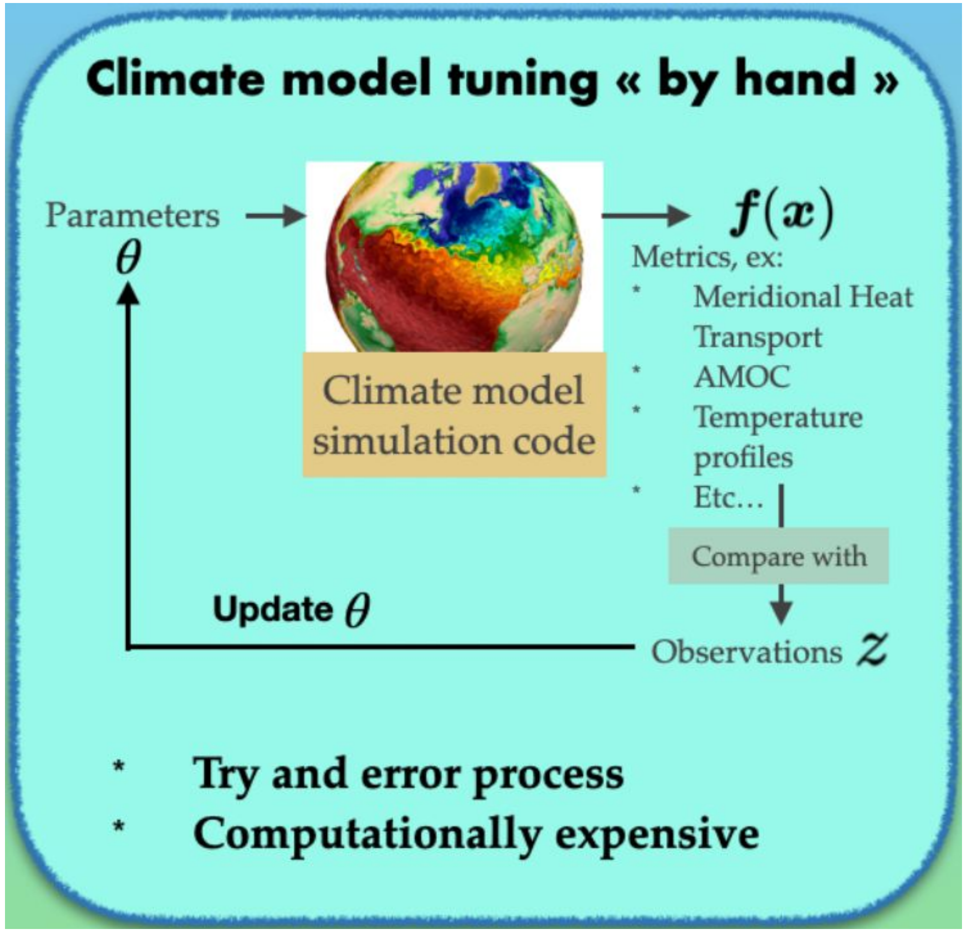
- Fine-tune the performance of AlphaGO (https://arxiv.org/abs/1812.06855)
- Allow Amazon Alexa learn how to speak with new voices (https://arxiv.org/abs/2002.01953)
- Efficiently find new molecules / genes (https://arxiv.org/abs/2010.00979)
- Fine-tune electric car engines
- Optimize large climate models

A great new reference for BO: **https://bayesoptbook.com/**

So, Climate model calibration?

Climate model calibration

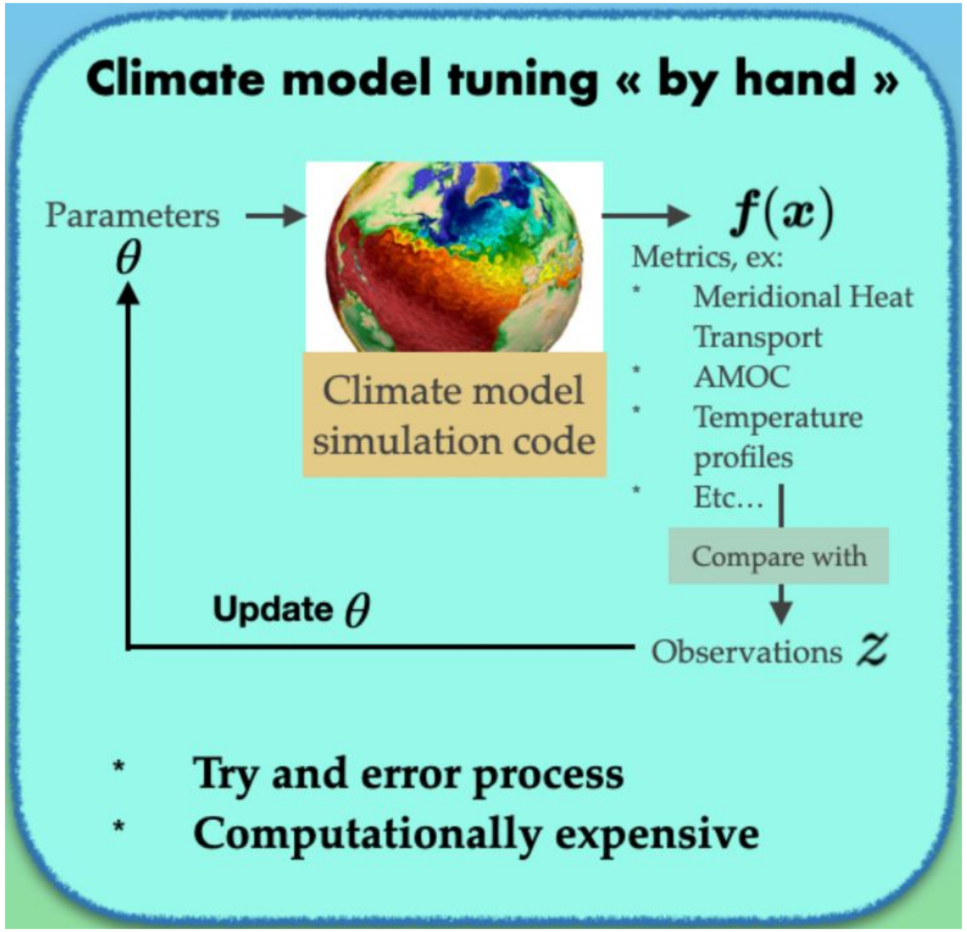
Identifying reasonable values for model parameters



Lguensat et al. 2022.

Climate model calibration

Identifying reasonable values for model parameters

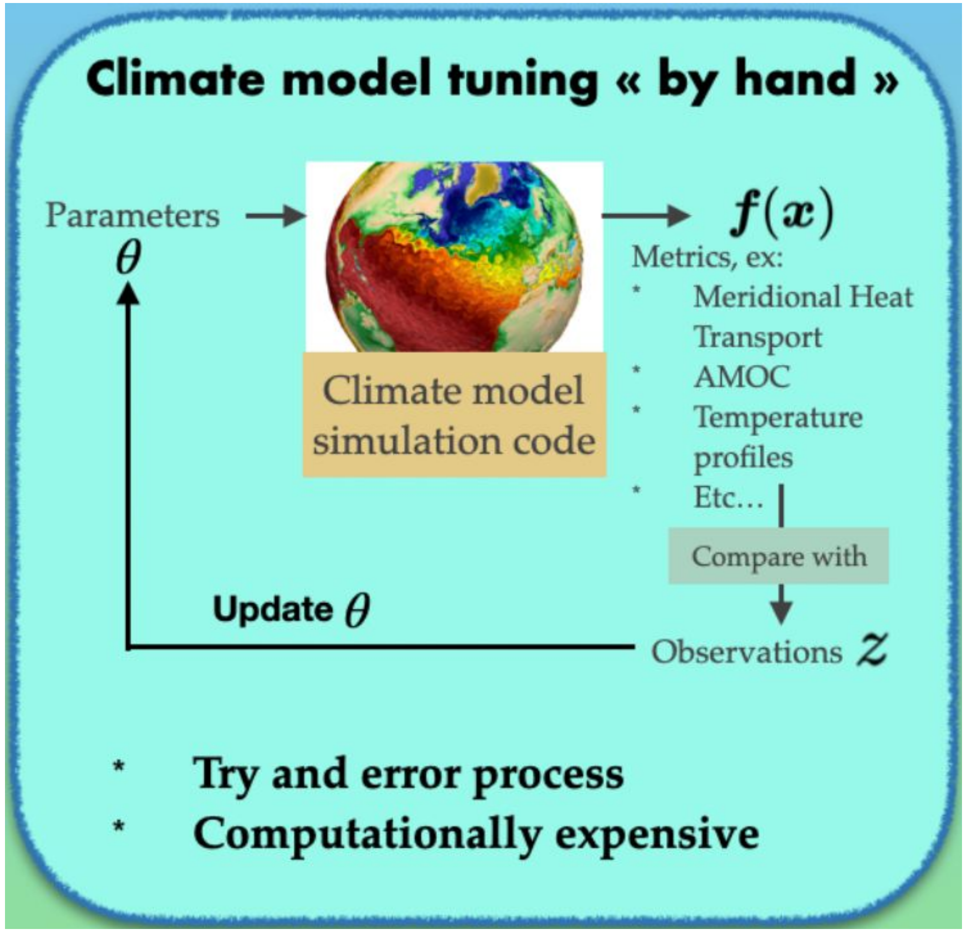


Lguensat et al. 2022.

- Need to find parameters that give high plausibility to historical data —----------> a **function maximisation** problem

Climate model calibration

Identifying reasonable values for model parameters

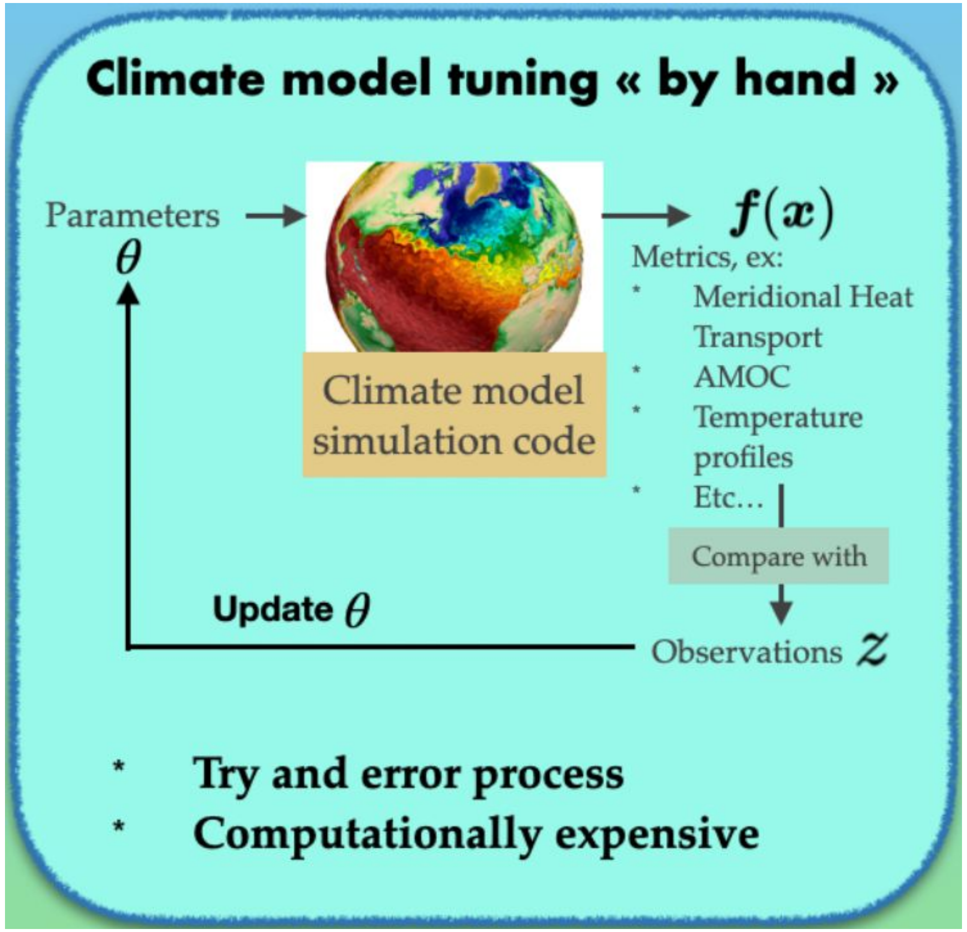


Lguensat et al. 2022.

- Need to find parameters that give high plausibility to historical data —----------> a **function maximisation** problem
- Climate models are expensive —----------------> can only afford a **limited number of evaluations** (no grid!)

Climate model calibration

Identifying reasonable values for model parameters



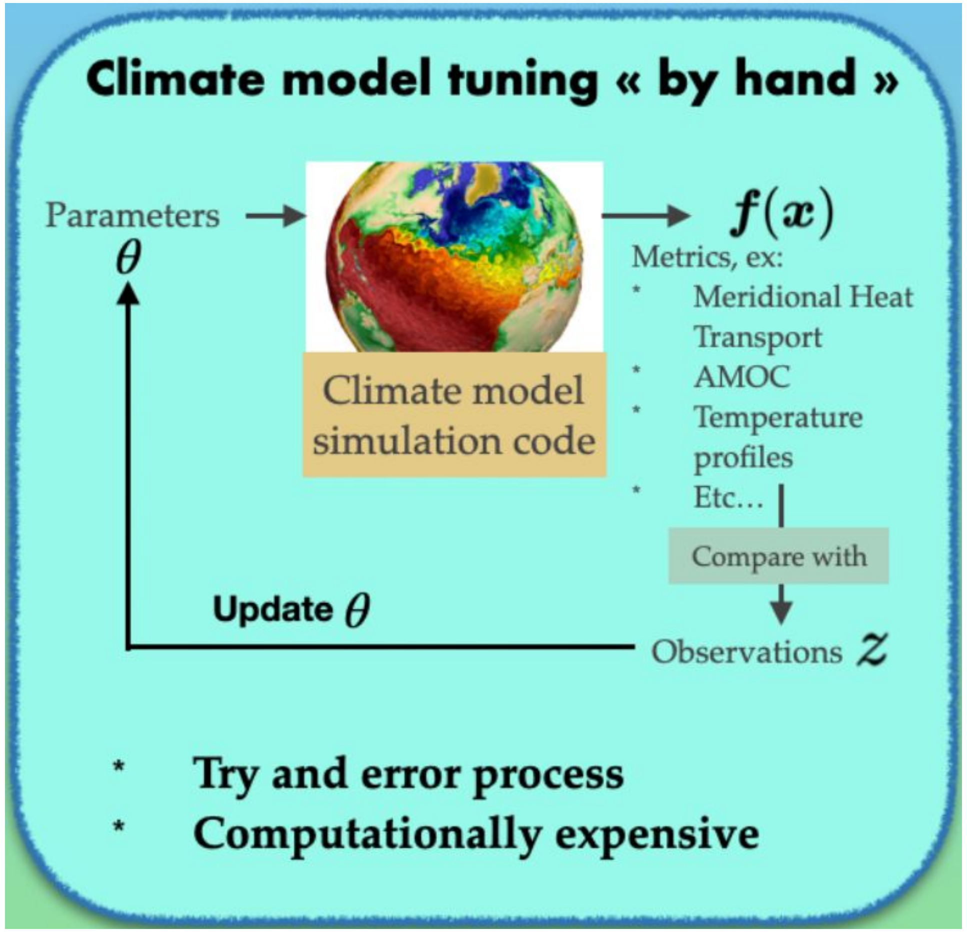
Lguensat et al. 2022.

- Need to find parameters that give high plausibility to historical data —----------> a **function maximisation** problem
- Climate models are expensive —----------------> can only afford a **limited number of evaluations** (no grid!)
- We do not have gradients (easily) and limited prior knowledge —----------------> a **black-box** objective function

Climate model calibration

Identifying reasonable values for model parameters

So we have a resource-constrained black-box function optimisation!



Lguensat et al. 2022.

- Need to find parameters that give high plausibility to historical data —---------> a **function maximisation** problem
- Climate models are expensive —----------------> can only afford a **limited number of evaluations** (no grid!)
- We do not have gradients (easily) and limited prior knowledge —----------------> a **black-box** objective function

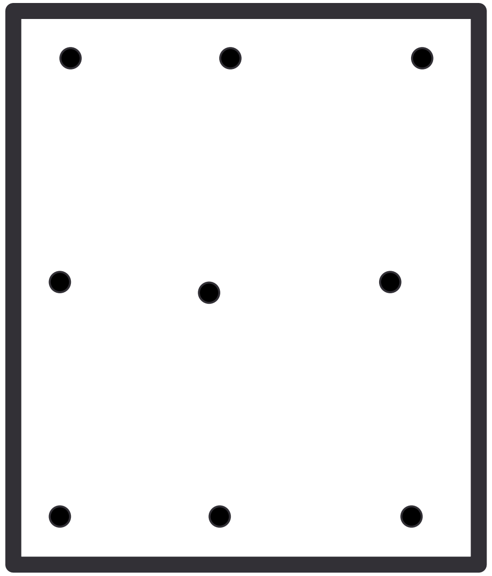
Climate model calibration by iteratively refocusing

sequentially whittle down the plausible region

Based on Lguensat et al. 2022.

Climate model calibration by iteratively refocusing

sequentially whittle down the plausible region

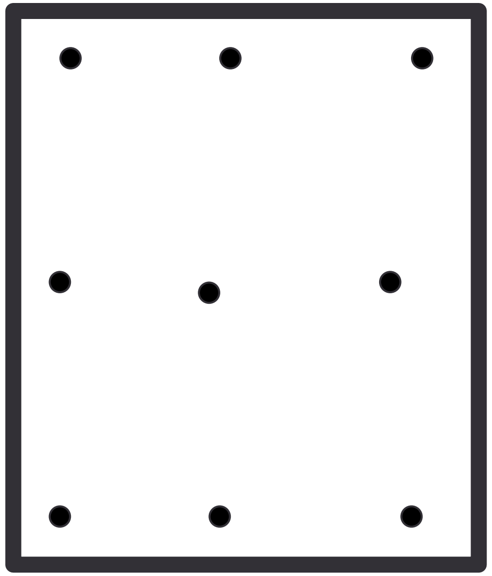


Initial Design

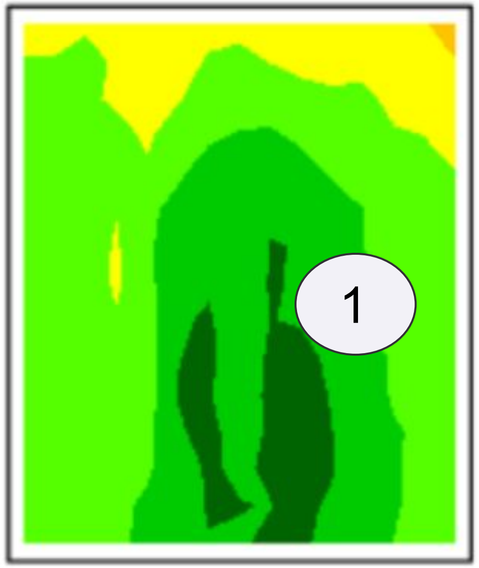
Based on Lguensat et al. 2022.

Climate model calibration by iteratively refocusing

sequentially whittle down the plausible region



Initial Design

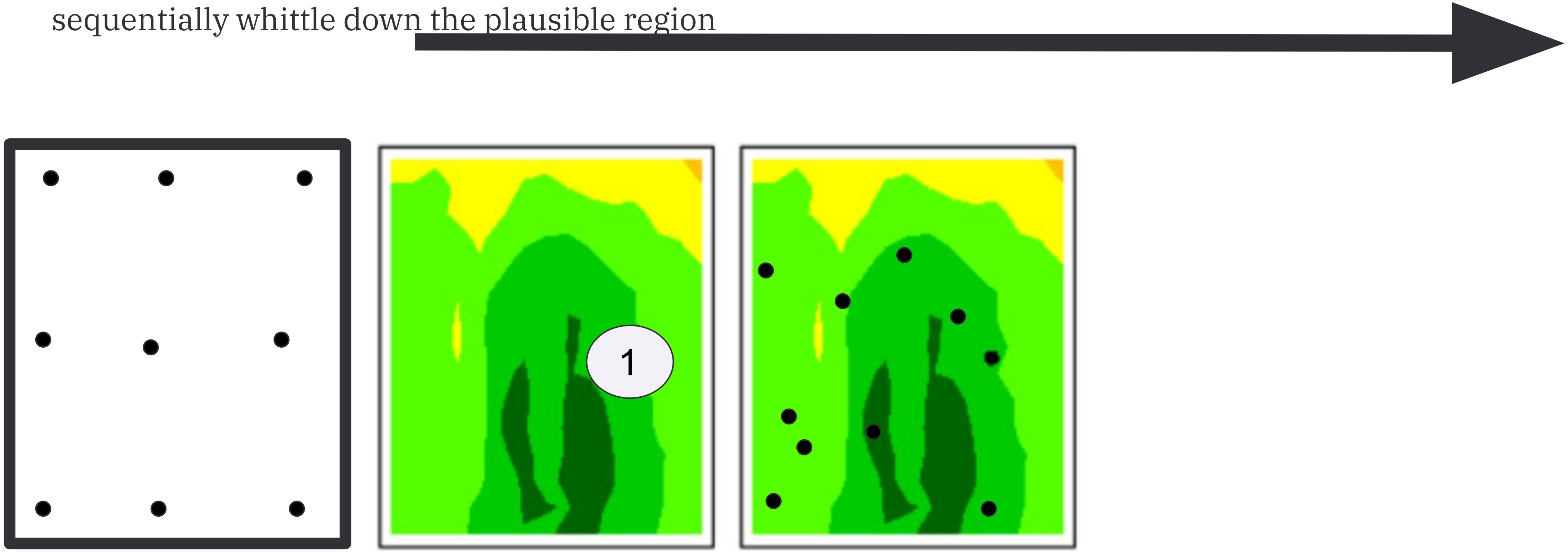


Predicted implausibility

Based on Lguensat et al. 2022.

Climate model calibration by iteratively refocusing

sequentially whittle down the plausible region



Initial Design

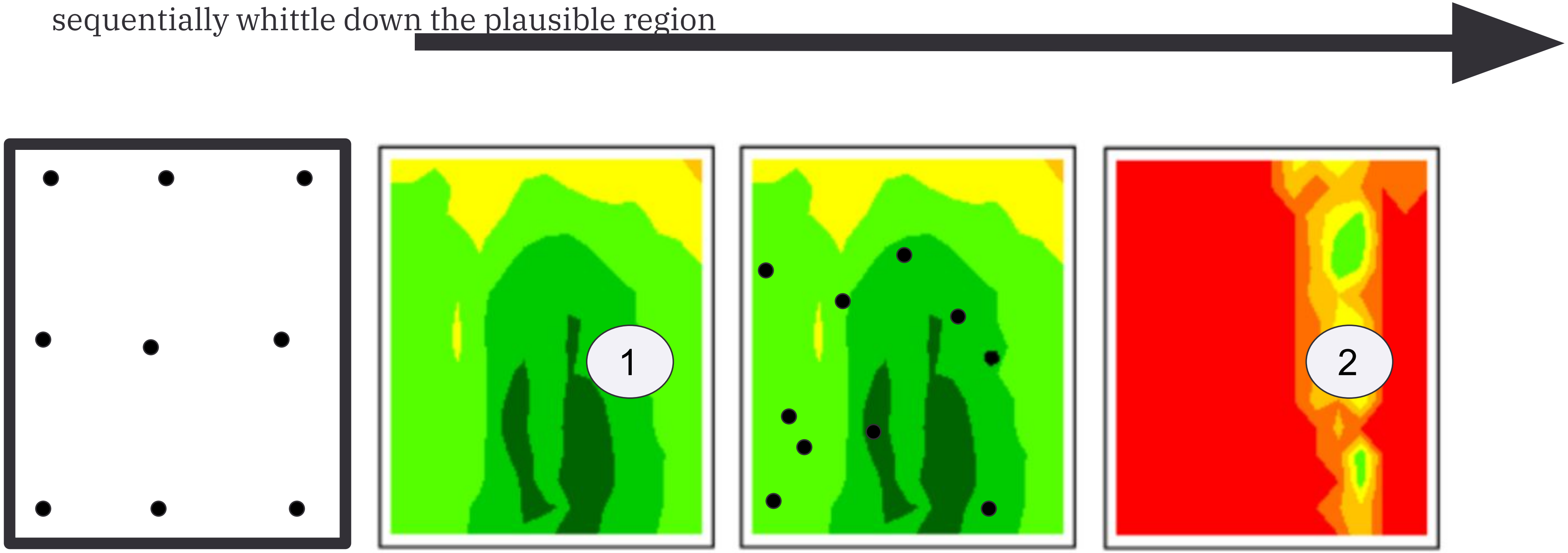
Predicted implausibility

1st set of evaluations

Based on Lguensat et al. 2022.

Climate model calibration by iteratively refocusing

sequentially whittle down the plausible region



Initial Design

Predicted implausibility

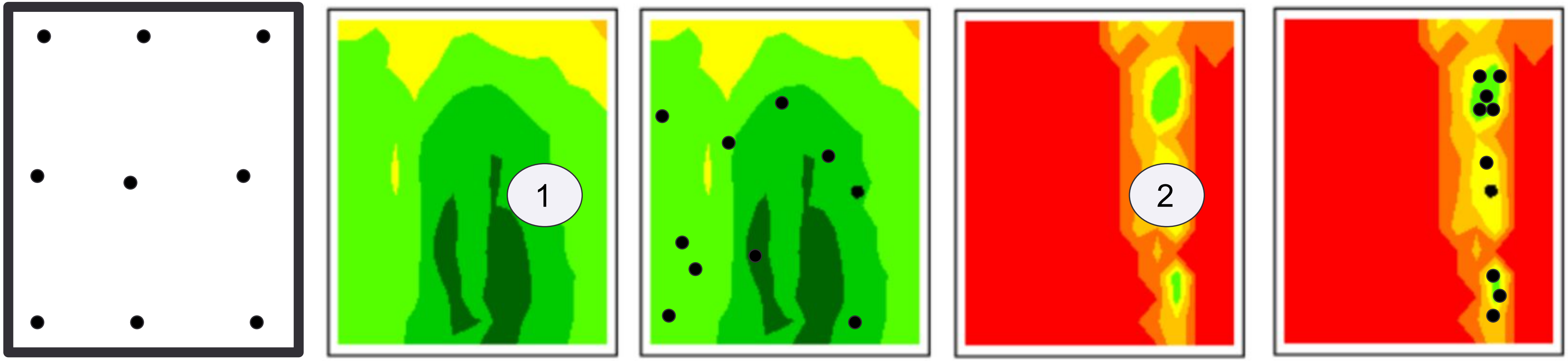
1st set of evaluations

Predicted implausibility

Based on Lguensat et al. 2022.

Climate model calibration by iteratively refocusing

sequentially whittle down the plausible region



Initial Design

Predicted implausibility

1st set of evaluations

Predicted implausibility

2nd set of evaluations

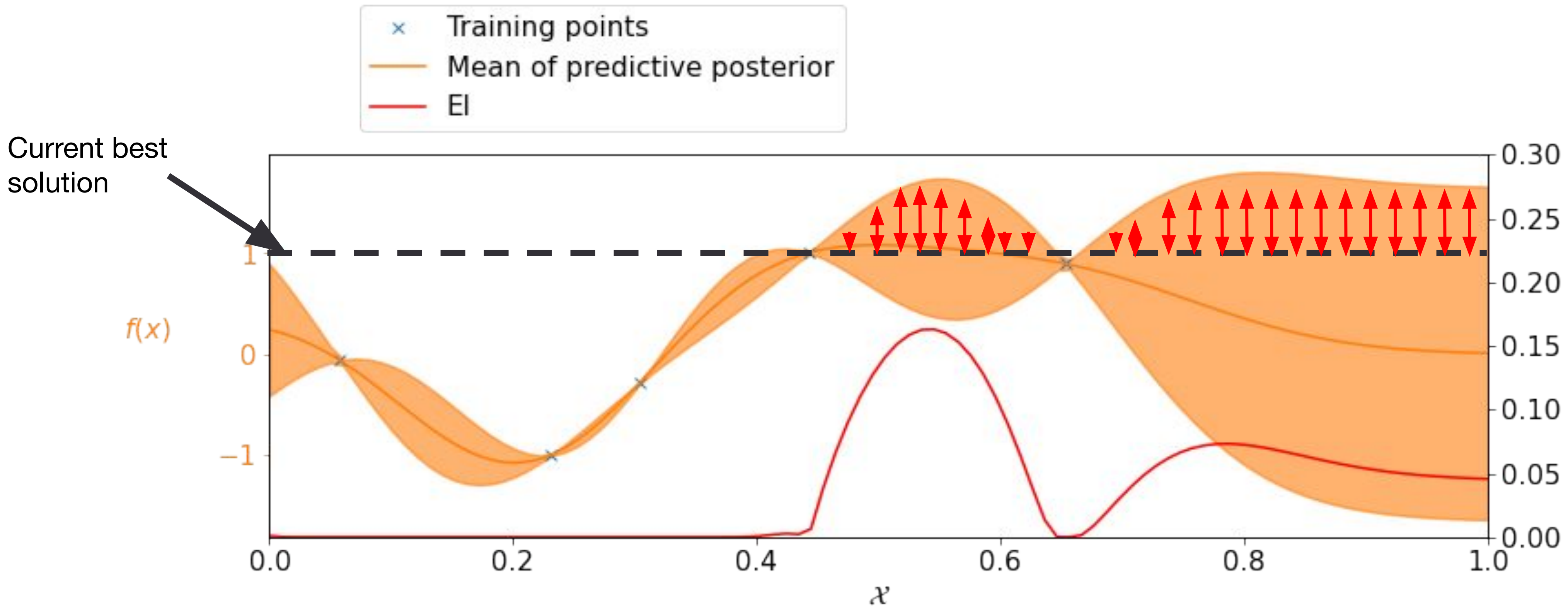
Based on Lguensat et al. 2022.

Back to molecular design

Large batches

Automatically choosing batches of points

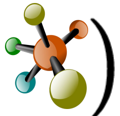
Using GP posteriors and utility functions



How to pick **3** points ?

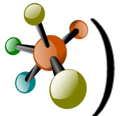
Automatically choosing batches of molecules

Using GP posteriors and utility functions

- $\alpha_{\mathrm{EI}}($  $) = \mathbb{E}_f[\max(f - f^\star, 0)]$ $\qquad f \sim \mathcal{N}(\mu, \sigma^2)$

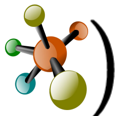
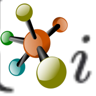
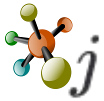
Automatically choosing batches of molecules

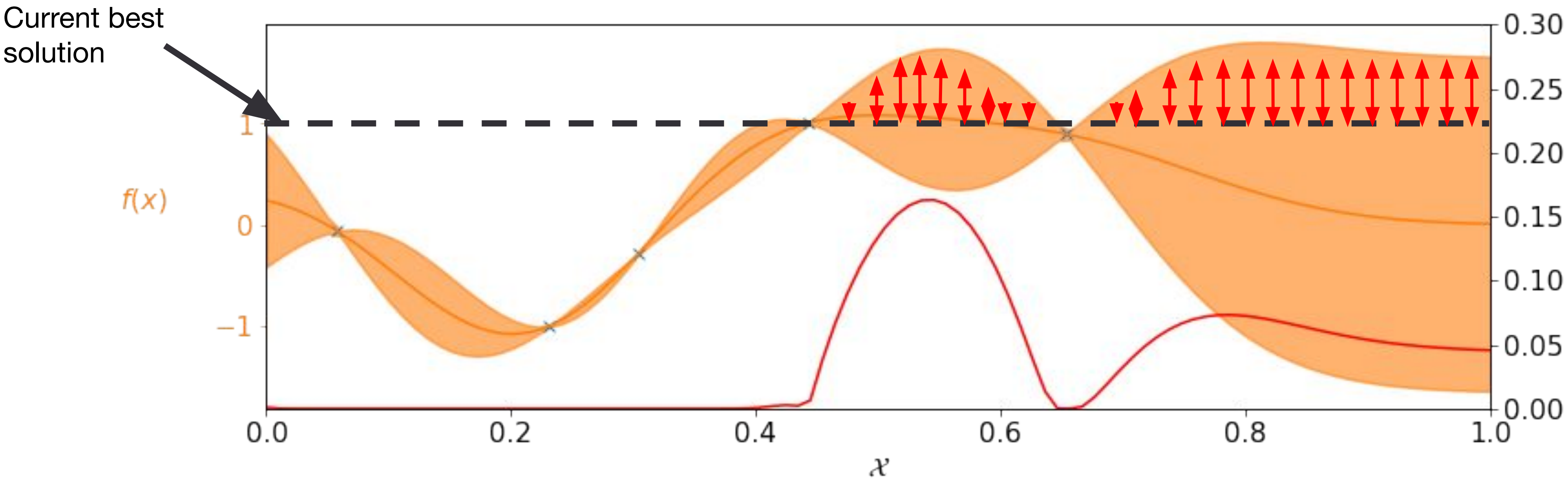
Using GP posteriors and utility functions

- $\alpha_{\mathrm{EI}}($  $) = \mathbb{E}_f[\max(f - f^\star, 0)]$
- $\alpha_{\mathrm{EI}}(\{$  $_i,$  $_j\}) = ???$

Automatically choosing batches of molecules

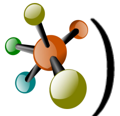
Using GP posteriors and utility functions

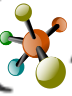
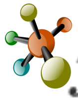
- $\alpha_{\mathrm{EI}}($  $) = \mathbb{E}_f[\max(f - f^\star, 0)]$
- $\alpha_{\mathrm{EI}}(\{$  $_i,$  $_j\}) = \mathbb{E}_{f_i,\, f_j}[\max(f_i - f^\star, f_j - f^\star, 0)]$



Automatically choosing batches of molecules

Using GP posteriors and utility functions

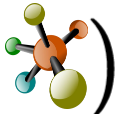
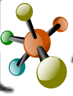
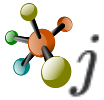
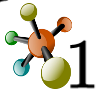
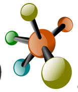
- $\alpha_{\mathrm{EI}}($  $) = \mathbb{E}_f[\max(f - f^\star, 0)]$

- $\alpha_{\mathrm{EI}}(\{$  $_i,$  $_j\}) = \mathbb{E}_{f_i,\, f_j}[\max(f_i - f^\star, f_j - f^\star, 0)]$

$$\begin{pmatrix} f_i \\ f_j \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_i \\ \mu_j \end{pmatrix}, \begin{pmatrix} \Sigma_{i,i} & \Sigma_{i,j} \\ \Sigma_{j,i} & \Sigma_{j,j} \end{pmatrix} \right)$$

Automatically choosing batches of molecules

Using GP posteriors and utility functions

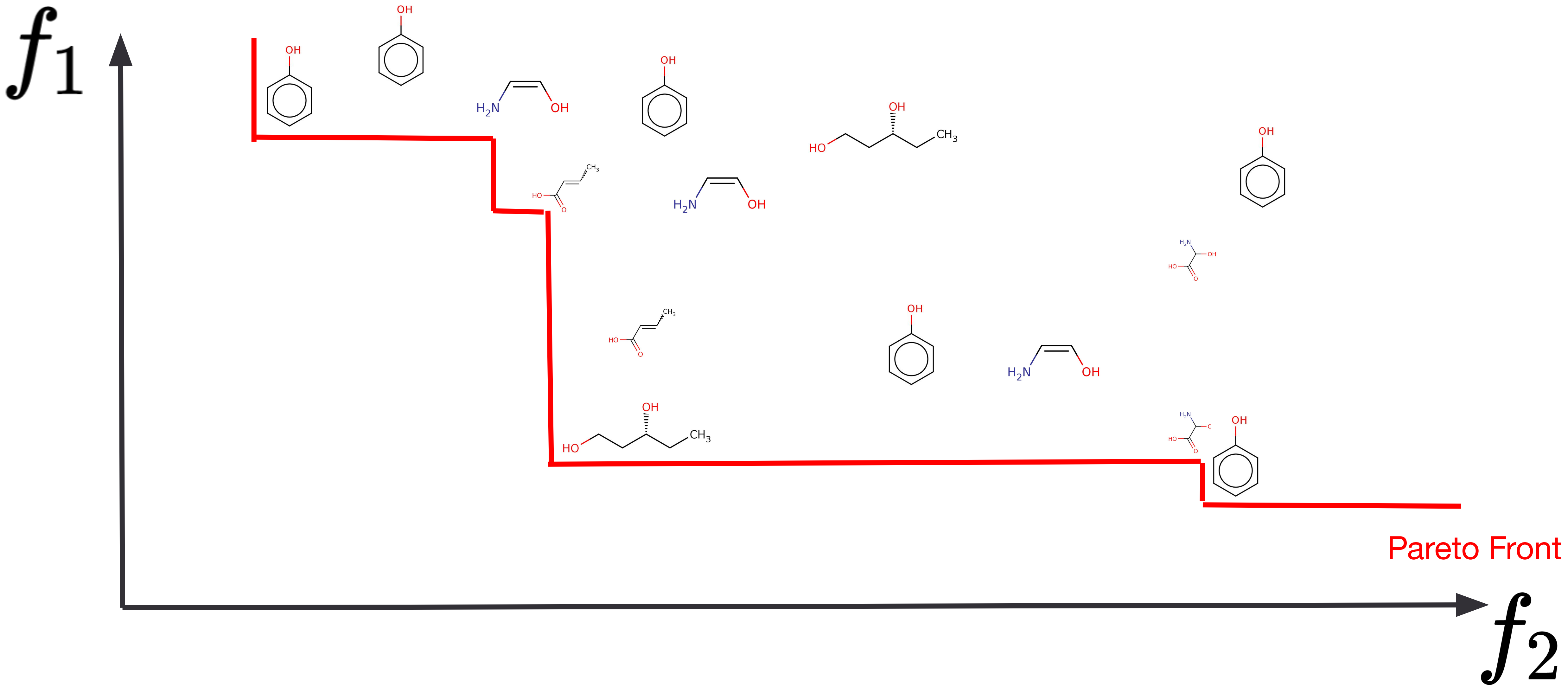
- $\alpha_{\mathrm{EI}}($$) = \mathbb{E}_f[\max(f - f^\star, 0)]$
- $\alpha_{\mathrm{EI}}(\{$$_i,$$_j\}) = \mathbb{E}_{f_i, f_j}[\max(f_i - f^\star, f_j - f^\star, 0)]$
- $\alpha_{\mathrm{EI}}(\{$$_1, \ldots,$$_B\}) = ???$

Back to molecular design

Multiple objectives

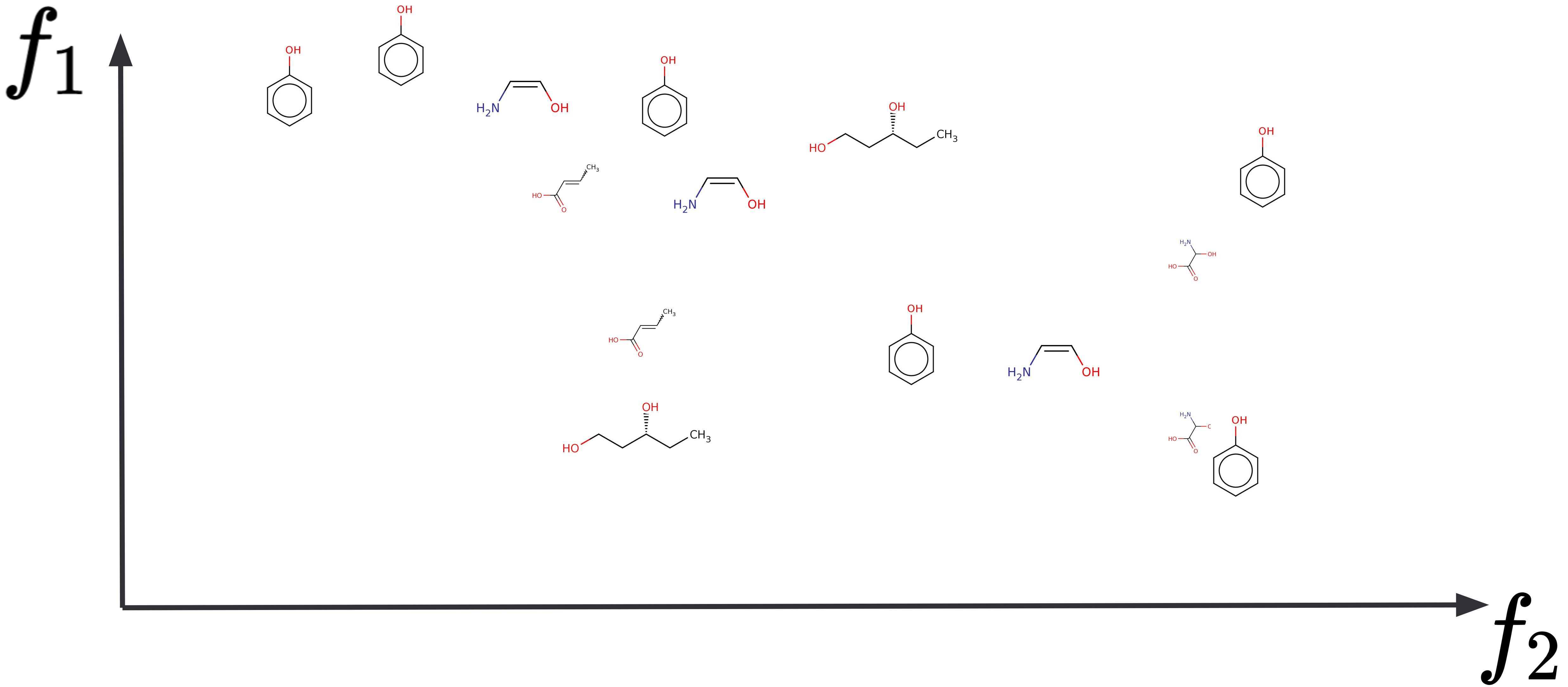
Multi-objective Optimisation

>1 competing objectives

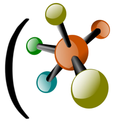
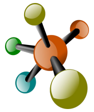


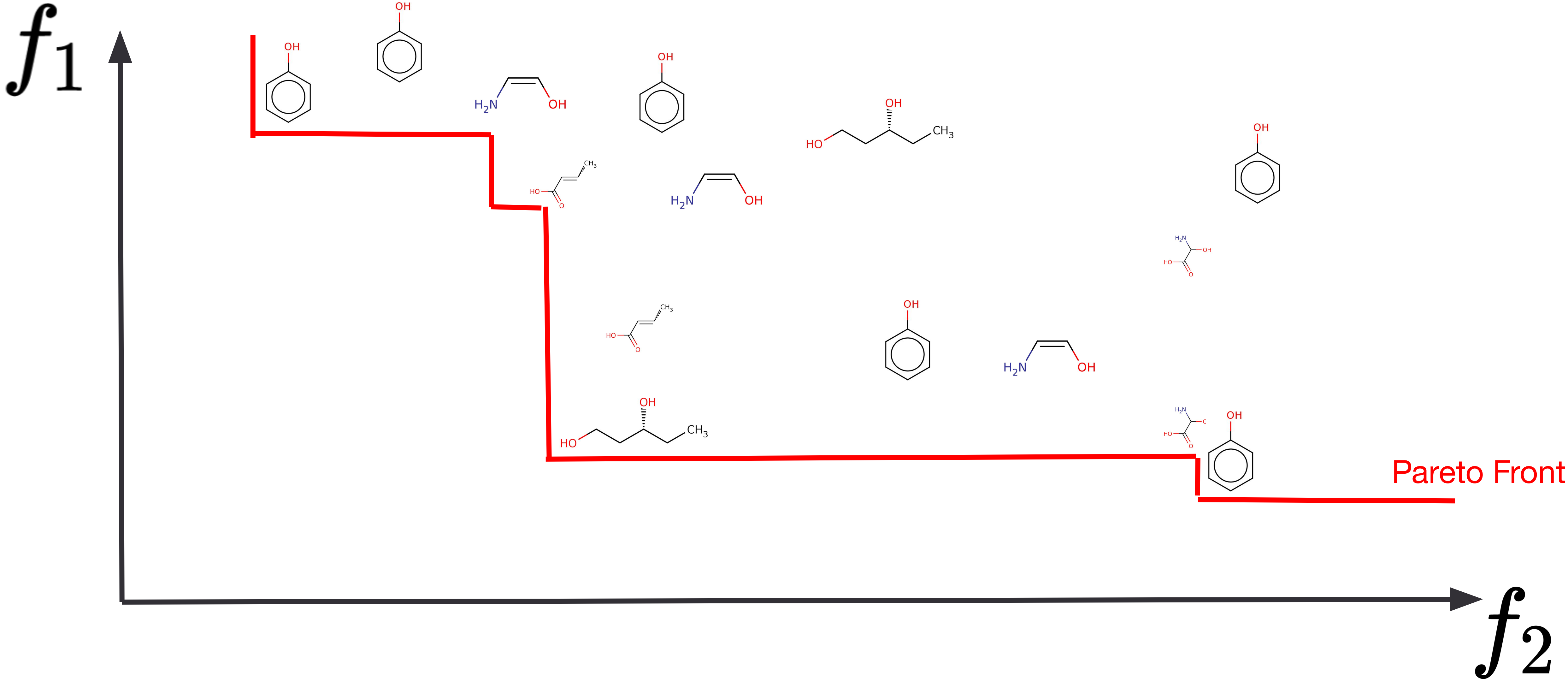
Multi-objective Optimisation

>1 competing objectives

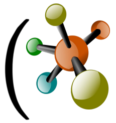
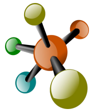


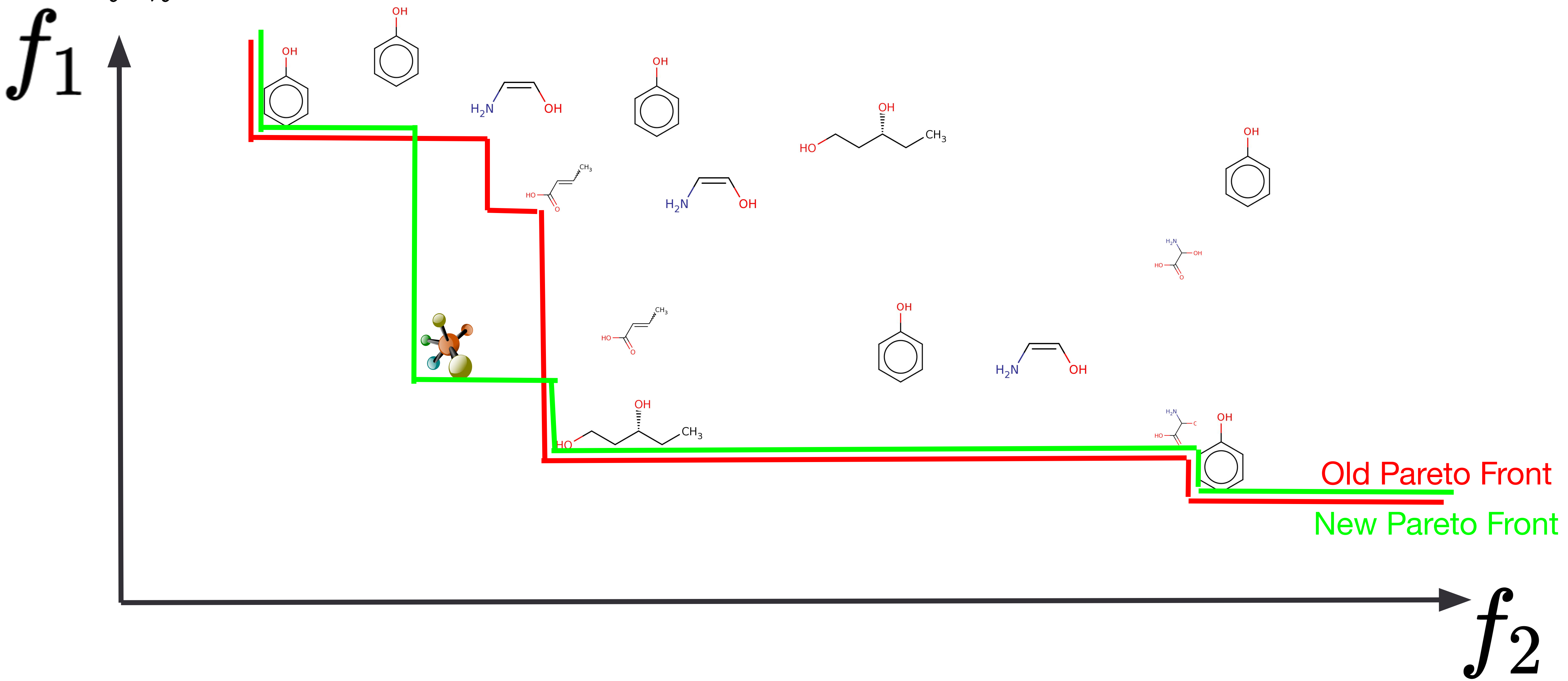
Multi-objective Optimisation

U_{f_1,f_2}(): what is the utility of evaluating  if it will return (f_1, f_2)



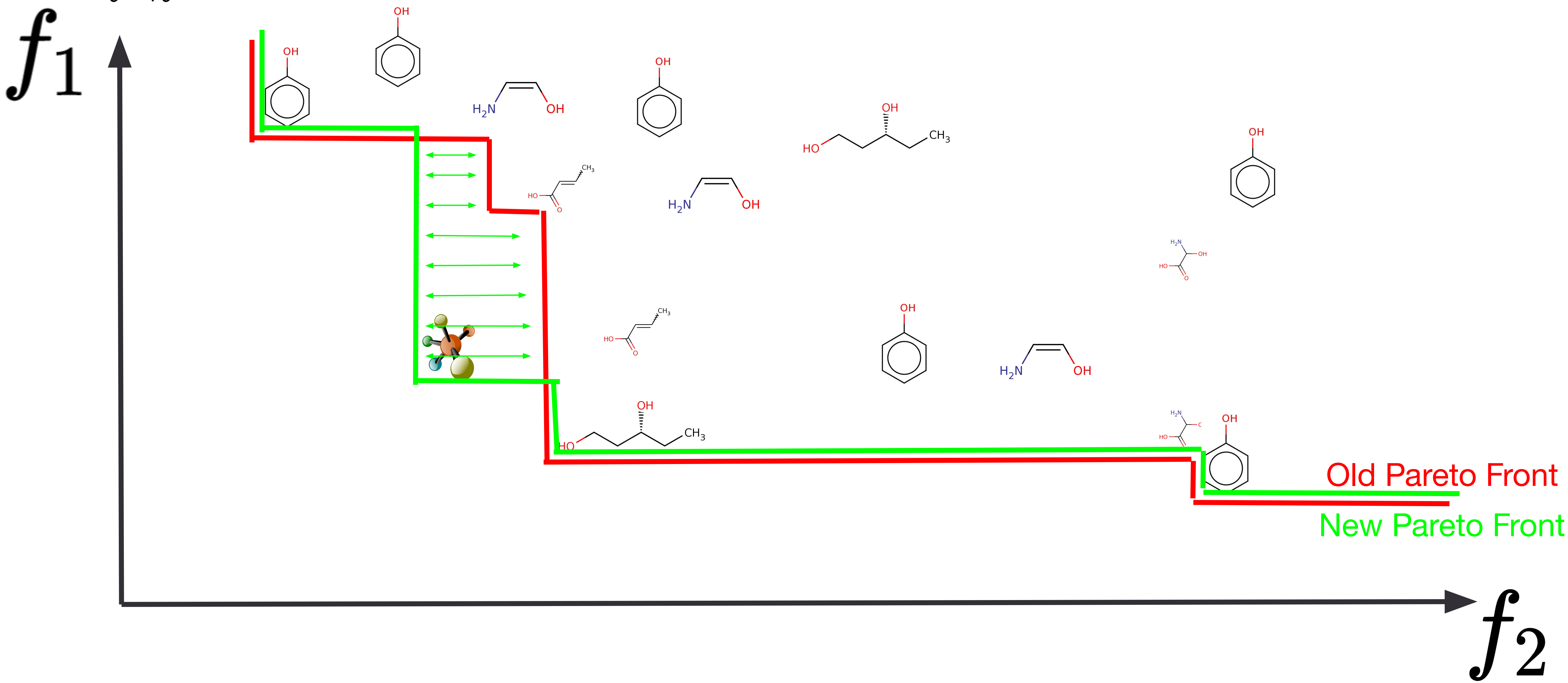
Multi-objective Optimisation

U_{f_1,f_2}() : what is the utility of evaluating  if it will return (f_1, f_2)

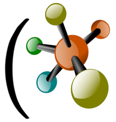
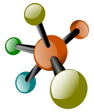


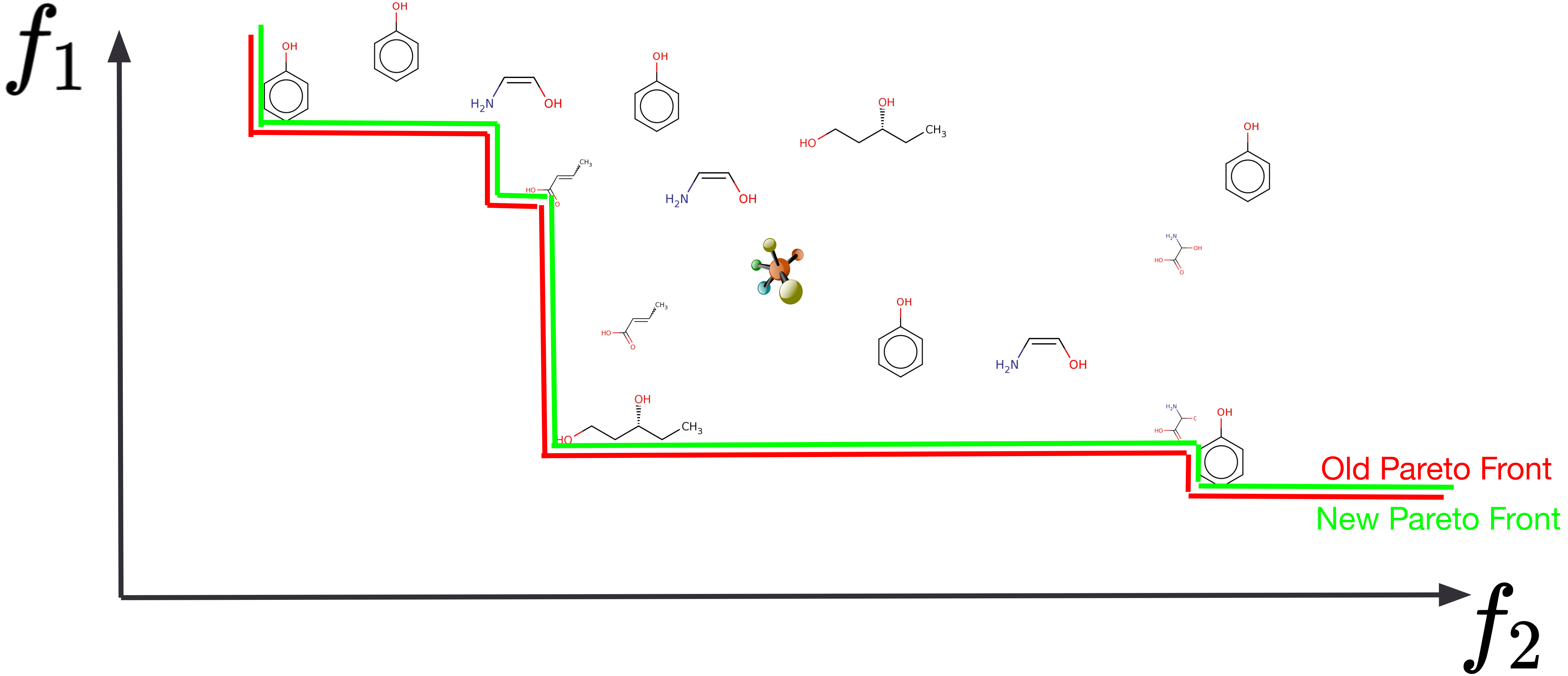
Multi-objective Optimisation

U_{f_1,f_2}( ) : what is the utility of evaluating if it will return (f_1, f_2)



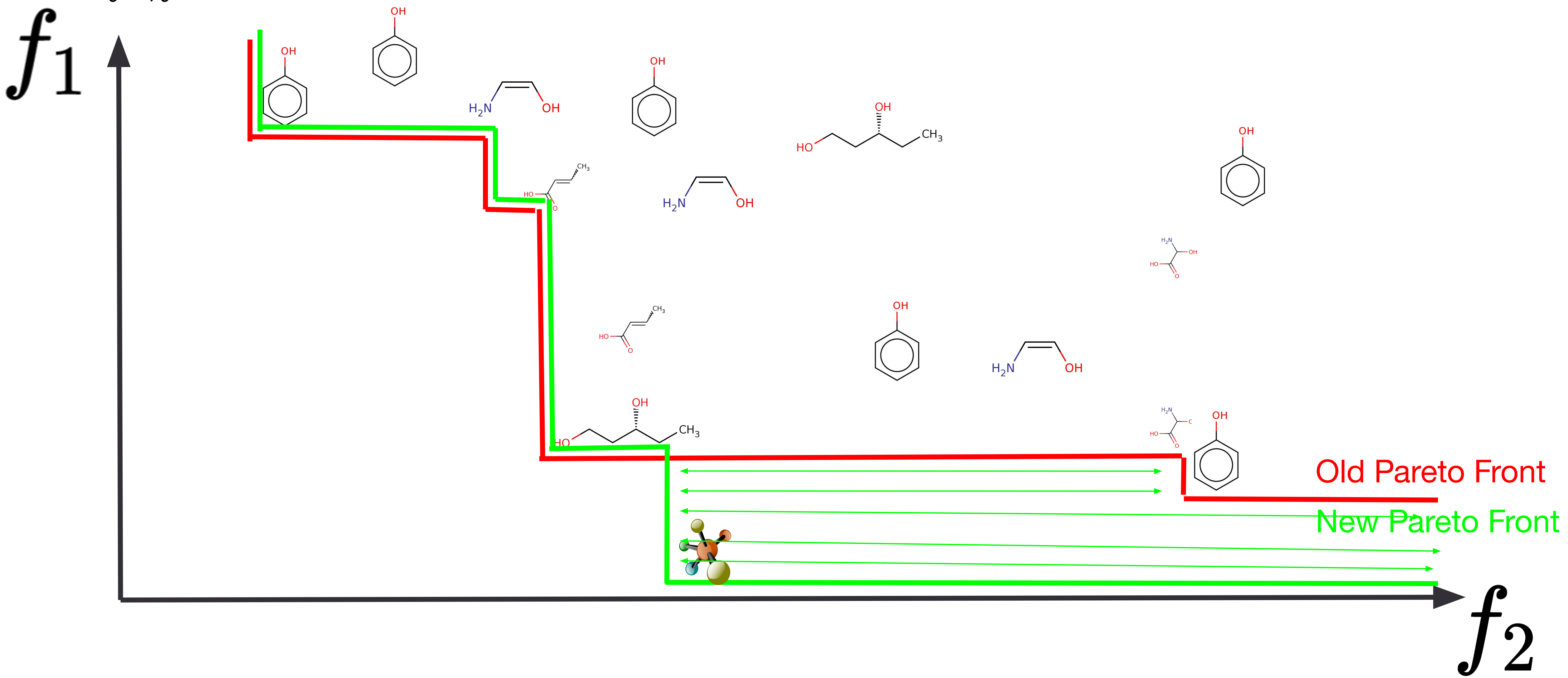
Multi-objective Optimisation

U_{f_1,f_2}(): what is the utility of evaluating  if it will return (f_1, f_2)

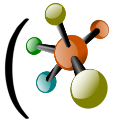
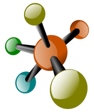


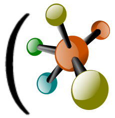
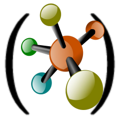
Multi-objective Optimisation

U_{f_1,f_2}( ): what is the utility of evaluating if it will return (f_1, f_2)



Multi-objective Optimisation

$U_{f_1,f_2}($  $)$: what is the utility of evaluating  if it will return (f_1, f_2)

- Use expected hyper-volume improvement $\alpha_{\mathrm{EHVI}}($  $) = \mathbb{E}_{f_1,f_2}(U_{f_1,f_2}($  $))$

$$f_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$$
$$f_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

Multi-objective Optimisation

$U_{f_1,f_2}(\;)$: what is the utility of evaluating if it will return (f_1, f_2)

- Use expected hyper-volume improvement $\alpha_{\text{EHVI}}(\;) = \mathbb{E}_{f_1,f_2}(U_{f_1,f_2}(\;))$

$$f_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$$
$$f_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

$$\alpha_{\text{EHVI}}(\{\;_i, \;_j\}) = ???$$

A more sophisticated acquisition function?

Entropy Search

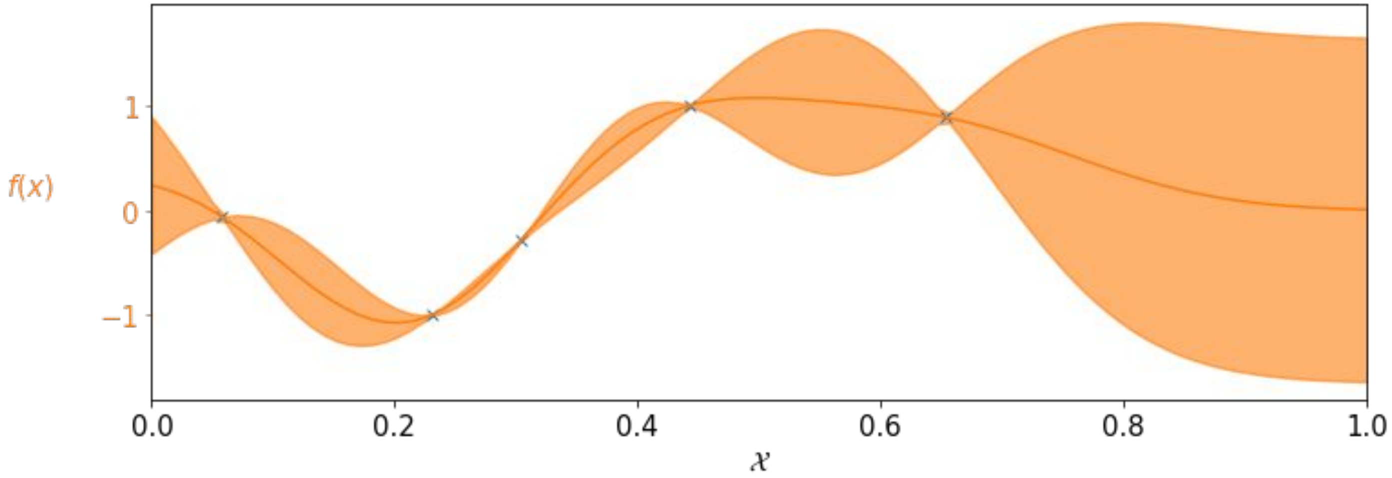
Quick Recap

$$\mathbf{x}^* = \arg\max_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$$

Quick Recap

$$\mathbf{x}^* = \arg\max_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$$

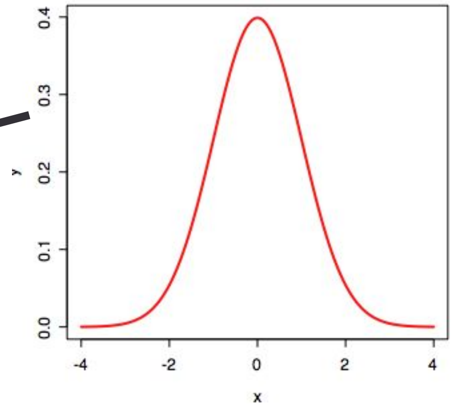
$$f(\mathbf{x}) | D \sim \mathcal{GP}$$



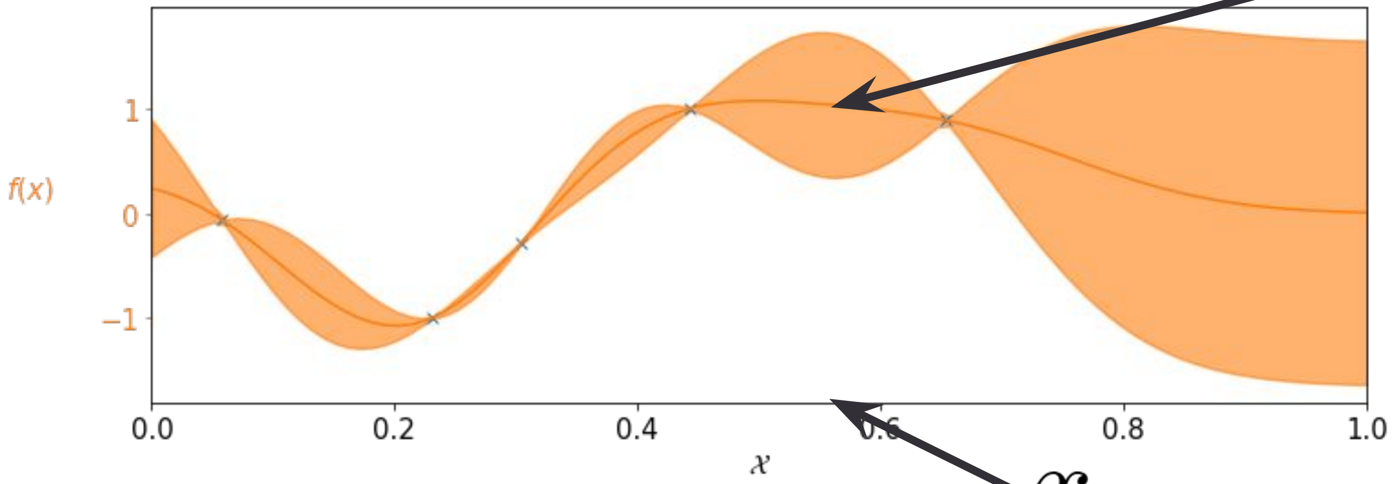
Quick Recap

$$\mathbf{x}^* = \arg\max_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x})$$

$$f(\mathbf{x}) | D \sim \mathcal{GP}$$



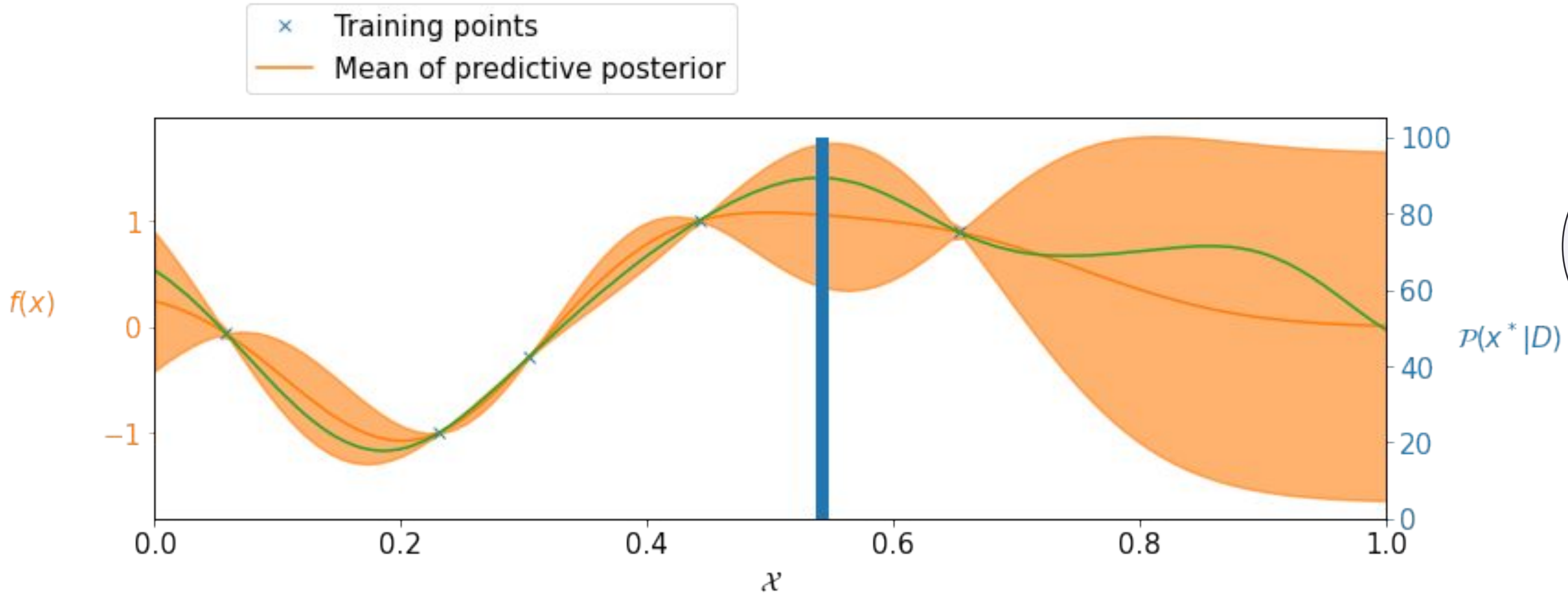
$$f(x) \sim N(\mu(x), \sigma^2(x))$$



$$x$$

What is our best guess for $\mathbf{x}^*$?

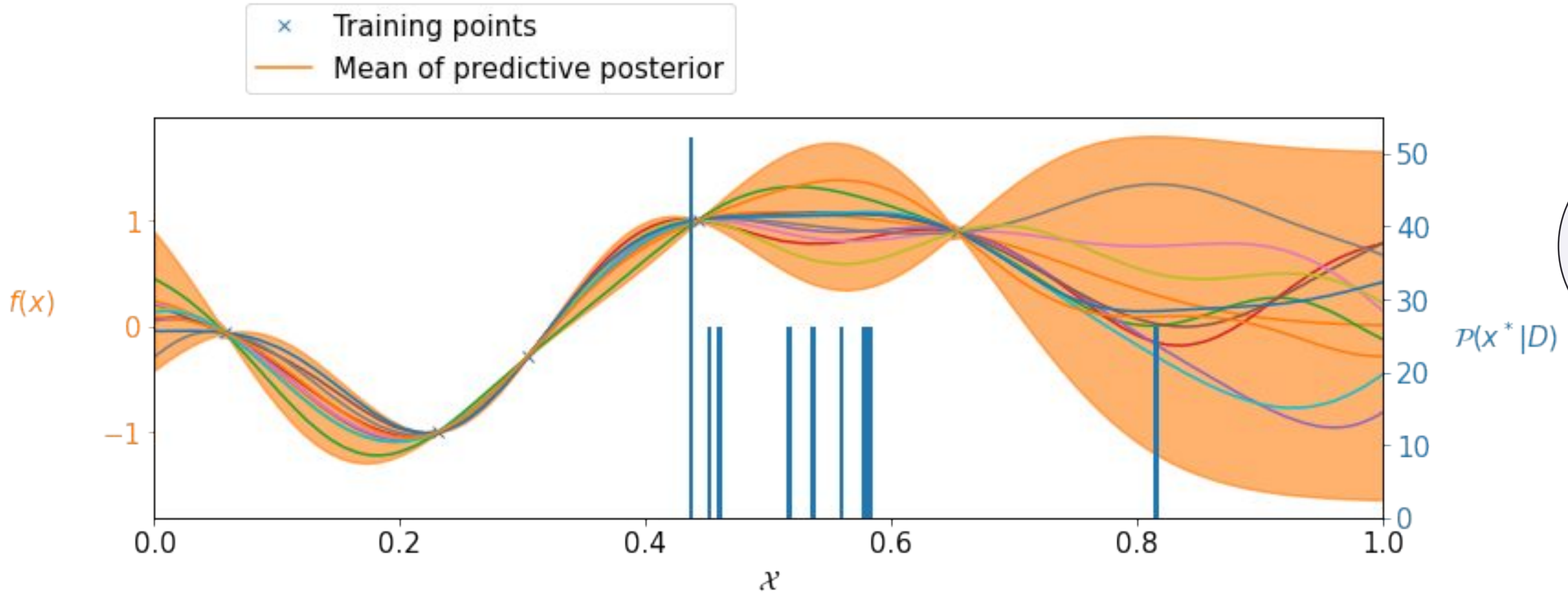
$P(\mathbf{x}^*|D)$ based on one sample



1 sample

What is our best guess for $\mathbf{x}^*$?

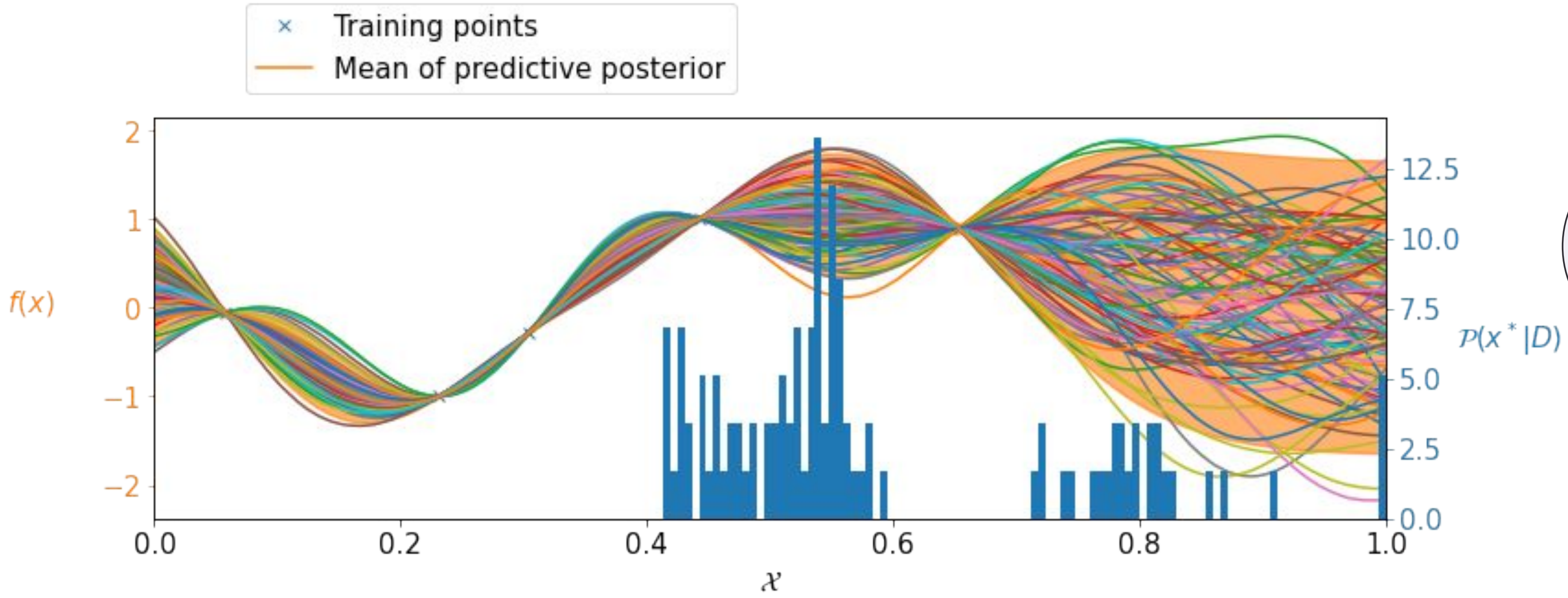
$P(\mathbf{x}^*|D)$ based on 10 samples



10 samples

What is our best guess for $\mathbf{x}^*$?

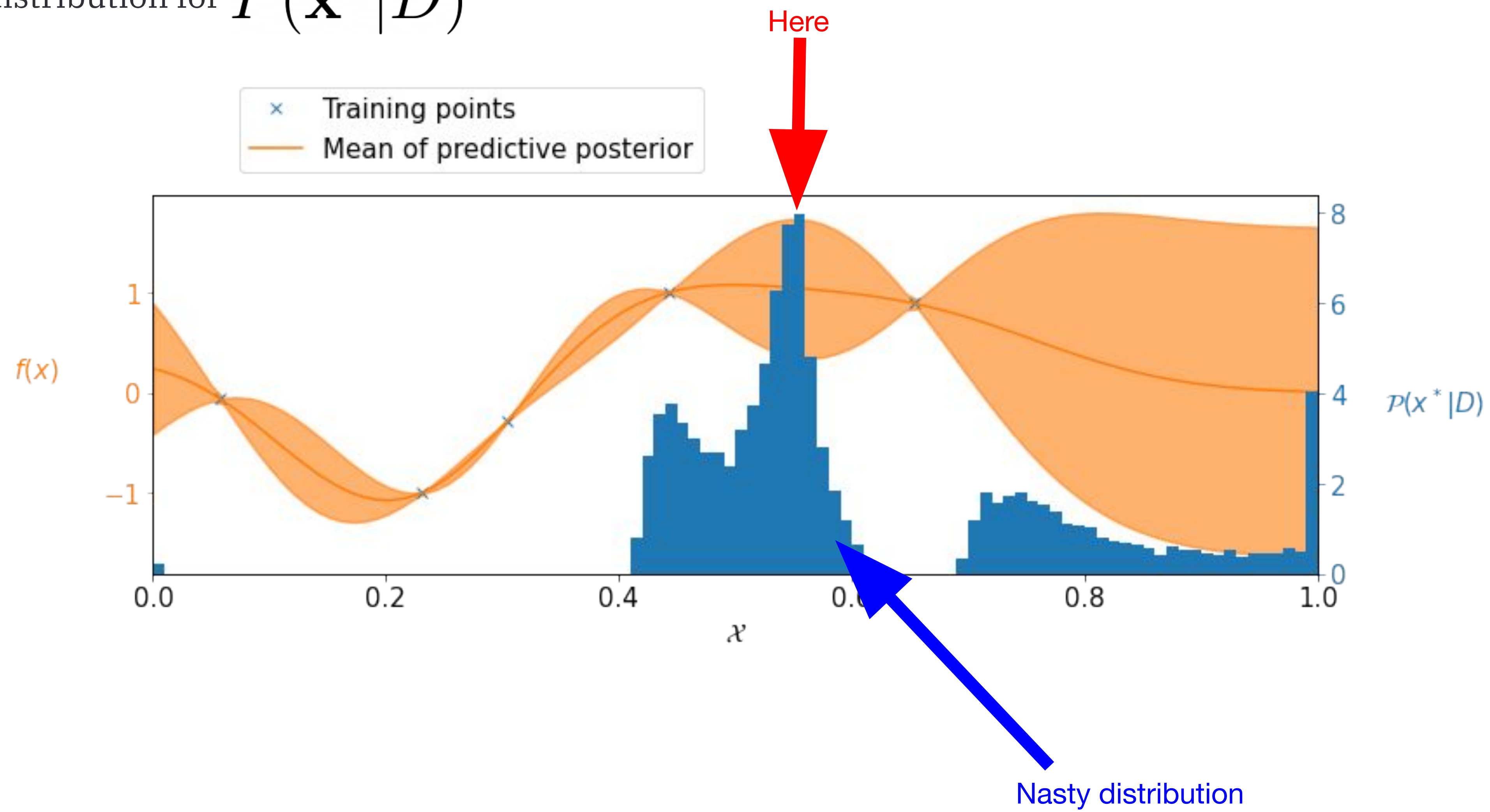
$P(\mathbf{x}^*|D)$ based on 100 samples



100 samples

What is our best guess for $\mathbf{x}^*$?

Empirical distribution for $P(\mathbf{x}^*|D)$



Where shall we evaluate next ?

We want to learn about $\mathbf{x}^*$

- Expected Improvement (EI) maximises $\alpha_{EI}(\mathbf{x}) = E[\max(f(\mathbf{x}) - f^*, 0)]$

Where shall we evaluate next ?

We want to learn about $\mathbf{x}^*$

- Expected Improvement (EI) maximises $\alpha_{EI}(\mathbf{x}) = E[\max(f(\mathbf{x}) - f^*, 0)]$

Only needs $f(\mathbf{x})|D$

Where shall we evaluate next ?

We want to learn about $\mathbf{x}^*$

- Expected Improvement (EI) maximises $\alpha_{EI}(\mathbf{x}) = E[\max(f(\mathbf{x}) - f^*, 0)]$

Only needs $f(\mathbf{x})|D$

Does not use full knowledge of $P(\mathbf{x}^*|D)$

Where shall we evaluate next ?

We want to learn about $\mathbf{x}^*$

- Expected Improvement (EI) maximises $\alpha_{EI}(\mathbf{x}) = E[\max(f(\mathbf{x}) - f^*, 0)]$

Only needs $f(\mathbf{x})|D$

Does not use full knowledge of $P(\mathbf{x}^*|D)$

Entropy search seeks to reduce our uncertainty in $P(\mathbf{x}^*|D)$

How to measure uncertainty?

How to measure uncertainty?

Variance or Differential Entropy?

$$\mathrm{Var}(X) = E\left[(X - \mu)^2\right]$$

How to measure uncertainty?

Variance or Differential Entropy?

$$\mathrm{Var}(X) = E\left[(X-\mu)^2\right]$$

$$\mathrm{H}(X) = E\left[-\log(p(X))\right]$$

How to measure uncertainty?

Variance or Differential Entropy?

$$\mathrm{Var}(X) = E\left[(X-\mu)^2\right]$$

$$\mathrm{H}(X) = E\left[-\log(p(X))\right]$$

	$\mathrm{Var}(X)$	$H(X)$
$X \sim \mathcal{N}(\mu, \sigma^2)$	σ^2	$\log(\sigma\sqrt{2\pi e})$

How to measure uncertainty?

Variance or Differential Entropy?

$$\mathrm{Var}(X) = E\left[(X-\mu)^2\right]$$

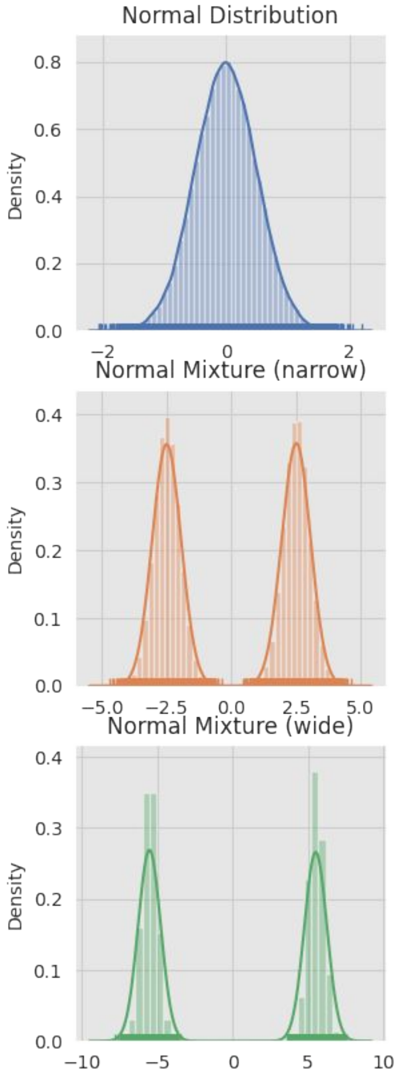
$$\mathrm{H}(X) = E\left[-\log(p(X))\right]$$

	$\mathrm{Var}(X)$	$H(X)$
$X \sim \mathcal{N}(\mu, \sigma^2)$	σ^2	$\log(\sigma\sqrt{2\pi e})$
$X \sim U(a, b)$	$\frac{(b-a)^2}{12}$	$\log(b-a)$

How to measure uncertainty?

Should we use entropy?

$$H(X) = E\left[-\log(p(X))\right]$$



H(X) = 0.7

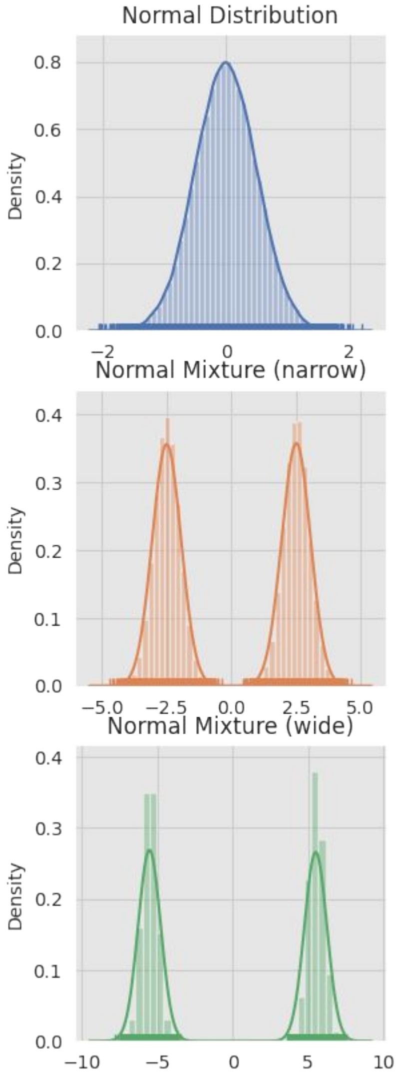
H(X) = 1.4

H(X) = 1.4

$$\mathrm{Var}(X) = E\left[(X-\mu)^2\right]$$

How to measure uncertainty?

Should we use variance (i.e. dispersion)?



Var(X) = 0.5

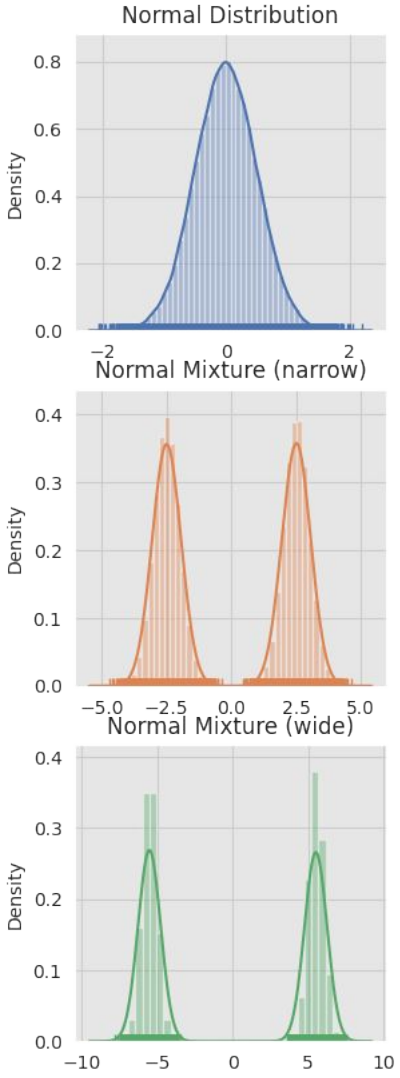
Var(X) = 6.5

Var(X) = 30.5

$$\mathrm{Var}(X) = E\left[(X-\mu)^2\right]$$

How to measure uncertainty?

Should we use variance (i.e. dispersion)?



Var(X) = 0.5

Var(X) = 6.5

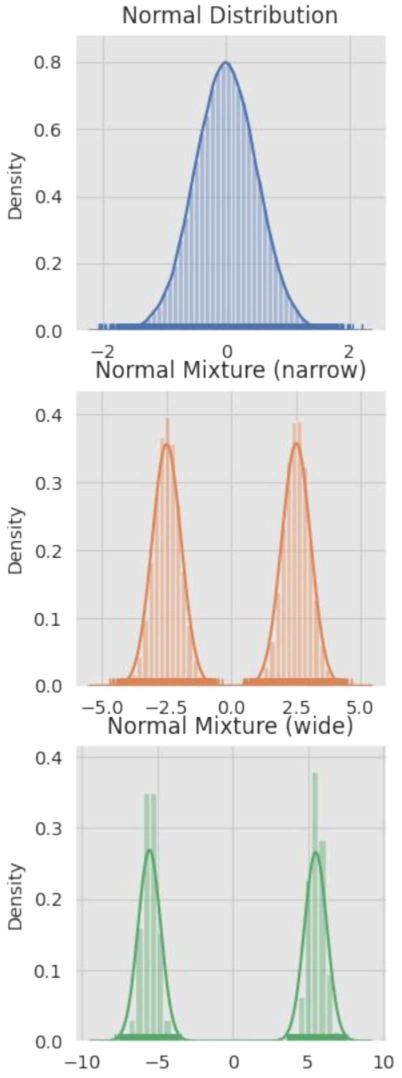
Var(X) = 30.5

Perhaps not good for multi-modal distributions ?

$$\mathrm{Var}(X) = E\left[(X-\mu)^2\right]$$

How to measure uncertainty?

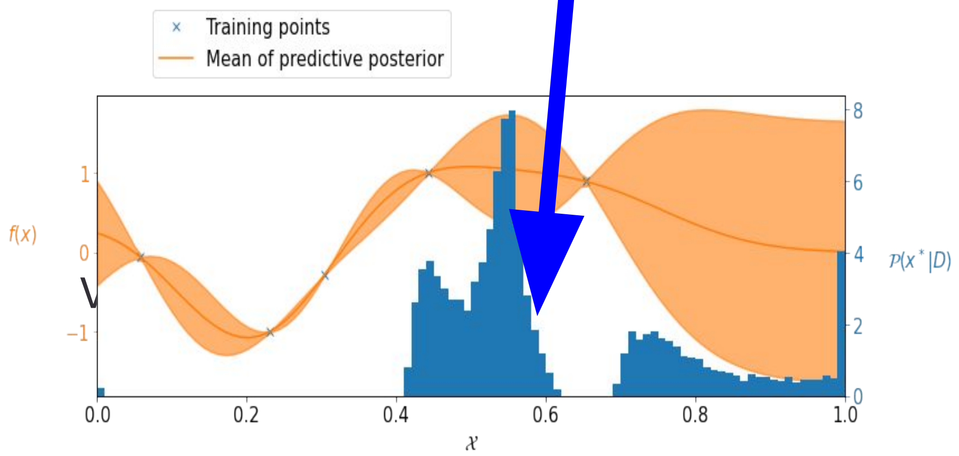
Should we use variance (i.e. dispersion)?



Var(X) = 0.5

Var(X) = 6.5

Perhaps not good for multi-modal distributions ?



Entropy Search

Reduce global uncertainty in $P(\mathbf{x}^*|D)$

Entropy Search

Reduce global uncertainty in $P(\mathbf{x}^* | D)$

How?

- Measure uncertainty by differential entropy $H(\mathbf{x}^* | D) = -E_{\mathbf{x} \sim \mathbf{x}^* | D}[\log(p(\mathbf{x}))]$

UNIVERSITY OF CAMBRIDGE Lancaster University

Entropy Search

Reduce global uncertainty in $P(\mathbf{x}^*|D)$

How?

- Measure uncertainty by differential entropy $H(\mathbf{x}^*|D) = -E_{\mathbf{x}\sim\mathbf{x}^*|D}[\log(p(\mathbf{x}))]$
- Make evaluation that provides the largest expected reduction in entropy

$$\alpha_{ES}(\mathbf{x}) = H(\mathbf{x}^*|D) - E_y[H(\mathbf{x}^*|D \cup \{y, \mathbf{x}\})]$$

Entropy Search

Reduce global uncertainty in $P(\mathbf{x}^*|D)$

How?

- Measure uncertainty by differential entropy $H(\mathbf{x}^*|D) = -E_{\mathbf{x}\sim\mathbf{x}^*|D}[\log(p(\mathbf{x}))]$
- Make evaluation that provides the largest expected reduction in entropy

$$\alpha_{ES}(\mathbf{x}) = H(\mathbf{x}^*|D) - E_y[H(\mathbf{x}^*|D \cup \{y, \mathbf{x}\})]$$

Current uncertainty

Expected uncertainty after collecting evaluation y at location $\mathbf{x}$

Entropy Search

Reduce global uncertainty in $P(\mathbf{x}^*|D)$

How?

- Measure uncertainty by differential entropy $H(\mathbf{x}^*|D) = -E_{\mathbf{x}\sim\mathbf{x}^*|D}[\log(p(\mathbf{x}))]$
- Make evaluation that provides the largest expected reduction in entropy

$$\alpha_{ES}(\mathbf{x}) = H(\mathbf{x}^*|D) - E_y[H(\mathbf{x}^*|D \cup \{y, \mathbf{x}\})]$$

Current uncertainty

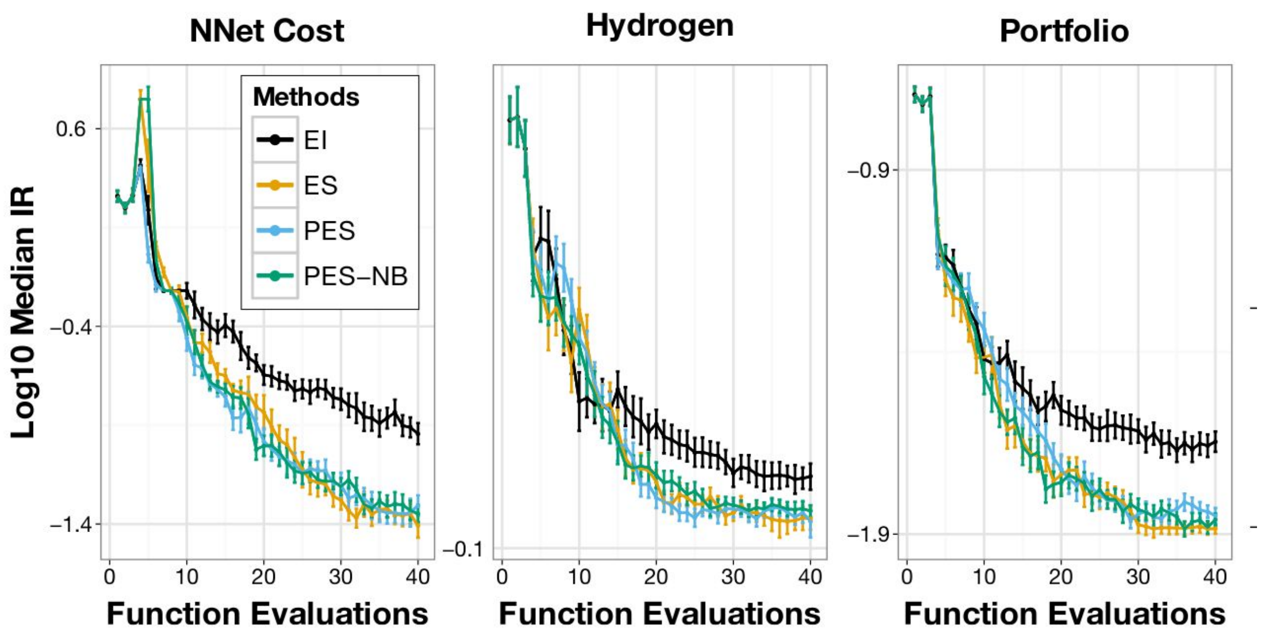
Expected uncertainty after collecting evaluation y at location $\mathbf{x}$

Fiendishly difficult to calculate!

- **What is** $H(\mathbf{x}^*|D)$ **?**
- **What is** $H(\mathbf{x}^*|D, \{y, \mathbf{x}\})$**???**

It can be worth calculating these horrible quantities

They can provide highly efficient optimization



For details see

- Entropy Search is $O(n^2 e^{2d} + e^{3d})$(Henning and Schuler, 2012)
- Predictive Entropy Search is $O(n^2 e^{2d} + n^3 e^d)$ (Hernandez-Lobato et al. 2014)

There is a better way!

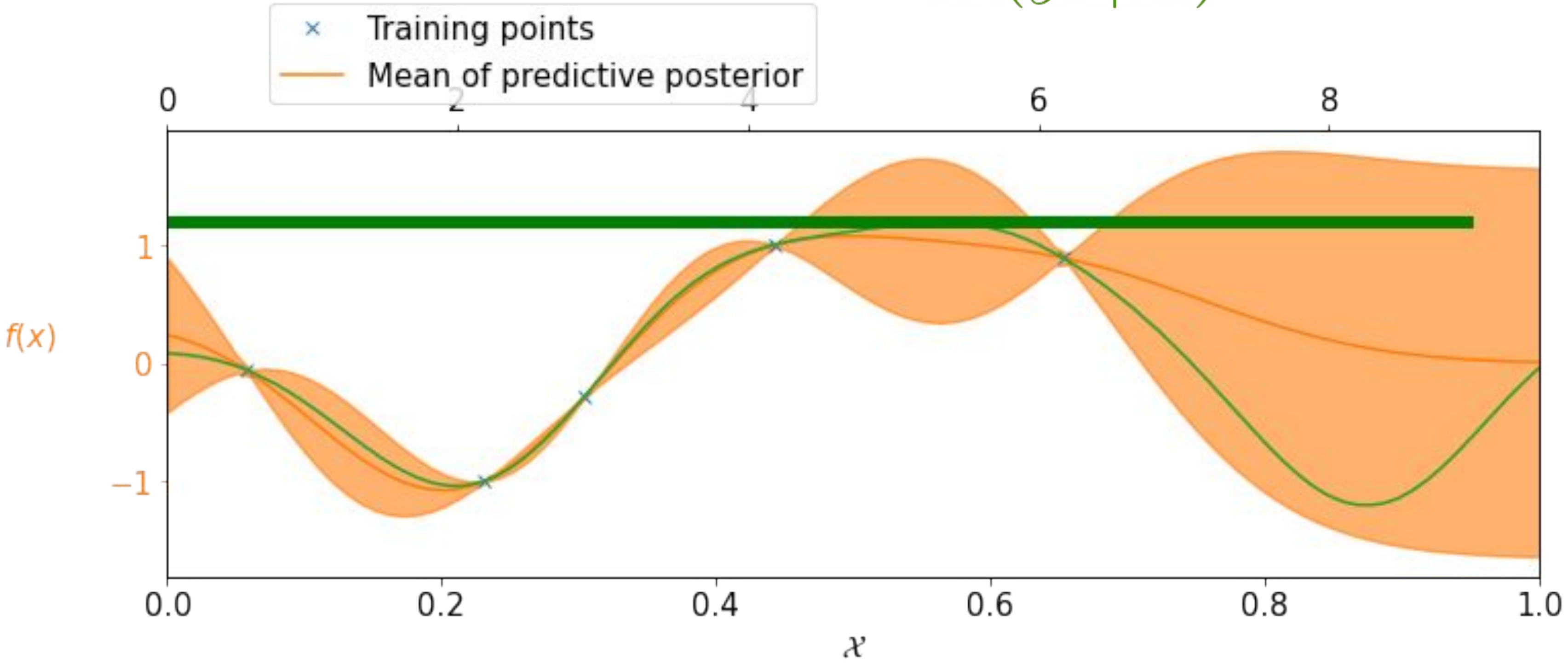
Min-value Entropy Search

Rather than reduce uncertainty in $H(\mathbf{x}^*|D)$, instead look at $H(y^*|D)$ where $y^* = f(\mathbf{x}^*)$

There is a better way!

Min-value Entropy Search

Rather than reduce uncertainty in $H(\mathbf{x}^*|D)$, instead look at $H(y^*|D)$ where $y^* = f(\mathbf{x}^*)$



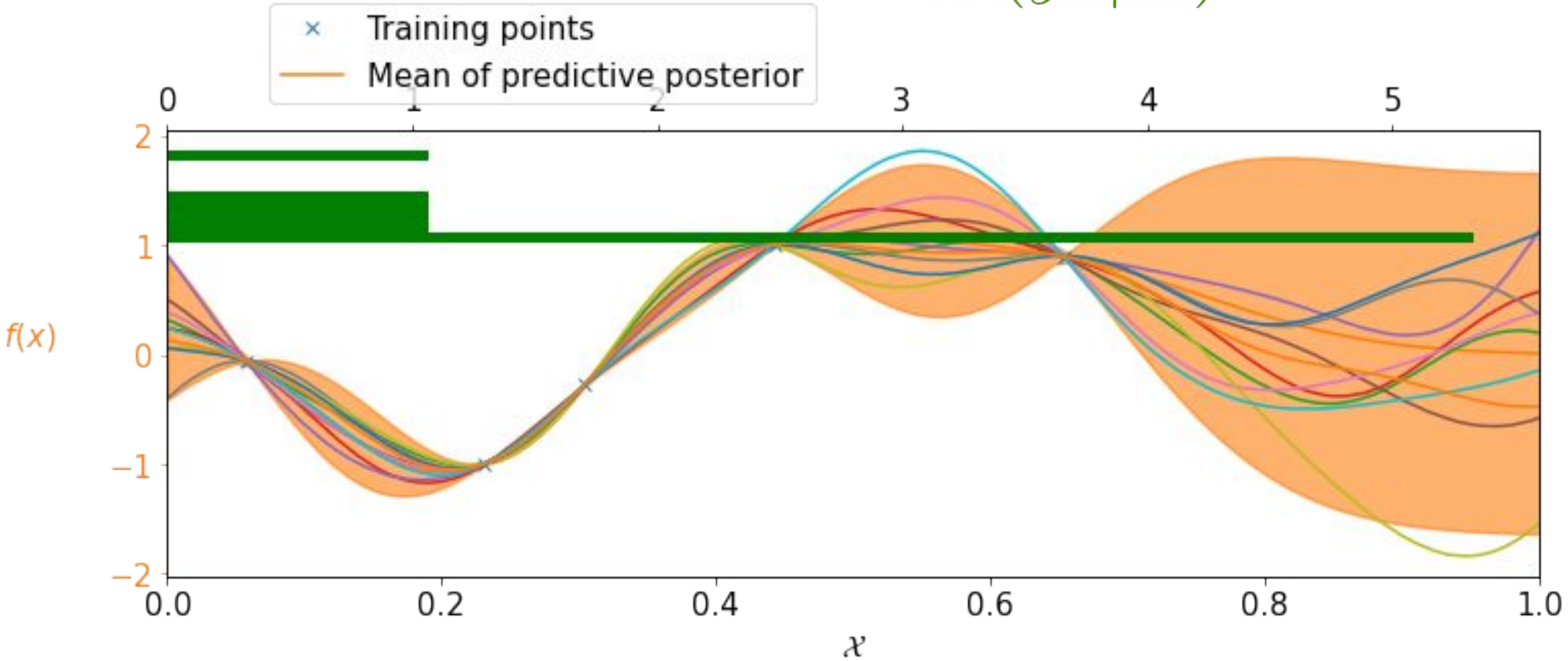
1 sample

UNIVERSITY OF CAMBRIDGE Lancaster University

There is a better way!

Min-value Entropy Search

Rather than reduce uncertainty in $H(\mathbf{x}^*|D)$, instead look at $H(y^*|D)$ where $y^* = f(\mathbf{x}^*)$

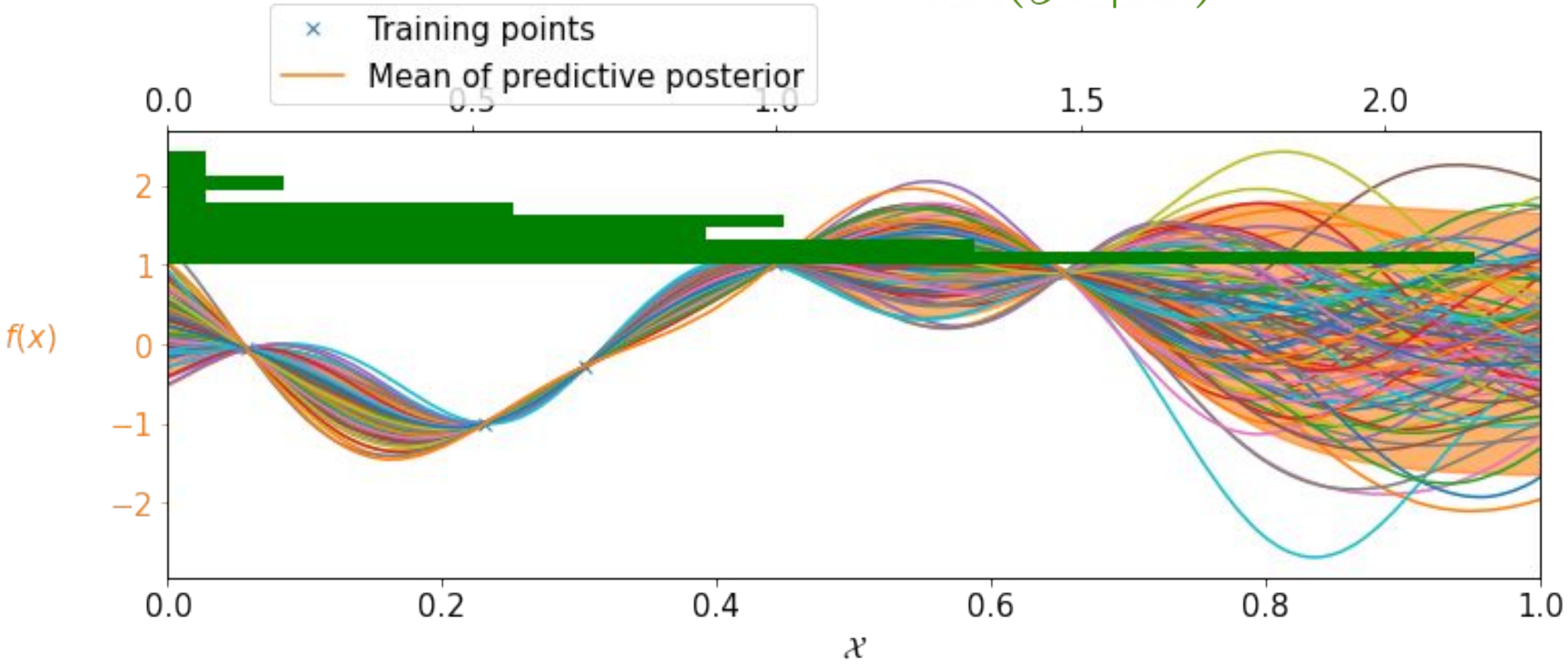


10 samples

There is a better way!

Min-value Entropy Search

Rather than reduce uncertainty in $H(\mathbf{x}^*|D)$, instead look at $H(y^*|D)$ where $y^* = f(\mathbf{x}^*)$

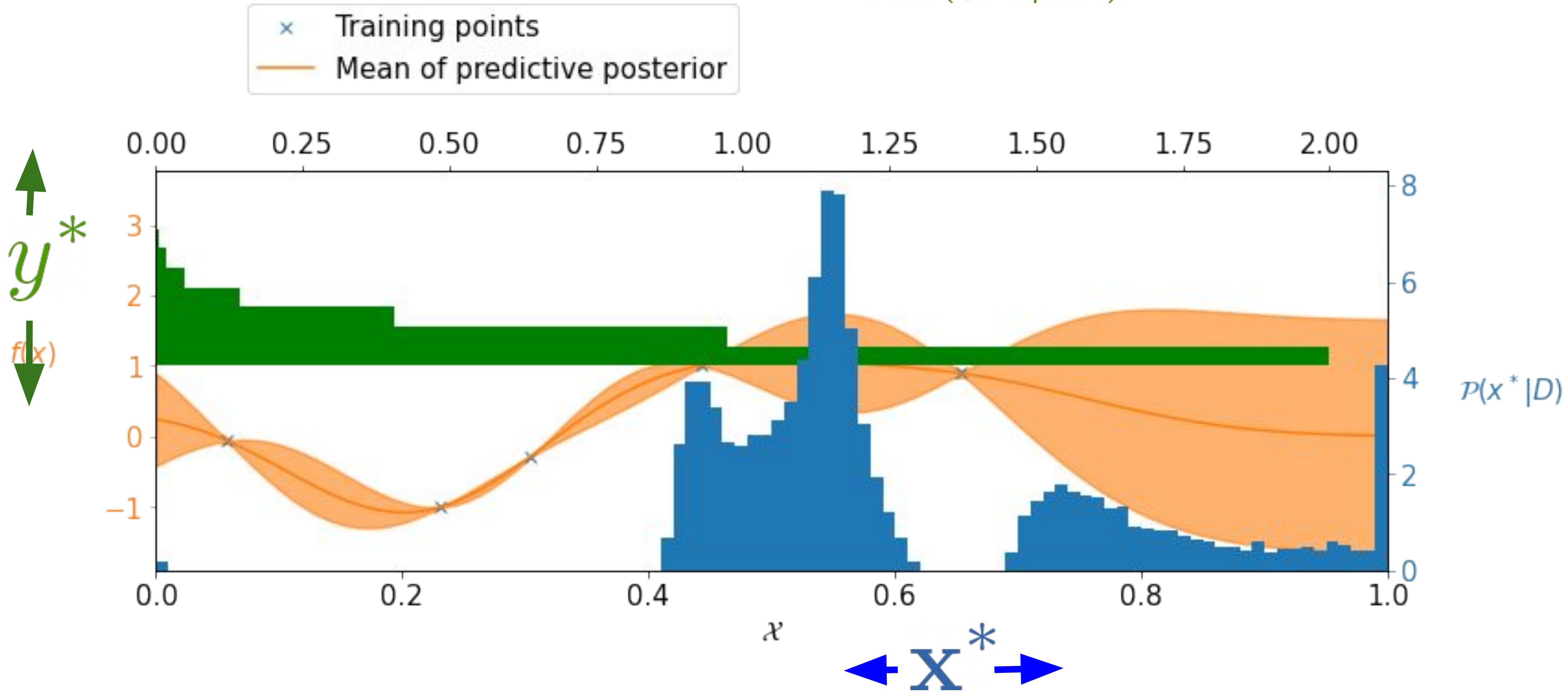


100 samples

There is a better way!

Min-value Entropy Search

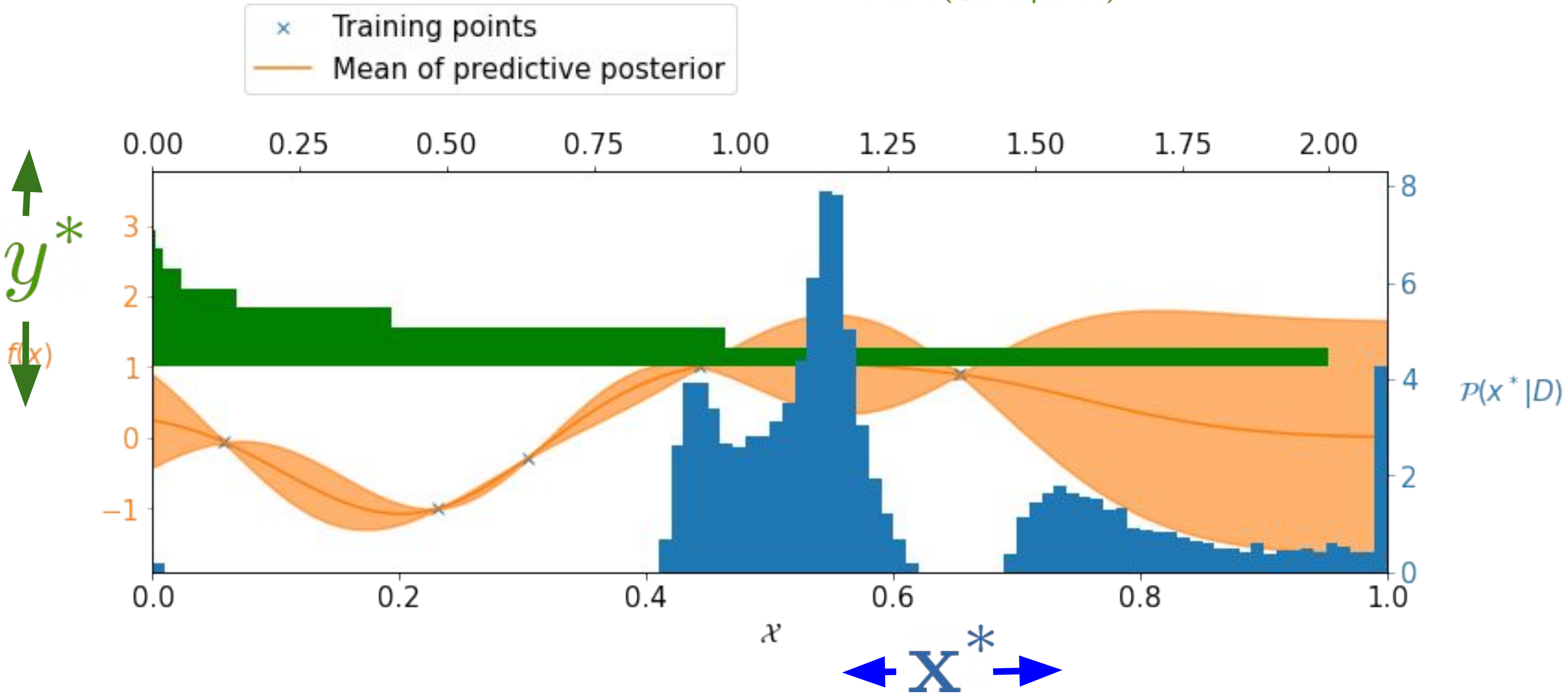
Rather than reduce uncertainty in $H(\mathbf{x}^*|D)$, instead look at $H(y^*|D)$ where $y^* = f(\mathbf{x}^*)$



There is a better way!

Min-value Entropy Search

Rather than reduce uncertainty in $H(\mathbf{x}^*|D)$, instead look at $H(y^*|D)$ where $y^* = f(\mathbf{x}^*)$



$$\alpha_{MES}(\mathbf{x}) = H(y|D) - E_{y^*|D}[y|D \cup y^*]$$

There is a better way!

Min-value Entropy Search

Rather than reduce uncertainty in $H(\mathbf{x}^*|D)$, instead look at $H(y^*|D)$ where $y^* = f(\mathbf{x}^*)$

$$\alpha_{\mathrm{MES}}(\mathbf{x}) = H(y^*|D) - E_{y|D}\left[H(y^*|D \bigcup (y, \mathbf{x}))\right]$$

Current uncertainty

Expected uncertainty after the evaluation

There is a better way!

Min-value Entropy Search

Rather than reduce uncertainty in $H(\mathbf{x}^*|D)$, instead look at $H(y^*|D)$ where $y^* = f(\mathbf{x}^*)$

$$\alpha_{\text{MES}}(\mathbf{x}) = H(y^*|D) - E_{y|D}\left[H(y^*|D \bigcup (y, \mathbf{x}))\right]$$

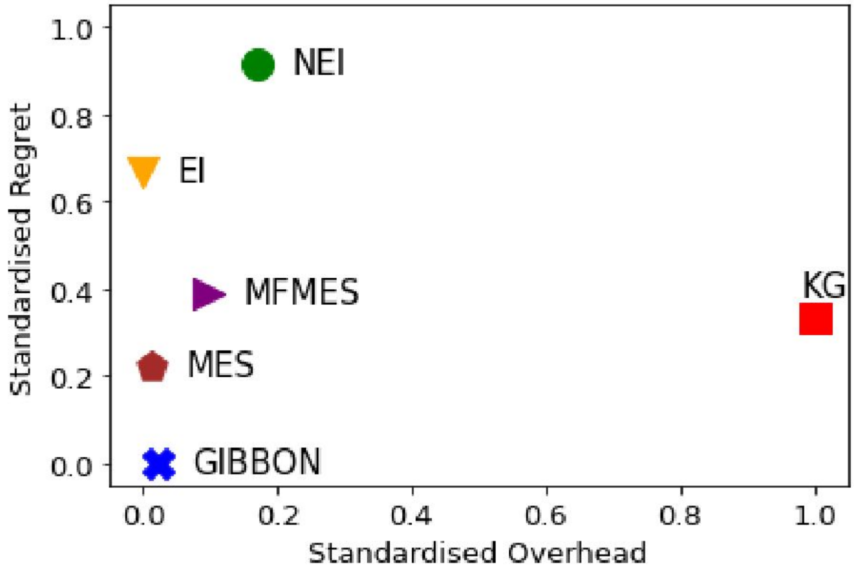
Current uncertainty

Expected uncertainty after the evaluation

Crucially $\mathbf{y}^* \in R$, whereas $\mathbf{x}^* \in R^d$

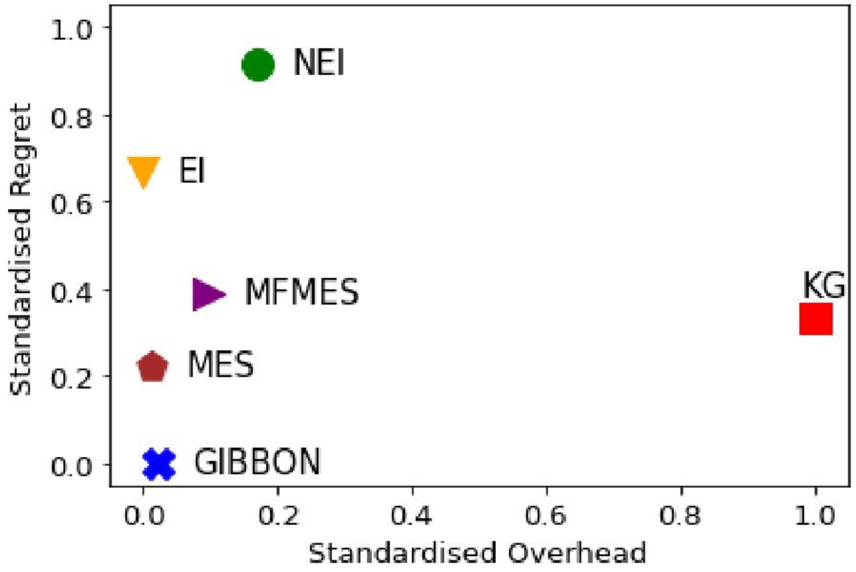
MES in practice

Highly effective optimization at low cost!



MES in practice

Highly effective optimization at low cost!



- Max-Value Entropy Search is $O(n^2e^d)$ for noiseless optimisation (Wang and Jegelka, 2017).
- MUMBO is $O(n^2e^d)$ for noisy optimisation (Moss et al., 2020)
- GIBBON is $O((n^2+B^2)e^d+B^3)$for batches of size B (Moss et al. 2021)

Thanks for listening

UNIVERSITY OF CAMBRIDGE

Lancaster University