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

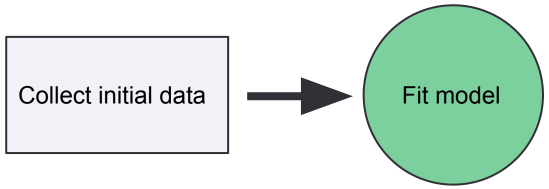
Sequential data collection

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

Collect initial data

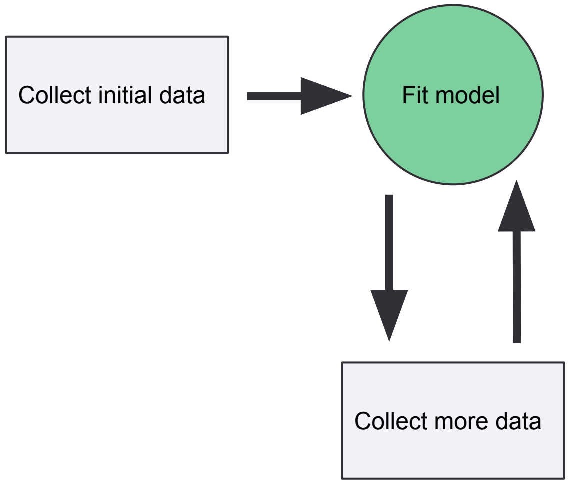
Sequential data collection

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



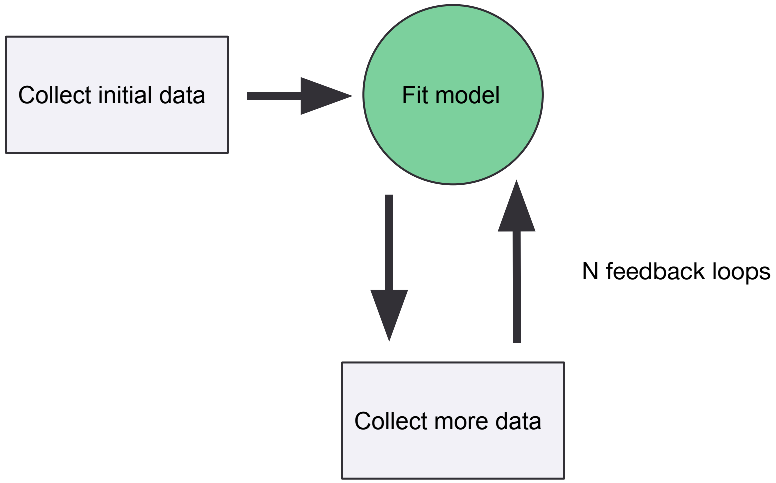
Sequential data collection

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



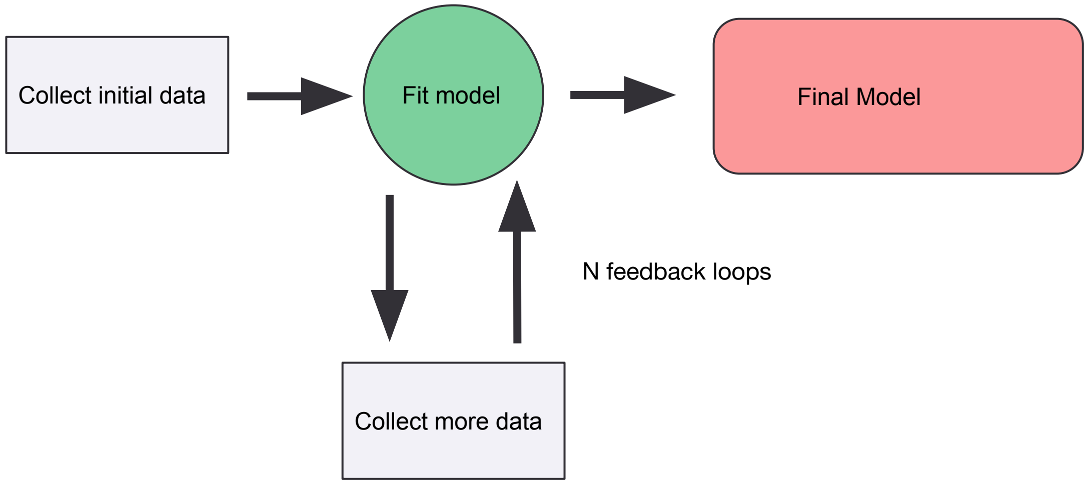
Sequential data collection

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



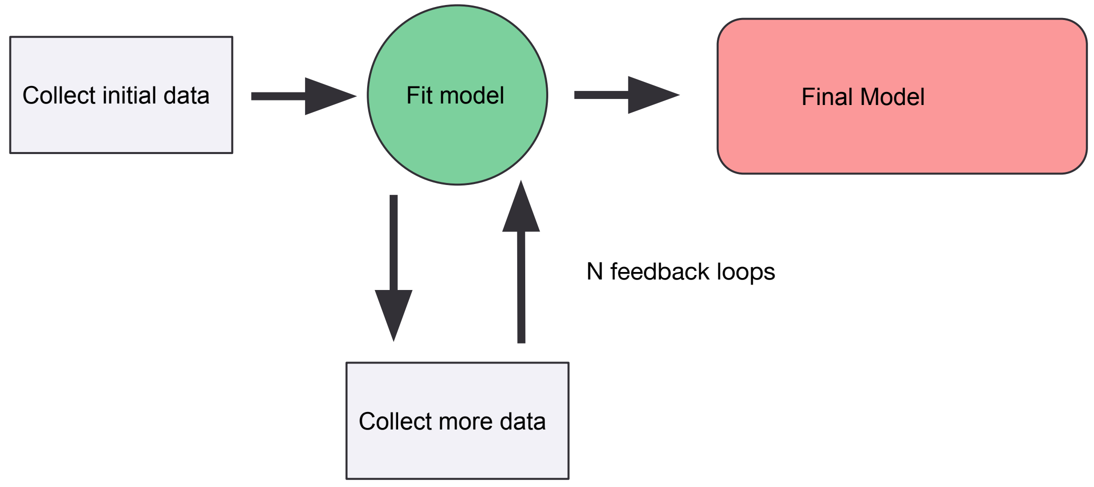
Sequential data collection

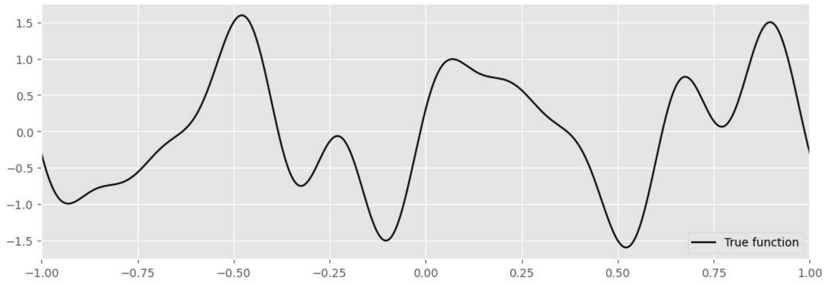
Let's make use of uncertainty estimates to make better models



Sequential data collection

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

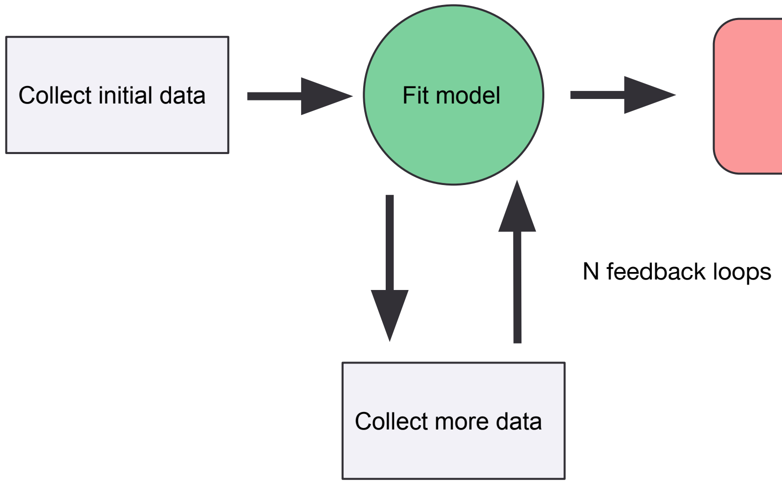


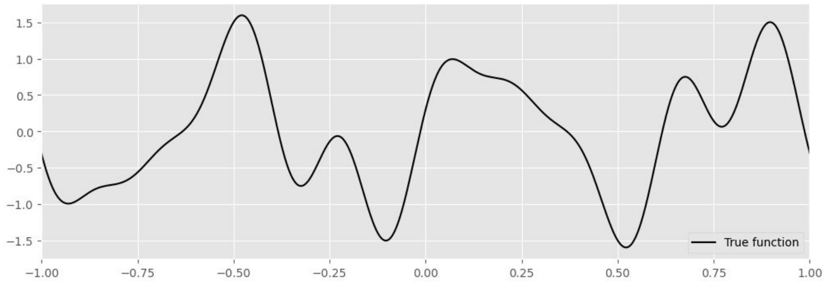


0

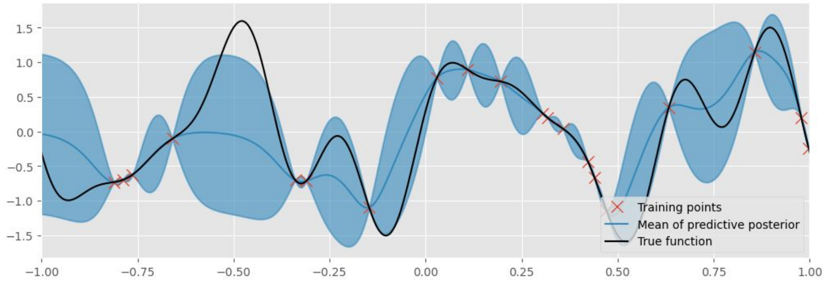
Sequential data collection

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





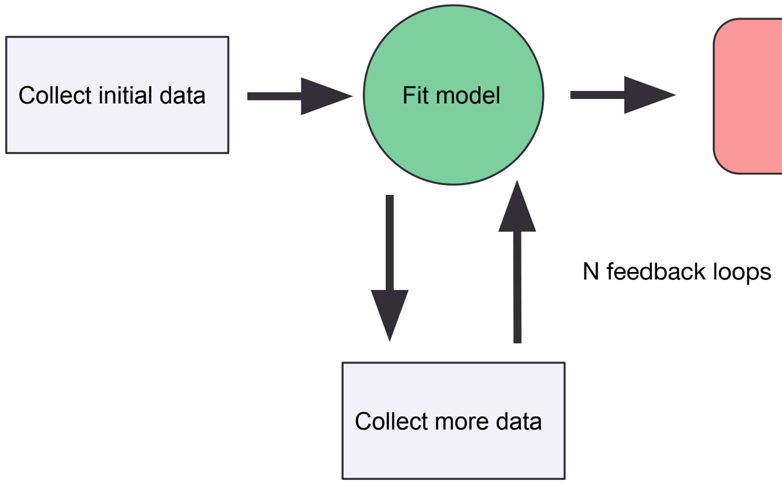
0

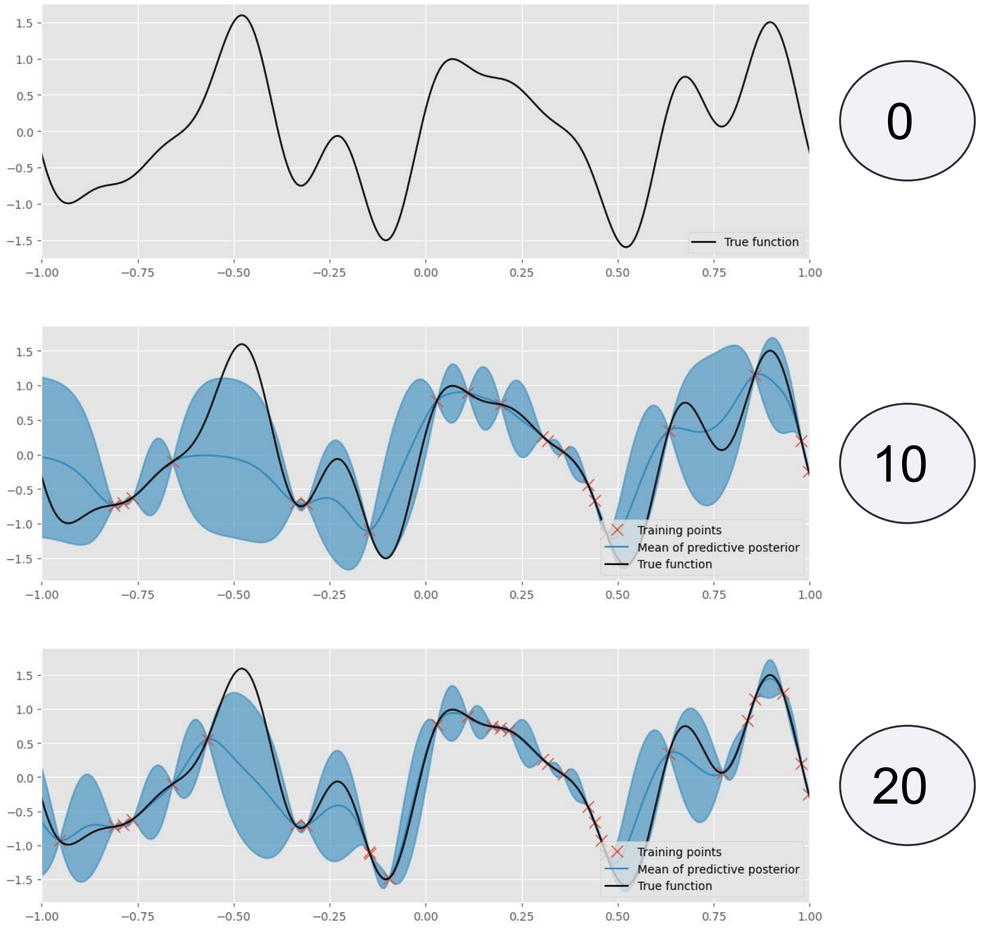


10

Sequential data collection

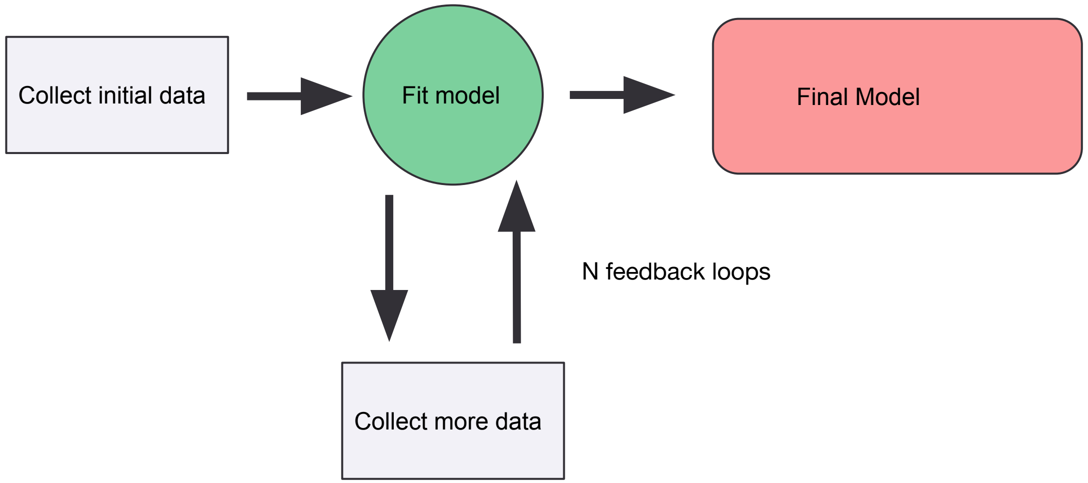
Let's make use of uncertainty estimates to make better models

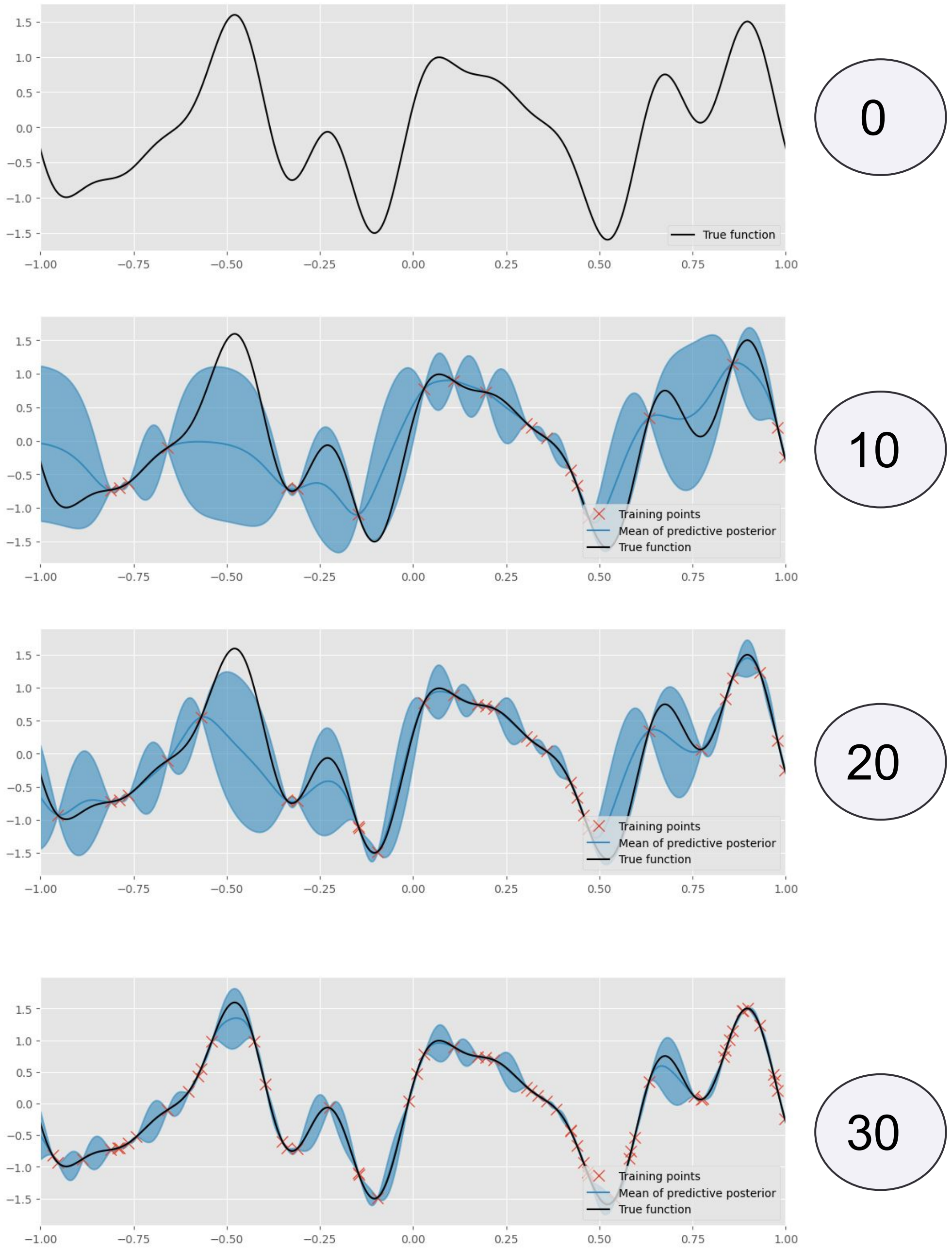




Sequential data collection

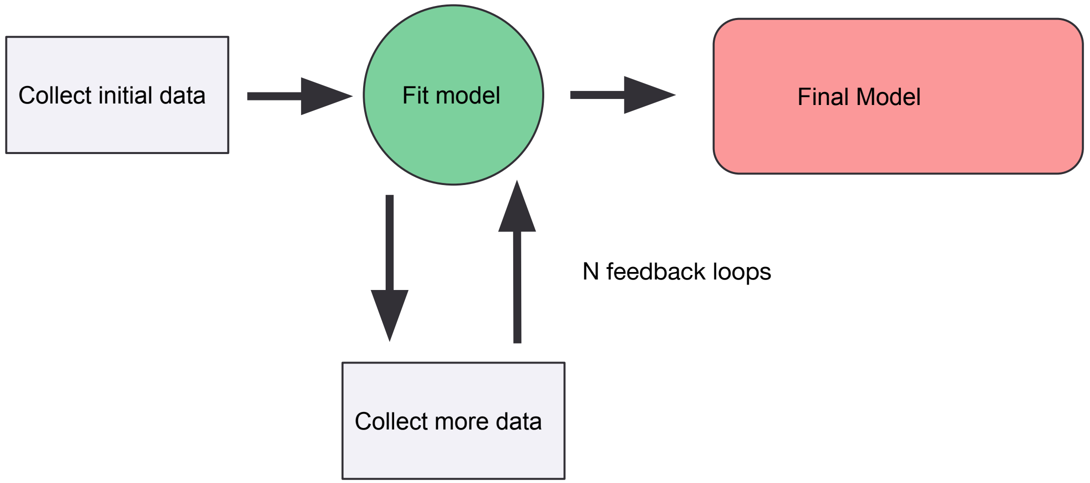
Let's make use of uncertainty estimates to make better models



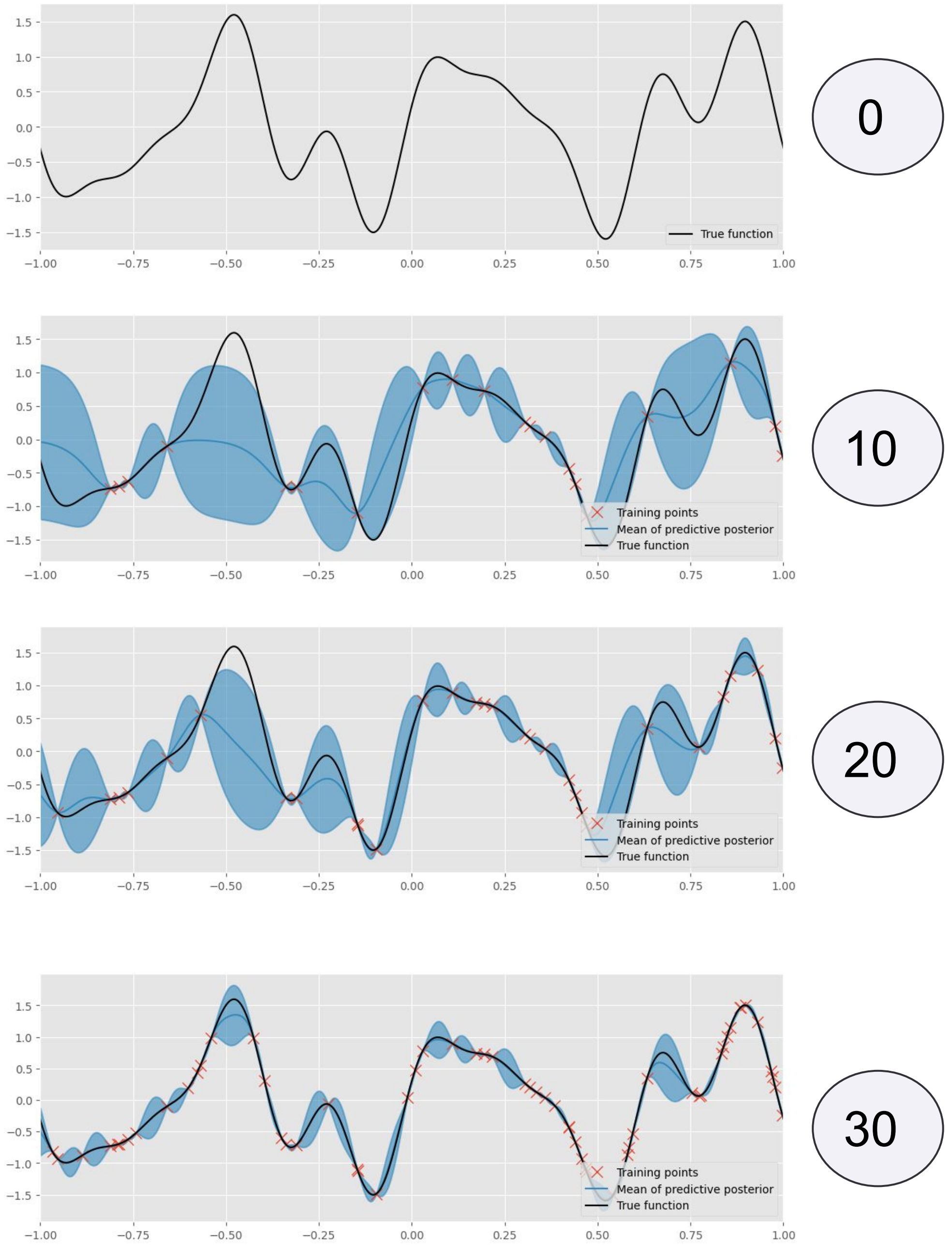


Sequential data collection

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

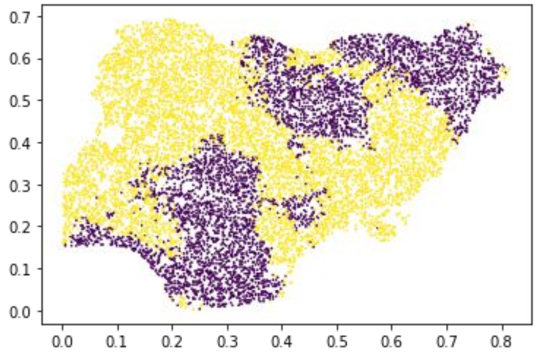
Sequentially collecting more data to improve your model for the task at hand

- 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

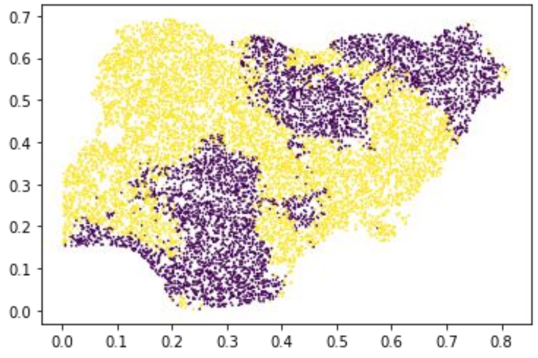
UNIVERSITY OF CAMBRIDGE

Lancaster University

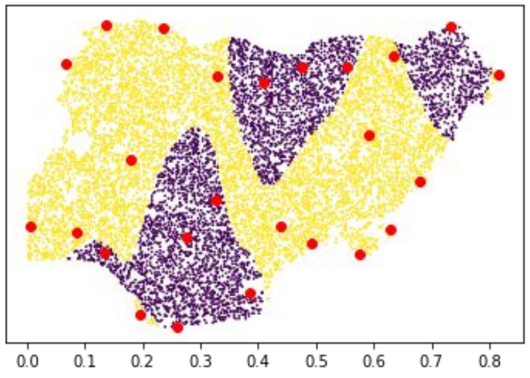
Active learning

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

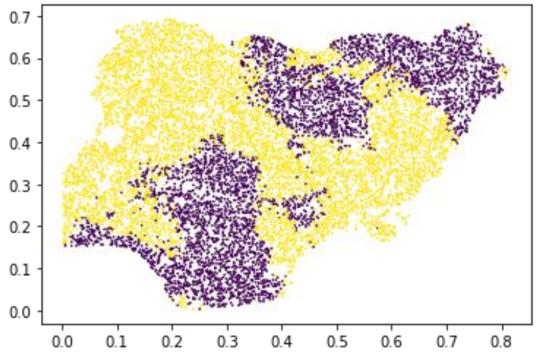


Model on Random data

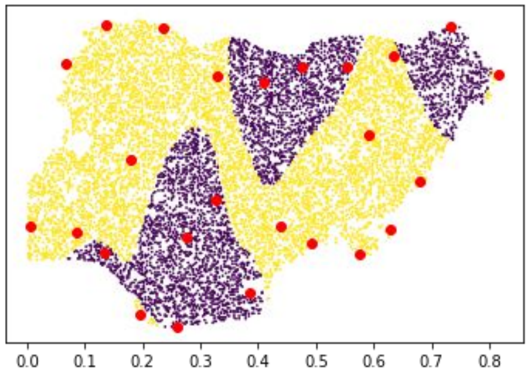
Active learning

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

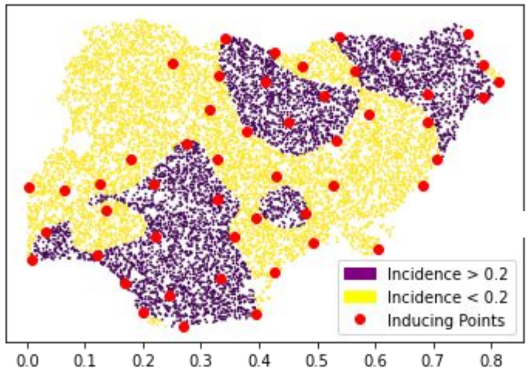
- I care about **regression** —> collect data to improve global model accuracy
- I care about the **maximum** value of my process —> collect data in promising regions (Bayesian Optimisation)
- I'm interested in **multiple objectives** -> populate the Pareto front (Multi-objective Bayesian Optimisation)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

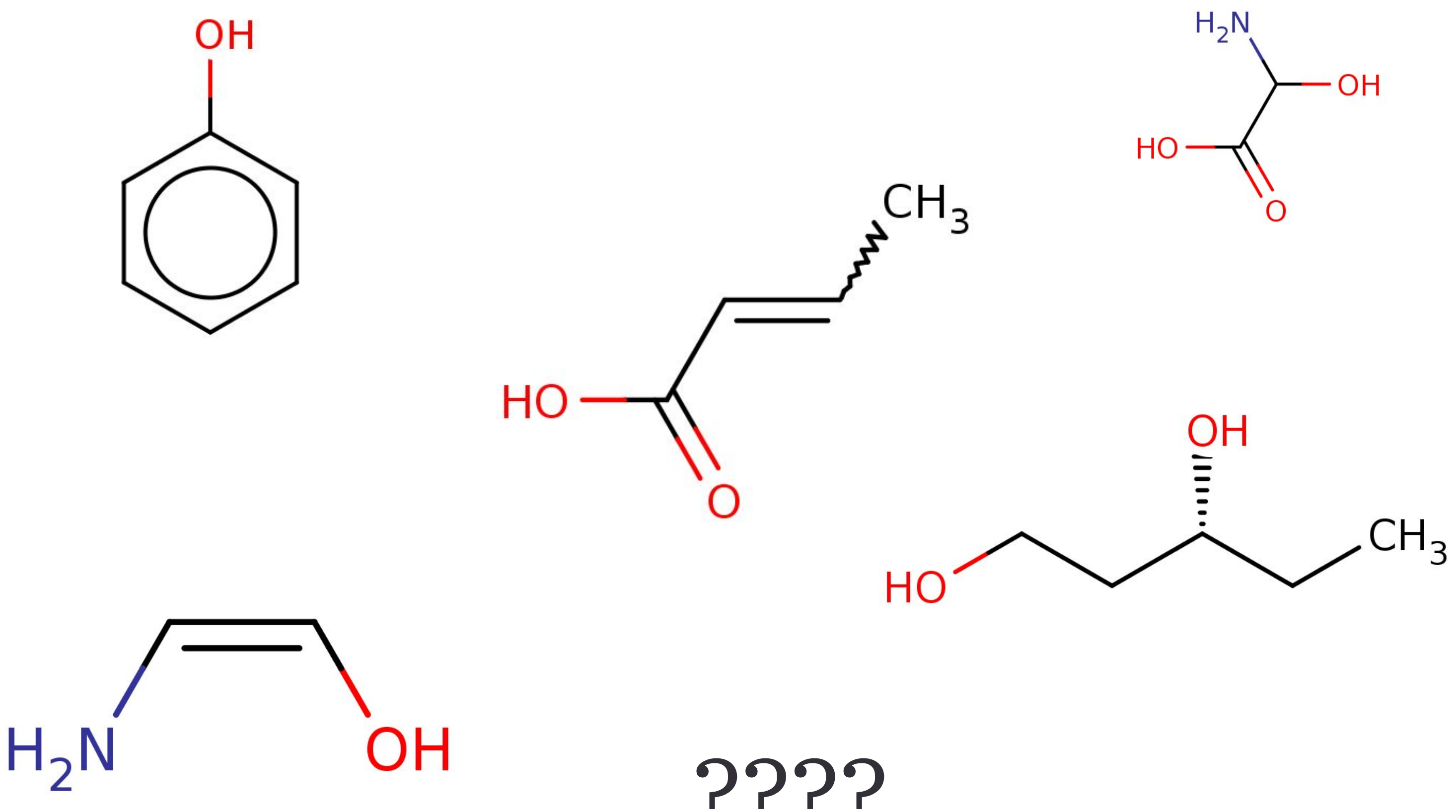
A molecular design pipeline

Efficiently explore molecule space

A molecular design pipeline

Efficiently explore molecule space

- **Large** library of candidates

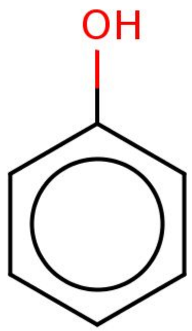


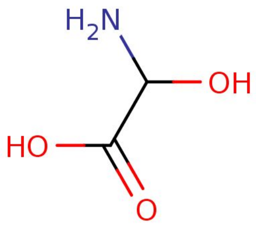
????

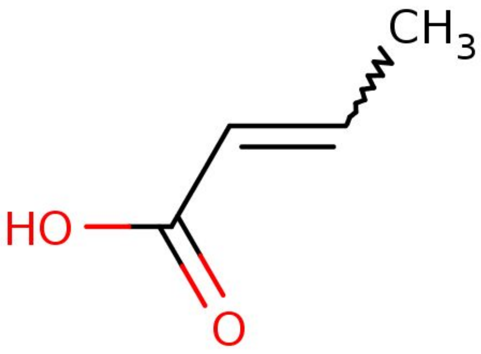
A molecular design pipeline

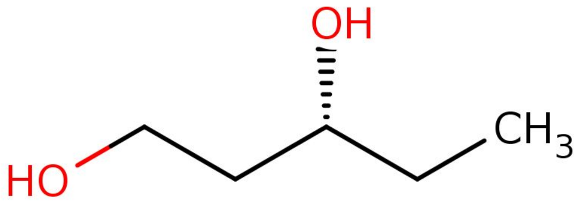
Efficiently explore molecule space

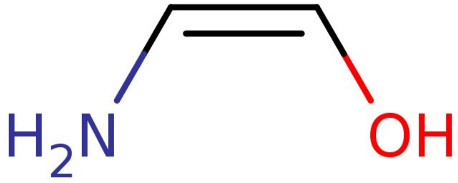
- **Large** library of candidates
- **Expensive** experiments (<10)









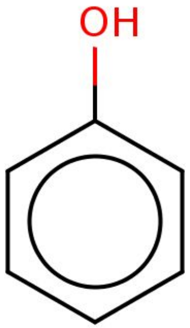


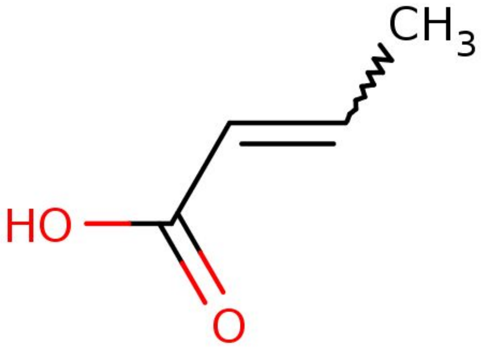
????

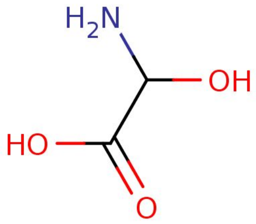
A molecular design pipeline

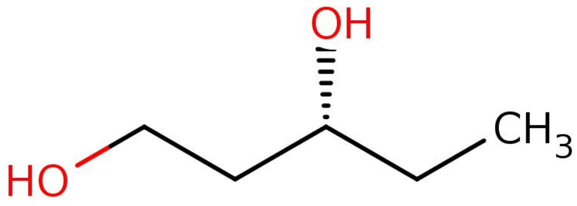
Efficiently explore molecule space

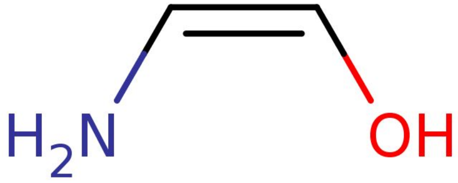
- **Large** library of candidates
- **Expensive** experiments (<10) (**IN A LAB !!!**)









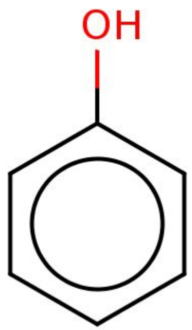


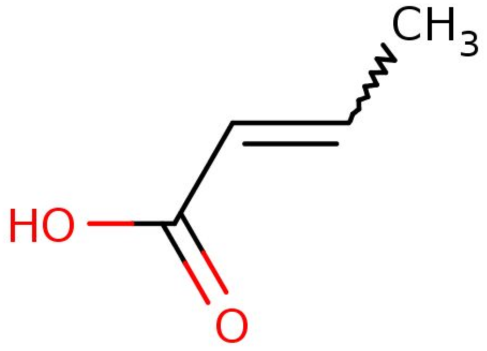
????

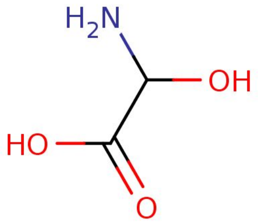
A molecular design pipeline

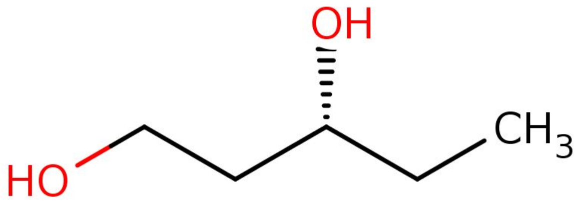
Efficiently explore molecule space

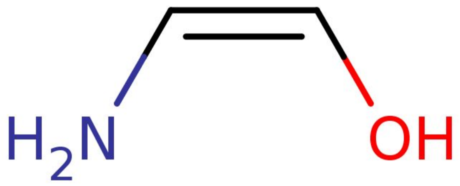
- **Large** library of candidates
- **Expensive** experiments (<10)
- High degree of **parallelism**









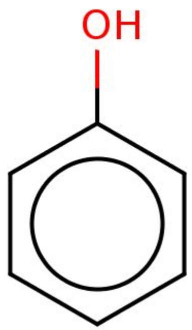


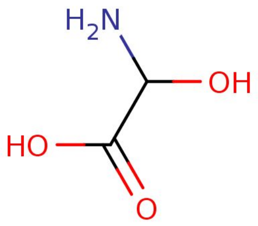
????

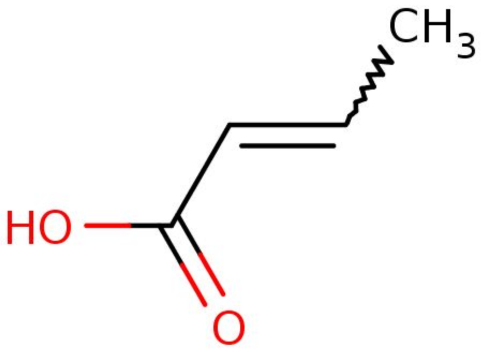
A molecular design pipeline

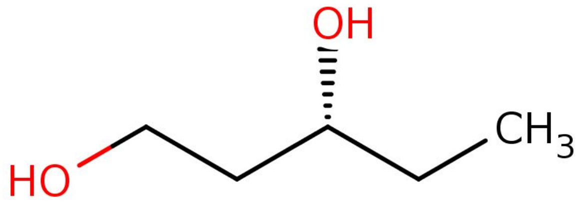
Efficiently explore molecule space

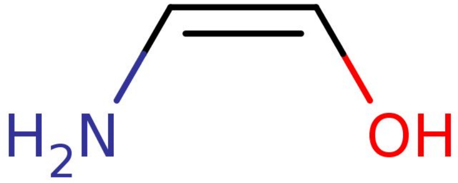
- **Large** library of candidates
- **Expensive** experiments (<10)
- High degree of **parallelism**
- Want molecules with high **affinity**









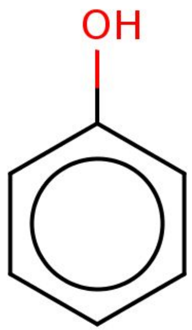


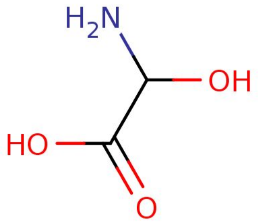
????

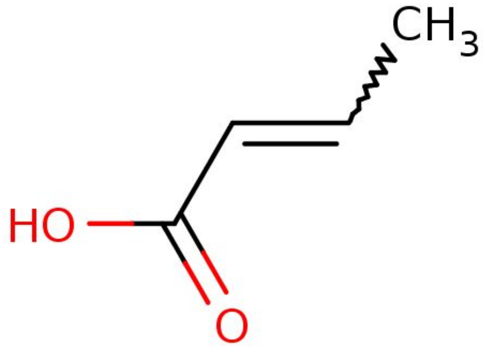
A molecular design pipeline

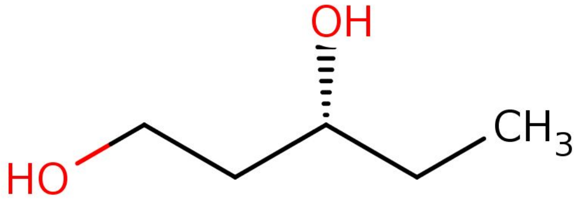
Efficiently explore molecule space

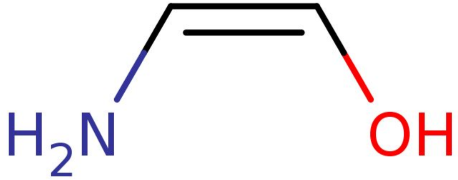
- **Large** library of candidates
- **Expensive** experiments (<10)
- High degree of **parallelism**
- Want molecules with high **affinity**
    - Also easy to make









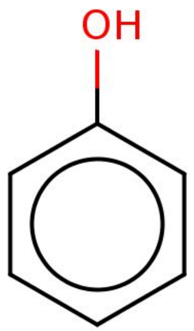


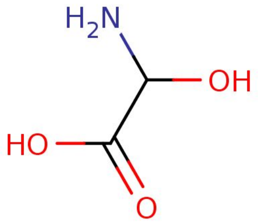
????

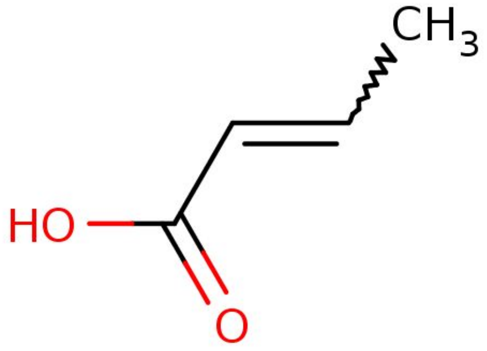
A molecular design pipeline

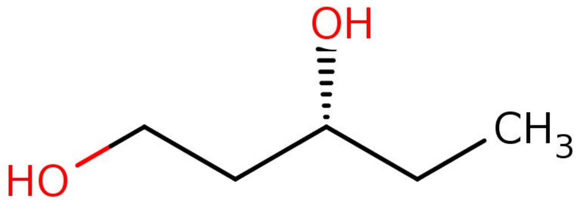
Efficiently explore molecule space

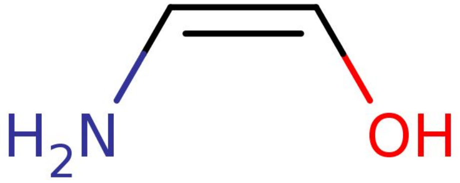
- **Large** library of candidates
- **Expensive** experiments (<10)
- High degree of **parallelism**
- Want molecules with high **affinity**
  - Also easy to make
  - Don't stick to themselves









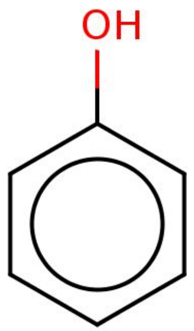


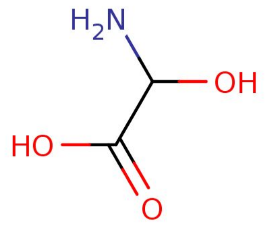
????

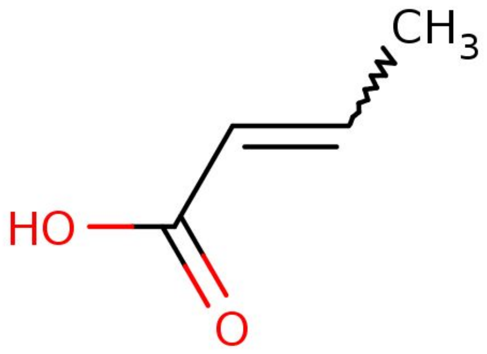
A molecular design pipeline

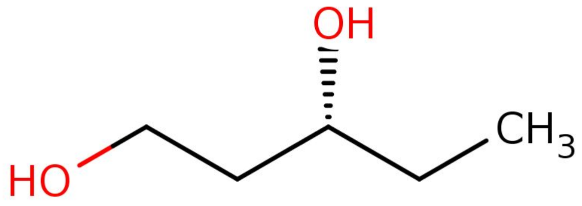
Efficiently explore molecule space

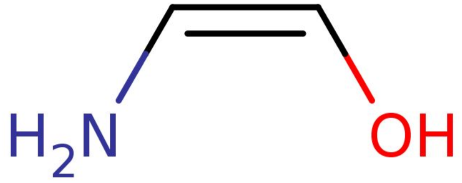
- **Large** library of candidates
- **Expensive** experiments (<10)
- High degree of **parallelism**
- Want molecules with high **affinity**
  - Also easy to make
  - Don’t stick to themselves
  - Stable









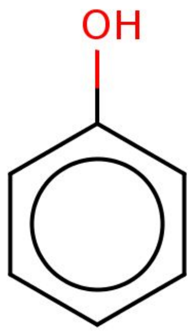


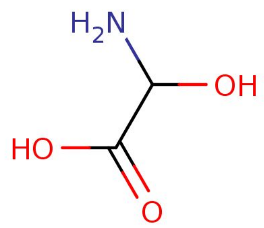
????

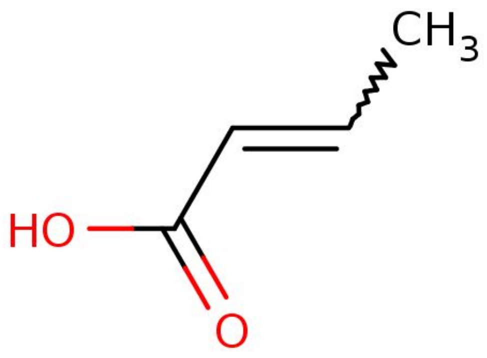
A molecular design pipeline

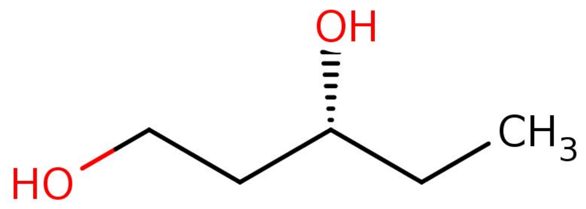
Efficiently explore molecule space

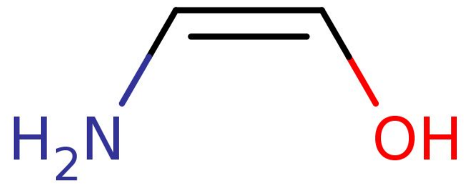
- **Large** library of candidates
- **Expensive** experiments (<10)
- High degree of **parallelism**
- Want molecules with high **affinity**
  - Also easy to make
  - Don’t stick to themselves
  - Stable
  - In a new area of “patent space”









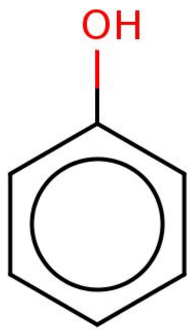


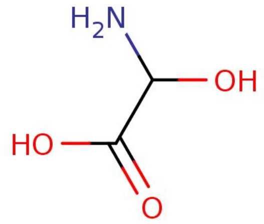
????

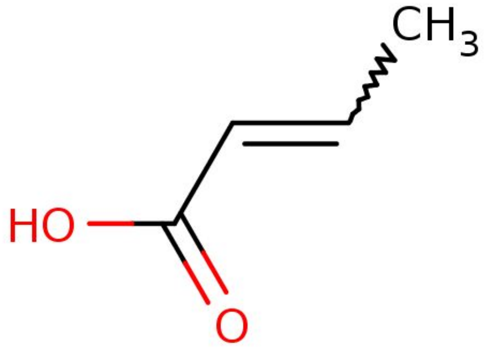
A molecular design pipeline

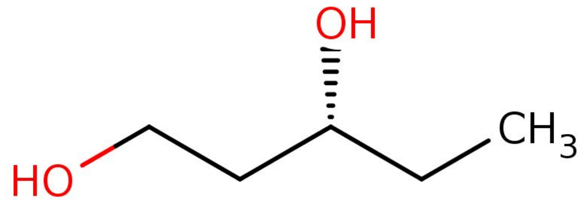
Efficiently explore molecule space

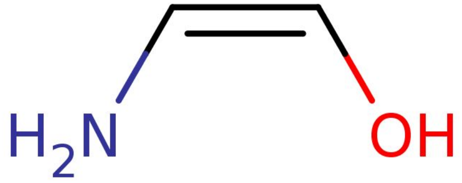
- **Large** library of candidates
- **Expensive** experiments (<10)
- High degree of **parallelism**
- Want molecules with high **affinity**
  - Also easy to make
  - Don’t stick to themselves
  - Stable
  - In a new area of “patent space”









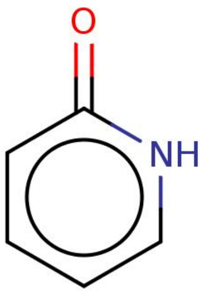


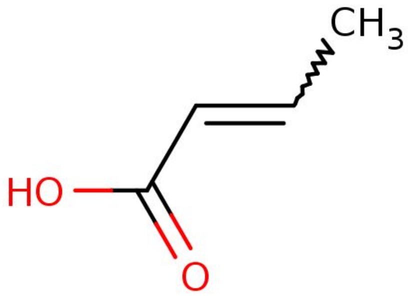
????

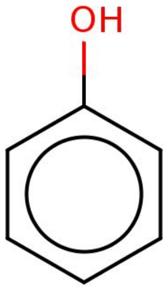
Any ideas?

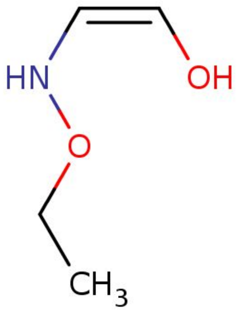
A Simpler Example

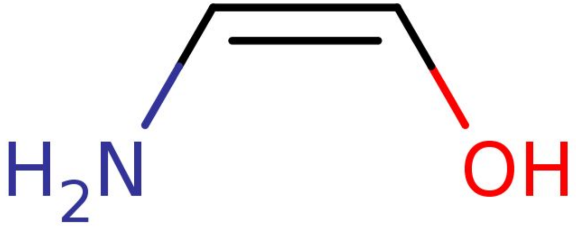
Can evaluate **at most** 4

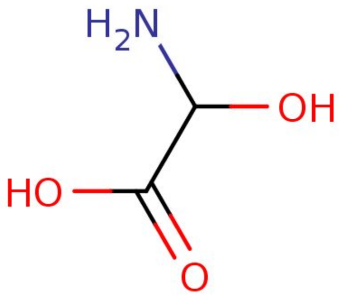






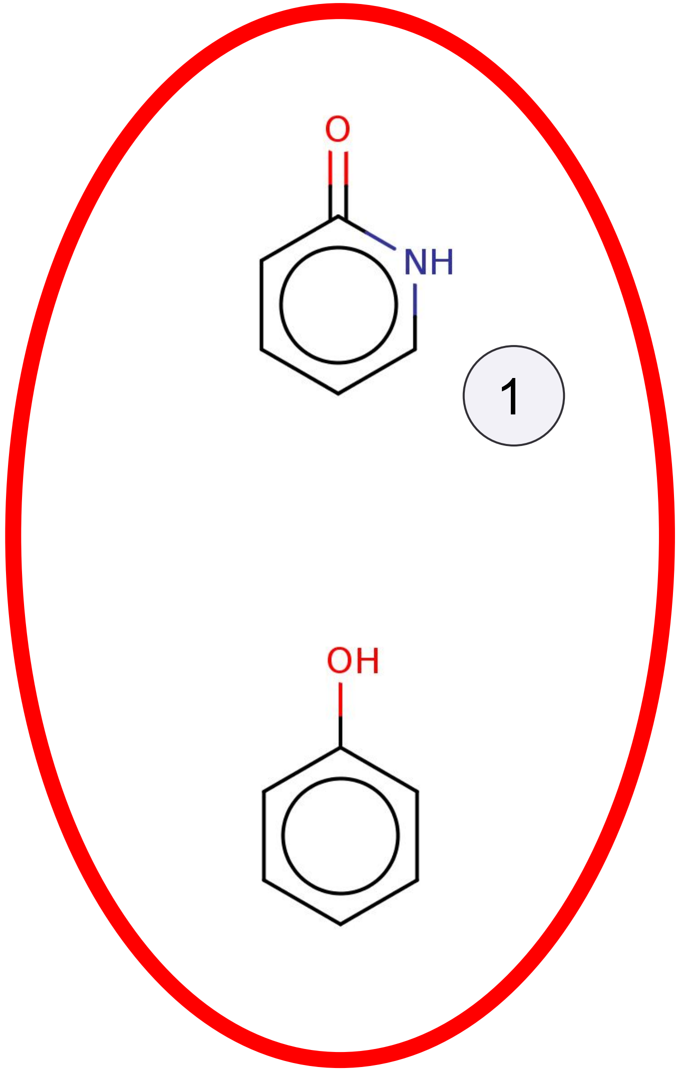




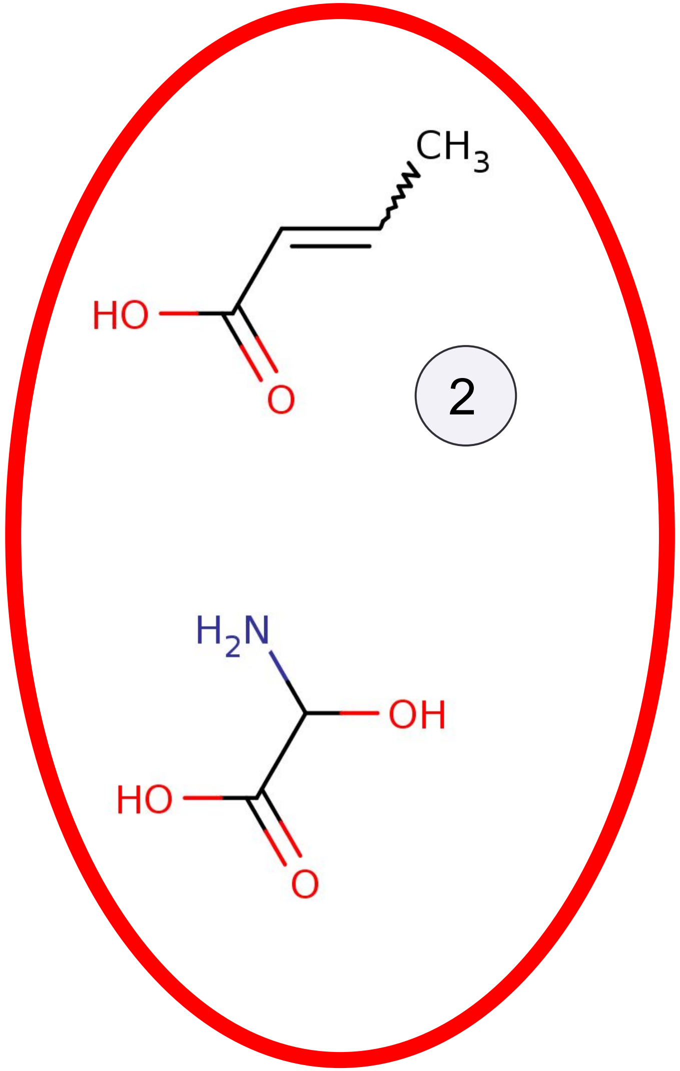


A Simpler Example (grouped)

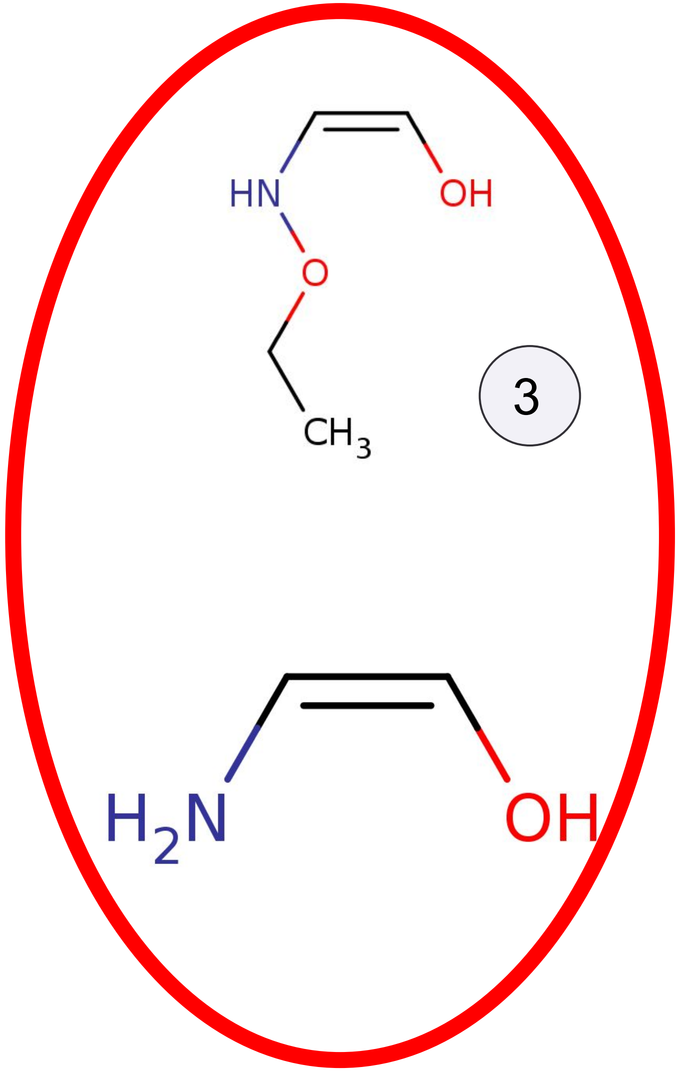
Can evaluate **at most** 4



1



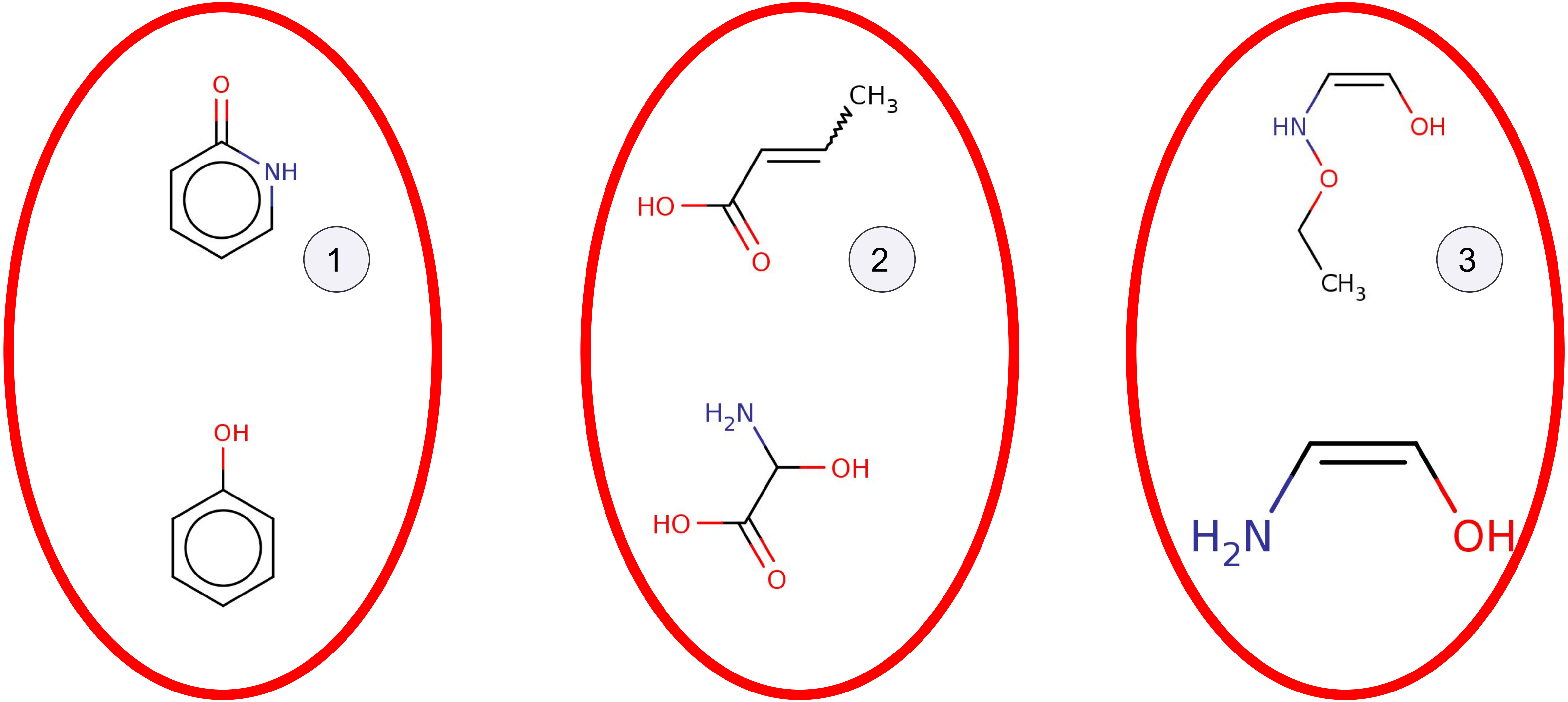
2



3

A Simpler Example (grouped)

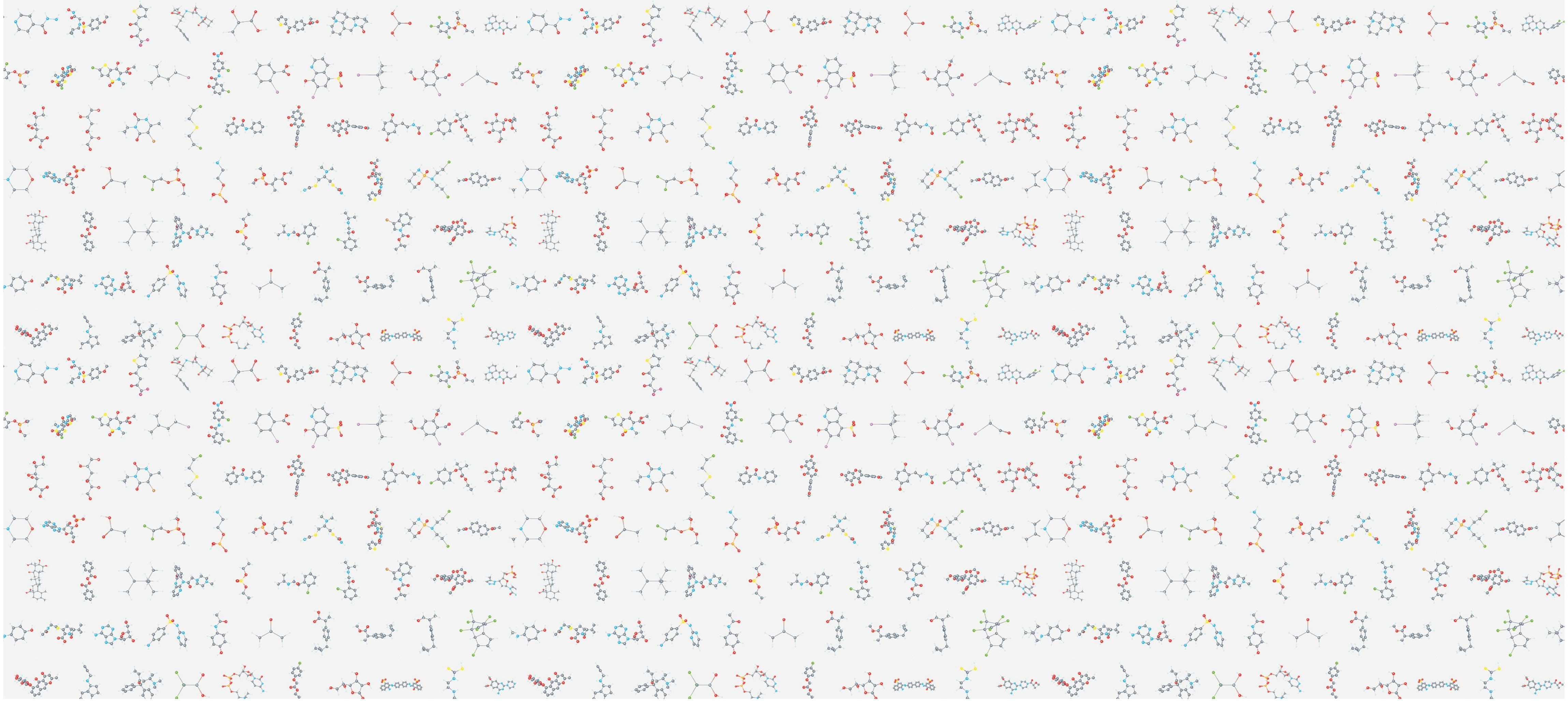
Can evaluate **at most** 4



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?

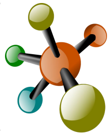
eek

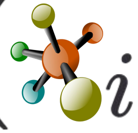
Use a GP!

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

An Aside: GPs for Molecules

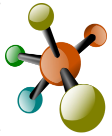
Structured Input Spaces

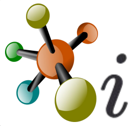
$y_i = f($  $_i) + \epsilon_i$

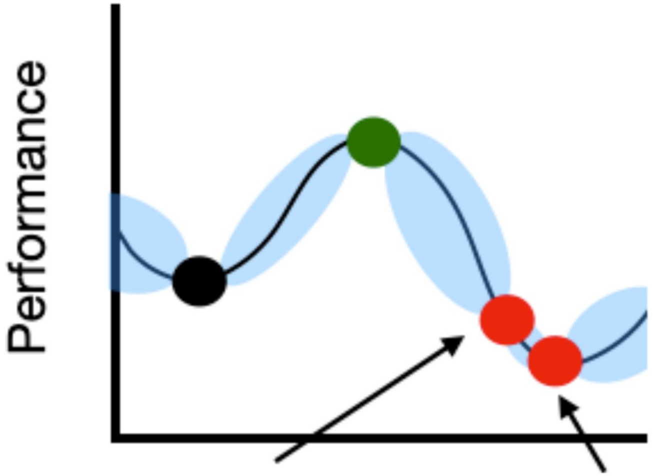
$D_N = \{($  $_i , y_i)\}_i^N$

An Aside: GPs for Molecules

Structured Input Spaces

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

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

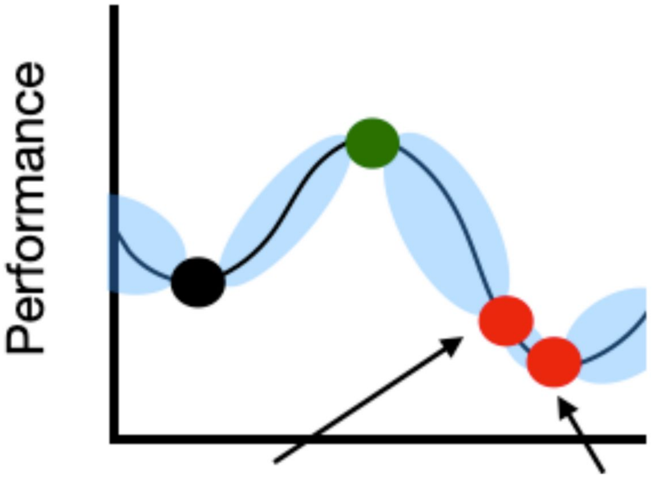


What do we require to define a GP?

An Aside: GPs for Molecules

Structured Input Spaces

$$y_i = f(\;_i) + \epsilon_i \qquad D_N = \{(\;_i, y_i)\}_i^N$$



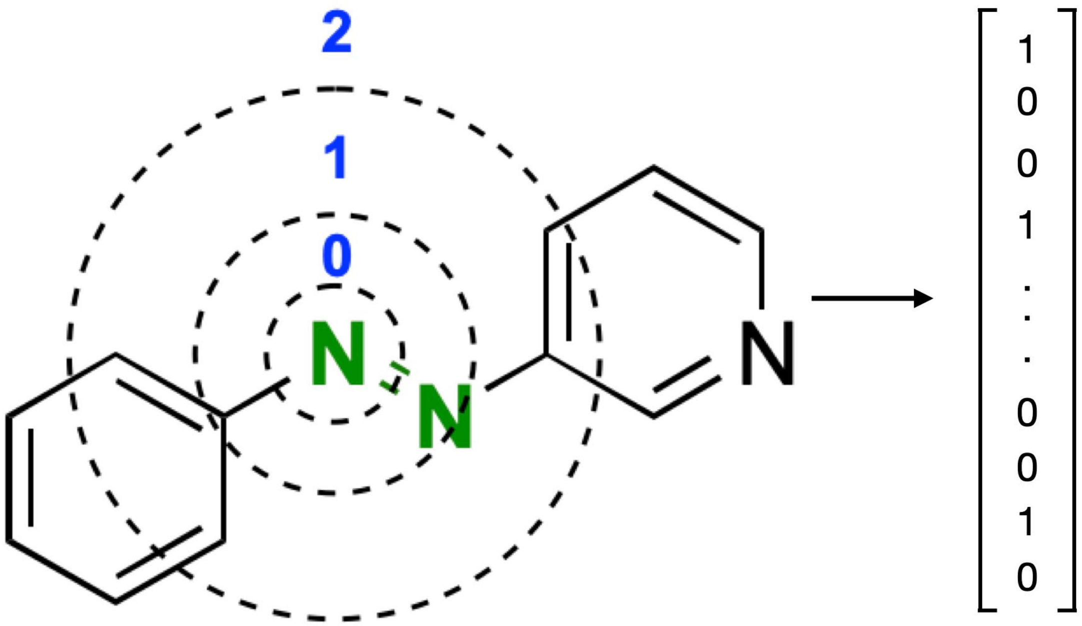
$$k(\;_i, \;_j) = ?$$

What do we require to define a GP?

An Aside: GPs for Molecules

Fingerprint Kernels

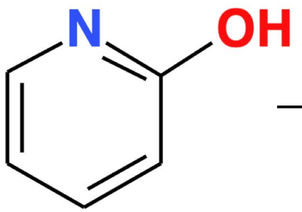
$$k(\ _i,\ _j) = k_{\mathrm{linear}}(\Phi(\ _i), \Phi(\ _j))$$



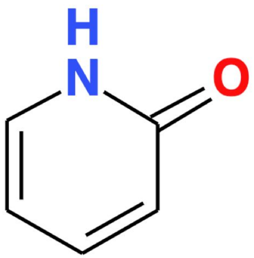
An Aside: GPs for Molecules

String kernels between SMILES strings

$$k(\;_i,\;_j) \;=\; k(str(\;_i),\; str(\;_j))$$



$\longrightarrow$ $Oc1ncccc1$



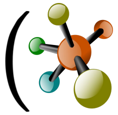
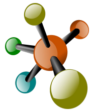
$\longrightarrow$ $O = c1[nH]cccc1$

Automatically choosing next molecules

Using GP posteriors and utility functions

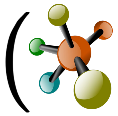
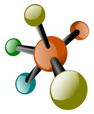
Automatically choosing next molecules

Using GP posteriors and utility functions

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

Automatically choosing next molecules

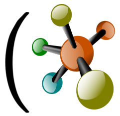
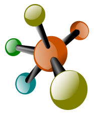
Using GP posteriors and utility functions

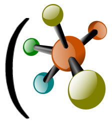
- $U_f($$)$: what is the utility of evaluating  (if it will return f)

  - $f^\star$ Is best so far

Automatically choosing next molecules

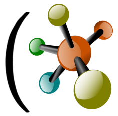
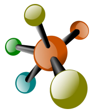
Using GP posteriors and utility functions

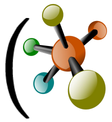
- $U_f($  $)$: what is the utility of evaluating  (if it will return f)

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

Automatically choosing next molecules

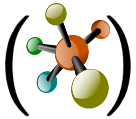
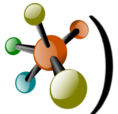
Using GP posteriors and utility functions

- $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)$

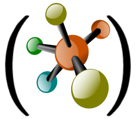
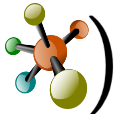
Automatically choosing next molecules

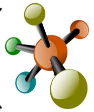
Using GP posteriors and utility functions

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

Automatically choosing next molecules

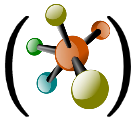
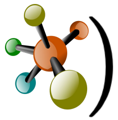
Using GP posteriors and utility functions

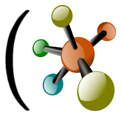
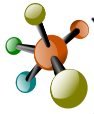
- $\alpha($  $) = \mathbb{E}_f[U_f($  $)]$: what utility is predicted by my model of f

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

Automatically choosing next molecules

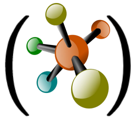
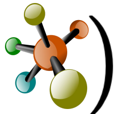
Using GP posteriors and utility functions

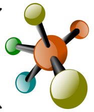
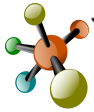
- $\alpha($  $) = \mathbb{E}_f[U_f($  $)]$: what utility is predicted by my model of f

  - 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

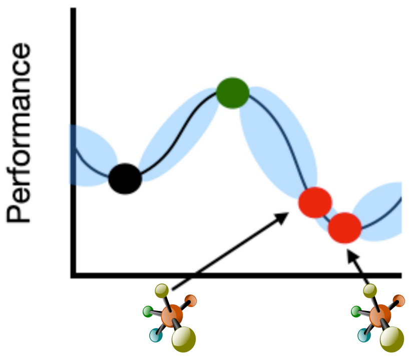
- $\alpha($  $) = \mathbb{E}_f[U_f($  $)]$: what utility is predicted by my model of f

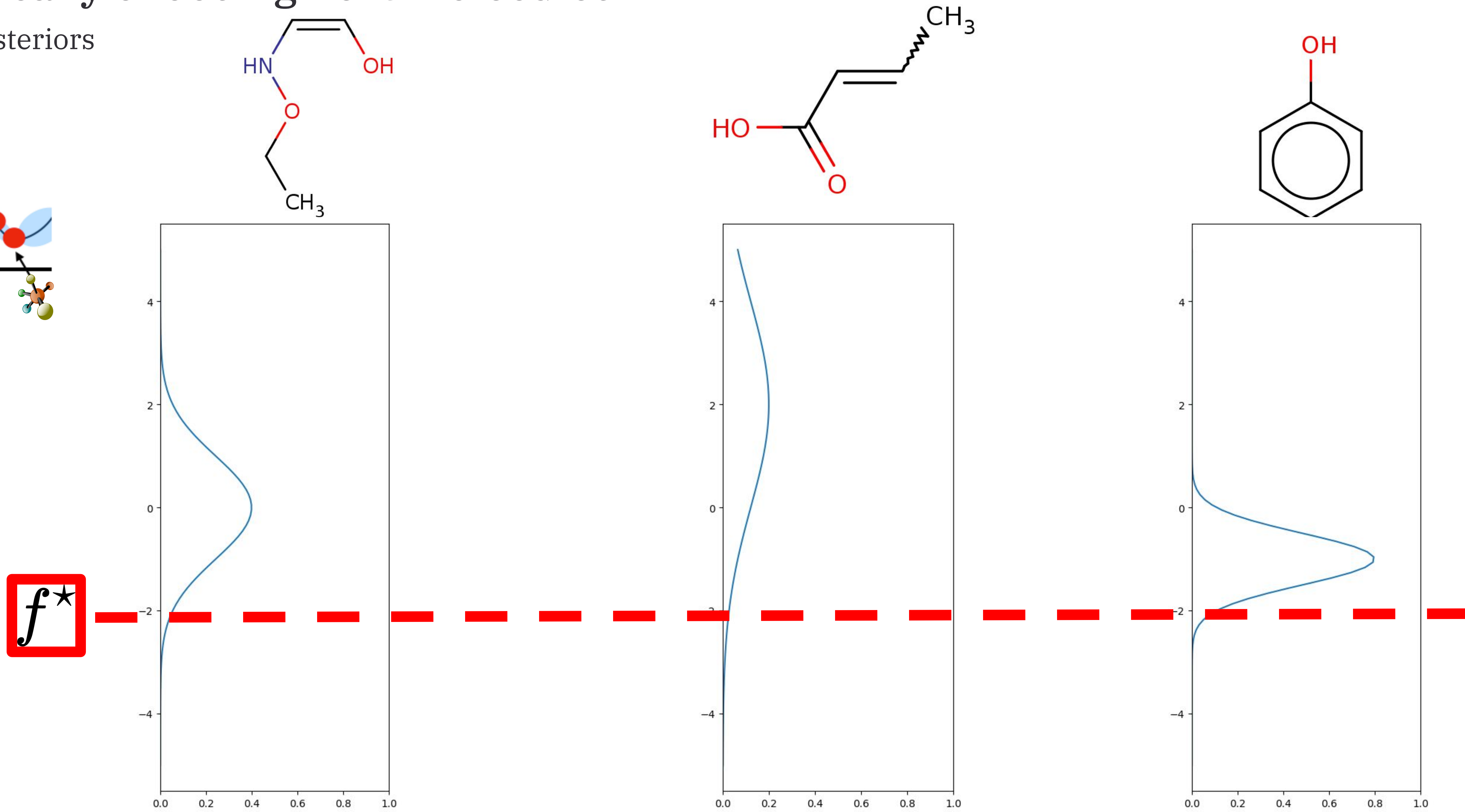
    - What the probability of improvement? $\alpha_{\mathrm{PI}}($  $) = \mathbb{E}_f\left[\mathbb{1}_{(f>f^\star)}\right]$
    - How much improvement do we expect? $\alpha_{\mathrm{EI}}($  $) = \mathbb{E}_f[\max(f - f^\star, 0)]$

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

Automatically choosing next molecules

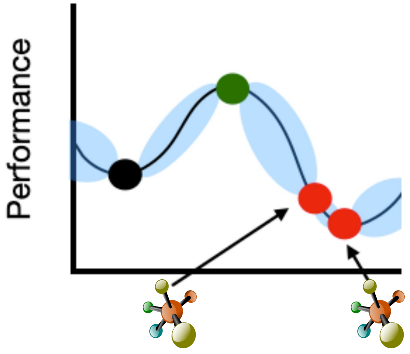
Using GP posteriors

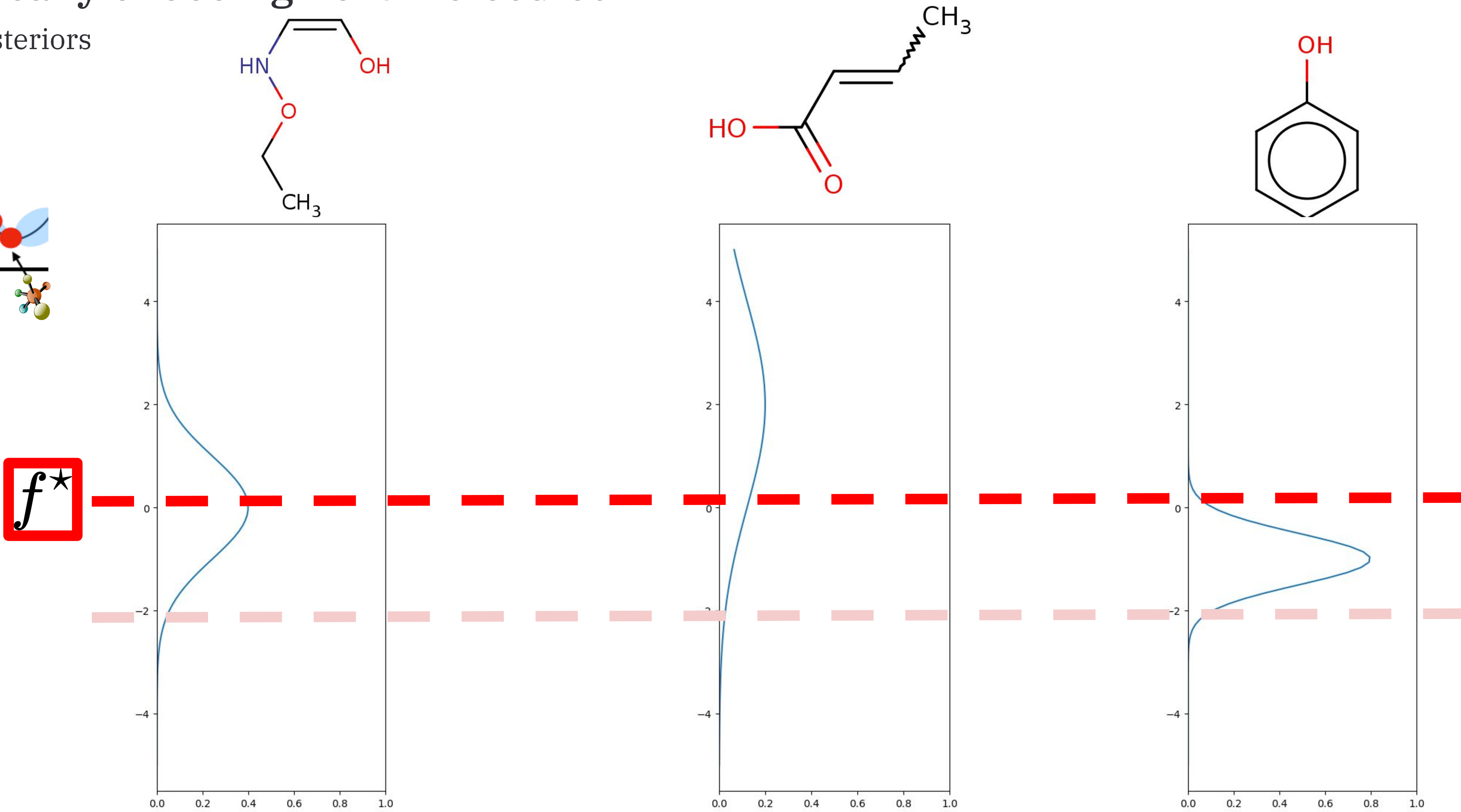




Automatically choosing next molecules

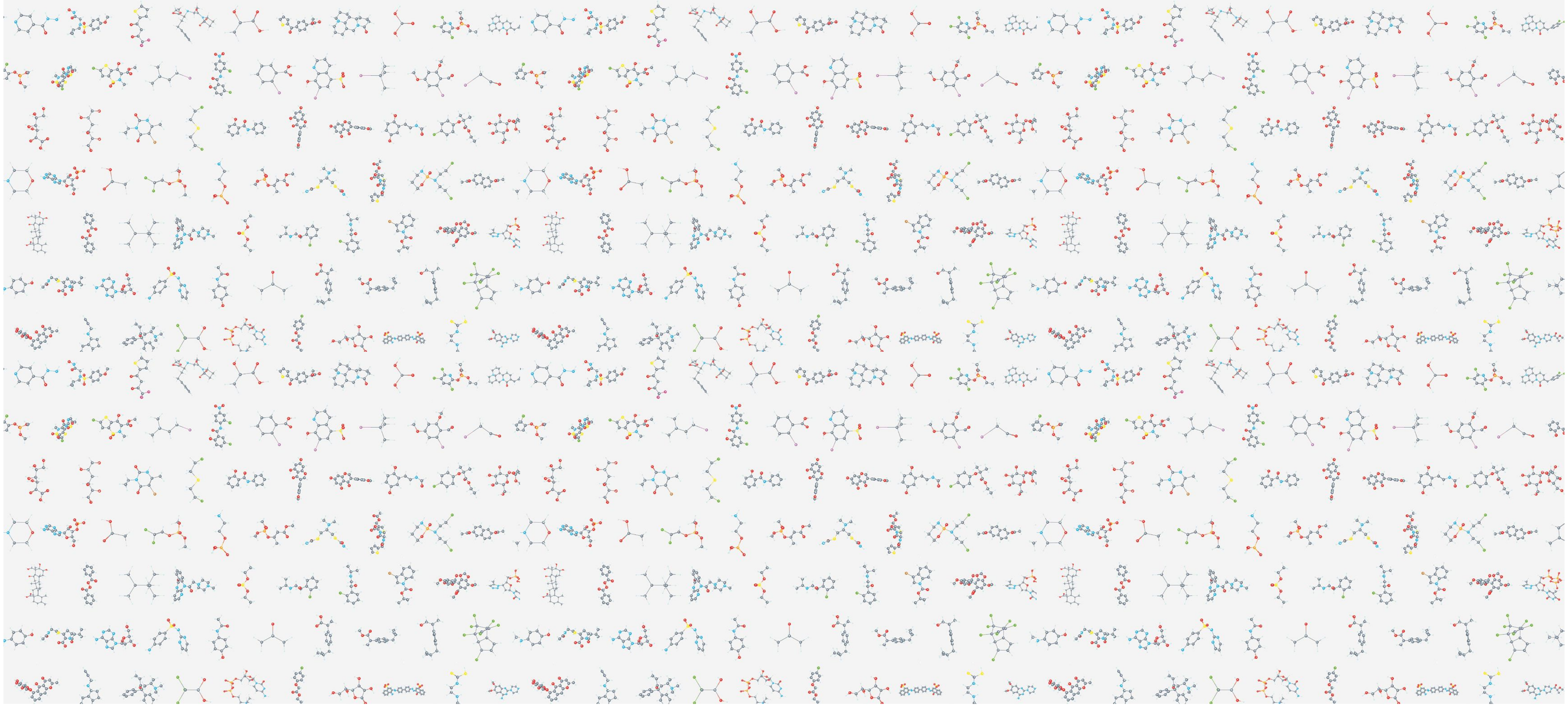
Using GP posteriors





Automatically choosing next molecules

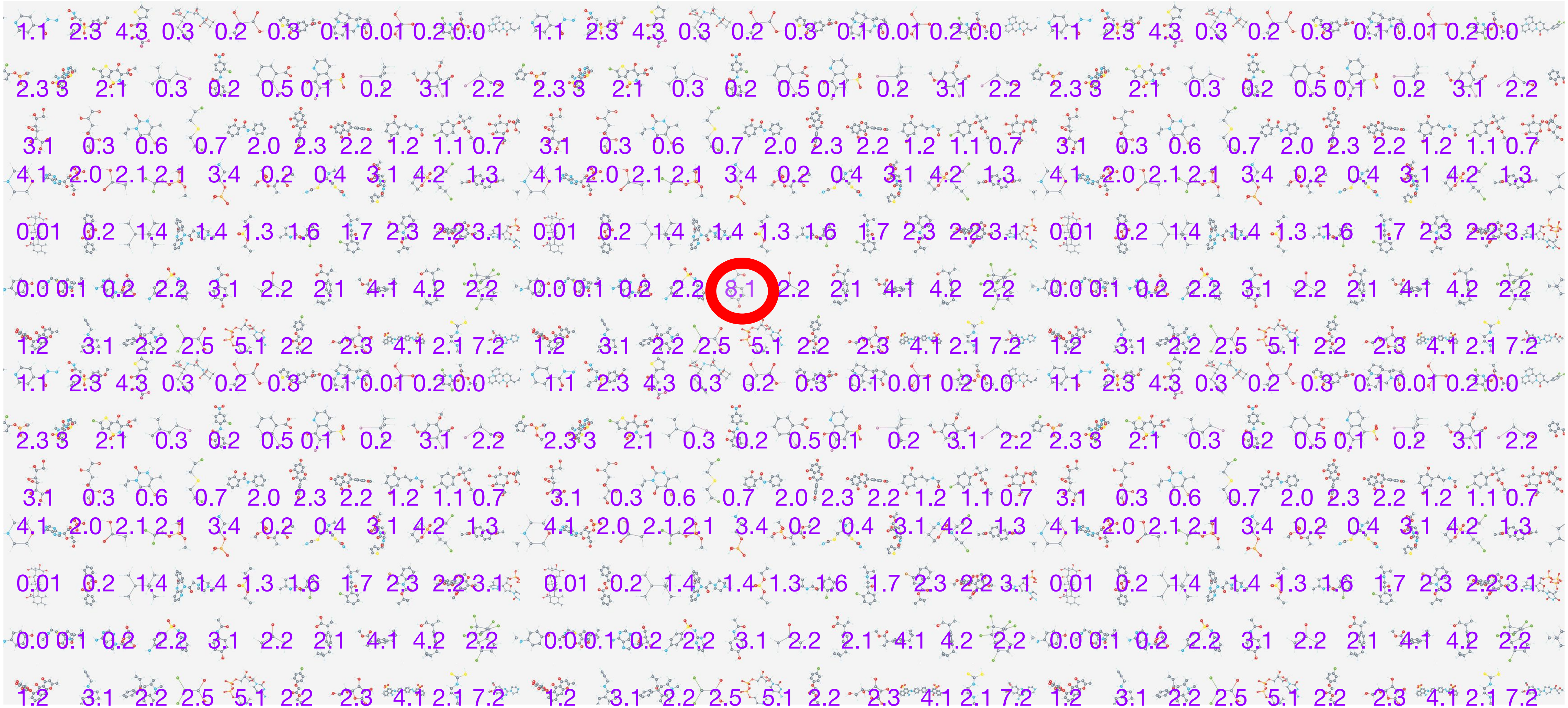
Calc acquisition function and pick best



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

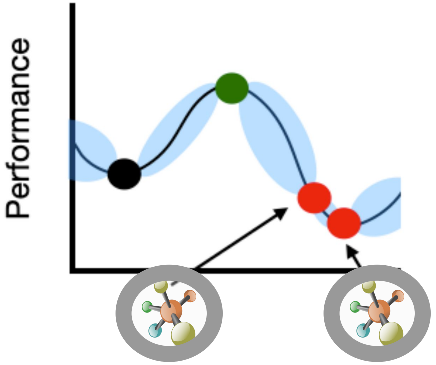
https://www.horizons-mag.ch/2018/09/06/the-universe-of-molecules/

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/

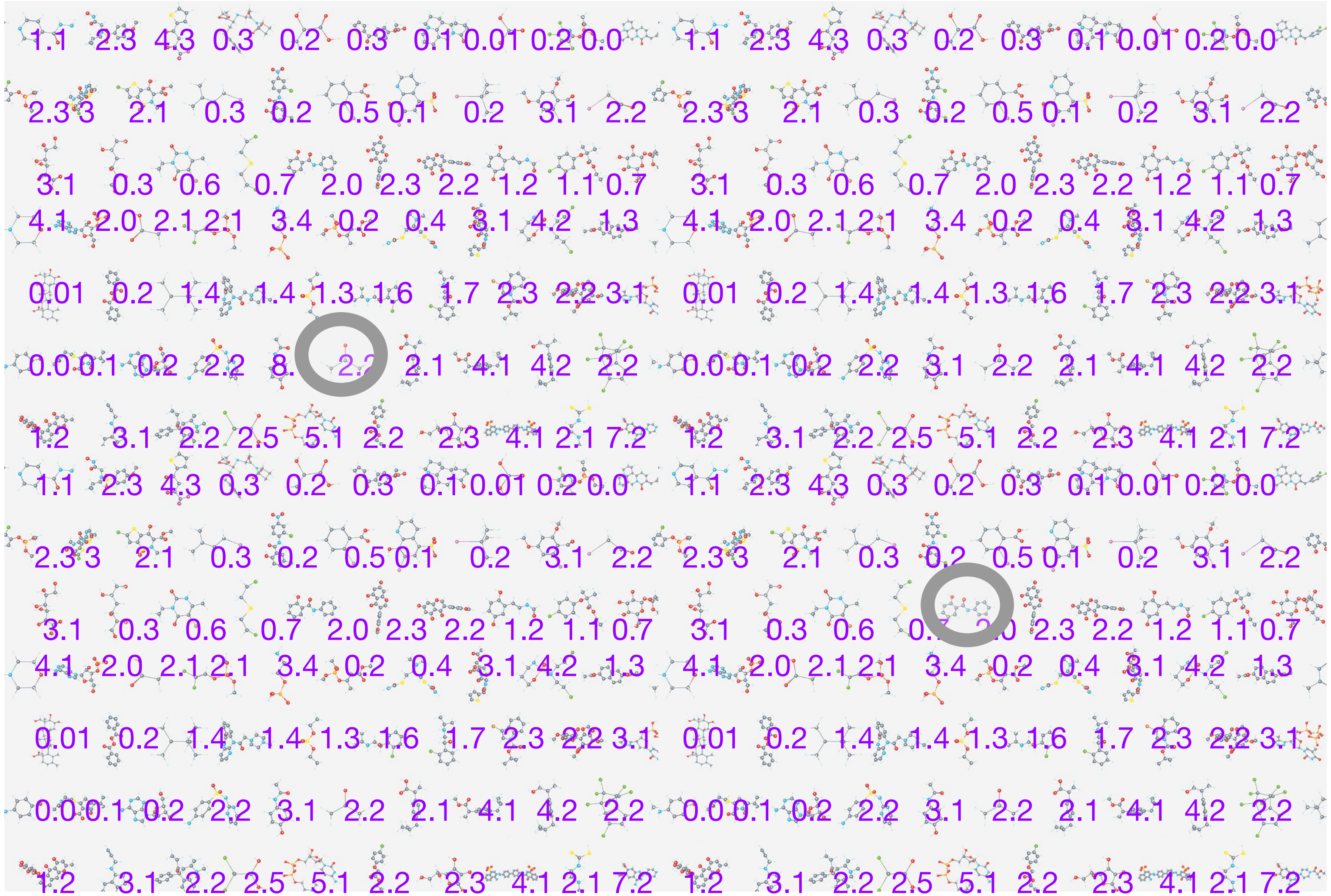
UNIVERSITY OF CAMBRIDGE

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

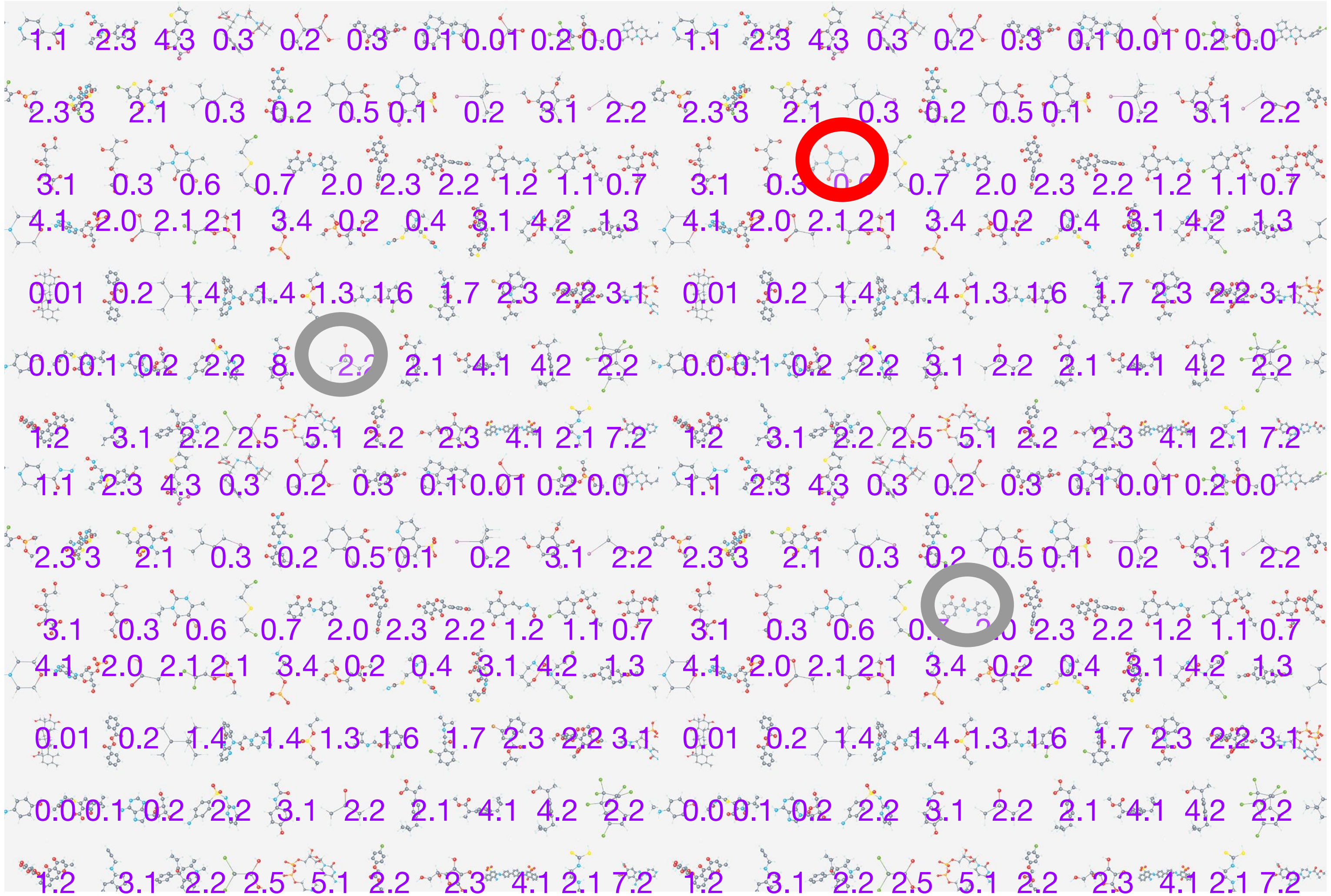


https://www.horizons-mag.ch/2018/09/06/the-universe-of-molecules/

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

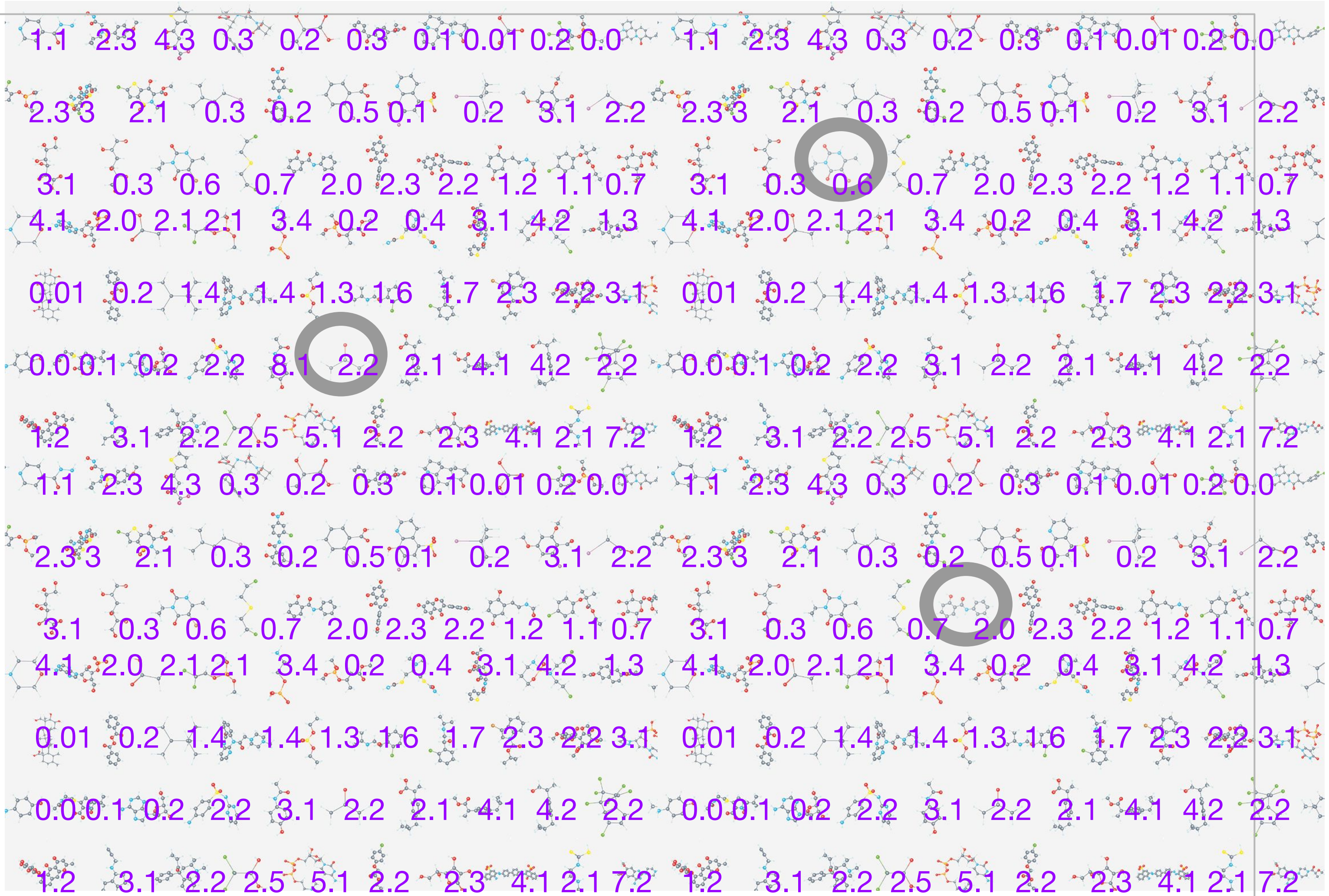
1. Evaluate 2 random molecules
2. Fit GP model to measurements
3. Calc acquisition function
4. Choose new molecule
5. Go to step 2.

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/

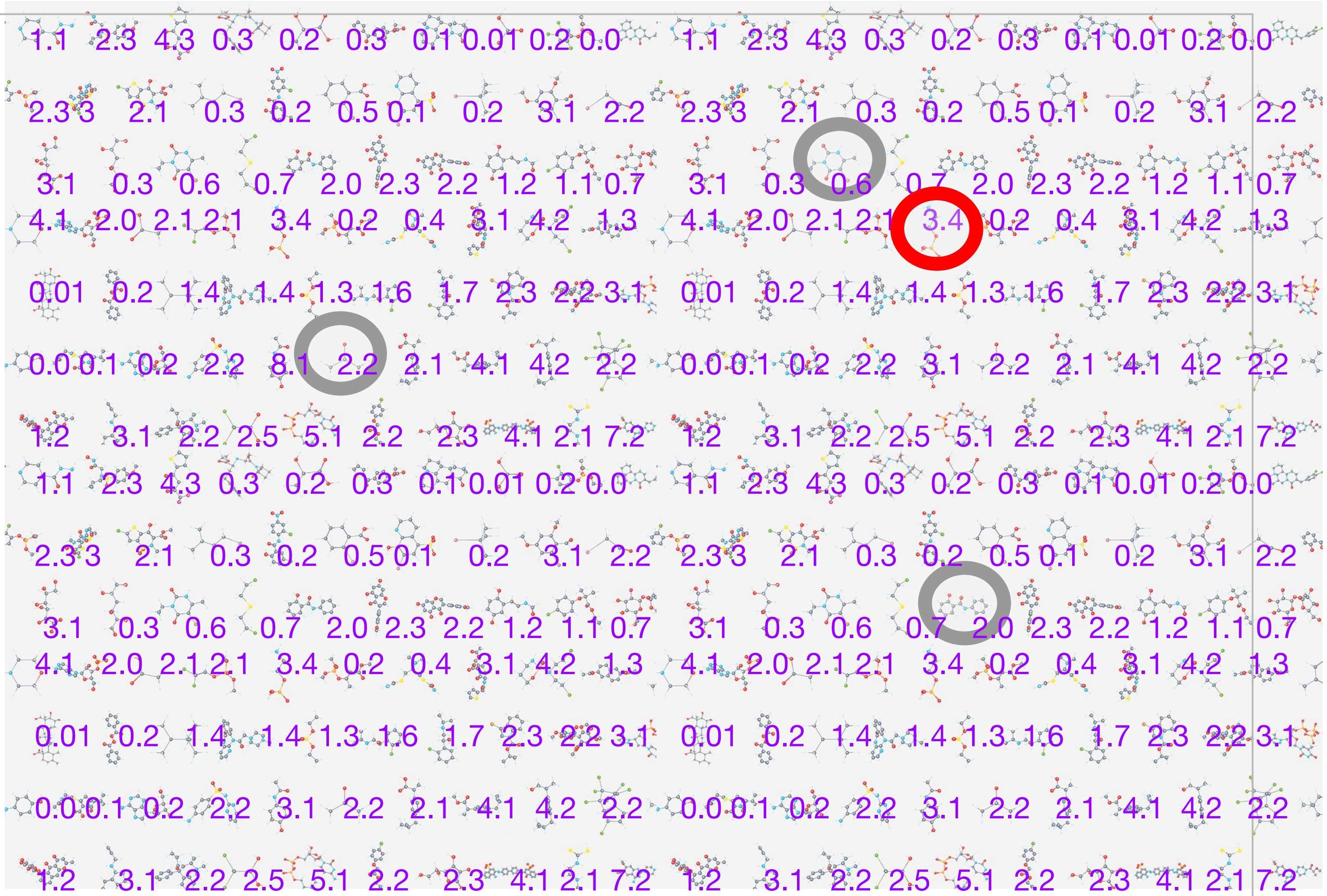
UNIVERSITY OF CAMBRIDGE

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.

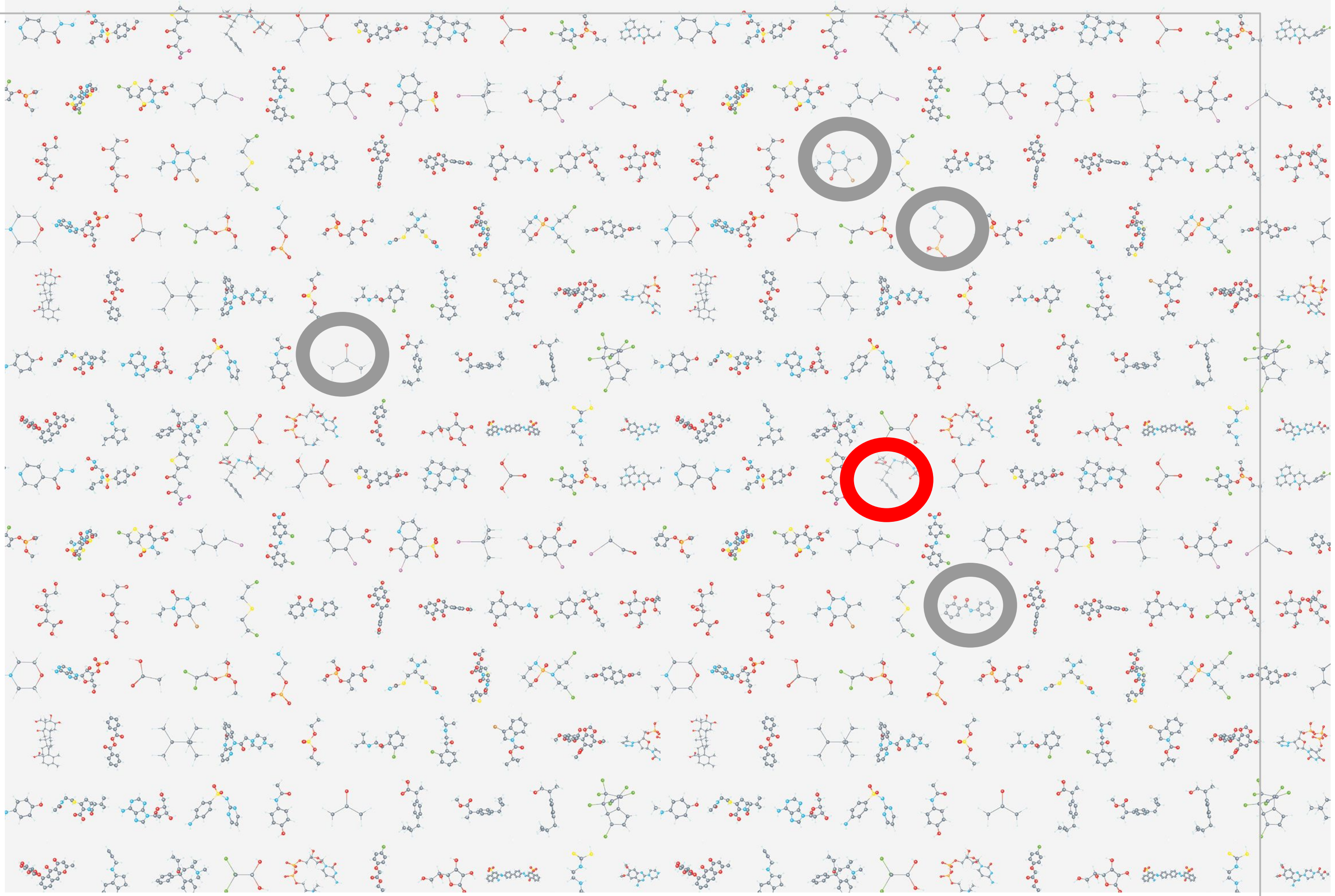


https://www.horizons-mag.ch/2018/09/06/the-universe-of-molecules/

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/

Automatically choosing next molecules

Full Bayesian optimisation loop

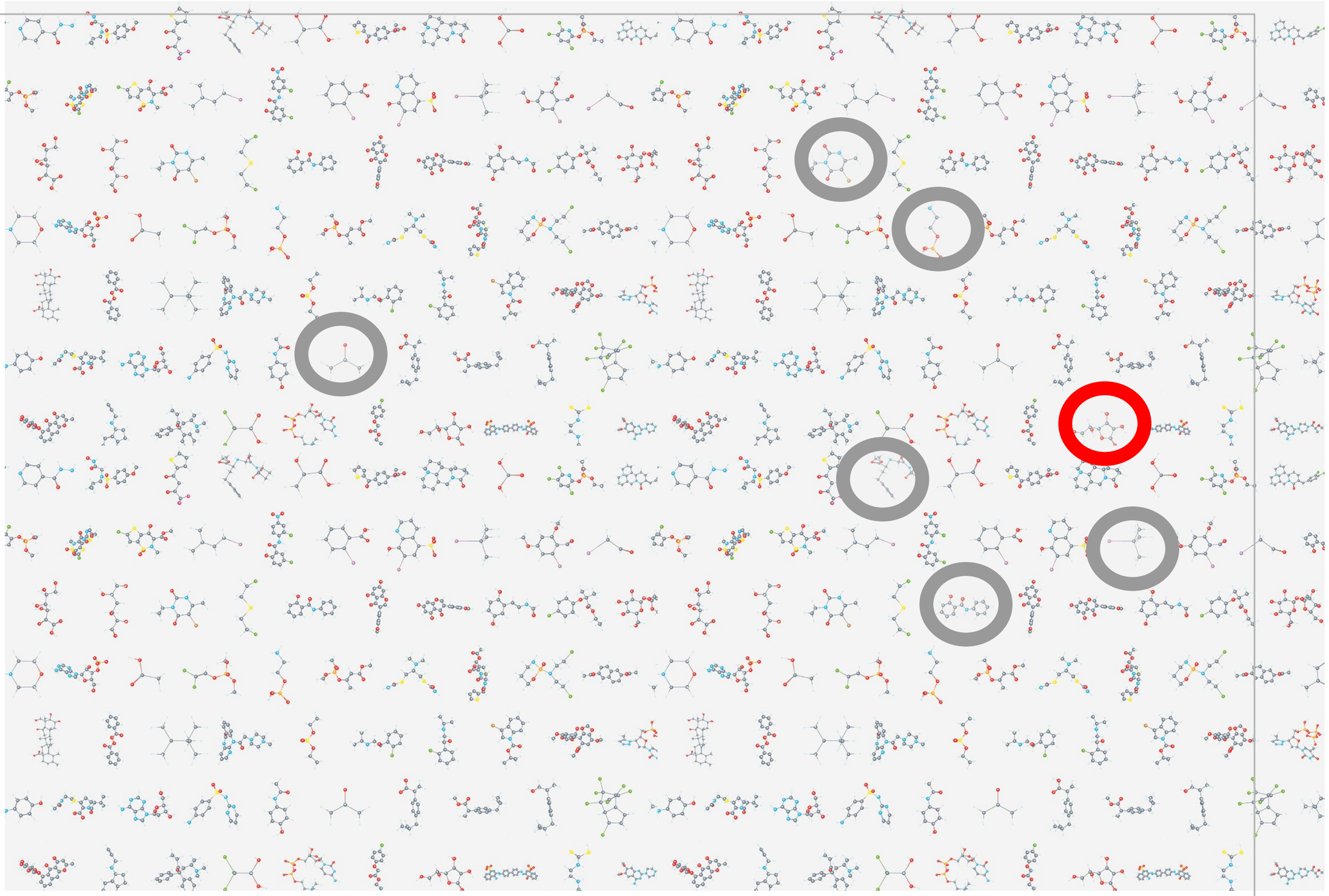
1. Evaluate 2 random molecules
2. Fit GP model to measurements
3. Calc new acquisition function
4. Choose new molecule
5. Go to step 2.

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.

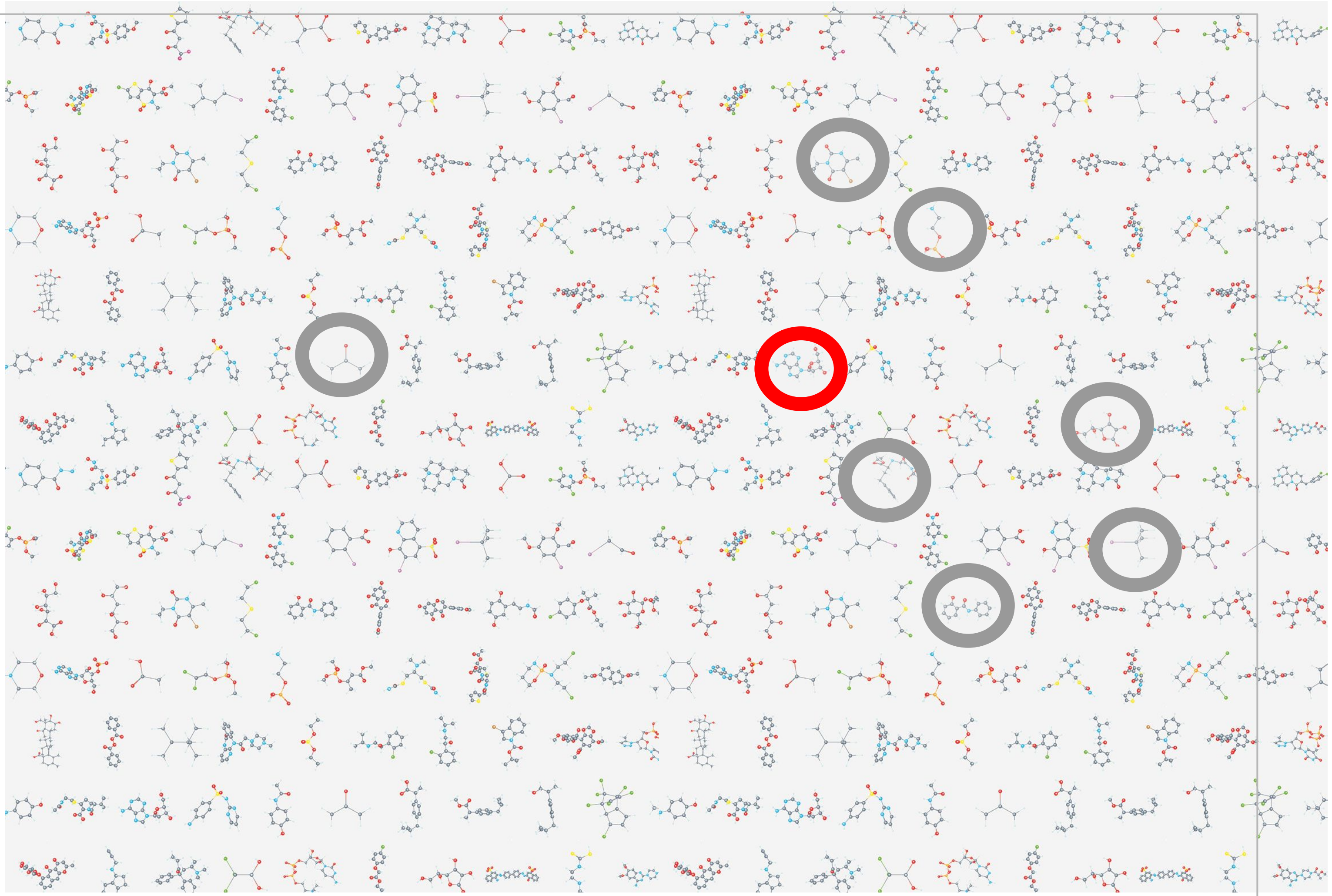


https://www.horizons-mag.ch/2018/09/06/the-universe-of-molecules/

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.



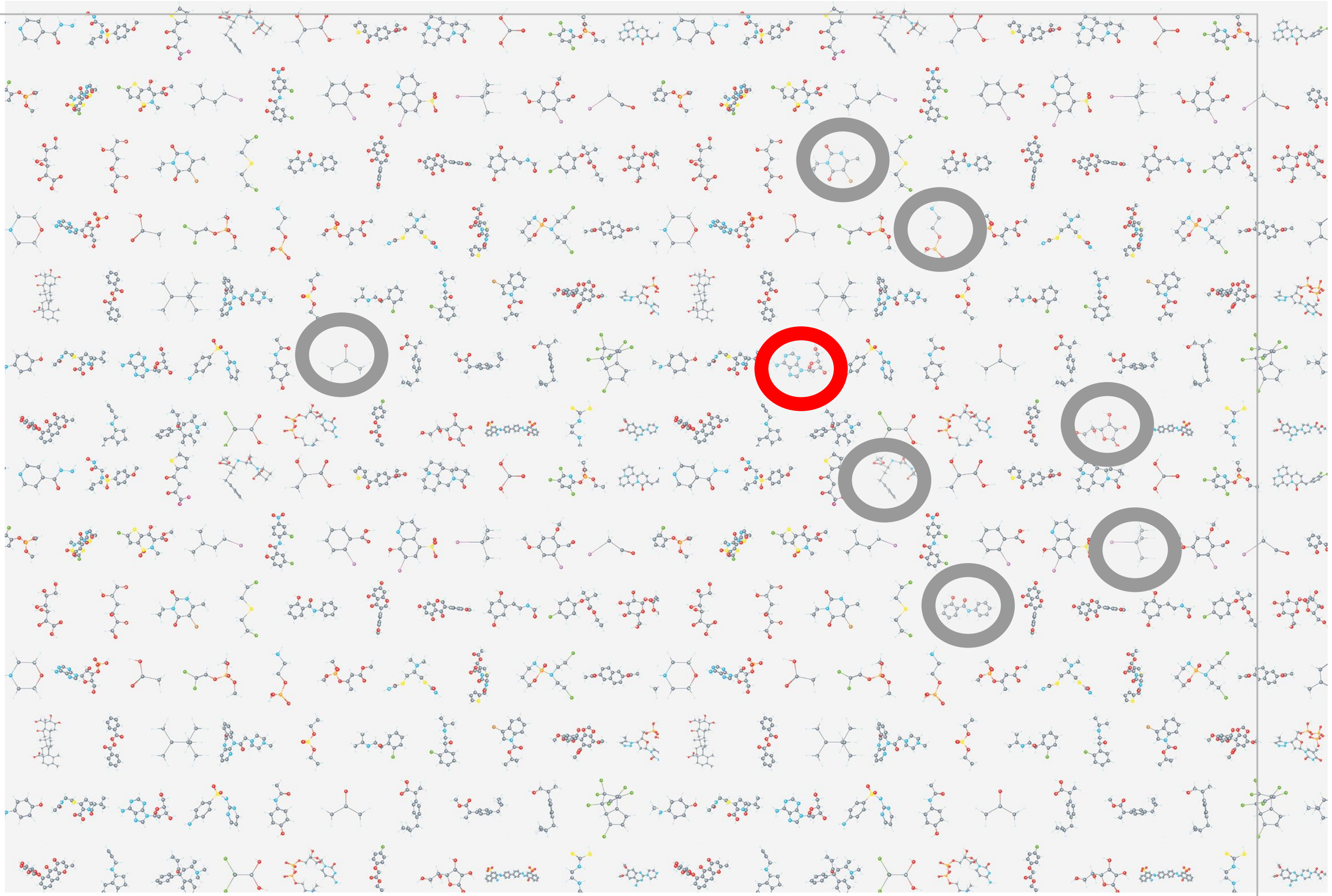
https://www.horizons-mag.ch/2018/09/06/the-universe-of-molecules/

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

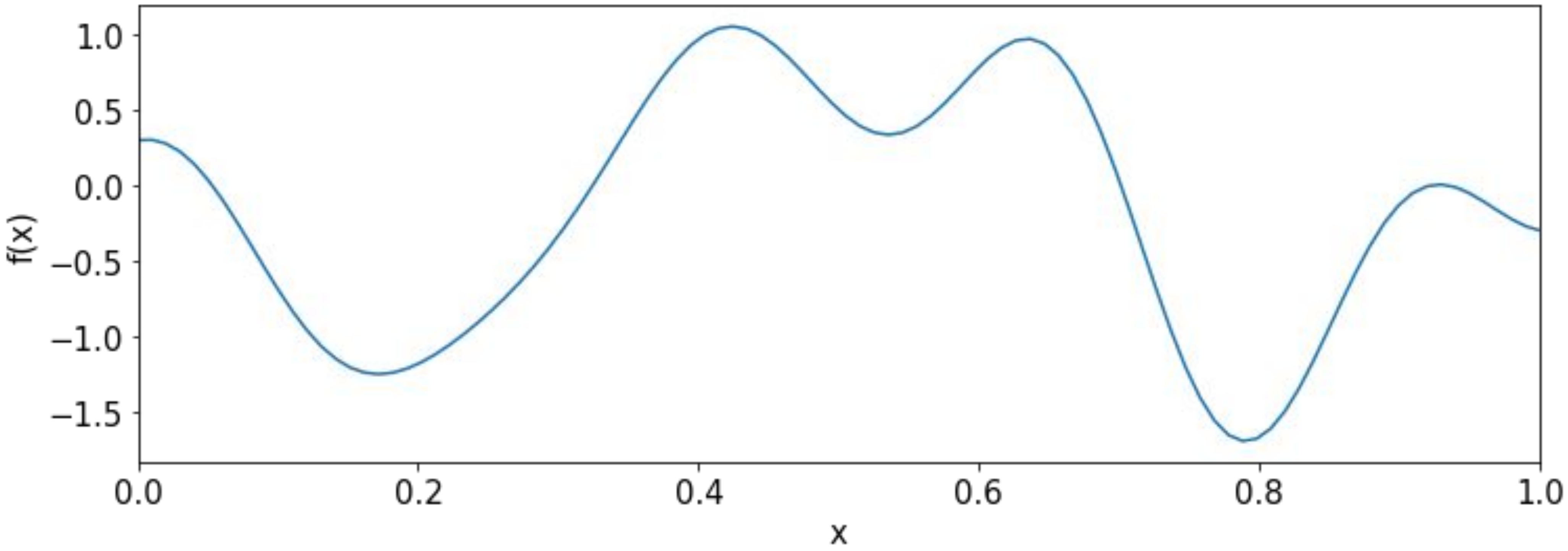
Let's find the maximum of a 1D function:

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

BO Demo

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

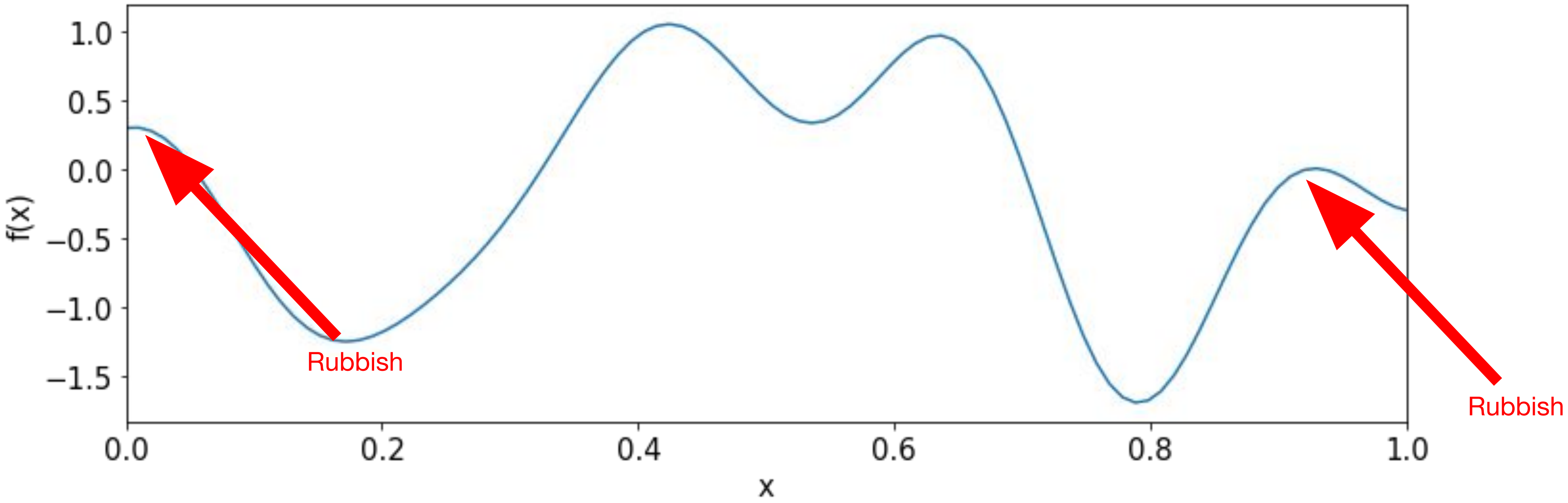
Using as **few** function evaluations as possible!



BO Demo

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

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

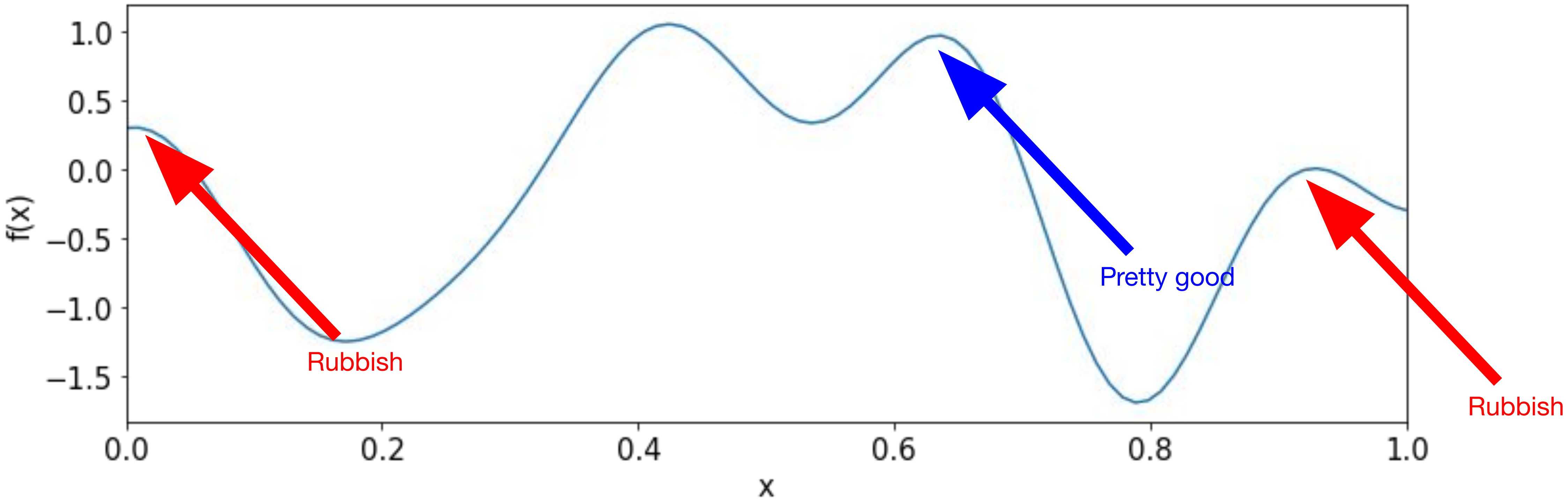


UNIVERSITY OF CAMBRIDGE Lancaster University

BO Demo

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

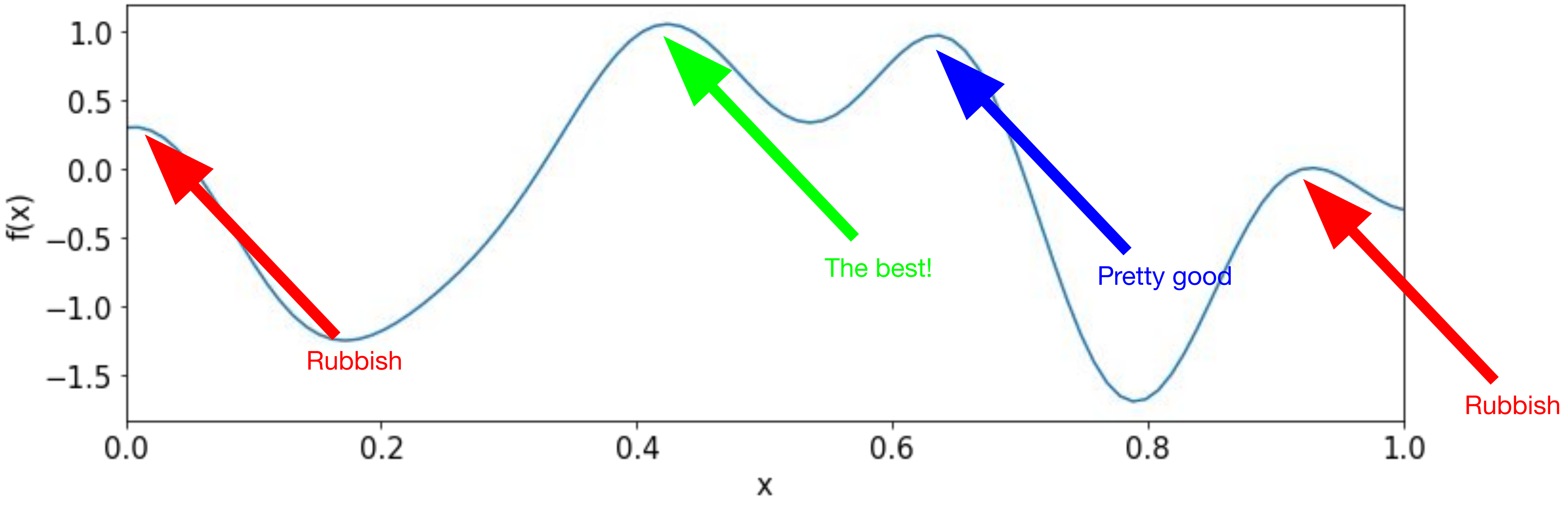
Using as **few** function evaluations as possible!



BO Demo

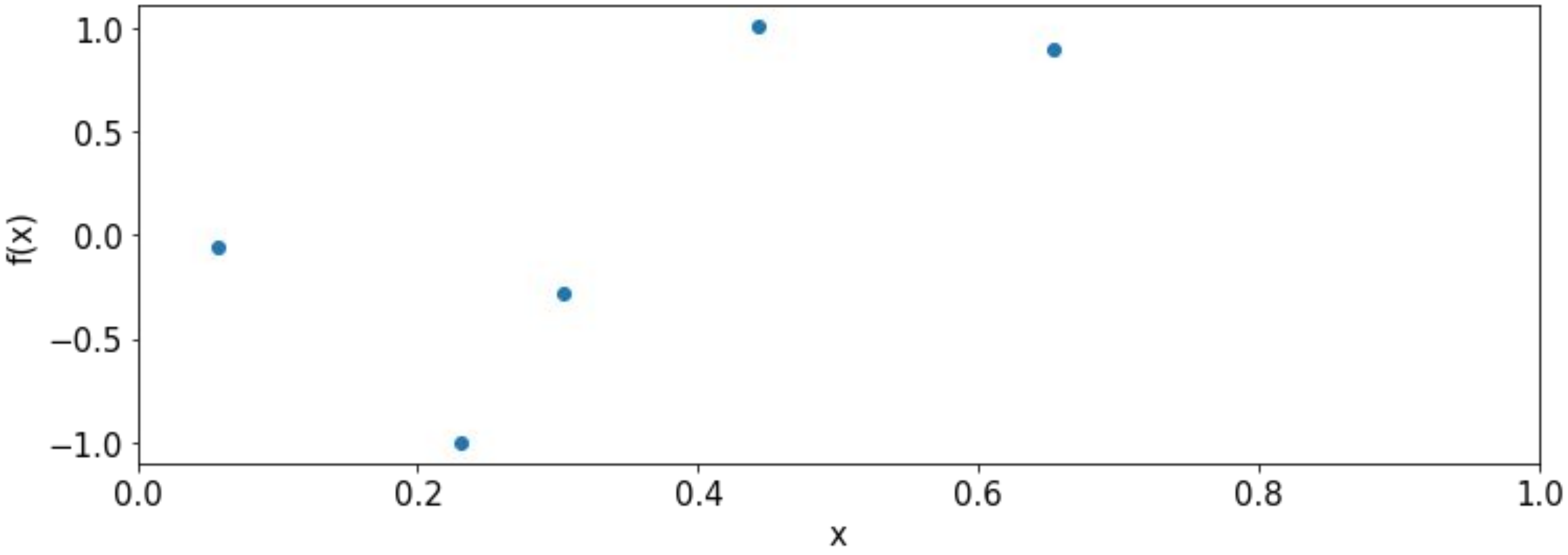
Let's find the maximum of a 1D function:

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



BO Demo

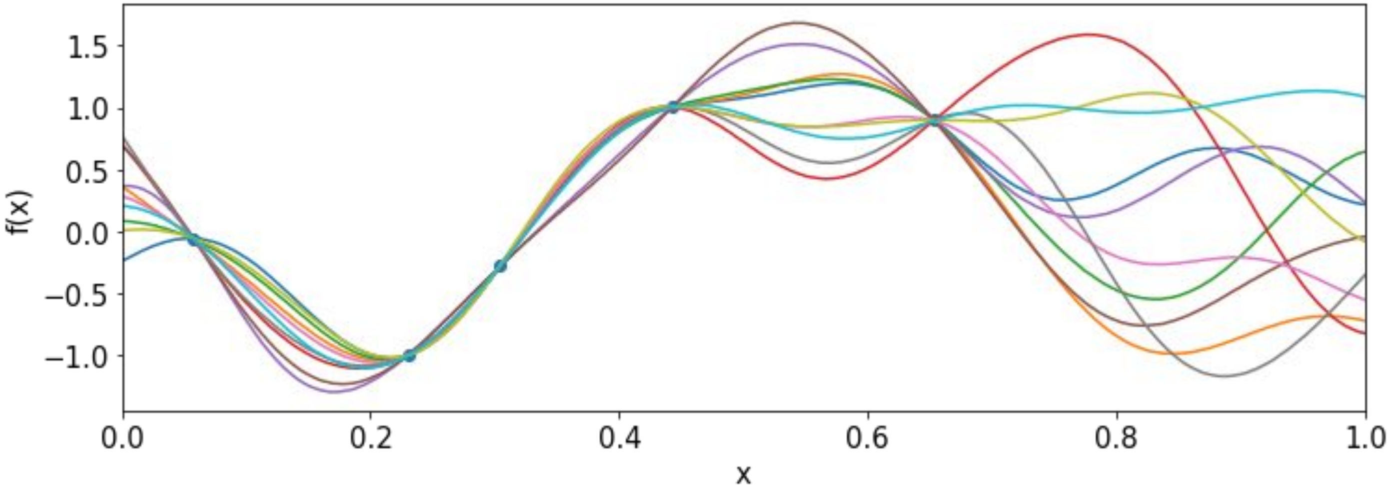
Suppose we make 5 evaluations

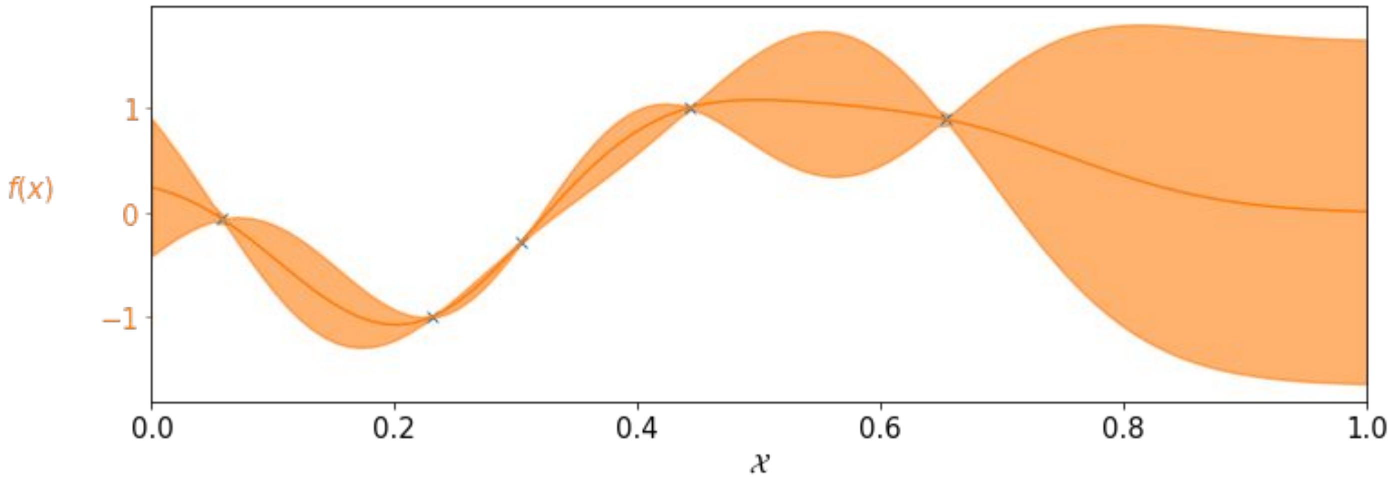


Where should we next evaluate? Explore/Exploit?

How to automate BO: step 1

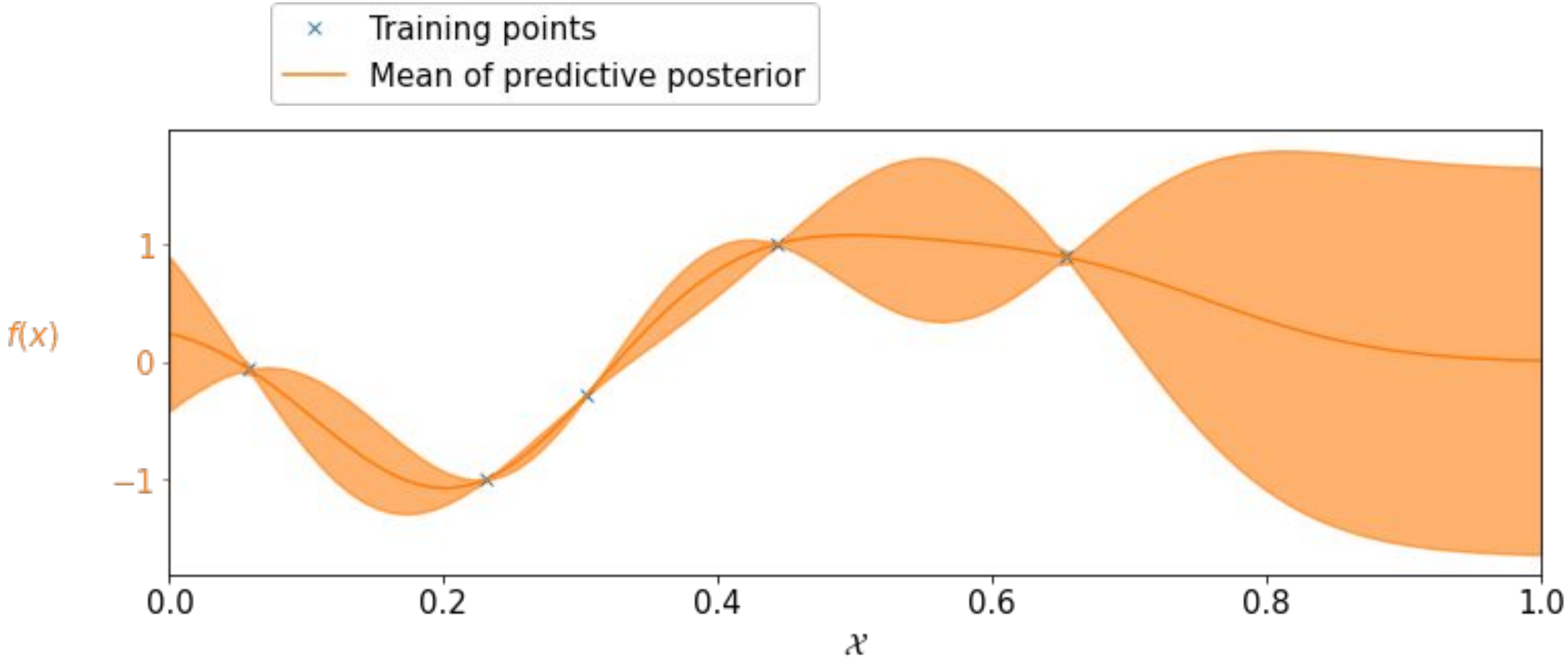
Use a statistical model like a Gaussian process





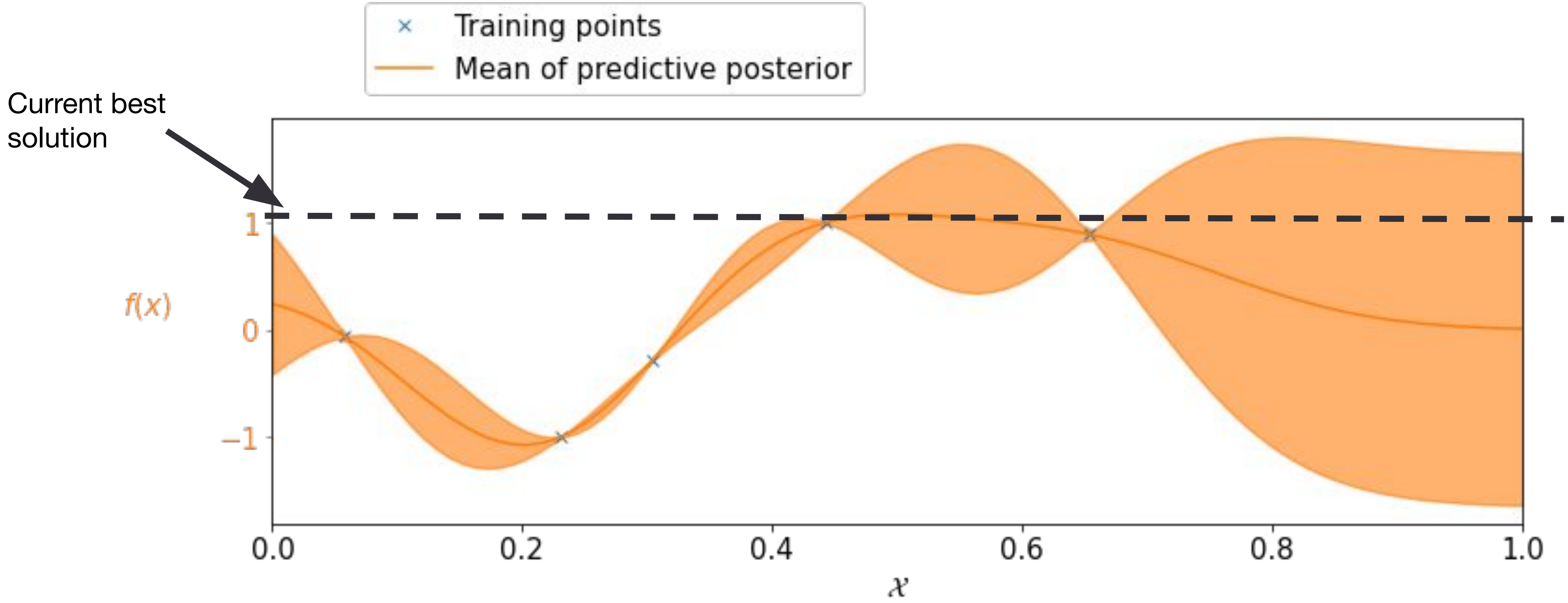
How to automate BO: step 2

Automated decision making via an acquisition function like expected improvement



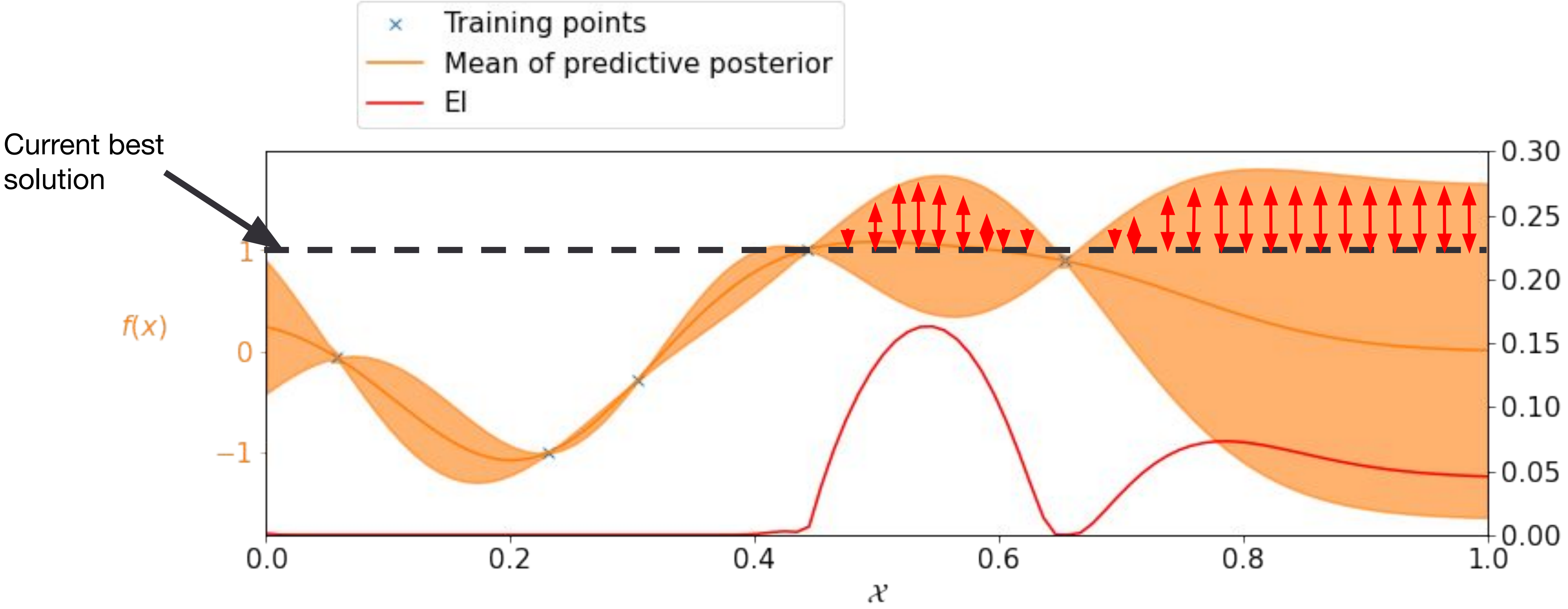
How to automate BO: step 2

Automated decision making via an acquisition function like expected improvement



How to automate BO: step 2

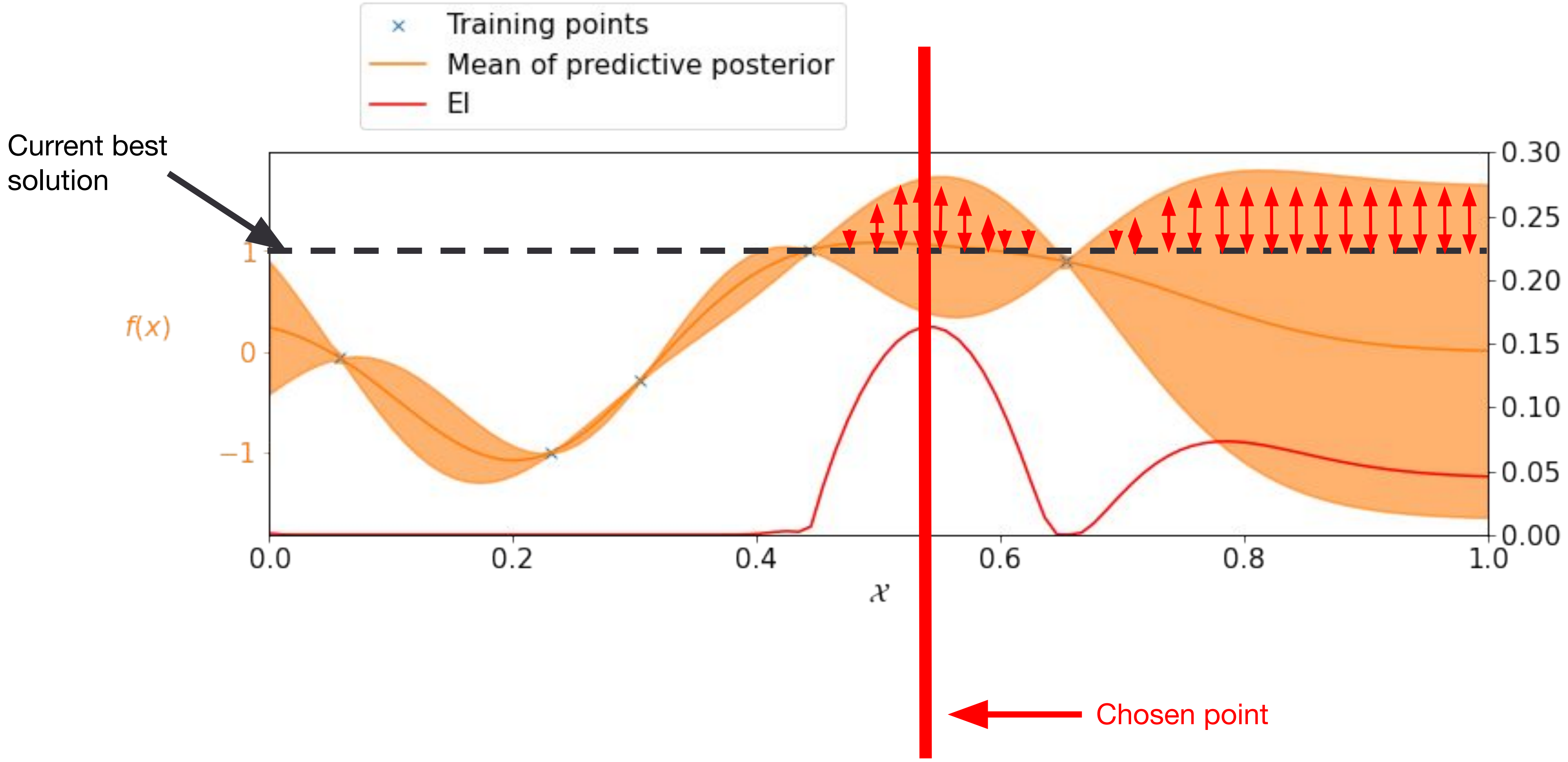
Automated decision making via an acquisition function like expected improvement



UNIVERSITY OF CAMBRIDGE Lancaster University

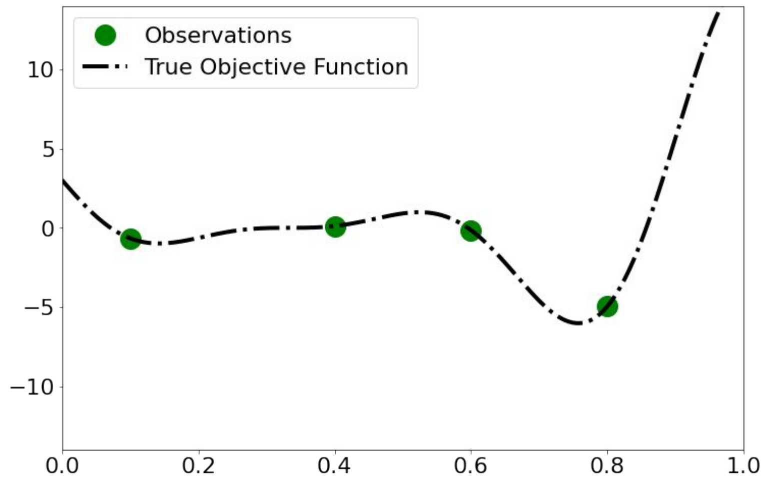
How to automate BO: step 2

Automated decision making via an acquisition function like expected improvement



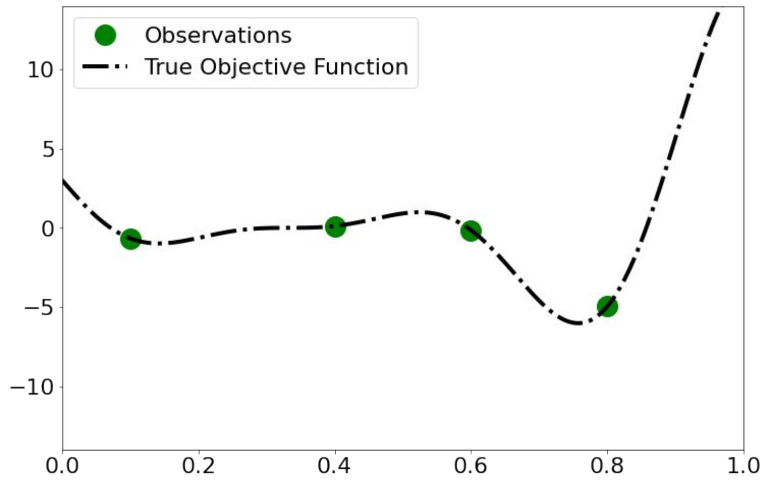
Expected Improvement

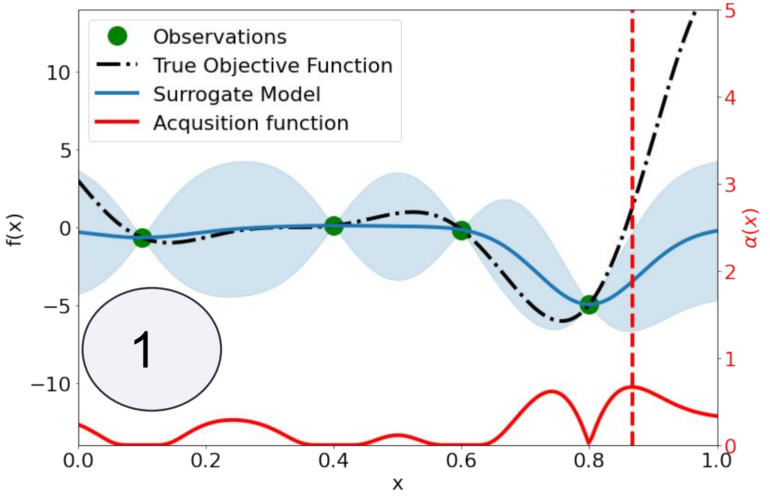
Demo BO loop



Expected Improvement

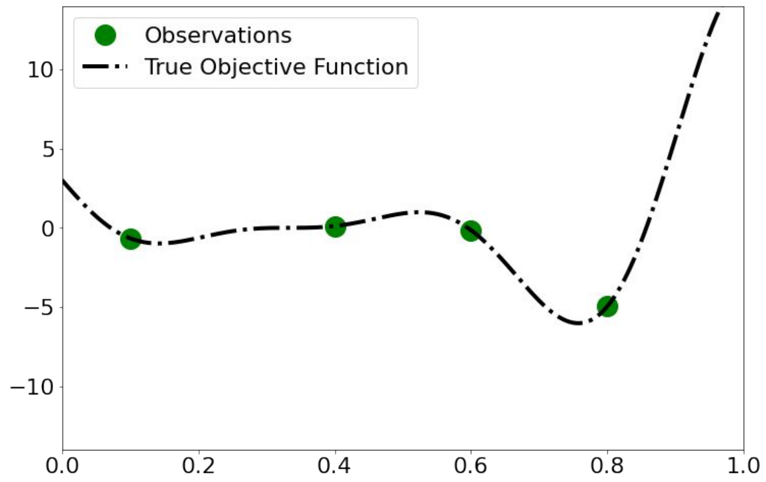
Demo BO loop

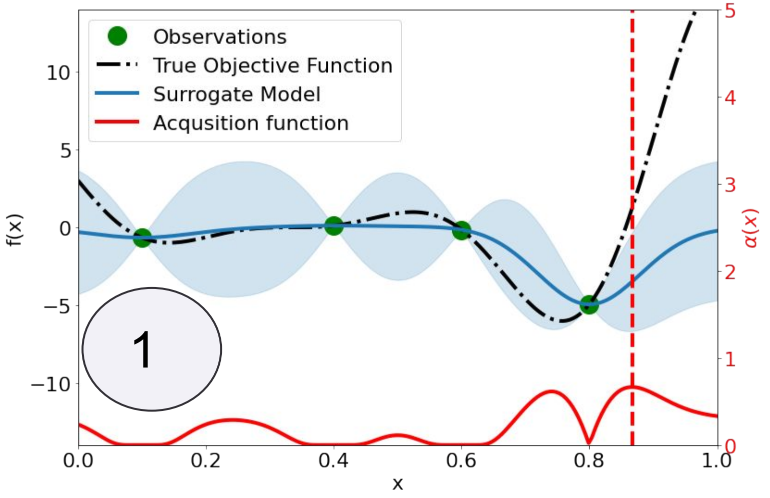


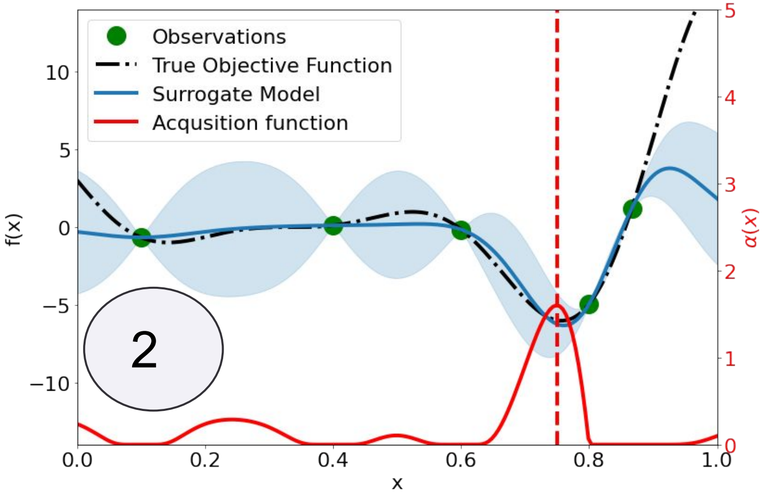


Expected Improvement

Demo BO loop

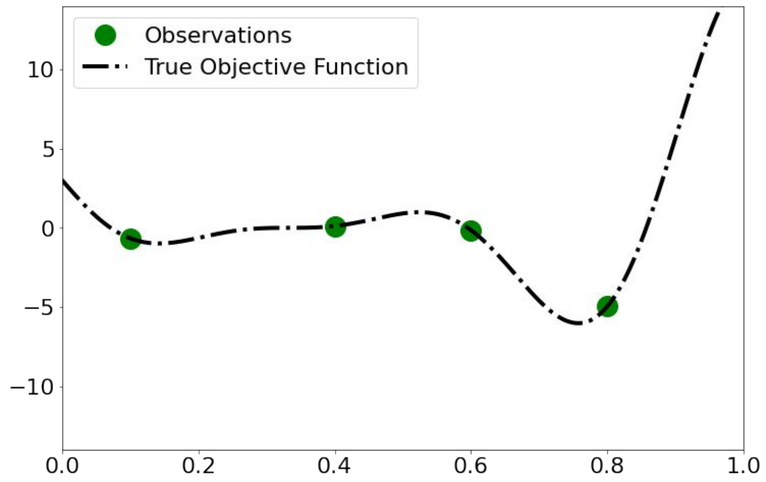


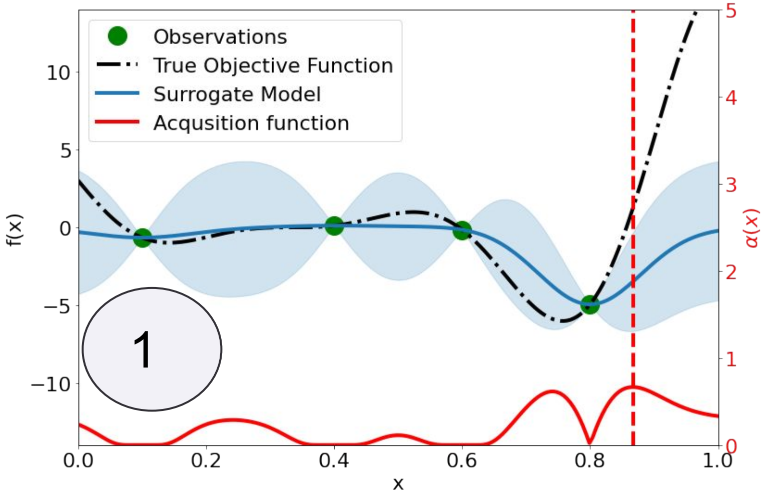


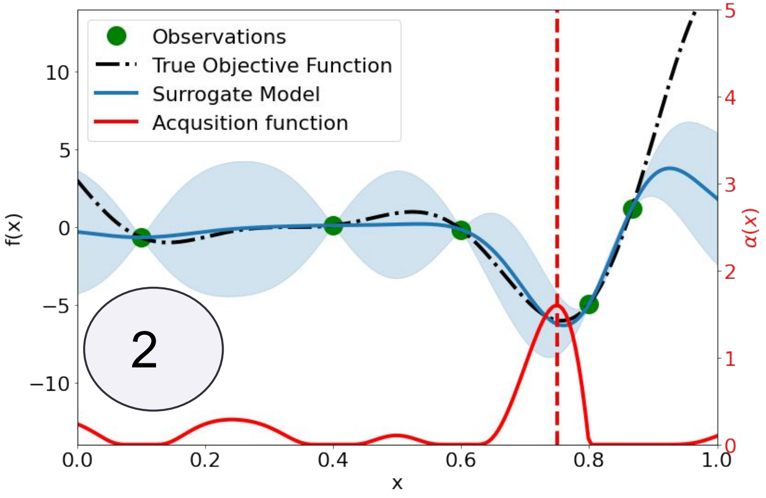


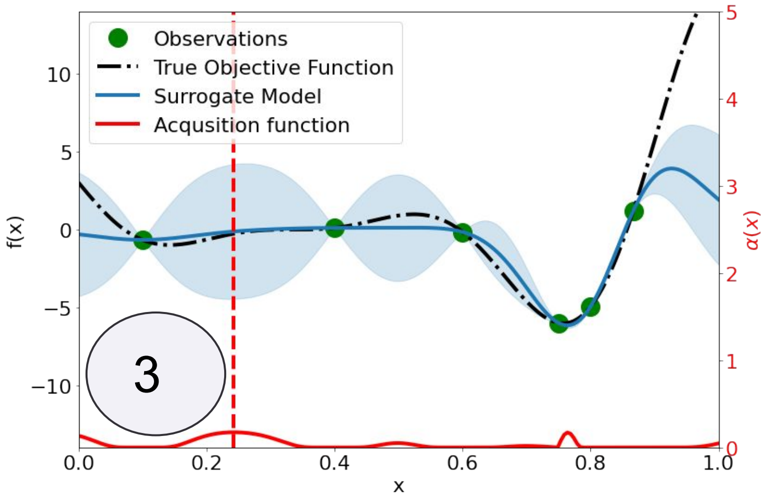
Expected Improvement

Demo BO loop



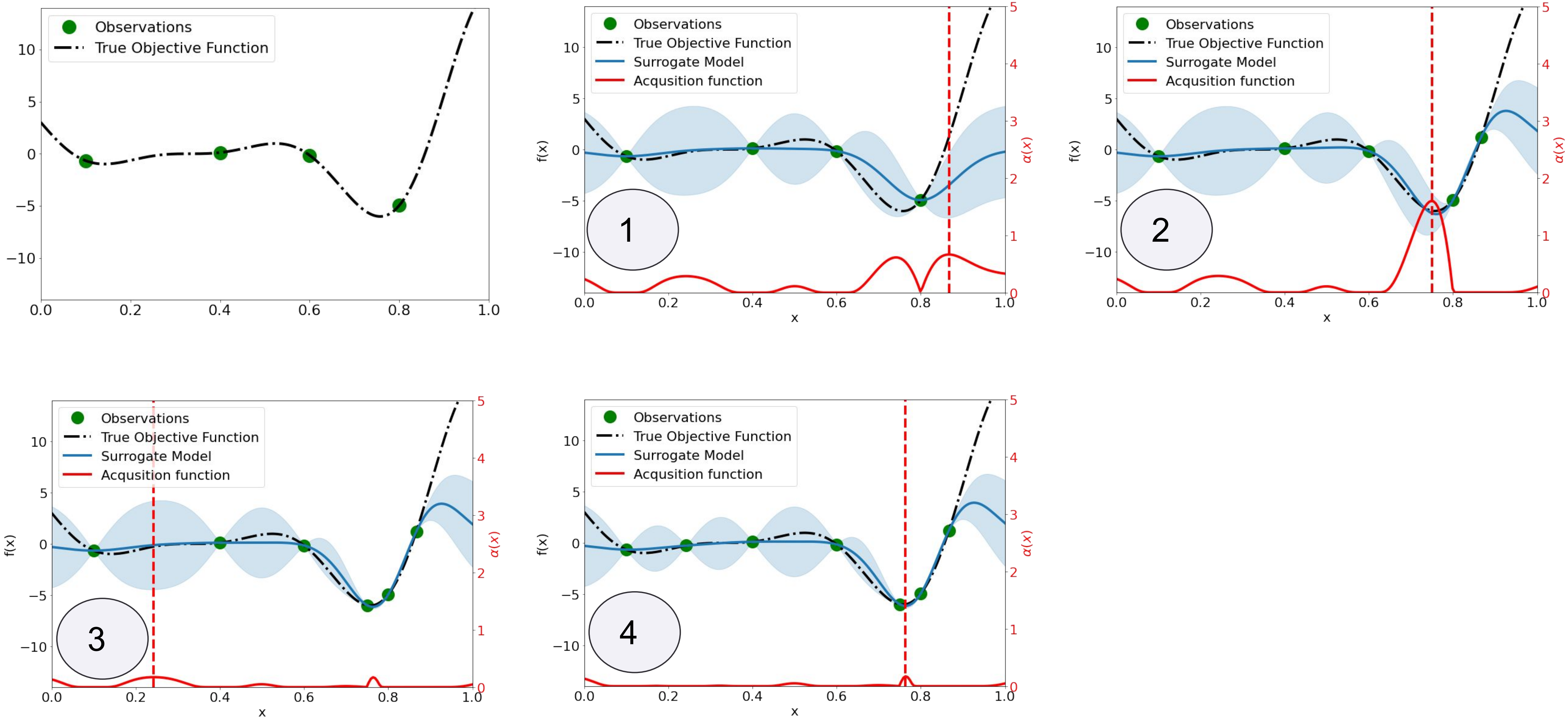






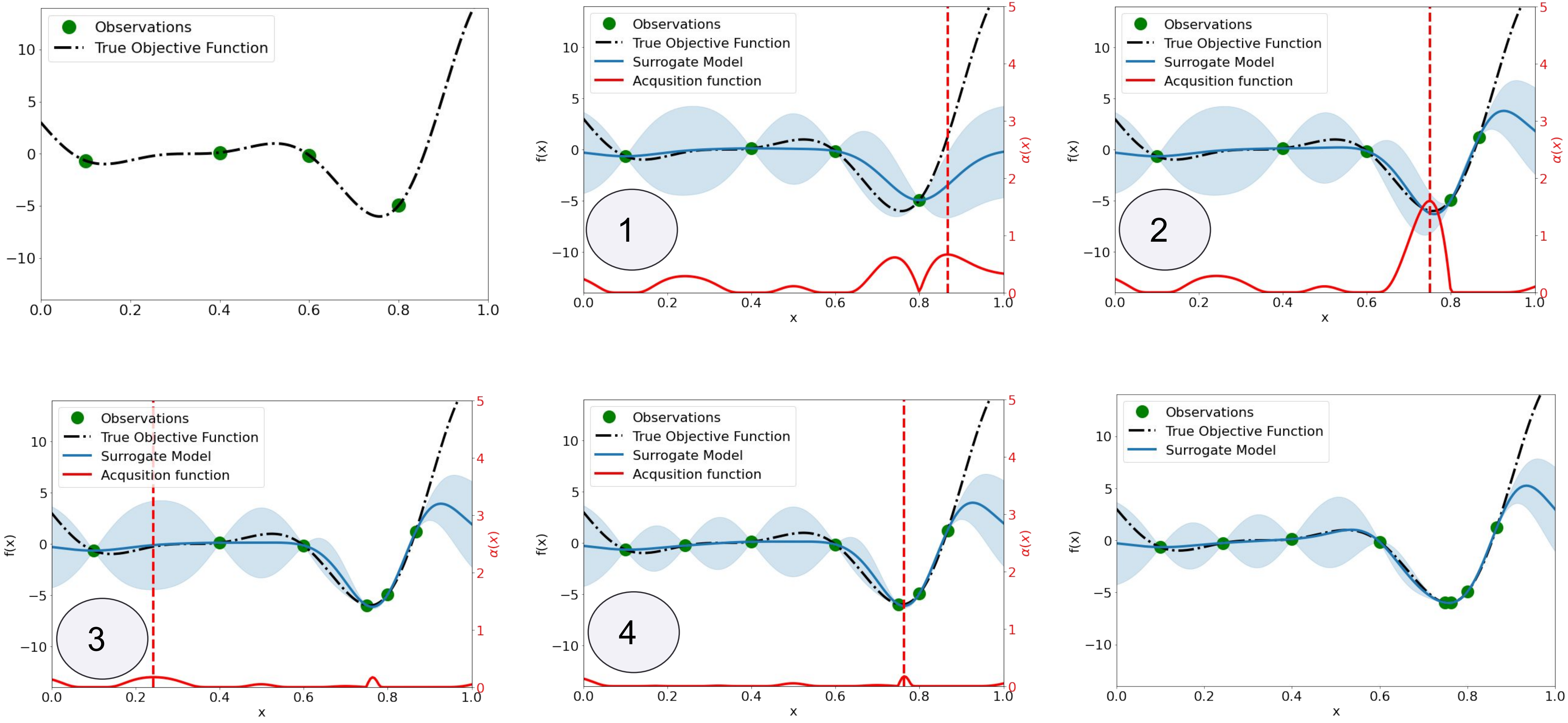
Expected Improvement

Demo BO loop



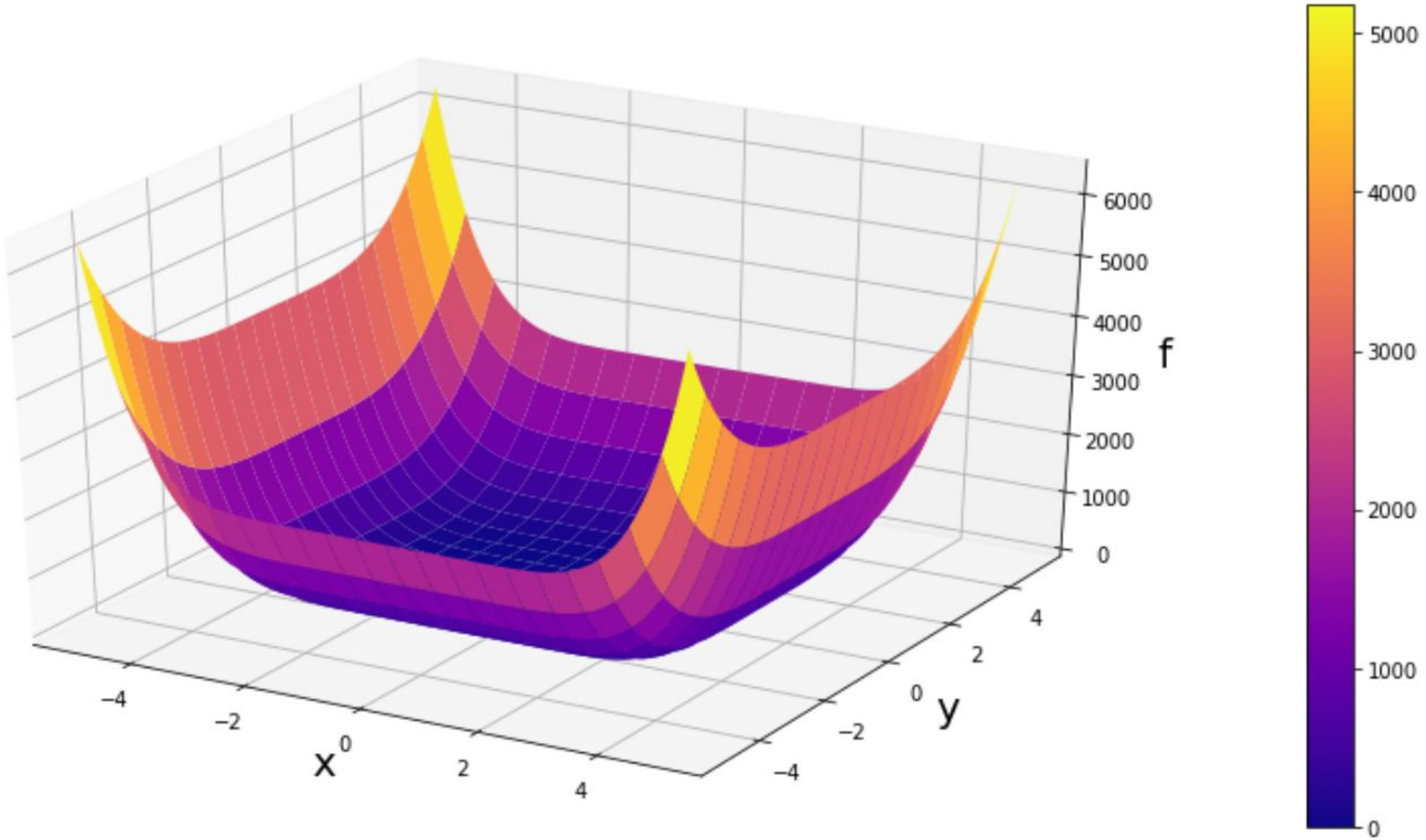
Expected Improvement

Demo BO loop



BO Demo 2

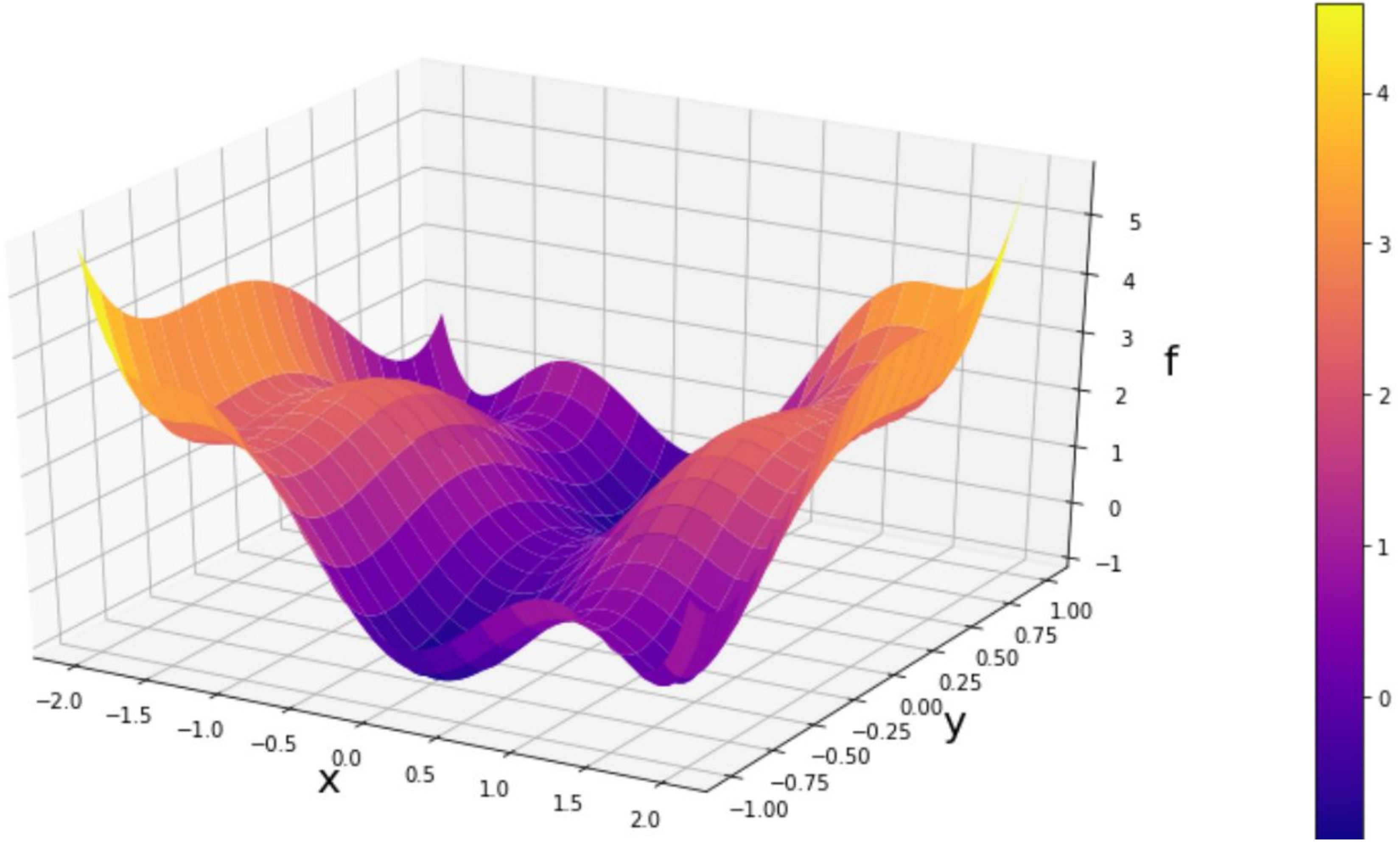
Let minimize the 6 Hump Camel function



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

BO Demo 2

Zoom in: Perhaps not quite as easy?

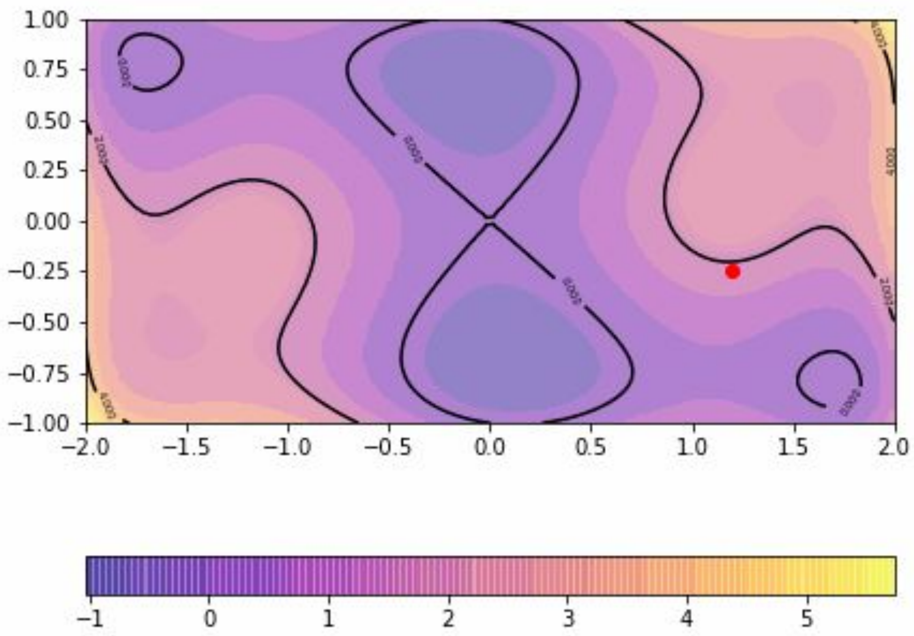


Looks like we **cannot** use a local optimizer!

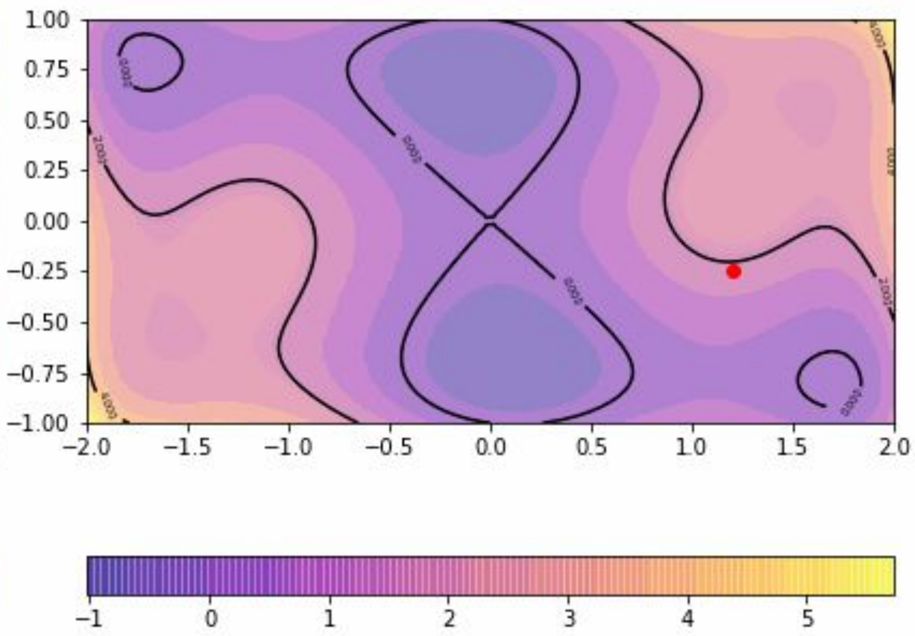
BO Demo 2

Bayesian optimization is a global optimizer

Bayesian optimization (global)

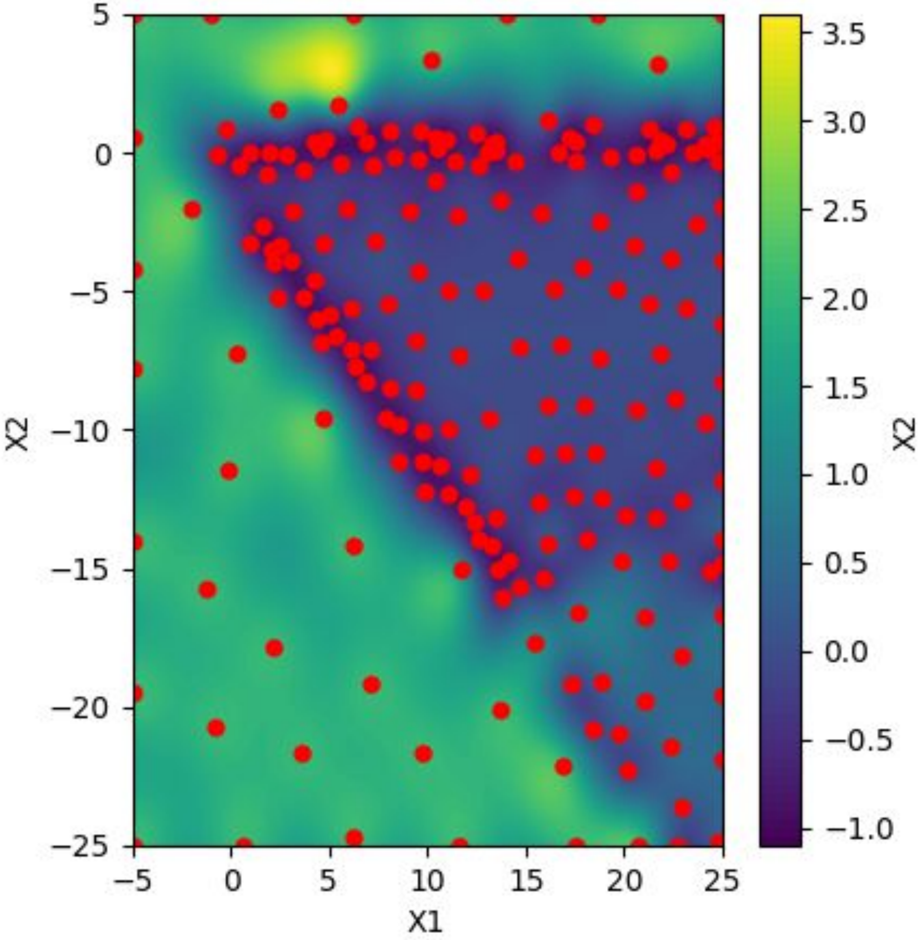


Gradient descent (local)



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

UNIVERSITY OF CAMBRIDGE Lancaster University

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/**

Thanks for listening

UNIVERSITY OF CAMBRIDGE

Lancaster University