Conditioning by adaptive sampling for robust design

David Brookes

Biophysics Graduate Group

University California, Berkeley

Jennifer Listgarten

EECS and Center for Computational Biology

University California, Berkeley

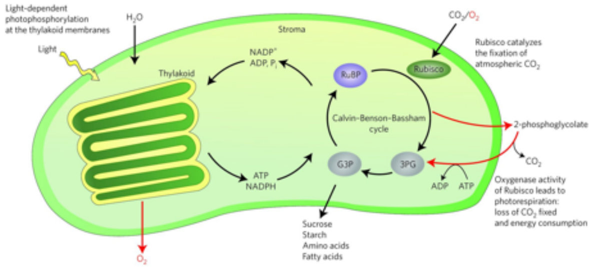
BERKELEY BIOPHYSICS

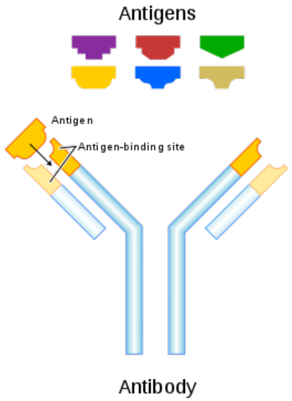
BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

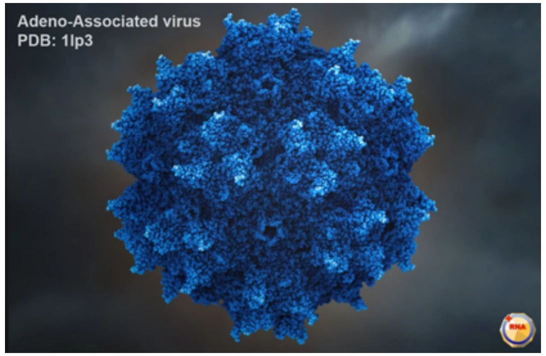
Berkeley EECS

Motivating problem: design protein sequences

- Proteins are made up of sequences of amino acids (20 possibilities)
- Huge variety of proteins whose function we would like to improve


Light-dependent photophosphorylation at the thylakoid membranes
H_2O
Stroma
CO_2/O_2
Light
Rubisco catalyzes the fixation of atmospheric CO_2
Thylakoid
$NADP^+$ ADP, P_i
RuBP
Rubisco
Calvin-Benson-Bassham cycle
2-phosphoglycolate
CO_2
G3P
3PG
ATP NADPH
ADP
ATP
Oxygenase activity of Rubisco leads to photorespiration: loss of CO_2 fixed and energy consumption
Sucrose
Starch
Amino acids
Fatty acids
O_2

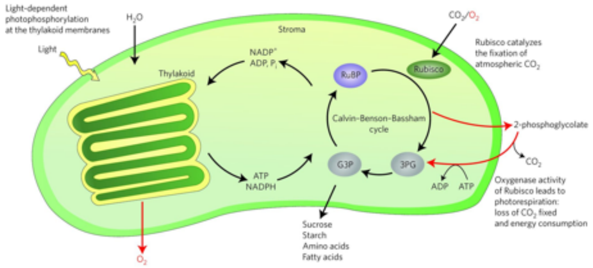


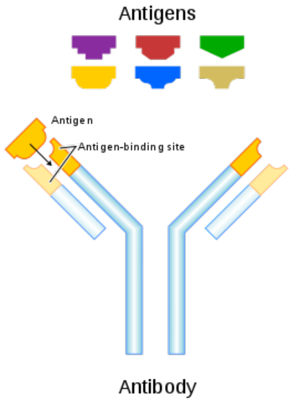


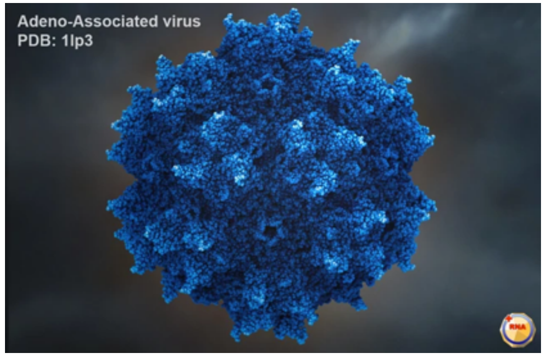
Motivating problem: design protein sequences

- Proteins are made up of sequences of amino acids (20 possibilities)
- Huge variety of proteins whose function we would like to improve

Proteins that fluoresce


Light-dependent photophosphorylation at the thylakoid membranes
H_2O
Stroma
CO_2/O_2
Rubisco catalyzes the fixation of atmospheric CO_2
Light
NADP$^+$
ADP, P_i
Thylakoid
RuBP
Rubisco
Calvin-Benson-Bassham cycle
2-phosphoglycolate
CO_2
G3P
3PG
ATP
NADPH
ADP
ATP
Oxygenase activity of Rubisco leads to photorespiration: loss of CO_2 fixed and energy consumption
Sucrose
Starch
Amino acids
Fatty acids
O_2

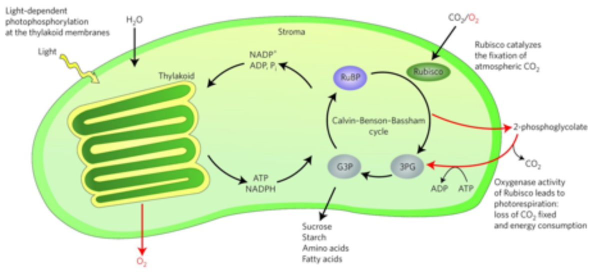


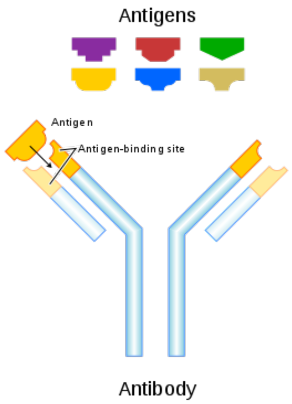


Motivating problem: design protein sequences

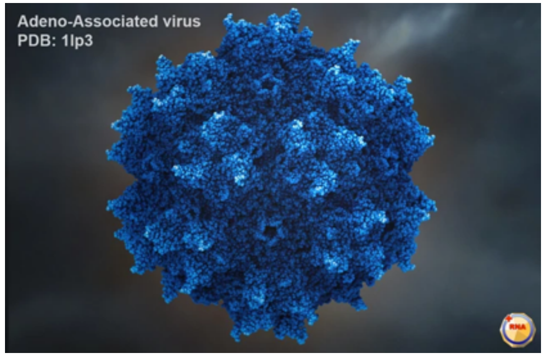
- Proteins are made up of sequences of amino acids (20 possibilities)
- Huge variety of proteins whose function we would like to improve

Proteins that fluoresce


Light-dependent photophosphorylation at the thylakoid membranes
H_2O
Stroma
CO_2/O_2
Light
Rubisco catalyzes the fixation of atmospheric CO_2
Thylakoid
$NADP^+$
ADP, P_i
RuBP
Rubisco
Calvin-Benson-Bassham cycle
2-phosphoglycolate
CO_2
G3P
3PG
ATP
NADPH
ADP
ATP
Oxygenase activity of Rubisco leads to photorespiration: loss of CO_2 fixed and energy consumption
Sucrose
Starch
Amino acids
Fatty acids
O_2



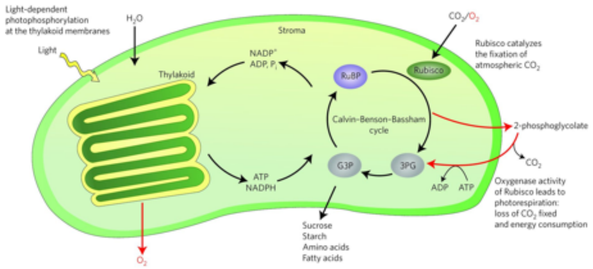
… that act as drugs



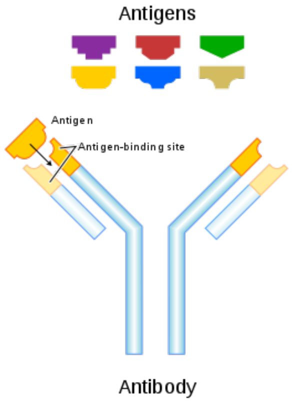
Motivating problem: design protein sequences

- Proteins are made up of sequences of amino acids (20 possibilities)
- Huge variety of proteins whose function we would like to improve

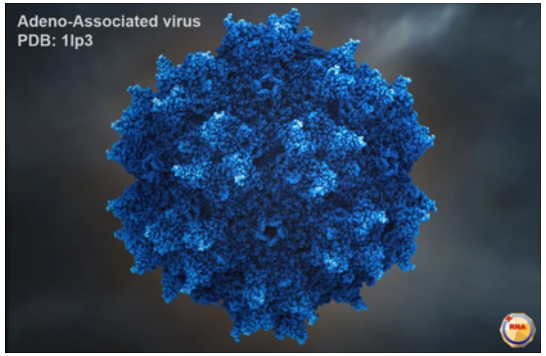
Proteins that fluoresce


Light-dependent photophosphorylation at the thylakoid membranes
H_2O
Stroma
CO_2/O_2
Light
Thylakoid
$NADP^+$ ADP, P_i
RuBP
Rubisco
Rubisco catalyzes the fixation of atmospheric CO_2
Calvin-Benson-Bassham cycle
2-phosphoglycolate
CO_2
G3P
3PG
ATP NADPH
ADP
ATP
Oxygenase activity of Rubisco leads to photorespiration: loss of CO_2 fixed and energy consumption
Sucrose
Starch
Amino acids
Fatty acids
O_2

… that fixate carbon in the atmosphere



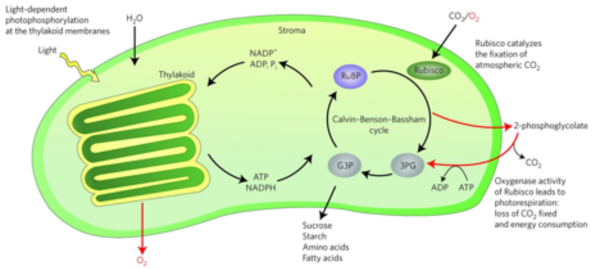
… that act as drugs



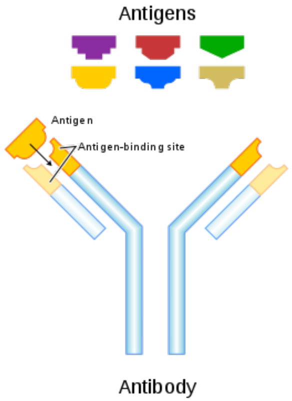
Motivating problem: design protein sequences

- Proteins are made up of sequences of amino acids (20 possibilities)
- Huge variety of proteins whose function we would like to improve

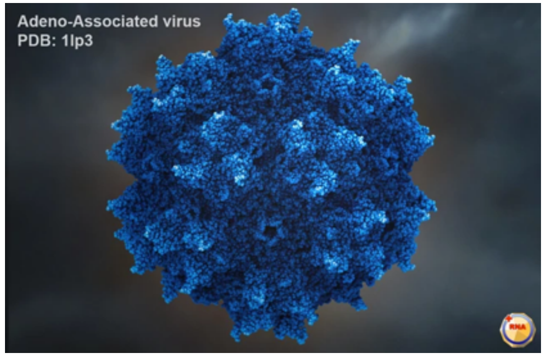
Proteins that fluoresce


Light-dependent photophosphorylation at the thylakoid membranes
H_2O
Stroma
CO_2/O_2
Rubisco catalyzes the fixation of atmospheric CO_2
Light
Thylakoid
$NADP^+$ ADP, P_i
RuBP
Rubisco
Calvin-Benson-Bassham cycle
2-phosphoglycolate
CO_2
G3P
3PG
ATP NADPH
ADP
ATP
Oxygenase activity of Rubisco leads to photorespiration: loss of CO_2 fixed and energy consumption
Sucrose
Starch
Amino acids
Fatty acids
O_2

... that fixate carbon in the atmosphere



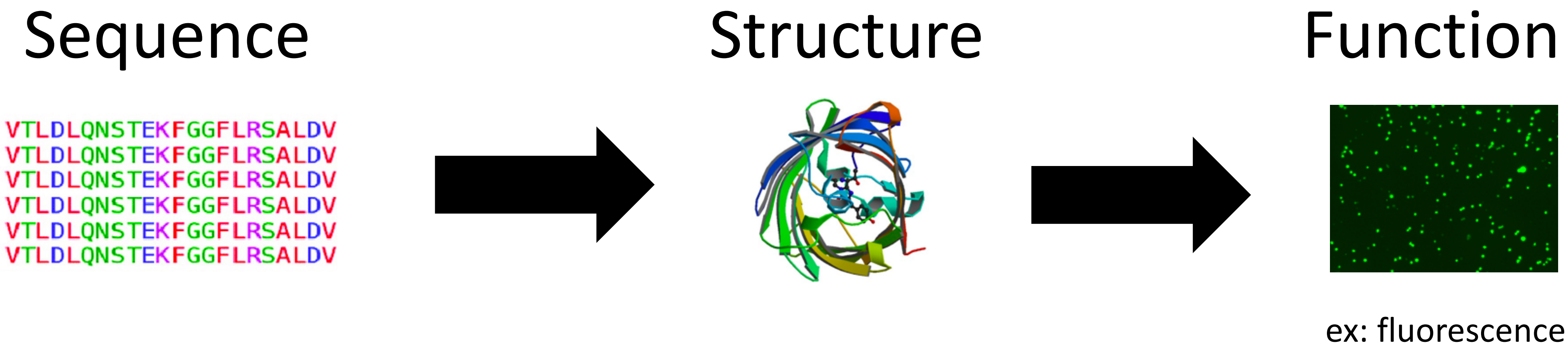
.... that act as drugs



... that deliver gene-editing tools to tissues

How to map sequence to function?

A law of molecular biology:



Sequence

VTLDLQNSTEKFGGFLRSALDV
VTLDLQNSTEKFGGFLRSALDV
VTLDLQNSTEKFGGFLRSALDV
VTLDLQNSTEKFGGFLRSALDV
VTLDLQNSTEKFGGFLRSALDV
VTLDLQNSTEKFGGFLRSALDV

Structure

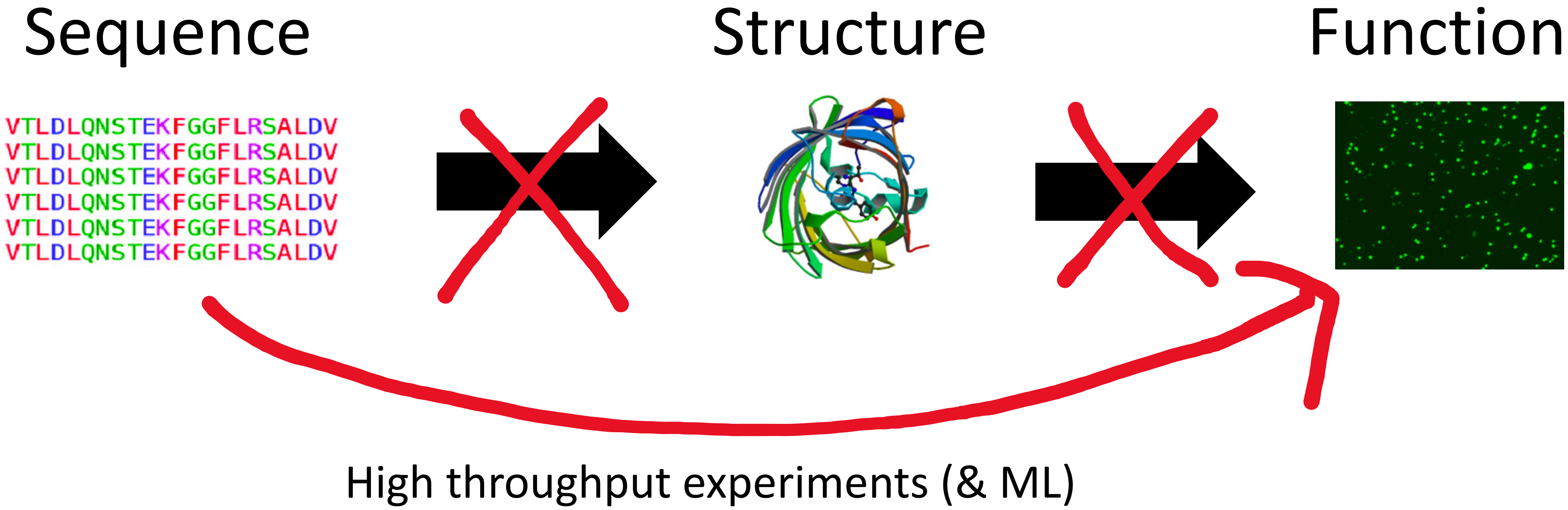
Function

ex: fluorescence

Hughes A, Mort M, Carlisle F, *et al* B04 Alternative Splicing In Htt *Journal of Neurology, Neurosurgery & Psychiatry* 2014;**85:**A10.
http://www.rcsb.org/structure/6FWW

Bypassing the structure relationships

A law of molecular biology:



Sequence

Structure

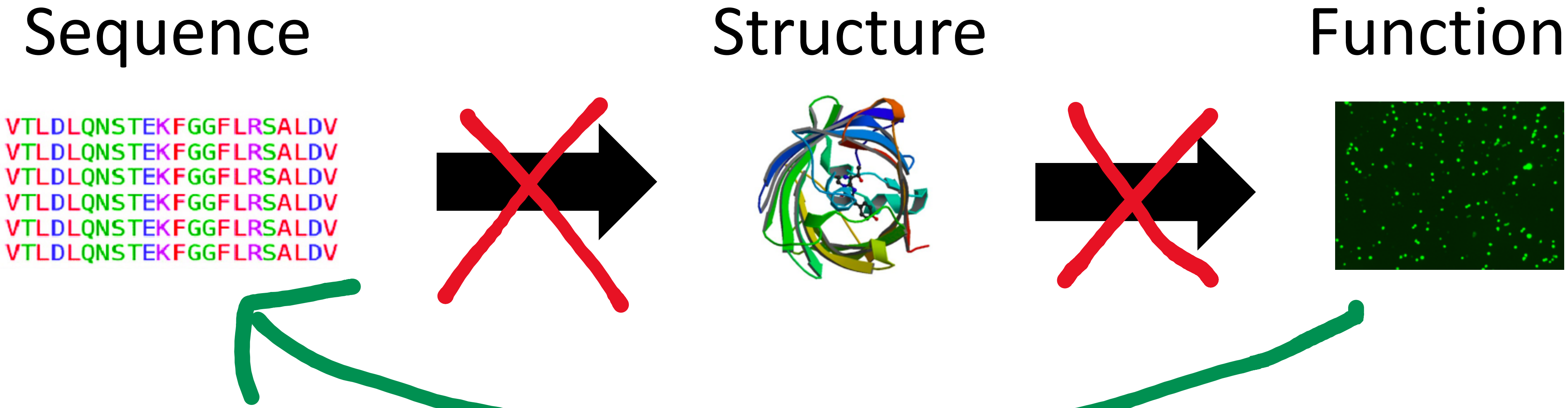
Function

High throughput experiments (& ML)

Hughes A, Mort M, Carlisle F, *et al* B04 Alternative Splicing In Htt *Journal of Neurology, Neurosurgery & Psychiatry* 2014;**85:**A10.
http://www.rcsb.org/structure/6FWW

Can we solve the inverse problem?

A law of molecular biology:



Sequence

Structure

Function

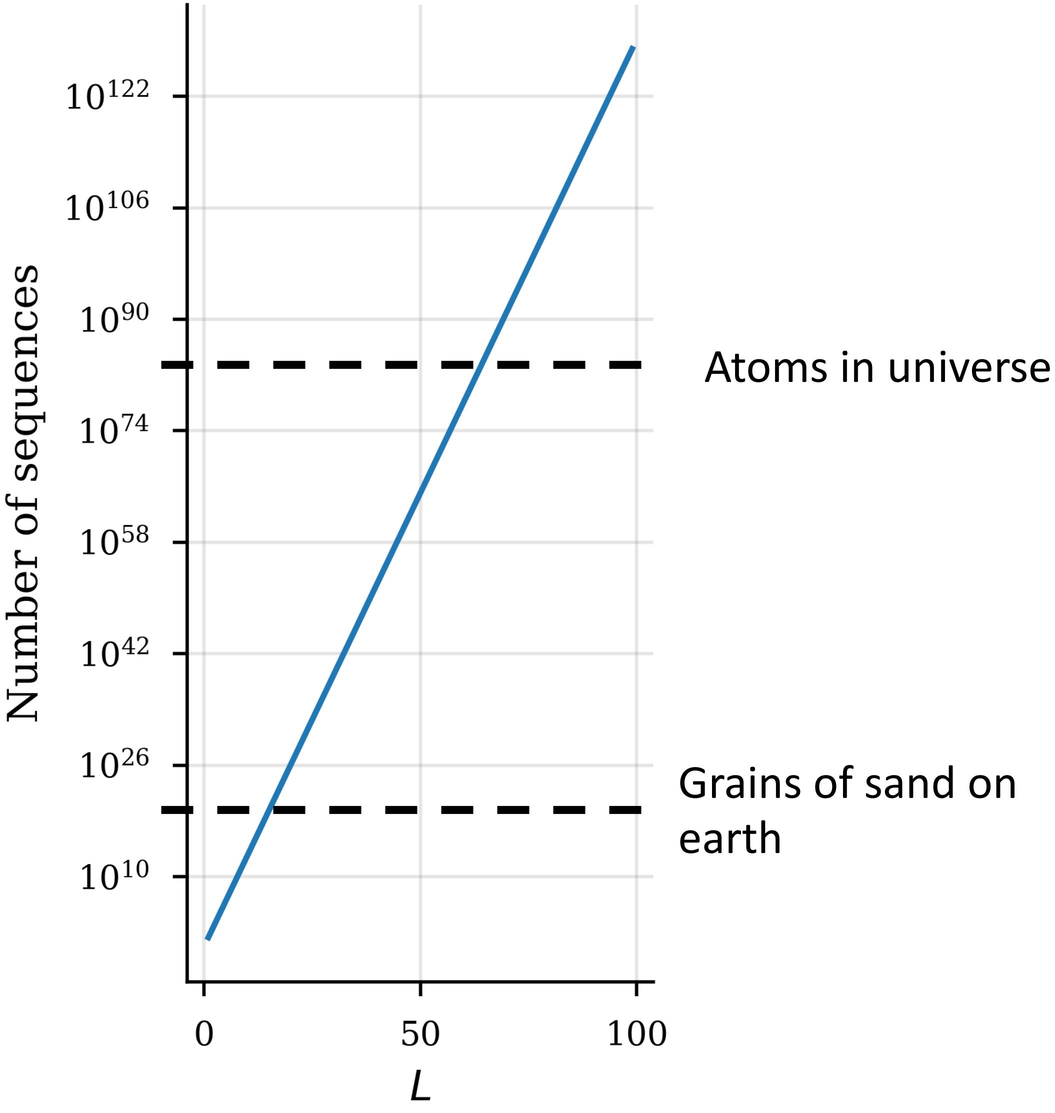
Design problem: Given a model, find sequences with desired function

Hughes A, Mort M, Carlisle F, *et al* B04 Alternative Splicing In Htt *Journal of Neurology, Neurosurgery & Psychiatry* 2014;**85**:A10.
http://www.rcsb.org/structure/6FWW

Why is protein design difficult?

- Huge, rugged search space

$\Longrightarrow$ size scales as 20^L

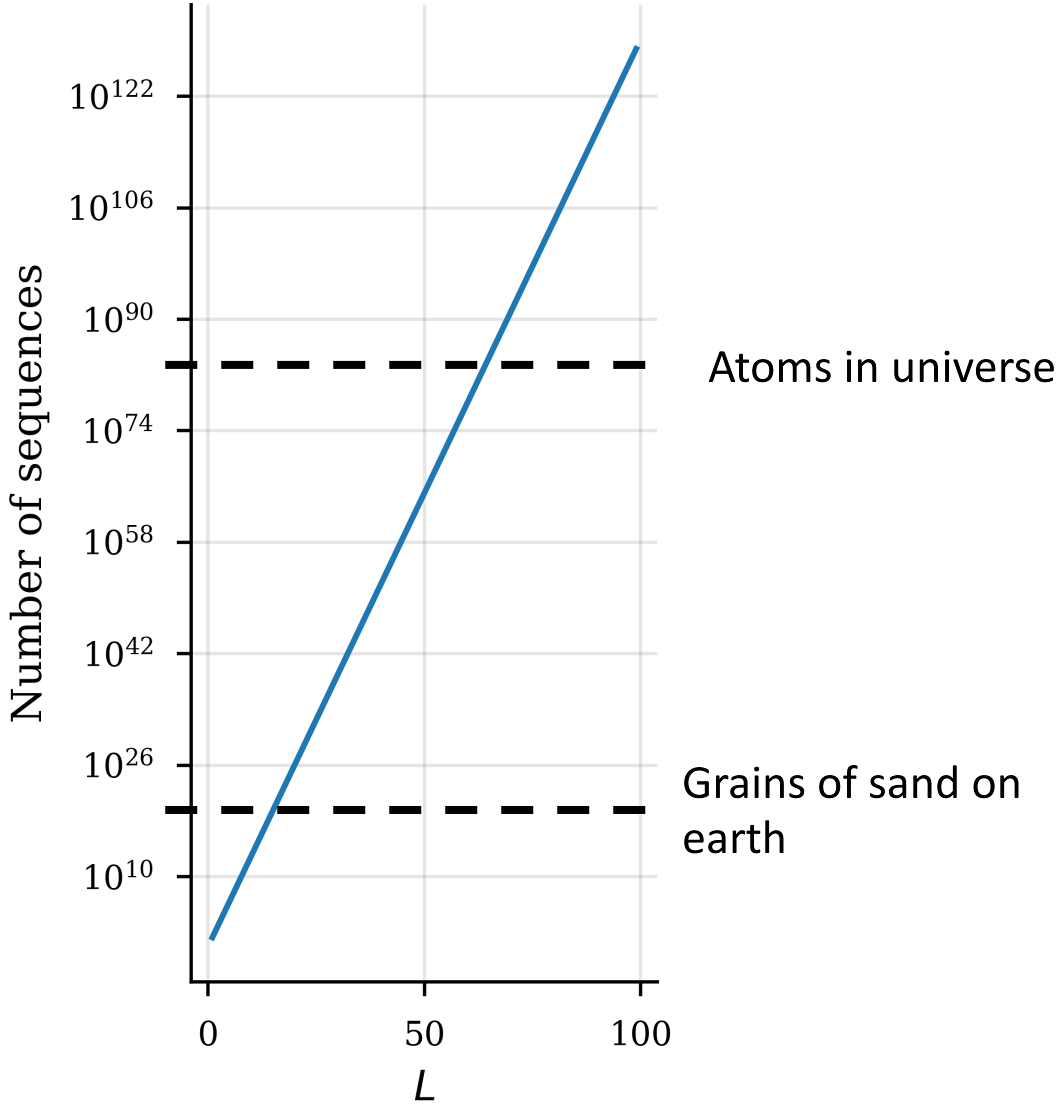


Why is protein design difficult?

- Huge, rugged search space

⟹ size scales as 20^L

- Discrete search space (no gradients)

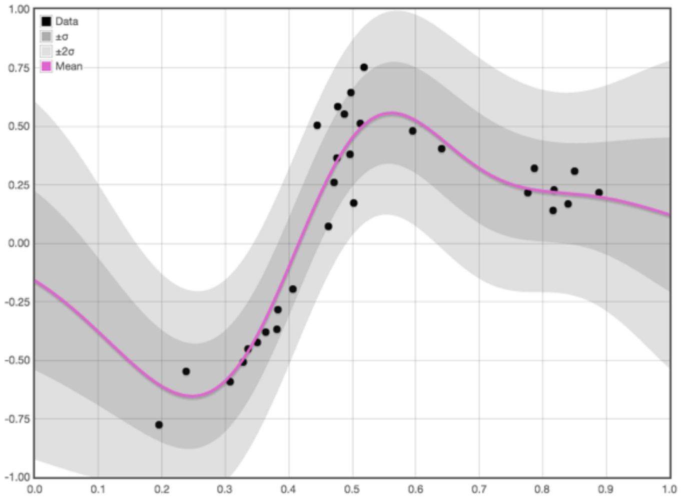


Why is protein design difficult?

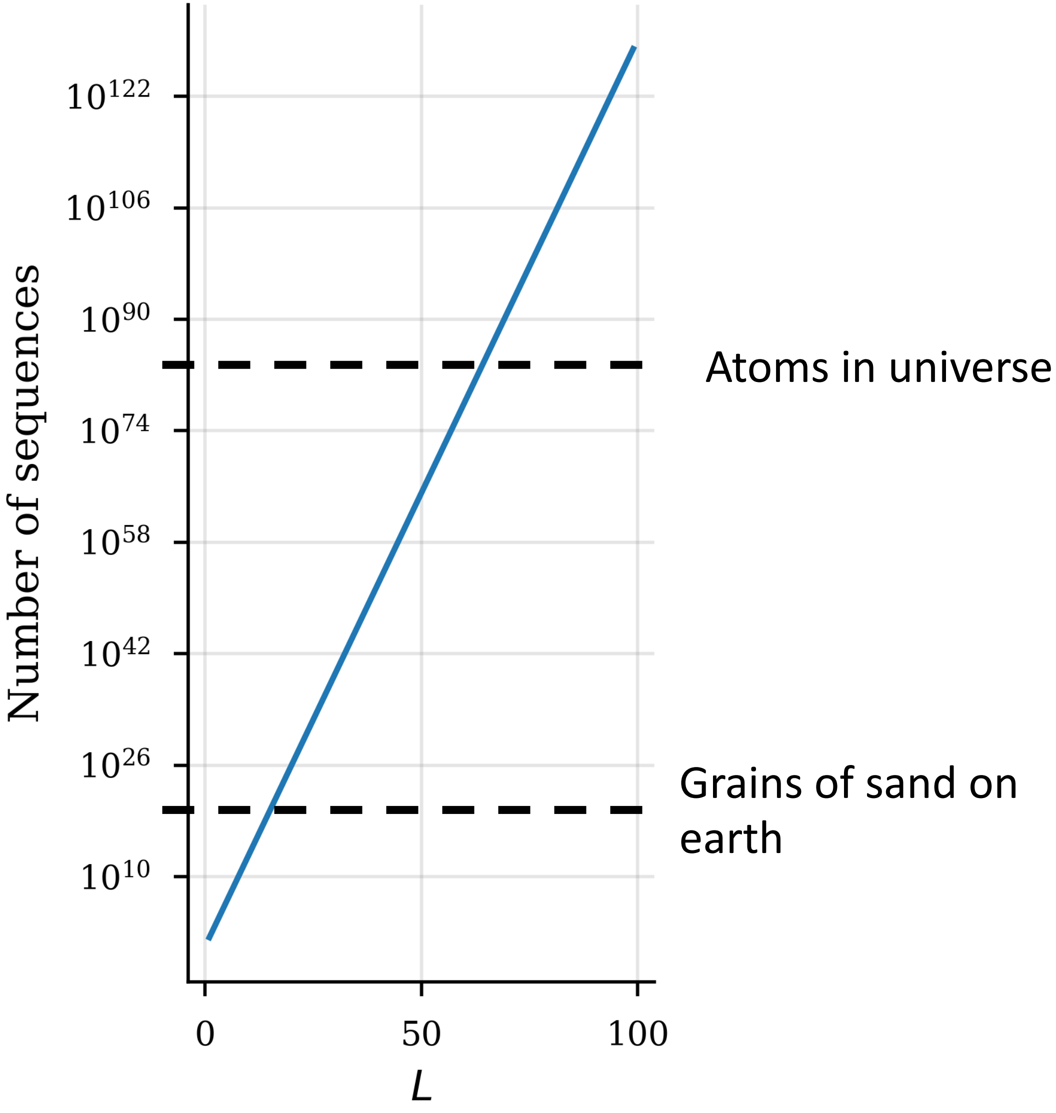
- Huge, rugged search space

⟹ size scales as 20^L

- Discrete search space (no gradients)
- Uncertainty in predictor



https://livingthing.danmackinlay.name/gaussian_processes.html69



Possible solution: model-based optimization (MBO)

Idea: replace the standard (hard) objective

$$\max_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x})$$

e.g. the space of sequences

Possible solution: model-based optimization (MBO)

Idea: replace the standard (hard) objective with a potentially easier one

$$\max_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}) \quad \Longrightarrow \quad \max_{\boldsymbol{\theta} \in \mathbb{R}^d} \mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta})}[f(\boldsymbol{x})]$$

the space of sequences

model over sequence space

Possible solution: model-based optimization (MBO)

Idea: replace the standard (hard) objective with a potentially easier one

$$\max_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}) \quad \Longrightarrow \quad \max_{\boldsymbol{\theta} \in \mathbb{R}^d} \mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta})}[f(\boldsymbol{x})]$$

Solution approach is to iterate:

1. Sample from “search model” $p(x|\theta)$
2. Evaluate samples on $f(x)$
3. Adjust θ so the model favors samples with large function evals

Possible solution: model-based optimization (MBO)

Idea: replace the standard (hard) objective with a potentially easier one

$$\max_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x}) \quad \Longrightarrow \quad \max_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\theta})}[f(\boldsymbol{x})]$$

Solution approach is to iterate:

1. Sample from “search model” $p(x|\theta)$
2. Evaluate samples on $f(x)$
3. Adjust θ so the model favors sequences with large function evals

- ✓ Model can sample broad areas of sequence space
- ✓ Does not require gradients of f
- ✓ Can incorporate uncertainty

First attempt at MBO for protein design:
Design by Adaptive Sampling (DbAS)

Our aim is solve the MBO objective:

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta})}\left[P(S|\mathbf{x})\right]$$

First attempt at MBO for protein design:
Design by Adaptive Sampling (DbAS)

Our aim is solve the MBO objective:

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta})} \left[P(S|\mathbf{x}) \right]$$

where

- $p(x|\theta)$ is the search model (VAE, HMM…)

First attempt at MBO for protein design:

Design by Adaptive Sampling (DbAS)

Our aim is solve the MBO objective:

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta})} \left[P(S|\mathbf{x}) \right]$$

where

- $p(x|\theta)$ is the search model (VAE, HMM…)
- S is desired set of property values

  → *e.g.* fluorescence > α

First attempt at MBO for protein design:
Design by Adaptive Sampling (DbAS)

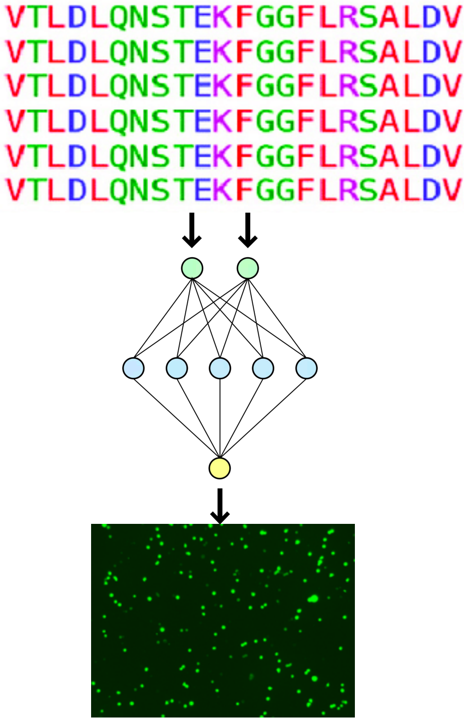
Our aim is solve the MBO objective:

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta})}\left[P(S|\mathbf{x})\right]$$

where

- $p(x|\theta)$ is the search model (VAE, HMM…)
- S is desired set of property values

  → *e.g.* fluorescence > α
- $P(S|x)$ is a stochastic predictive model (“oracle”) that maps sequences to property



Design by Adaptive Sampling (cont.)

Two issues:

1. ***θ* is in the expectation distribution.**

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log \mathbb{E}_{p(\mathbf{x}|\boxed{\boldsymbol{\theta}})} \left[P(S|\mathbf{x}) \right]$$

Design by Adaptive Sampling (cont.)

maximize a lower bound

Two issues:

1. ~~**θ is in the expectation distribution.**~~

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta})}\left[P(S|\mathbf{x})\right],$$

$$\Downarrow \quad \geq$$

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta}^{(t)})}\left[P(S|\mathbf{x}) \log p(\mathbf{x}|\boldsymbol{\theta})\right].$$

Design by Adaptive Sampling (cont.)

Two issues:

1. ~~**$\boldsymbol{\theta}$ is in the expectation distribution.**~~
2. **MC estimates for rare events.**

maximize a lower bound

$$\operatorname*{argmax}_{\boldsymbol{\theta}} \log \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta})}\left[P(S|\mathbf{x})\right],$$

$$\Downarrow \quad \geq$$

$$\operatorname*{argmax}_{\boldsymbol{\theta}} \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta}^{(t)})}\left[P(S|\mathbf{x}) \log p(\mathbf{x}|\boldsymbol{\theta})\right].$$

Design by Adaptive Sampling (cont.)

Two issues:

1. ~~**θ is in the expectation distribution.**~~
2. ~~**MC estimates for rare events.**~~

maximize a lower bound

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta})}\left[P(S|\mathbf{x})\right],$$

↓ ≥

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta}^{(t)})}\left[P(S|\mathbf{x}) \log p(\mathbf{x}|\boldsymbol{\theta})\right].$$

anneal a sequence of relaxations:

$S^t \rightarrow S$, where $S^t \supset S^{t+1}$

Design by Adaptive Sampling (cont.)

Two issues:

1. ~~**$\boldsymbol{\theta}$ is in the expectation distribution.**~~
2. ~~**MC estimates for rare events.**~~

maximize a lower bound

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta})} \left[P(S|\mathbf{x}) \right],$$

↓ $\geq$

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \, \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta}^{(t)})} \left[P(S|\mathbf{x}) \log p(\mathbf{x}|\boldsymbol{\theta}) \right]$$

↓ Anneal and MC

$$\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{M} P(S^{(t)}|\mathbf{x}_i^{(t)}) \log p(\mathbf{x}_i^{(t)}|\boldsymbol{\theta})$$

Design by Adaptive Sampling (cont.)

maximize a lower bound

Two issues:

1. ~~θ **is in th**~~
   ~~**distribut**~~
2. ~~**MC estimates for rare events.**~~

Assumes oracle is unbiased and has good uncertainty estimates

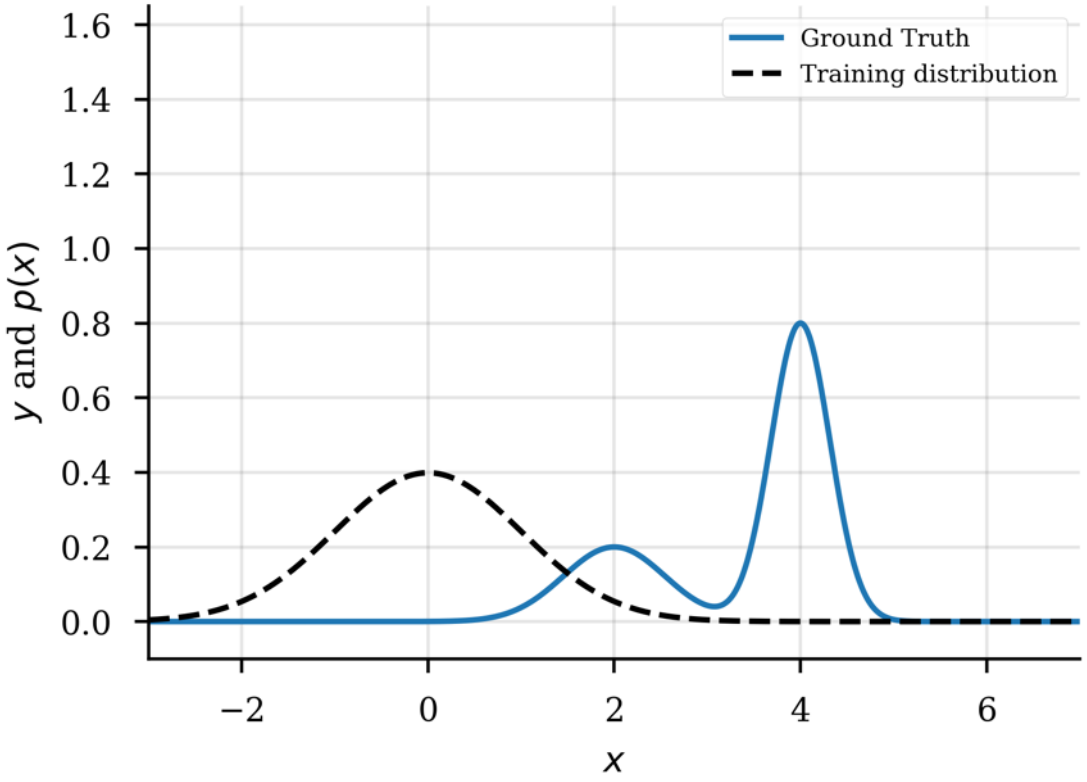
$$\mathrm{argmax} \log \mathbb{E} \; [P(S|\mathbf{x})] \,,$$

$$\max_{\boldsymbol{\theta}} \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta}^{(t)})} [P(S|\mathbf{x}) \log p(\mathbf{x}|\boldsymbol{\theta})]$$

↓ Anneal and MC

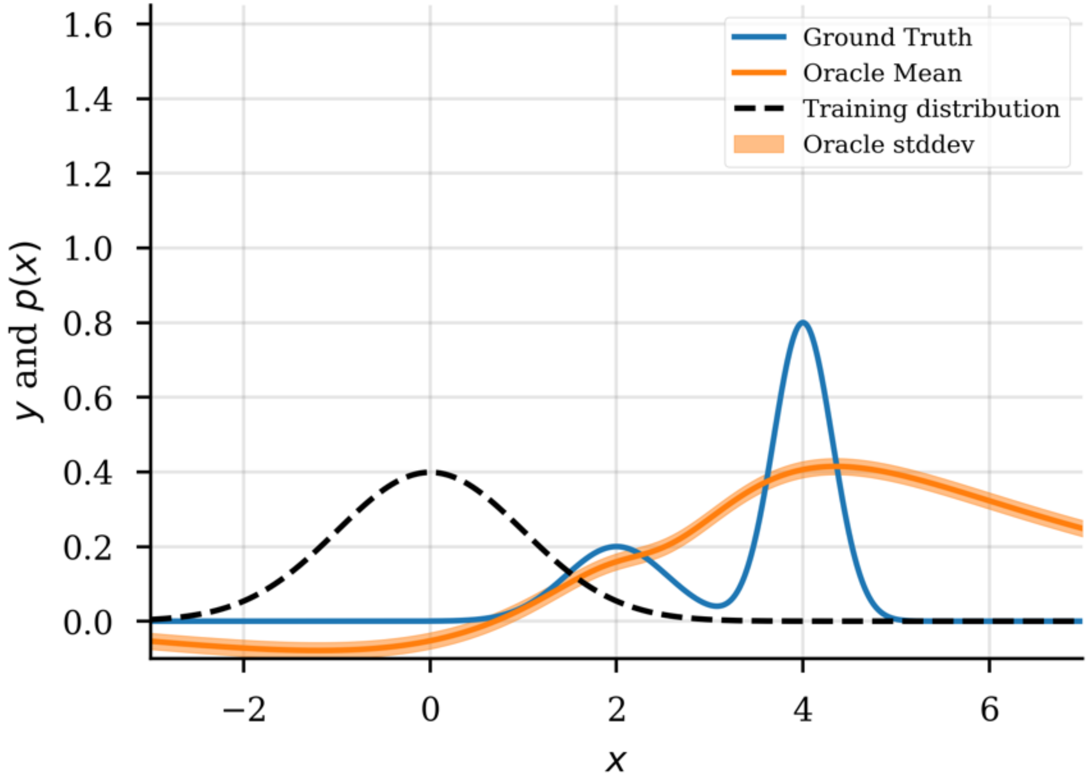
$$\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\mathrm{argmax}} \sum_{i=1}^{M} P(S^{(t)}|\mathbf{x}_i^{(t)}) \log p(\mathbf{x}_i^{(t)}|\boldsymbol{\theta})$$

How pathological oracles lead you astray



How pathological oracles lead you astray

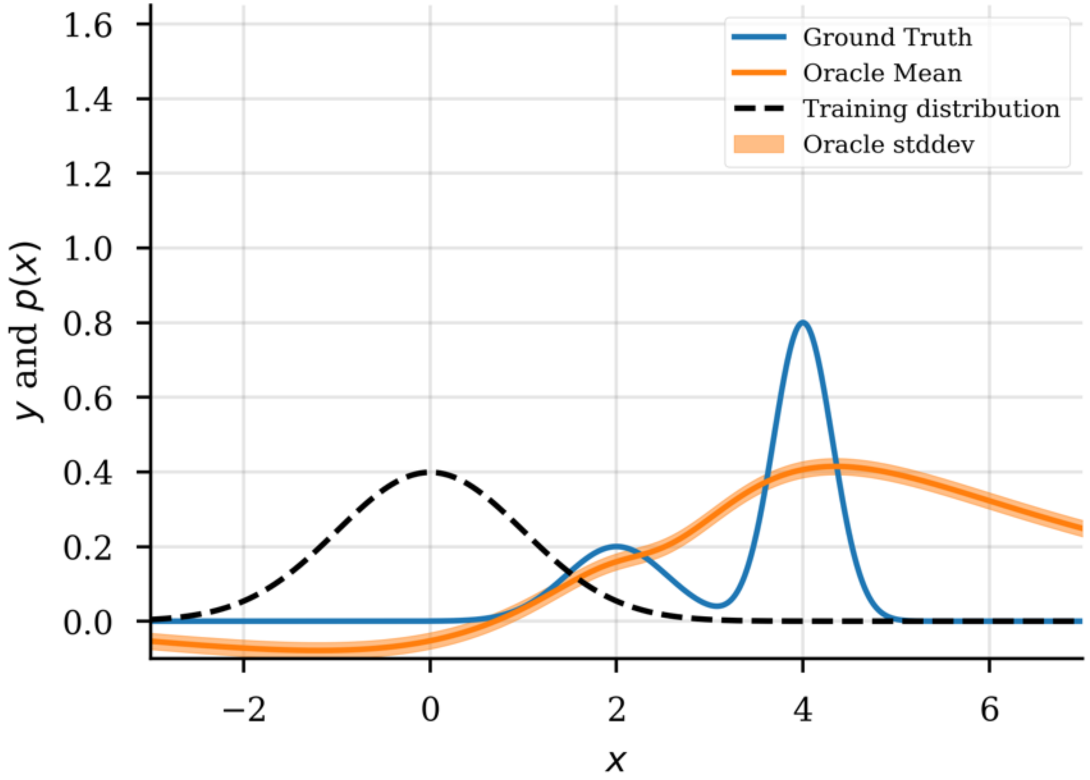
Acceptable



Many training examples

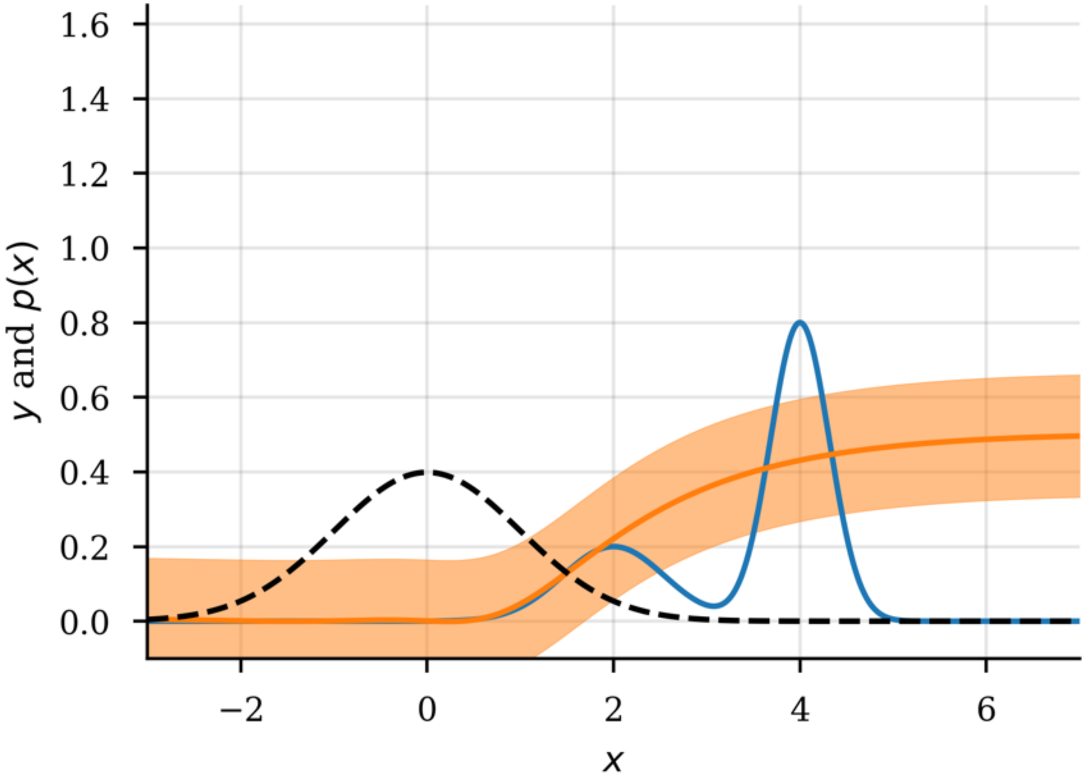
How pathological oracles lead you astray

Acceptable



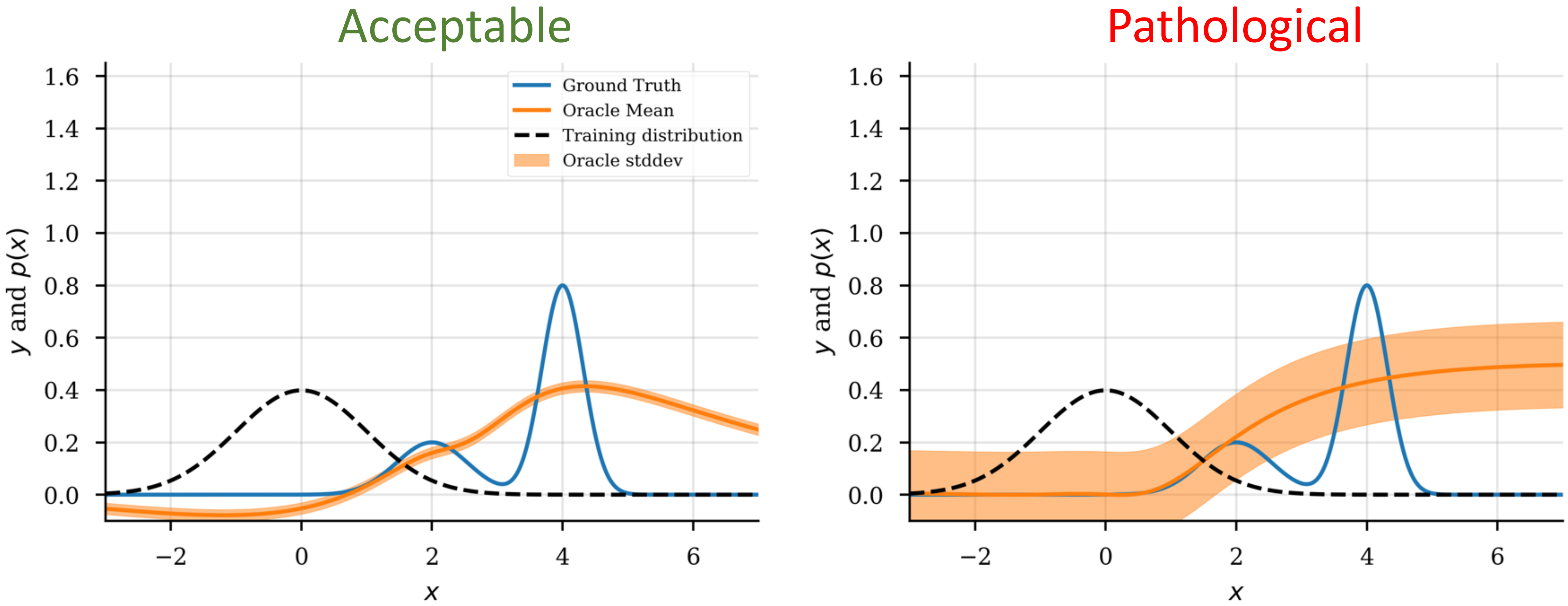
Many training examples

Pathological



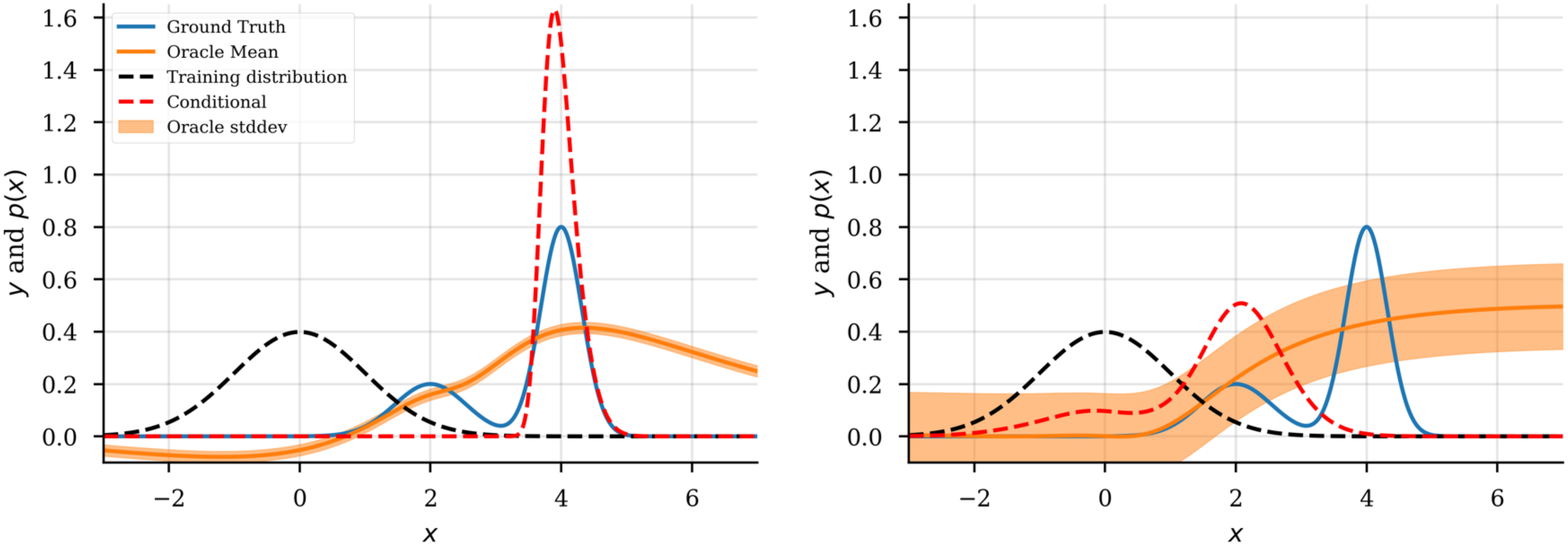
Fewer training examples

How pathological oracles lead you astray



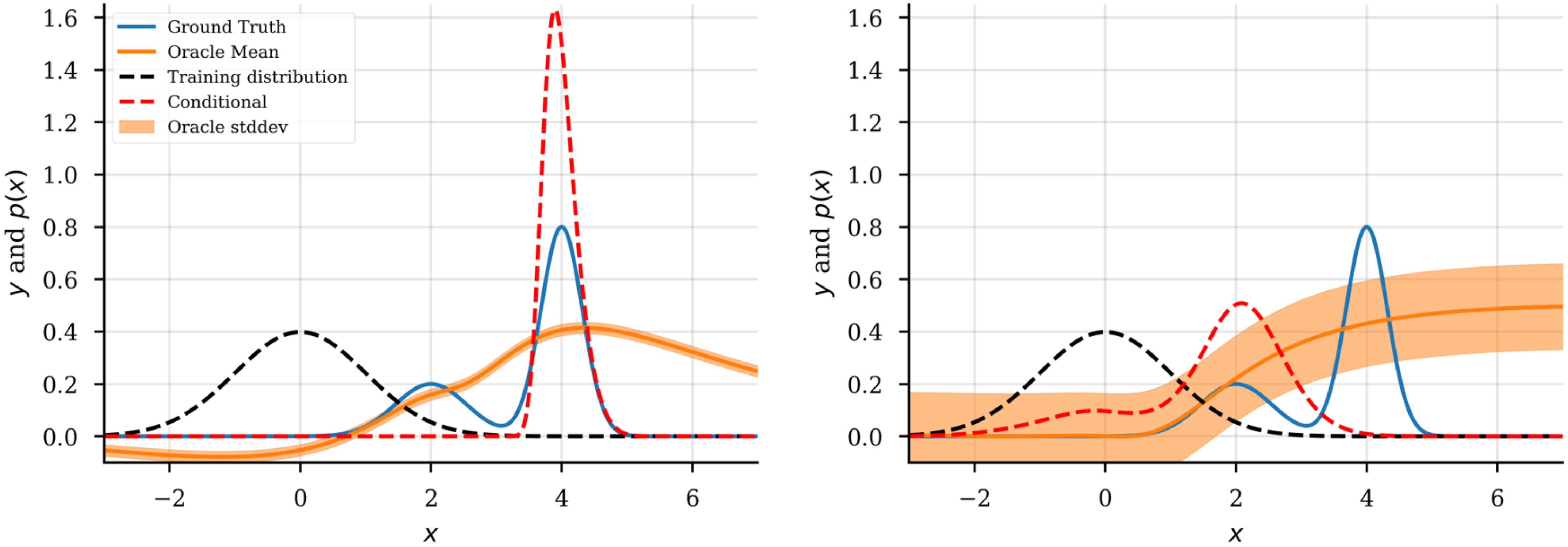
Idea: estimate training distribution of x *conditioned* on high values of oracle

Fixing pathological oracles w/ conditioning



Idea: estimate training distribution of x *conditioned* on high values of oracle

Fixing pathological oracles w/ conditioning



Idea: estimate training distribution of x *conditioned* on high values of oracle

Don’t have access to training distribution, but can build a model $p\left(\boldsymbol{x}\middle|\boldsymbol{\theta}^{(0)}\right)$ to approximate it

Conditioning by Adaptive Sampling (CbAS)

Previous formulation:

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta})}\left[P(S|\mathbf{x})\right]$$

↓ ≥

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta}^{(t)})}\left[P(S|\mathbf{x}) \log p(\mathbf{x}|\boldsymbol{\theta})\right]$$

↓ Anneal and MC

$$\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{M} P(S^{(t)}|\mathbf{x}_i^{(t)}) \log p(\mathbf{x}_i^{(t)}|\boldsymbol{\theta})$$

New formulation:

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, D_{KL}\left(p(\mathbf{x}|S, \boldsymbol{\theta}^{(0)})||p(\mathbf{x}|\boldsymbol{\theta})\right)$$

$p(\boldsymbol{x}|\boldsymbol{\theta}^{(\mathbf{0})})$ models the training distribution

Conditioning by Adaptive Sampling (CbAS)

Previous formulation:

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta})} \left[P(S|\mathbf{x})\right]$$

$$\Big\downarrow \quad \geq$$

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \, \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta}^{(t)})} \left[P(S|\mathbf{x}) \log p(\mathbf{x}|\boldsymbol{\theta})\right]$$

$$\Big\downarrow \quad \text{Anneal and MC}$$

$$\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{M} P(S^{(t)}|\mathbf{x}_i^{(t)}) \log p(\mathbf{x}_i^{(t)}|\boldsymbol{\theta})$$

New formulation:

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}} \, D_{KL} \left(p(\mathbf{x}|S, \boldsymbol{\theta}^{(0)}) || p(\mathbf{x}|\boldsymbol{\theta}) \right)$$

$$\Big\downarrow \quad =$$

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \, \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta}^{(0)})} \left[P(S|\mathbf{x}) \log p(\mathbf{x}|\boldsymbol{\theta})\right]$$

Conditioning by Adaptive Sampling (CbAS)

Previous formulation:

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta})} \left[P(S|\mathbf{x}) \right]$$

$$\Big\downarrow \quad \geq$$

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \, \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta}^{(t)})} \left[P(S|\mathbf{x}) \log p(\mathbf{x}|\boldsymbol{\theta}) \right]$$

$$\Big\downarrow \quad \text{Anneal and MC}$$

$$\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{M} P(S^{(t)}|\mathbf{x}_i^{(t)}) \log p(\mathbf{x}_i^{(t)}|\boldsymbol{\theta})$$

New formulation:

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}} \, D_{KL} \left(p(\mathbf{x}|S, \boldsymbol{\theta}^{(0)}) || p(\mathbf{x}|\boldsymbol{\theta}) \right)$$

$$\Big\downarrow \quad =$$

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \, \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta}^{(0)})} \left[P(S|\mathbf{x}) \log p(\mathbf{x}|\boldsymbol{\theta}) \right]$$

Can't anneal when sampling dist. doesn't change!

Conditioning by Adaptive Sampling (CbAS)

Previous formulation:

$$=$$

↓ $\geq$

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta}^{(t)})}\left[P(S|\mathbf{x}) \log p(\mathbf{x}|\boldsymbol{\theta})\right]$$

↓ Anneal and MC

$$\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{M} P(S^{(t)}|\mathbf{x}_i^{(t)}) \log p(\mathbf{x}_i^{(t)}|\boldsymbol{\theta})$$

New formulation:

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, D_{KL}\left(p(\mathbf{x}|S,\boldsymbol{\theta}^{(0)})||p(\mathbf{x}|\boldsymbol{\theta})\right)$$

↓ $=$

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta}^{(0)})}\left[P(S|\mathbf{x}) \log p(\mathbf{x}|\boldsymbol{\theta})\right]$$

↓ $=$

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta}^{(t)})}\left[\frac{p(\mathbf{x}|\boldsymbol{\theta}^{(0)})}{p(\mathbf{x}|\boldsymbol{\theta}^{(t)})} P(S|\mathbf{x}) \log p(\mathbf{x}|\boldsymbol{\theta})\right]$$

Importance sampling
proposal dist.

Conditioning by Adaptive Sampling (CbAS)

Previous formulation:

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \log \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta})} \left[P(S|\mathbf{x}) \right]$$

$$\Big\downarrow \quad \geq$$

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \, \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta}^{(t)})} \left[P(S|\mathbf{x}) \log p(\mathbf{x}|\boldsymbol{\theta}) \right]$$

↓ Anneal and MC

$$\boldsymbol{\theta}^{(t+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{M} P(S^{(t)}|\mathbf{x}_i^{(t)}) \log p(\mathbf{x}_i^{(t)}|\boldsymbol{\theta})$$

New formulation:

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}} \, D_{KL} \left(p(\mathbf{x}|S, \boldsymbol{\theta}^{(0)}) || p(\mathbf{x}|\boldsymbol{\theta}) \right)$$

$$\Big\downarrow \quad =$$

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \, \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta}^{(0)})} \left[P(S|\mathbf{x}) \log p(\mathbf{x}|\boldsymbol{\theta}) \right]$$

$$\Big\downarrow \quad =$$

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \, \mathbb{E}_{p(\mathbf{x}|\boldsymbol{\theta}^{(t)})} \left[\frac{p(\mathbf{x}|\boldsymbol{\theta}^{(0)})}{p(\mathbf{x}|\boldsymbol{\theta}^{(t)})} P(S|\mathbf{x}) \log p(\mathbf{x}|\boldsymbol{\theta}) \right]$$

↓ Anneal and MC

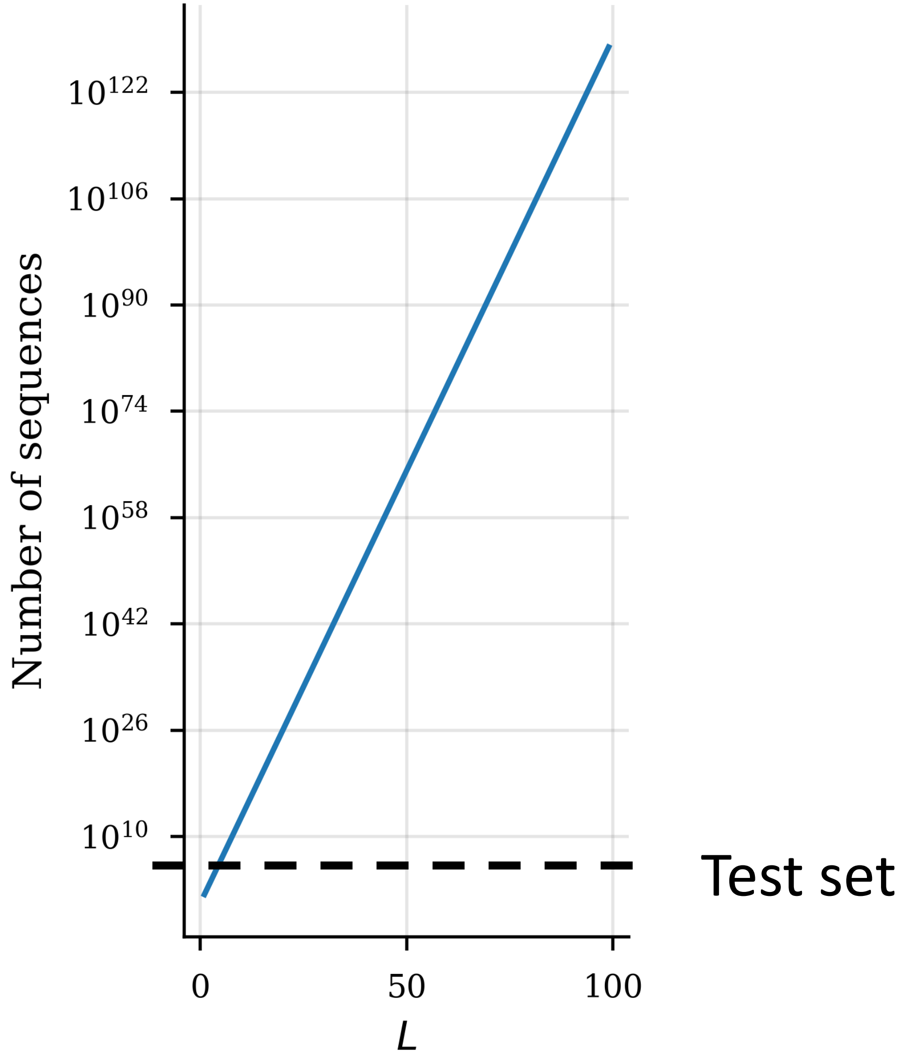
$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{M} \frac{p(\mathbf{x}_i^{(t)}|\boldsymbol{\theta}^{(0)})}{p(\mathbf{x}_i^{(t)}|\boldsymbol{\theta}^{(t)})} P(S^{(t)}|\mathbf{x}_i^{(t)}) \log p(\mathbf{x}_i^{(t)}|\boldsymbol{\theta})$$

Testing is fundamentally different

- We don’t trust our oracle and generally can’t query the ground truth

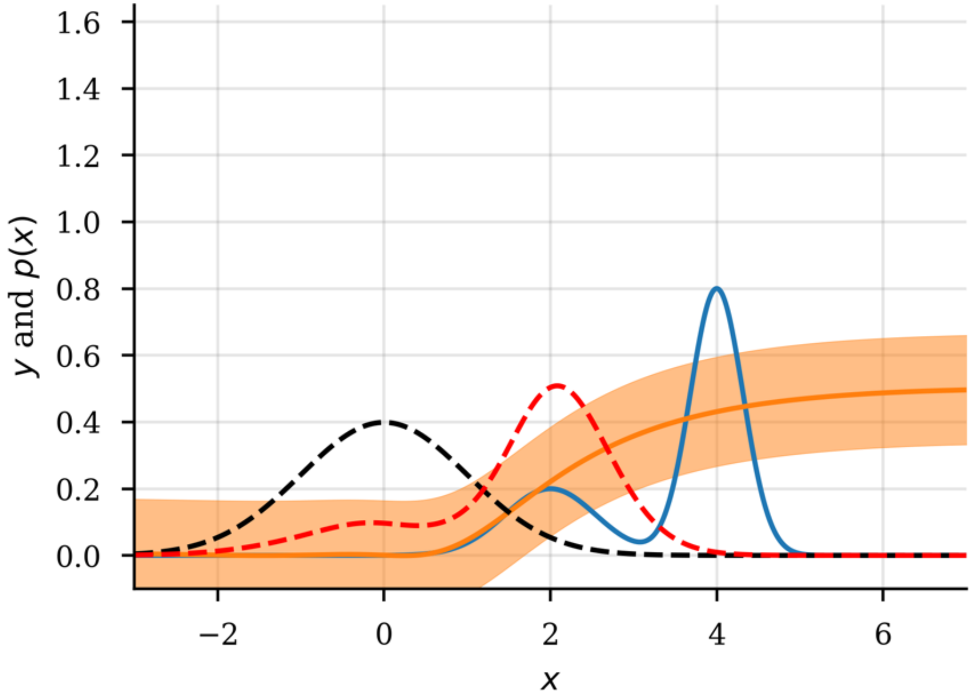
Testing is fundamentally different

- We don't trust our oracle and generally can't query the ground truth
- We can't hold-out a test set of good sequences
  - Near-zero chance of any of these sequences being found by the method



Testing is fundamentally different

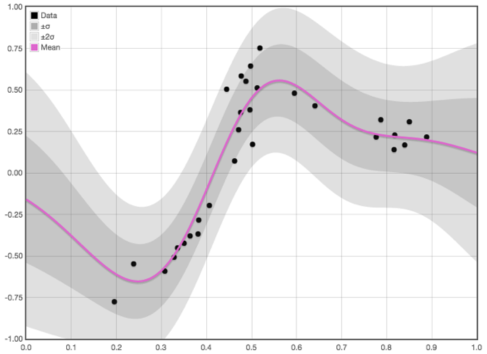
- We don’t trust our oracle and generally can’t query the ground truth
- We can’t hold-out a test set of good sequences
    - Near-zero chance of any of these sequences being found by the method
- We can’t use some canonical test function as the oracle
    - In our problem it is untrustworthy



Testing strategy

- Simulate a ground truth based on real data

  → “Ground truth” is a GP mean function

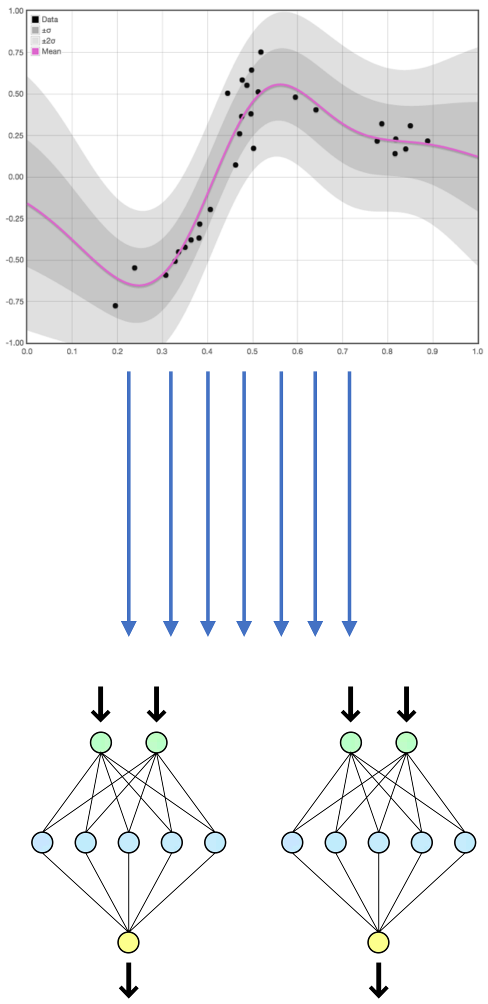


Ground truth GP

Testing strategy

- Simulate a ground truth based on real data

  → "Ground truth" is a GP mean function
- Ground truth vales values are sampled from the GP for given sequences
- Use these input-output pairs to train oracles.



Ground truth GP

Training data

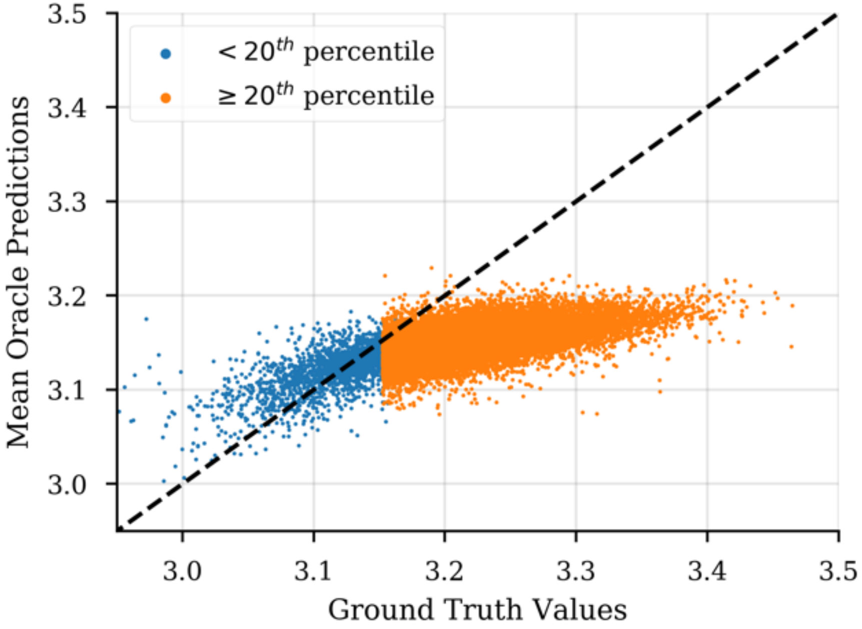
Oracles

Testing strategy

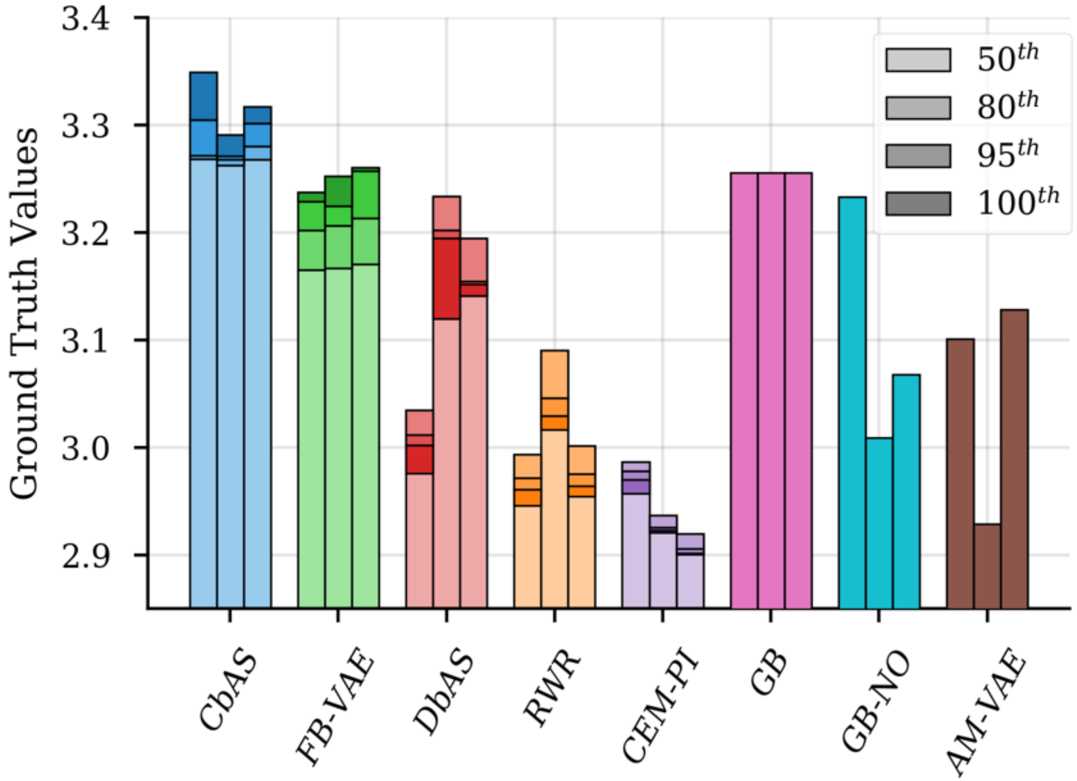
- Simulate a ground truth based on real data

  → "Ground truth" is a GP mean function

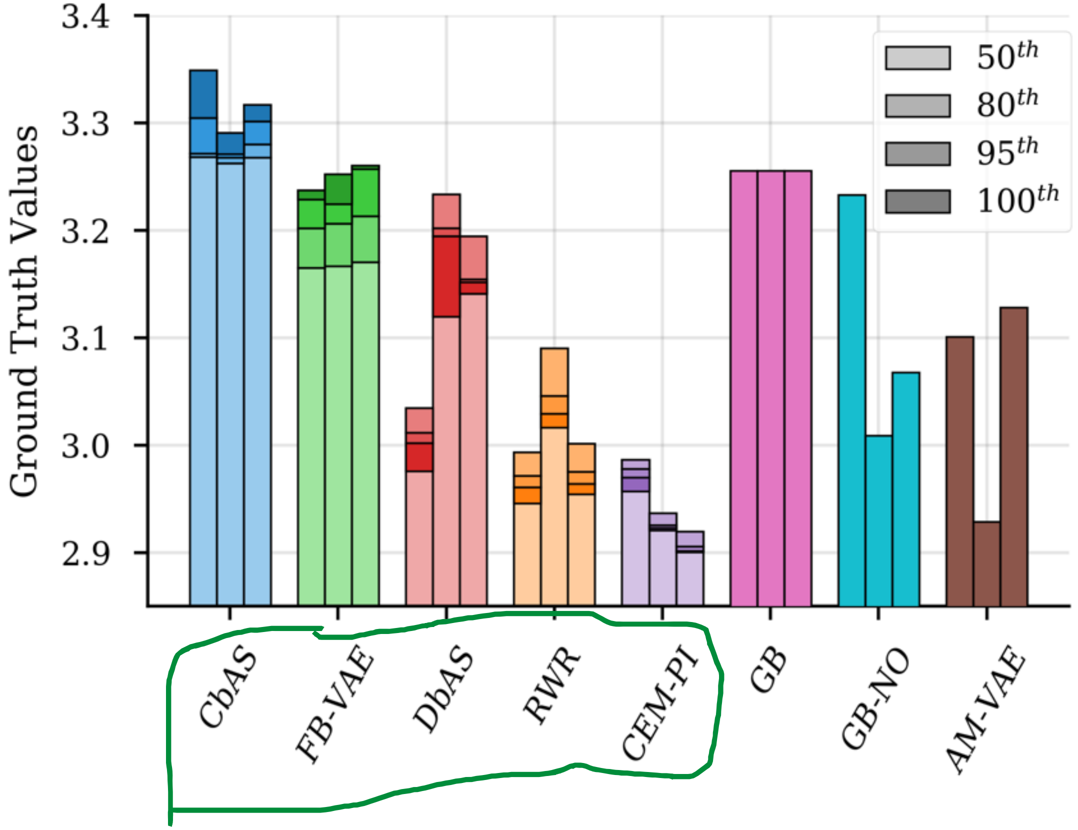
- Ground truth vales values are sampled from the GP for given sequences
- Use these input-output pairs to train oracles
- Coerce training set so these oracles exhibit pathologies



Results


Ground Truth Values
3.4
3.3
3.2
3.1
3.0
2.9
50th
80th
95th
100th
CbAS
FB-VAE
DbAS
RWR
CEM-PI
GB
GB-NO
AM-VAE

Results

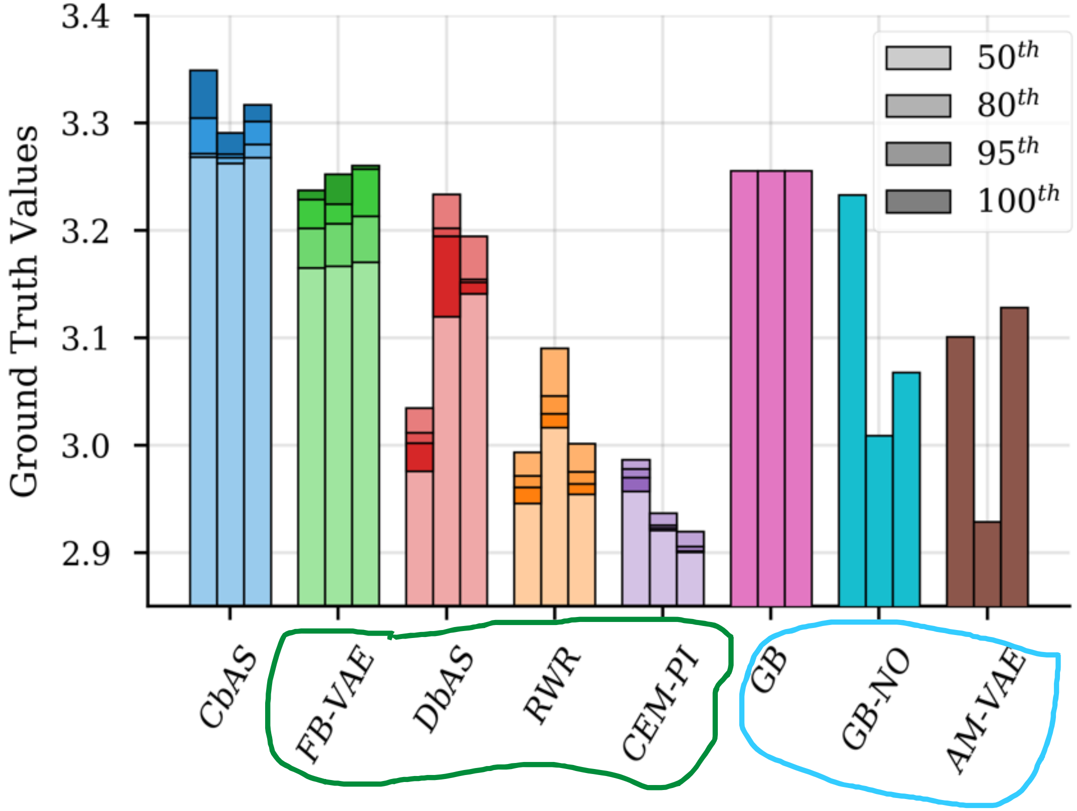


Model-based optimizations

Use weighted ML updates with weights:

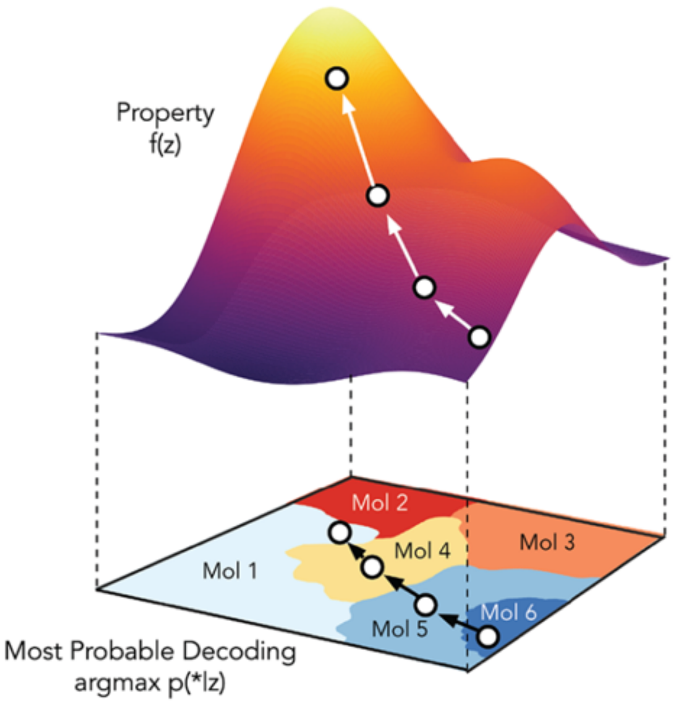
- *CbAS*: $\frac{p(x|\theta^{(0)})}{p(x|\theta^{(t)})} P(S^{(t)}|x)$
- *DbAS*: $P(S^{(t)}|x)$
- *RWR*: $e^{\alpha f(x)}$
- CEM-PI: $\mathbb{I}_{\{PI(x)>\gamma^{(t)}\}}(x)$
- FB-VAE: $\mathbb{I}_{\{f(x)>\gamma^{(t)}\}}(x)$ w/ additional considerations

Results

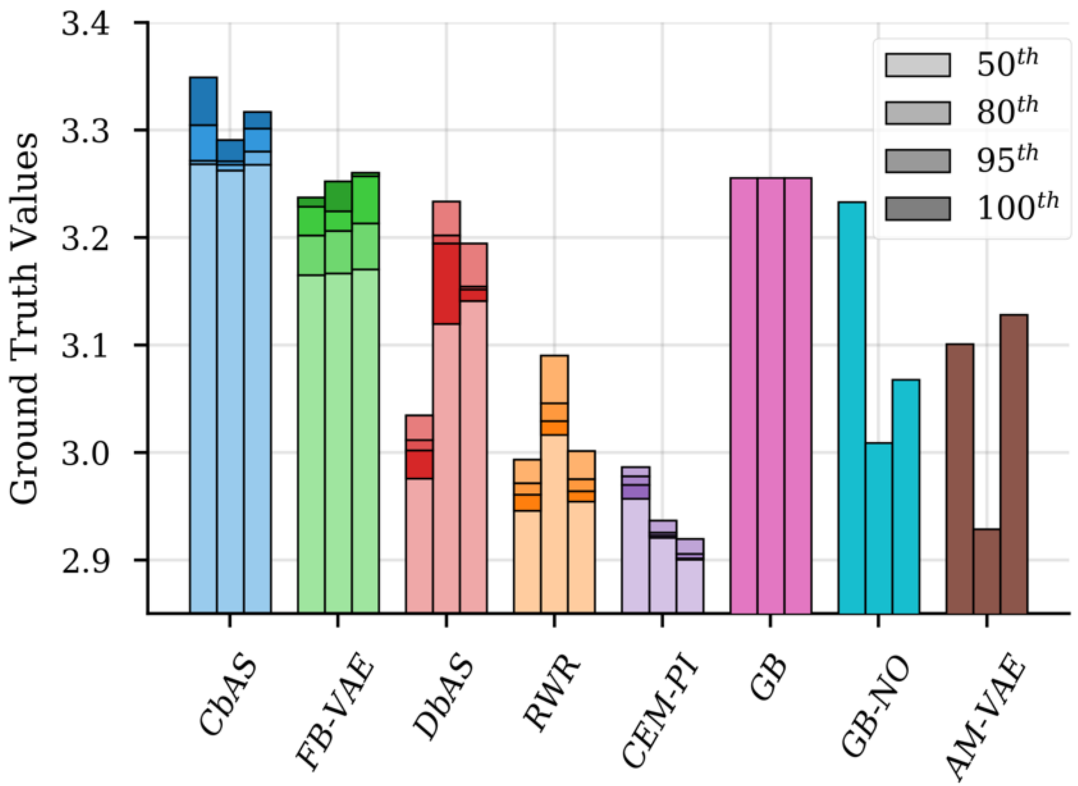


Model-based optimizations

Gradient descent on latent spaces



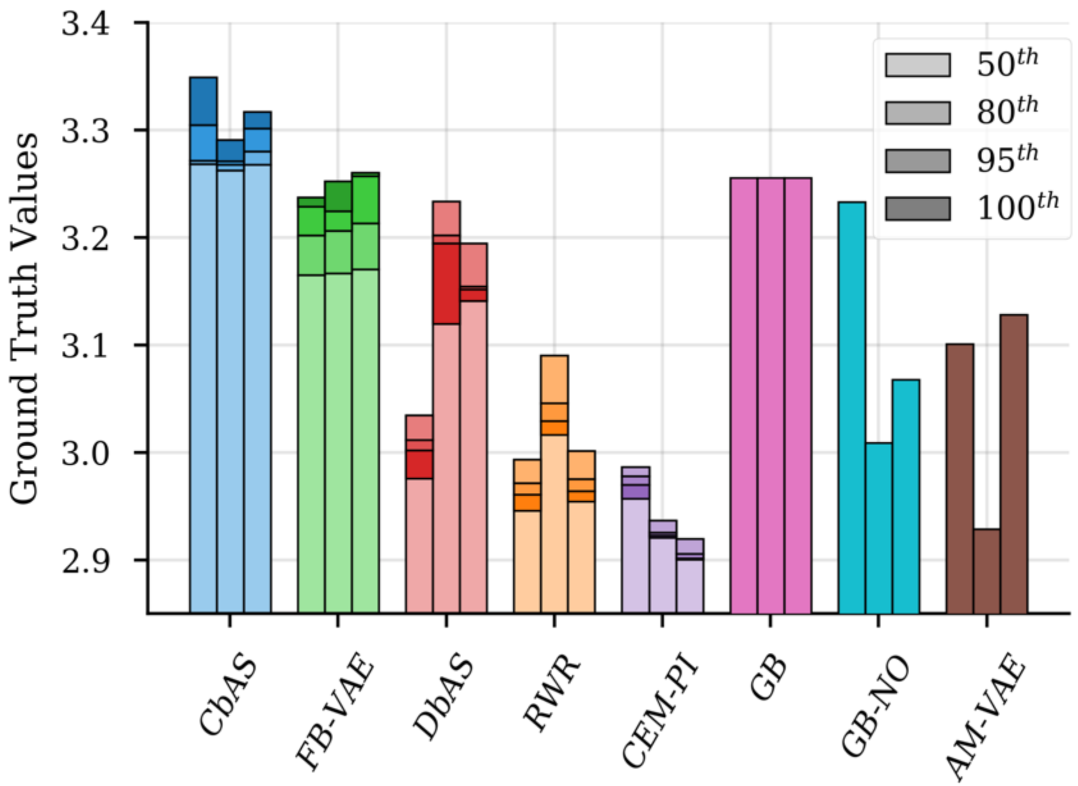
Results



What does each bar represent?

	Oracle	Ground truth
Seq1	🟩	🟨
Seq2	🟩	🟩
Seq3	🟩	🟩
Seq4	🟩	🟩
Seq5	🟨	🟨
Seq6	🟨	🟧
Seq7	🟧	🟨
Seq8	🟧	🟧
Seq9	🟧	🟧

Results



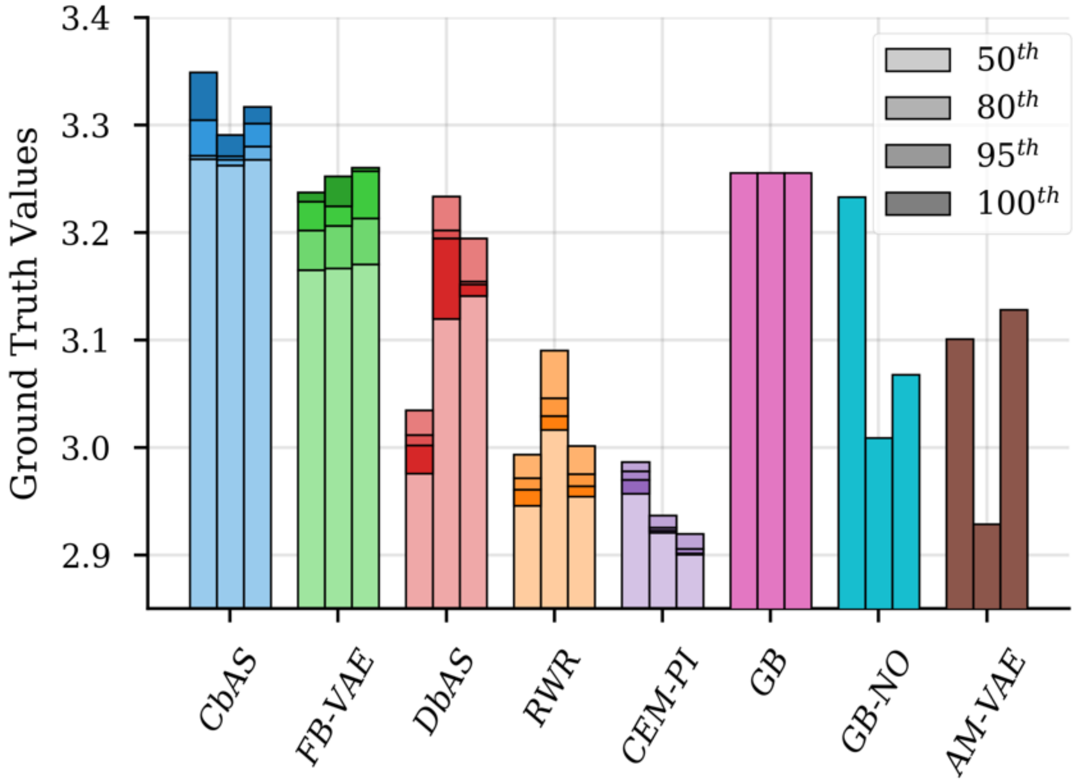
What does each bar represent?

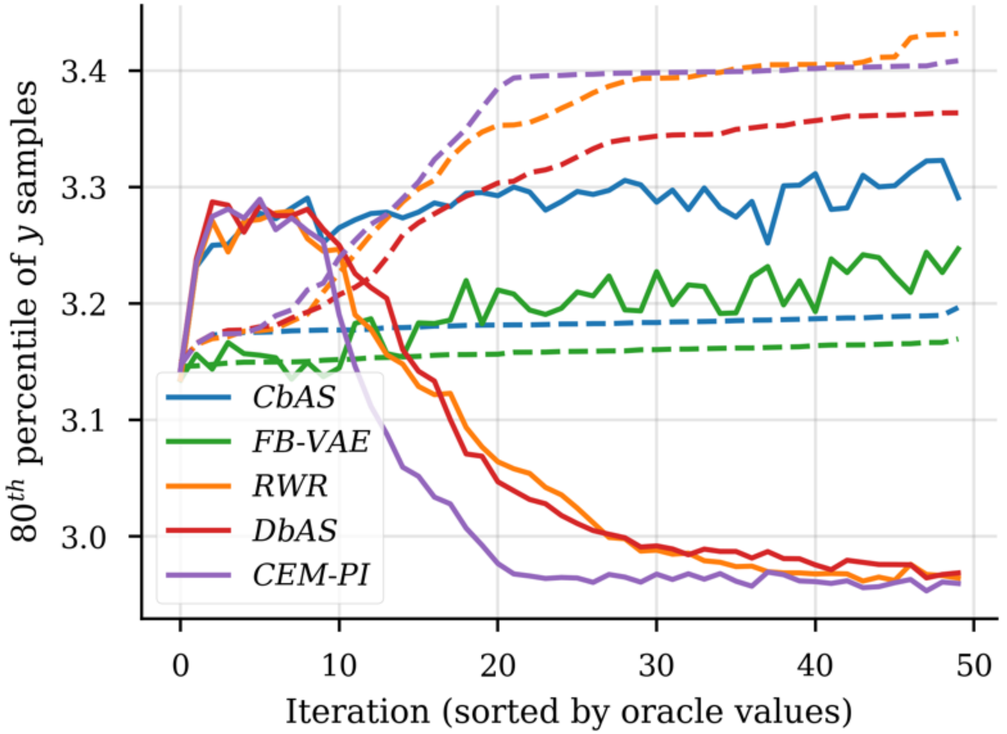
	Oracle	Ground truth
Seq1	🟩	🟨
Seq2	🟩	🟩
Seq3	🟩	🟩
Seq4	🟩	🟩
Seq5	🟨	🟨
Seq6	🟨	🟧
Seq7	🟧	🟨
Seq8	🟧	🟧
Seq9	🟧	🟧

Greater than i^{th} percentile

Mean

Results


Ground Truth Values
3.4
3.3
3.2
3.1
3.0
2.9
50th
80th
95th
100th
CbAS
FB-VAE
DbAS
RWR
CEM-PI
GB
GB-NO
AM-VAE


80th percentile of y samples
3.4
3.3
3.2
3.1
3.0
CbAS
FB-VAE
RWR
DbAS
CEM-PI
0
10
20
30
40
50
Iteration (sorted by oracle values)

Wrap-up

- Introduced a new model-based optimization method that is robust to pathological oracles
- Specifically targeted for discrete design problems
- Ongoing work to move beyond proof-of-principle:
  - Collaboration with wet-lab to perform end-to-end validation

Thanks!

Funding:

BERKELEY BIOPHYSICS

Berkeley EECS

NIH