SAND2023-02613C

Bayesian Optimal Experimental Designs for High-dimensional Physics-based Models

Jim Oreluk, Leonid Sheps, Habib N. Najm

joreluk@sandia.gov
Sandia National Laboratories, Livermore, CA, USA

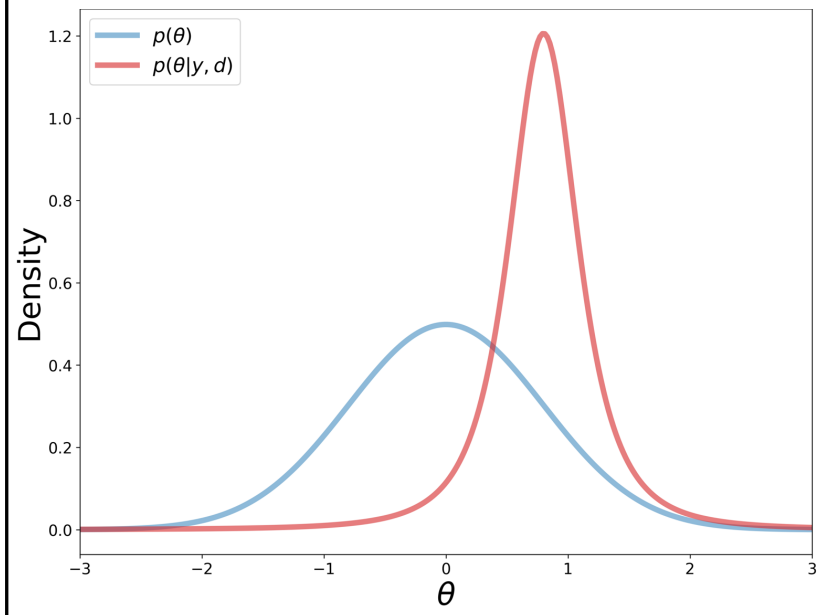
DATAWorks 2023

Sandia National Laboratories
U.S. DEPARTMENT OF ENERGY NNSA

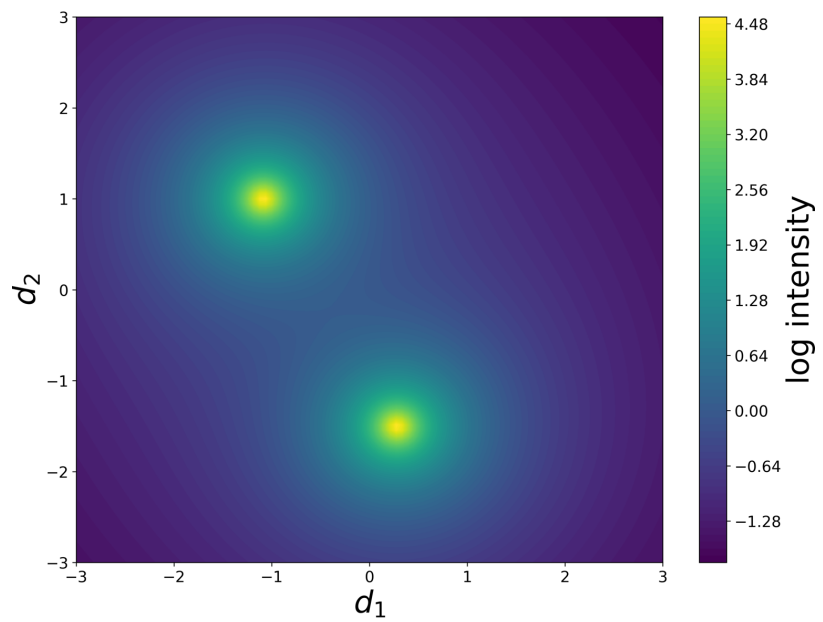
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Motivation

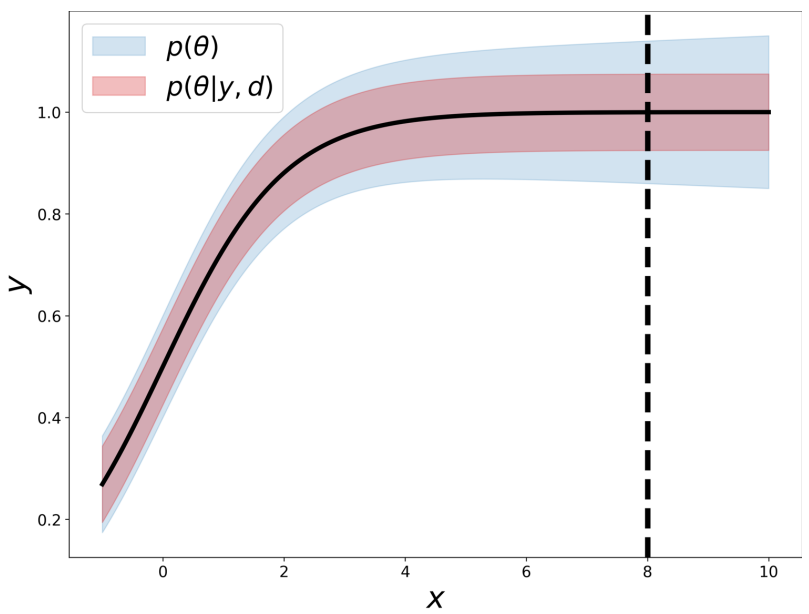
Experimental data helps us improve our understanding of a problem



Learning unknown parameter values



Identifying source locations in a field

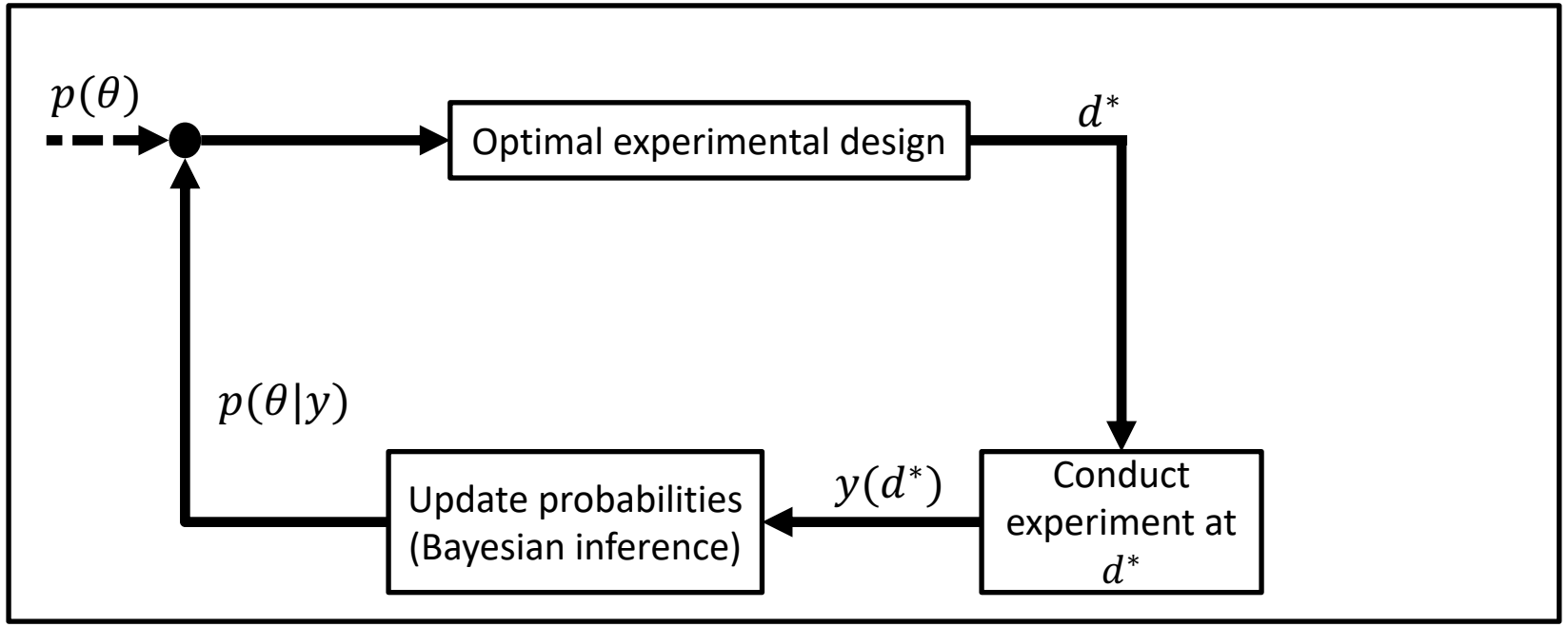


Improving predictivity of a QOI

Decision making problem

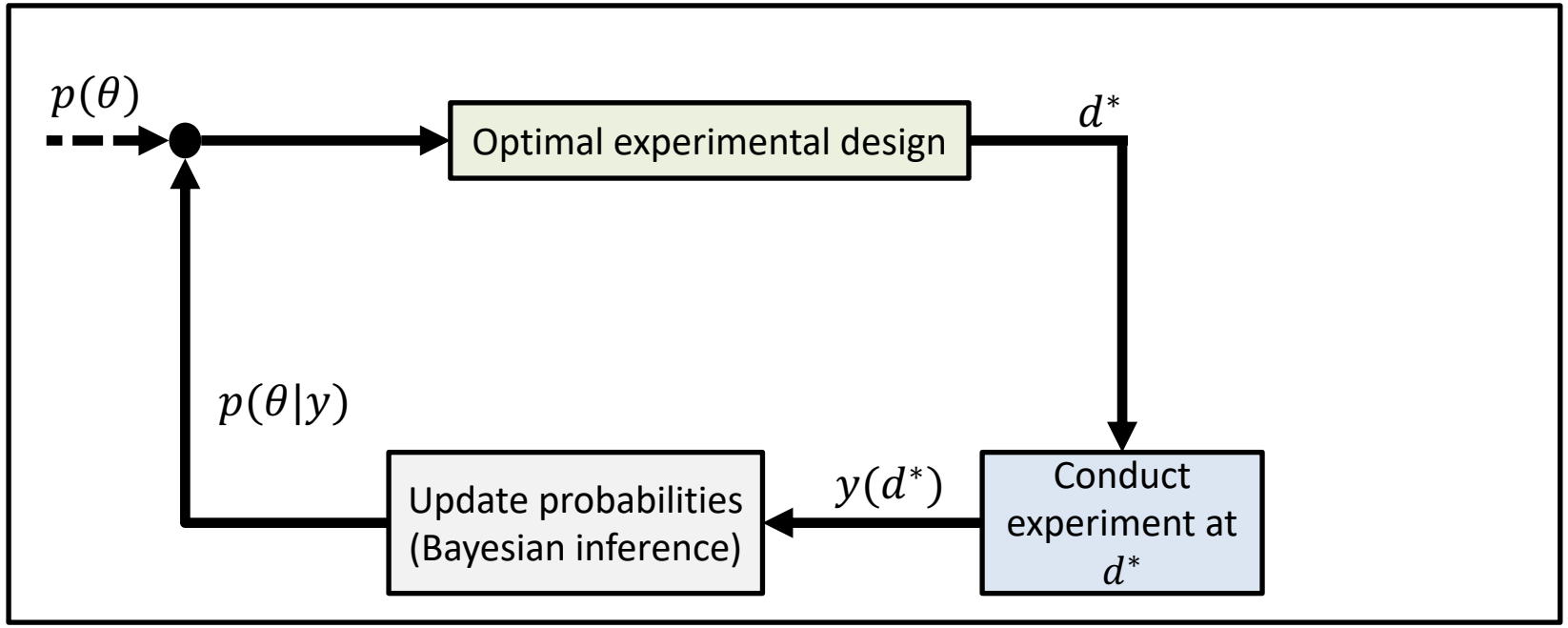
What experiment and/or test should we conduct that is **most** beneficial?

Optimal experimental design loop



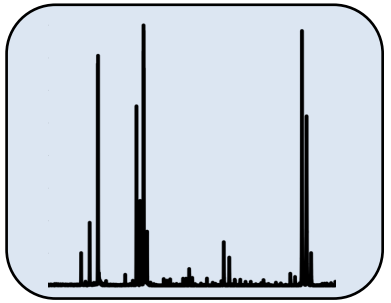
Optimal experimental design loop

$$d^* = \arg\max_{d \in \mathcal{D}} U(d)$$



Bayes' theorem

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$



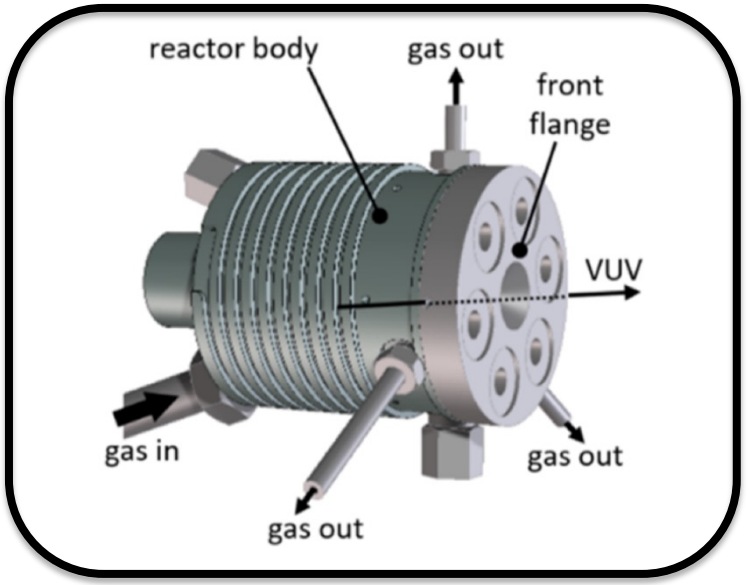
Experimental overview

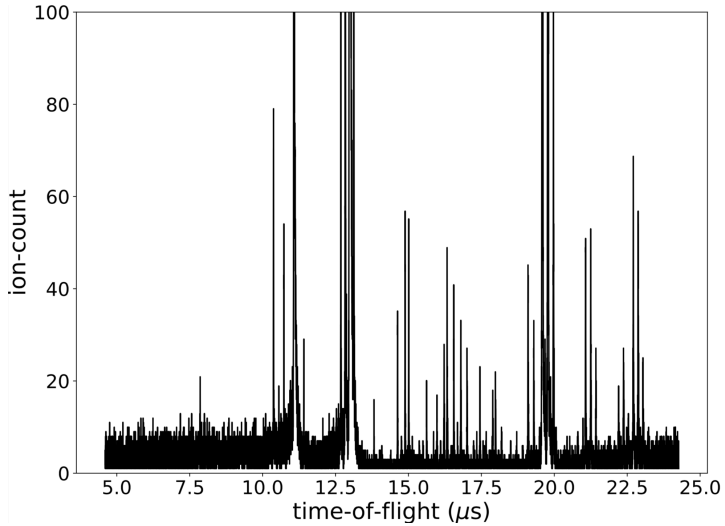
Objective: study the low-temperature oxidation of propane

- Measure the evolution of highly reactive intermediates and products

High-pressure photolysis reactor (HPR) experiment

- Premixed mixture flows into a constant pressure reactor
- Photolysis laser fires instantaneously irradiating the gas mixture
  - Chemical precursor breaks down initiating reactions
- Gas mixture exhausts out, sampled by a synchrotron tunable vacuum-ultraviolet (VUV) photoionization mass spectrometer
  - Measurement of time-of-flight mass spectrum





Time-of-flight mass spectrum at a fixed VUV energy (11.3 eV) and kinetic time (60 ms)

L. Sheps, I. Antonov, K. Au. Sensitive mass spectrometer for time-resolved gas-phase chemistry studies at high pressures. *The Journal of Physical Chemistry A* 123.50 (2019) 10804-10814.

Motivation

- Operation of the real experiment is costly and laborious
  - Potential setup time for the apparatus
  - Calibration of the apparatus and instrumentation
- Limited time to run experiments
  - Advanced Light Source, Lawrence Berkeley National Laboratory

Can we use a computational model to identify which experimental conditions are expected to be most informative?

Modeling the HPR experiment

Data model:

$$y(\boldsymbol{d}, \boldsymbol{x}) = \xi(\boldsymbol{d}, \boldsymbol{x}) + \epsilon(\boldsymbol{d}, \boldsymbol{x})$$

$$y(\boldsymbol{d}, \boldsymbol{x}) = f(\boldsymbol{\theta}, \boldsymbol{d}, \boldsymbol{x}) + \delta(\boldsymbol{x}) + \epsilon(\boldsymbol{d}, \boldsymbol{x})$$

$$\boldsymbol{x} = [\tau, t, E], \boldsymbol{\theta} \in \mathbb{R}^{1151}$$

$$\delta(\boldsymbol{x}) \sim GP(\mu_\delta(\boldsymbol{x}), \Sigma_\delta(\boldsymbol{x}, \boldsymbol{x}')), \epsilon(\boldsymbol{x}) \sim \mathcal{N}\left(0, s(\boldsymbol{x})^2\right)$$

$\boldsymbol{d}$: design conditions

$\boldsymbol{\theta}$: model parameters

$\boldsymbol{x}$: spatial/temporal coordinates

$y(\boldsymbol{d}, \boldsymbol{x})$: ion-count data

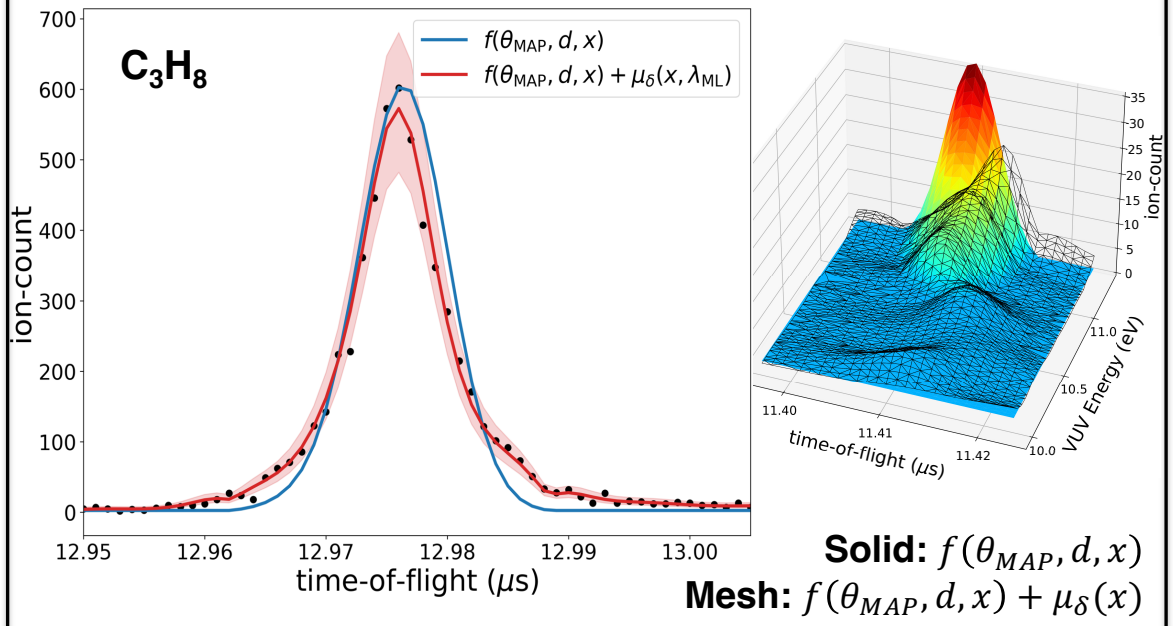
$\xi(\boldsymbol{d}, \boldsymbol{x})$: true physical process

$f(\boldsymbol{\theta}, \boldsymbol{d}, \boldsymbol{x})$: physics and instrument model

$\delta(\boldsymbol{x})$: model error

$\epsilon(\boldsymbol{d}, \boldsymbol{x})$: observation noise

- **Physics model**
  - Batch reactor model
  - Photolysis laser and ionization model [1]
- **Chemical model**
  - C0 - C3 chemical mechanism
  - 173 species / 1146 reactions [2]
- **Instrument model**
  - Maps concentrations to ion counts
  - Peaks idealized as Gaussian distributions



Solid: $f(\theta_{MAP}, d, x)$

Mesh: $f(\theta_{MAP}, d, x) + \mu_\delta(x)$

Predictions with model error provides additional fidelity to the physics model while being more inline with the data

[1] Oreluk, et al. *Combustion Theory and Modelling*, 2022.

[2] Miller, et al. *Progress in Energy and Combustion Science*, 2021.

Bayesian optimal experimental design

Objective

Find a set of experimental conditions that maximizes the expected utility $U(d)$

- Utility function models and compares the desirability of outcomes
- Target experiments to learn specific chemical rate constant measurements

$$d^* = \arg\max_{d \in \mathcal{D}} U(d)$$

$$\text{where,} \quad U(d) = \int_{y \in \mathcal{Y}} \int_{\theta \in \Theta} u(y, d, \theta) p(\theta, y|d) \mathrm{d}\theta \mathrm{dy}$$

$$= \int_{y \in \mathcal{Y}} \int_{\theta \in \Theta} u(y, d, \theta) p(\theta|y, d) p(y|d) \mathrm{d}\theta \mathrm{dy}$$

Notation

d : design conditions

θ : model parameters

y : data

Expected utility involves a double integral over data and model parameters.
Necessitates a model for the physical system and for the experimental instrument.
Uses Bayesian posterior on model parameters, and Bayesian model evidence on data.

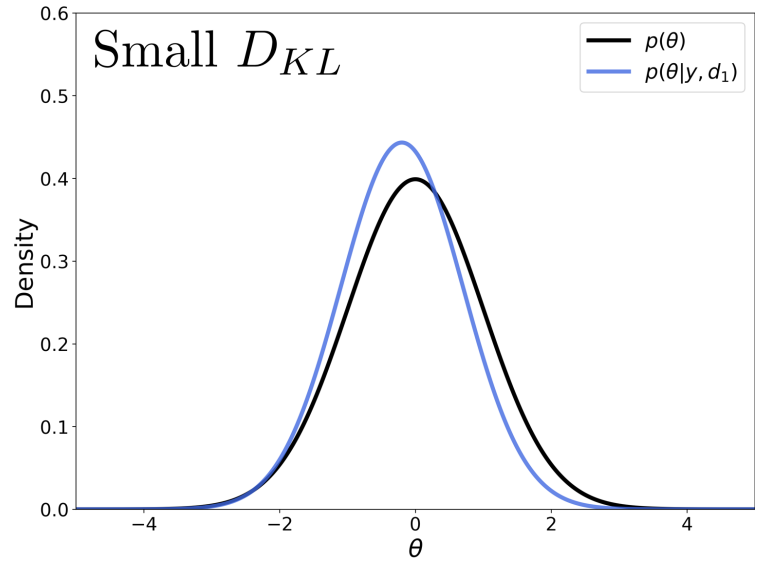
D.V. Lindley, On a measure of the information provided by an experiment, *The Annals of Mathematical Statistics* 27 (1956) 986–1005.
K. Chaloner and I. Verdinelli, Bayesian experimental design: A review. *Statistical Science* (1995) 273-304.

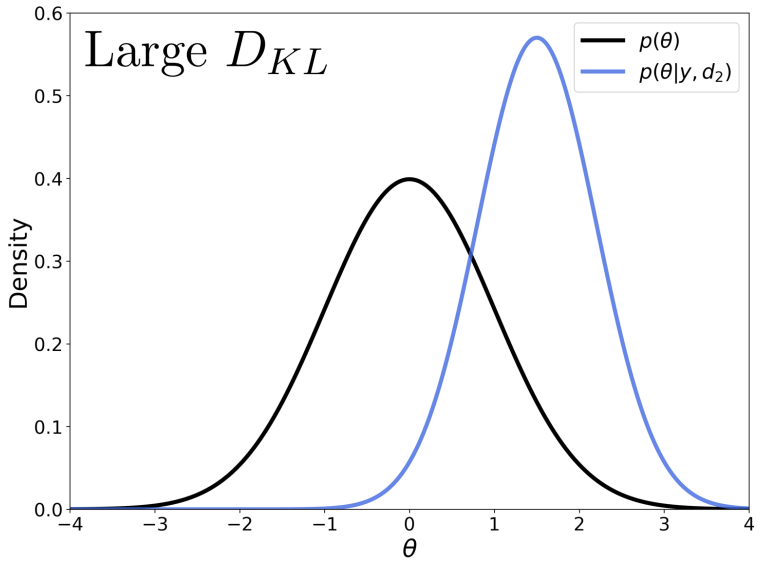
Selecting a utility function, $u(y, d)$

Utility function should *reflect the goal or purpose* of the experiment

1) Improve the estimate of a unknown model parameter
   - Information gain of an experiment is closely related to minimizing the parameter uncertainty

$$u(y,d) = D_{\mathrm{KL}}(p(\theta|y,d)||p(\theta)) = \int p(\theta|y,d) \log \left[\frac{p(\theta|y,d)}{p(\theta)} \right] \mathrm{d}\theta$$





Selecting a utility function, $u(y, d)$

Utility function should *reflect the goal or purpose* of the experiment

1) Improve the estimate of a unknown model parameter
   - Information gain of an experiment is closely related to minimizing the parameter uncertainty

$$u(y,d) = D_{\mathrm{KL}}(p(\theta|y,d)||p(\theta)) = \int p(\theta|y,d)\log\left[\frac{p(\theta|y,d)}{p(\theta)}\right]\mathrm{d}\theta$$

2) Improve the predictivity of a quantity of interest, q
   - Measure divergence between prior predictive and posterior predictive distributions

$$p(q) = \int_{\theta\in\Theta} p(q|\theta)p(\theta)\mathrm{d}\theta$$

$$p(q|y,d) = \int_{\theta\in\Theta} p(q|\theta)p(\theta|y,d)\mathrm{d}\theta$$

$$\begin{aligned} u(y,d) &= D_{\mathrm{KL}}(p(q|y,d)||p(q)) \\ &= \int p(q|y,d)\log\left[\frac{p(q|y,d)}{p(q)}\right]\mathrm{d}\theta \end{aligned}$$

Selecting a utility function, $u(y, d)$

Utility function should *reflect the goal or purpose* of the experiment

1) Improve the estimate of a unknown model parameter
   - Information gain of an experiment is closely related to minimizing the parameter uncertainty

$$u(y,d) = D_{\mathrm{KL}}(p(\theta|y,d)||p(\theta)) = \int p(\theta|y,d) \log\left[\frac{p(\theta|y,d)}{p(\theta)}\right] \mathrm{d}\theta$$

$$U(d) = \int_{\mathcal{Y}} \left(\int_{\Theta} p(\theta|y,d) \log\left[\frac{p(\theta|y,d)}{p(\theta)}\right] d\theta \right) p(y|d) dy$$

$$U(d) = \int_{\mathcal{Y}} \int_{\Theta} \left[\log p(y|\theta,d) - \log p(y|d)\right] p(y|\theta,d) p(\theta) d\theta dy$$

Approximating the expected utility

Numerical approximation:

$$U(d) \approx \frac{1}{N}\sum_{i=0}^{N}\left[\log p(y^{(i)}|\theta^{(i)},d) - \log \underbrace{p(y^{(i)}|d)}_{???}\right] \quad \text{where, } \theta^{(i)} \sim p(\theta),\ y^{(i)} \sim p(y|\theta^{(i)},d)$$

Nested Monte Carlo

$$U(d) \approx \frac{1}{N}\sum_{i=0}^{N}\left[\log p(y^{(i)}|\theta^{(i)},d) - \log\left(\frac{1}{M}\sum_{j=0}^{M} p(y^{(i)}|\theta^{(j)},d)\right)\right]$$

$$d^* = \arg\max_{d\in\mathcal{D}} U(d)$$

- *Costly to evaluate* $U(d)$
- $\frac{\mathrm{d}U}{\mathrm{d}d} = \ ?$

T. Rainforth et al., On nesting Monte Carlo estimators, *International Conference on Machine Learning*. PMLR, 2018.
K.J. Ryan, Estimating expected information gains for experimental designs with application to the random fatigue-limit model, *Journal of Computational and Graphical Statistics* 12 (2003) 585–603.

Maximizing the expected utility, $U(d)$

$$d^* = \arg\max_{d \in \mathcal{D}} U(d)$$

Bayesian Optimization

- Construct a GP of the unknown objective function

$$U(d) \sim \mathcal{N}\left(\mu\left(d\right), K\left(d, d'\right)\right)$$

- Use *acquisition function* $\alpha(d)$ to select new samples
  - Gaussian Process upper confidence bound (UCB)

$$\alpha_t\left(d\right) = \mu_{t-1}\left(d\right) + \sqrt{\beta}_t \sigma_{t-1}\left(d\right)$$

$$\sigma_{t-1}\left(d\right) = \sqrt{K\left(d, d\right)}$$

  - Exploits regions with a high mean and explores regions of high uncertainty

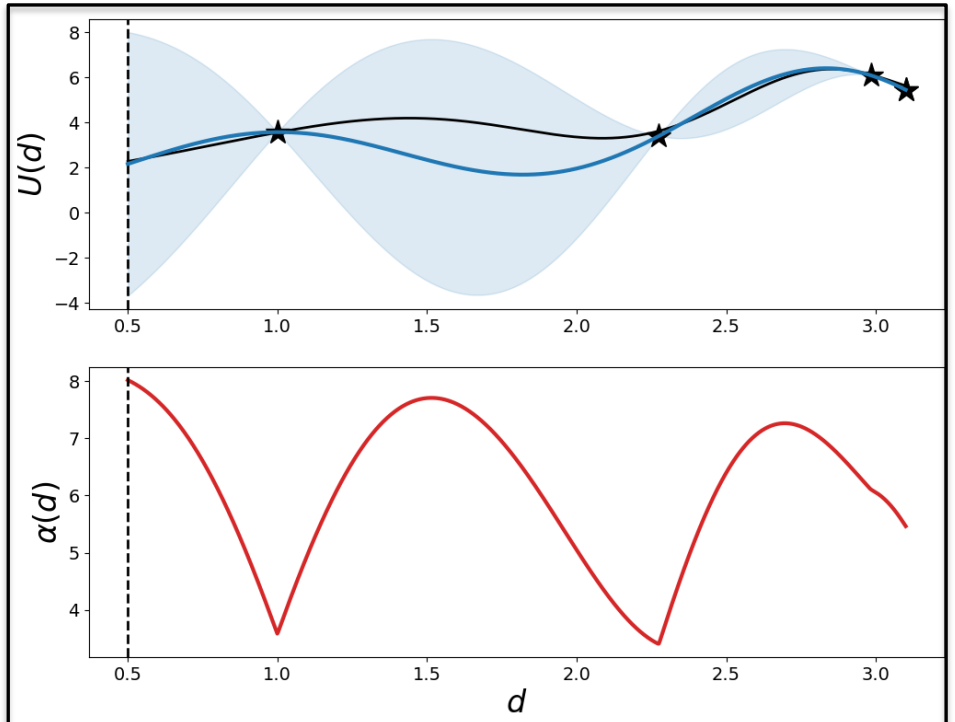


Illustration of an iteration of Bayesian optimization, unknown objective function (top), acquisition function (bottom)

- Select next optimization iteration at: $d_t = \arg\max_{d \in \mathcal{D}} \alpha_t\left(d\right)$
- Evaluate utility function at $U(d_t)$

N. Srinivas, A. Krause, S. Kakade, & M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv:0912.3995* (2009).

Challenges

High-dimensional model output

- High-fidelity physics-based simulations are be expensive to evaluate
- Evaluating $U(d) \sim \mathcal{O}(NM)$ forward solves

$$U(d) \approx \frac{1}{N} \sum_{i=0}^{N} \left[\log p(y^{(i)}|\theta^{(i)}, d) - \log \left(\frac{1}{M} \sum_{j=0}^{M} p(y^{(i)}|\theta^{(j)}, d) \right) \right]$$

- Constructing a surrogate model addresses the costly run-time

$$f(\boldsymbol{\theta}, \boldsymbol{d}, x_i) \approx g(\boldsymbol{\theta}, \boldsymbol{d}, x_i) = \sum_{p=1}^{P} c_p \Psi_p(\xi)$$

  - Total **number of outputs** remains problematic $y(\boldsymbol{d}, \boldsymbol{x}) \in \mathbb{R}^{510{,}000{,}000}$

Challenges

High-dimensional model output

- High-fidelity physics-based simulations are be expensive to evaluate
- Evaluating $U(d) \sim \mathcal{O}(NM)$ forward solves

$$U(d) \approx \frac{1}{N} \sum_{i=0}^{N} \left[\log p(y^{(i)}|\theta^{(i)}, d) - \log \left(\frac{1}{M} \sum_{j=0}^{M} p(y^{(i)}|\theta^{(j)}, d) \right) \right]$$

- Constructing a surrogate model addresses the costly run-time

$$f(\boldsymbol{\theta}, \boldsymbol{d}, x_i) \approx g(\boldsymbol{\theta}, \boldsymbol{d}, x_i) = \sum_{p=1}^{P} c_p \Psi_p(\xi)$$

Can we find *low-dimensional* representations of the high-dimensional output?

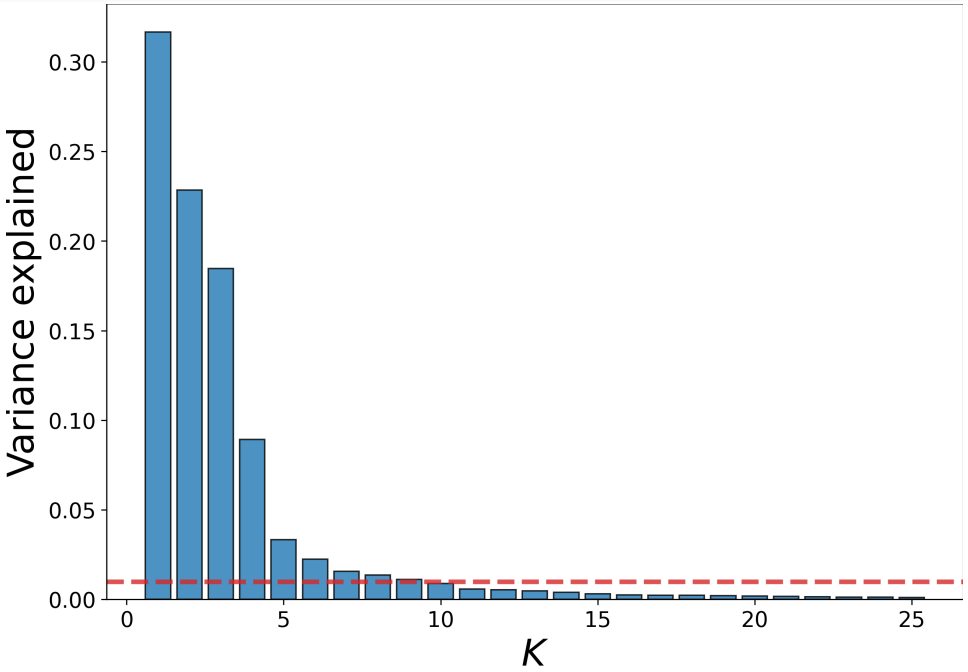
Reducing output dimensionality

Goal: Map high-dimensional model output to a lower-dimensional space while minimizing loss of information

Truncated SVD

At a fixed design d,

- Draw n samples, where $\boldsymbol{\theta}^{(i)} \sim p(\boldsymbol{\theta})$
- Evaluate the model $f(\boldsymbol{\theta}, \boldsymbol{d}, \boldsymbol{x}) + \delta(\boldsymbol{\lambda}_{\mathrm{ML}}, \boldsymbol{x}) + \epsilon(\boldsymbol{x})$
- Construct output matrix $\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T$, where $\boldsymbol{Y} \in \mathbb{R}^{n \times J}$, $J = 5.1 \times 10^8$
- Retain only top K singular values of $\boldsymbol{S}$
- Low-rank approximation: $\boldsymbol{Y}_K = \boldsymbol{U}_K \boldsymbol{S}_K \boldsymbol{V}_K^T$



93% of the total variance is explained by a small number ($K = 10$) components at 620 K, 2 bar, $n_{\mathrm{C3H8}} = 2\mathrm{e}14, n_{\mathrm{O2}} = 8\mathrm{e}18,\ n_{\mathrm{Cl}} = 1\mathrm{e}13$.

Transformation:

$$q(\boldsymbol{\theta}, \boldsymbol{d}, \boldsymbol{x}) = y(\boldsymbol{\theta}, \boldsymbol{d}, \boldsymbol{x})\, \boldsymbol{V}_K$$

$$\underbrace{q(\boldsymbol{\theta}, \boldsymbol{d}, \boldsymbol{x})}_{(1 \times K)} = \underbrace{[f(\boldsymbol{\theta}, \boldsymbol{d}, \boldsymbol{x}) + \delta(\boldsymbol{\lambda}_{\mathrm{ML}}, \boldsymbol{x}) + \epsilon(\boldsymbol{x})]}_{(1 \times J)} \underbrace{\boldsymbol{V}_K}_{(J \times K)}$$

Challenges: Target *specific* model parameters

Expected utility measures difference between $p(\theta)$ and $p(\theta|y)$, in other words, difference in ***all uncertain*** *parameters*

- Only interested in designs that improve *particular chemical rates*
  - chemical model contains 1146 uncertain chemical rates

Set of targeted reaction rates

Reaction ID	
	n-R + O2 ⟶ QOOH3
1040	O2 + CH3CH2CH2 ⟶ CH3CHCH2OOH
	n-R + O2 ⟶ QOOH1
1041	O2 + CH3CH2CH2 ⟶ CH2CH2CH2OOH
1042	O2 + CH3CH2CH2 ⟶ CH2CH2CH2OOH
	n-R + O2 ⟶ propene
1043	O2 + CH3CH2CH2 ⟶ HO2 + CH3CHCH2
	n-R + O2 ⟶ propylene oxide
1044	O2 + CH3CH2CH2 ⟶ OH + C-CH2OCH(CH3)
	n-RO2 ⟶ products
1045	CH3CH2CH2OO ⟶ CH3CHCH2OOH
1046	CH3CH2CH2OO ⟶ CH2CH2CH2OOH
1047	CH3CH2CH2OO ⟶ CH2CH2CH2OOH
1048	CH3CH2CH2OO ⟶ HO2 + CH3CHCH2
1049	CH3CH2CH2OO ⟶ OH + C-CH2OCH(CH3)
	QOOH3 ⟶ products
1051	CH3CHCH2OOH ⟶ HO2 + CH3CHCH2
1052	CH3CHCH2OOH ⟶ OH + C-CH2OCH(CH3)
	QOOH1 ⟶ propene
1053	CH2CH2CH2OOH ⟶ HO2 + CH3CHCH2
	i-R+O2 ⟶ products
1056	O2 + CH3CHCH3 ⟶ CH3CH(OOH)CH2
1057	O2 + CH3CHCH3 ⟶ HO2 + CH3CHCH2
1058	O2 + CH3CHCH3 ⟶ OH + C-CH2OCH(CH3)
	i-RO2 ⟶ products
1059	CH3CH(OO)CH3 ⟶ CH3CH(OOH)CH2
1060	CH3CH(OO)CH3 ⟶ HO2 + CH3CHCH2
1061	CH3CH(OO)CH3 ⟶ OH + C-CH2OCH(CH3)
	QOOH2 ⟶ products
1062	CH3CH(OOH)CH2 ⟶ HO2 + CH3CHCH2
1063	CH3CH(OOH)CH2 ⟶ OH + C-CH2OCH(CH3)
	O2 + QOOH1 ⟶ products
1065	O2 + CH2CH2CH2OOH ⟶ OHOCH2CH2CH2OO
	O2 + QOOH3 ⟶ products
1069	O2 + CH3CHCH2OOH ⟶ CH3CH(OO)CH2OOH
1071	O2 + CH3CHCH2OOH ⟶ CH2CH(OOH)CH2OOH
	O2 + QOOH2 ⟶ products
1080	O2 + CH3CH(OOH)CH2 ⟶ CH3CH(OOH)CH2OO
1081	O2 + CH3CH(OOH)CH2 ⟶ CH2CH(OOH)CH2OOH
	3−hydroperoxy−propylperoxyradical ⟶ KHP
1088	OHOCH2CH2CH2OO ⟶ OH + OCHCH2CH2OOH
	KHP ⟶ products
1112	OCHCH2CH2OOH ⟶ OCHCH2CH2O + OH

Challenges: Target *specific* model parameters

Expected utility measures difference between $p(\theta)$ and $p(\theta|y)$, in other words, difference in ***all uncertain*** *parameters*

- Only interested in designs that improve *particular chemical rates*
  - chemical model contains 1146 uncertain chemical rates

Solution: Reformulate utility to marginalize over nuisance parameters

Let θ be the parameters of interest and η be the nuisance parameters

Nested Monte Carlo estimator becomes:

$$U(d) \approx \frac{1}{N}\sum_{i=1}^{N}\left[\log\left(\frac{1}{K}\sum_{k=1}^{K} p\left(y^{(i)}|\theta^{(i)},\eta^{(k)},d\right)\right) - \log\left(\frac{1}{M}\sum_{j=1}^{M} p\left(y^{(i)}|\theta^{(j)},\eta^{(j)},d\right)\right)\right]$$

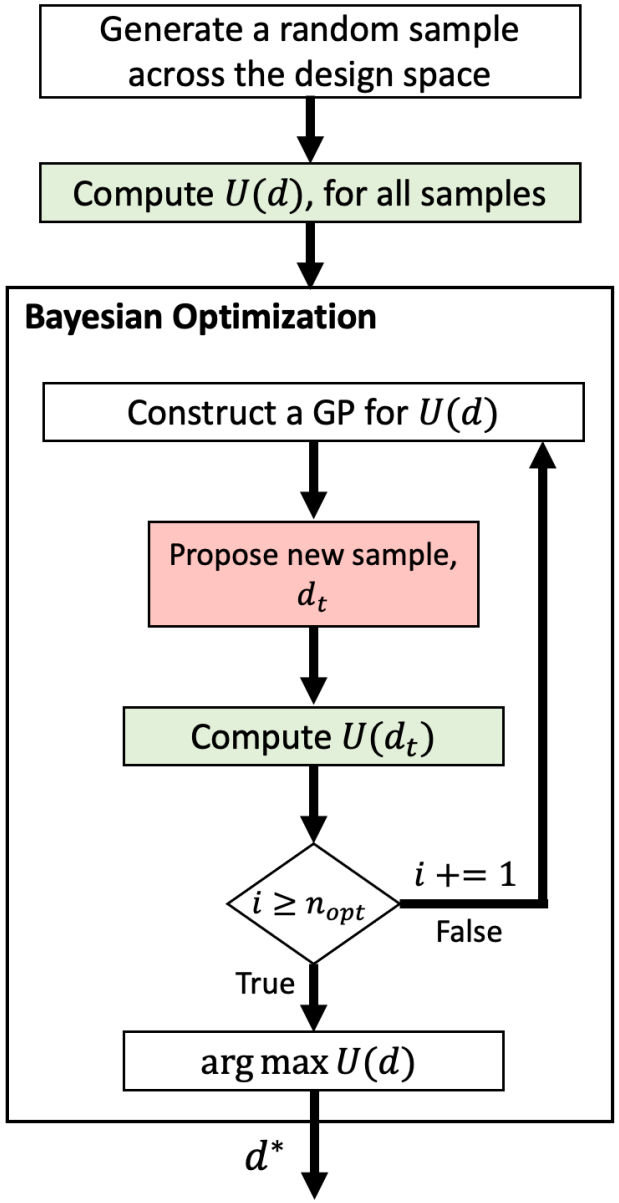
Additional loop in likelihood to marginalize over nuisance parameters

Set of targeted reaction rates

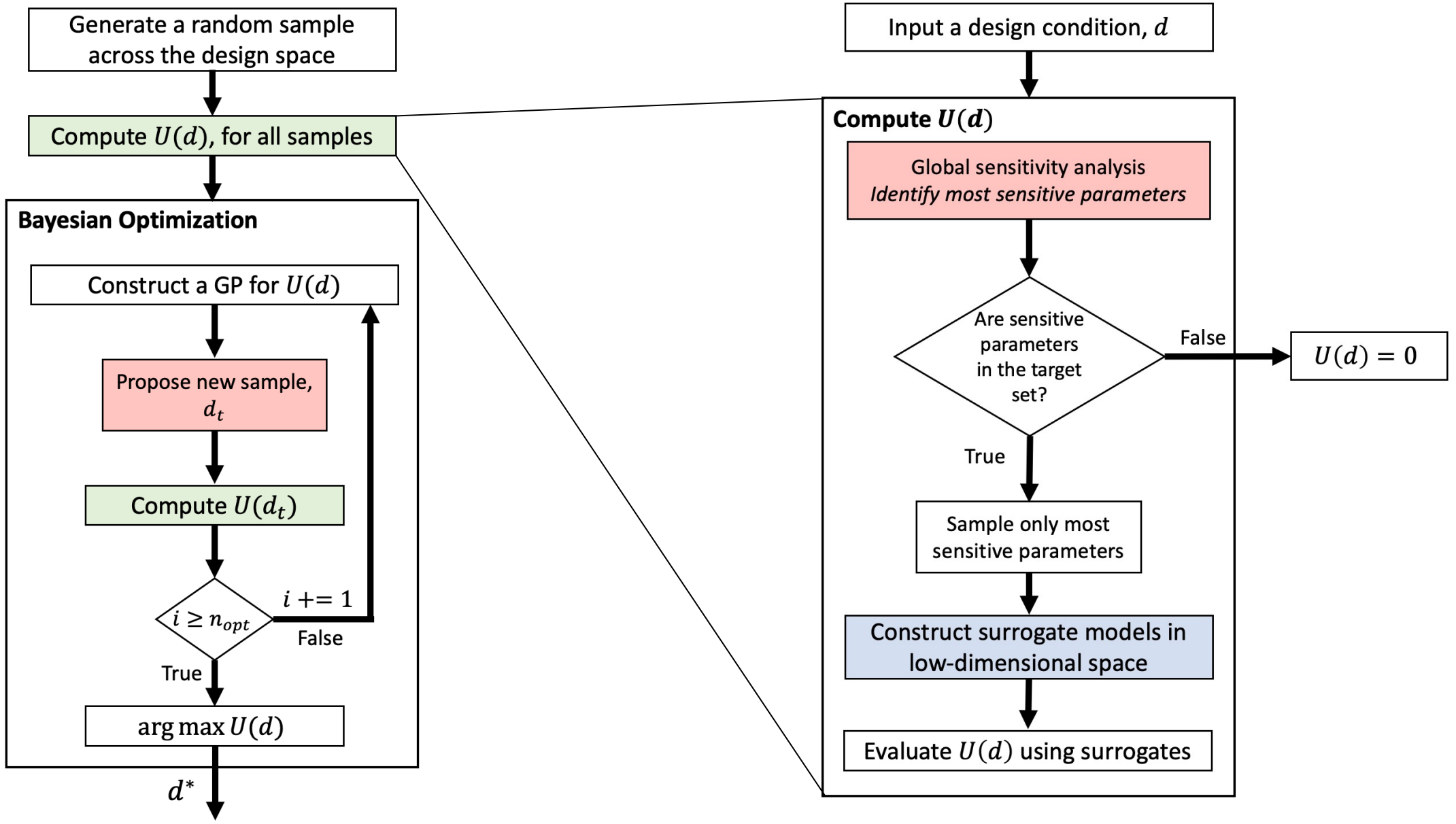
Reaction ID	
	n-R + O2 ⟶ QOOH3
1040	O2 + CH3CH2CH2 ⟶ CH3CHCH2OOH
	n-R + O2 ⟶ QOOH1
1041	O2 + CH3CH2CH2 ⟶ CH2CH2CH2OOH
1042	O2 + CH3CH2CH2 ⟶ CH2CH2CH2OOH
	n-R + O2 ⟶ propene
1043	O2 + CH3CH2CH2 ⟶ HO2 + CH3CHCH2
	n-R + O2 ⟶ propylene oxide
1044	O2 + CH3CH2CH2 ⟶ OH + C-CH2OCH(CH3)
	n-RO2 ⟶ products
1045	CH3CH2CH2OO ⟶ CH3CHCH2OOH
1046	CH3CH2CH2OO ⟶ CH2CH2CH2OOH
1047	CH3CH2CH2OO ⟶ CH2CH2CH2OOH
1048	CH3CH2CH2OO ⟶ HO2 + CH3CHCH2
1049	CH3CH2CH2OO ⟶ OH + C-CH2OCH(CH3)
	QOOH3 ⟶ products
1051	CH3CHCH2OOH ⟶ HO2 + CH3CHCH2
1052	CH3CHCH2OOH ⟶ OH + C-CH2OCH(CH3)
	QOOH1 ⟶ propene
1053	CH2CH2CH2OOH ⟶ HO2 + CH3CHCH2
	i-R+O2 ⟶ products
1056	O2 + CH3CHCH3 ⟶ CH3CH(OOH)CH2
1057	O2 + CH3CHCH3 ⟶ HO2 + CH3CHCH2
1058	O2 + CH3CHCH3 ⟶ OH + C-CH2OCH(CH3)
	i-RO2 ⟶ products
1059	CH3CH(OO)CH3 ⟶ CH3CH(OOH)CH2
1060	CH3CH(OO)CH3 ⟶ HO2 + CH3CHCH2
1061	CH3CH(OO)CH3 ⟶ OH + C-CH2OCH(CH3)
	QOOH2 ⟶ products
1062	CH3CH(OOH)CH2 ⟶ HO2 + CH3CHCH2
1063	CH3CH(OOH)CH2 ⟶ OH + C-CH2OCH(CH3)
	O2 + QOOH1 ⟶ products
1065	O2 + CH2CH2CH2OOH ⟶ OHOCH2CH2CH2OO
	O2 + QOOH3 ⟶ products
1069	O2 + CH3CHCH2OOH ⟶ CH3CH(OO)CH2OOH
1071	O2 + CH3CHCH2OOH ⟶ CH2CH(OOH)CH2OOH
	O2 + QOOH2 ⟶ products
1080	O2 + CH3CH(OOH)CH2 ⟶ CH3CH(OOH)CH2OO
1081	O2 + CH3CH(OOH)CH2 ⟶ CH2CH(OOH)CH2OOH
	3−hydroperoxy−propylperoxyradical ⟶ KHP
1088	OHOCH2CH2CH2OO ⟶ OH + OCHCH2CH2OOH
	KHP ⟶ products
1112	OCHCH2CH2OOH ⟶ OCHCH2CH2O + OH

Feng, Chi, and Youssef M. Marzouk. "A layered multiple importance sampling scheme for focused optimal Bayesian experimental design." *arXiv preprint arXiv:1903.11187* (2019).

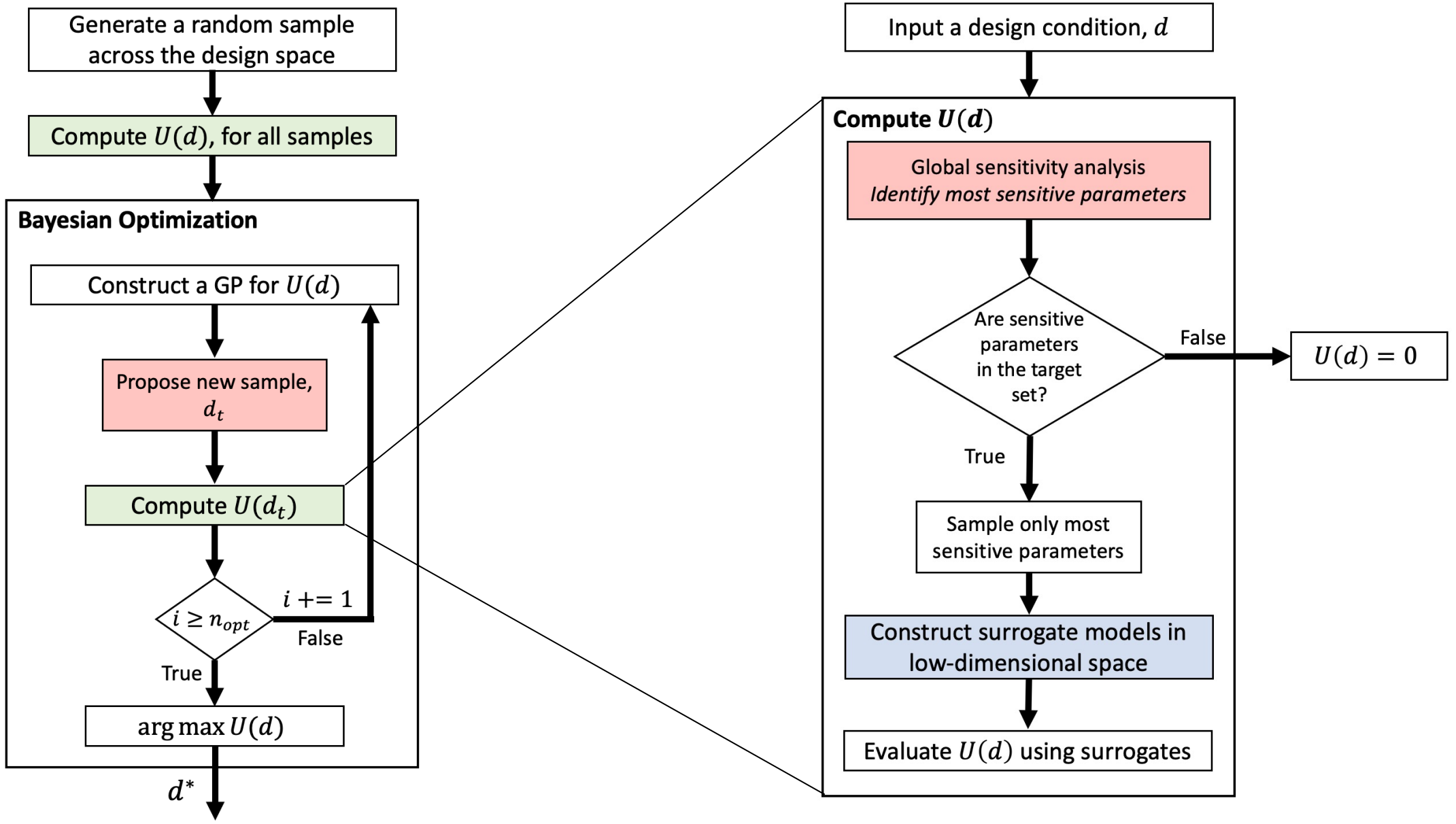
Targeted OED workflow



Targeted OED workflow



Targeted OED workflow



Results

Original data model,

$$y(\boldsymbol{d}, \boldsymbol{x}) = \xi(\boldsymbol{d}, \boldsymbol{x}) + \epsilon(\boldsymbol{d}, \boldsymbol{x})$$
$$y(\boldsymbol{d}, \boldsymbol{x}) = f(\boldsymbol{\theta}, \boldsymbol{d}, \boldsymbol{x}) + \delta(\boldsymbol{x}) + \epsilon(\boldsymbol{d}, \boldsymbol{x})$$
$$\boldsymbol{x} = [\tau, t, E], \boldsymbol{\theta} \in \mathbb{R}^{1151}$$
$$\delta(\boldsymbol{x}) \sim GP(\mu_\delta(\boldsymbol{x}), \Sigma_\delta(\boldsymbol{x}, \boldsymbol{x}')), \epsilon(\boldsymbol{x}) \sim \mathcal{N}\left(0, s(\boldsymbol{x})^2\right)$$

Design space: $\boldsymbol{d} = [T, p, n_{Cl}, n_{O2}]$

$$\Sigma_\delta = \begin{bmatrix} \Sigma_1 & & & & & \\ & \ddots & & & & \\ & & \Sigma_D & & & \\ & & & \ddots & & \\ & & & & \Sigma_{Dt_e+1} & \\ & & & & & \ddots \\ & & & & & & \Sigma_{D(t_e+1)} \end{bmatrix}$$

Assumptions:

- **Model error $\boldsymbol{\delta(x)}$ independent of $\boldsymbol{d}$**
- Local GPs constructed independently for each kinetic time
- Considering all 1151 model parameters as uncertain
  - Designs are constructed to target specific sets of the model parameters
- Prior parameter uncertainty is specified by literature

Results

Original data model,

$$y(\boldsymbol{d}, \boldsymbol{x}) = \xi(\boldsymbol{d}, \boldsymbol{x}) + \epsilon(\boldsymbol{d}, \boldsymbol{x})$$

$$y(\boldsymbol{d}, \boldsymbol{x}) = f(\boldsymbol{\theta}, \boldsymbol{d}, \boldsymbol{x}) + \delta(\boldsymbol{x}) + \epsilon(\boldsymbol{d}, \boldsymbol{x})$$

$$\boldsymbol{x} = [\tau, t, E], \boldsymbol{\theta} \in \mathbb{R}^{1151}$$

$$\delta(\boldsymbol{x}) \sim GP(\mu_\delta(\boldsymbol{x}), \Sigma_\delta(\boldsymbol{x}, \boldsymbol{x}')), \epsilon(\boldsymbol{x}) \sim \mathcal{N}\left(0, s(\boldsymbol{x})^2\right)$$

Design space: $\boldsymbol{d} = [T, p, n_{Cl}, n_{O2}]$

$$\Sigma_\delta = \begin{bmatrix} \Sigma_1 & & & & \\ & \ddots & & & \\ & & \Sigma_D & & \\ & & & \ddots & \\ & & & & \Sigma_{Dt_e+1} & & \\ & & & & & \ddots & \\ & & & & & & \Sigma_{D(t_e+1)} \end{bmatrix}$$

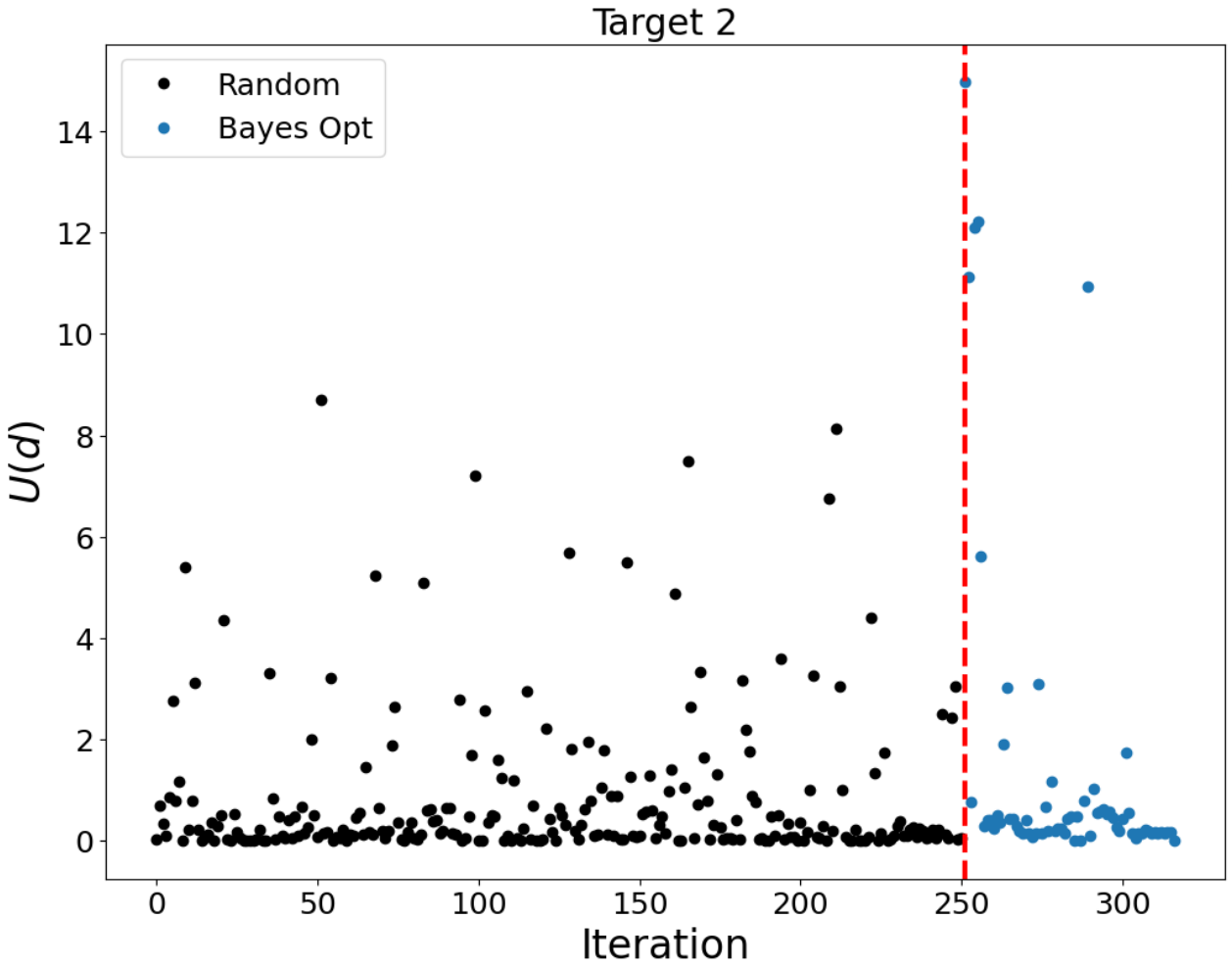
$t = 0$

$t = t_e$

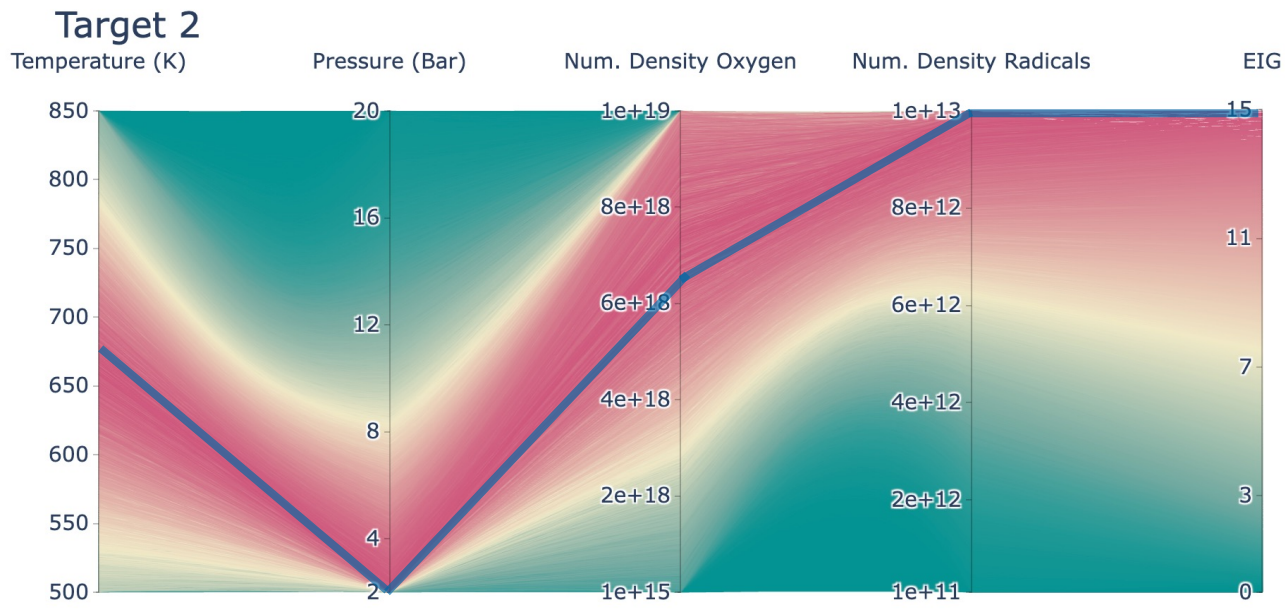
Assumptions:

- Model error $\delta(x)$ independent of d
- **Local GPs constructed independently for each kinetic time**
- Considering all 1151 model parameters as uncertain
  - Designs are constructed to target specific sets of the model parameters
- Prior parameter uncertainty is specified by literature

Results

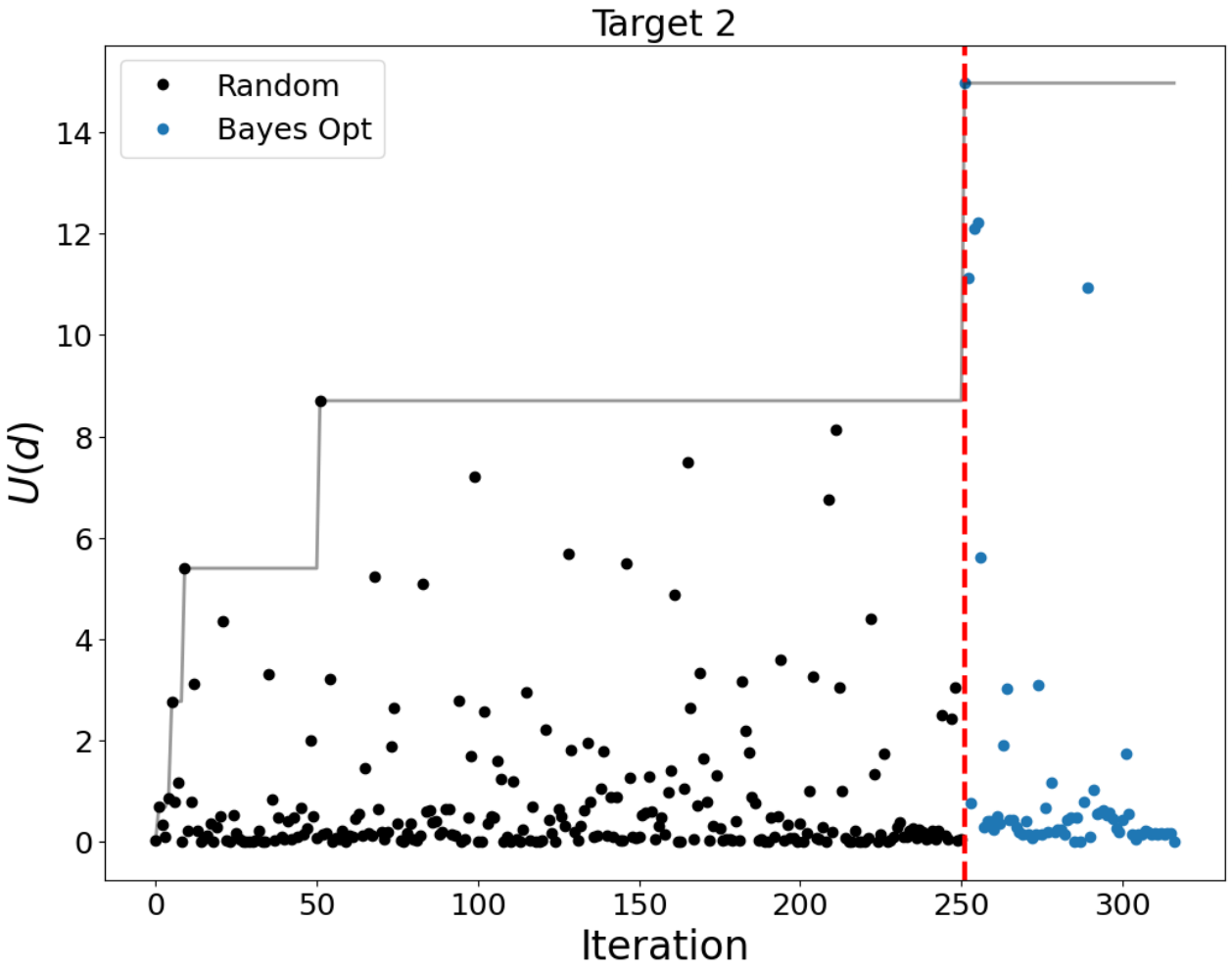


Optimization iterations for maximizing the expected utility. A number of random samples (black points) are evaluated in parallel prior to Bayesian Optimization (blue points) , which quickly improves on the previous maxima.

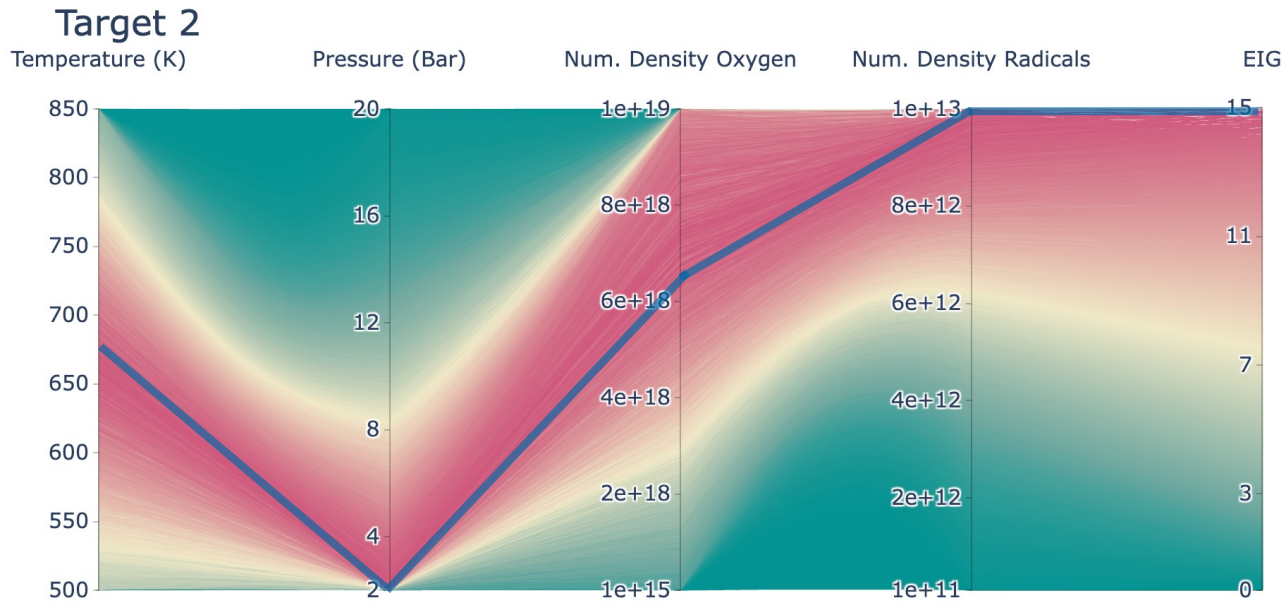


Parallel coordinate plot shows the range of design variables associated with high $U(d)$ in red. Trajectory of optimal design is highlighted in blue.

Results

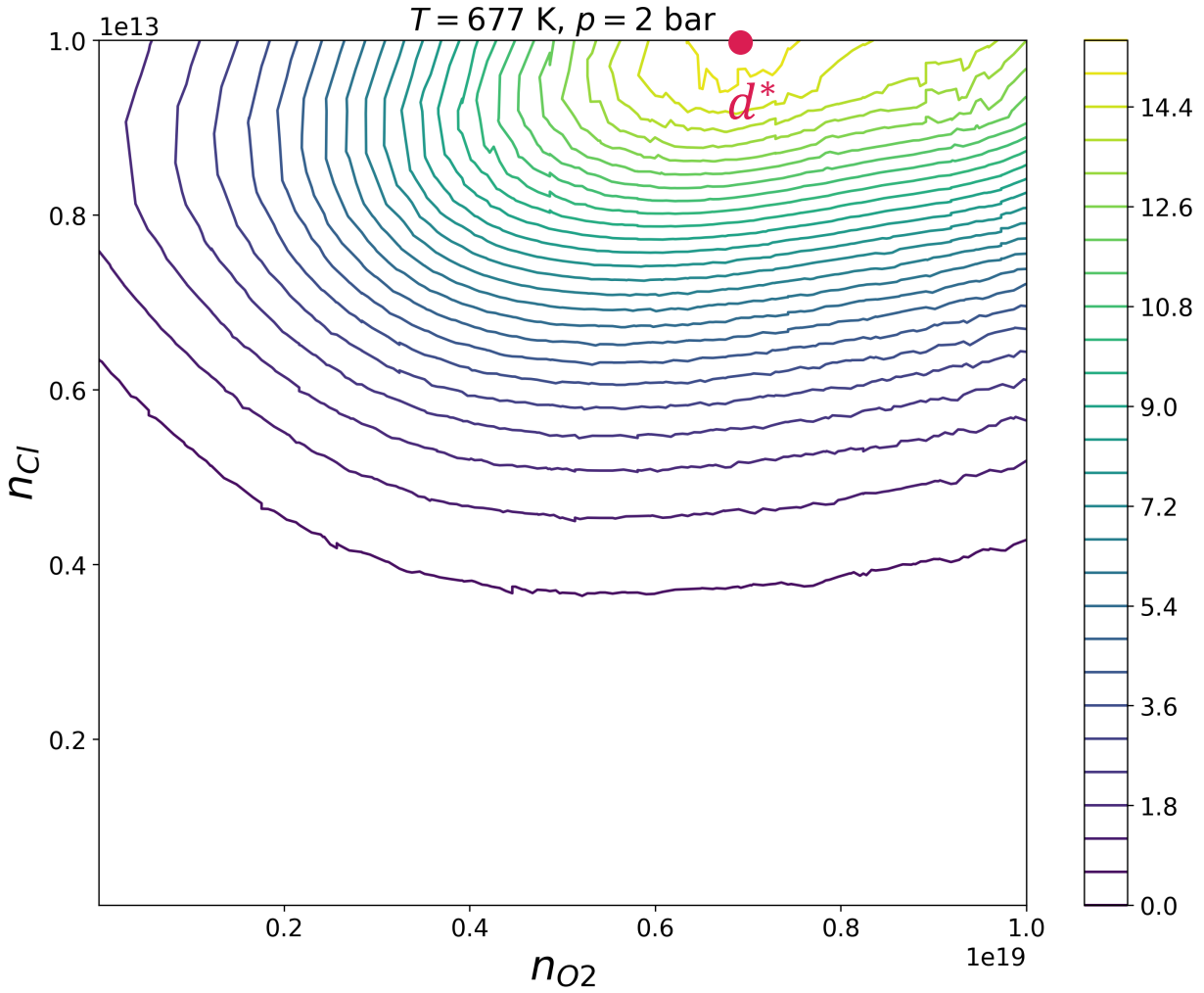


Optimization iterations for maximizing the expected utility. A number of random samples (black points) are evaluated in parallel prior to Bayesian Optimization (blue points) , which quickly improves on the previous maxima.



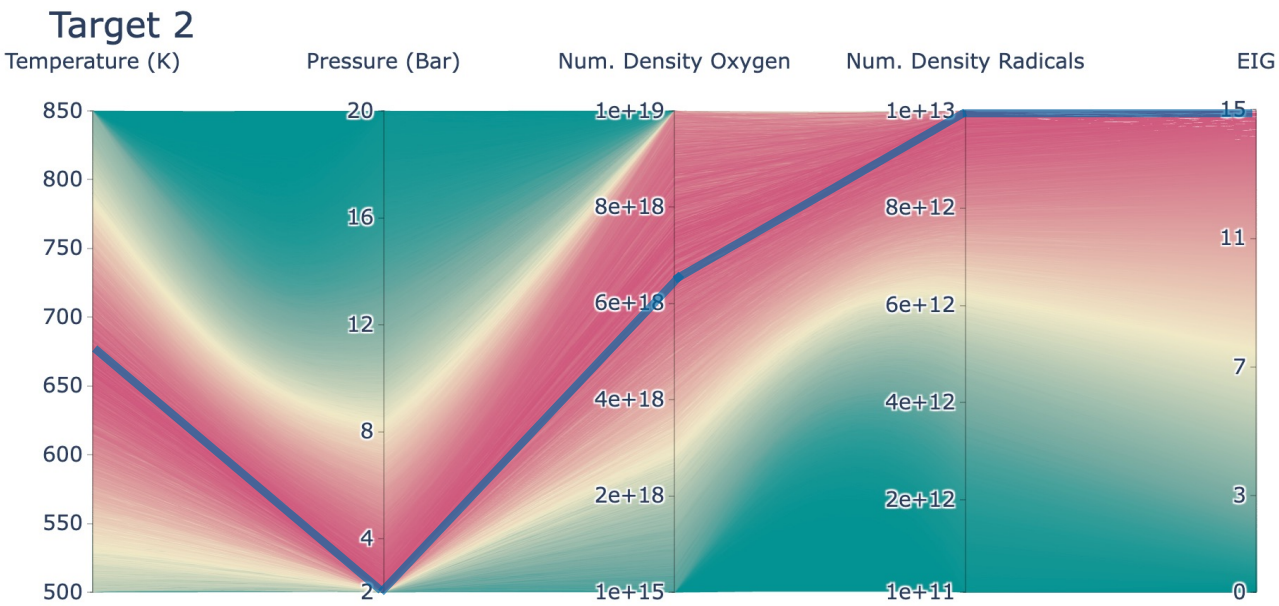
Parallel coordinate plot shows the range of design variables associated with high $U(d)$ in red. Trajectory of optimal design is highlighted in blue.

Results



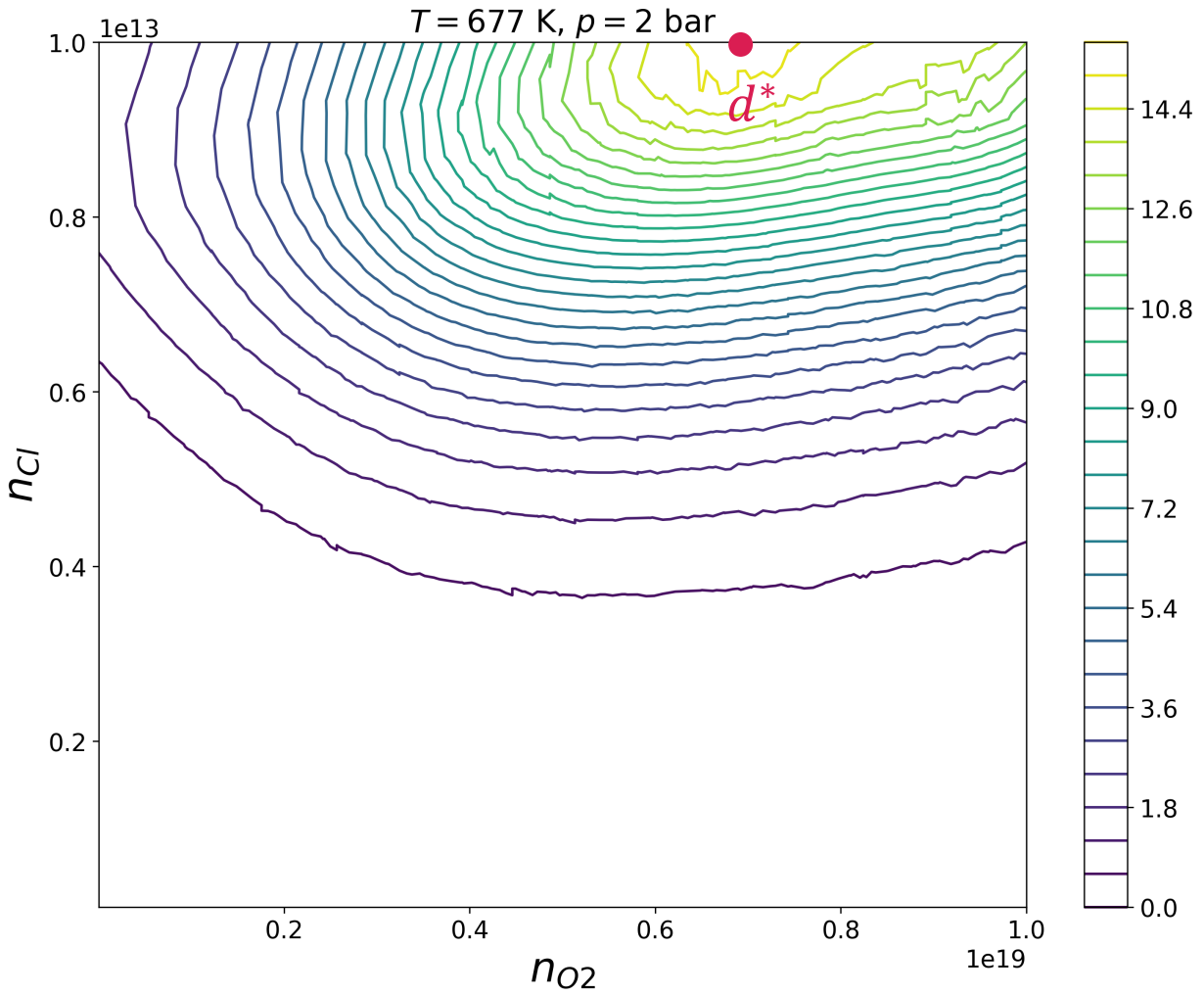
Cross-section of $U(d)$ for a fixed temperature and pressure.

Heatmap enables selection of a suboptimal design which can be desirable if resources are scarce or there are safety concerns



Parallel coordinate plot shows the range of design variables associated with high $U(d)$ in red. Trajectory of optimal design is highlighted in blue.

Results



Cross-section of $U(d)$ for a fixed temperature and pressure.

Heatmap enables selection of a suboptimal design which can be desirable if resources are scarce or there are safety concerns

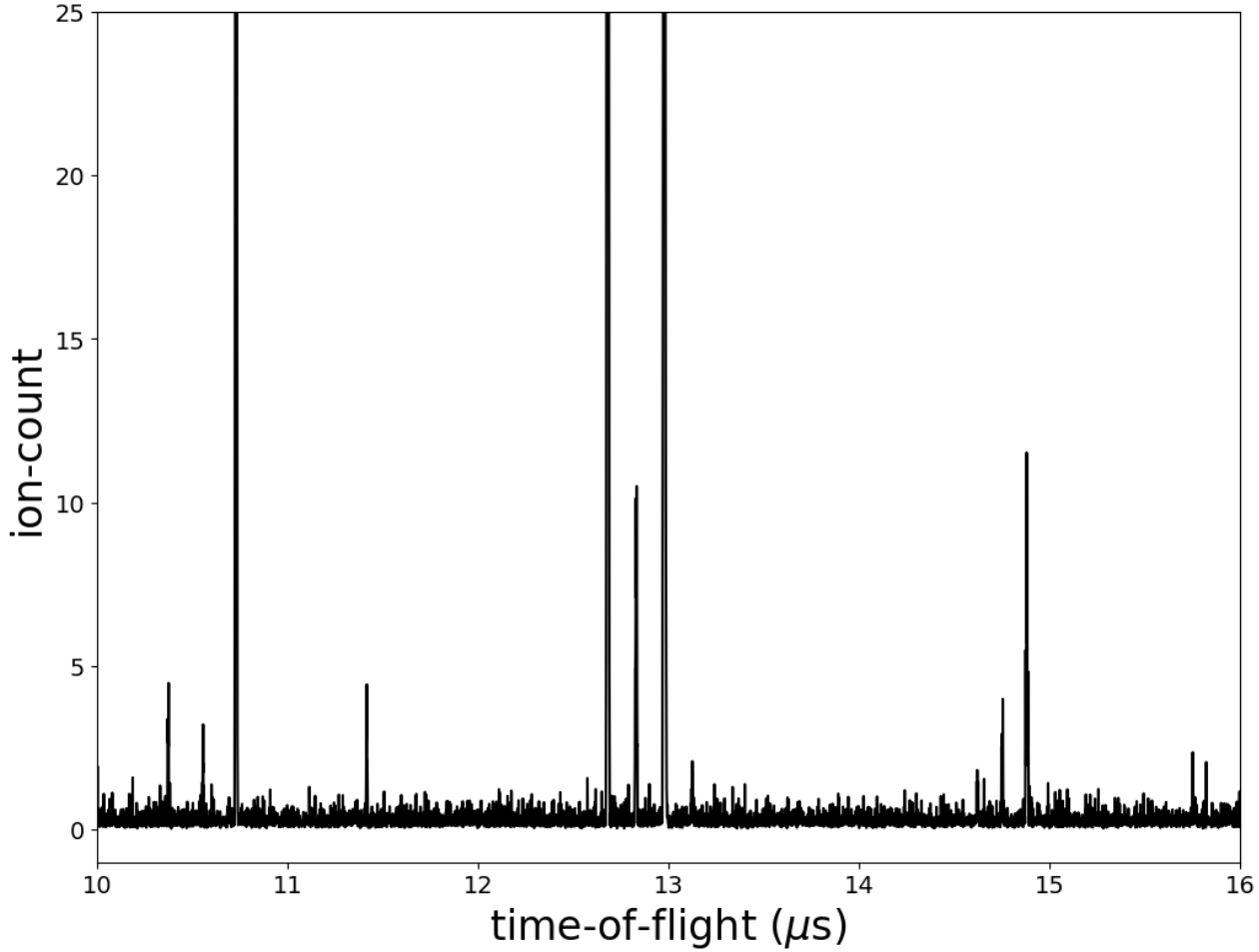
What does the posterior distribution look like after measuring data at d^*?

- Generate datasets using the probabilistic model at d^*
- Plausible posterior distributions are aggregated using logarithmic pooling:

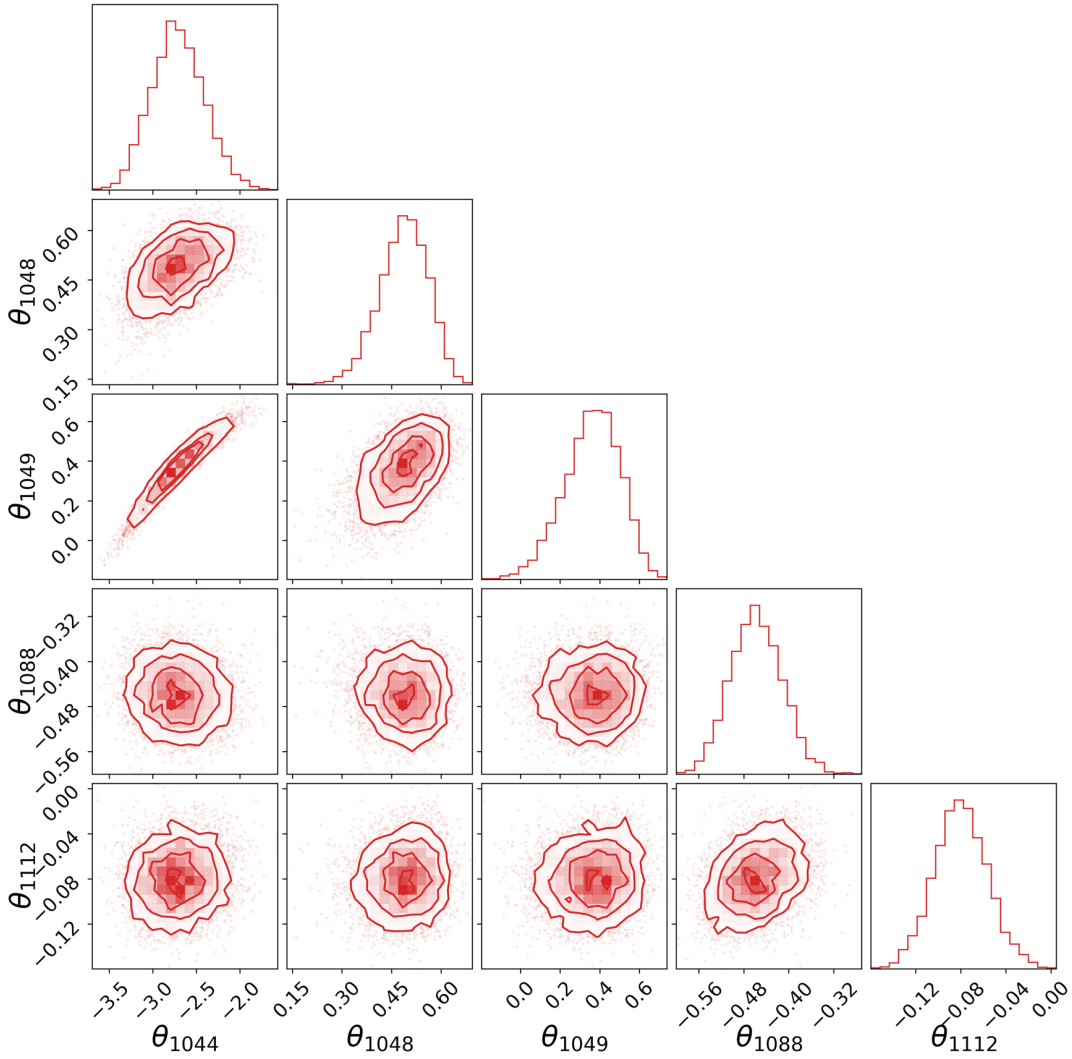
$$p(\theta)_{\log} \approx \left[\prod_{i=1}^{K} p(\theta|y_i)\right]^{1/K}$$

$$p(\theta|y_i) \propto p(\theta)p(y_i|\theta)$$

Results

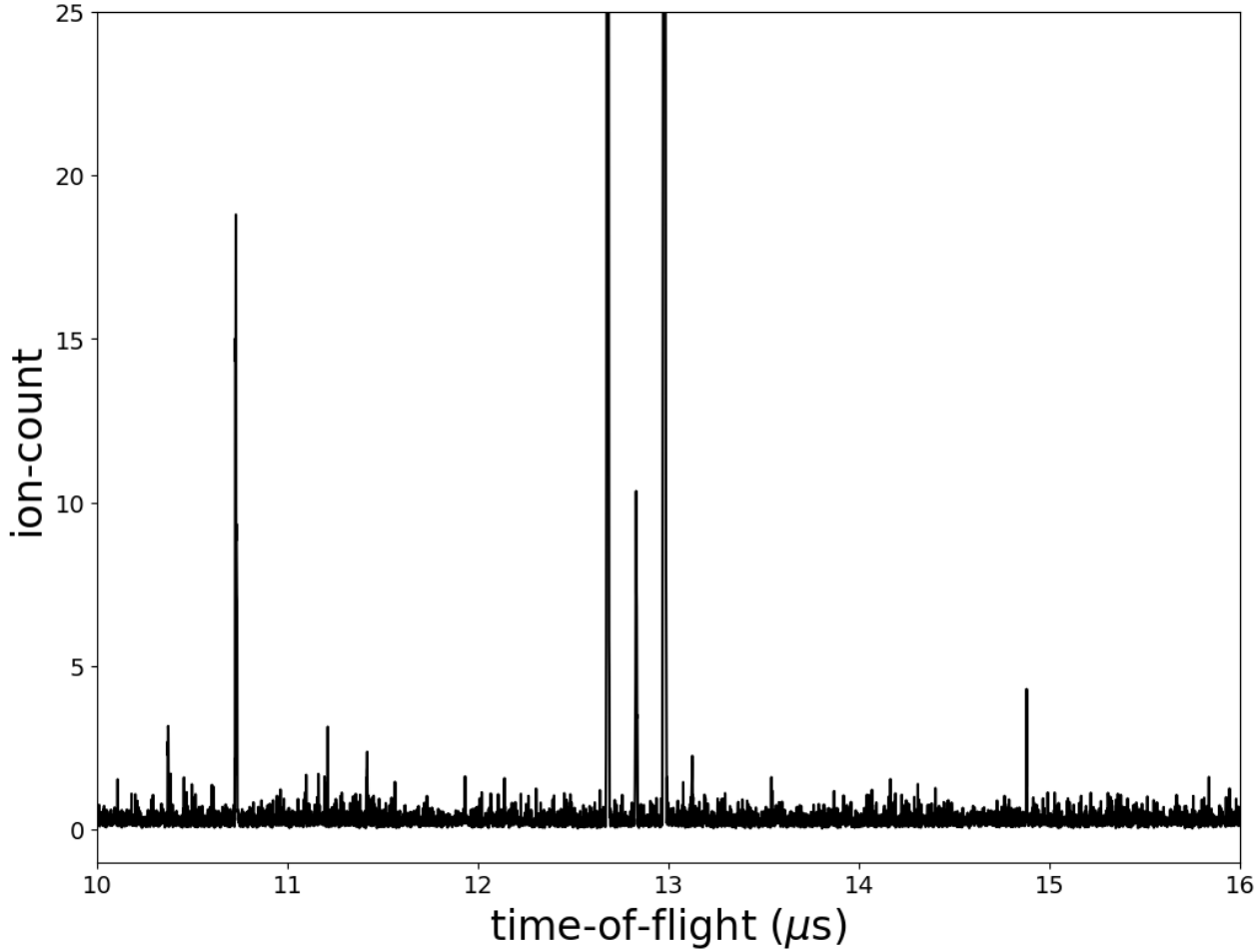


Visualization of the range of plausible datasets generated by the probabilistic model at d^* both time and VUV energy are fixed at 60ms and 11.3 eV, respectively. New species appear with significant changes to the peak magnitudes.

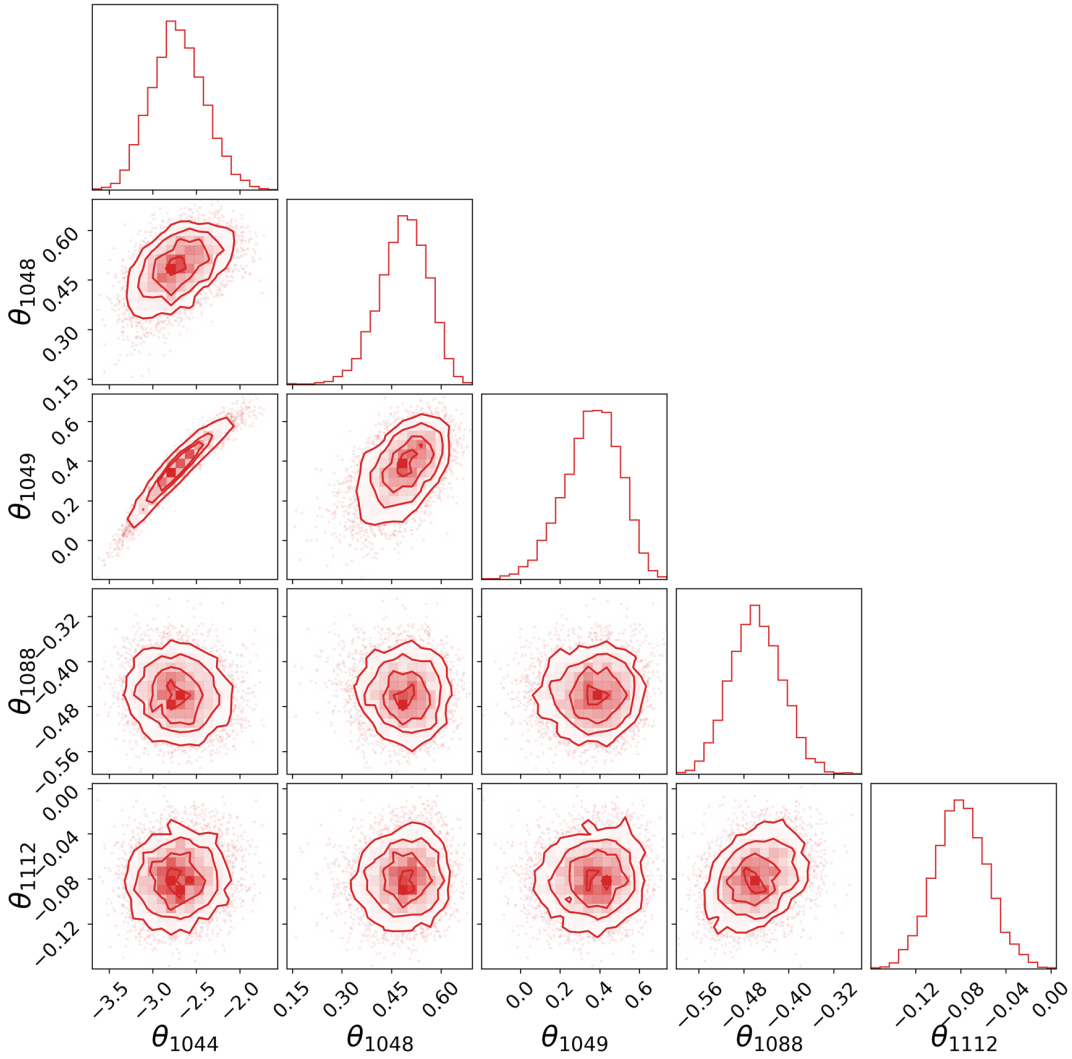


Estimated posterior after log pooling with $K = 10^3$ plausible datasets generated at d^*. Prior distributions are a standard normal.

Results

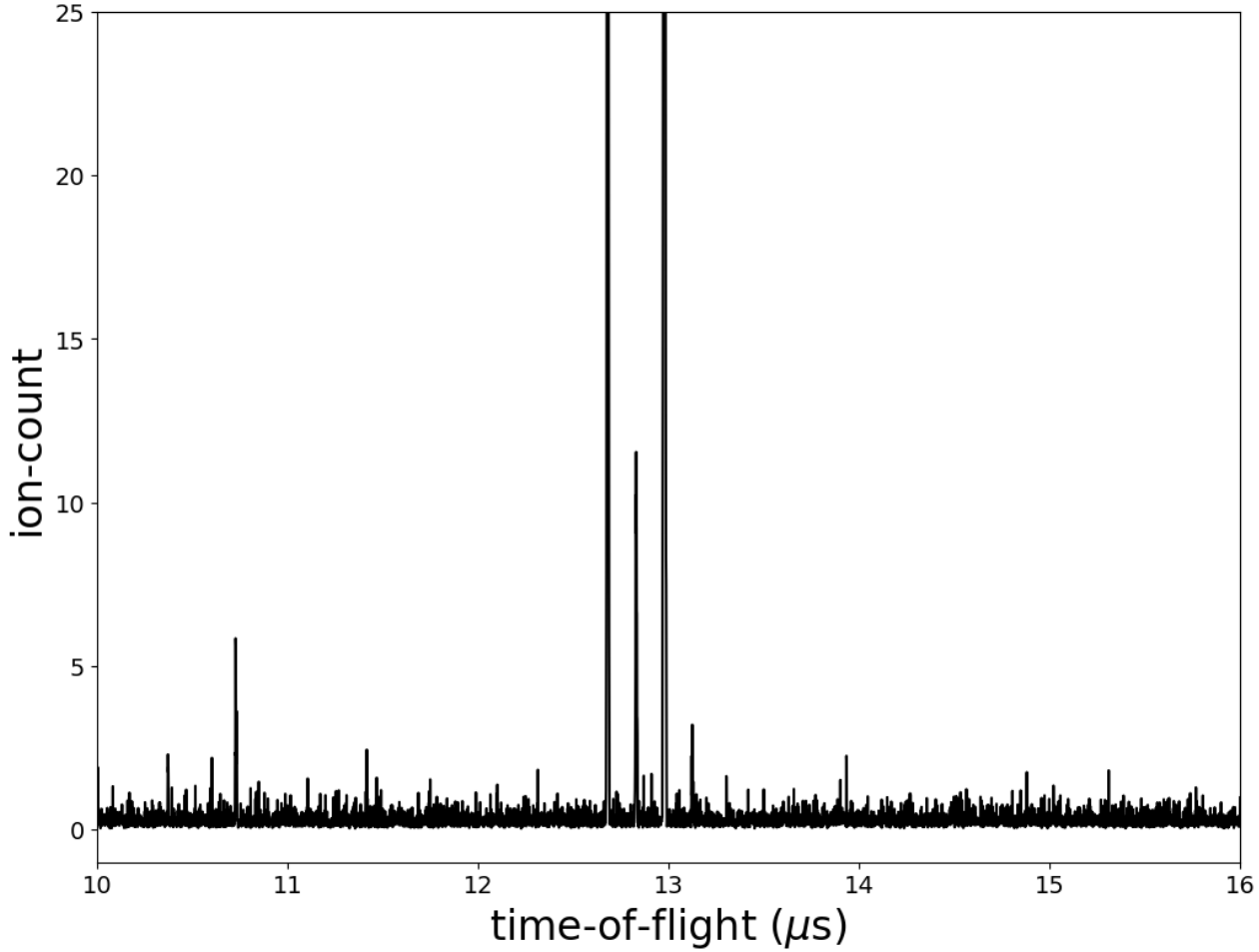


Visualization of the range of plausible datasets generated by the probabilistic model at d^* both time and VUV energy are fixed at 60ms and 11.3 eV, respectively. New species appear with significant changes to the peak magnitudes.

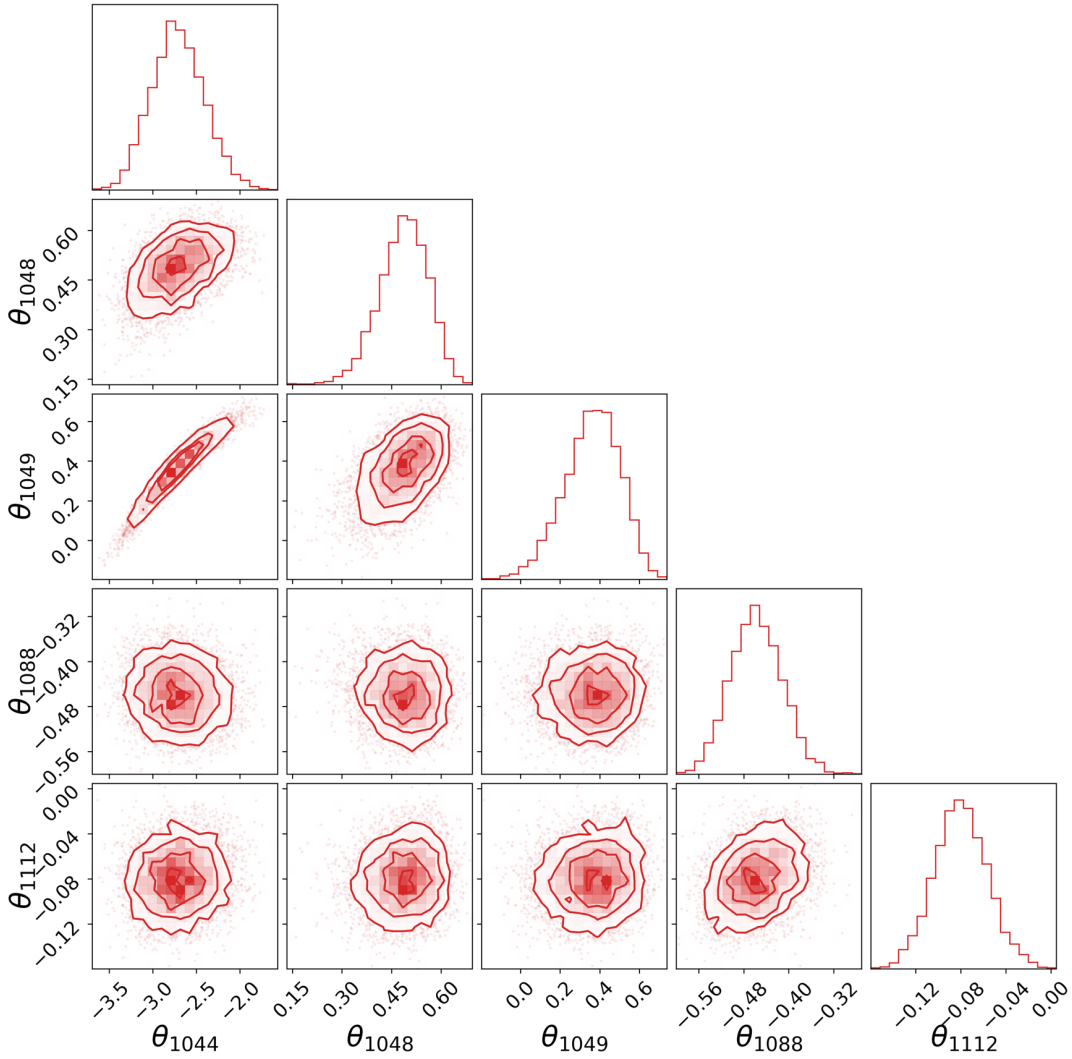


Estimated posterior after log pooling with $K = 10^3$ plausible datasets generated at d^*. Prior distributions are a standard normal.

Results

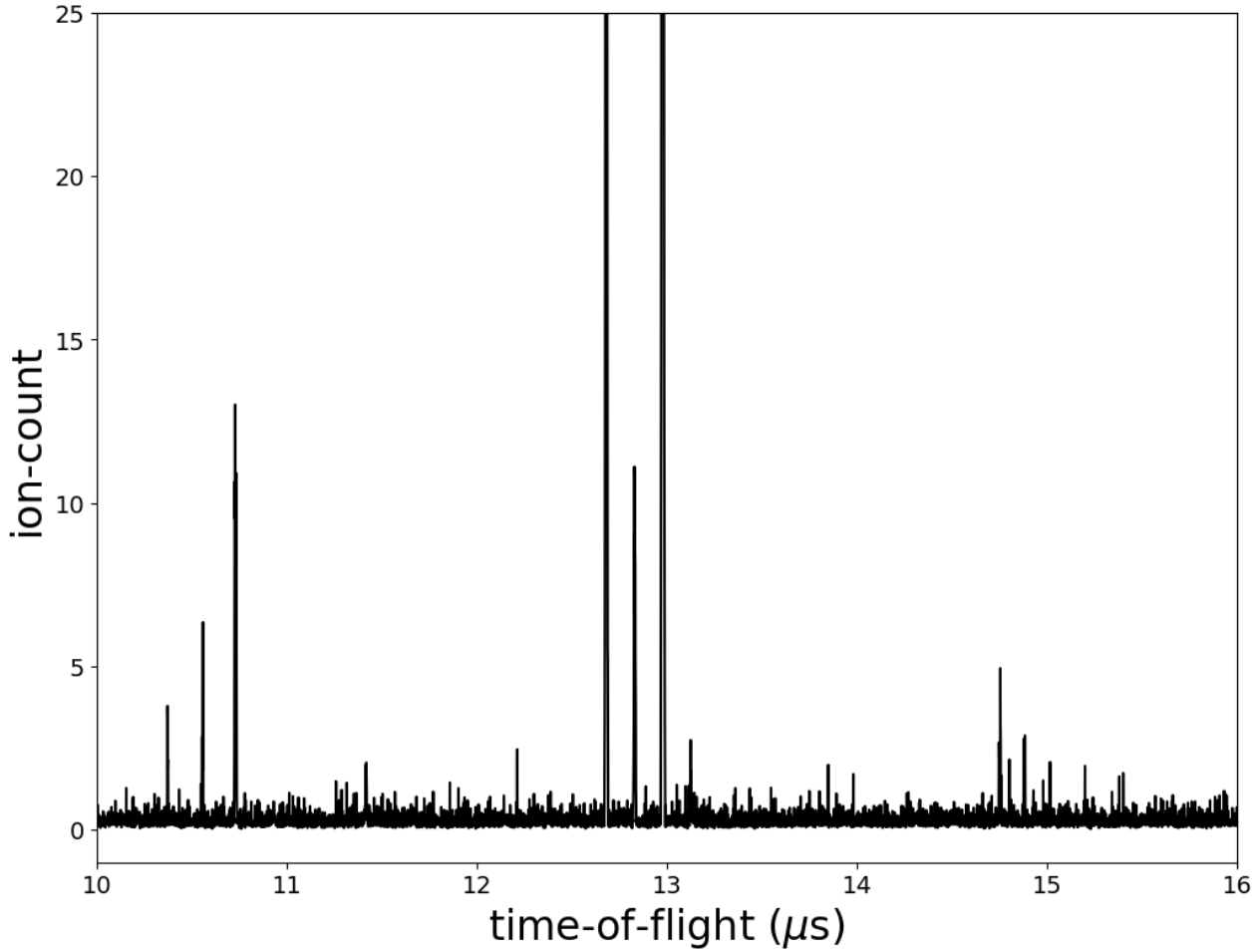


Visualization of the range of plausible datasets generated by the probabilistic model at d^* both time and VUV energy are fixed at 60ms and 11.3 eV, respectively. New species appear with significant changes to the peak magnitudes.

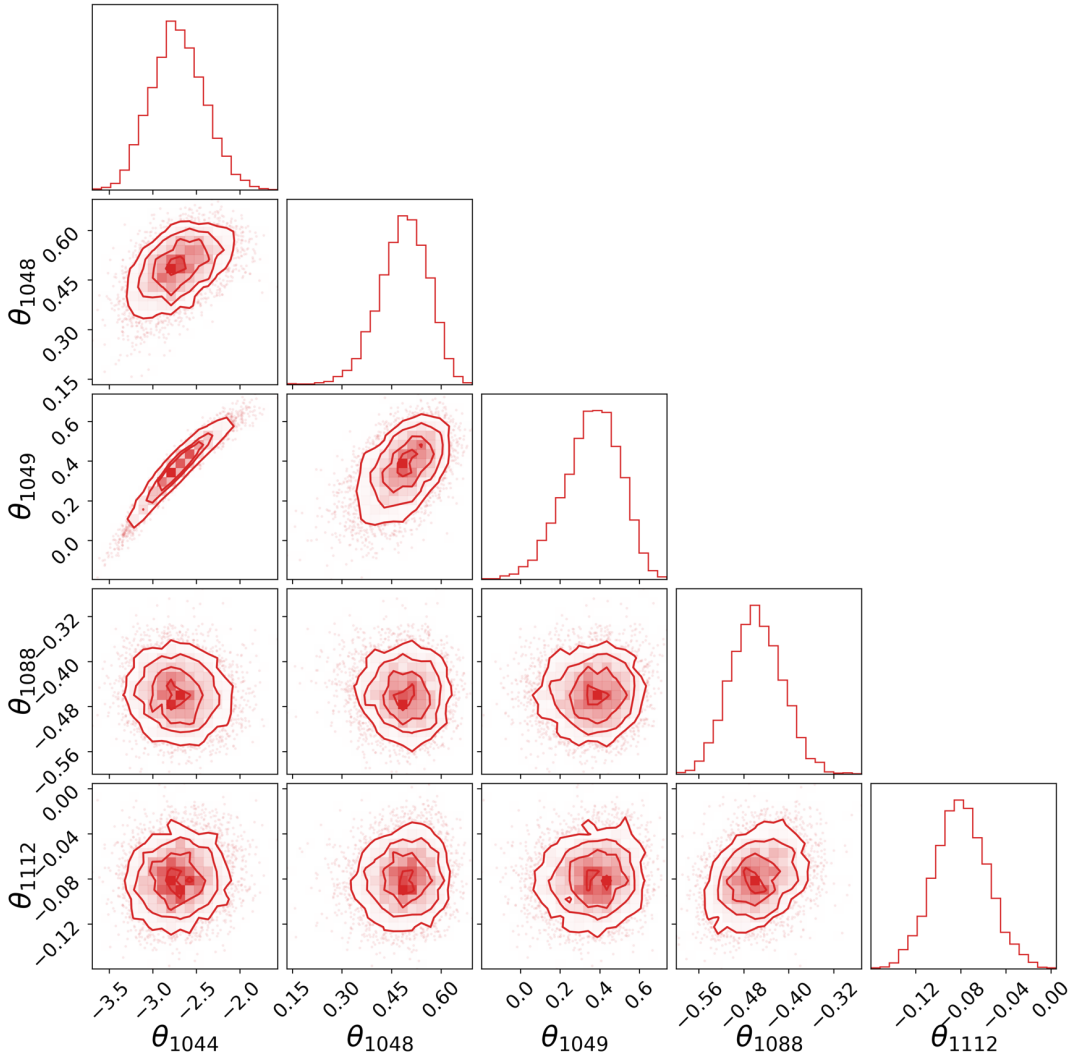


Estimated posterior after log pooling with $K = 10^3$ plausible datasets generated at d^*. Prior distributions are a standard normal.

Results



Visualization of the range of plausible datasets generated by the probabilistic model at d^* both time and VUV energy are fixed at 60ms and 11.3 eV, respectively. New species appear with significant changes to the peak magnitudes.

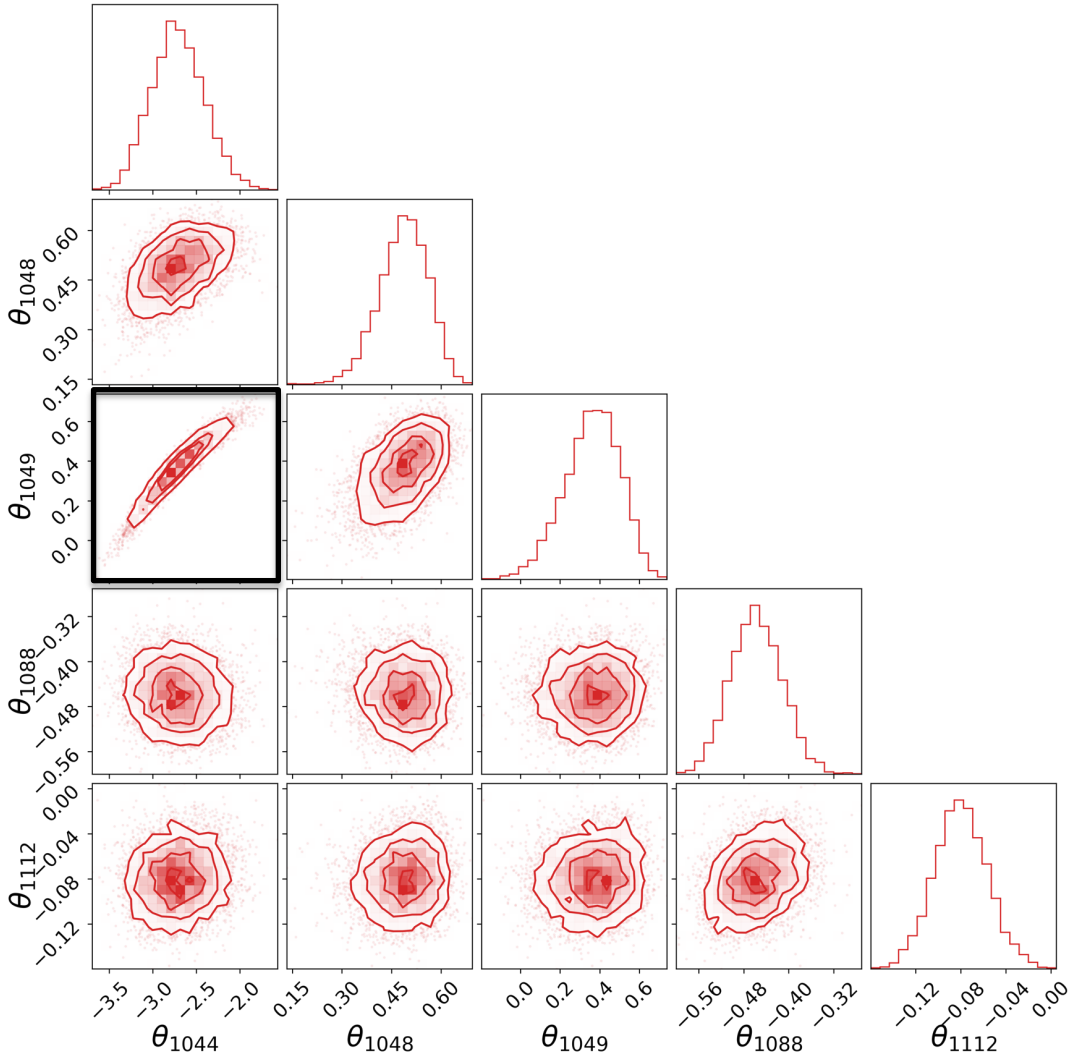


Estimated posterior after log pooling with $K = 10^3$ plausible datasets generated at d^*. Prior distributions are a standard normal.

Results

	n-R + O2 ⟶ propylene oxide
1044	O2 + CH3CH2CH2 ⟶ OH + C-CH2OCH(CH3)
	n-RO2 ⟶ products
1049	CH3CH2CH2OO ⟶ OH + C-CH2OCH(CH3)

Posterior distribution provides correlations between two of the sensitive reaction rates, both sharing propylene oxide as a product.



Estimated posterior after log pooling with $K = 10^3$ plausible datasets generated at d^*. Prior distributions are a standard normal.

Conclusion & Future Work

- Demonstrated feasibility and workflow for applying OED for high-dimensional physics-based models
- Bayesian optimization enables efficient exploration of noisy utility functions
- Low-dimensional surrogates of model output provide necessary computational savings for assessing complex models

Future Work

- Demonstrate improvement in the posterior density using measurements at optimal design compared to random designs
- Update initial prior estimates utilizing available measurement data

Acknowledgements

We would like to thank Oscar Diaz-Ibarra, Kyungjoo Kim, Arun Hegde, Cosmin Safta, and Khachik Sargsyan for helpful conversations about this work.

This work was supported by the US Department of Energy (DOE), Office of Basic Energy Sciences (BES) Division of Chemical Sciences, Geosciences, and Biosciences.

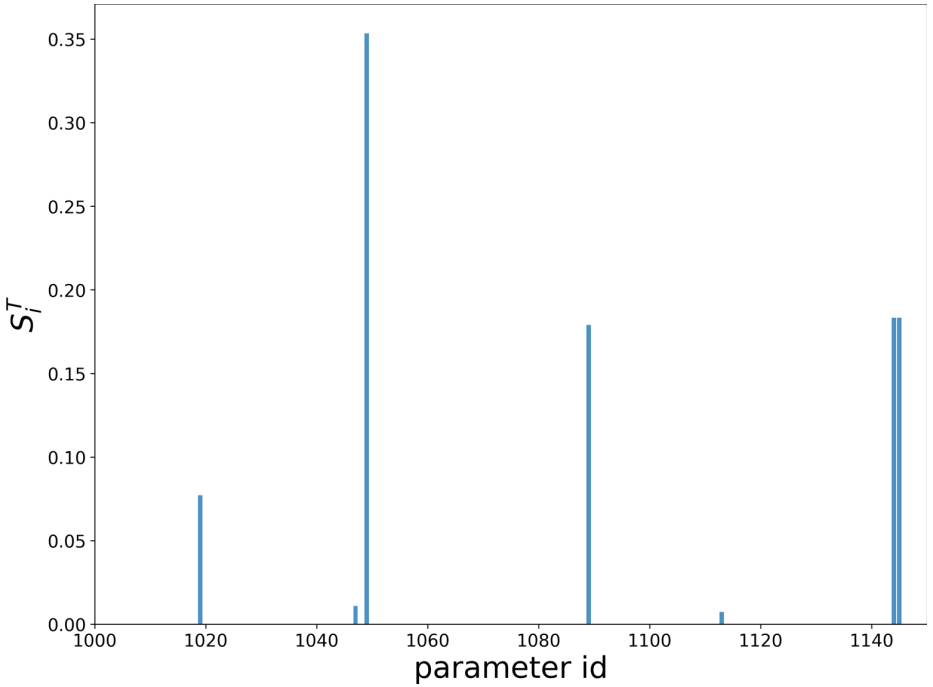
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525.

Additional Slides

Challenges

Large number of uncertain model parameters

- Physics and instrument model contain 1151 uncertain model parameters
- Only a *small fraction* have a significant impact on the spectrum at each design



Only a small number of nonzero total Sobol indices for design conditions 620 K, 2 bar, $n_{\mathrm{C3H8}} = 2\mathrm{e}14$, $n_{\mathrm{O2}} = 8\mathrm{e}18$, $n_{\mathrm{Cl}} = 1\mathrm{e}13$.

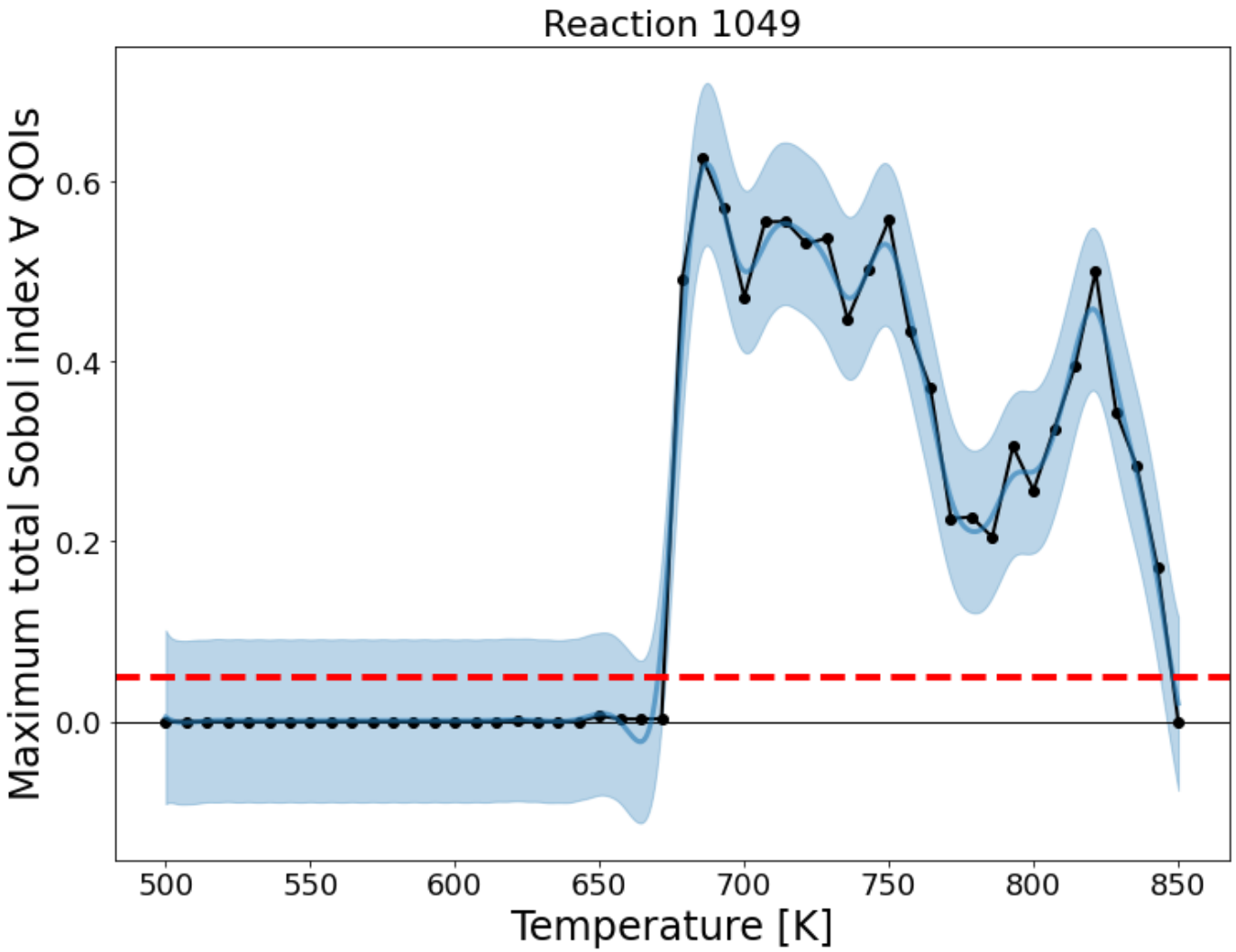
Variance-based Global Sensitivity Analysis

$$S_i^T = 1 - \frac{\mathrm{Var}_{\theta_{\sim i}}\left[\mathbb{E}_{\theta_i}\left(f\left(\boldsymbol{\theta}\right)|\theta_{\sim i}\right)\right]}{\mathrm{Var}_{\theta}\left[f(\boldsymbol{\theta})\right]}$$

- $\ell 1$ regularization of the polynomial chaos expansion coefficients

$$\boldsymbol{c}^{CS} = \arg\min_{\boldsymbol{c}} ||y - \boldsymbol{c}\boldsymbol{\Psi}||_2 + \gamma||\boldsymbol{c}||_1$$

Bruno Sudret. Global sensitivity analysis using polynomial chaos expansions. *Reliability engineering & system safety* 93, no. 7 (2008): 964-979


Reaction 1049
Maximum total Sobol index ∀ QOIs
0.0
0.2
0.4
0.6
500
550
600
650
700
750
800
850
Temperature [K]

Reducing output dimensionality

Goal: Map high-dimensional model output to a lower-dimensional space while minimizing loss of information

$$g : \mathbb{R}^J \mapsto \mathbb{R}^K$$

Autoencoder

Encoder network: $q = g_\psi(y)$ $\quad q \in \mathbb{R}^K$

Decoder network: $\widehat{y} = h_\phi(g_\psi(y))$ $\quad y \in \mathbb{R}^J$

$K \ll J$

Minimize reconstruction loss

$$L(\phi, \psi) = \min_{\phi,\psi} ||y - h_\phi(g_\psi(y))||_2^2$$

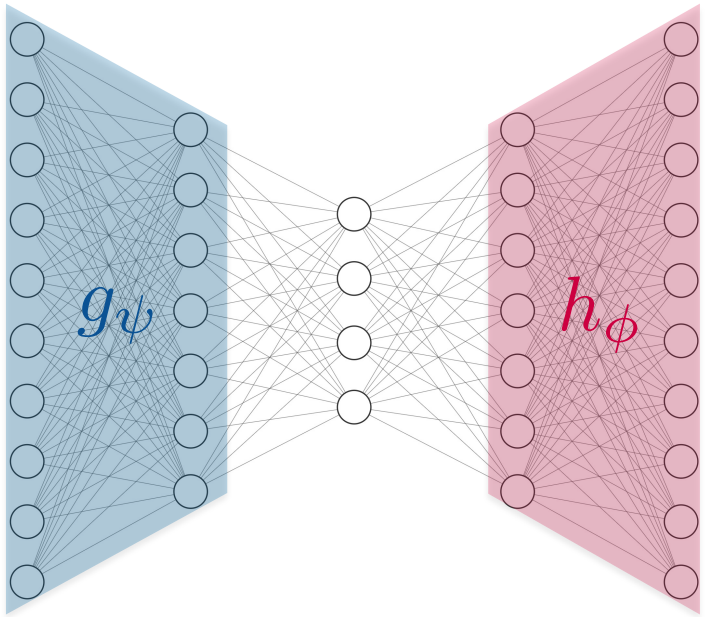
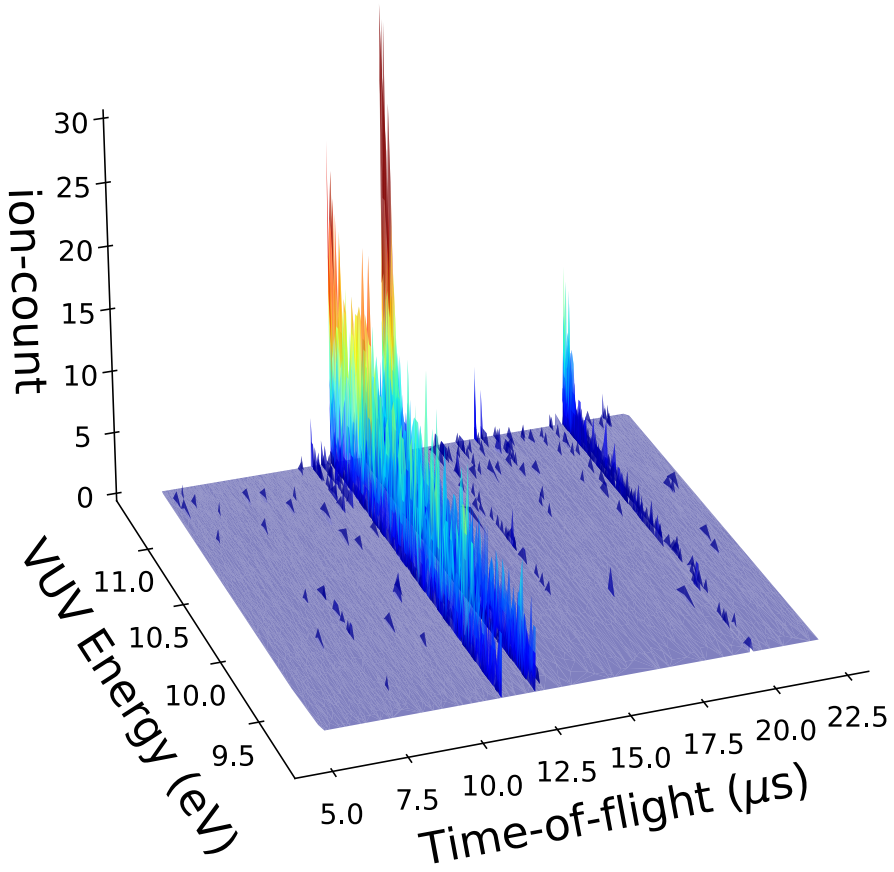
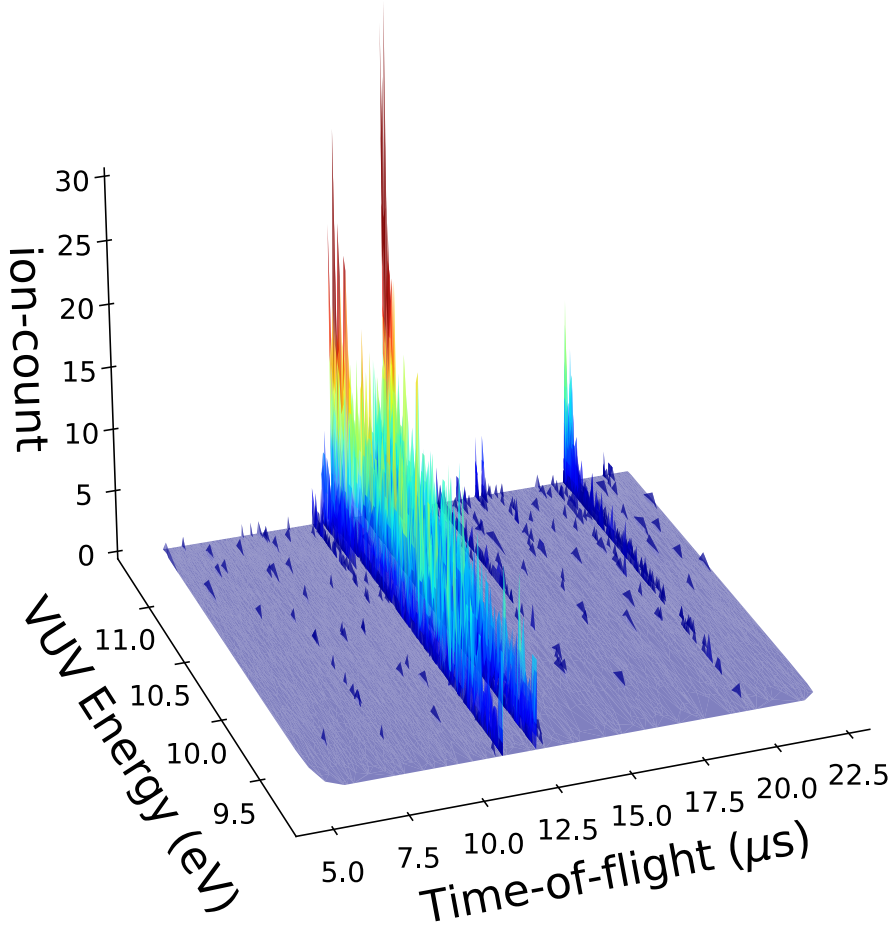


Illustration of an autoencoder architecture

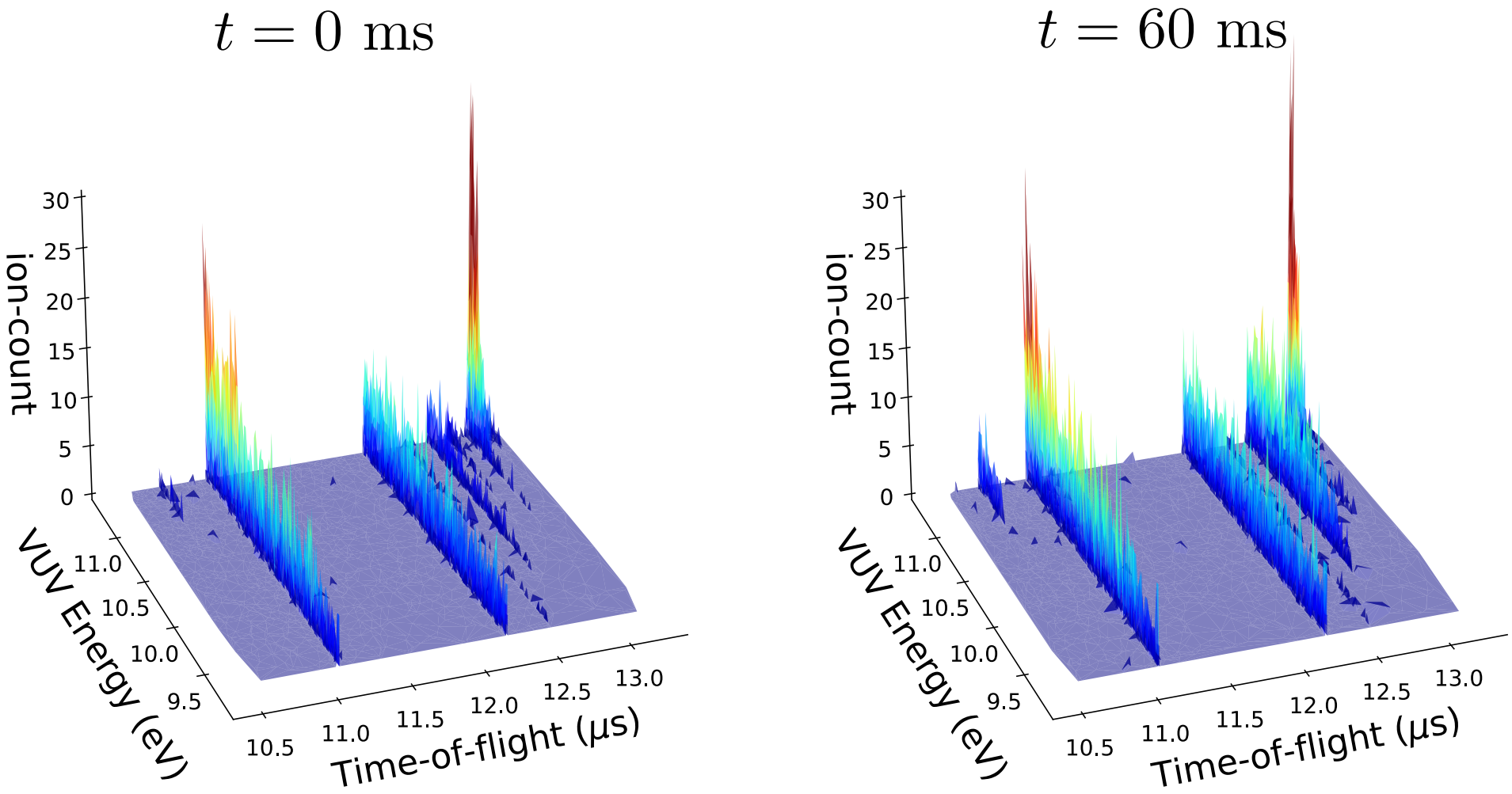
$t = 0$ ms



$t = 60$ ms

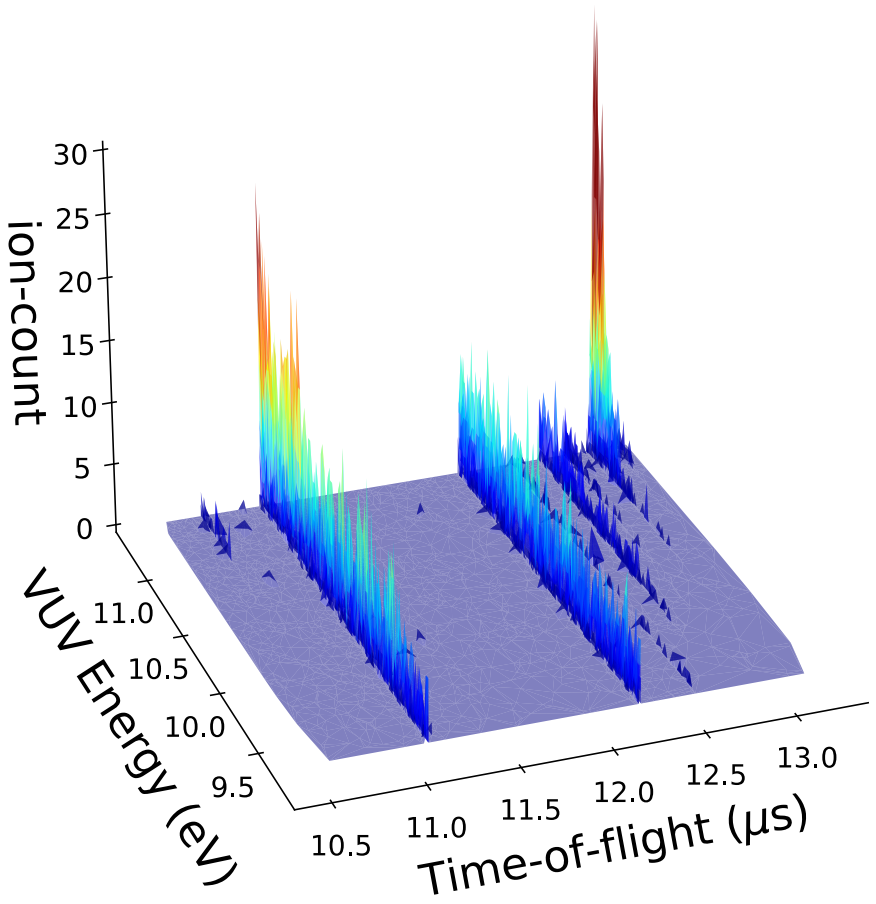


Experiment conducted at $T = 700$ K, p $= 10$ bar, $\chi_{C3H8} = 9.7 \times 10^{-7}$, $\chi_{O_2} = 2.9 \times 10^{-2}$, $\chi_{pre} = 2.2 \times 10^{-4}$ with a helium bath gas

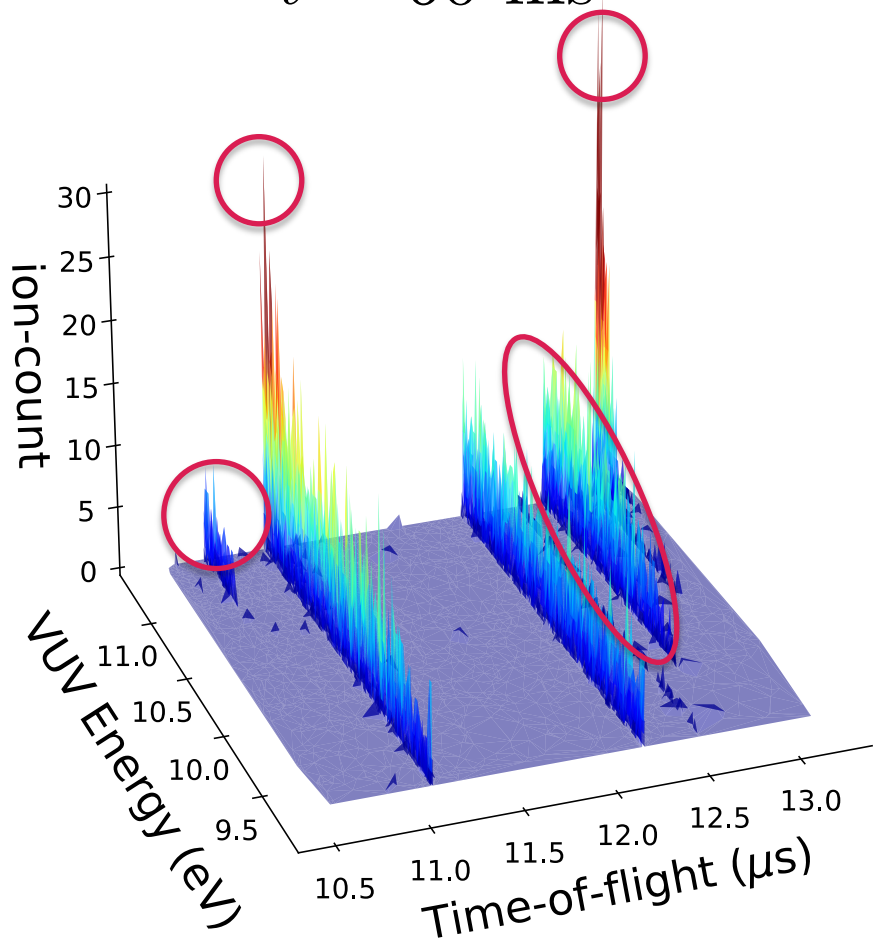


Experiment conducted at $T = 700$ K, p $= 10$ bar, $\chi_{C3H8} = 9.7 \times 10^{-7}$, $\chi_{O_2} = 2.9 \times 10^{-2}$, $\chi_{pre} = 2.2 \times 10^{-4}$ with a helium bath gas

$t = 0$ ms

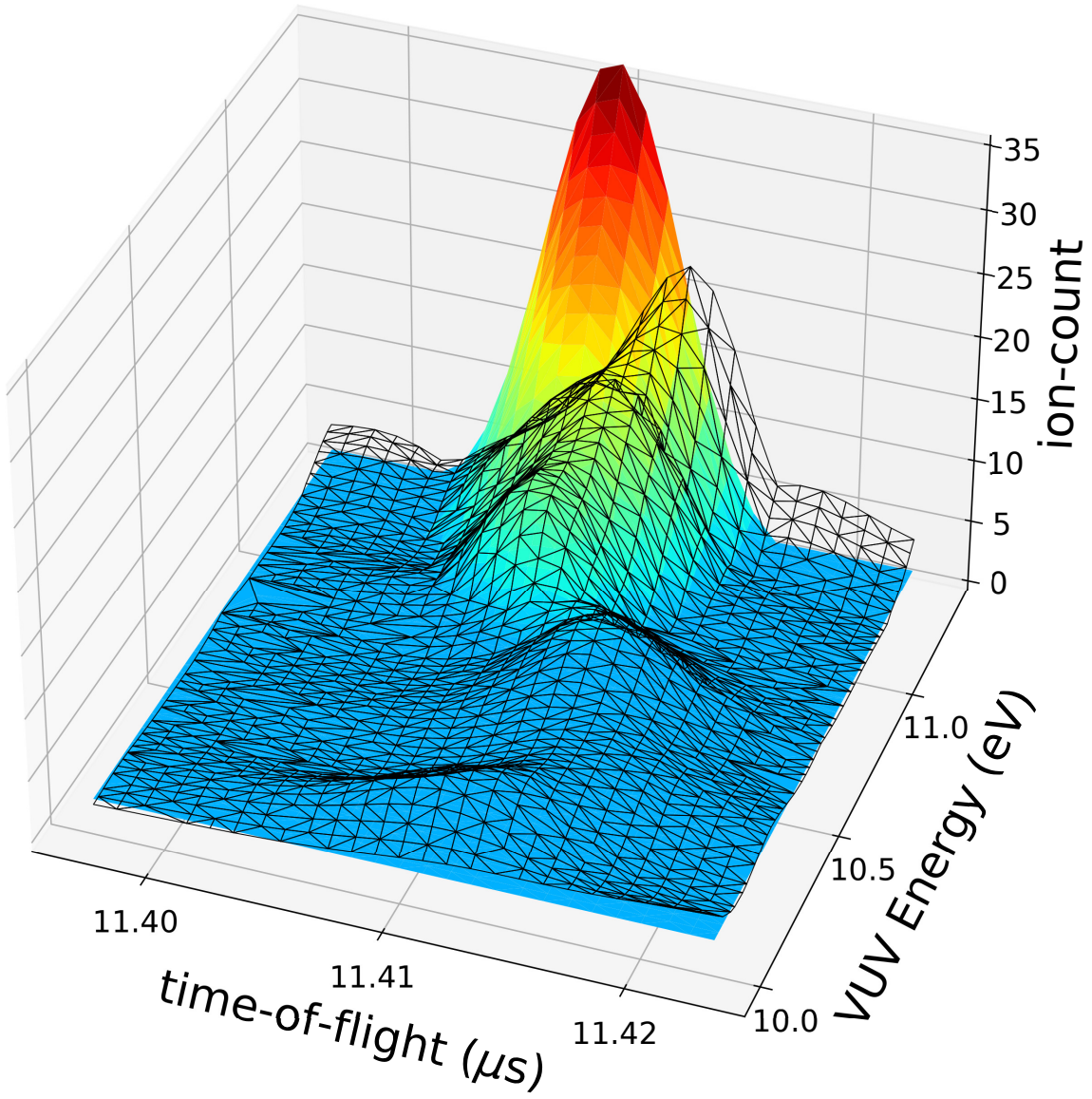


$t = 60$ ms



Experiment conducted at $T = 700$ K, p $= 10$ bar, $\chi_{C3H8} = 9.7 \times 10^{-7}$, $\chi_{O_2} = 2.9 \times 10^{-2}$, $\chi_{pre} = 2.2 \times 10^{-4}$ with a helium bath gas

Modeling the HPR experiment



Solid surface is the prediction of $f(\theta_{MAP}, d, x)$ for one of the peaks in the time-of-flight spectrum (H_2O_2).
Mesh surface shows the prediction with model error, $f(\theta_{MAP}, d, x) + \mu_\delta(x)$ which increases the fidelity of the predictive model.

Numerical approximation

$$u(y, d, \theta) = D_{KL}(p(\theta|y, d)||p(\theta)) = \int_\Theta p(\theta|y, d) \log \left[\frac{p(\theta|y, d)}{p(\theta)} \right] d\theta = u(y, d)$$

$$\begin{aligned} U(d) &= \int_\mathcal{Y} \int_\Theta u(y, d) p(\theta|y, d) d\theta p(y|d) dy \\ &= \int_\mathcal{Y} u(y, d) p(y|d) dy \\ &= \int_\mathcal{Y} \left(\int_\Theta p(\theta|y, d) \log \left[\frac{p(\theta|y, d)}{p(\theta)} \right] d\theta \right) p(y|d) dy \end{aligned}$$

Using $p(\theta|y, d) = p(y|\theta, d)p(\theta)/p(y|d)$,

$$\begin{aligned} U(d) &= \int_\mathcal{Y} \int_\Theta \log \left[\frac{p(y|\theta, d)}{p(y|d)} \right] p(y|\theta, d) p(\theta) d\theta dy \\ &= \int_\mathcal{Y} \int_\Theta \left[\log p(y|\theta, d) - \log p(y|d) \right] p(y|\theta, d) p(\theta) d\theta dy \end{aligned}$$