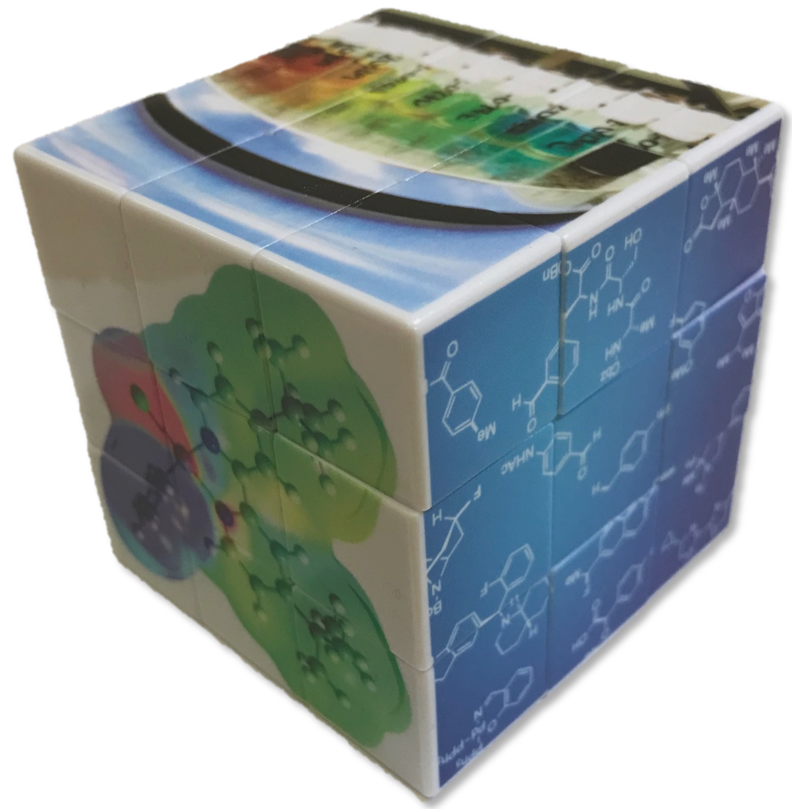
Development of Bayesian Optimization for Chemical Synthesis

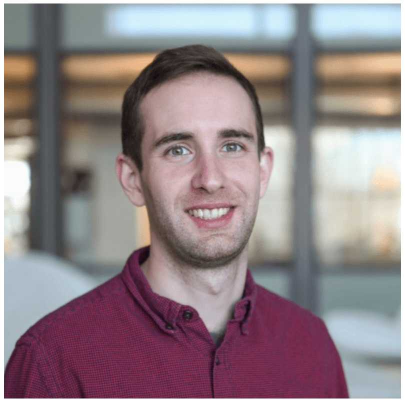
Abigail Doyle | UCLA

CAMLC25 Workshop, September 18th, 2025



Contributing new approaches to chemical synthesis and catalysis by embracing diverse skills and perspectives, and by building a research culture that enables everyone to develop to their highest potential.

Bayesian Optimization: the Team

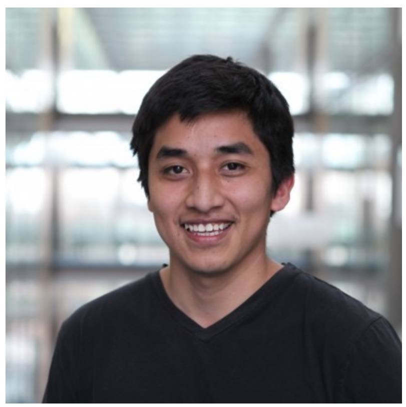

Benjamin Shields


Jason Stevens


Jun Li


Marvin Parasram


Farhan Damani


Jesus Martinez-Alvarado


Jacob Janey


Ryan Adams

C-CAS Center for Computer Assisted Synthesis

Bristol-Myers Squibb

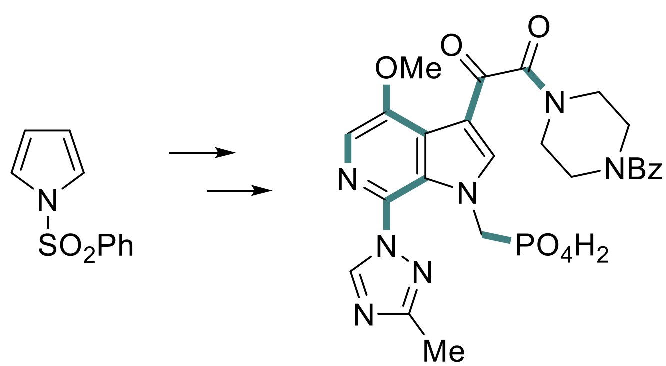
Princeton Catalysis Initiative

NSF

Reaction optimization is ubiquitous in chemistry

Process Optimization

Late-stage synthesis and preparation scale



Experimental HIV inhibitor

Optimize yield for a single target molecule

Important Experimental Parameters

Categorical

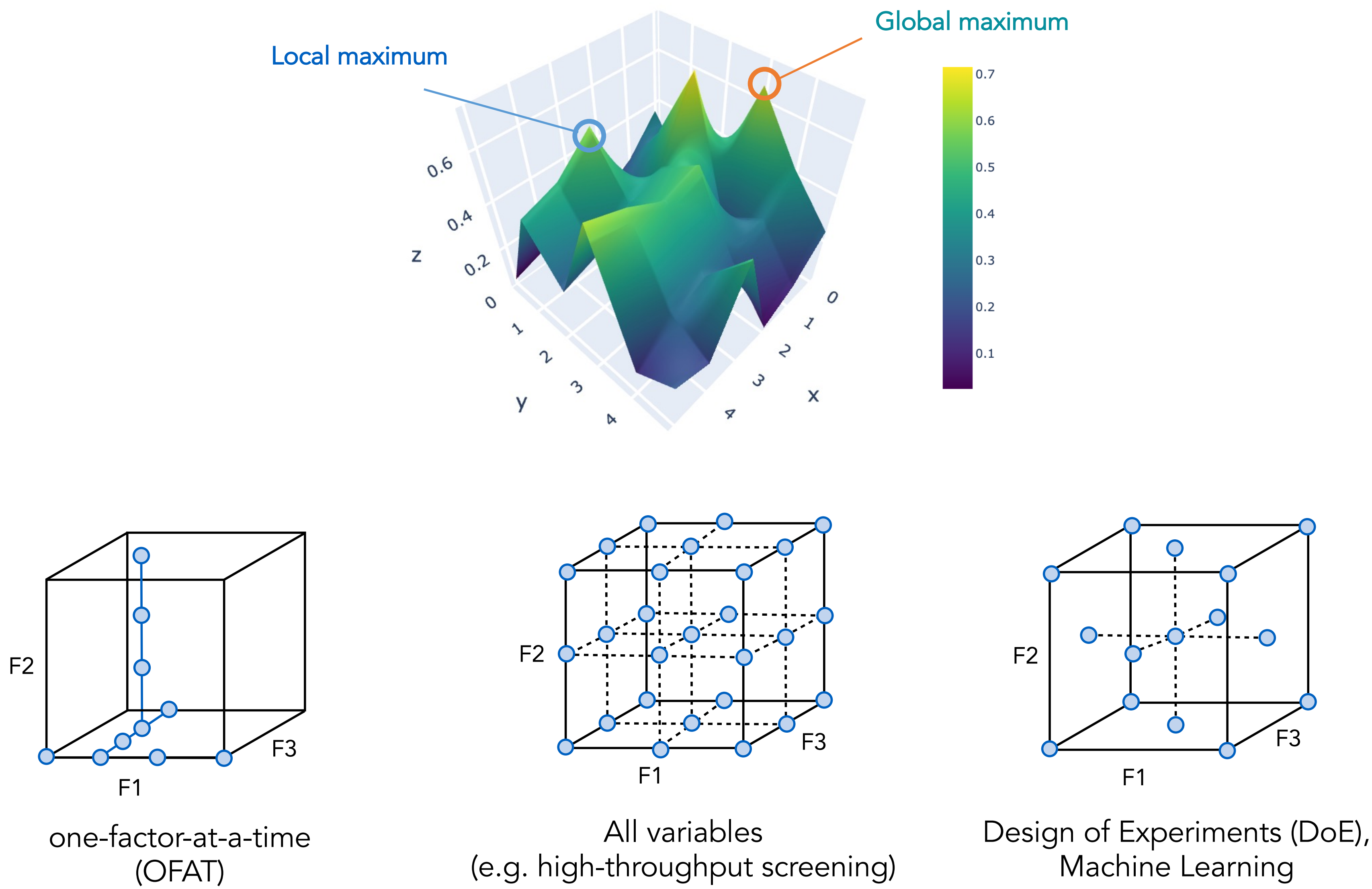
Catalyst (precat./ligand)
Acid/Base
Oxidant/Reductant
Solvent
Additive
...

Continuous

[Concentration]
Equivalents
Temperature
Pressure
Time
...

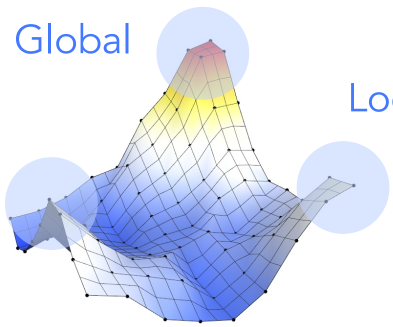
Chemists rely on literature precedent, intuition, & mechanistic understanding

Reaction Optimization in Synthetic Chemistry



Introduction to Bayesian Optimization

Bayesian optimization is a general approach to global optimization of expensive black-box functions



Global optimization with min experiments

$$\boldsymbol{EI(x)} = \begin{cases} (\boldsymbol{\mu(x)} - \boldsymbol{f^{+}} - \boldsymbol{\delta})\boldsymbol{\Phi} + \underbrace{\boldsymbol{\sigma(x)}}_{SD}\boldsymbol{\varphi} \\ \boldsymbol{0} \end{cases}$$

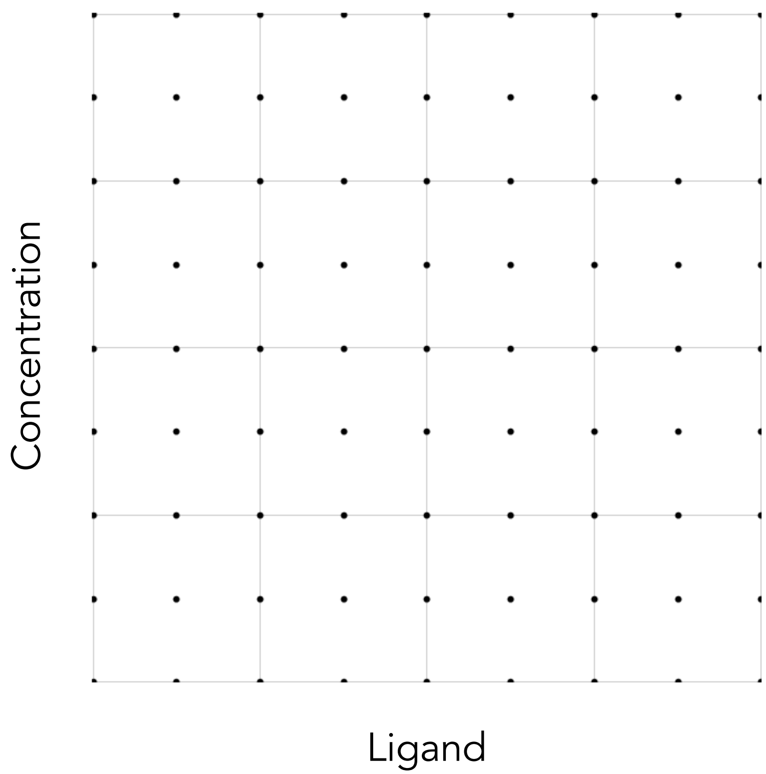
$\uparrow SD \rightarrow \uparrow Expected\ Improvement$

Statistical uncertainty drives optimization

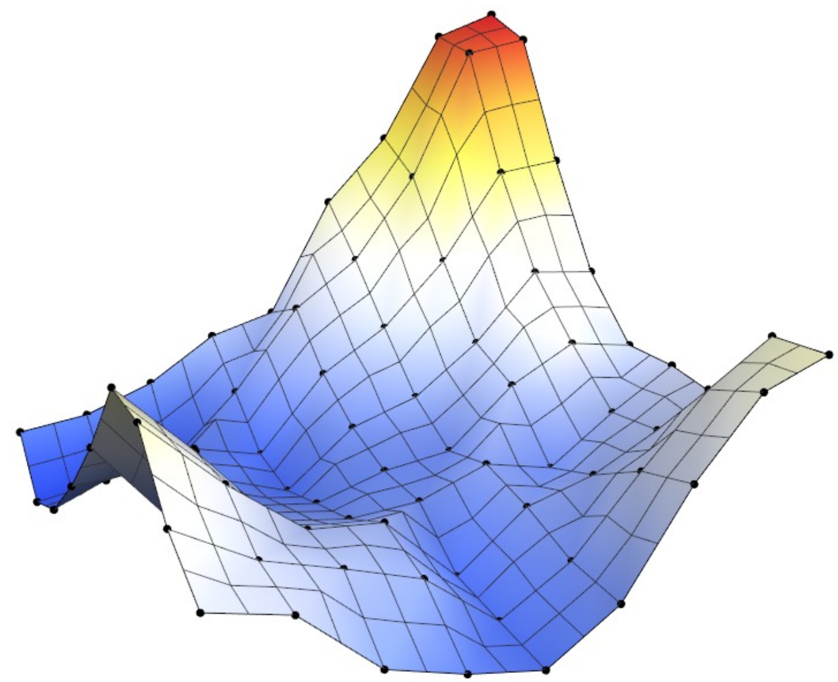
Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; de Freitas, N. *Proceedings of the IEEE*, **2016**, *104*, 148
Practical Bayesian Optimization of Machine Learning Algorithms. *NIPS Proceedings*, 2012

Sequential decision making with Bayesian optimization

Search Space

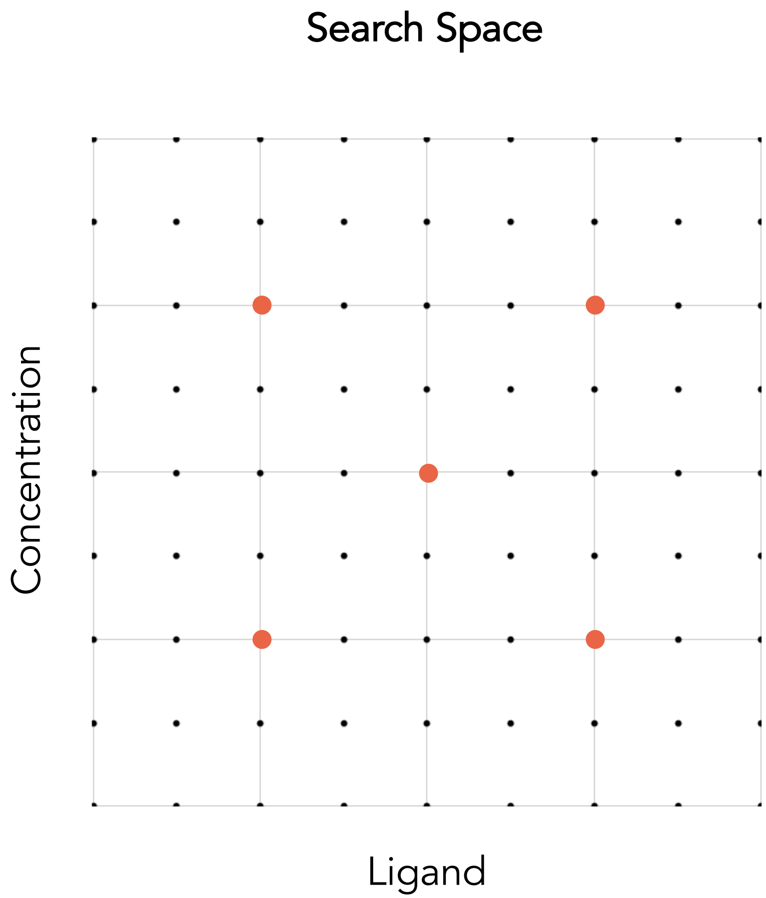


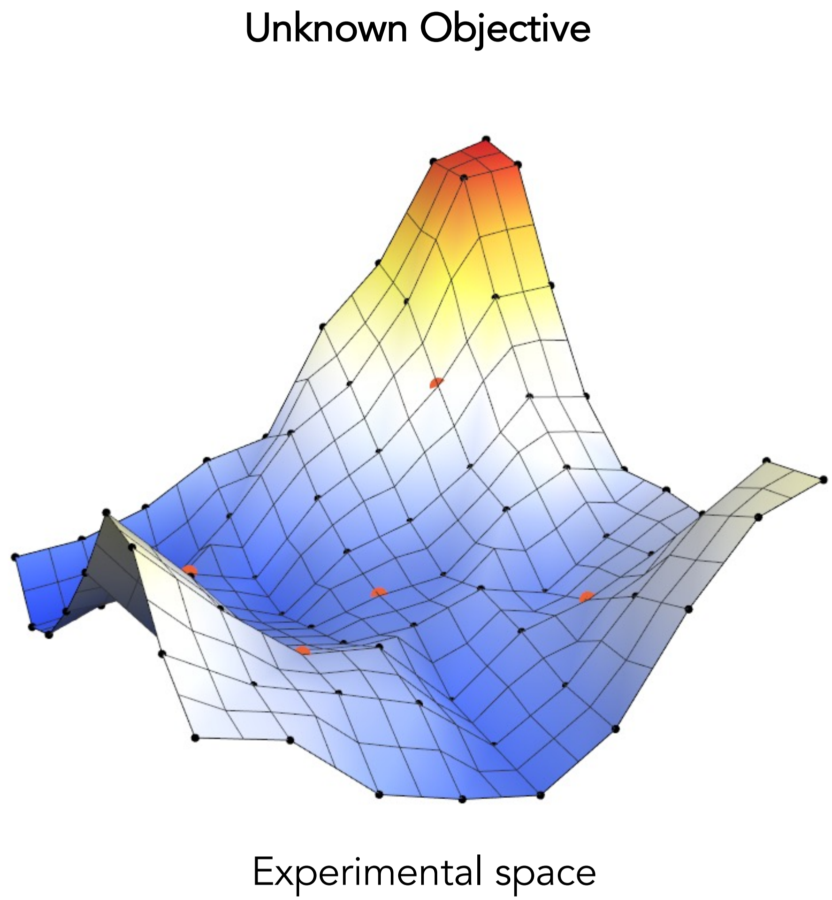
Unknown Objective



We start by defining a search space over which we seek to maximize an experimental outcome

Sequential decision making with Bayesian optimization

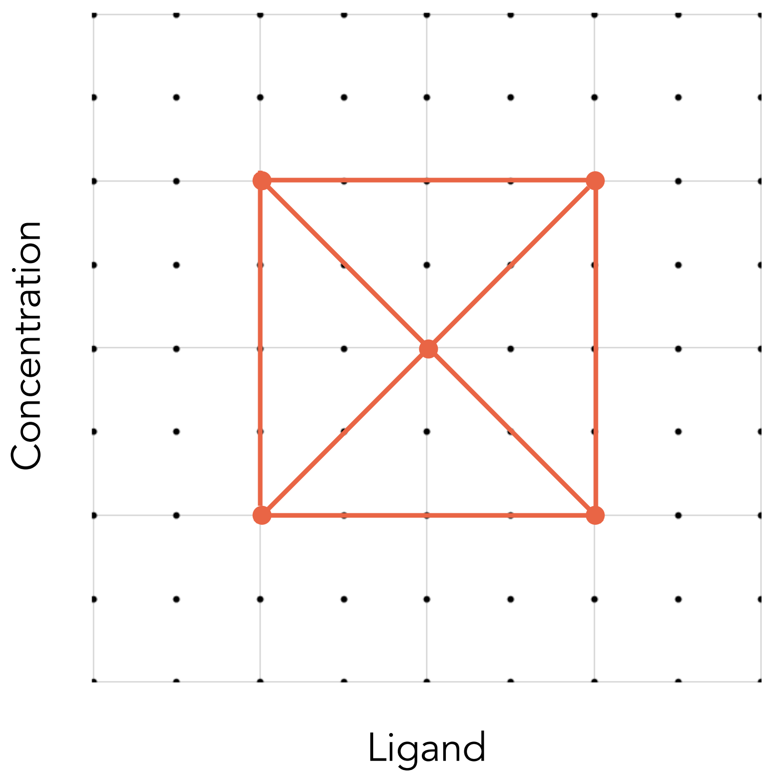




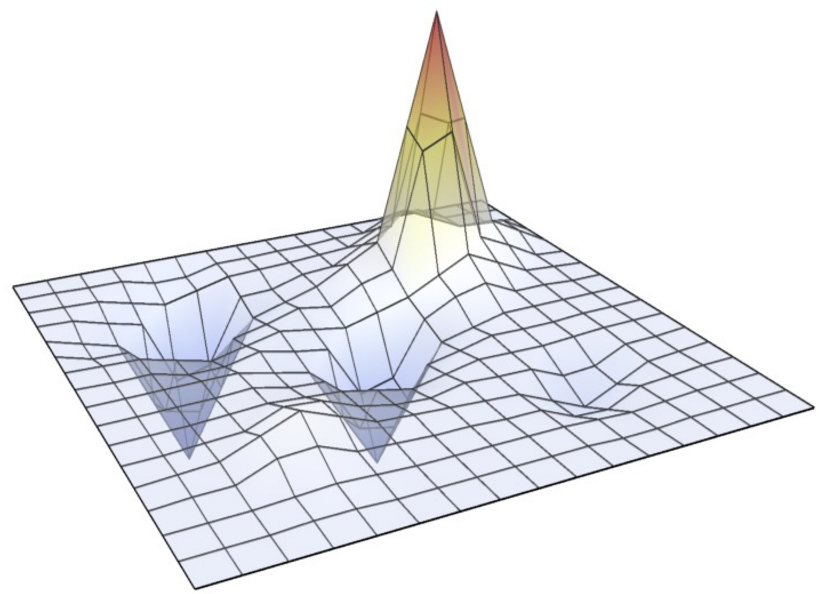
The optimizer is initialized by collecting data using some experimental design (or by using existing results)

Sequential decision making with Bayesian optimization

Search Space



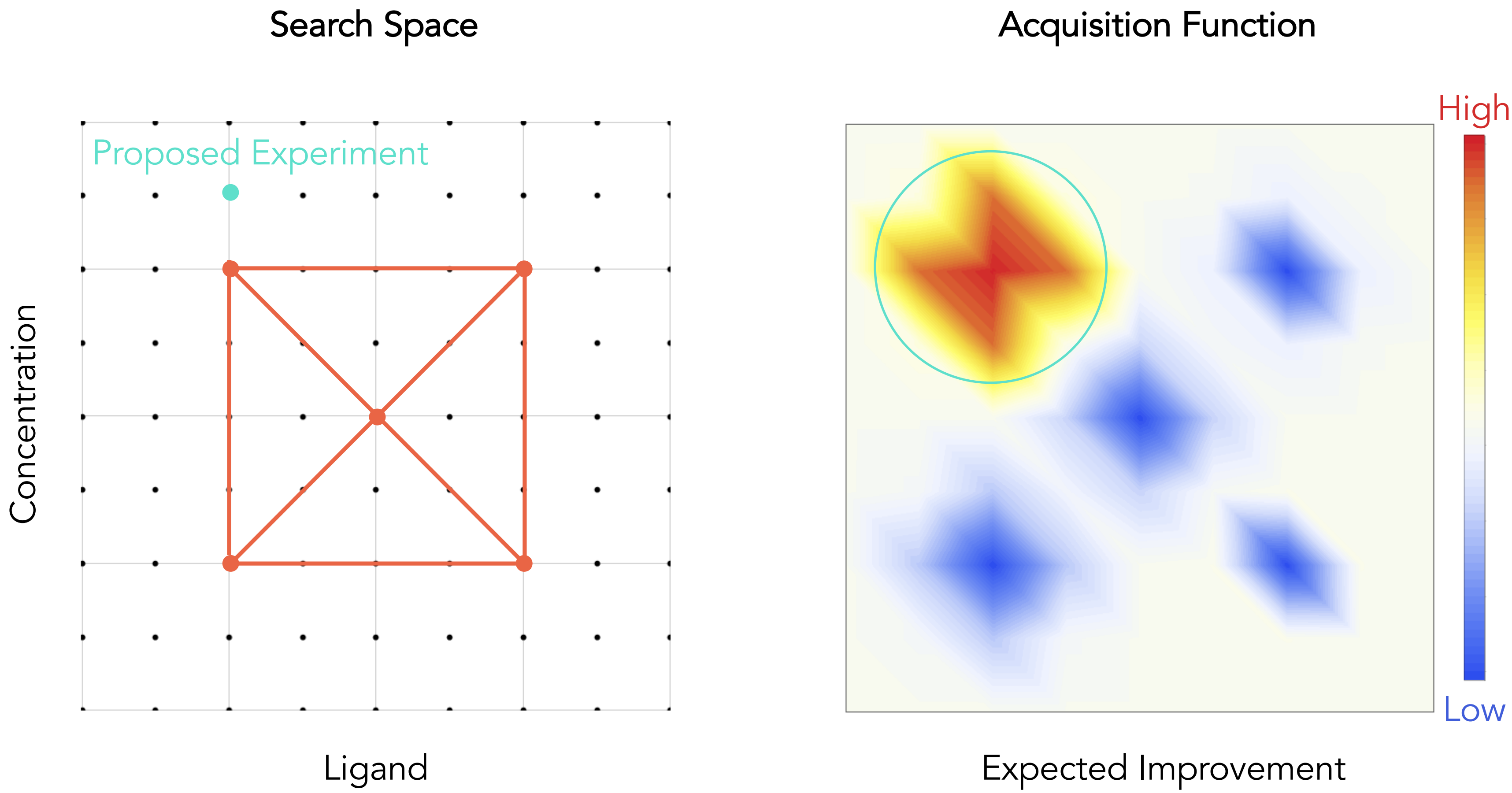
Surrogate Model



Gaussian process

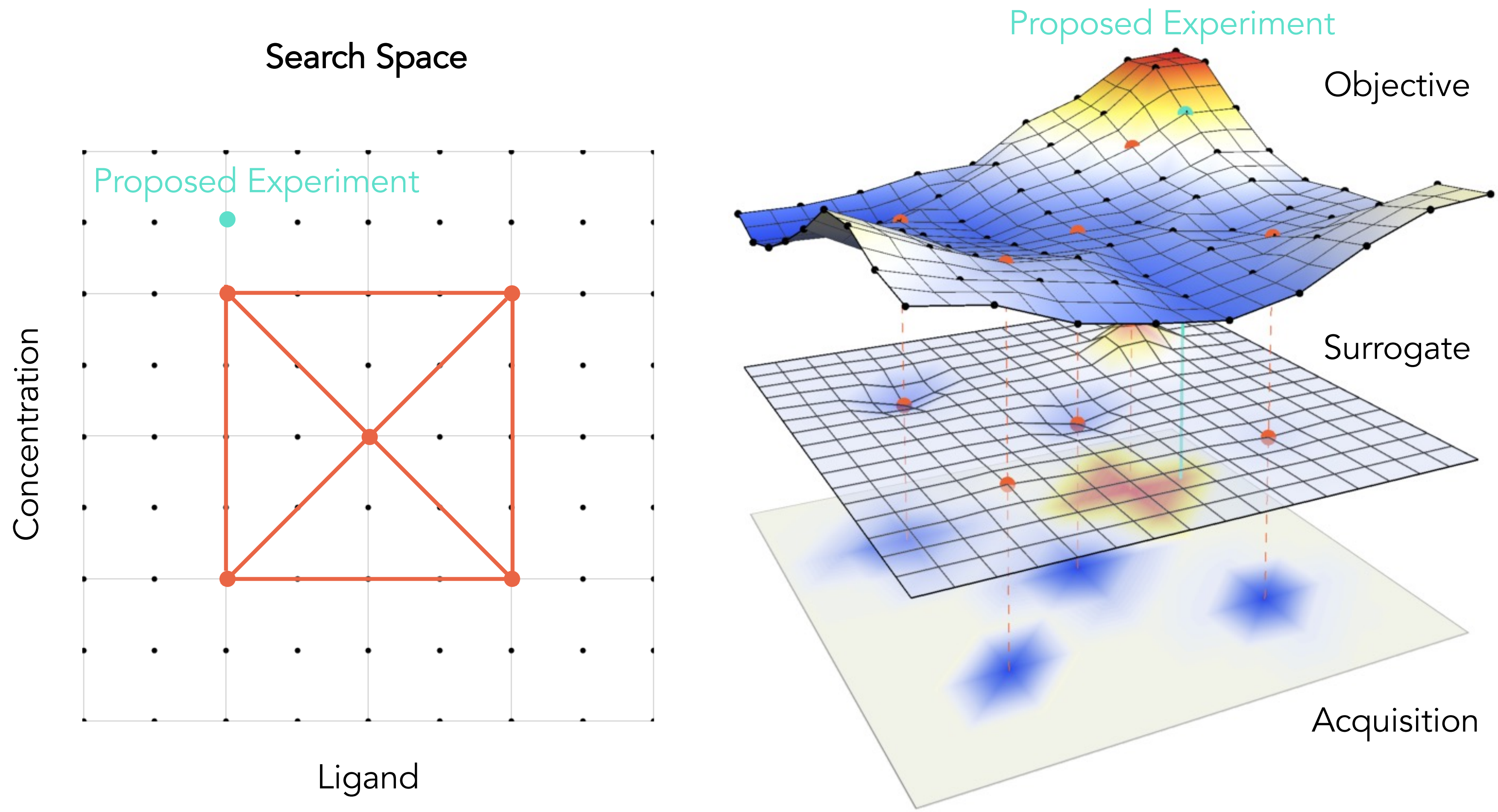
A model is then fit to the data – key model requirements are predictions and estimates of variance

Sequential decision making with Bayesian optimization



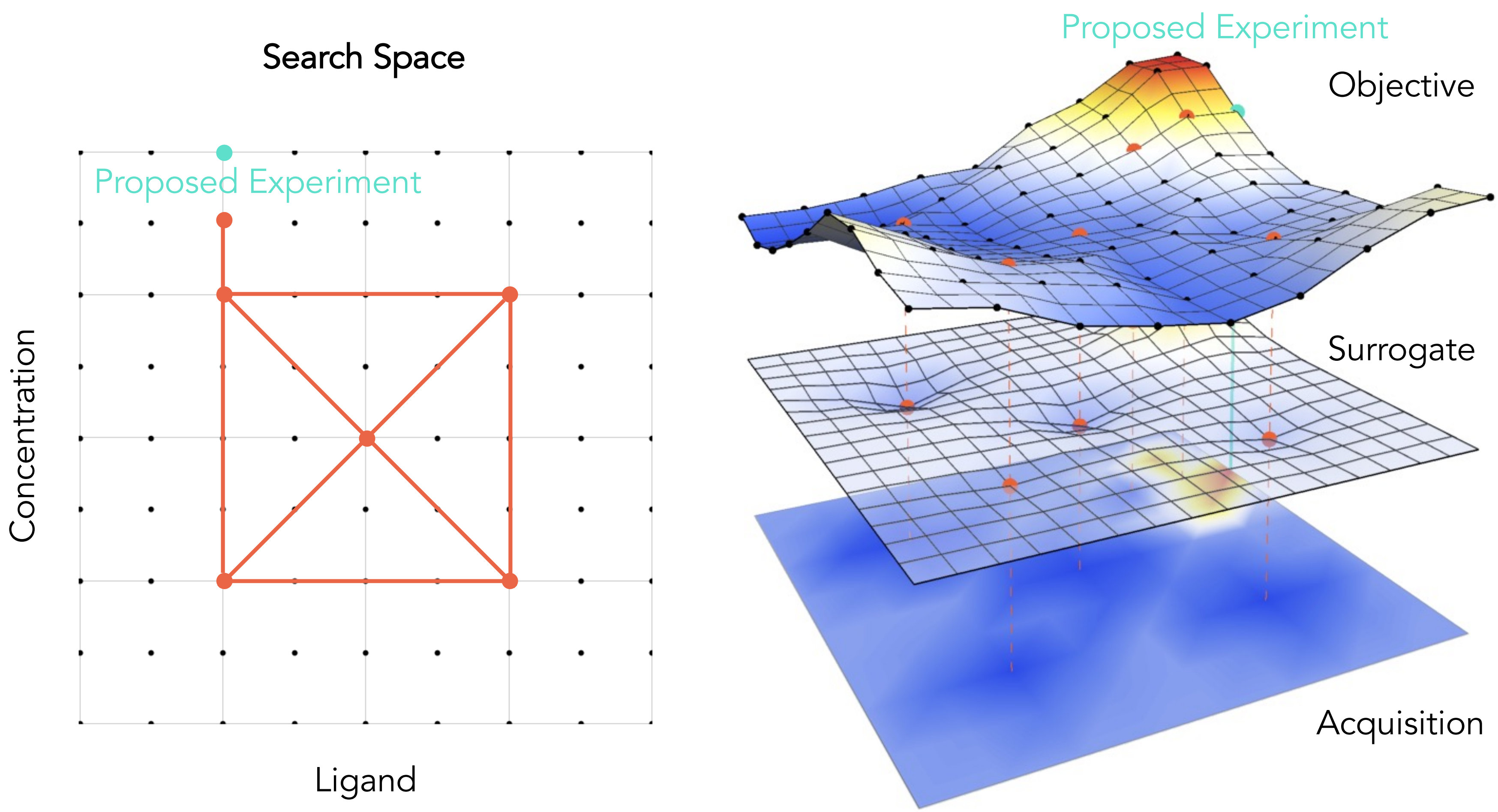
An acquisition function is derived from the surrogate model to select the next experiment(s).

Sequential decision making with Bayesian optimization



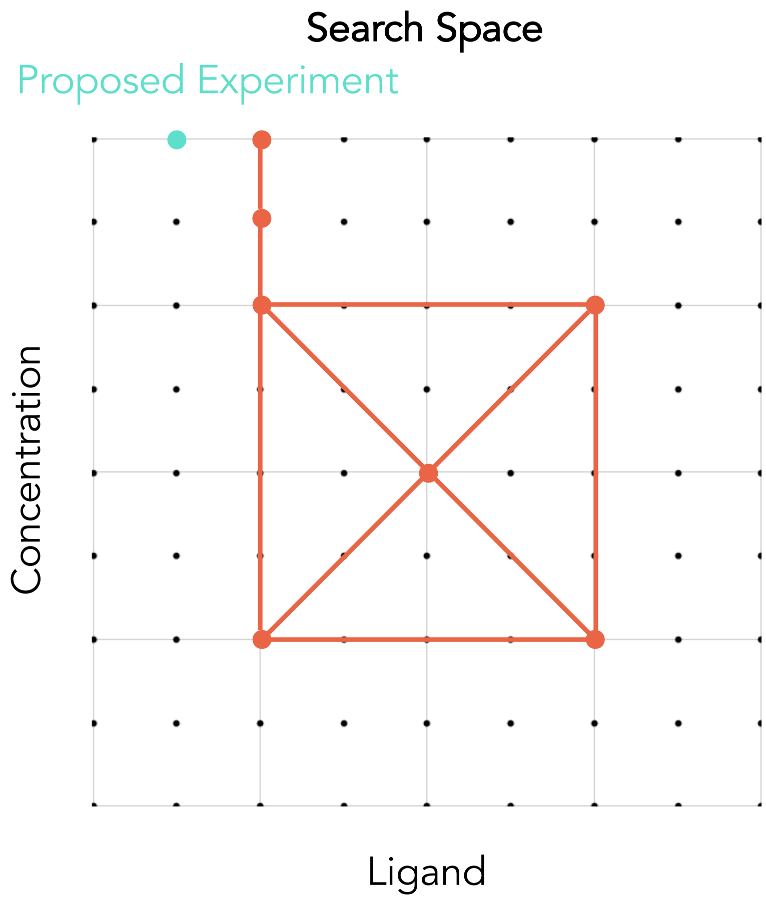
This process is iterated until optimization is complete or resources are depleted

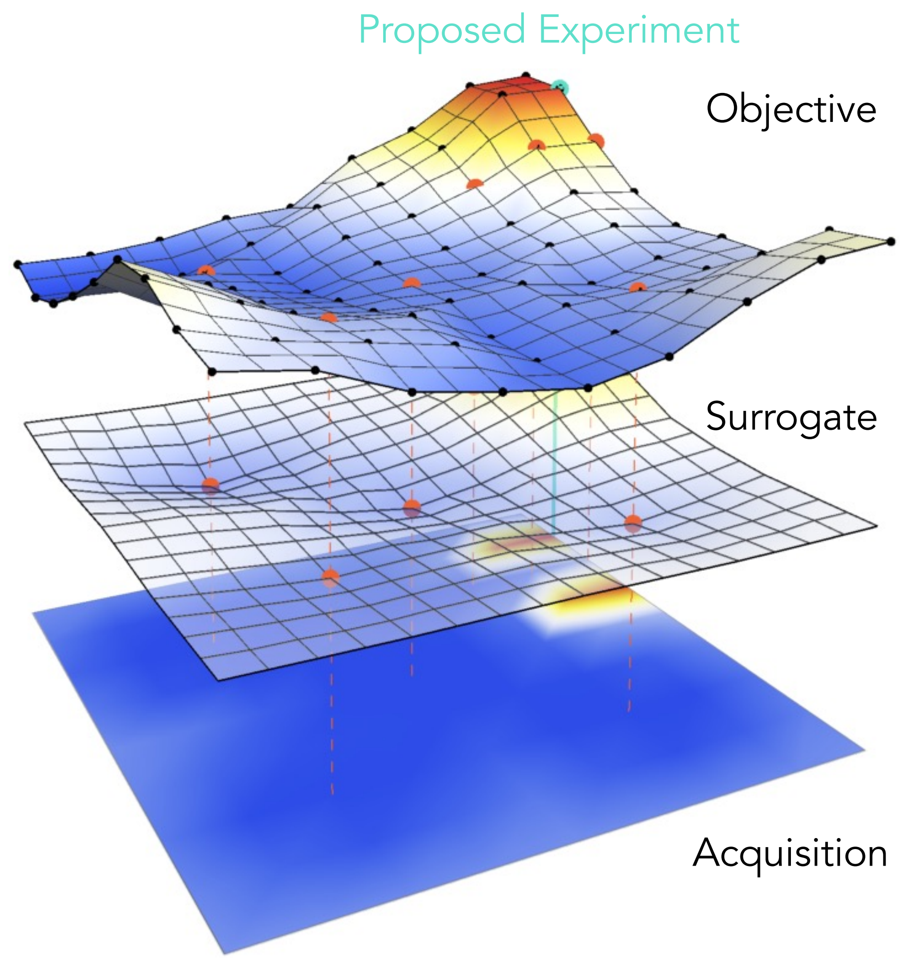
Sequential decision making with Bayesian optimization



This process is iterated until optimization is complete or resources are depleted

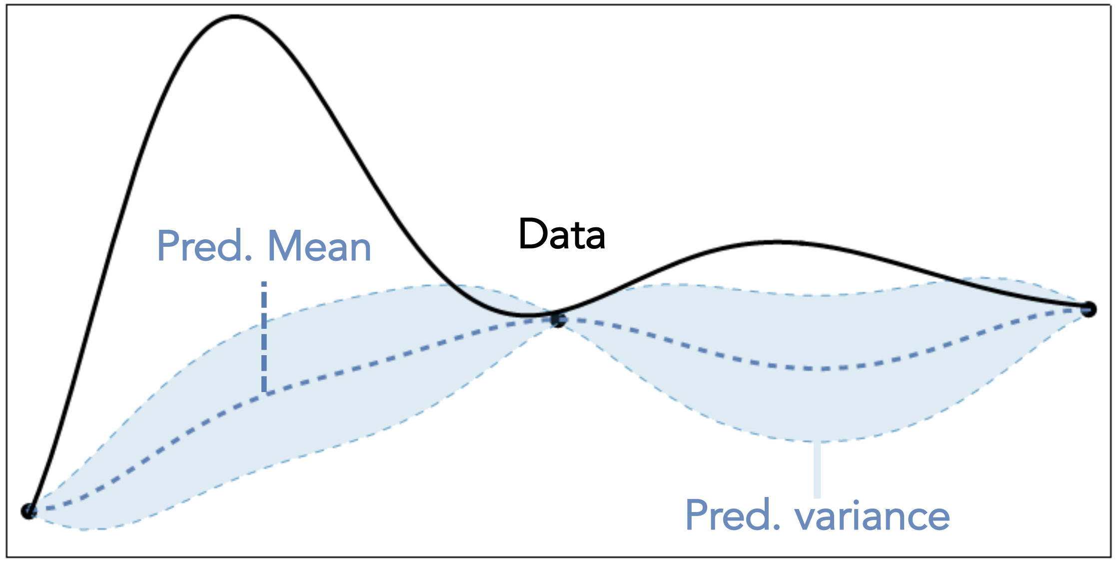
Sequential decision making with Bayesian optimization





This process is iterated until optimization is complete or resources are depleted

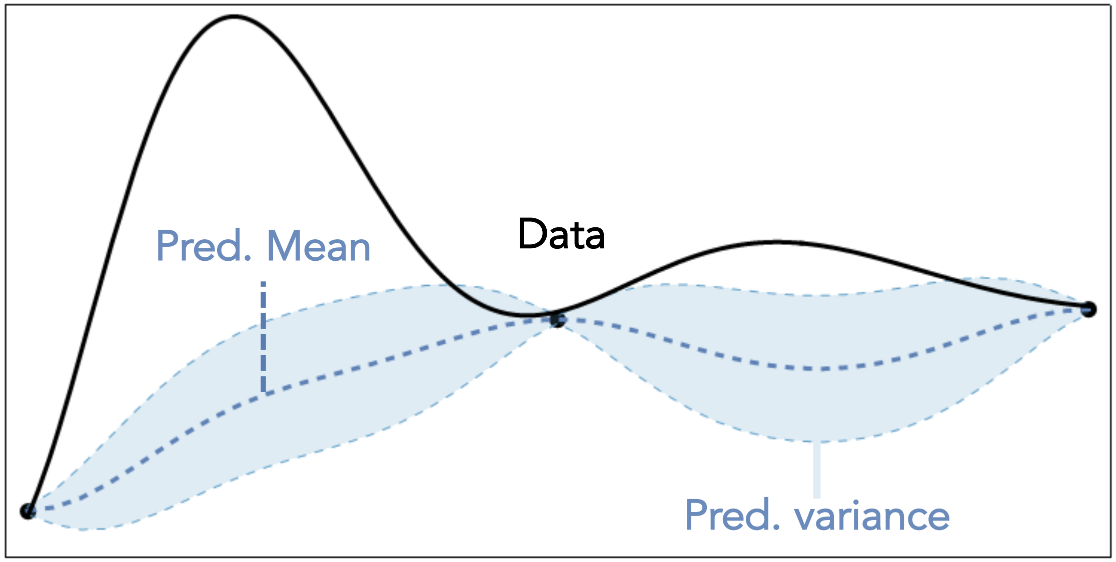
Acquisition functions: Decision making algorithms



Balancing exploration and exploitation	
Exploration of new areas of the space	***High variance***
Exploitation of information	***High mean***

These algorithms seek to make decision which maximize desired outcome and/or minimize uncertainty

Acquisition functions: Decision making algorithms

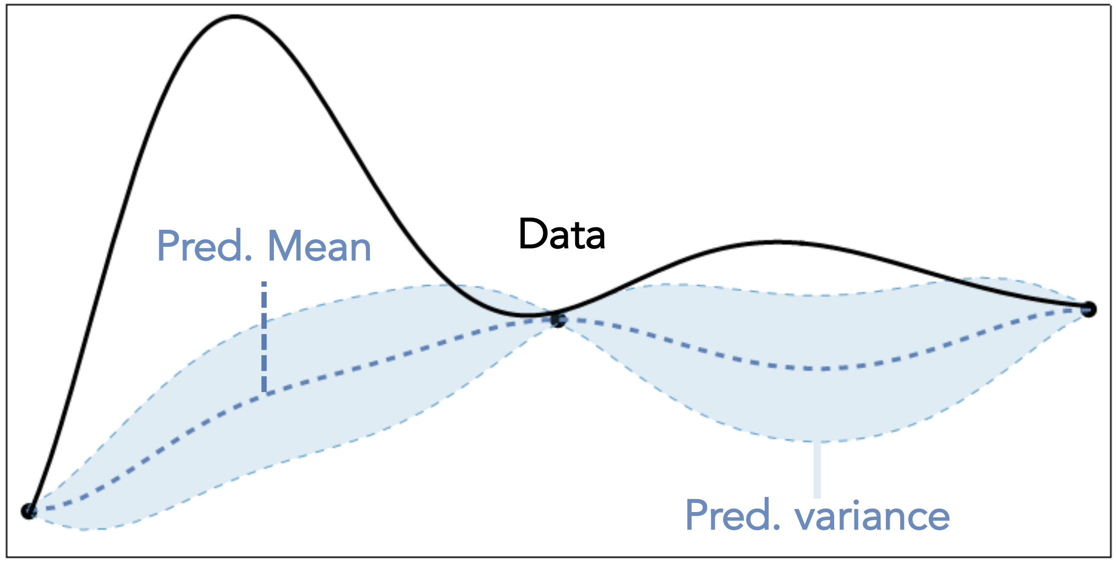


Example: Expected Improvement

$$EI(x) = \begin{cases} (\mu(x) - f(x^+) - \delta)\Phi + \sigma(x)\varphi & if\ \sigma(x) > 0 \\ 0 & if\ \sigma(x) = 0 \end{cases}$$

EI quantifies the average (expected) improvement over the best observed value

Acquisition functions: Decision making algorithms



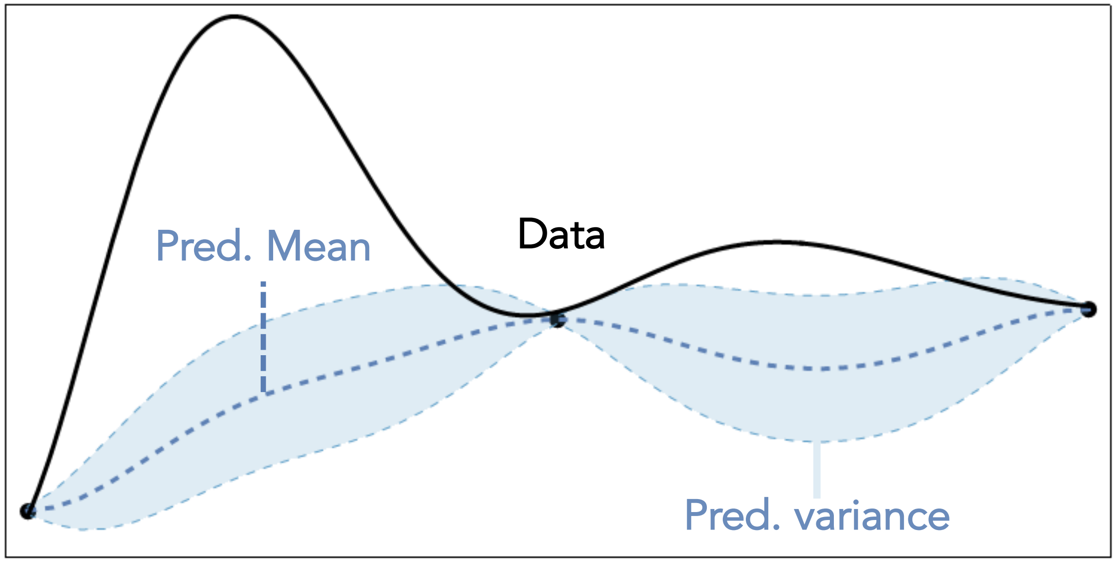
Example: Expected Improvement

Mean   Best y

$$EI(x) = \begin{cases} (\mu(x) - f(x^+) - \delta)\Phi + \sigma(x)\varphi & if\ \sigma(x) > 0 \\ 0 & if\ \sigma(x) = 0 \end{cases}$$

EI increases when models posterior mean is greater than the best observed value

Acquisition functions: Decision making algorithms

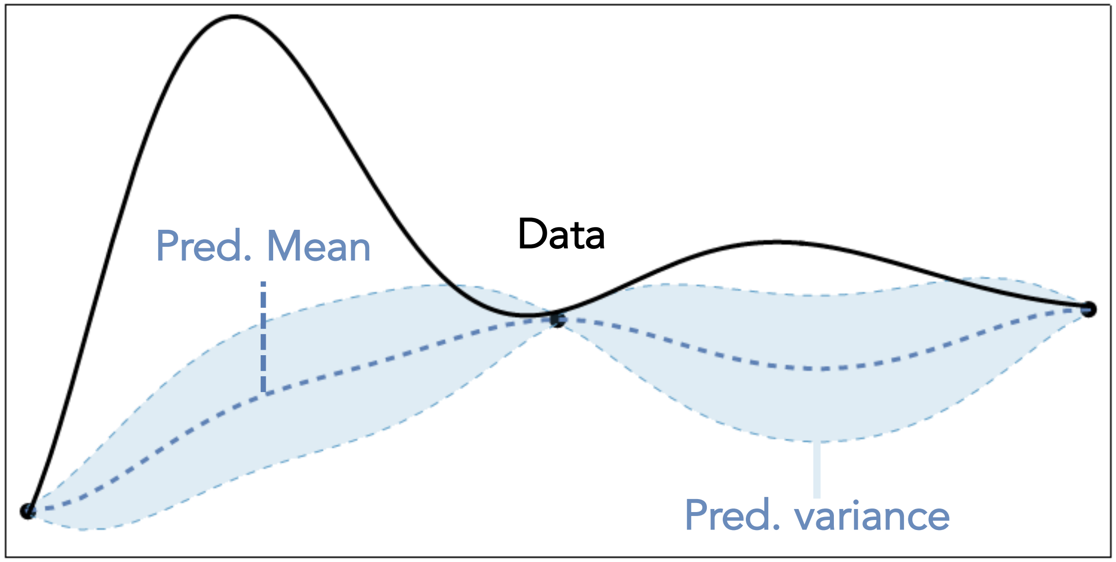


Example: Expected Improvement

$$EI(x) = \begin{cases} (\underbrace{\mu(x)}_{\text{Mean}} - \underbrace{f(x^+)}_{\text{Best y}} - \delta)\Phi + \underbrace{\sigma(x)}_{\text{SD}}\varphi & if\ \sigma(x) > 0 \\ 0 & if\ \sigma(x) = 0 \end{cases}$$

EI increases as the model posterior standard deviation increases

Acquisition functions: Decision making algorithms



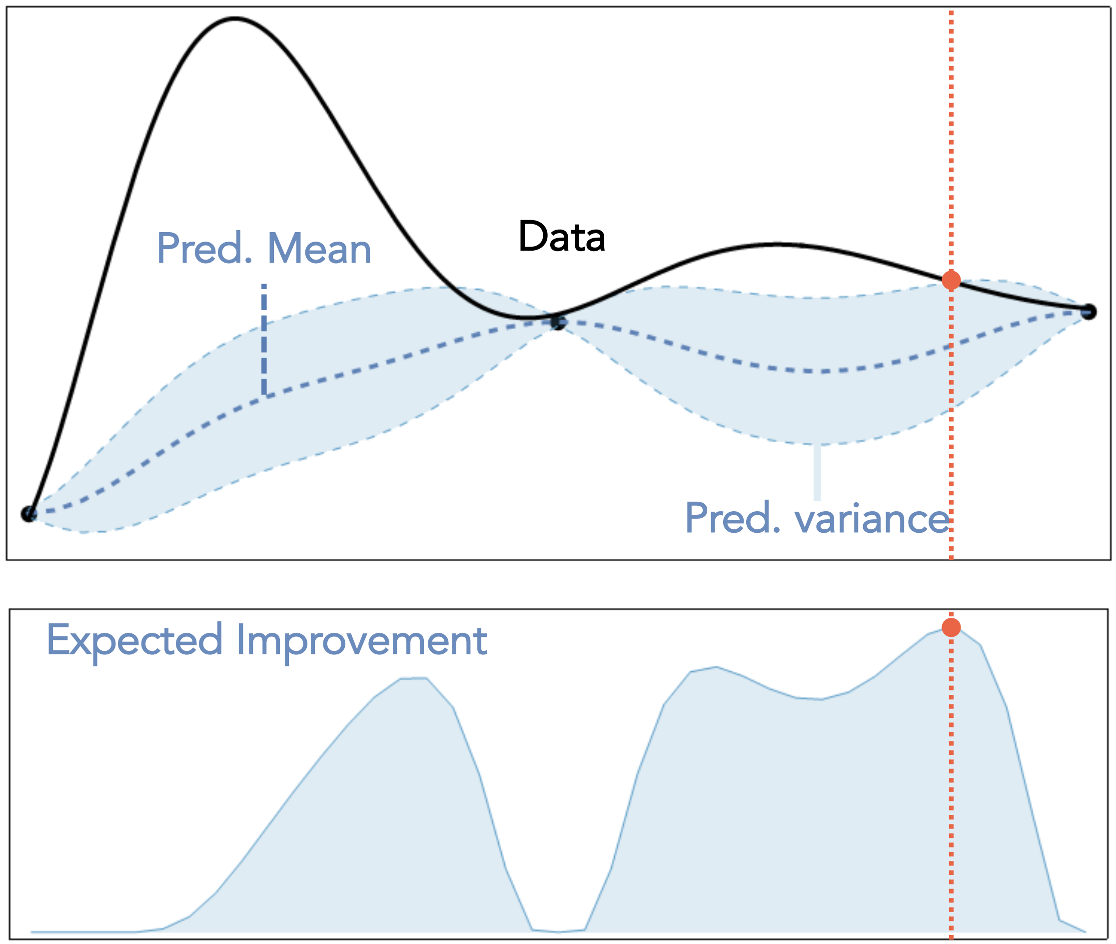
Example: Expected Improvement

Exploitation   Exploration

$$EI(x) = \begin{cases} (\mu(x) - f(x^{+}) - \delta)\Phi + \sigma(x)\varphi & if\ \sigma(x) > 0 \\ 0 & if\ \sigma(x) = 0 \end{cases}$$

Thus balancing the exploitation of information with the exploration of areas of high variance

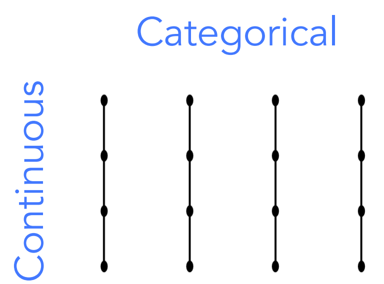
Acquisition functions: Decision making algorithms

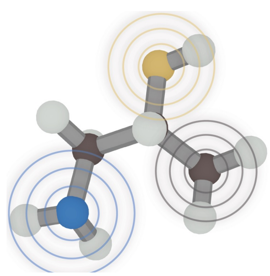


We choose to run the next experiment at an x value which maximizes EI

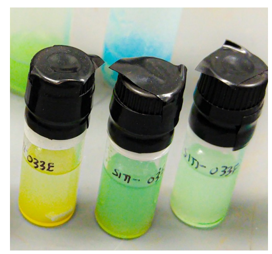
Reaction optimization using active learning: prior art

Applicable to diverse search spaces





Benchtop applications



arbitrary batch sizes

Statistical assessment

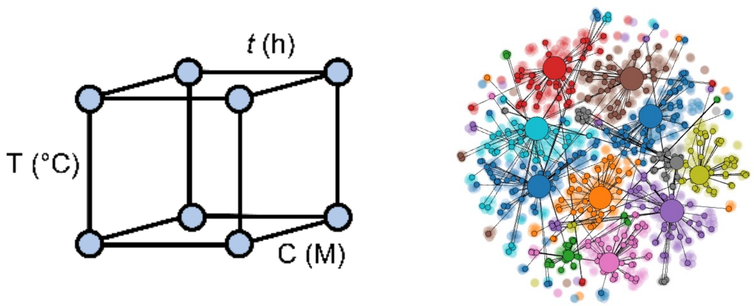
27	24	48	2	48	50	27	72	74	18	21	21	49	14	76
11	15	18	2	6	5	8	14	15	9	15	20	8	10	13
4	35	43	0	29	38	1	59	71	35	38	51	4	32	49
22	23	35	1	17	20	5	54	57	23	29	27	21	21	31
43	47	47	1	52	50	15	76	64	56	74	72	60	76	73
15	19	17	3	16	22	8	28	27	30	36	37	11	26	29
6	51	45	0	42	49	1	71	66	66	66	66	6	68	73
45	46	45	1	45	43	3	73	65	57	76	73	36	69	69
11	37	41	0	38	53	3	77	67	65	80	74	11	47	60
7	10	10	1	13	14	4	19	23	25	28	33	7	15	15
1	28	40	0	40	42	1	62	67	77	92	66	3	47	52
10	33	40	0	43	45	1	69	65	64	76	73	4	41	40
5	35	41	1	52	58	6	80	72	67	73	76	18	57	69
7	9	10	2	12	17	4	20	25	16	21	24	8	14	14

How does BO perform compared to chemists?
Does it find global max?

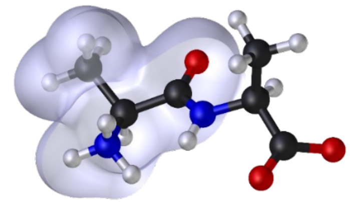
Baumgartner, L. M.; Coley, C. C.; Reizman, B. J.; Gao, K. W.; Jensen, K. F. *React. Chem. Eng.* **2018**, *3*, 303.
Häse, F.; Roch, L. M.; Aspuru-Guzik, A. *ArXiv200312127* (2020)
Reker, D.' Hoyt, E. A.; Bernardes, G. J. L.; Rodrigues, T. *Cell Rep. Phys. Sci.* **2020**, *1*, 100247

Bayesian optimization of chemical processes – Anatomy

Search Space

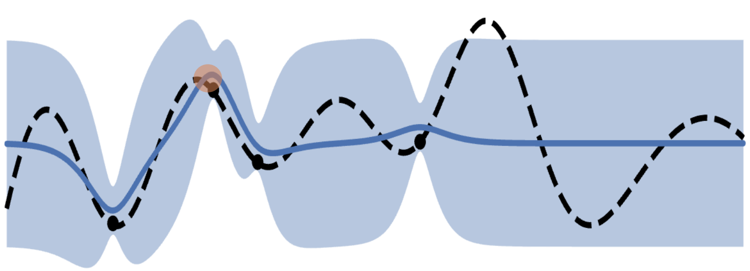


Reaction Encoding

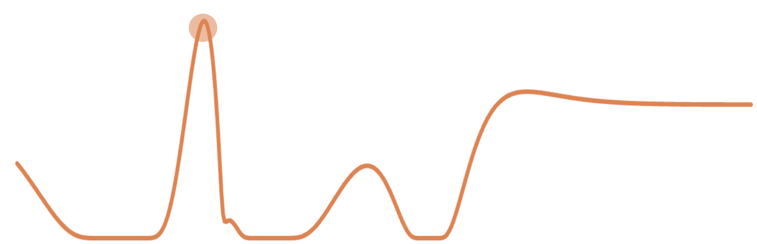


```
Boltzmann_weight
ligand_stoichiometry
ligand_number_of_atoms
ligand_charge
ligand_multiplicity
ligand_convergence_criteria
ligand_dipole
ligand_molar_mass
ligand_molar_volume
ligand_electronic_spatial_extent
ligand_homo_energy
ligand_lumo_energy
```

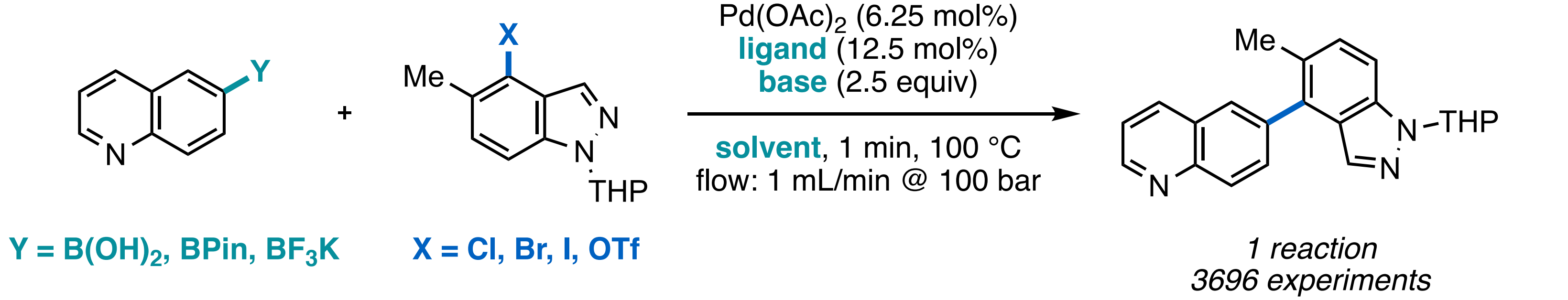
Response Surface Model

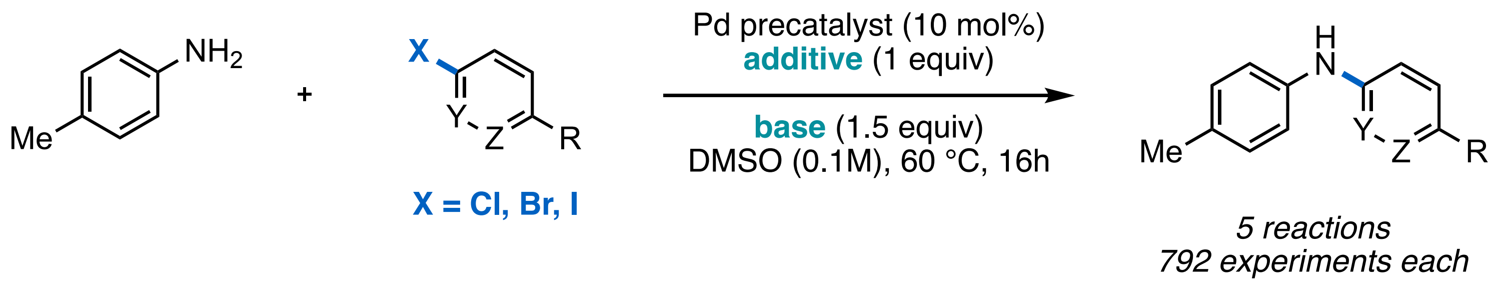


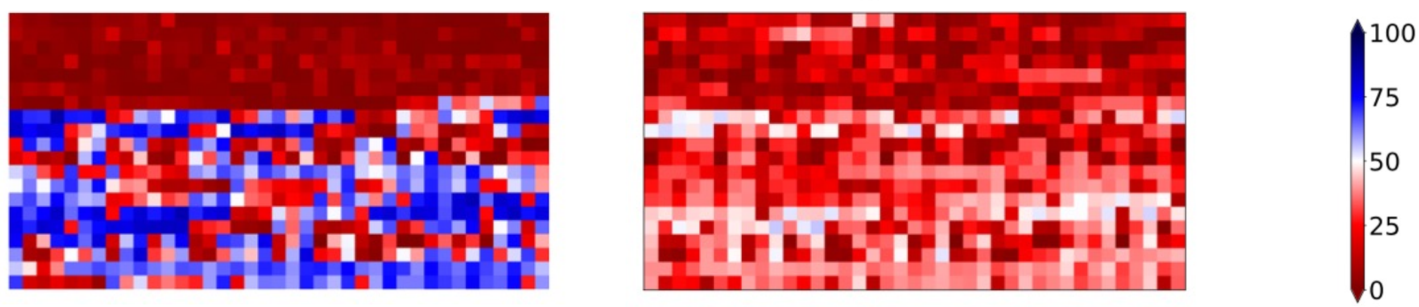
Acquisition Function



Bayesian optimization of chemical processes – Training





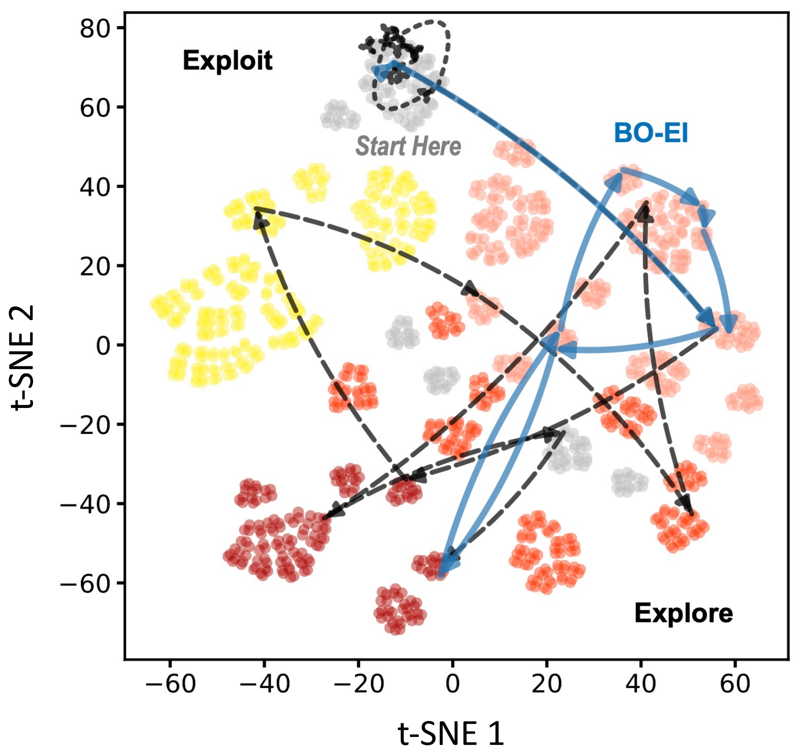


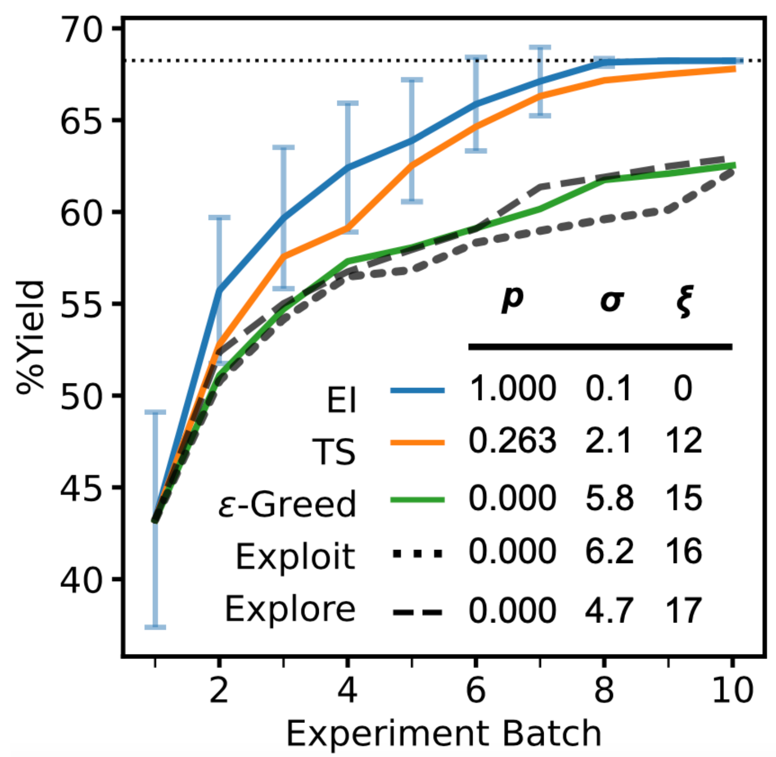
Perera, D.; Tucker, J. W.; Brahmbatt, S.; Helal, C. J.; Chong, A.; Farrell, W.; Richardson, P.; Sach, N. W. *Science*, **2018**, *359*, 429.

Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. *Science* **2018**, *360*, 186.

Bayesian optimization of chemical processes – Training

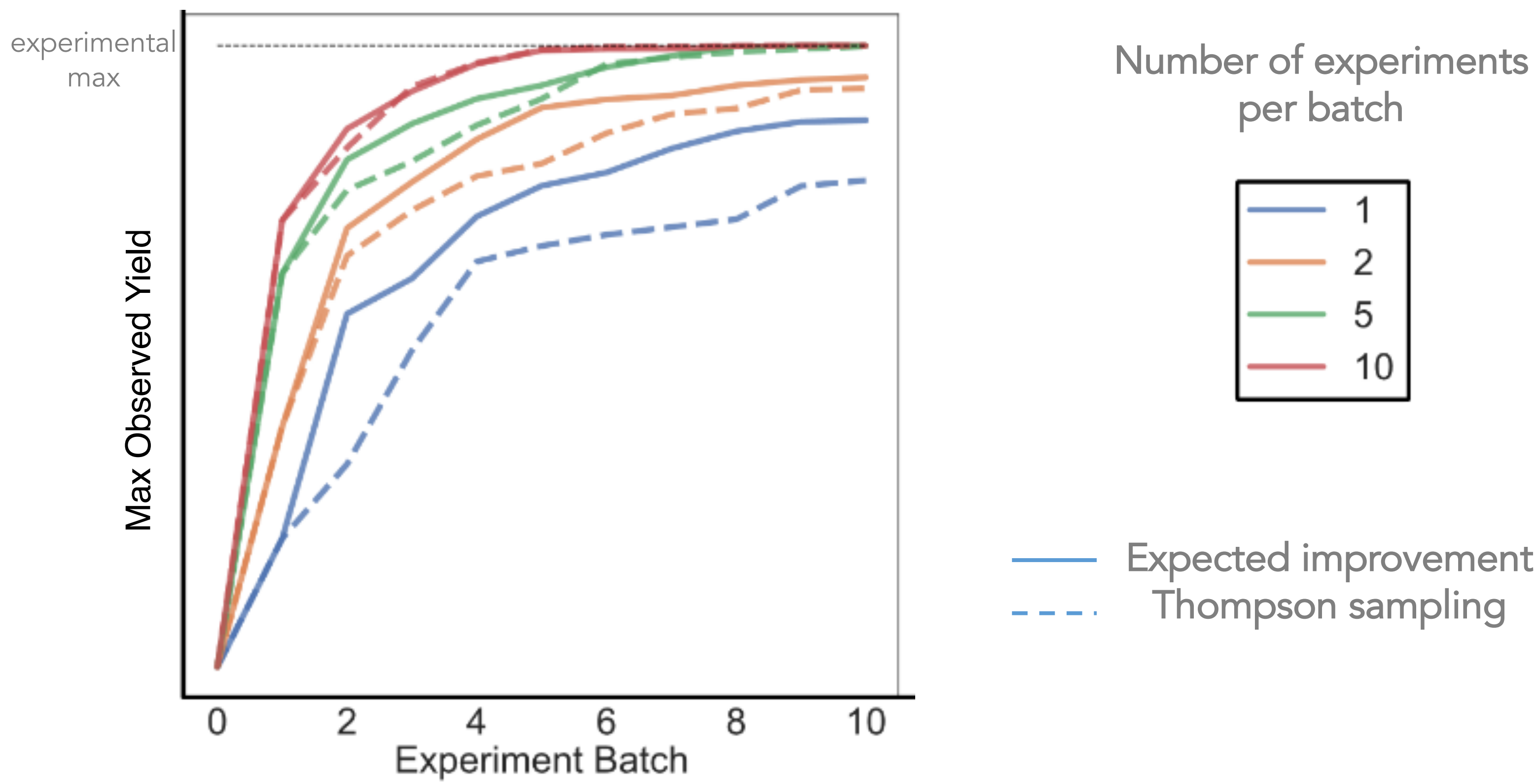
Featurization	Response surface model	Acquisition function	DoE
DFT (AutoQchem)	Gaussian process	Expected improvement	Generalized subset
Mordred	Random Forest	Thompson sampling	D-Optimal
One-hot	Bayesian Linear	e-greedy, exploit, explore	





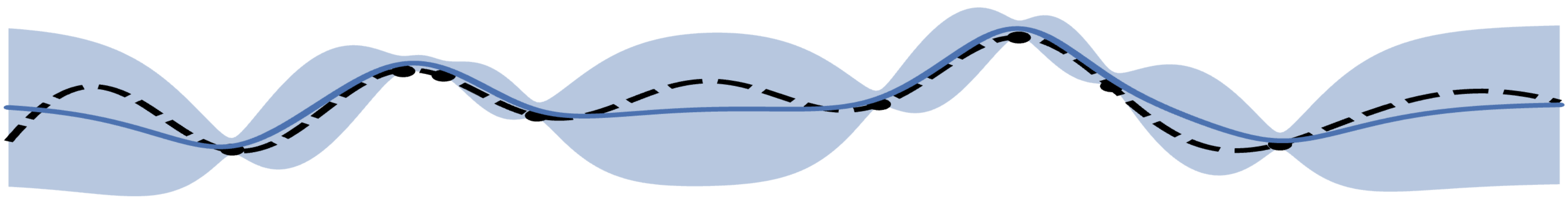
Bayesian optimization of chemical processes – Training

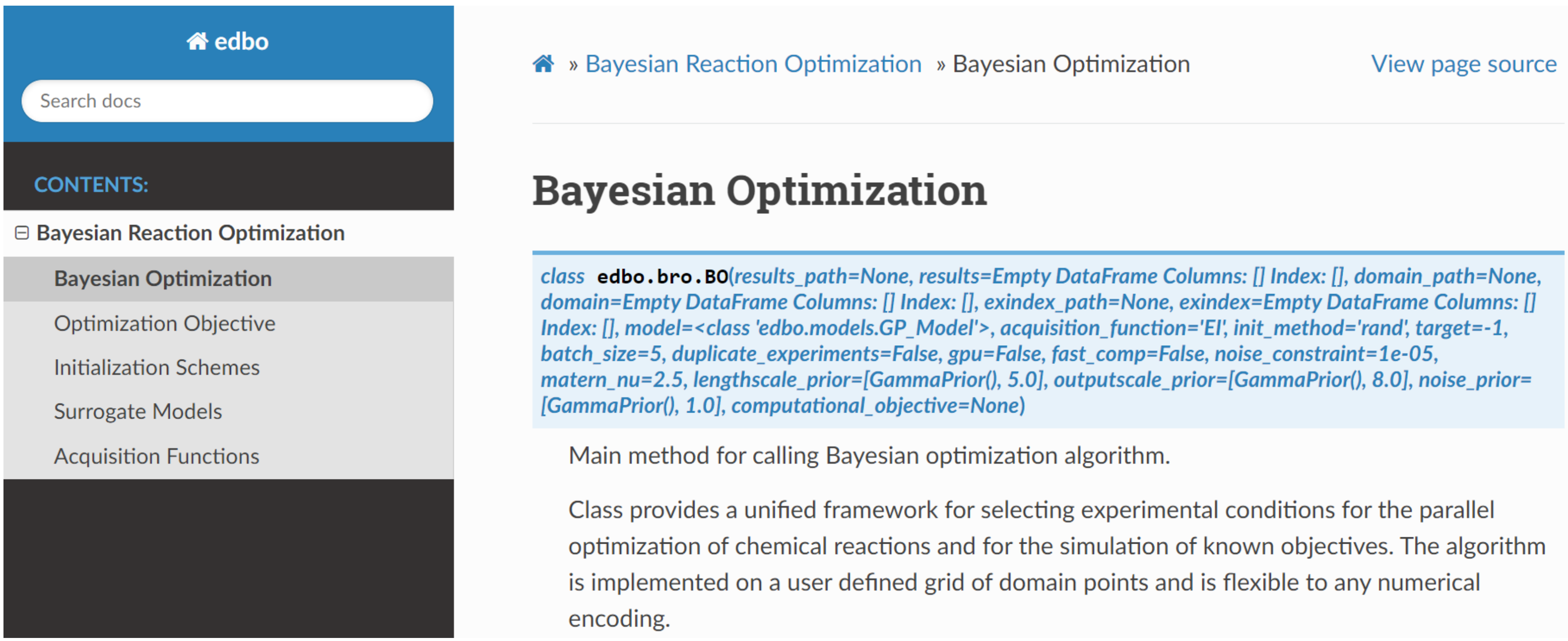
Kriging believer for parallel experimentation



Parallel optimization can save time, assuming it takes the same amount of time to execute 5 experiments as it does 1. However, parallel optimization consumes more resources (more experiments).

Bayesian reaction optimizer: Software





edbo

Search docs

CONTENTS:

- Bayesian Reaction Optimization
  - **Bayesian Optimization**
  - Optimization Objective
  - Initialization Schemes
  - Surrogate Models
  - Acquisition Functions

» Bayesian Reaction Optimization » Bayesian Optimization   View page source

Bayesian Optimization

class **edbo.bro.BO**(*results_path=None, results=Empty DataFrame Columns: [] Index: [], domain_path=None, domain=Empty DataFrame Columns: [] Index: [], exindex_path=None, exindex=Empty DataFrame Columns: [] Index: [], model=<class 'edbo.models.GP_Model'>, acquisition_function='EI', init_method='rand', target=-1, batch_size=5, duplicate_experiments=False, gpu=False, fast_comp=False, noise_constraint=1e-05, matern_nu=2.5, lengthscale_prior=[GammaPrior(), 5.0], outputscale_prior=[GammaPrior(), 8.0], noise_prior=[GammaPrior(), 1.0], computational_objective=None*)

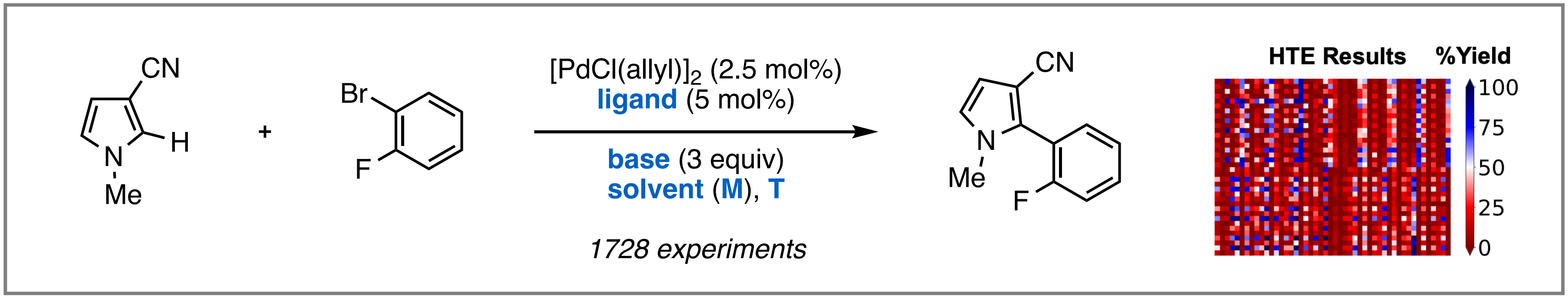
Main method for calling Bayesian optimization algorithm.

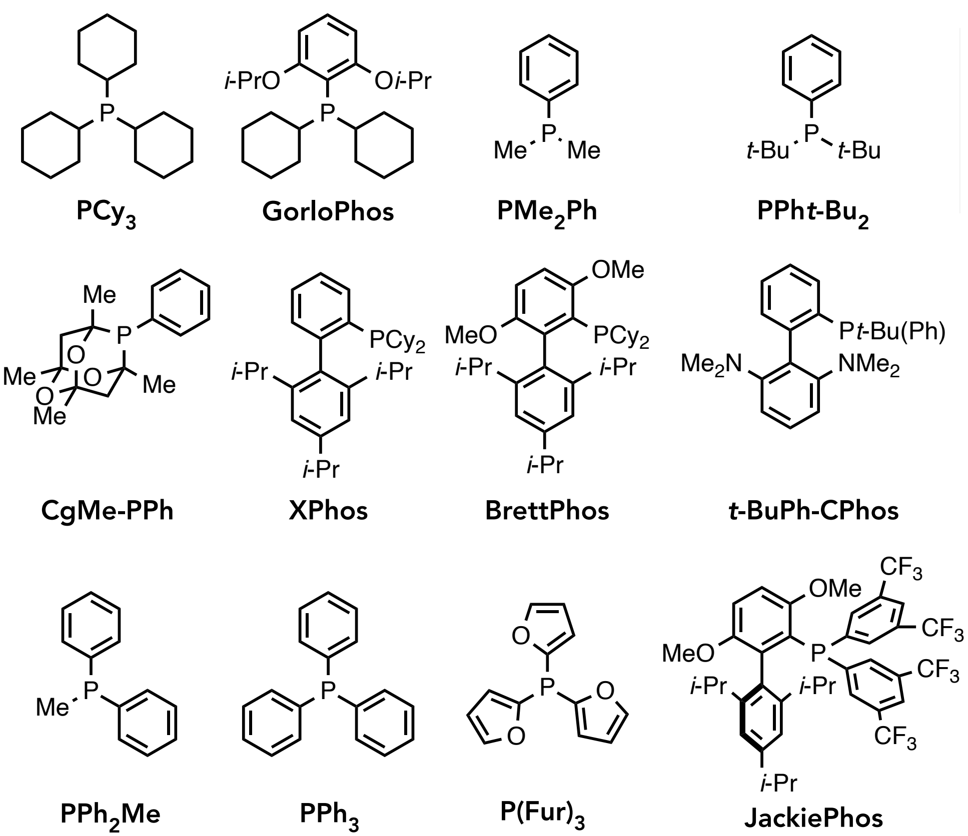
Class provides a unified framework for selecting experimental conditions for the parallel optimization of chemical reactions and for the simulation of known objectives. The algorithm is implemented on a user defined grid of domain points and is flexible to any numerical encoding.

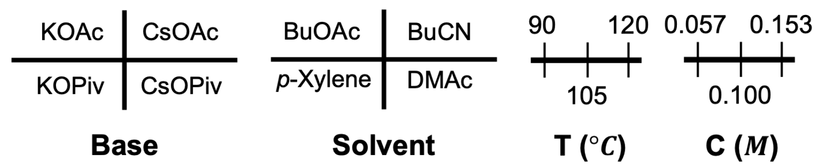
We developed a flexible python API to facilitate using BO at the bench.

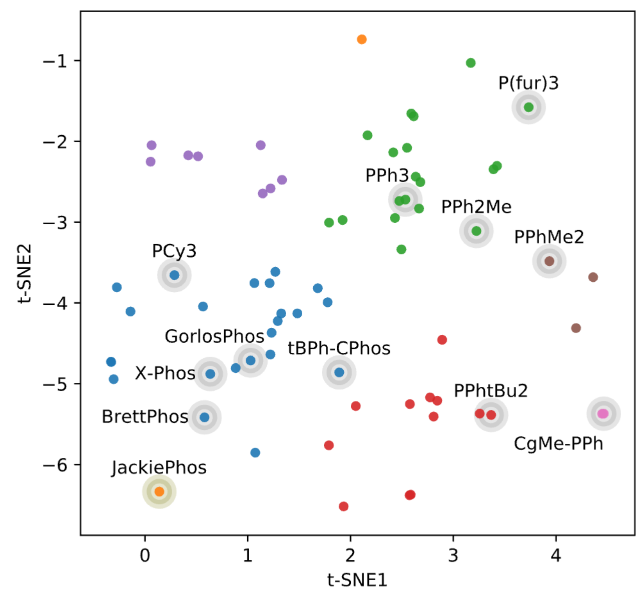
https://b-shields.github.io/edbo/index.html

Bayesian optimization of chemical processes – Test



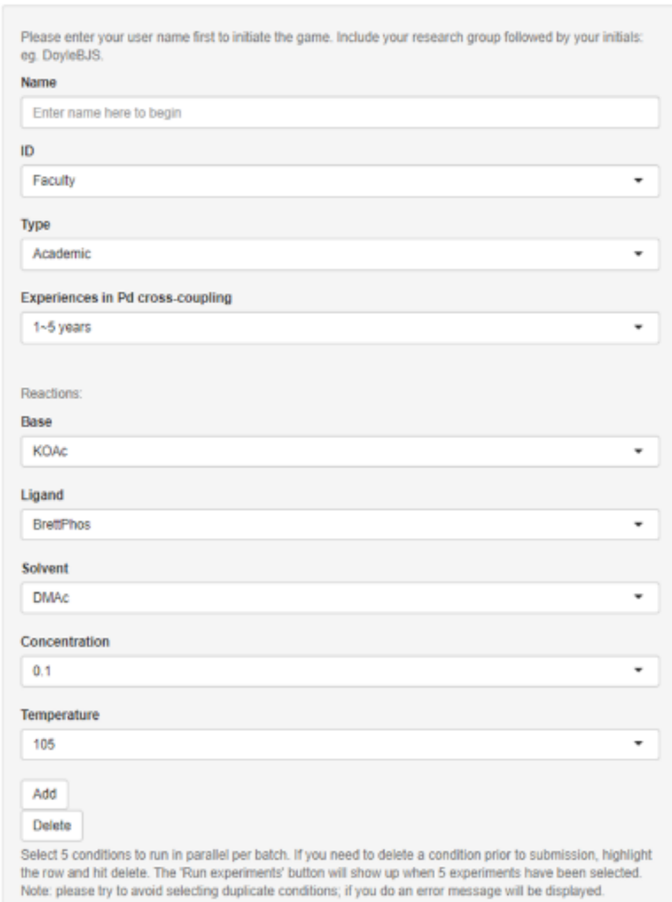


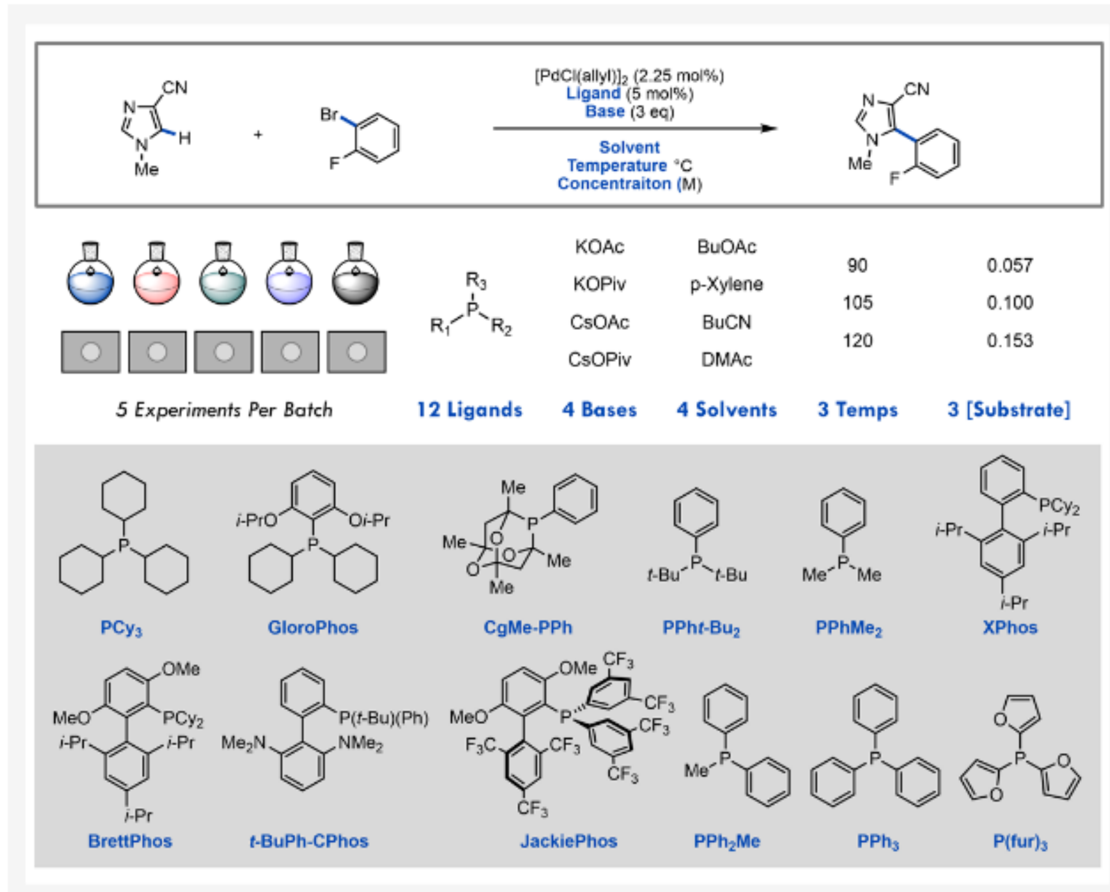




Benchmarking optimizer performance

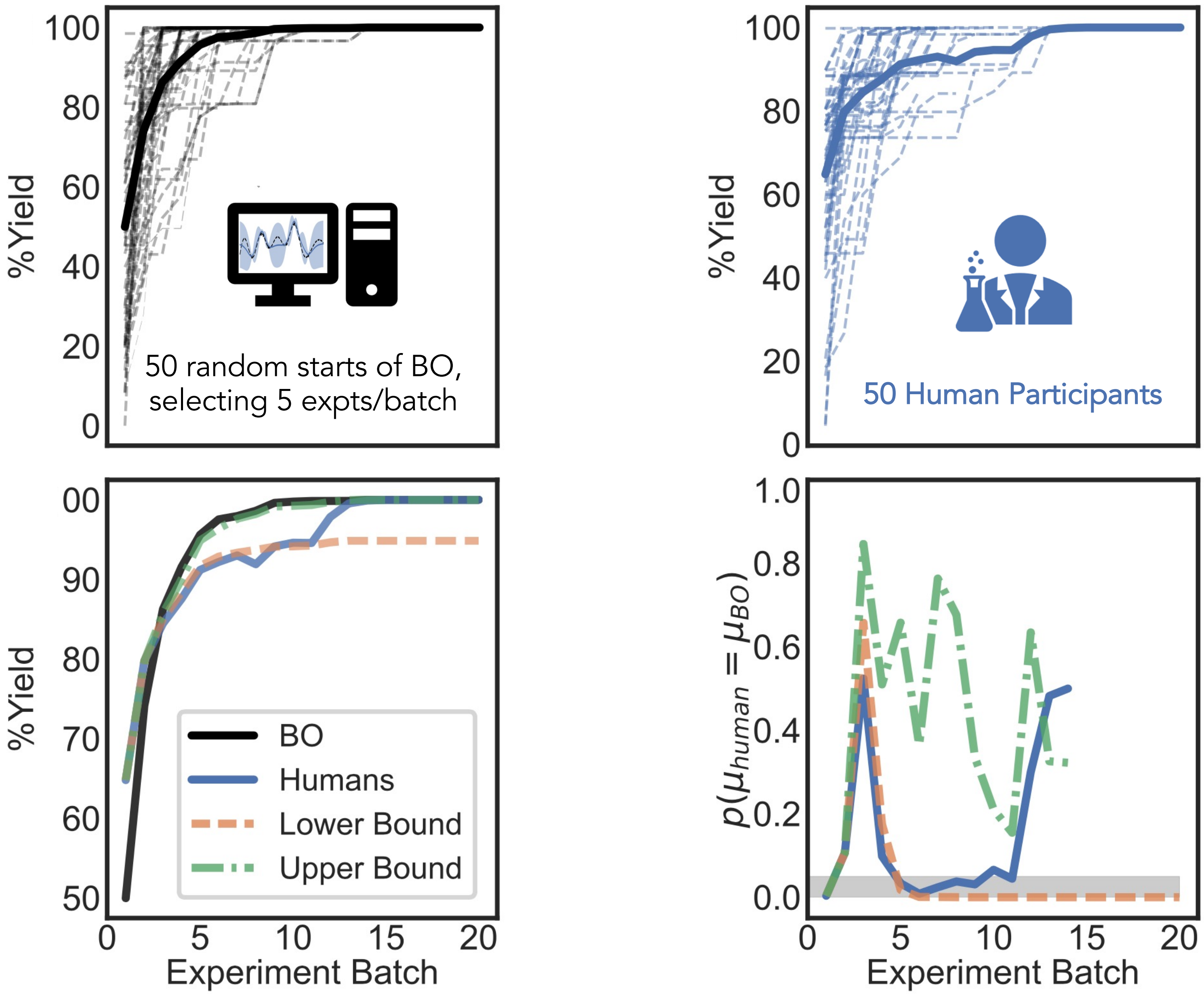
The real benchmark of success is against the decisions of human scientists when optimizing the same reaction



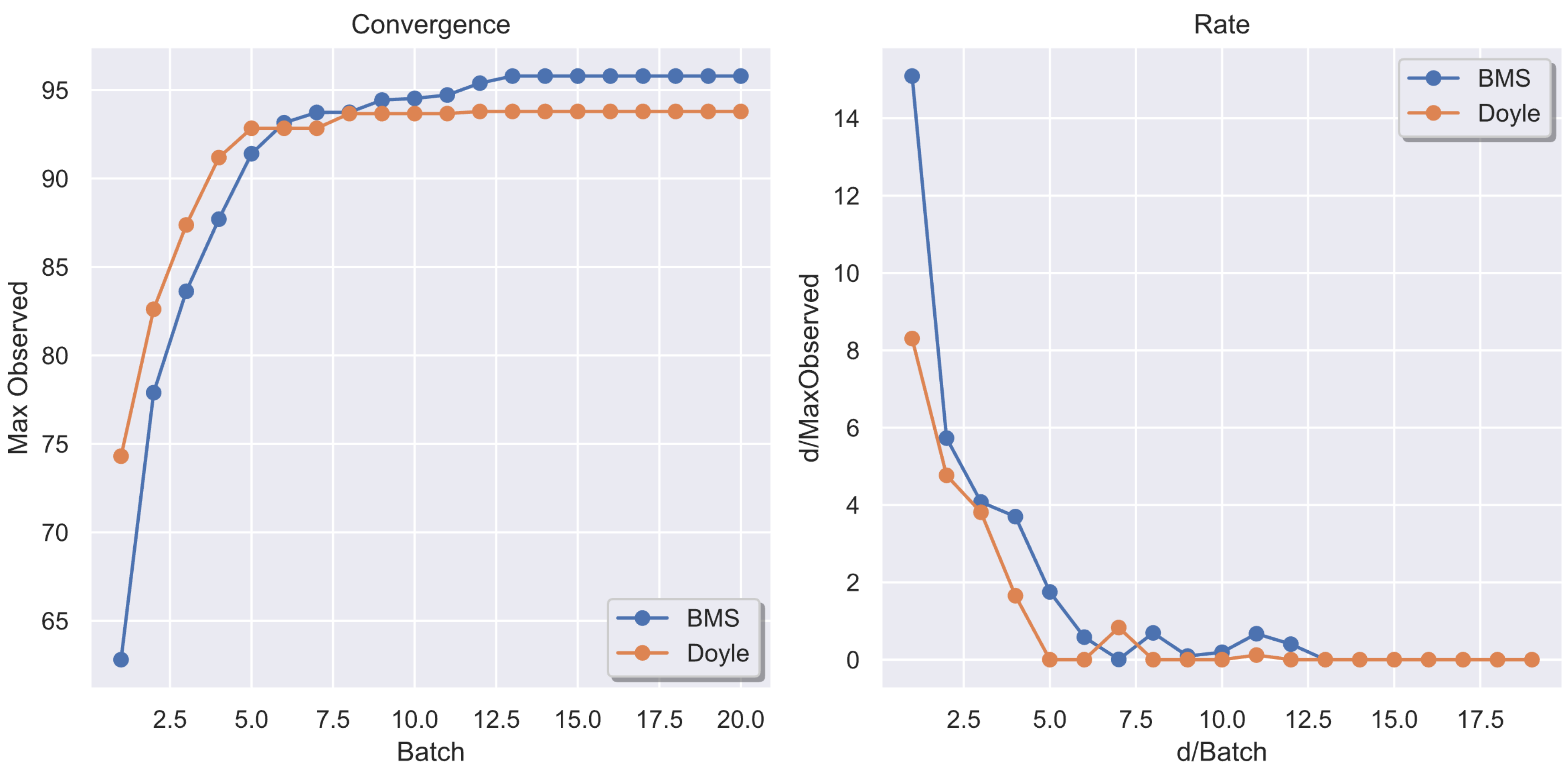


https://github.com/b-shields/EvML

Bayesian optimization of chemical processes – results

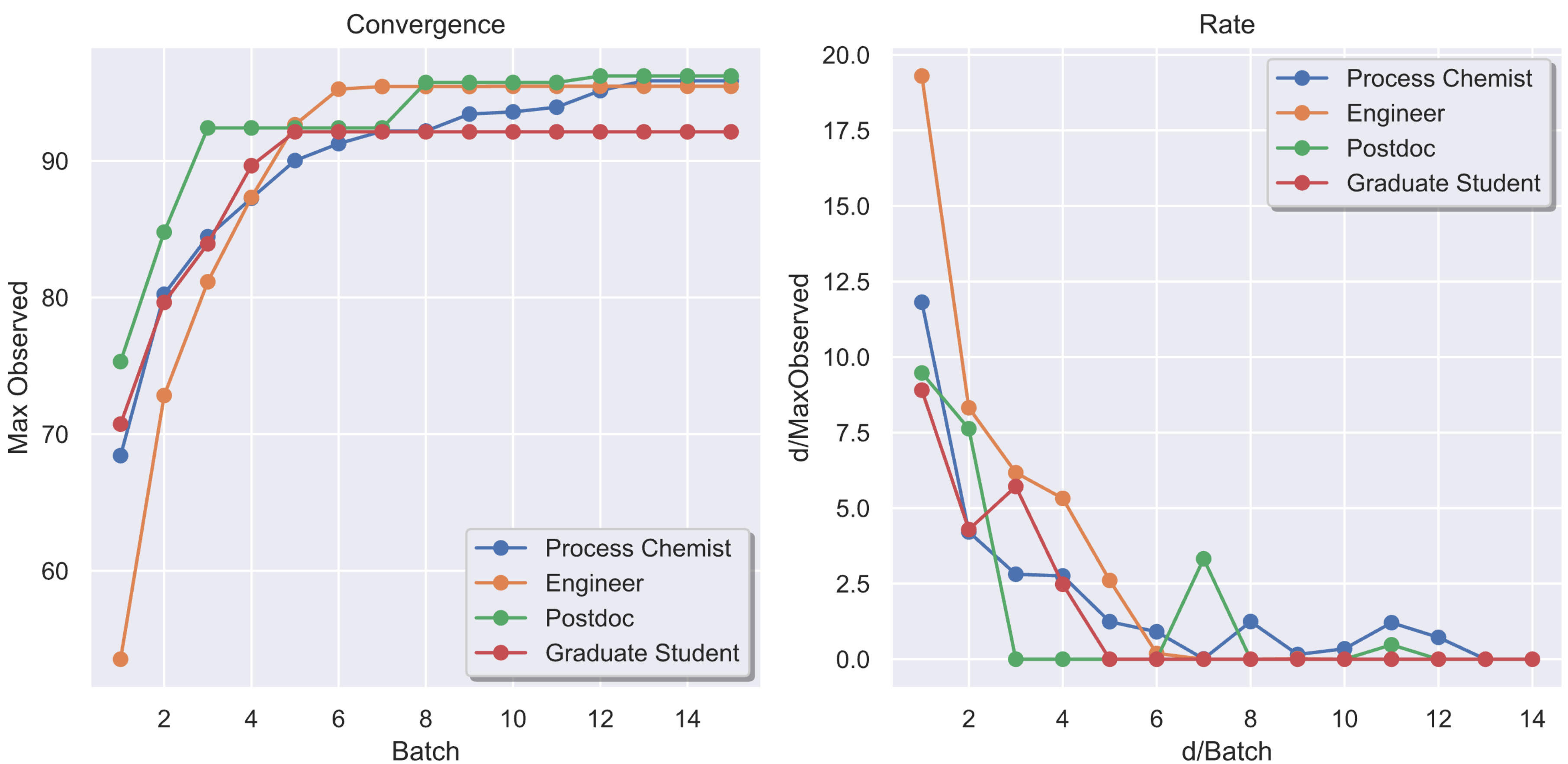


Industry versus academia



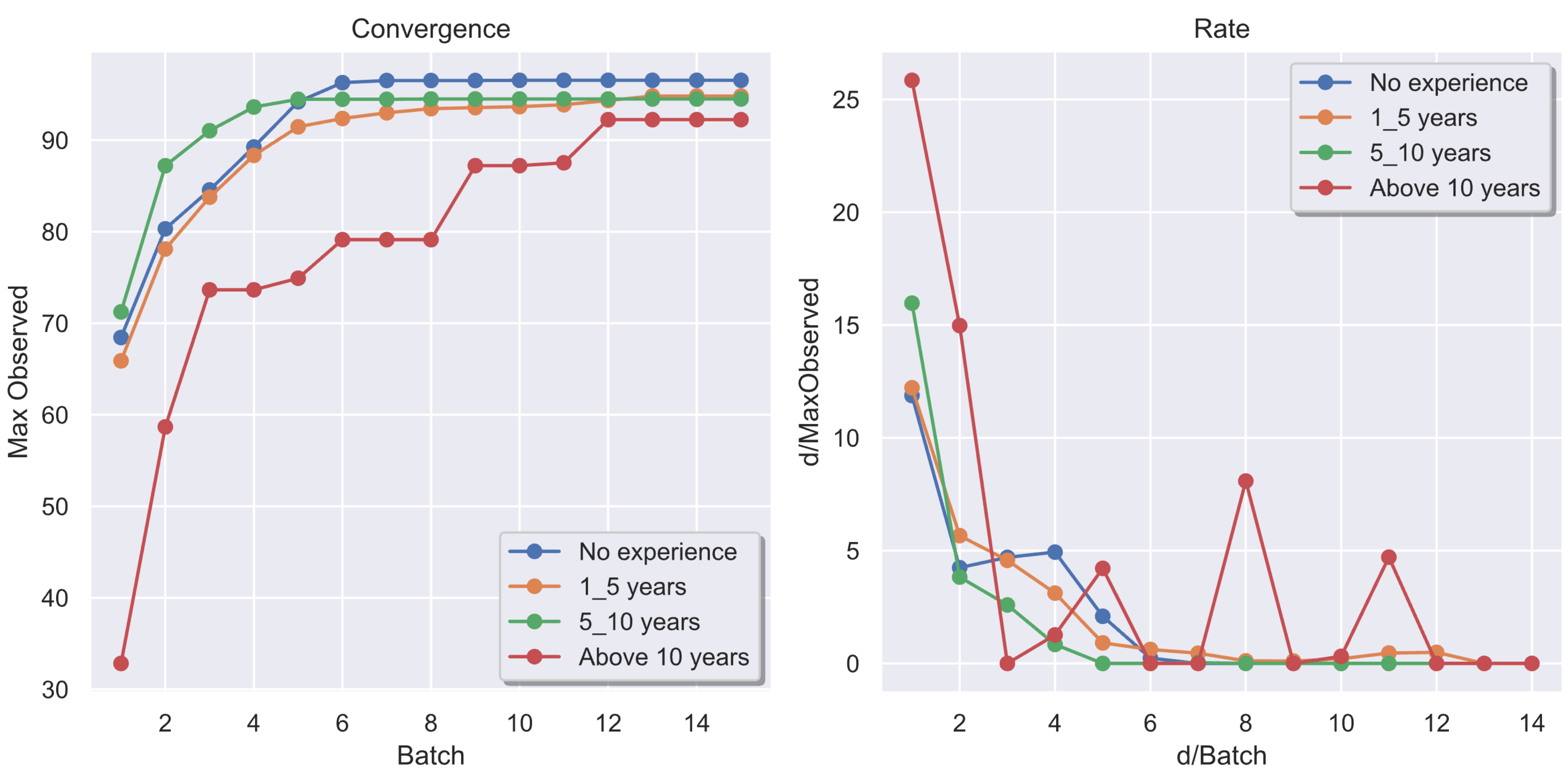
The Doyle lab made better initial decisions but was ultimately overtaken by BMS

Domain expertise



Chemists provide better initial choices but engineers rapidly regain the ground

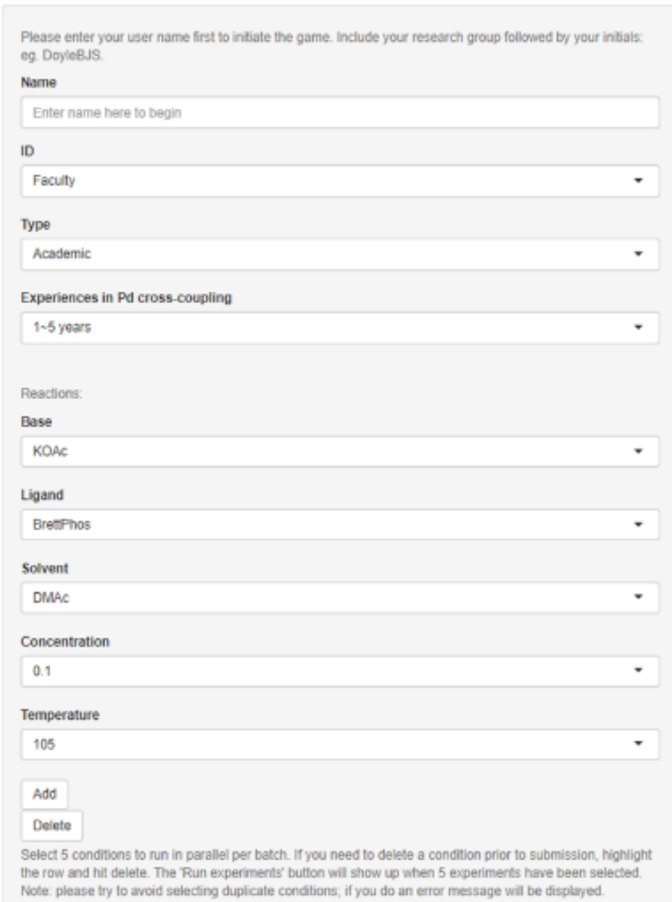
Importance of experience

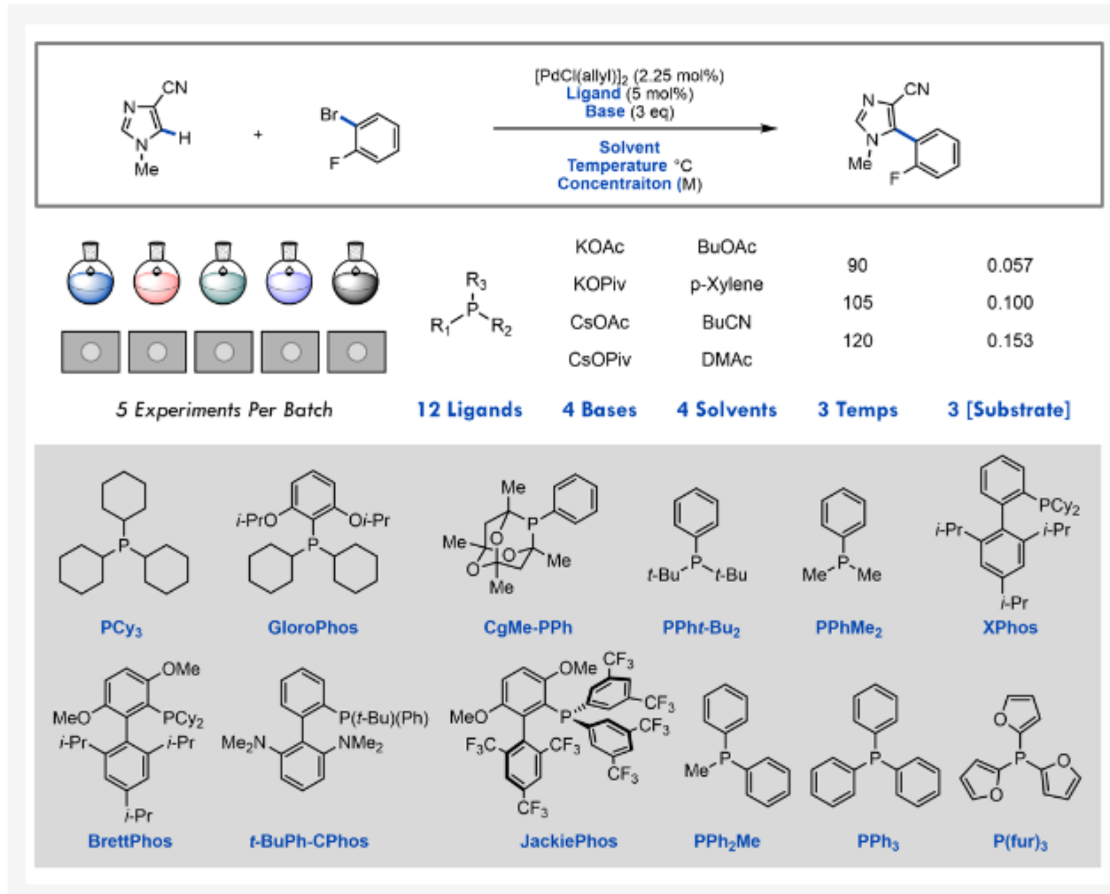


Interestingly, those with no experience in Pd-catalysis achieved the highest average yield in the long run

Benchmarking optimizer performance

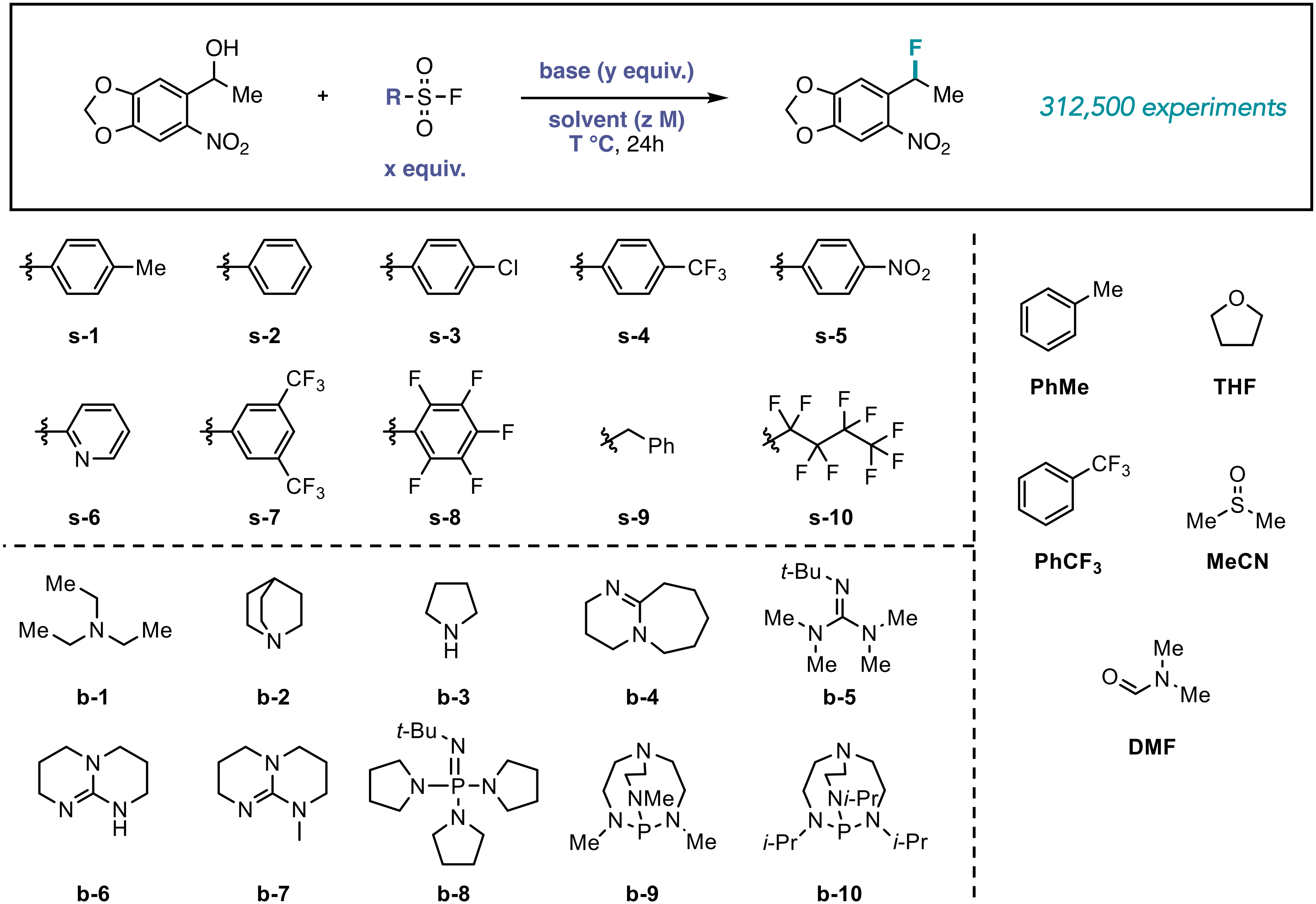
The real benchmark of success is against the decisions of human scientists when optimizing the same reaction



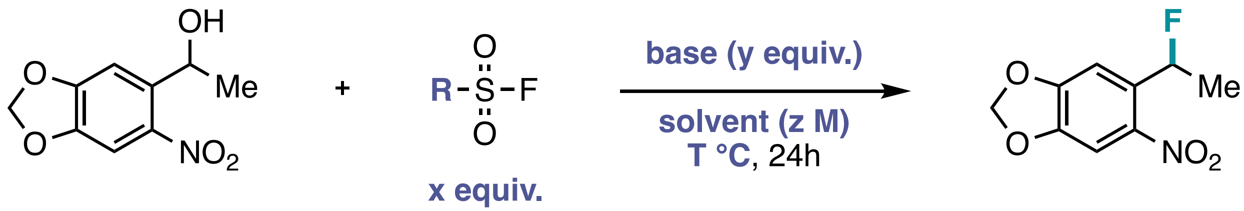


https://github.com/b-shields/EvML

Real world test case – Deoxyfluorination reaction



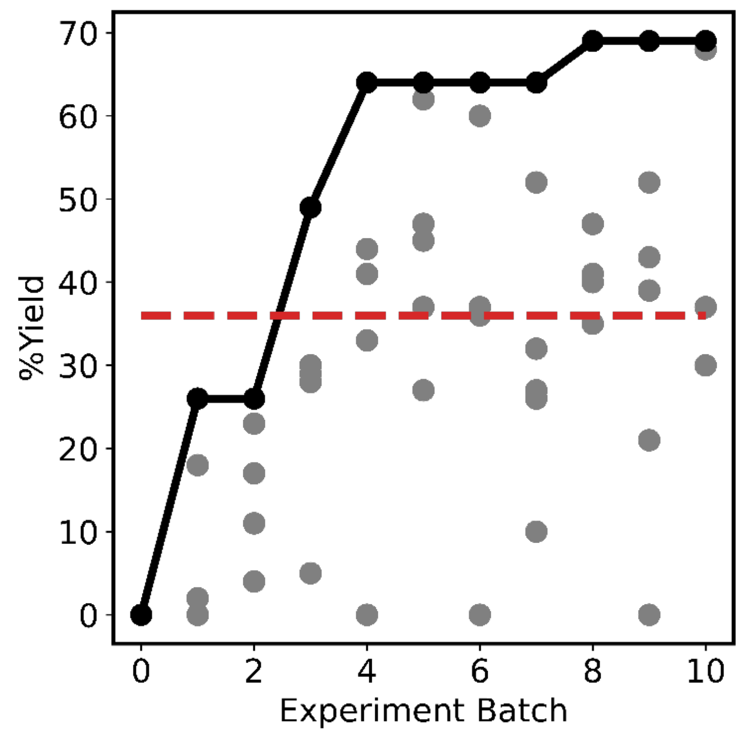
Real world test case – Deoxyfluorination reaction



Standard Conditions
PyFluor (1.1 equiv)
DBU (2 equiv)
PhMe (0.5M), rt, 24h
36% yield

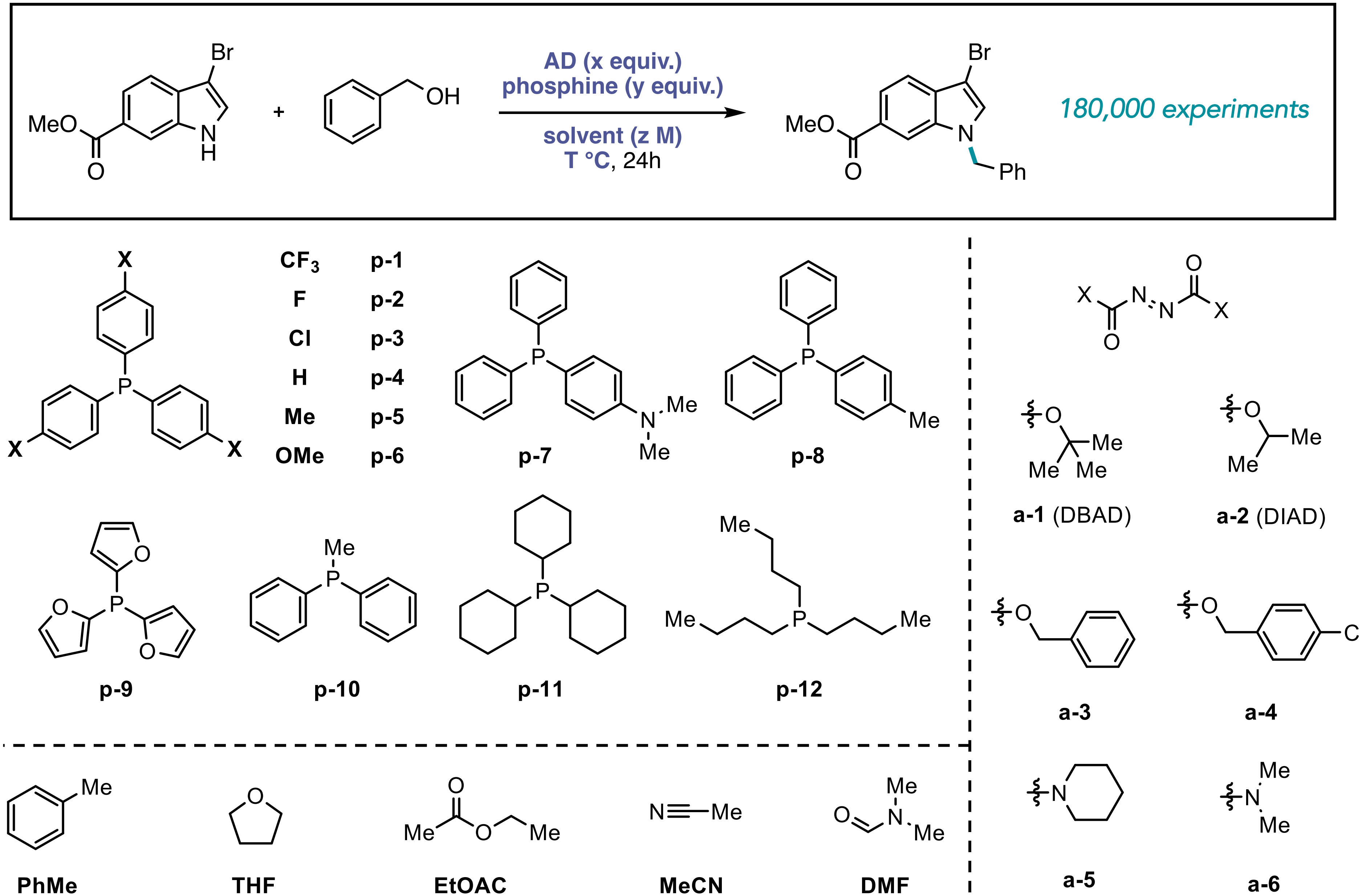
R	x (eq)	base	y (eq)	solvent	C (M)	T (°C)	yield
4-CF_3-Ph	1.9	BTMG	1.9	$PhCF_3$	0.5	20	**69% yield**
3,5,-dCF_3-Ph	1.1	BTMG	1.9	PhMe	0.1	60	**69% yield**
4-NO_2-Ph	1.9	BTMG	1.9	PhMe	0.5	20	**68% yield**

Within 3 batches of 5 experiments (out of 312,500 experiments) BO identified improved parameters over standard conditions.

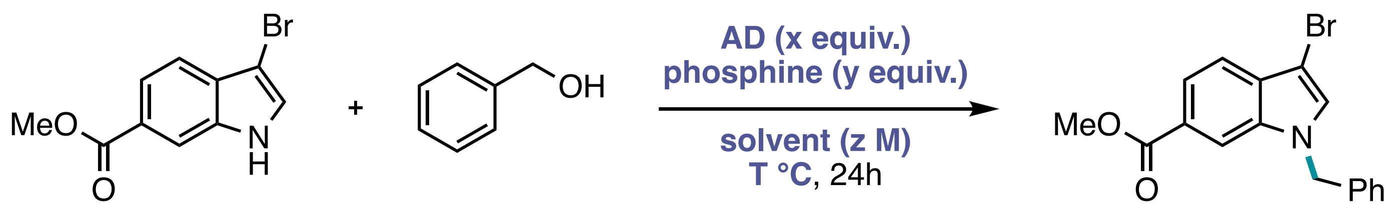


DFT encoding, GP model, EI, batch size = 5

Real world test case – Mitsunobu reaction

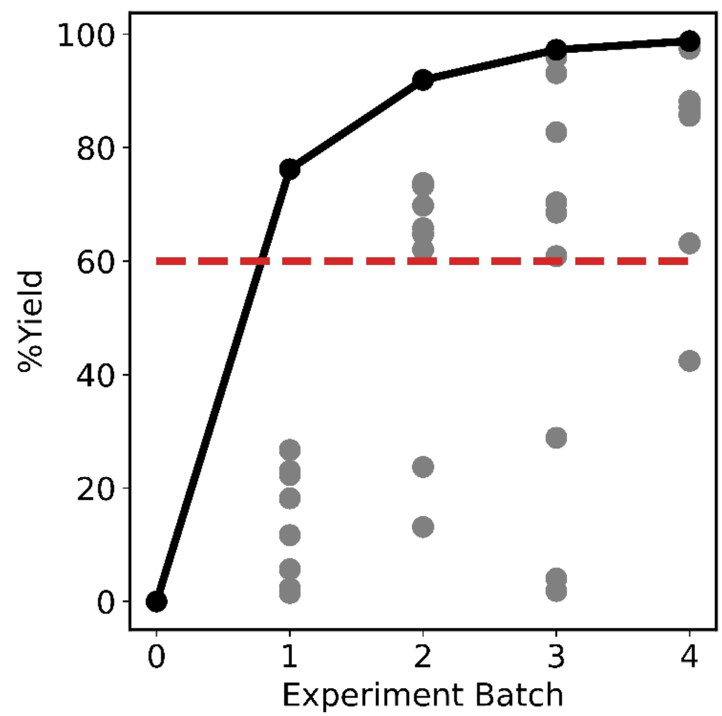


Real world test case – Mitsunobu reaction



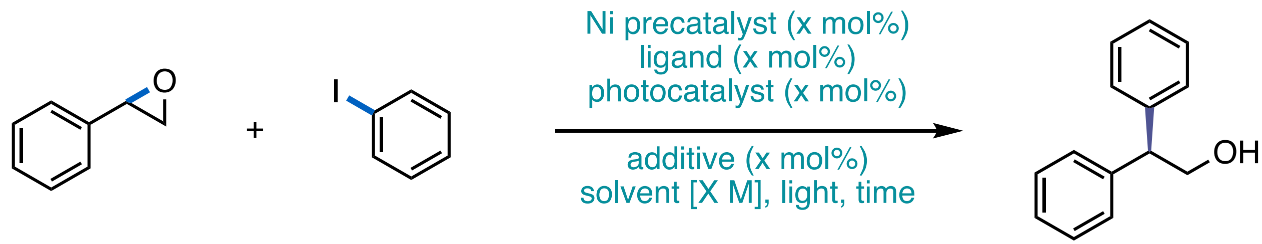
AD	x (eq)	phosphine	y (eq)	solvent	C (M)	T (°C)	yield
a-5	1.9	p-10	1.9	THF	0.2	25	**99% yield**
a-6	1.9	p-10	1.9	MeCN	0.15	45	**99% yield**
a-5	1.9	p-10	1.9	THF	0.15	45	**99% yield**

Within 4 batches of 10 experiments BO identified parameters which give quantitative yield.

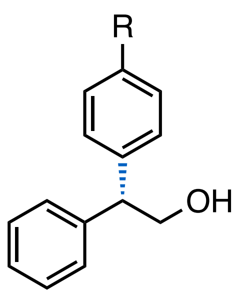


DFT encoding, GP model, EI, batch size = 10

Application to “real” optimization problems

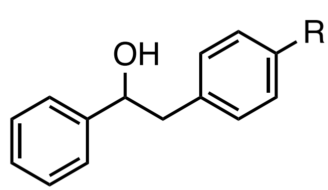


stereoselectivity



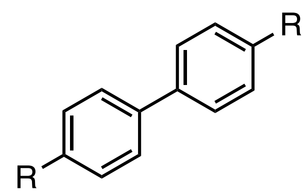
R enantiomer

regioselectivity

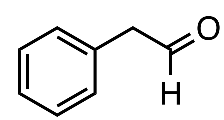


linear product

chemoselectivity



homocoupling

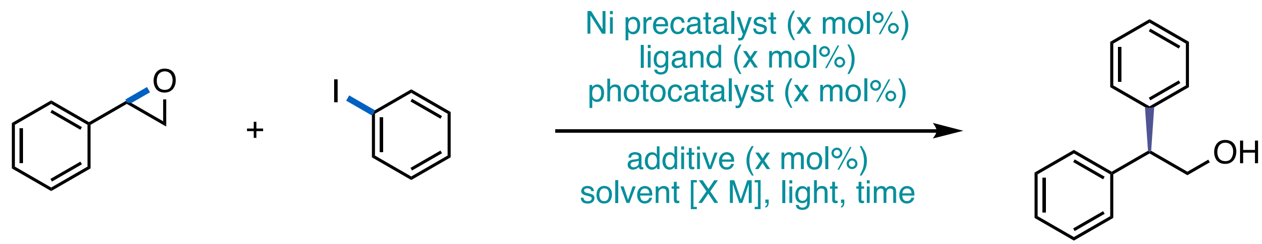


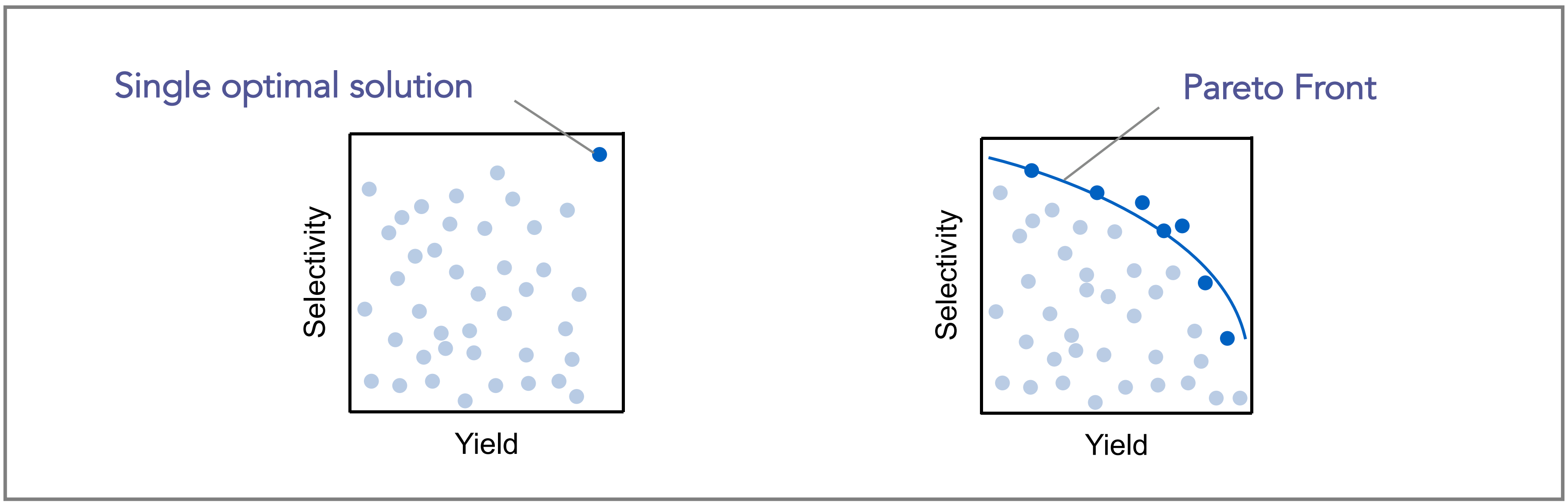
β-hydride elimination

Optimization problems in chemistry are typically multi-objective.

Will Lau, Jose Garrido Torres

Multi-objective optimization





Optimization problems in chemistry are typically multi-objective.

Will Lau, Jose Garrido Torres

EDBO+: the team

Jose Garrido Torres

Pranay Anchuri

Ryan Adams

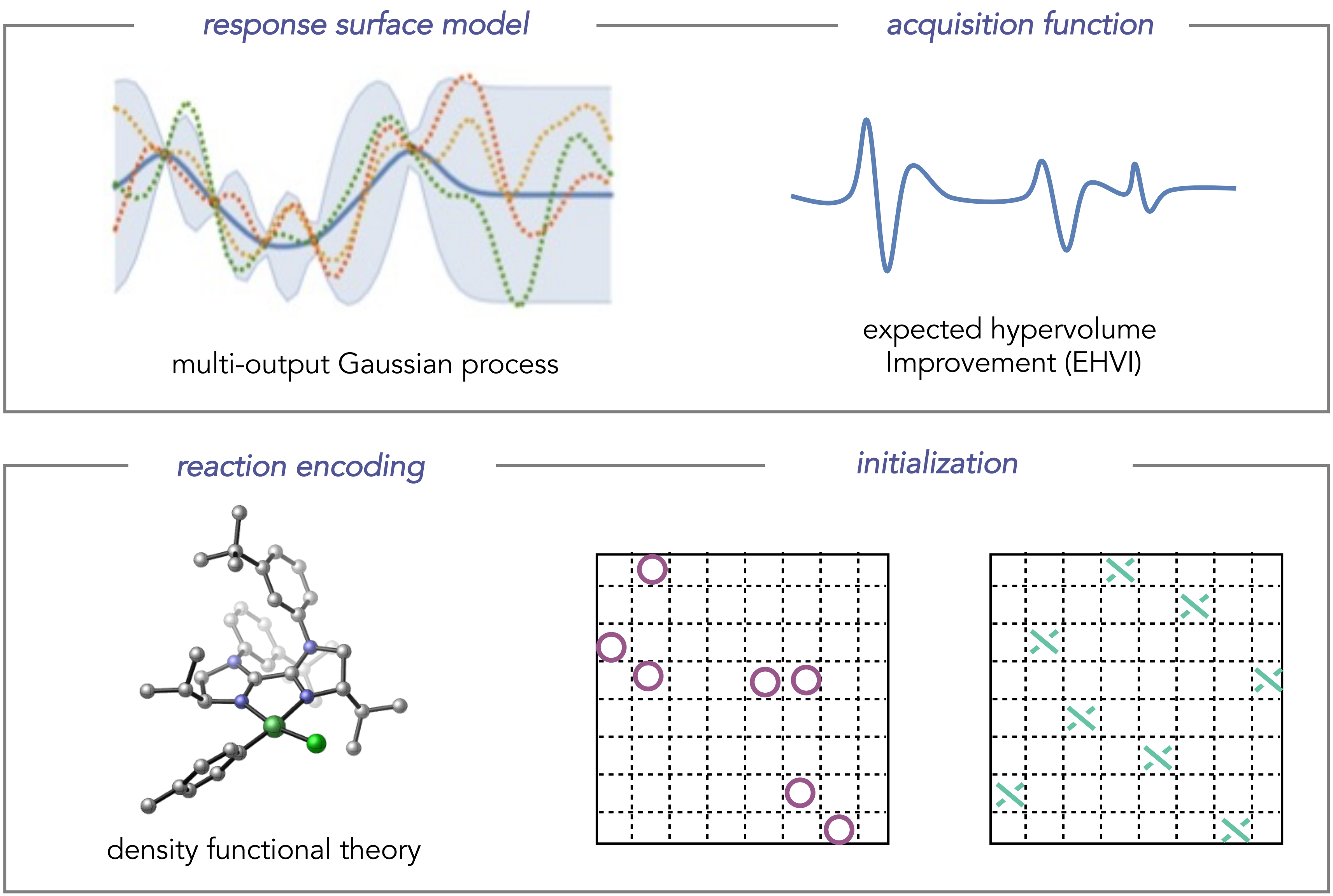
Jun Li Jason Stevens Jacob Janey Jose Tabora Alina Borovika Bristol Myers Squibb™

C-CAS Center for Computer Assisted Synthesis

Princeton Catalysis Initiative

NSF

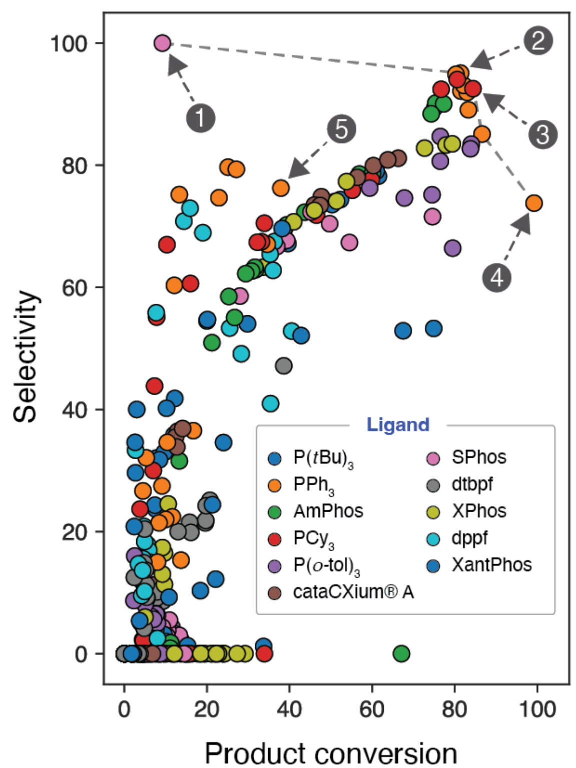
Algorithm optimization



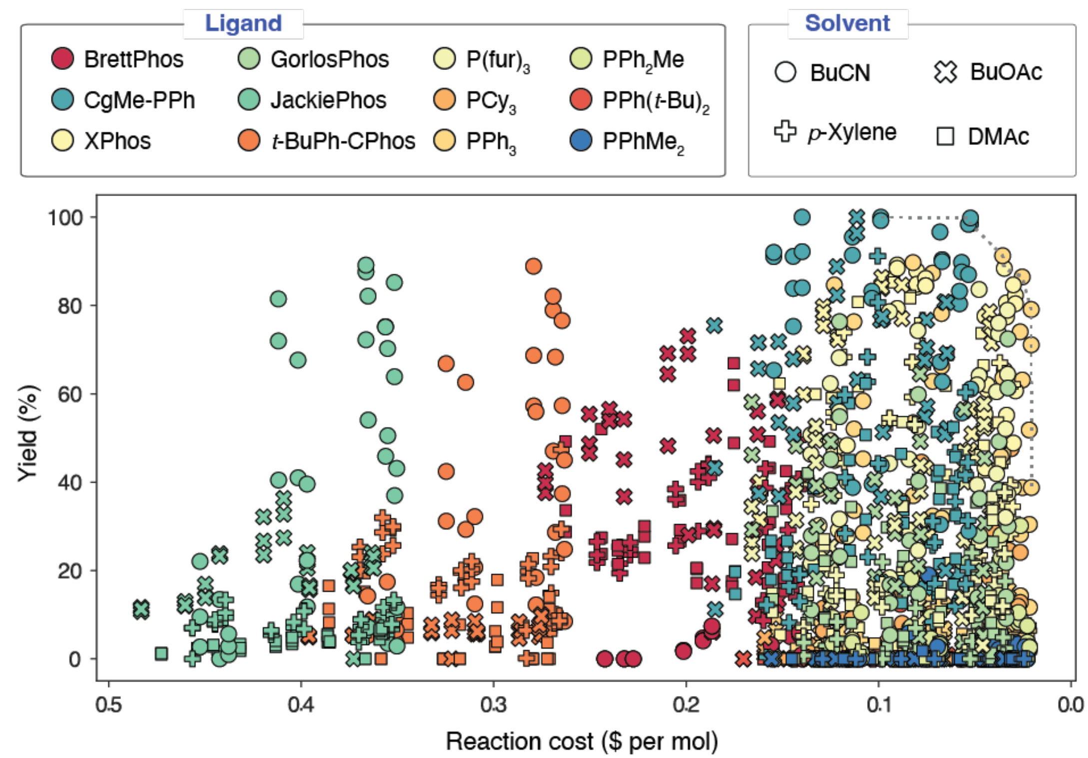
Will Lau, Jose Garrido Torres

Multi-objective optimization of chemical processes – Training

Pfizer Suzuki dataset

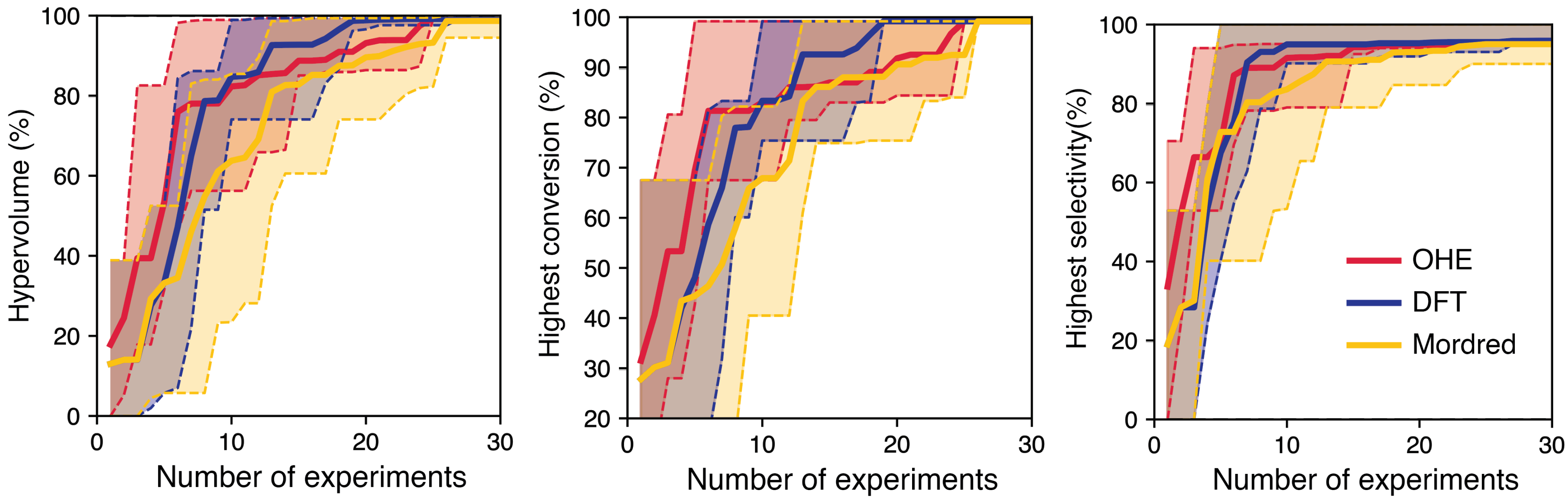


Pd-catalyzed C(sp²)–H arylation



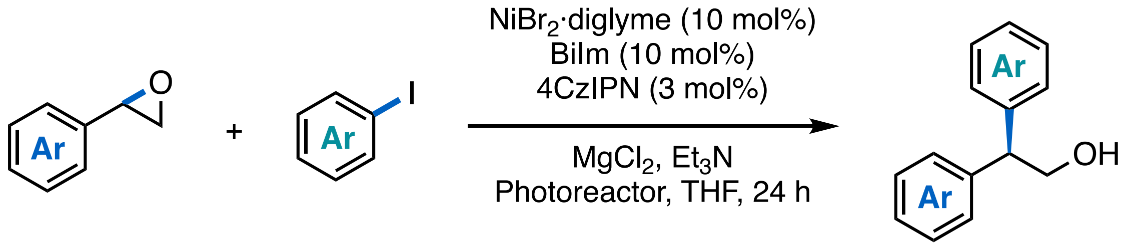
Richardson, P. & Sach, N. W. *unpublished*.
Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Martinez Alvarado, J. I.; Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021**, *590*, 89-96.

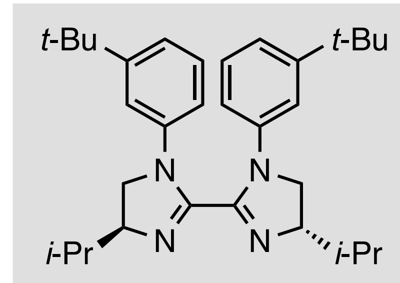
EDBO+: Evaluation of reaction encoding



10 simulations using a Gaussian process surrogate model and q-Expected HyperVolume Improvement (q-EHVI) as the acquisition function

EDBO+: Enantioselective arylation of epoxides





Will Lau

Mia Borden

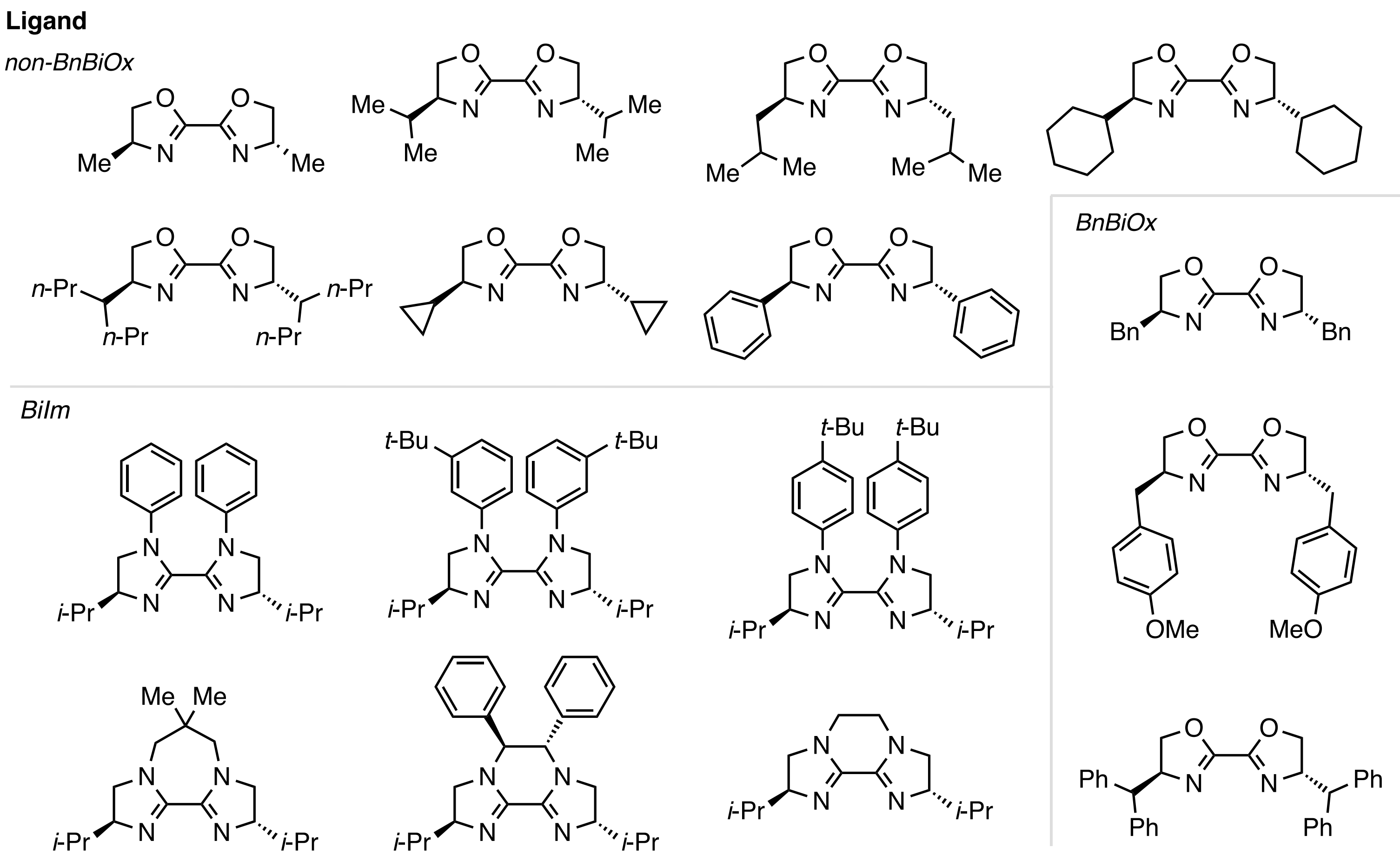
Talia Steiman

Lucy Wang (UG)

~500 experiments to optimize this enantioselective reaction to 91% ee and 70% yield (model reaction), primarily screening one variable at a time.

J. Am. Chem. Soc. **2021**, *143*, 15873

EDBO+: Enantioselective arylation of epoxides



Ni precatalyst
$NiBr_2$diglyme,
$NiCl_2$glyme,
$Ni(cod)_2$

Additive
none
$MgCl_2$

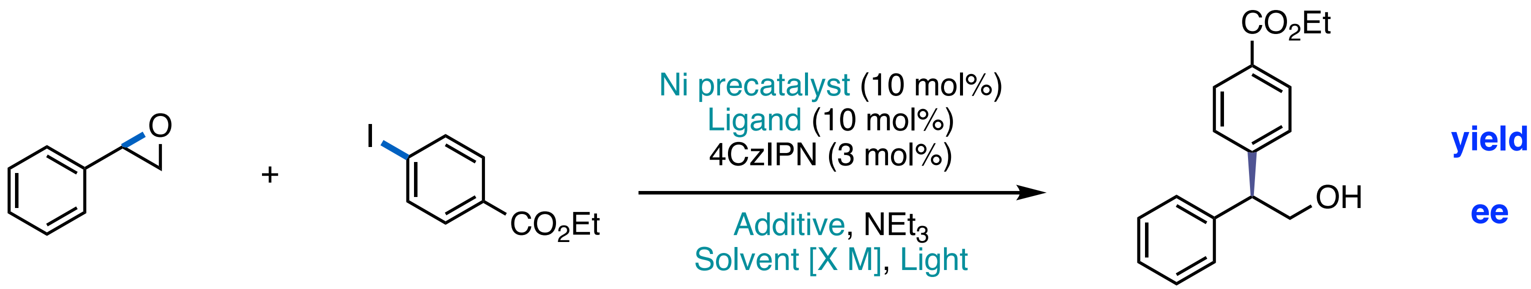
Solvent
THF
dioxane
MeCN

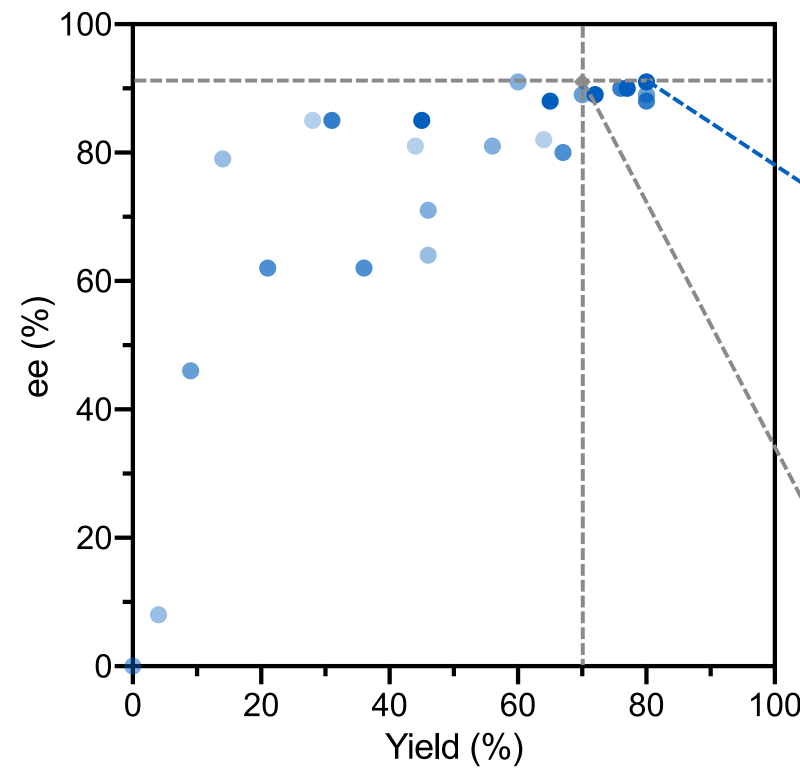
Light source
BlueLEDs
Photoreactor (PR)

Concentration (M)
0.033 | 0.05 | 0.1

The search space included categorical and continuous variables and spanned a fully combinatorial space of ~60,000 potential.

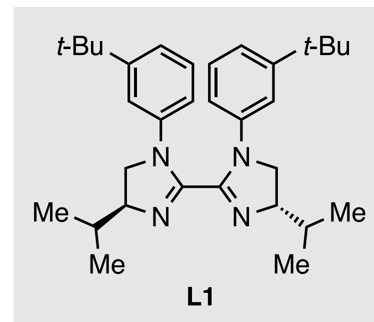
Multi-objective Bayesian optimization in the same search space





Round 8:
80% yield
91% ee

$NiBr_2$diglyme,
L1, $MgCl_2$,
MeTHF,
0.033M, PR

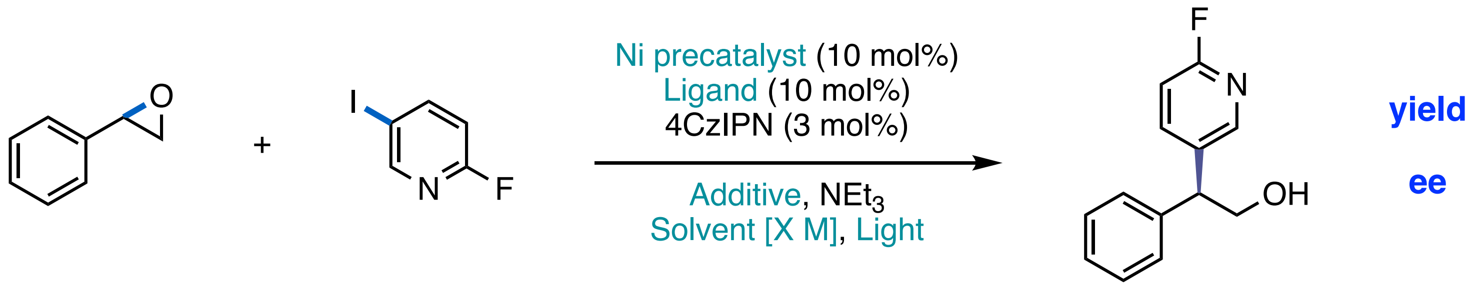


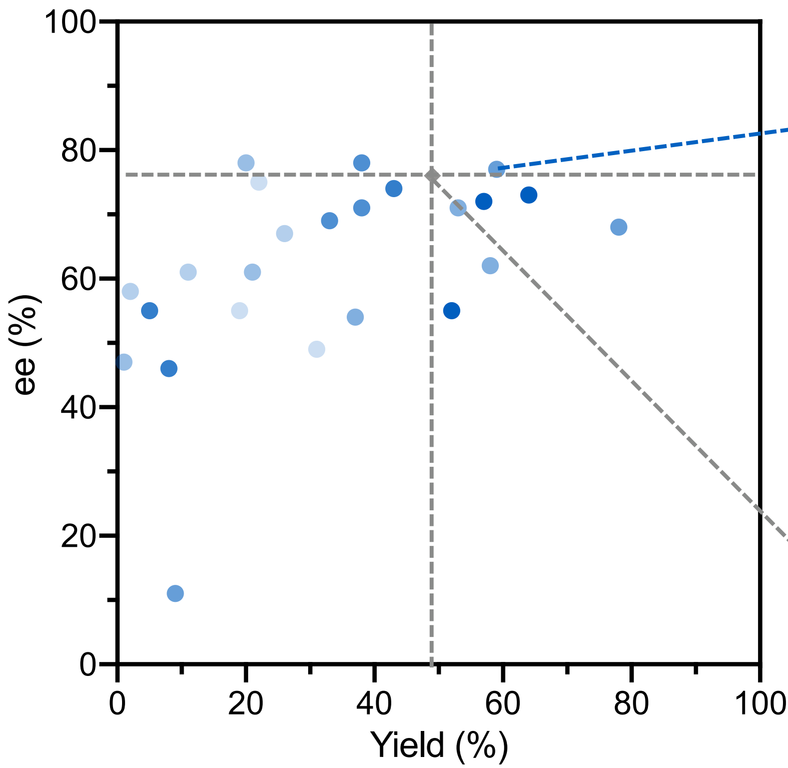
Optimized manually:
70% yield
91% ee

Ni(cod)$_2$,
L1, $MgCl_2$,
THF,
0.05M, PR

- *Batch size = 3*
- *DFT featurization of ligands*
- *Total # of experiments = 24*

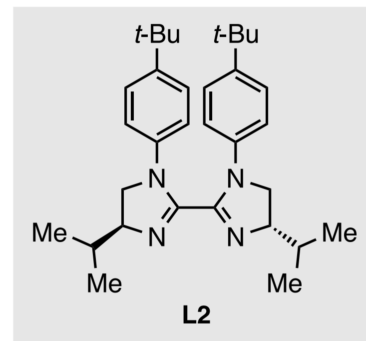
Optimization with a challenging substrate





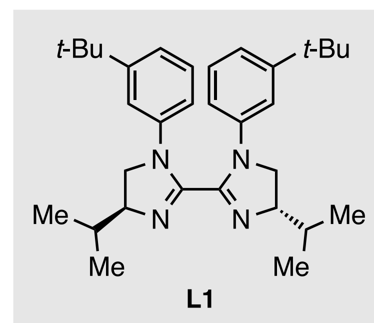
Round 5:
59% yield
77% ee

$NiCl_2$glyme,
L2, $MgCl_2$,
dioxane,
0.1M, BLEDs



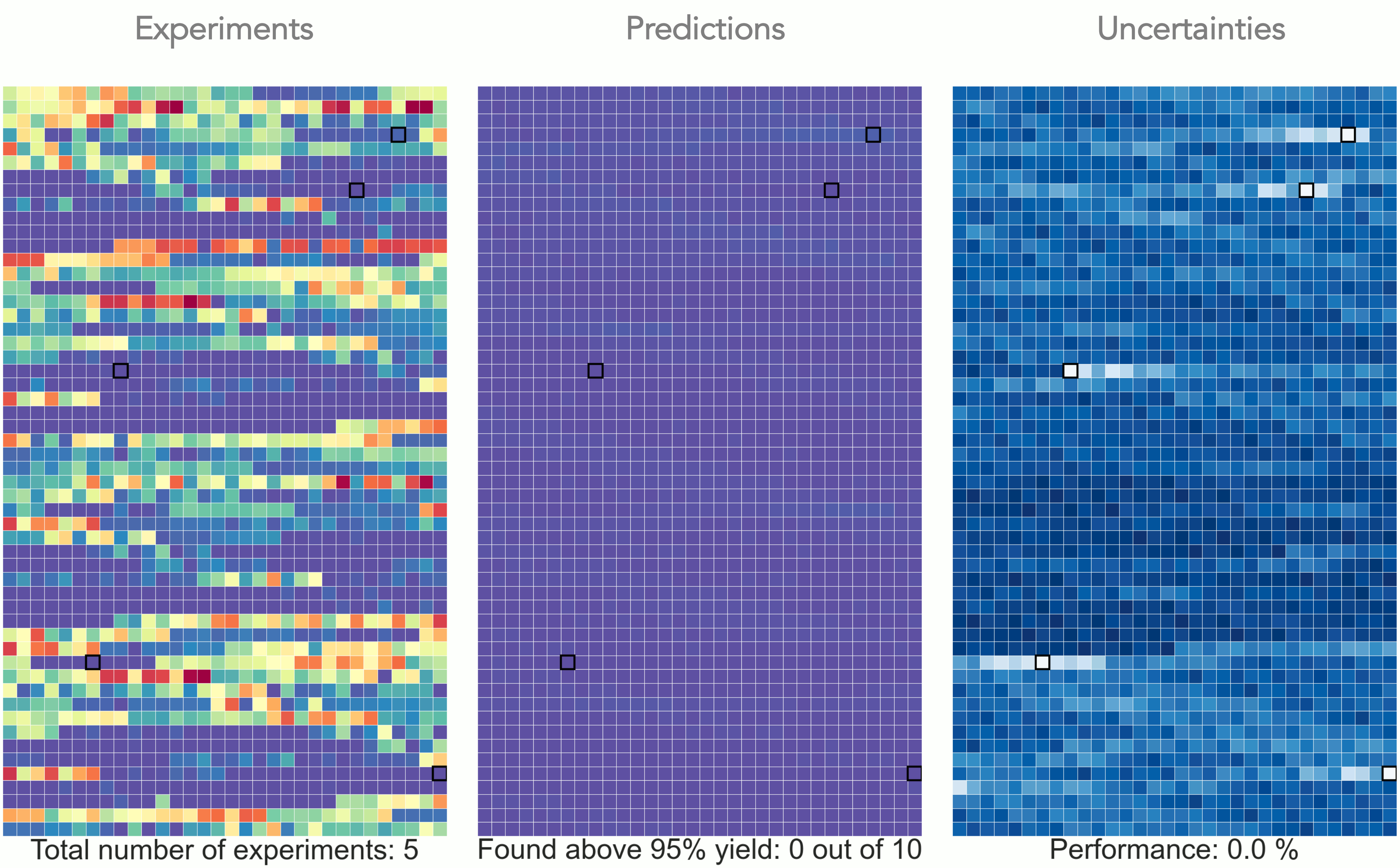
Optimized
manually:
49% yield
76% ee

$NiBr_2$diglyme,
L1, $MgCl_2$,
THF,
0.05M, PR



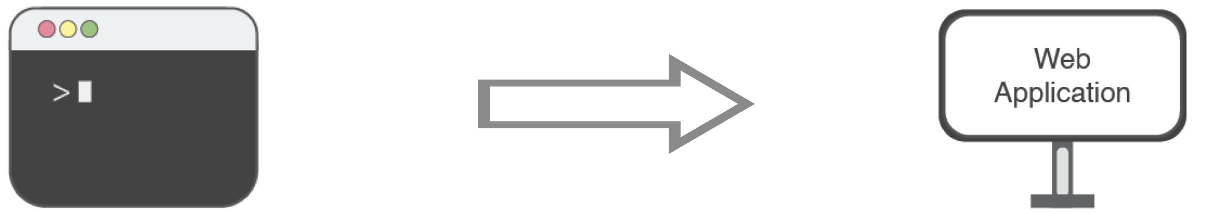
Will Lau, Jose Garrido Torres

Bayesian optimization of chemical processes – Visualization



Jose Garrido Torres

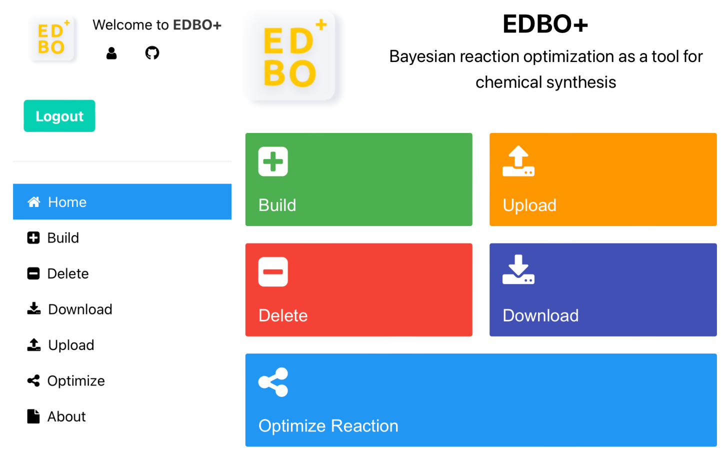
EDBO+ GUI



Command line interface (CLI) → Graphical user interface (GUI)

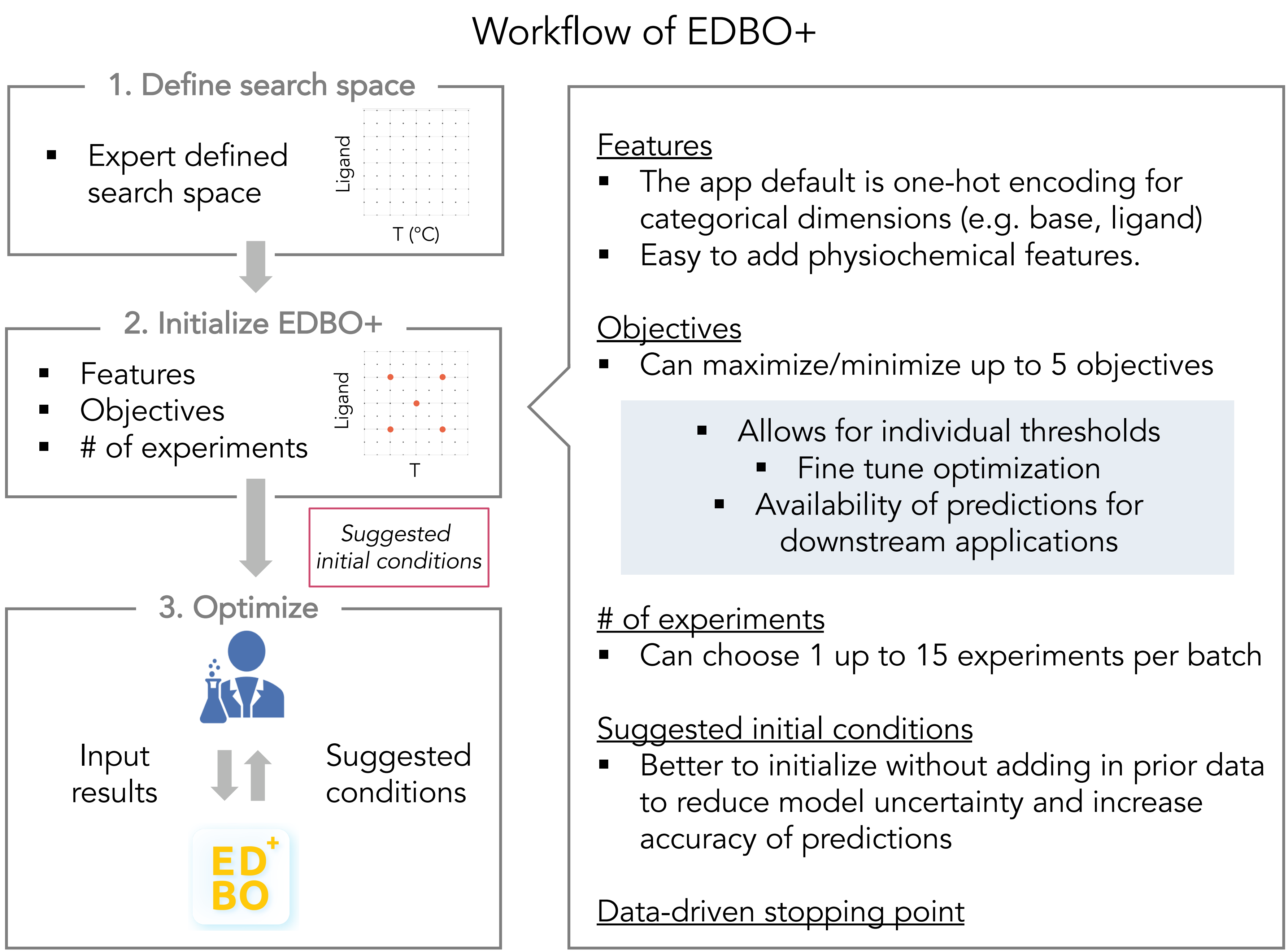
https://www.edbowebapp.com
https://github.com/doyle-lab-ucla/edboplus

```python
from edbo.plus.optimizer import EDBOplus

bo = EDBOplus()

components = {
    'ligand': ['L1', 'L2', 'L3'],
    'solvent': ['DMSO', 'THF'],
    'temperature': [25, 50, 60, 75],
    'time': [0, 6, 12, 24],
}

bo.generate_reaction_scope(
    filename='my_reaction.csv',
    components=components
)
```

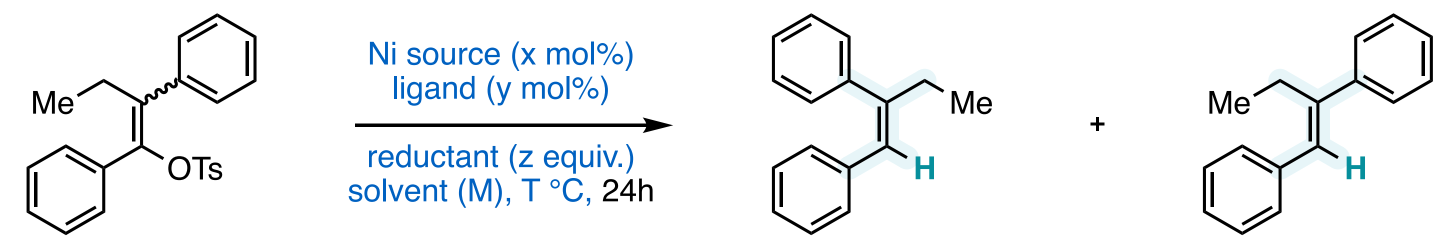


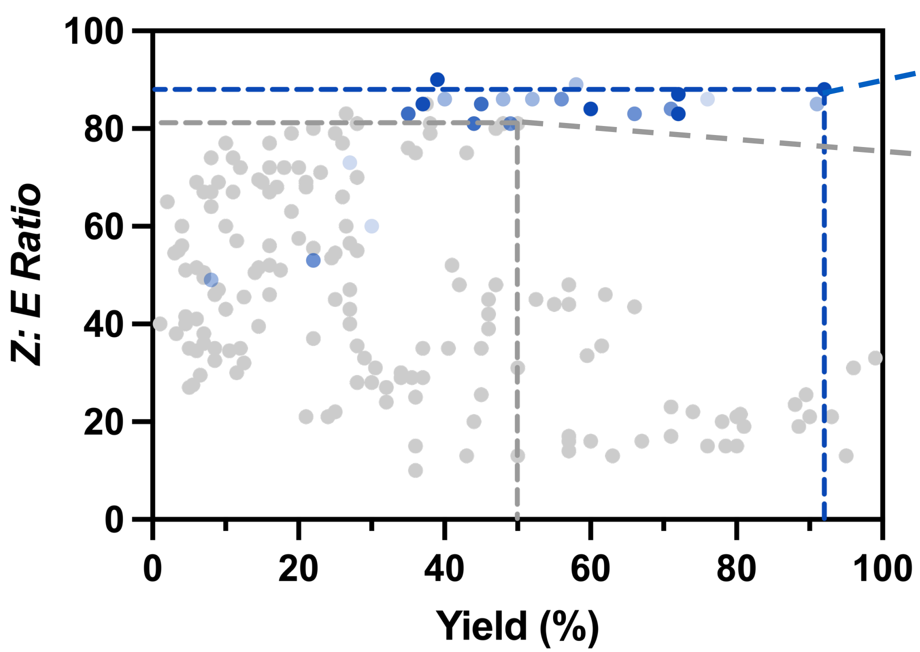
Computations are carried out in the cloud such that no software is necessary to use the optimizer web app.

Jose Garrido, Pranay Anchuri


Workflow of EDBO+
1. Define search space
Expert defined search space
Ligand
T (°C)
2. Initialize EDBO+
Features
Objectives
# of experiments
Ligand
T
Suggested initial conditions
3. Optimize
Input results
Suggested conditions
ED+ BO
Features
The app default is one-hot encoding for categorical dimensions (e.g. base, ligand)
Easy to add physiochemical features.
Objectives
Can maximize/minimize up to 5 objectives
Allows for individual thresholds
Fine tune optimization
Availability of predictions for downstream applications
# of experiments
Can choose 1 up to 15 experiments per batch
Suggested initial conditions
Better to initialize without adding in prior data to reduce model uncertainty and increase accuracy of predictions
Data-driven stopping point

Interaction effects in high dimensional space





<u>HTE Optimization</u>
81:19 Z:E
50% yield

$Ni(COD)_2$ (5.0 mol%)
CataCXiumPCy (15 mol%)
n-BuZnBr (2.0 equiv)
THF, 70 °C

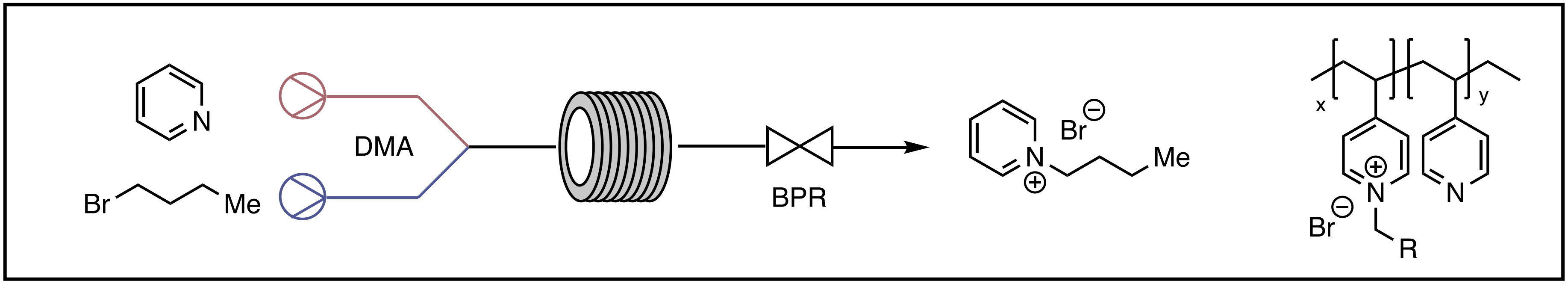
EDBO+
5 experiments/batch
GP surrogate
qEHVI acquisition
DFT/cheminformatics encoding

<u>Round 5</u>
88:12 Z:E
92% yield

$Ni(acac)_2$ (5.0 mol%)
CataCXiumPCy (15 mol%)
n-PrZnBr (3.0 equiv)
CPME, 50 °C

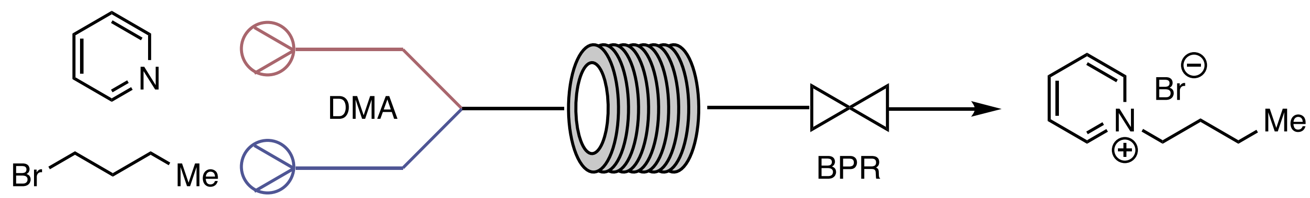
Natalie Romer (Sigman), Daniel Min (Doyle), & Daniel Zell (+ Genentech Team)

Data-driven stopping point – polymer functionalization



Luke Baldwin (AFRL), Jose Torres (Doyle) & team

Data-driven stopping point – polymer functionalization

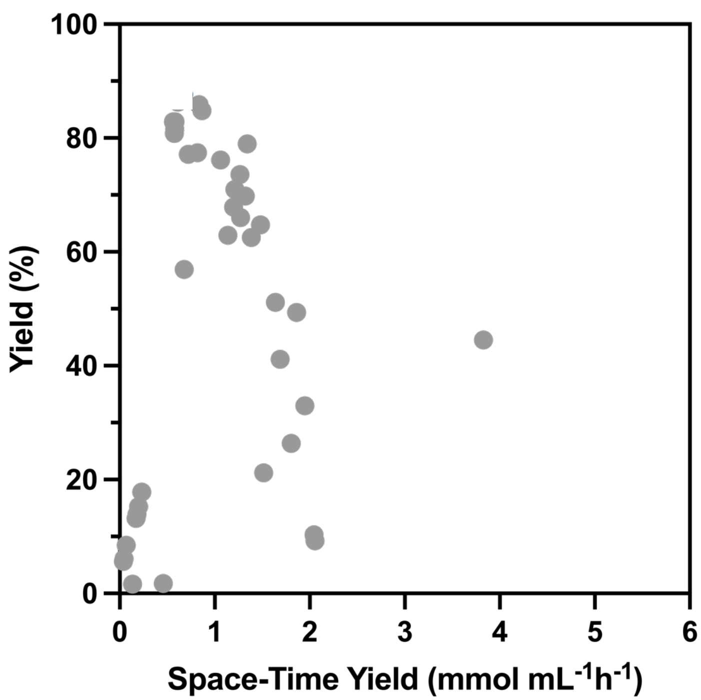


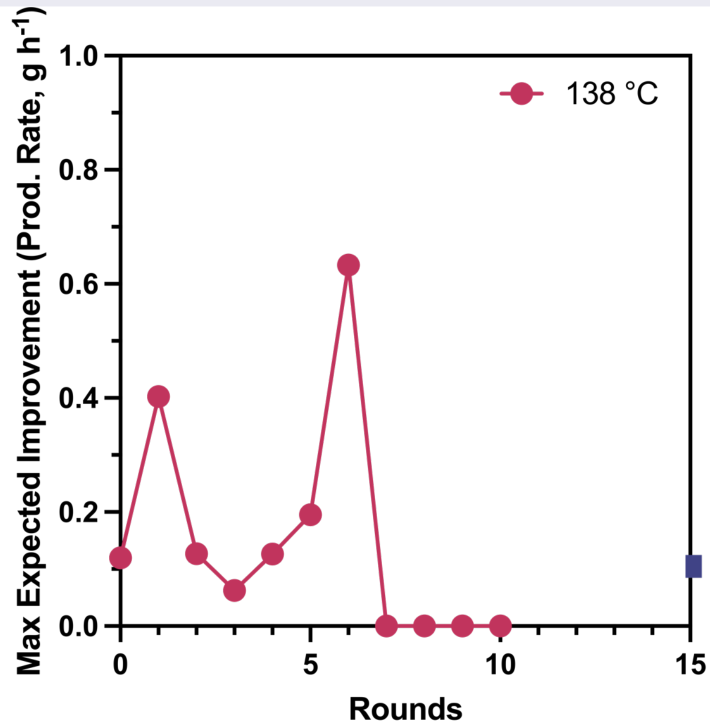
τ_{res}: 1 – 43min

T °C: 30 – 138 °C

χ_{pyr}: 0.33 – 0.66

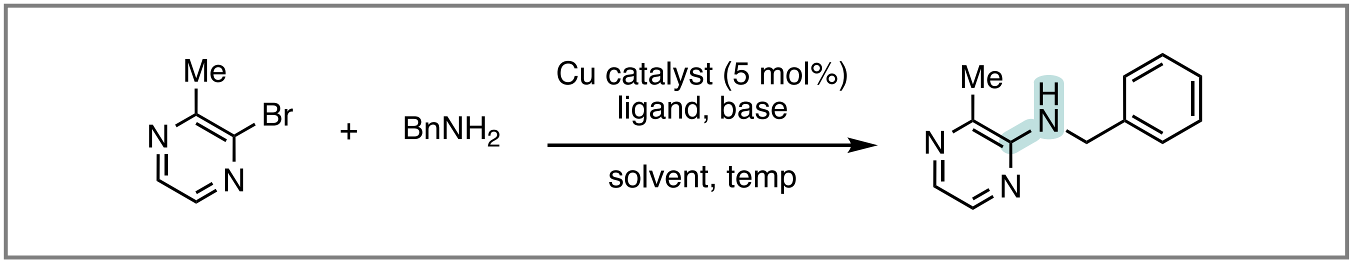
EDBO+: 3 experiments/round, GP model, qEHVI acquisition

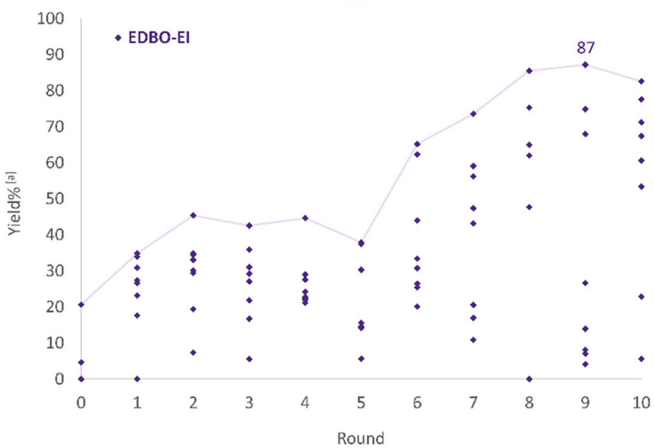


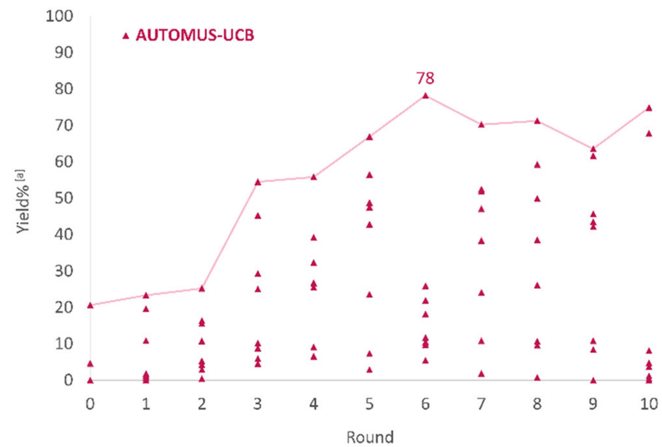


Luke Baldwin (AFRL), Jose Torres (Doyle) & team

Comparisons: Syngenta







#	BO	[Cu]	ligand (mol%)	base (equiv)	solvent (M)	*T* (°C)	equiv of 2a	yield (%)
5	EDBO-EI	CuBr	L7 (10)	K_3PO_4 (2)	DMSO (1)	80	1.5	87
6 b		CuBr	L7 (5)	K_3PO_4 (2)	EtOH (1)	80	1.5	86
9	AUTOMUS-UCB	Cu_2O	L12 (10)	K_2CO_3 (2)	DMSO (1)	80	1.5	78
10		CuI	L12 (5)	K_2CO_3 (2)	DMSO (1)	80	1.5	75

Braconi, E.; Godineau, E. *ACS Sustainable Chem. Eng.* **2023**, *11*, 10545

Comparisons: “A Chemist’s Guide to Multi-Objective Optimization”

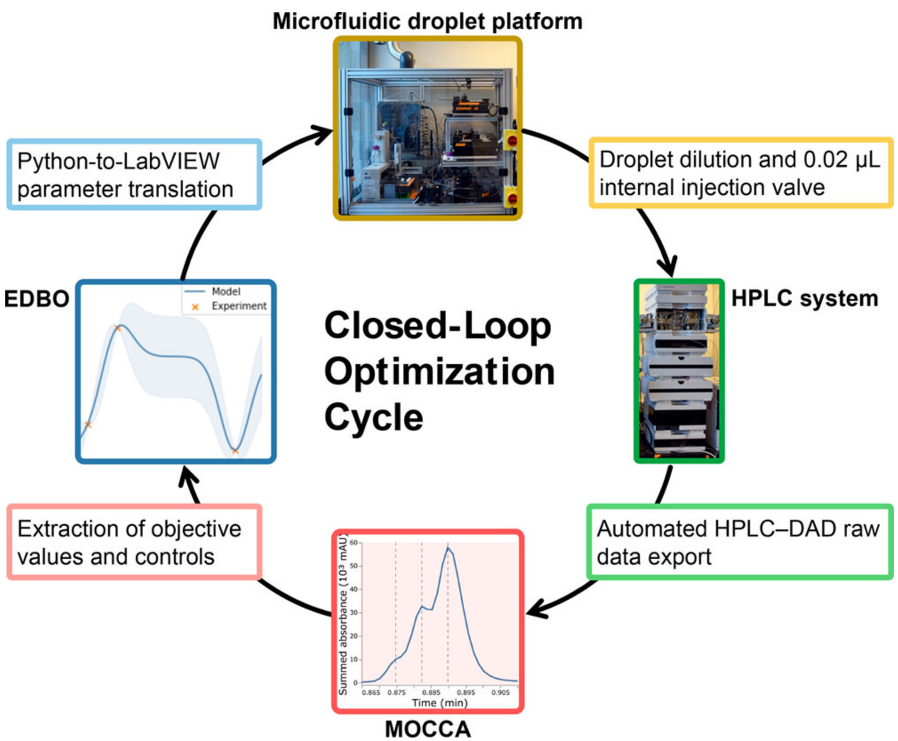
Solver	Continuous variables	Discrete Numeric	Categorical variables	Batch/Parallel Evaluation	Constraints
MVMOO	✓		✓		
EDBO+		✓	✓	✓	
Dragonfly	✓	✓	✓	✓	✓
TSEMO	✓			✓	
EIM-EGO	✓				

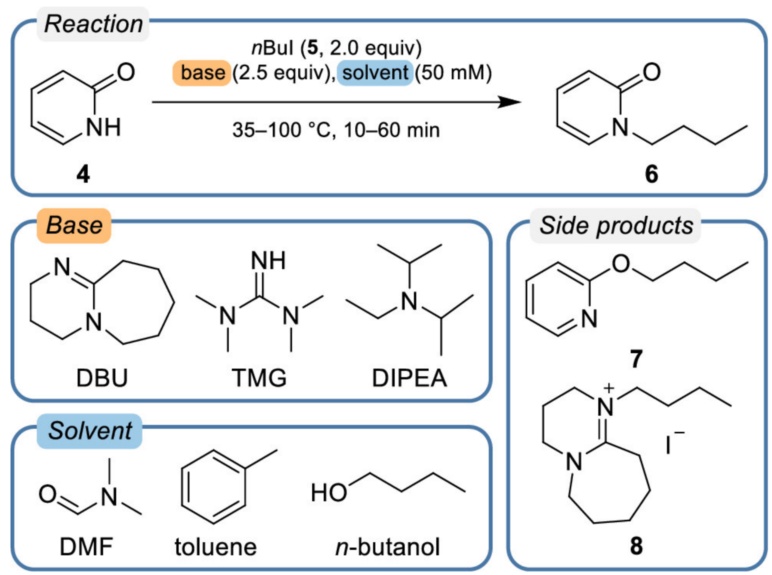
“Declaring a single solver as the definitive winner is difficult because each solver comes with its unique set of features. However, EDBO+ and MVMOO stand out as the best options, as they efficiently handle both continuous and categorical variables and consistently deliver strong performance across all problems.”

Vel, A. S.; Cortés-Borda, D.; Felpin, F.-X. *Reaction Chem. Eng.* **2024**, *9*, 2882

Manson, J. A.; Chamberlain, T. W.; Bourne, R. A. *J. Glob. Optim.*, **2021**, *80*, 865

Applications: a few examples from others – Jensen



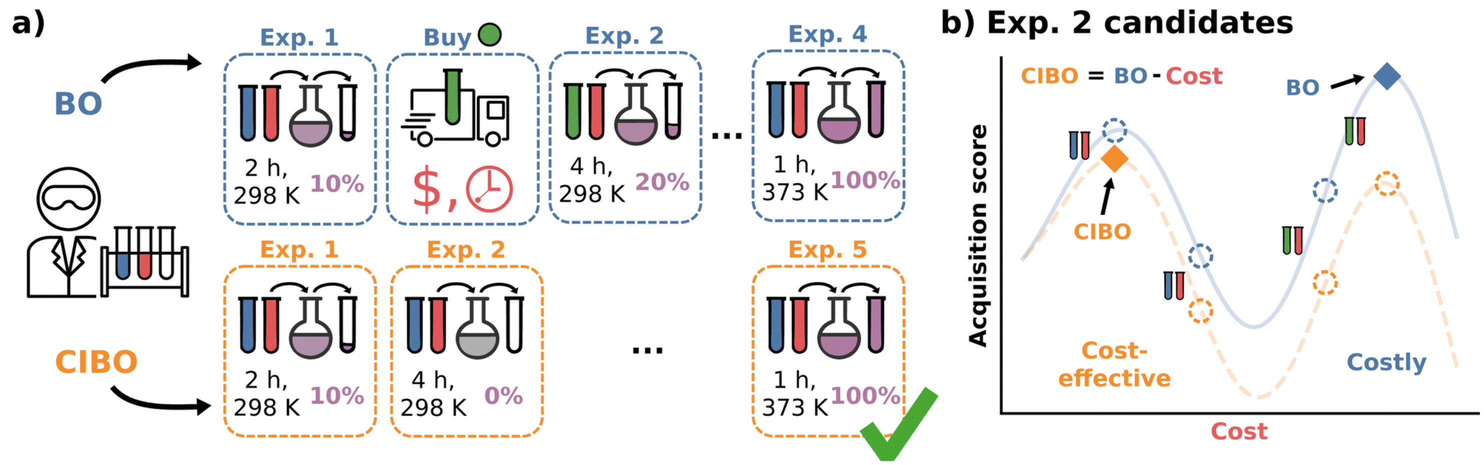


Open-source Python project called MOCCA for the analysis of HPLC–DAD (photodiode array detector) raw data

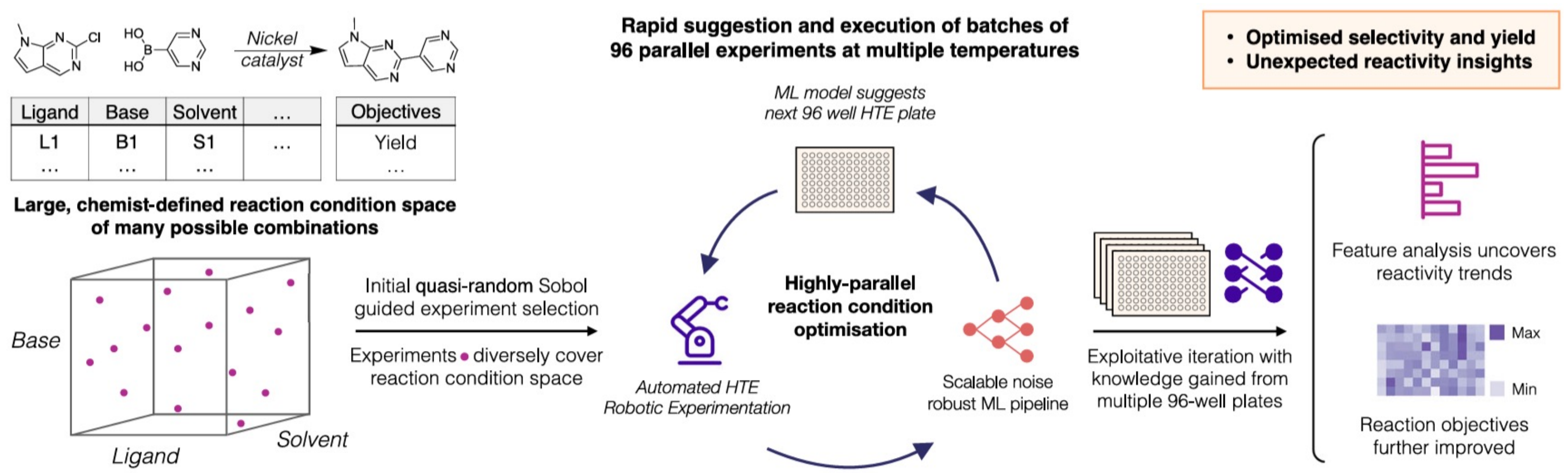
Hillenbrand, J.; Volpin, G.; Jensen, K. F. *et al. ACS Cent. Sci.* **2023**, *9*, 307
Bonciolini, S.; Pulcinella, A.; Noël, T. *J. Am. Chem. Soc.* **2025**, *147*, 28523

Advances: a few examples from others

Cost-Informed Bayesian Optimization - Corminboeuf & Waser



Highly parallel Bayesian Optimization - Schwaller



Schoepfer, A. A.; Weinreich, J.; Laplaza, R.; Waser, J.; Corminboeuf, C. *Digital Discovery*, **2024,** *3*, 2289

Sin, J. W.; Chau, S. L.; Burwood, R. P.; Püntener, K.; Bigler, R.; Schwaller, P. *Nature Commun*. **2025**, *16*, 6464

Graduate students: Jason Wang, Maddy Ruos, Judah Raab, Erin Bucci, Daniel Min, Winston Gee, Paris Dee, Flora Fan, Braden Chaffin, Neyci Gutierrez-Valencia, Alex Maertens

Postdoctoral fellows: Melecio Perea, Rajat Maji, Alex Cusumano, Emily Wearing

Undergraduate students: Giselle Brown, Gabe Jones-Thomson

Thank you!

Doyle Group Alumni

Collaborators

Ryan Adams (Princeton), Matt Sigman (Utah), Jason Stevens, Jun Li, Jose Tabora, Kristine Golden (BMS), Sarah Reisman (Caltech), Ryan Hadt (Caltech), Matt Bird (NREL), Connor Coley (MIT), Iulia Strambeanu, Kristi Leonard, Jordan Compton, Santeri Aikonen, Paolo Neves (Janssen), Francis Gosselin, Allen Hong, Julie Deichert (Genentech), Dipannita Kalyani (Merck), Shashank Shekhar, Michael Haibach (Abbvie), Chong Liu (UCLA), Luke Baldwin (AFRL), Ken Houk (UCLA), Anastassia Alexandrova (UCLA), Steven Lopez (Northwestern), Osvaldo Gutierrez (UCLA)

Funding

NIH (R35 GM126986), NIH S10OD034287, NSF CCI CCAS, NSF (CHE-1148750), NSF HDR ID4 Institute, DOE EFRC BioLEC, Camille & Henry Dreyfus Foundation, Bristol Myers Squibb, Genentech, Merck, Janssen, Novartis, Bill & Melinda Gates Foundation, BioPacific MIP

NIH NIGMS · U.S. DEPARTMENT OF ENERGY · NSF · C-CAS Center for Computer Assisted Synthesis · bioLEC · BILL & MELINDA GATES foundation

BioPACIFIC NSF Materials Innovation Platform DMR-1933487 · NSF ID4 Institute for Data Driven Dynamical Design · SCHMIDT FUTURES