Traditional or Adaptive Experimental Design? A Comparison of Statistical Design of Experiments and Bayesian Optimization for a Chemical Synthesis Problem

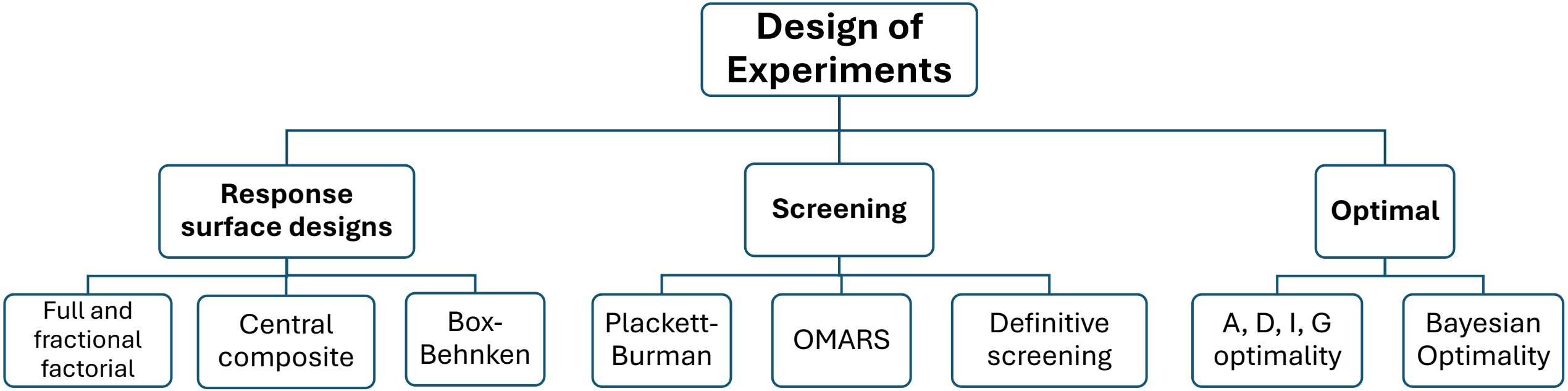
João P. L. Coutinho*, You Peng**, Ricardo Rendall**, Caterina Rizzo**, Kaiwen Ma**, Swee-Teng Chin**, Ivan Castillo**, Marco S. Reis*

* University of Coimbra, CERES, Department of Chemical Engineering, Portugal

** The Dow Chemical Company

Introduction

1. Introduction



- Although designs are static, several stages are reccomended for DOE
- Classic designs tipically assumes the system can be described by a first and second order polynomial model

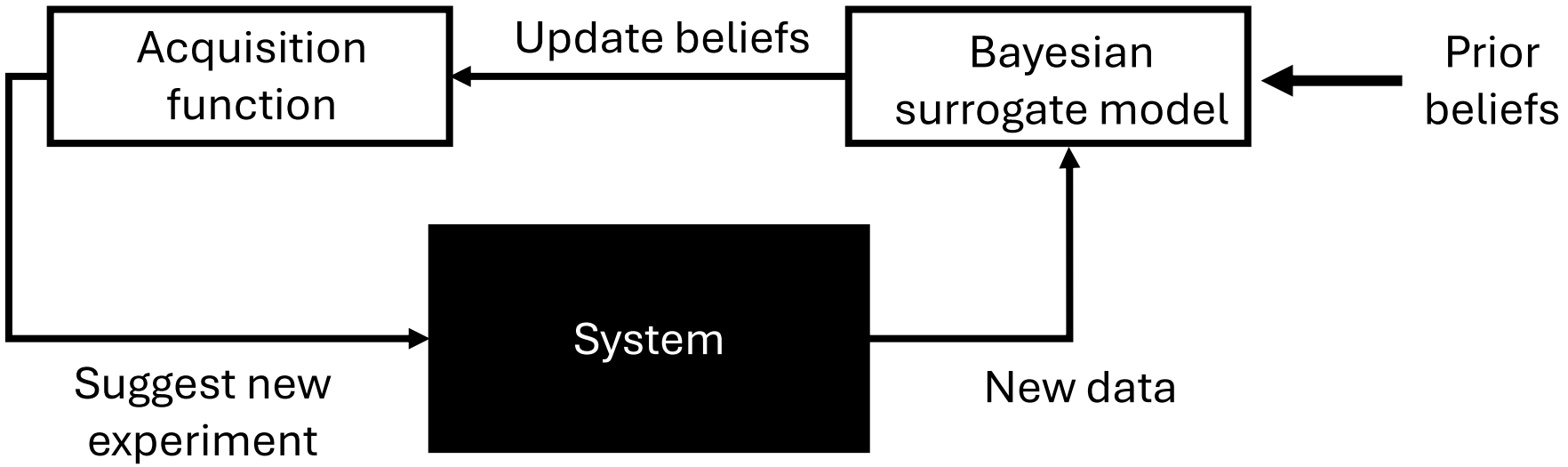
$$y = \beta_0 + \underbrace{\sum_i \beta_i x}_{\text{Main effects}} + \underbrace{\sum_j \sum_{i<j} \beta_{i,j} x_i x_j}_{\text{Interactions}} + \underbrace{\sum_i \beta_{ii} x^2}_{\text{Second order effects}}$$

1. Introduction

- Recently, Bayesian Optimization (BO) has attracted attention as an alternative approach for experimental design

- BO is a sequential Bayesian experimental design framework with two main components:
  - Bayesian surrogate model that reflects **prior beliefs** and provides uncertainty estimates given data

  - Acquisition function balances **exploration** with **exploitation** to search for the optimum

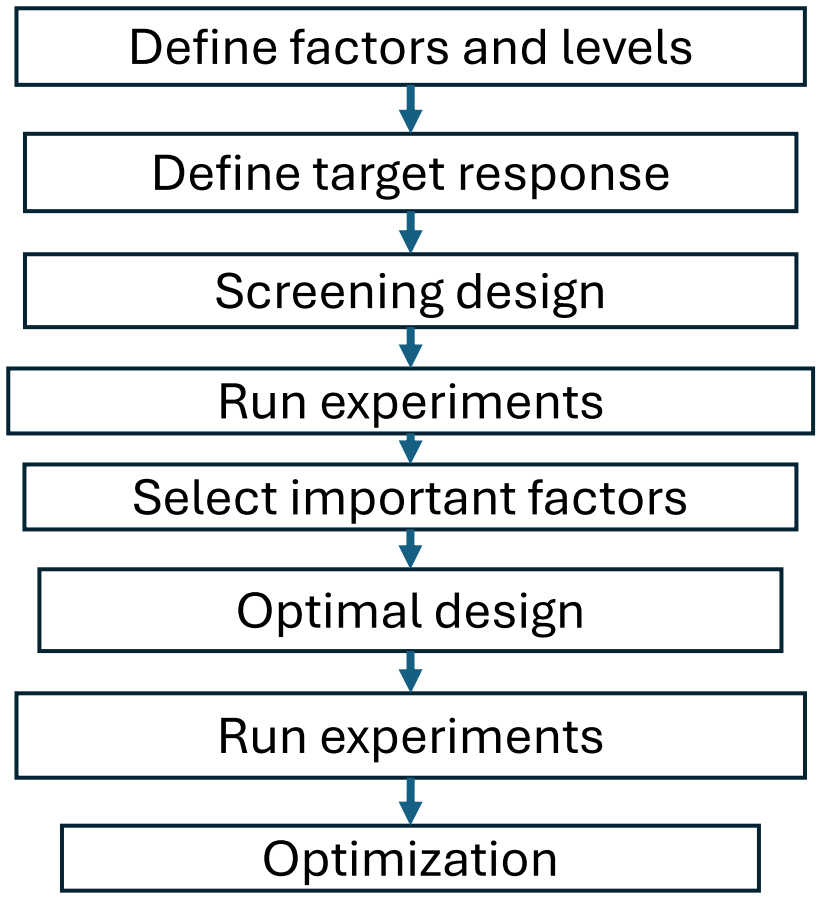


- Despite recent popularity, the principles of BO are not new
  - Thompson sampling was suggested in 1930's as an heuristic to solve the exploration-exploitation dillema
  - BO was formalized in the 1960's and had a recent resurgence in the context of hyperparameter tuning

1. Introduction

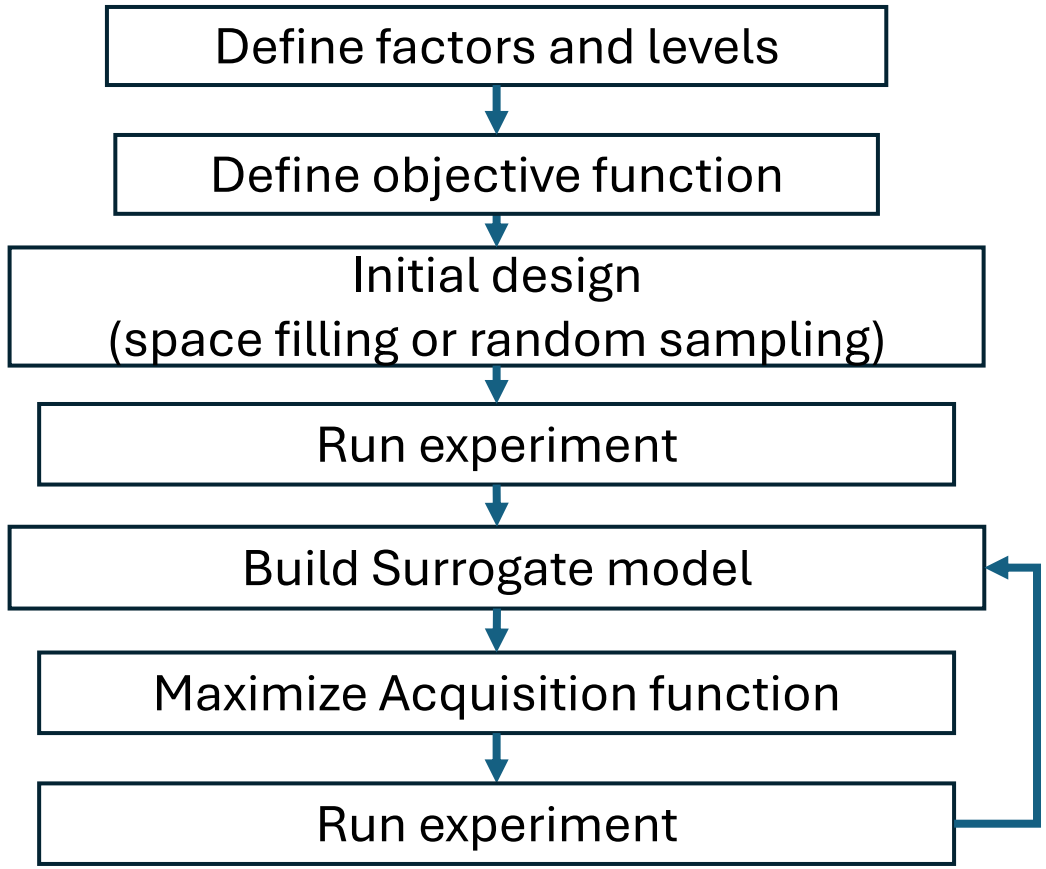
- We aim to compare sequential DOE with BO for a complex real-world chemical synthesis problem

Sequential DOE



Design focus:
reducing overall **model uncertainty**

Bayesian Optimization



Design focus:
reducing **optimum uncertainty**

Case study

2. Case study

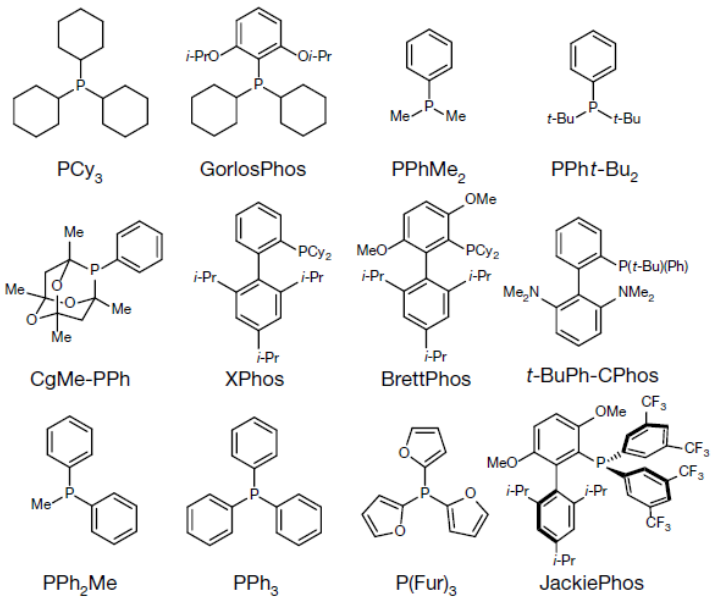
- Openly available chemical reaction dataset used as case study [1]

- Problem has 3 categorical factors, 2 continuous factors and a single response (reaction yield)

4 bases

KOAc	CsOAc
KOPiv	CsOPiv

12 ligands



4 solvents

BuOAc	BuCN
p-Xylene	DMAc

2 continuous variables: Temperature and concentration

90	120	0.057	0.153
105		0.100	
T (°C)		*C* (M)	

- Experimental data from full factorial design (1728 samples) used as proxy for reaction system

- **Objective:** find the combinations of the factors that maximize reaction yield

[1] Shields et. al, 2021, *Bayesian reaction optimization as a tool for chemical synthesis*, Nature

2. Case study

- One-hot-encoding (OHE) is the standard approach to model categorical factors
- In chemical synthesis, we can also use chemical descriptors as covariates for each individual categorical factor
- **291 covariates** are obtained from **Density Functional Theory (DFT)** simulations
  - Due to high dimensionality, descriptors cannot be used within standard DOE
  - BO is used with both types of encodings (OHE and DFT)

One-hot-encoding (OHE)

Base	Level 1	Level 2	Level 3
1	1	0	0
2	0	1	0
3	0	0	1

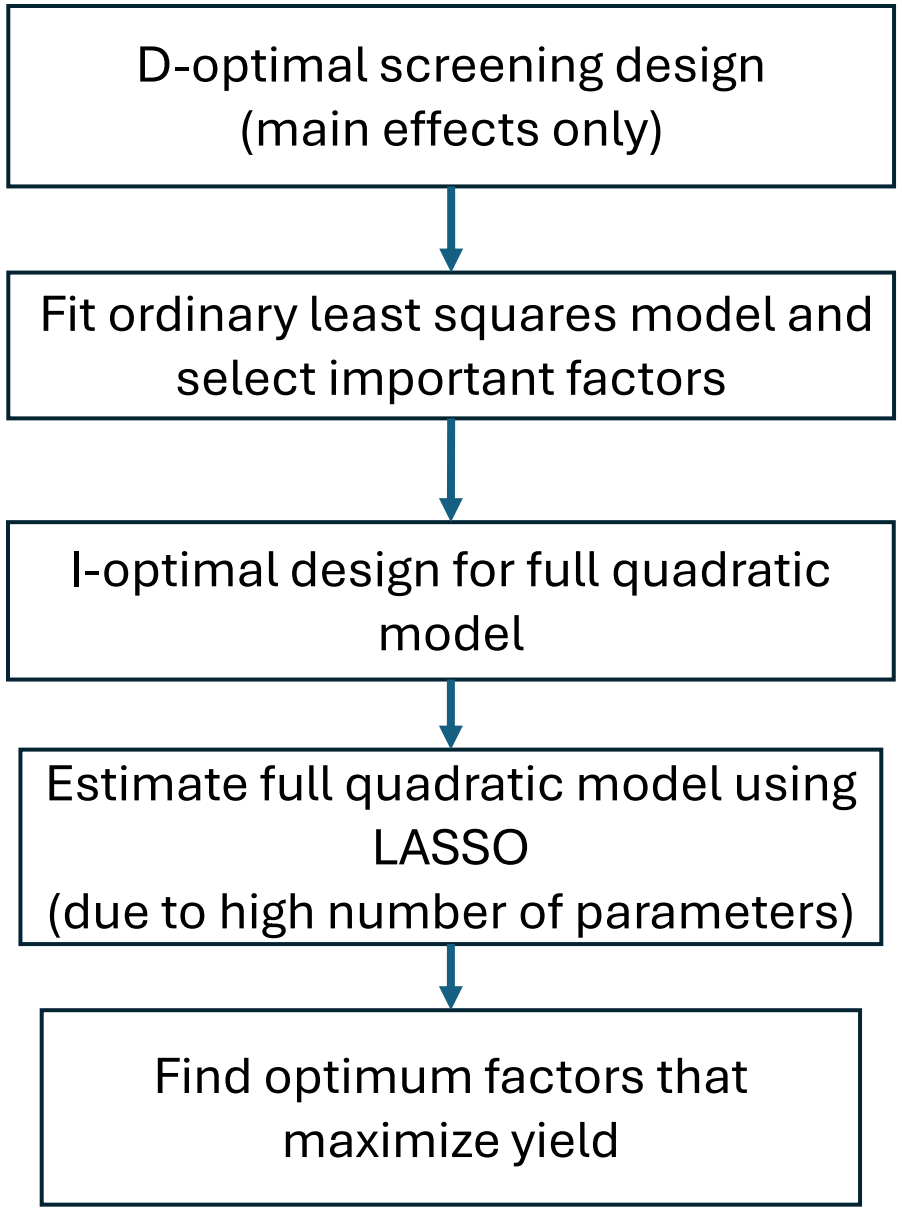
Chemical descriptors

Base	Descriptor 1	Descriptor 2	Descriptor 3	Descriptor 4
1	10	30	100	30
2	7	2	30	2
3	2	0,1	11	0,1

Categorical levels are mapped to 291 molecular features

Methodology

3. Methodology



- There are 192 total possible combinations for categorical factors
- D-optimal design for main effects requires 48 experiments
- Factors are selected according to p-values and magnitude of regression coeficients
- I-optimal design is used to obtain minimum prediction variance over design space
- JMP-PRO 18 used for design and modeling

3. Methodology

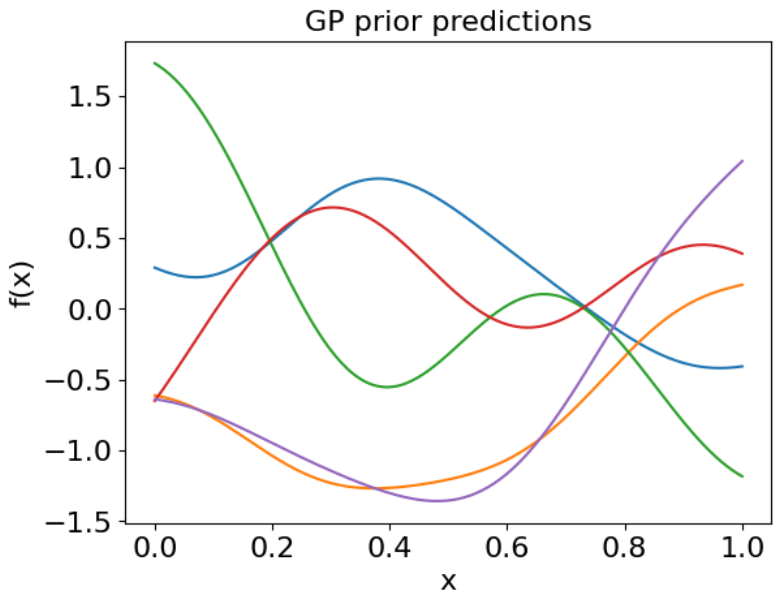
- Gaussian processes (GP) allow to obtain a **posterior distribution over functions**

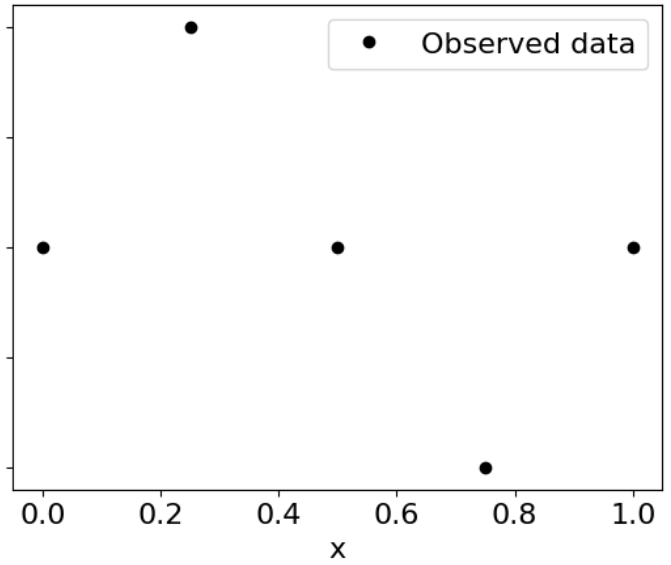
$$\text{Posterior} = \frac{\text{Likelihood x Prior}}{\text{Marginal Likelihood}}$$

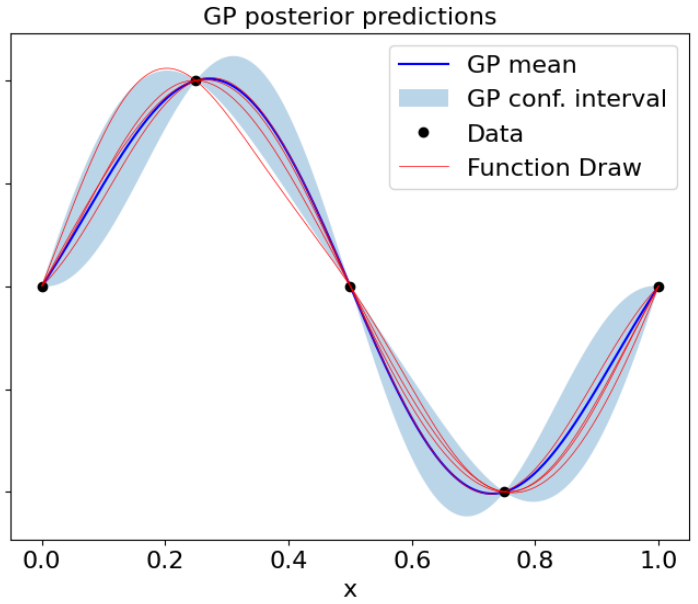
- GP mean and kernel encode the **prior belief about the smoothness and overall shape** of the function

- GPs provide uncertainty estimates and can flexibly model both midly and highly non-linear functions

$$p(f \mid \theta) = \mathcal{GP}(\mu_\theta(x), K_\theta(x, x'))$$





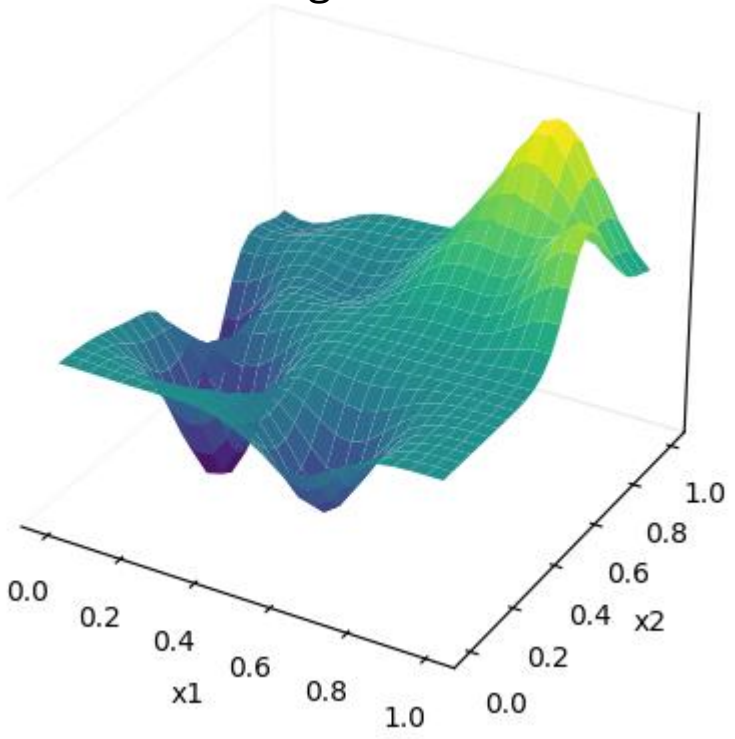


3. Methodology

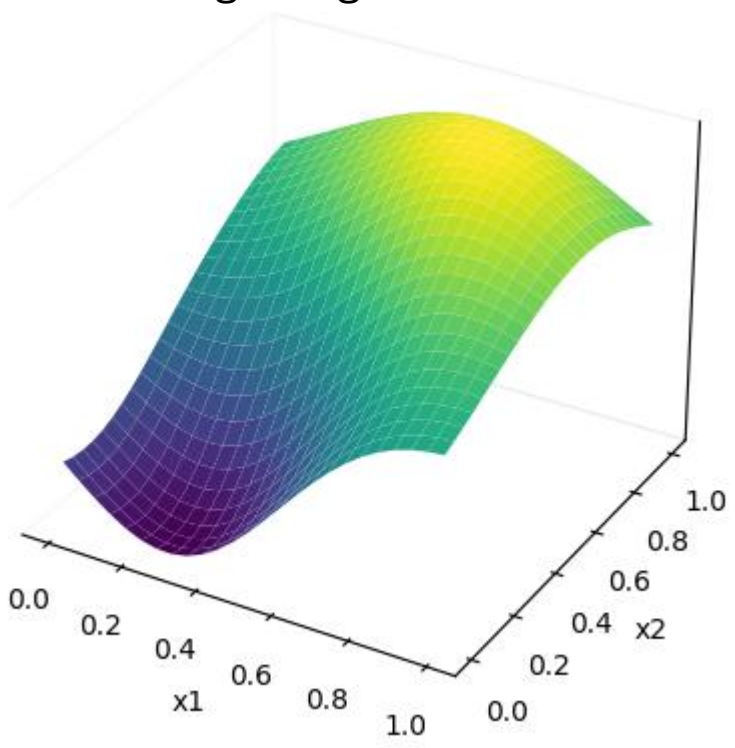
- The kernel lengthscale controls the complexity of the response surface given the distance between points

$$K(x, x') = \exp\left(-\frac{|| x - x' ||^2}{l^2}\right) \propto \frac{\text{Distance between points}}{\text{Lengthscale}}$$

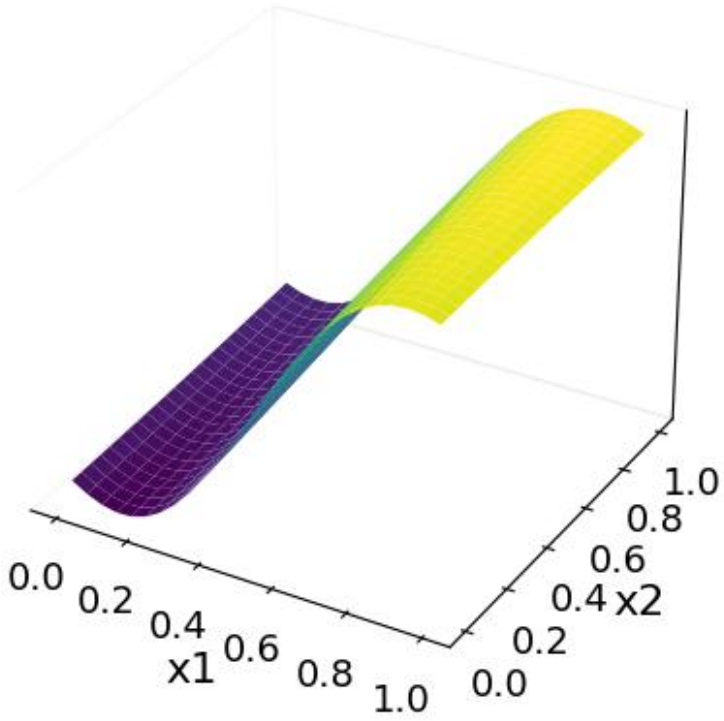
GP mean prediction with small lengthscales



GP mean prediction with large lengthscales



Anisotropic GP with large lengthscale for input 2 (input 2 is irrelevant)



3. Methodology

- GP hyperparameters are generally estimated by maximizing the marginal likelihood (Empirical Bayes)

$$\hat{\theta}_{MLE} = \arg\max -\frac{1}{2}\mathbf{y}[\mathbf{K}+\sigma_n^2\mathbf{I}]^{-1}\mathbf{y}^T - \frac{1}{2}\log|\mathbf{K}+\sigma_n^2\mathbf{I}| - \frac{N}{2}\log 2\pi$$

Data fit **Complexity**

- Being more Bayesian, GP hyperparameters can be treated as random variables with a prior distribution

- We can use **Maximum A Posteriori (MAP)** estimation as a way to encode more prior knowledge

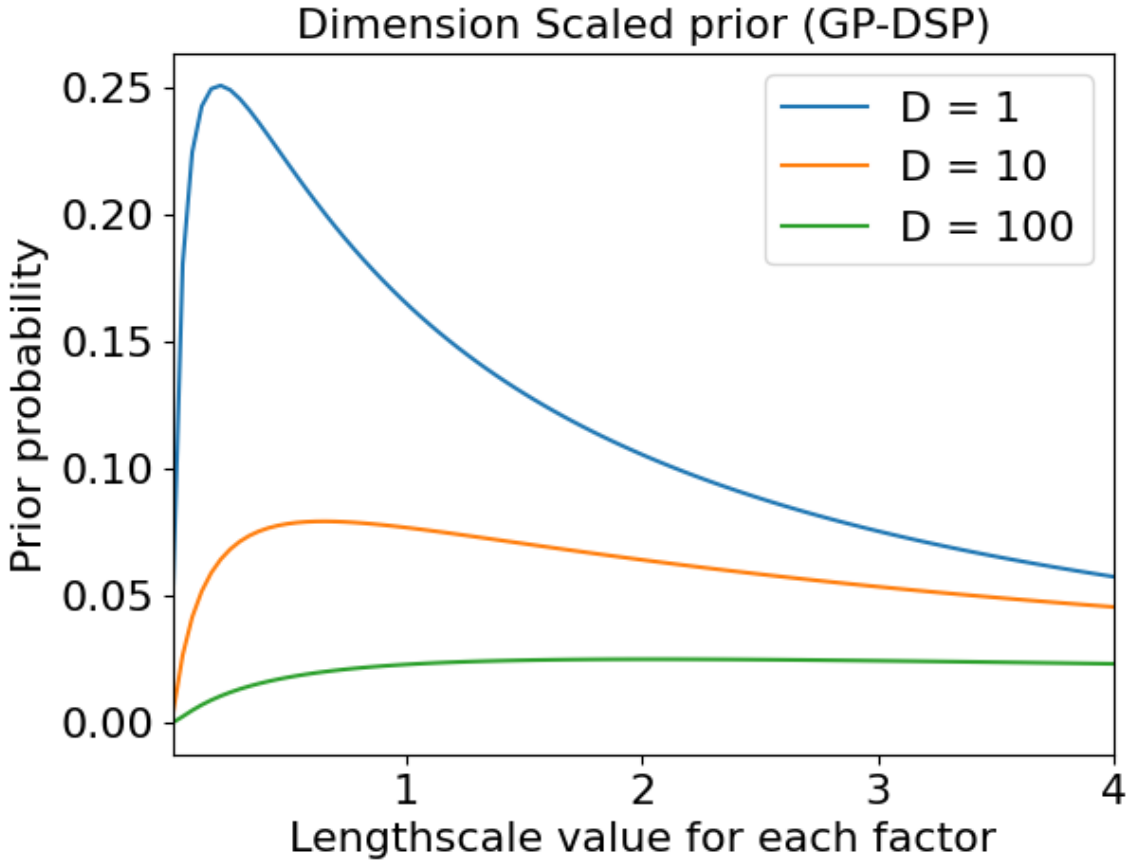
$$\hat{\theta}_{MAP} = \arg\max -\frac{1}{2}\mathbf{y}[\mathbf{K}+\sigma_n^2\mathbf{I}]^{-1}\mathbf{y}^T - \frac{1}{2}\log|\mathbf{K}+\sigma_n^2\mathbf{I}| - \frac{N}{2}\log 2\pi + \log p(\theta)$$

Data fit **Complexity** **Prior hyper-parameter probability**

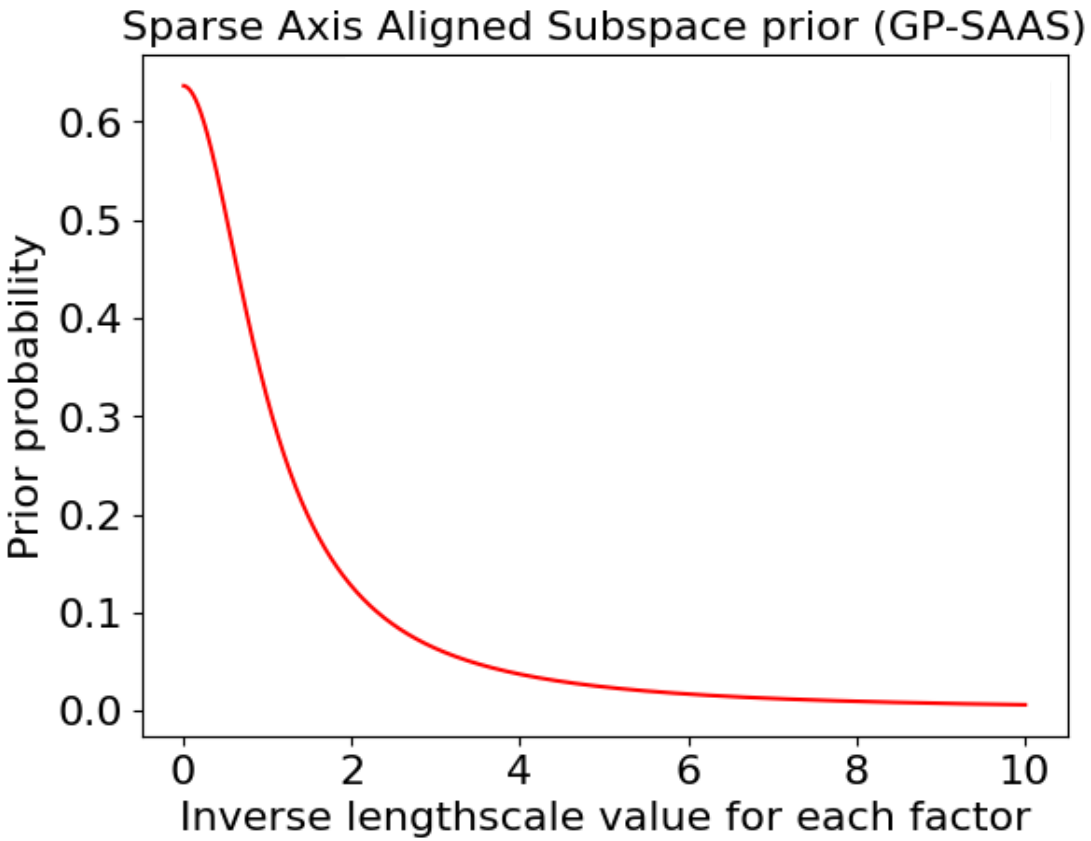
- Being **fully Bayesian**, we sample from the hyperparameter prior using Markov Chain Monte Carlo (MCMC)

3. Methodology

- Given the large number of factors, we can make two prior assumptions about the system under study [2,3]



Assumption: Non-linearity "decreases" with dimensionality
Inference: MAP estimation



Assumption: Only a subset of inputs is important (sparsity)
Inference: MCMC

[2] Hvarfner et al, 2024, *Vanilla Bayesian Optimization Performs Great in High Dimensions*, ICML
[3] Eriksson, Jankowiak, 2021, *High-Dimensional Bayesian Optimization with Sparse Axis-Aligned Subspaces*, UAI

3. Methodology

- Acquisition functions quantify the value of information (regarding optimum), conditioned on prior beliefs and data
- **Exploration** (gather new information) is combined with **exploitation** (optimize given current knowledge)
- In this work, we rely on the log-Expected Improvement acquisition function, using Botorch Python package [4]

$$\text{Log EI}(x) = \log \mathbb{E}\left[\max\{f(x) - y_{max}\}\right]$$

$$= \log\left(\frac{\mu(x) - y_{max}}{\sigma(x)}\right) + \log \sigma(x)$$

Predicted improvement over best value **(Exploitation)**

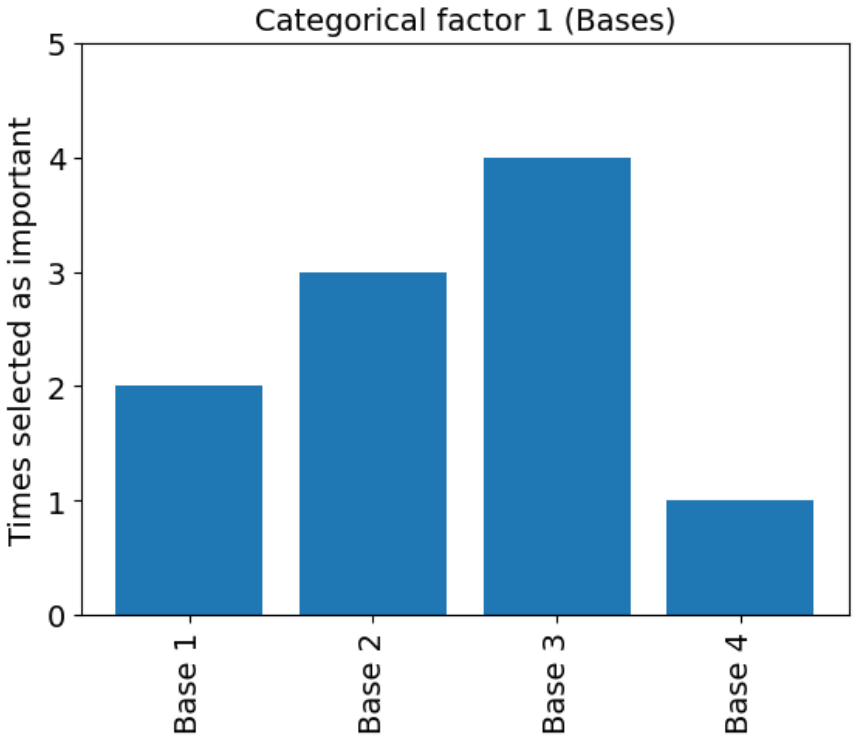
Standard deviation of prediction **(Exploration)**

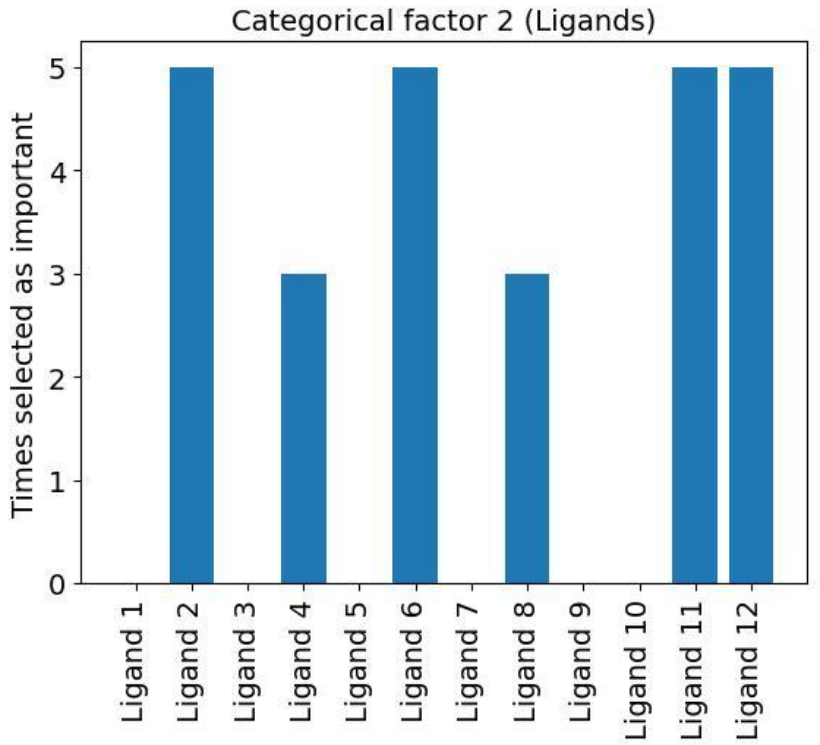
[4] Balandat et al 2020, BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization, NEURIPS

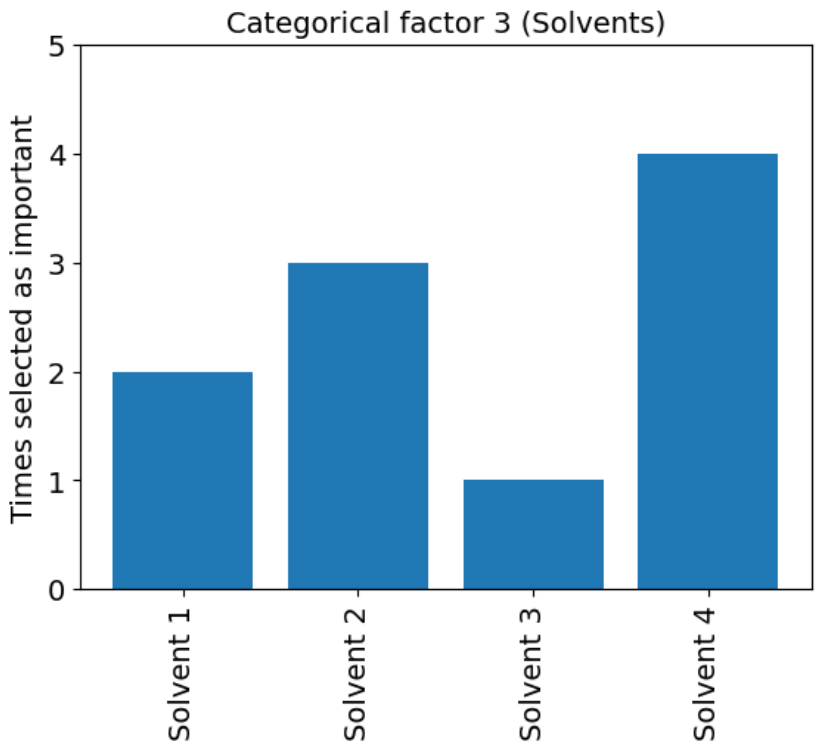
Results

4. Results

- D-optimal screening design can lead to different categorical level combinations with identical D-optimality
- To assess variability, 5 different repetions of sequential DOE are obtained by generating different D-optimal designs
- Different designs lead to different conclusions about factor importance
- I optimal design requires 39 to 44 additional runs for a total of 83 to 92 runs

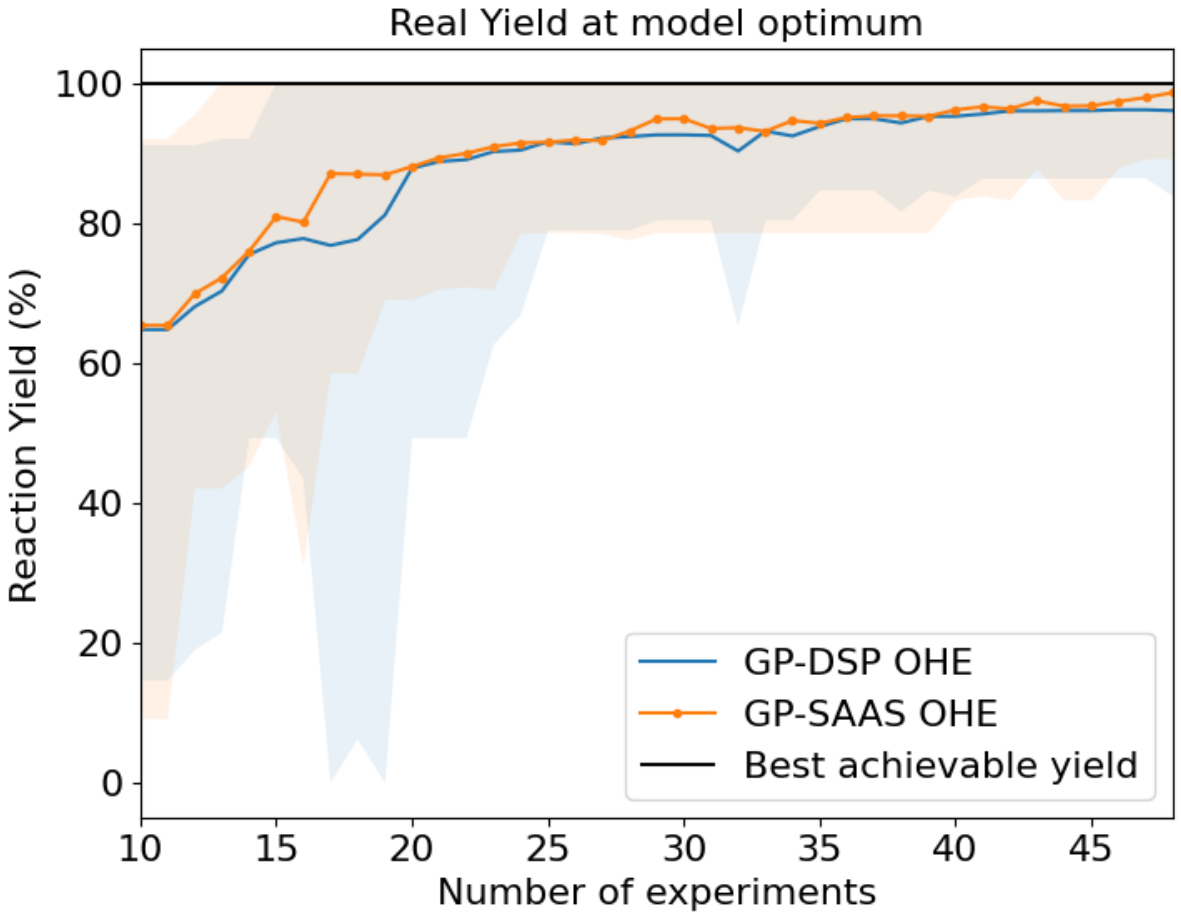


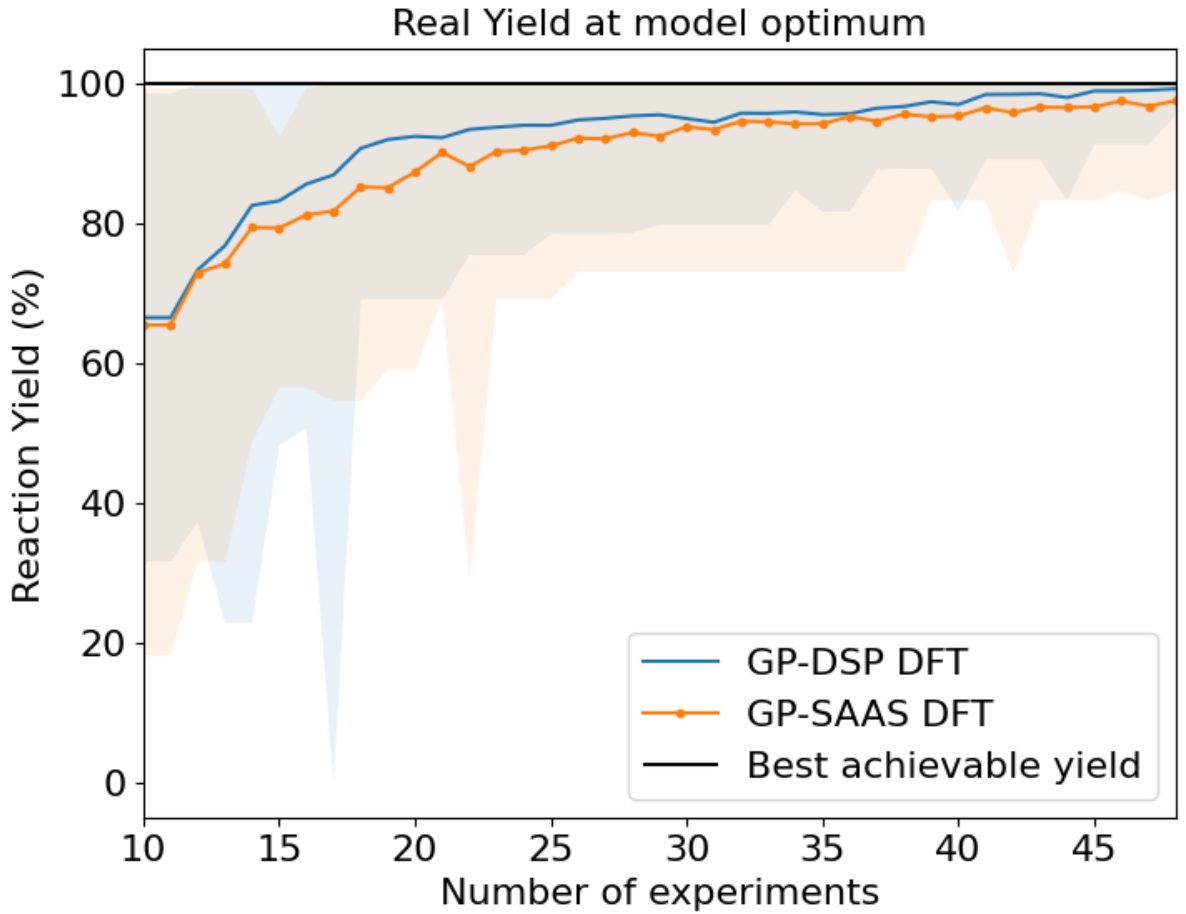




4. Results

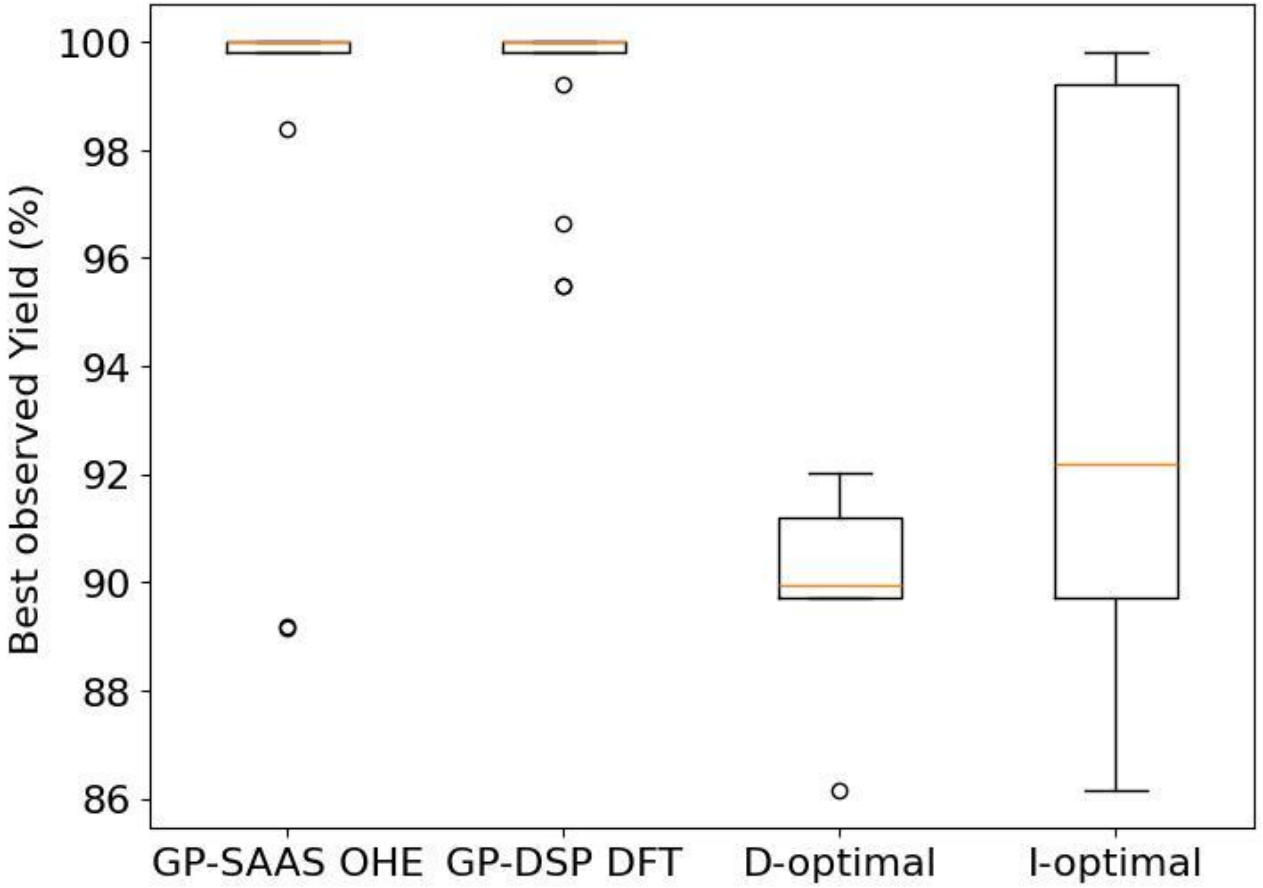
- 20 different BO repetions with 48 experiments, each with 10 initial experiments selected using random sampling
- All algorithms quickly converge near to the global optimum
- Best models: GP-SAAS for OHE, GP-DSP for DFT encodings

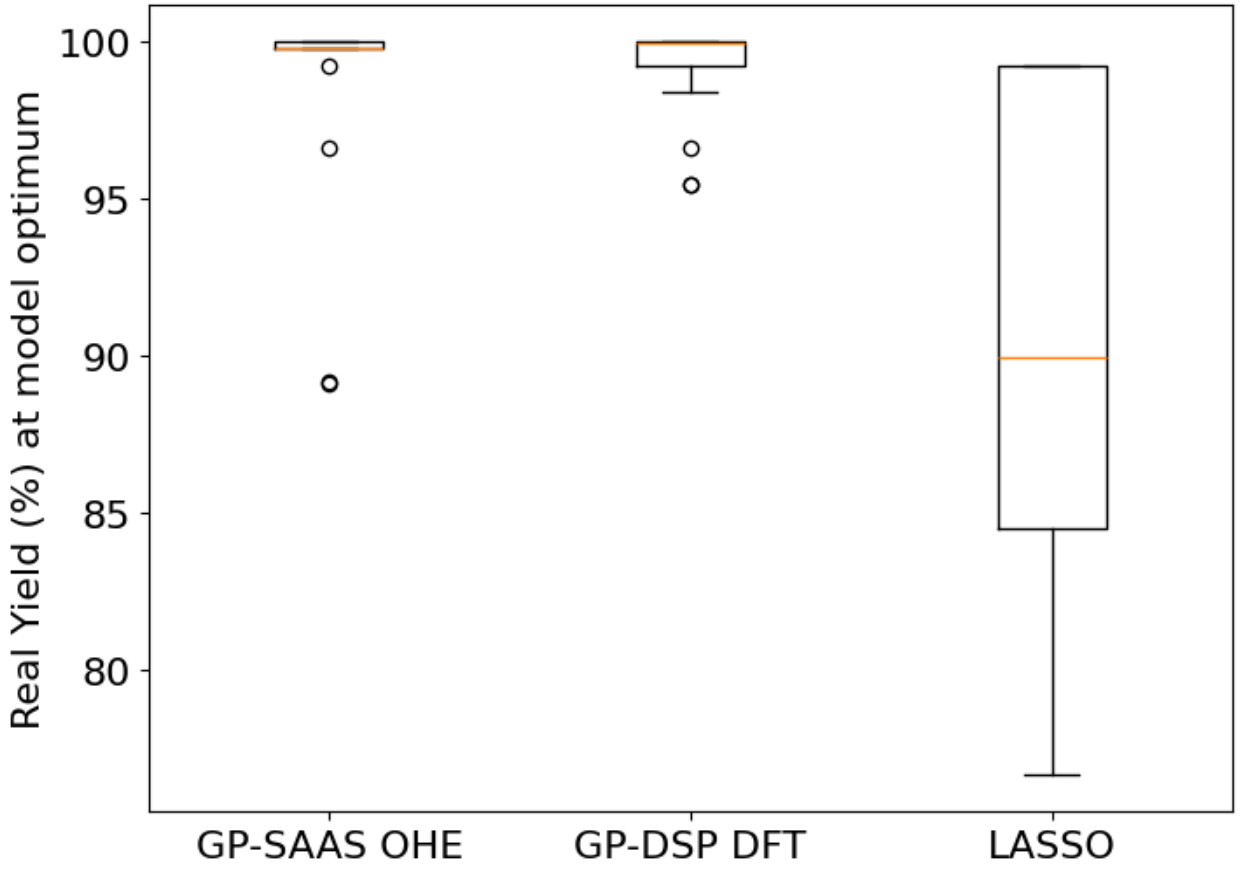




4. Results

- BO algorithms lead to better observed yields than both D and I optimal designs
- None of the 5 DOE repetions find the global optimum of 100% yield
- BO models lead to better results than final DOE models, while using only 48 runs (same as D-optimal screening design)

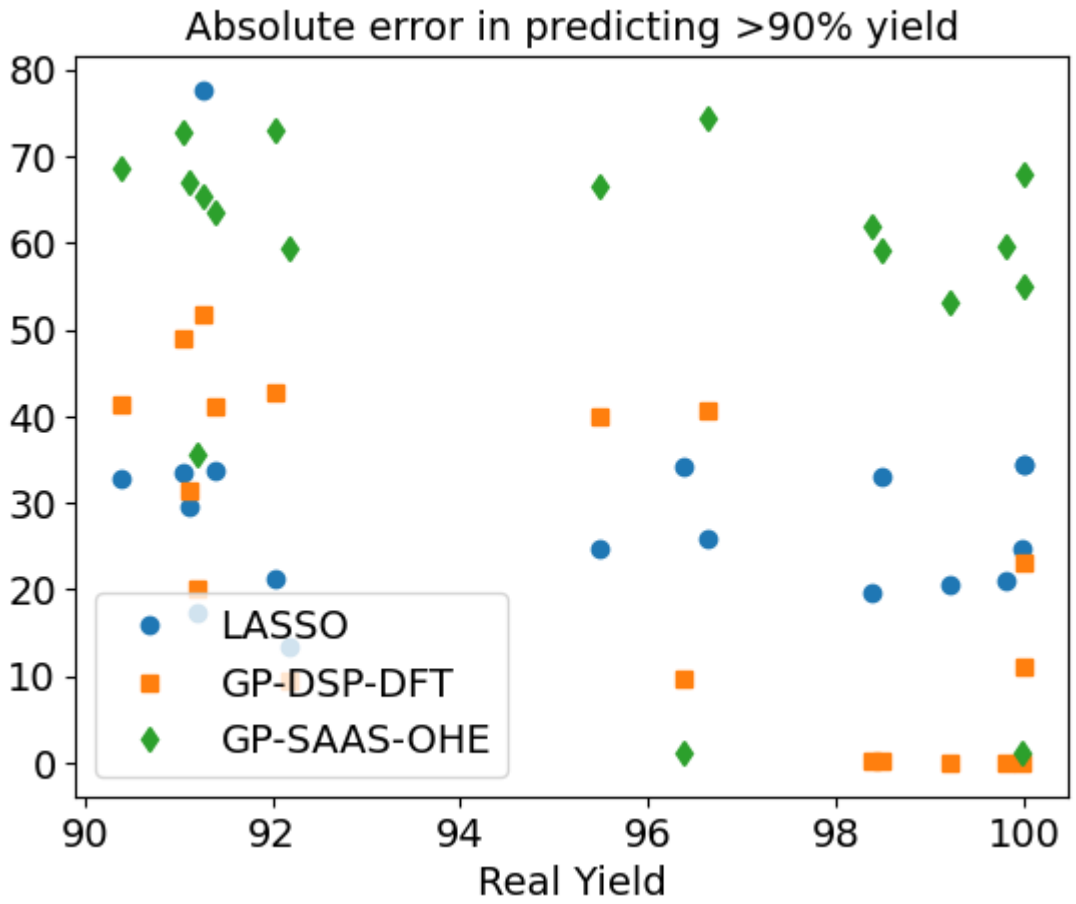




4. Results

- Full dataset (1728 samples) is used to assess global prediction quality in test scenarios
- GPs use all categorical factors, LASSO model only uses a subset of factors
- GP-DSP-DFT leads to slighly better overall predictions and better predictions around the optimum

Metric for global prediction quality (1728 samples)	LASSO (44 training samples)	GP-DSP-DFT (48 training samples)	GP-SAAS-OHE (48 training samples)
R squared	0,32	0,34	0,36
Root mean squared error	20,2	19,5	16,0



Conclusions

5. Conclusions

- BO leads to faster convergence than sequential DOE on the same modeling basis (OHE)

- DFT encodings can improve optimization performance – **291 covariates are efficiently used within BO**

- BO is an efficient appproach to experimental design with many categorical factors:
  - Faster convergence than sequential DOE
  - Avoid variability and uncertainty in the assessing important factors
  - Easier human implementation by avoiding screening stage and restrictive modeling assumptions

- We are currently exploring ways to include DOE principles that could improve BO even further

Thank you for your attention

João P. L. Coutinho*, You Peng**, Ricardo Rendall**, Caterina Rizzo**, Kaiwen Ma**, Swee-Teng Chin**, Ivan Castillo**, Marco S. Reis*

* University of Coimbra, CERES, Department of Chemical Engineering, Portugal

** The Dow Chemical Company