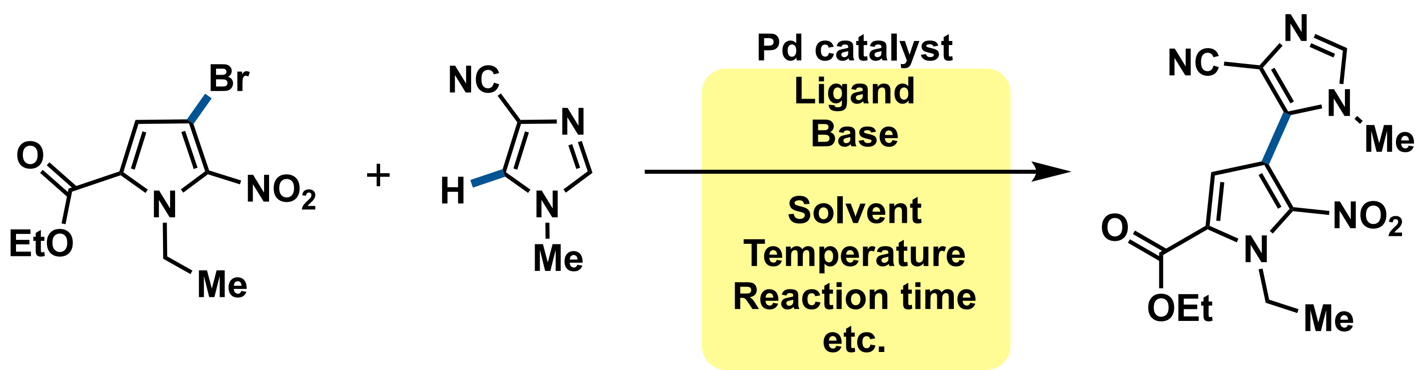
Bayesian Optimization for the Exploration of Reaction Conditions

Literature seminar

M2 Yu Irie

2023/9/28 (Thu)

Optimization of chemical reaction



>10^5 configurations!!

In typical laboratory, time and materials are limited…

- Scour chemical literature for similar reactions
- Experience
- Mechanistic understanding

…

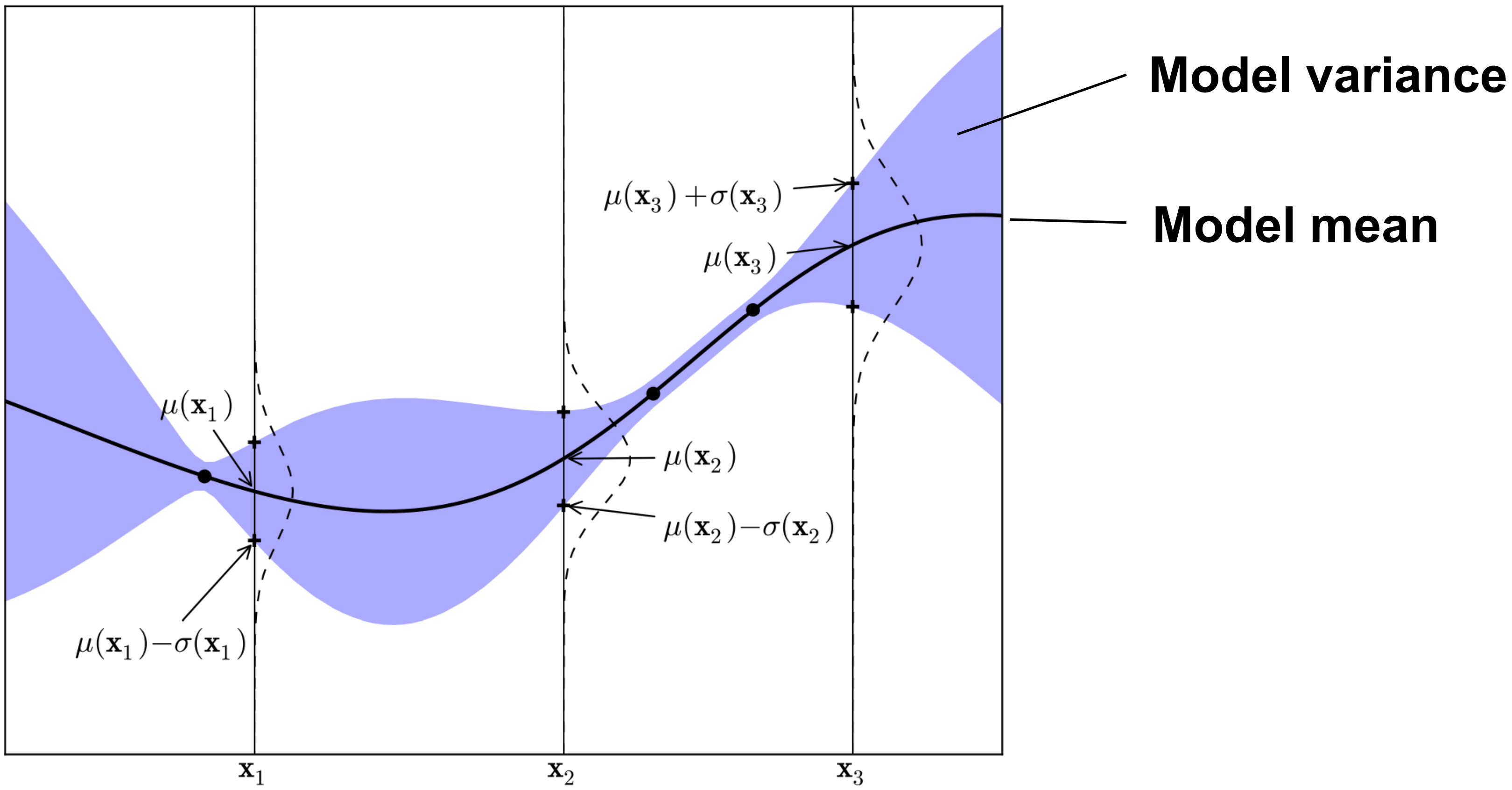
Reaction optimization in machine learning

$$\boldsymbol{x} \longrightarrow \boxed{\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})} \longrightarrow \mathrm{y}$$

Blackbox function

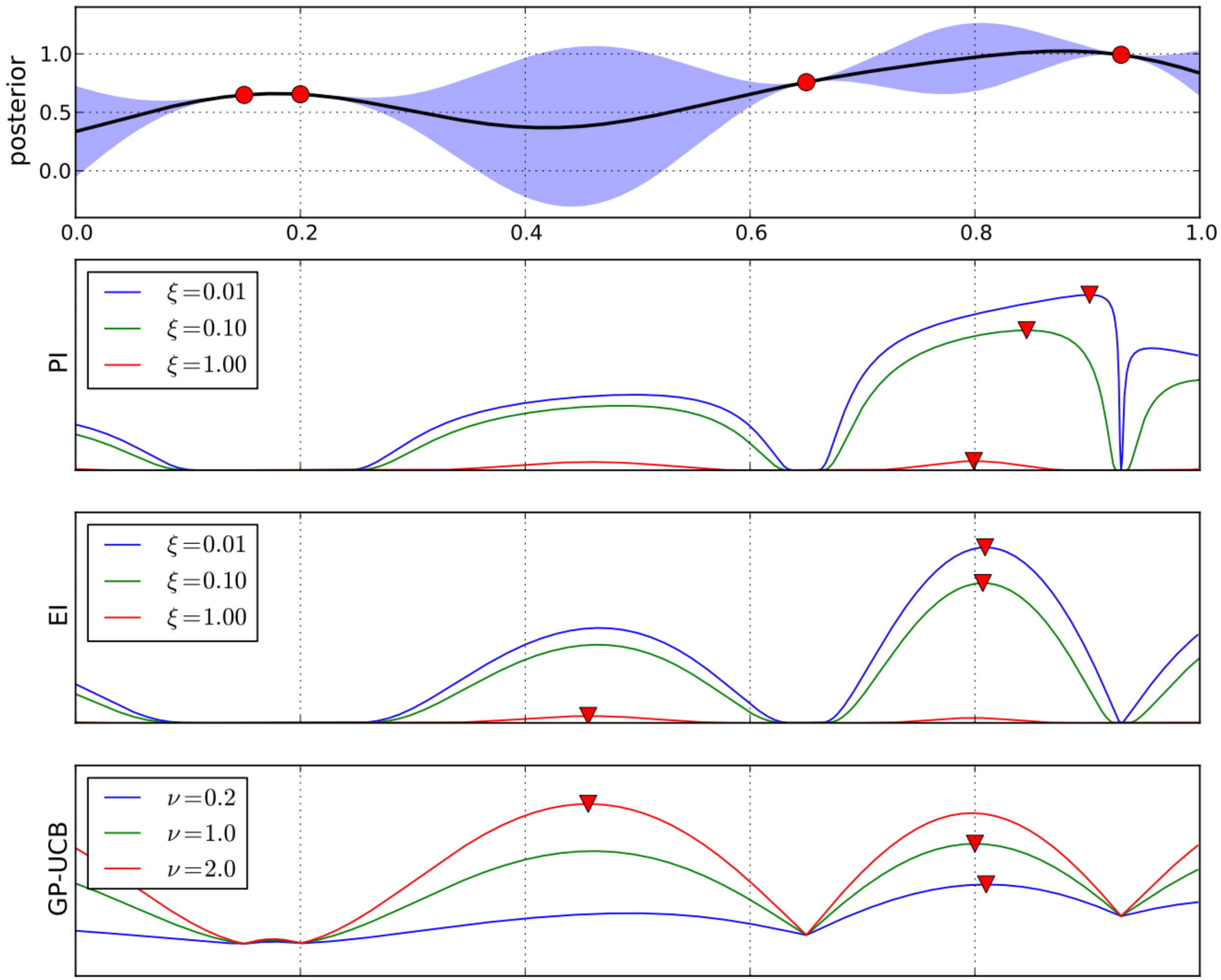
- **Automated approaches to algorithm optimization**
  - **Bayesian optimization**
    - ✓ **High-quality configurations in fewer evaluations**

Gaussian process



E. Brochu, V. M. Cora, N. de Freitas, https://doi.org/10.48550/arXiv.1012.2599

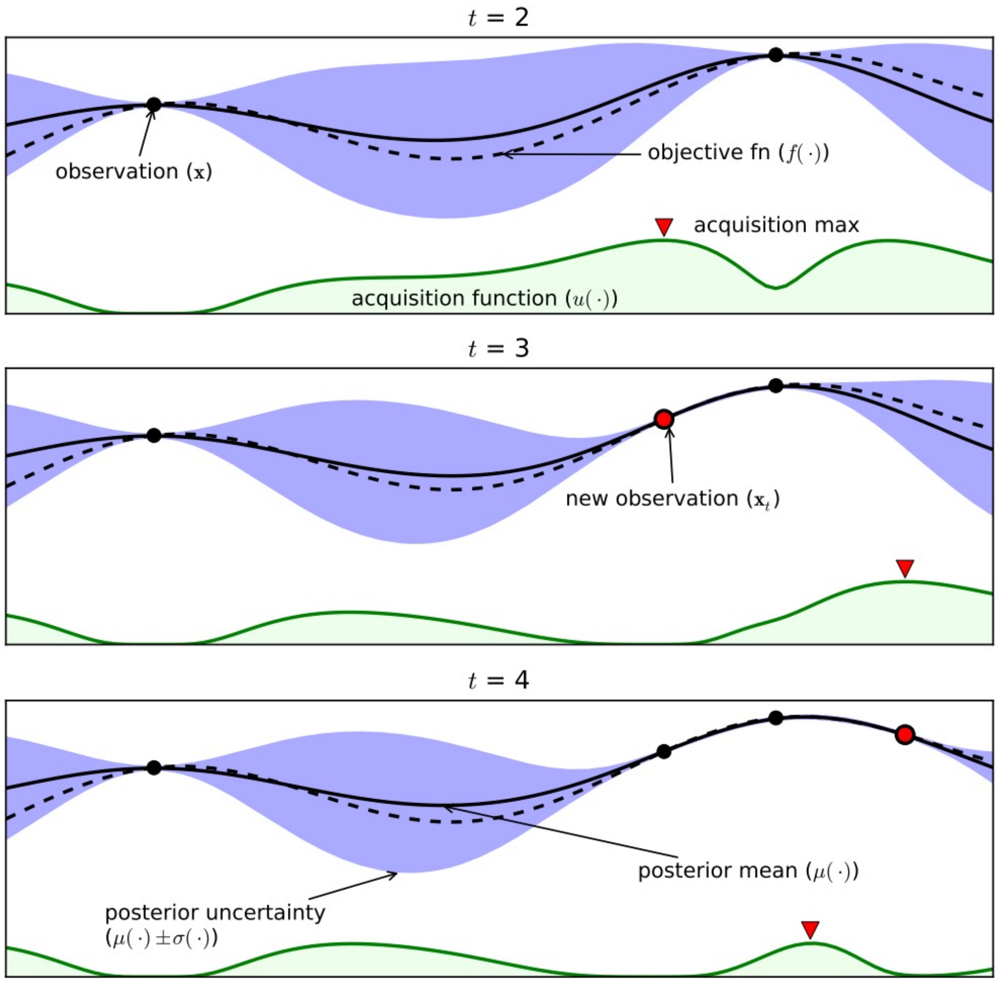
Acquisition function



Exploitation
vs.
Exploration

E. Brochu, V. M. Cora, N. de Freitas, https://doi.org/10.48550/arXiv.1012.2599

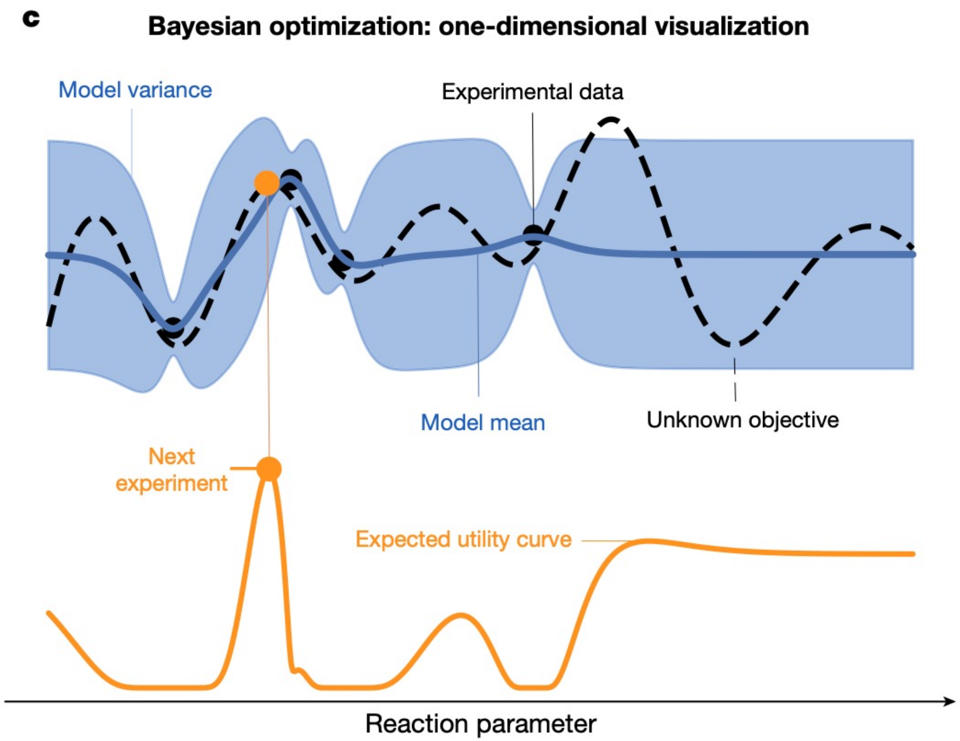
Bayesian optimization

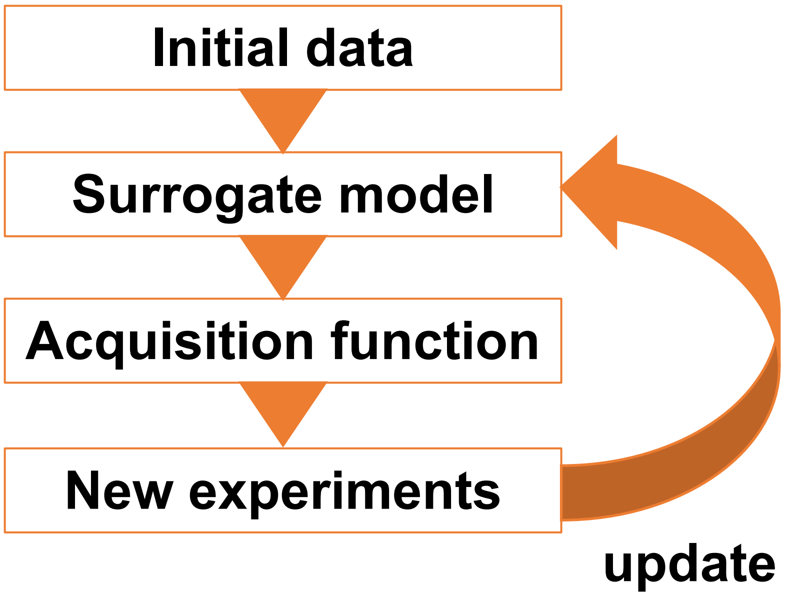


- ✓ **Find the objective function**
- ✓ **Uncertainty is gradually minimized**
- ✓ **Where to observe next is determined automatically**

E. Brochu, V. M. Cora, N. de Freitas, https://doi.org/10.48550/arXiv.1012.2599

Bayesian optimization





- Application to diverse search spaces
- Selection of multiple experiments in parallel

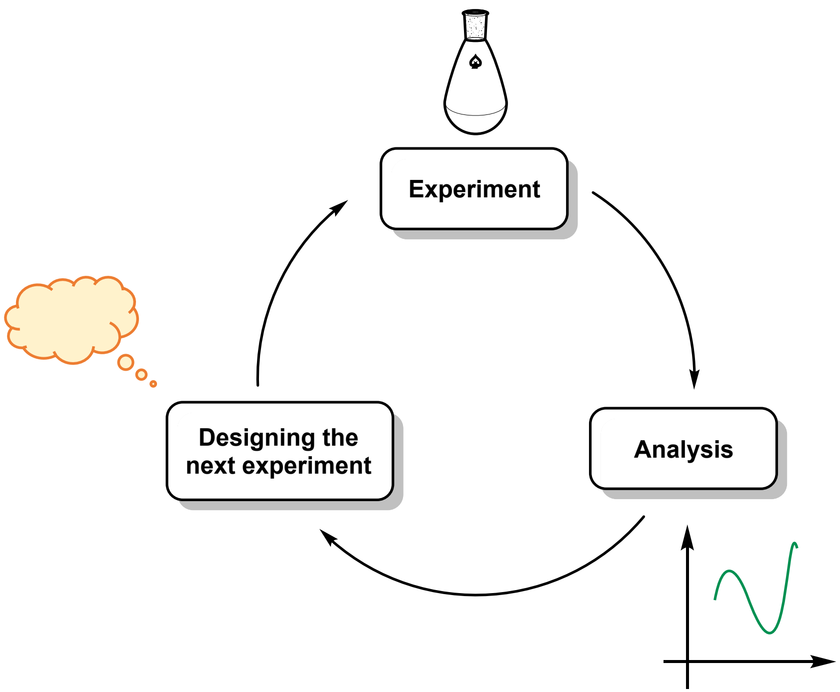
➢ **Optimization of chemical processes**

Table of contents

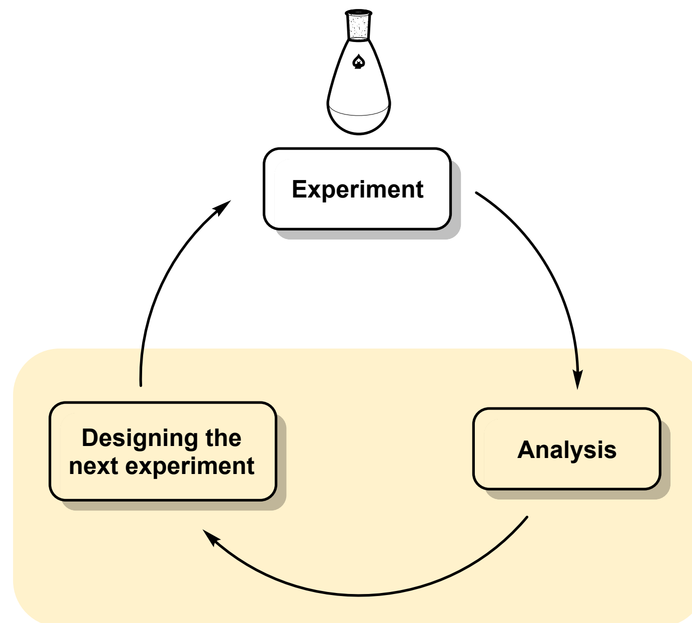
Table of contents

Application to chemical reactions

Conventional reaction optimization



- **Acceleration of the optimization of synthetic reactions**

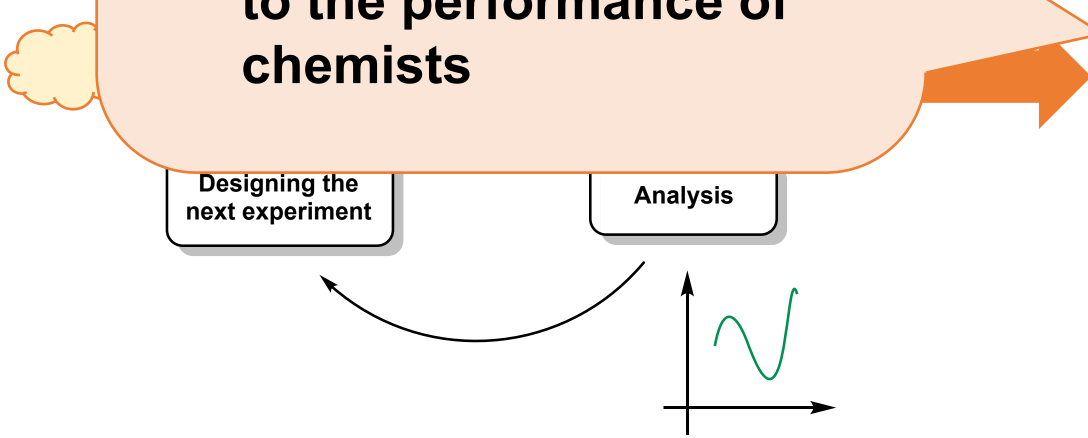


Bayesian optimization (BO)

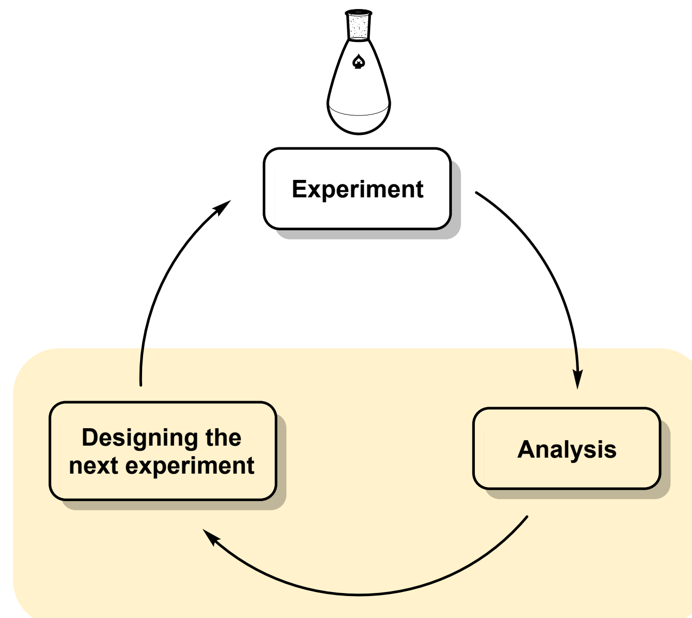
B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. M. Alvarado, J. M. Janey, R. P. Adams, A. G. Doyle, *Nature* **2021**, *590*, 89–96.

Application to chemical reactions

1. Application to typical batch chemistry
2. General-purpose software platforms
3. Systematic comparisons to the performance of chemists



➢ **Acceleration of the optimization of synthetic reactions**

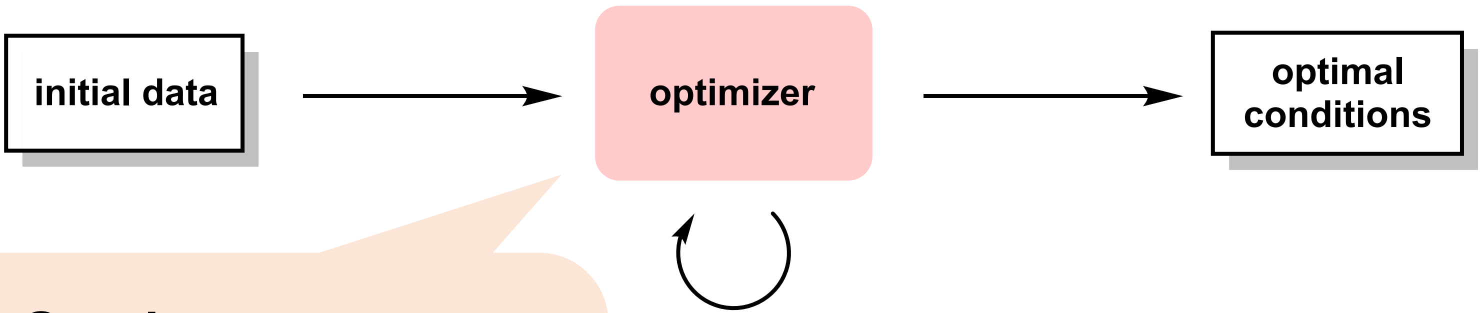


Bayesian optimization (BO)

B. J. Shields, J. Stevens, J. Li, M. Parasram, F. Damani, J. I. M. Alvarado, J. M. Janey, R. P. Adams, A. G. Doyle, *Nature* **2021**, *590*, 89–96.

Table of contents

Tuning of the algorithm components



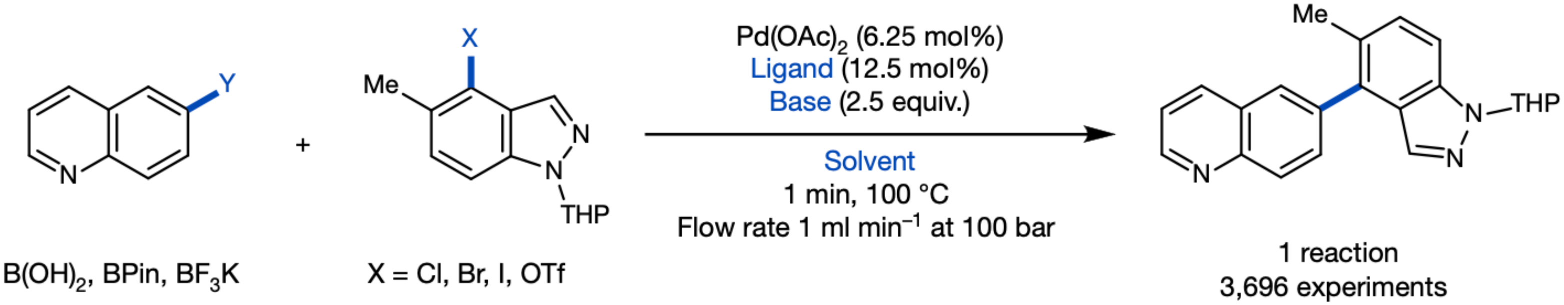
- ✓ Good average performance
- ✓ Outcomes with low variance with respect to the initial data

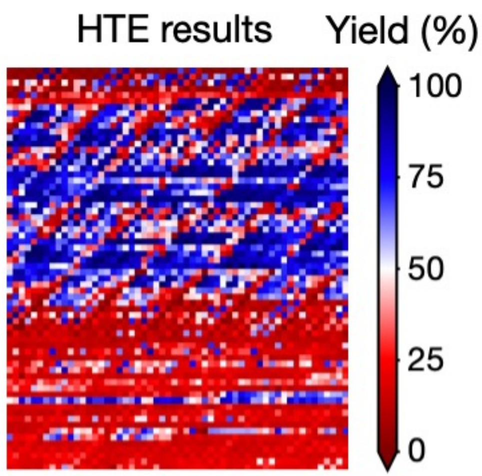
Reaction optimizations with different random initial starting data

R. P. Adams, A. G. Doyle, *et al. Nature* **2021**, *590*, 89–96.

Training data used to select BO parameters

1

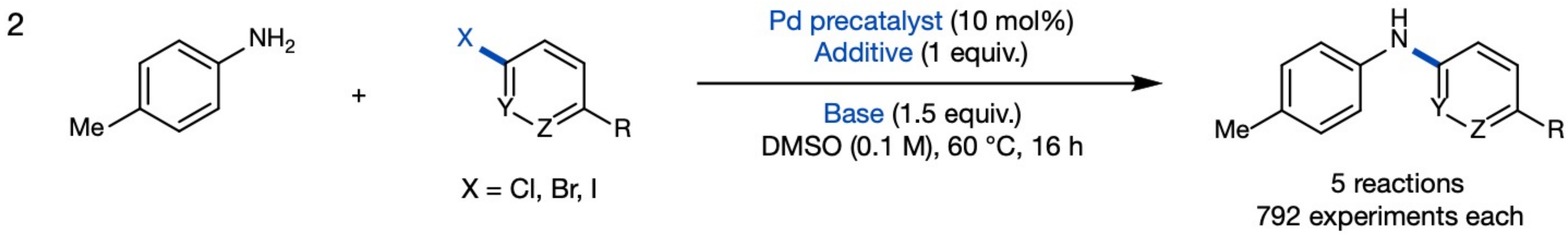




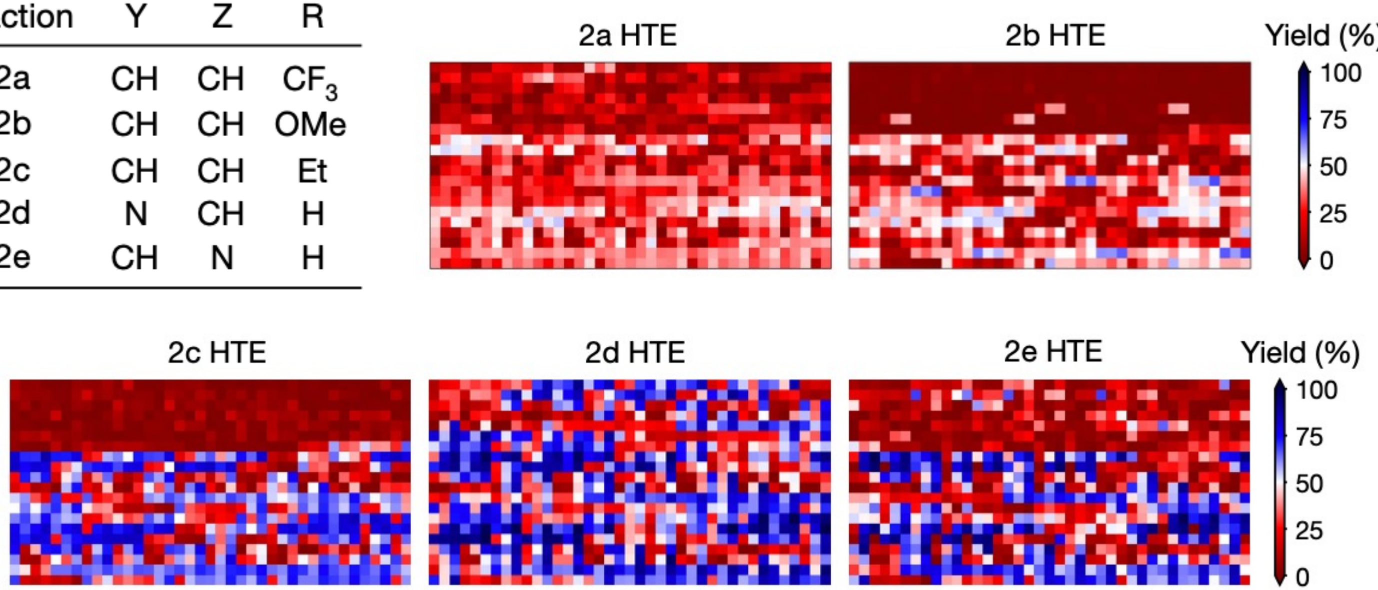
- **Published results of HTE (High Throughput Experiments) of 6 reactions in total were used for model optimization.**

P. Richardson, N. W. Sach, *et al. Science* **2018**, *359*, 429–434.
R. P. Adams, A. G. Doyle, *et al. Nature* **2021**, *590*, 89–96.

Training data used to select BO parameters

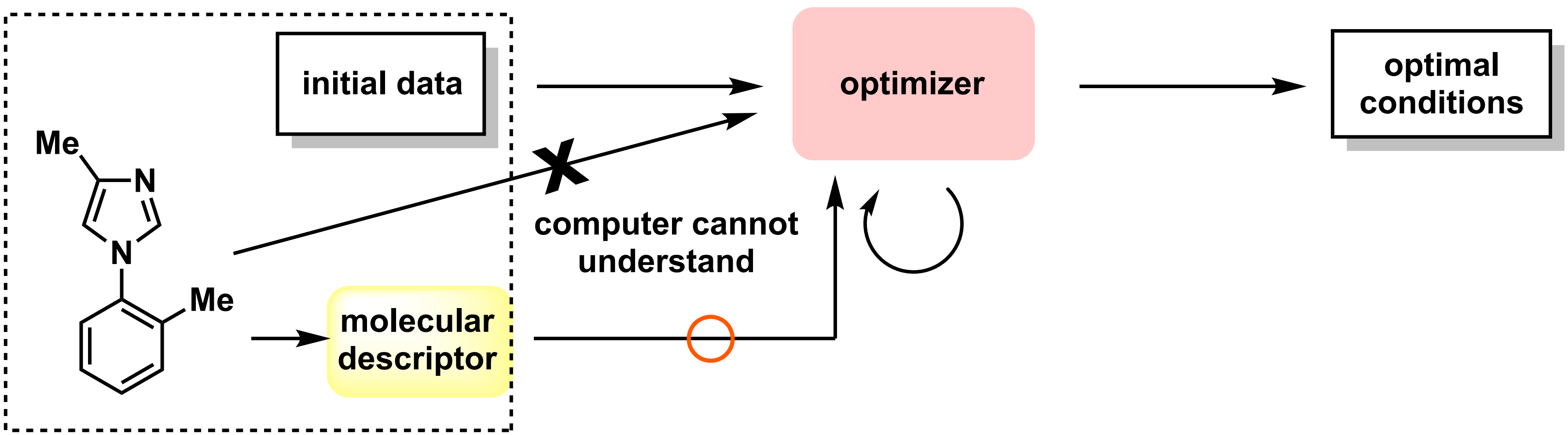


Reaction	Y	Z	R
2a	CH	CH	CF_3
2b	CH	CH	OMe
2c	CH	CH	Et
2d	N	CH	H
2e	CH	N	H



S. D. Dreher, A. G. Doyle, *et al. Science* **2018**, *360*, 186–190.
R. P. Adams, A. G. Doyle, *et al. Nature* **2021**, *590*, 89–96.

Selection of descriptors

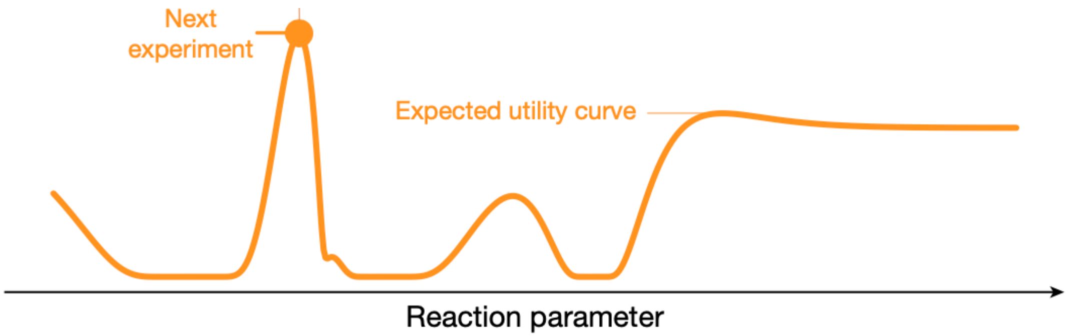


In this report, Density Functional Theory Descriptors was the best.

Contains: global (e.g. LUMO energy and dipole moment)/ local (e.g. atomic NMR shift and charge for labeled atoms) electronic and global steric (e.g. molar volume) descriptors etc.

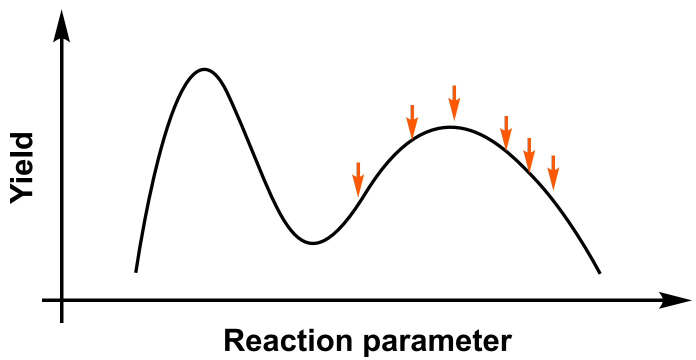
R. P. Adams, A. G. Doyle, *et al. Nature* **2021**, *590*, 89–96.

Optimization of an acquisition function



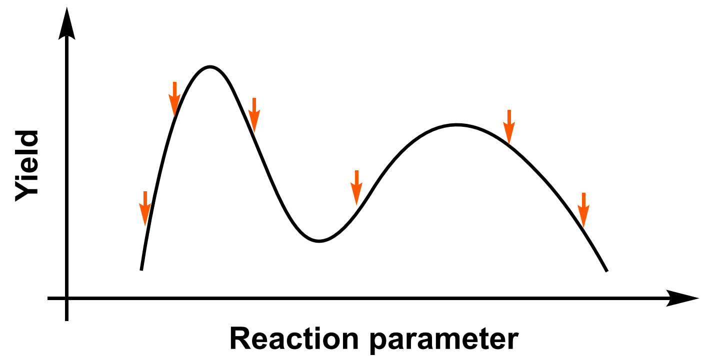
✓ **utilization of both information and uncertainty to drive optimization**

➢ **Pure exploitation**



- **Could become trapped in local maxima**

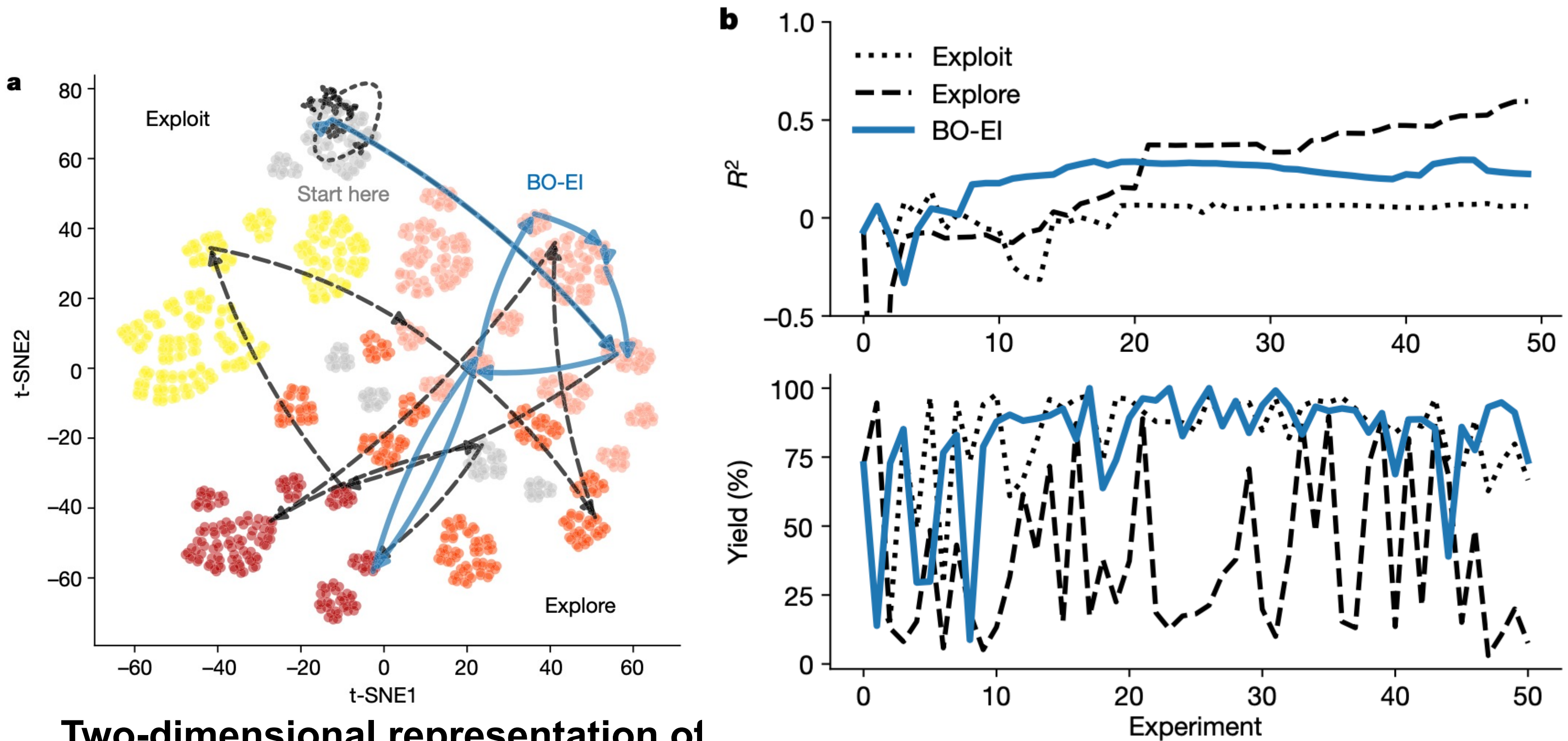
➢ **Pure exploration**



- **May achieve the best global understanding**

R. P. Adams, A. G. Doyle, *et al. Nature* **2021**, *590*, 89–96.

Balancing exploration and exploitation



Two-dimensional representation of reaction 1

Model fit score

Explorer: Better fit, low yield
Exploiter: High yield

Balancing

Expected improvement

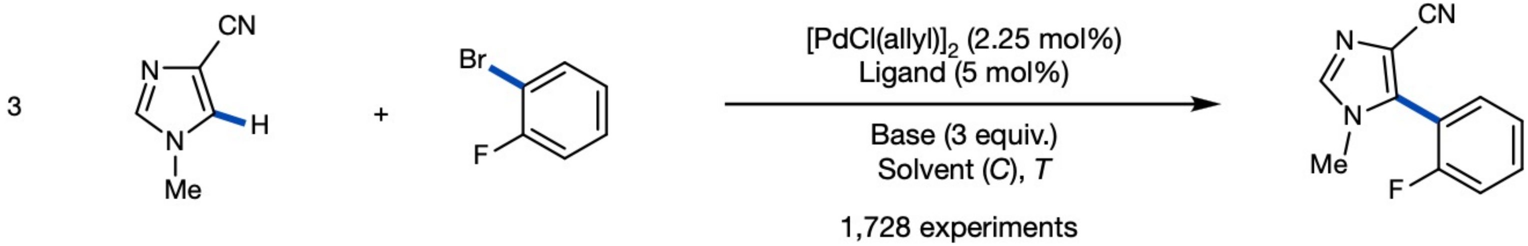
R. P. Adams, A. G. Doyle, *et al. Nature* **2021**, *590*, 89–96.

Table of contents

Test in a new reaction space

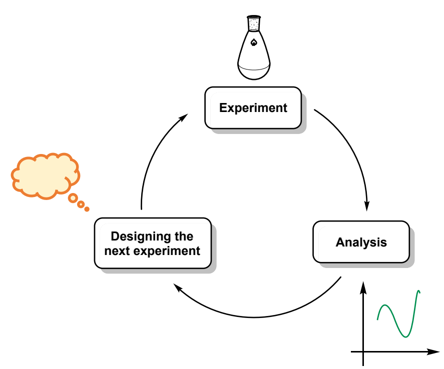
Palladium-catalyzed C–H functionalization

a

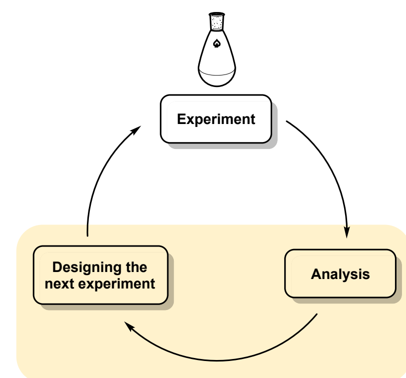


High-throughput experimentation

Data collection



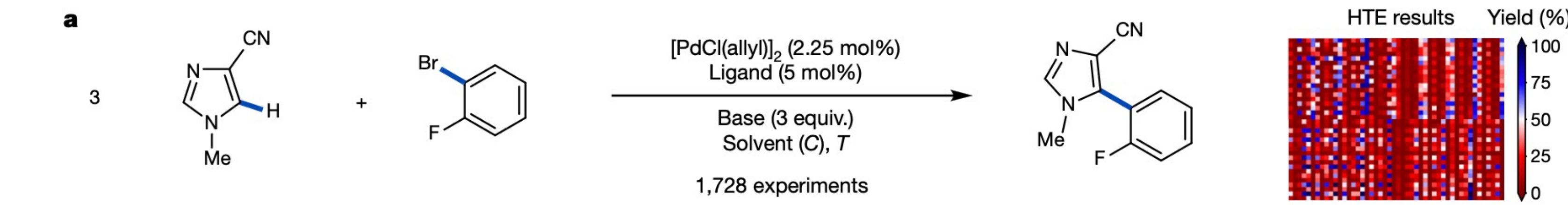
vs.

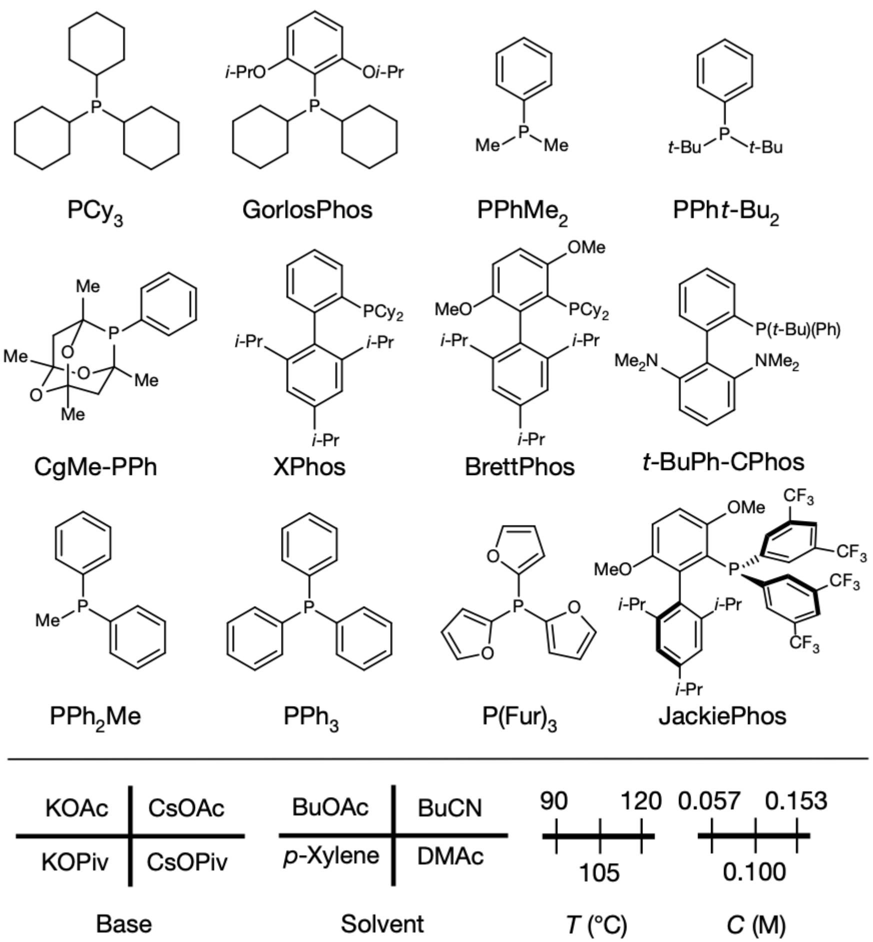


Bayesian optimization (BO)

R. P. Adams, A. G. Doyle, *et al. Nature* **2021**, *590*, 89–96.

BO vs. human experts (Data collection)

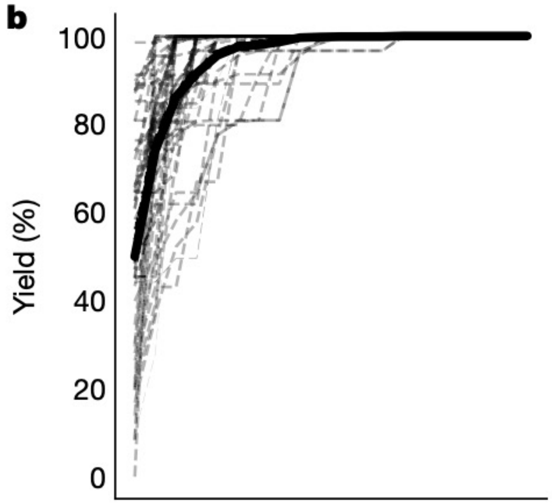




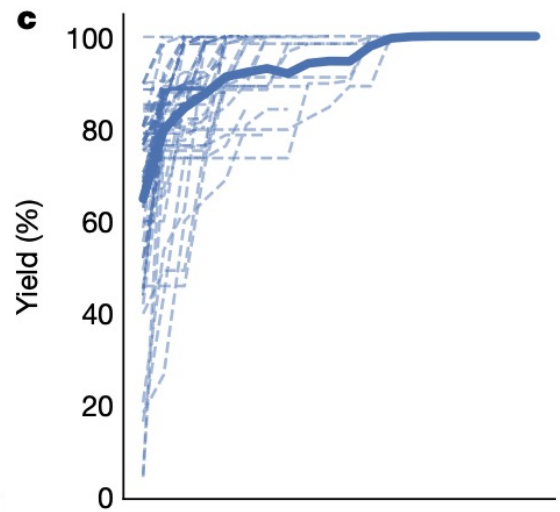
- **Experiments that were expected to give a satisfactory distribution across the larger search space was selected.**
- **Collected experimental results for the entire search space via HTE.**

R. P. Adams, A. G. Doyle, *et al. Nature* **2021**, *590*, 89–96.

BO vs. human experts (Game results)



ML performance

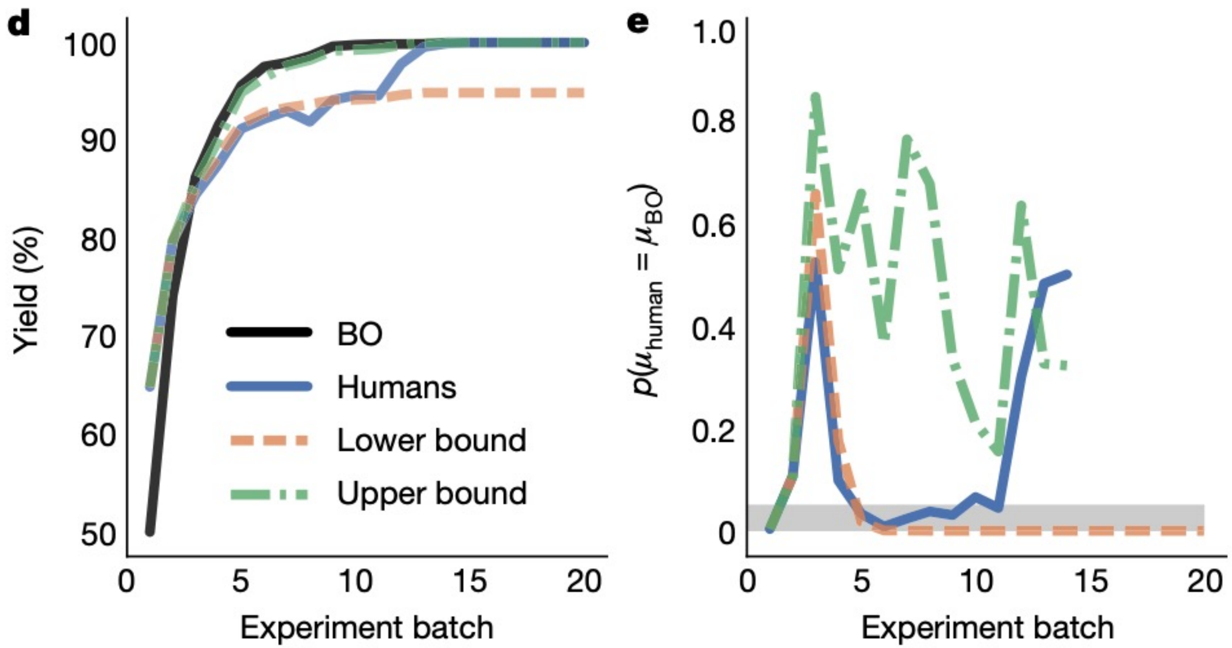


human performance

	Optimizer	Humans
Initial choices	random	significantly better selection
Final results	○ (the average performance within three batches of five experiments)	△

R. P. Adams, A. G. Doyle, *et al. Nature* **2021**, *590*, 89–96.

BO vs. human experts (Game results)

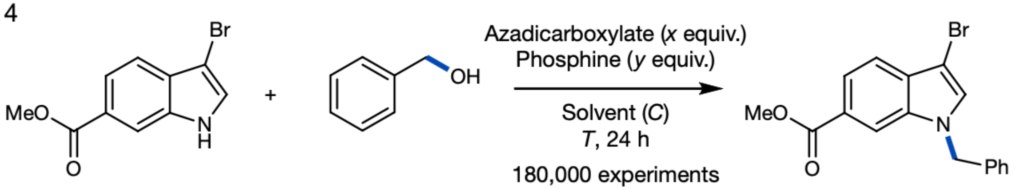


- **Best- and worst-case bounds for average human performance**
  - ➢ **Bayesian reaction optimization on average outperformed human experts.**

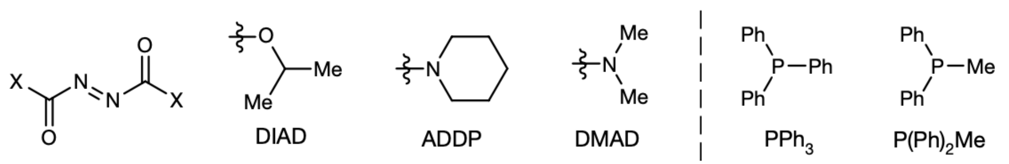
R. P. Adams, A. G. Doyle, *et al. Nature* **2021**, *590*, 89–96.

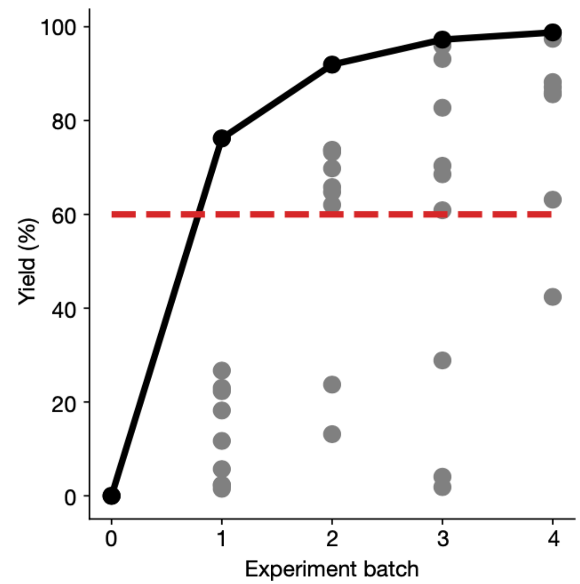
Optimization of a Mitsunobu reaction

4



Entry	AD	x (equiv.)	Phosphine	y (equiv.)	Solvent	C (M)	T (°C)	Yield
1	DIAD	1.1	PPh_3	1.1	THF	0.1	25	60%
2	ADDP	1.9	$P(Ph)_2Me$	1.9	THF	0.2	25	99%
3	DMAD	1.9	$P(Ph)_2Me$	1.9	MeCN	0.15	45	99%
4	ADDP	1.9	$P(Ph)_2Me$	1.9	THF	0.15	45	99%





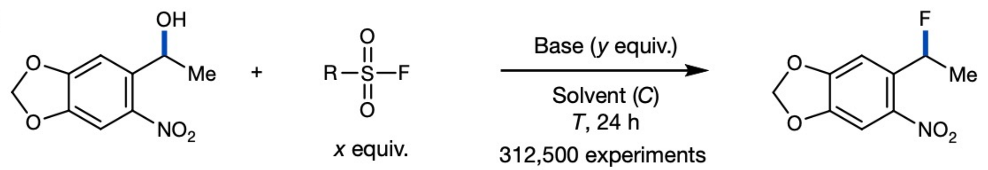
➢ **From 180,000 possible configurations in total…**

- ✓ **The optimizer quickly surpassed the benchmark result.**
- ✓ **Three distinct sets of reaction conditions (99% yield) were identified in only four rounds of ten experiments.**

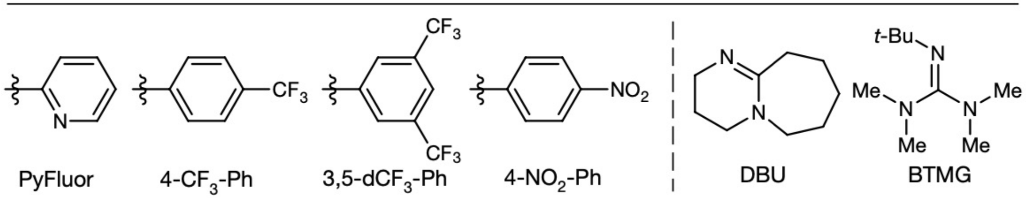
R. P. Adams, A. G. Doyle, *et al. Nature* **2021**, *590*, 89–96.

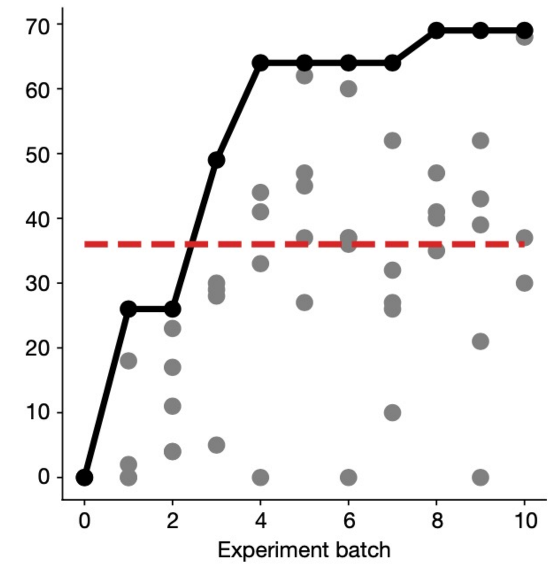
Optimization of a deoxyfluorination reaction

5



Entry	R	x (equiv.)	Base	y (equiv.)	Solvent	C (M)	T (°C)	Yield
1	PyFluor	1.1	DBU	2.0	THF	0.5	20	36%
2	4-CF_3-Ph	1.9	BTMG	1.9	$PhCF_3$	0.5	20	69%
3	3,5-dCF_3-Ph	1.1	BTMG	1.9	PhMe	0.1	60	69%
4	4-NO_2-Ph	1.9	BTMG	1.9	PhMe	0.5	20	68%





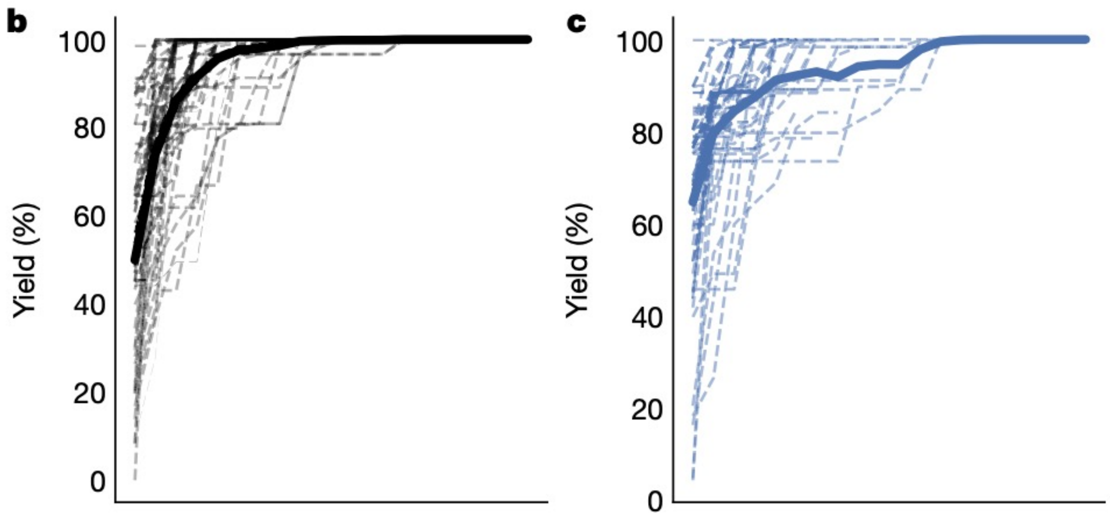
➢ **From 312,500 possible configurations in total…**

✓ **The optimizer surpassed the benchmark result within three rounds of five experiments.**

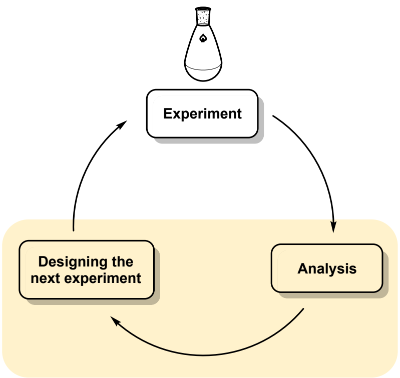
✓ **Reaction conditions that produced TM in 69% yield were identified in ten rounds of experiments.**

R. P. Adams, A. G. Doyle, *et al. Nature* **2021**, *590*, 89–96.

Short summary



ML performance **human performance**



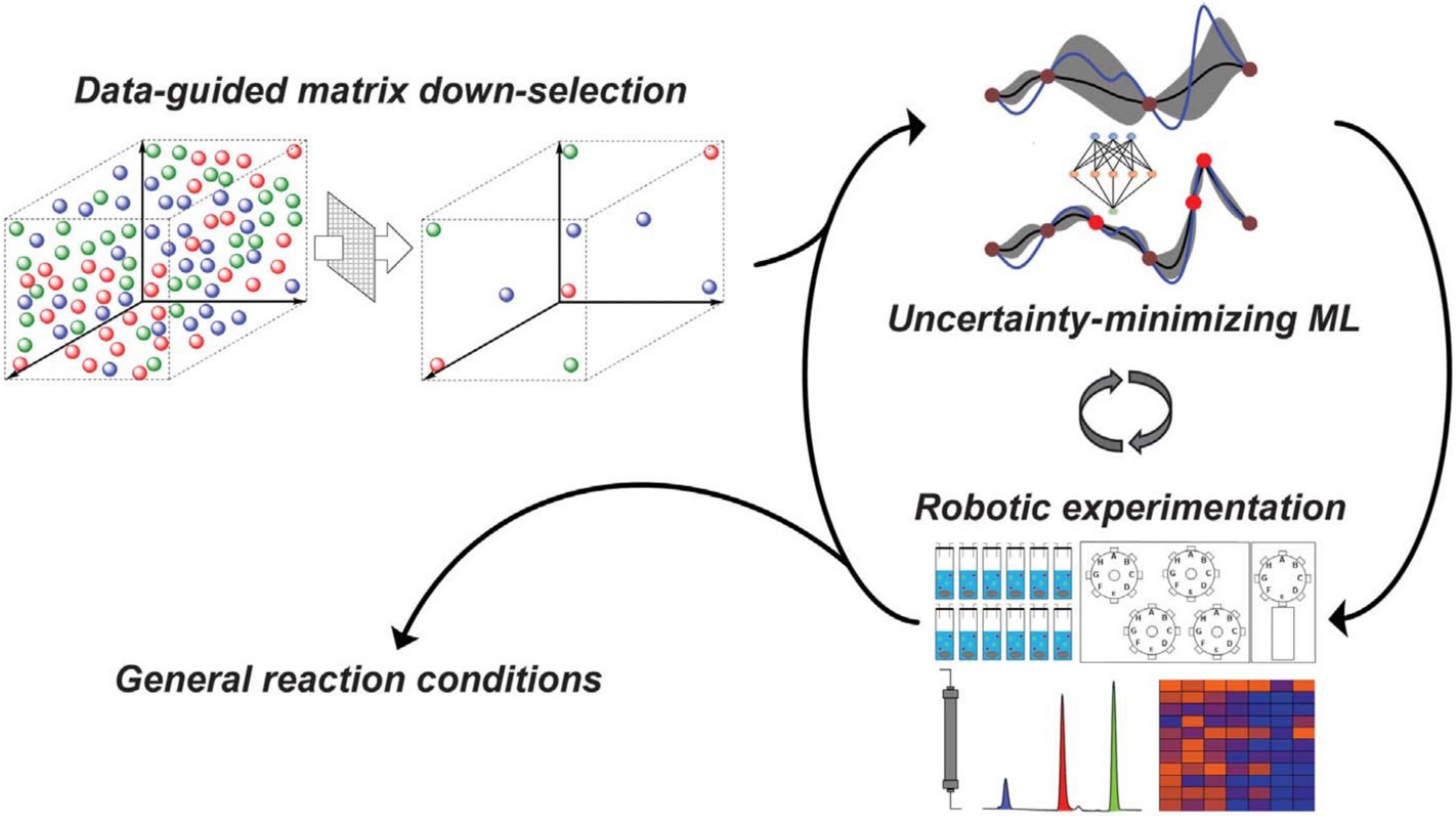
Bayesian optimization (BO)

✓ **An efficient and solid method to optimize reaction conditions**

R. P. Adams, A. G. Doyle, *et al.* *Nature* **2021**, *590*, 89–96.

Table of contents

Optimization of general reaction conditions

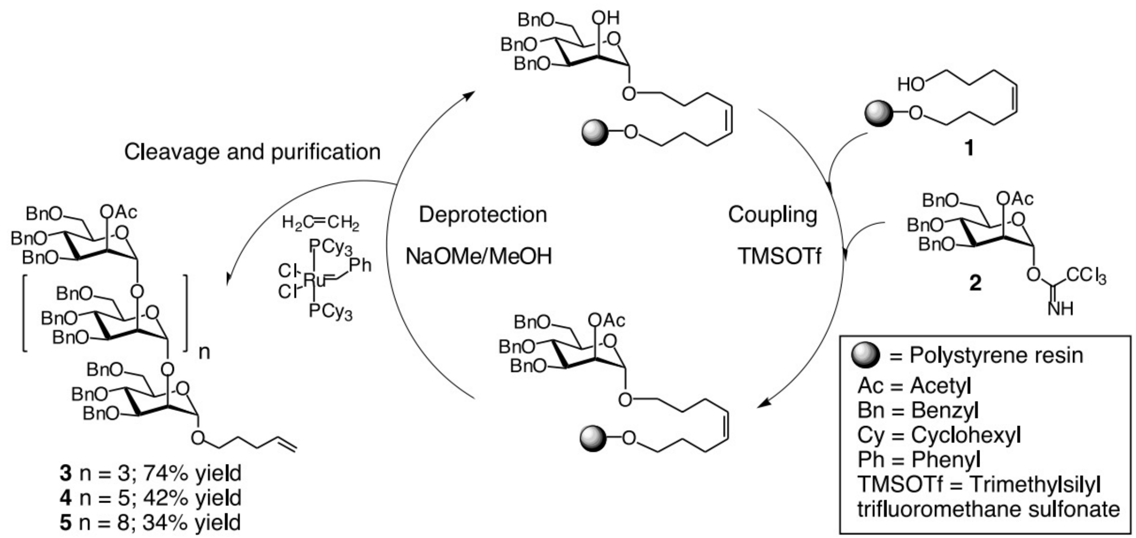


N. H. Angello, V. Rathore, W. Beker, A. Wołos, E. R. Jira, R. Roszak, T. C. Wu, C. M. Schroeder, A. Aspuru-Guzik, B. A. Grzybowski, M. D. Burke, *Science* **2022**, *378*, 399–405.

Importance of general reaction conditions

Automated synthesis methods for peptides, nucleic acids, and polysaccharides

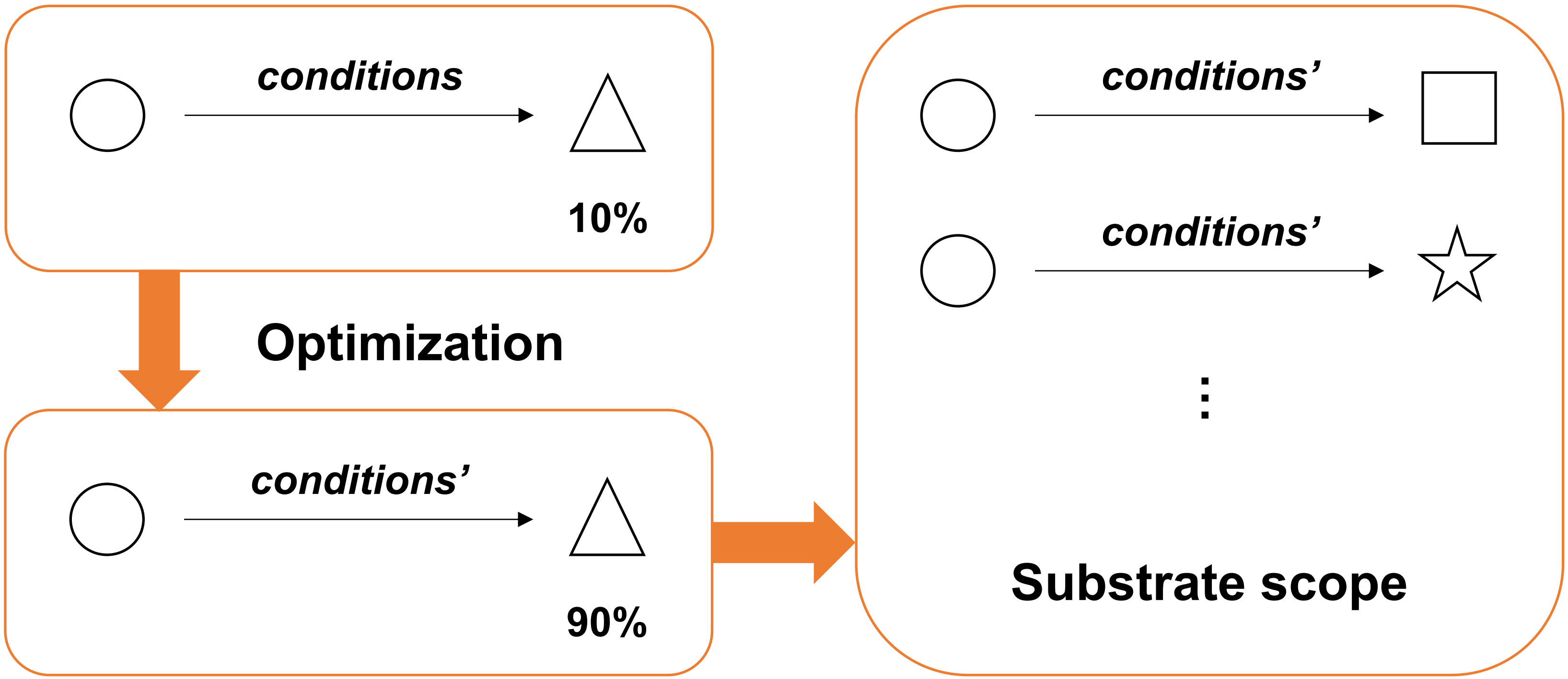
Highly general reaction conditions



Automated oligosaccharide synthesis

P. H. Seeberger, *et al. Science* **2001**, *291*, 1523–1527.

Reaction conditions for small organic synthesis



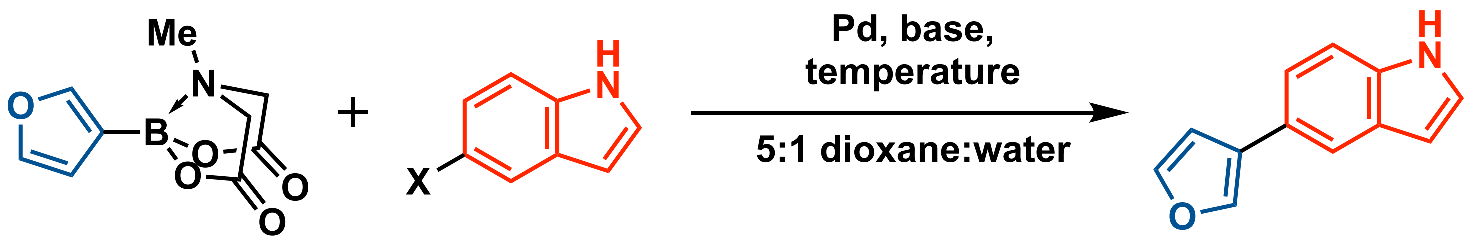
To identify general conditions:
all possible combinations of substrates
× all possible combinations of reaction conditions

- **Difficult to navigate via standard approaches**

B. A. Grzybowski, M. D. Burke, *et al. Science* **2022**, *378*, 399–405.

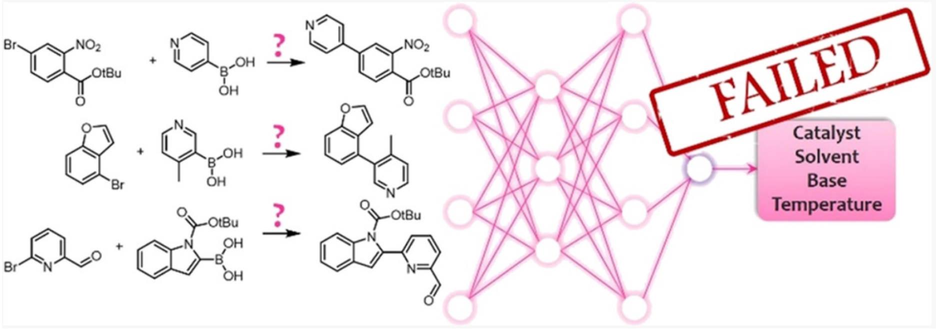
(Hetero)aryl Suzuki-Miyaura cross-coupling

A promising method to synthesize heteroaryl molecular fragments: SMC (Suzuki-Miyaura cross-coupling)



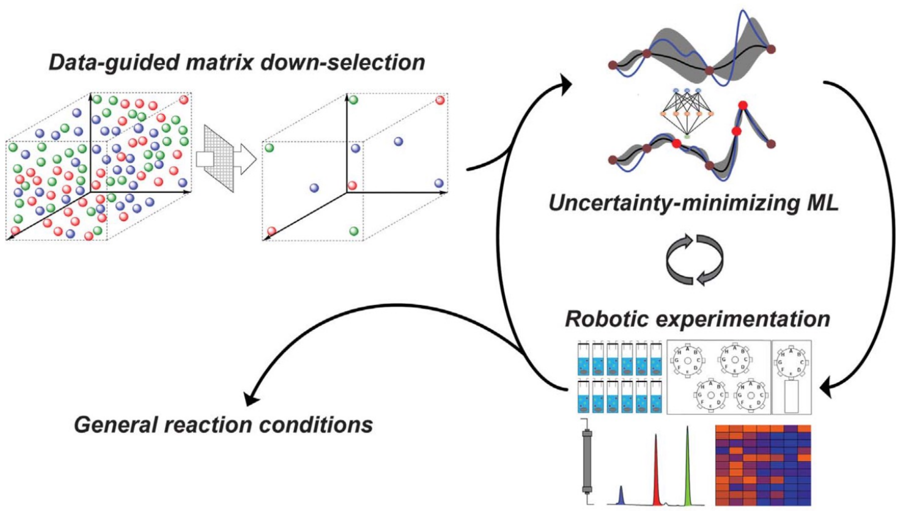
➢ **Finding general conditions**

- **This had been attempted, but failed, to discover them by mining the extensive chemical literature**



M. D. Burke, B. A. Grzybowski, *et al. J. Am. Chem. Soc.* **2022**, *144*, 4819–4827.

Overview of the method



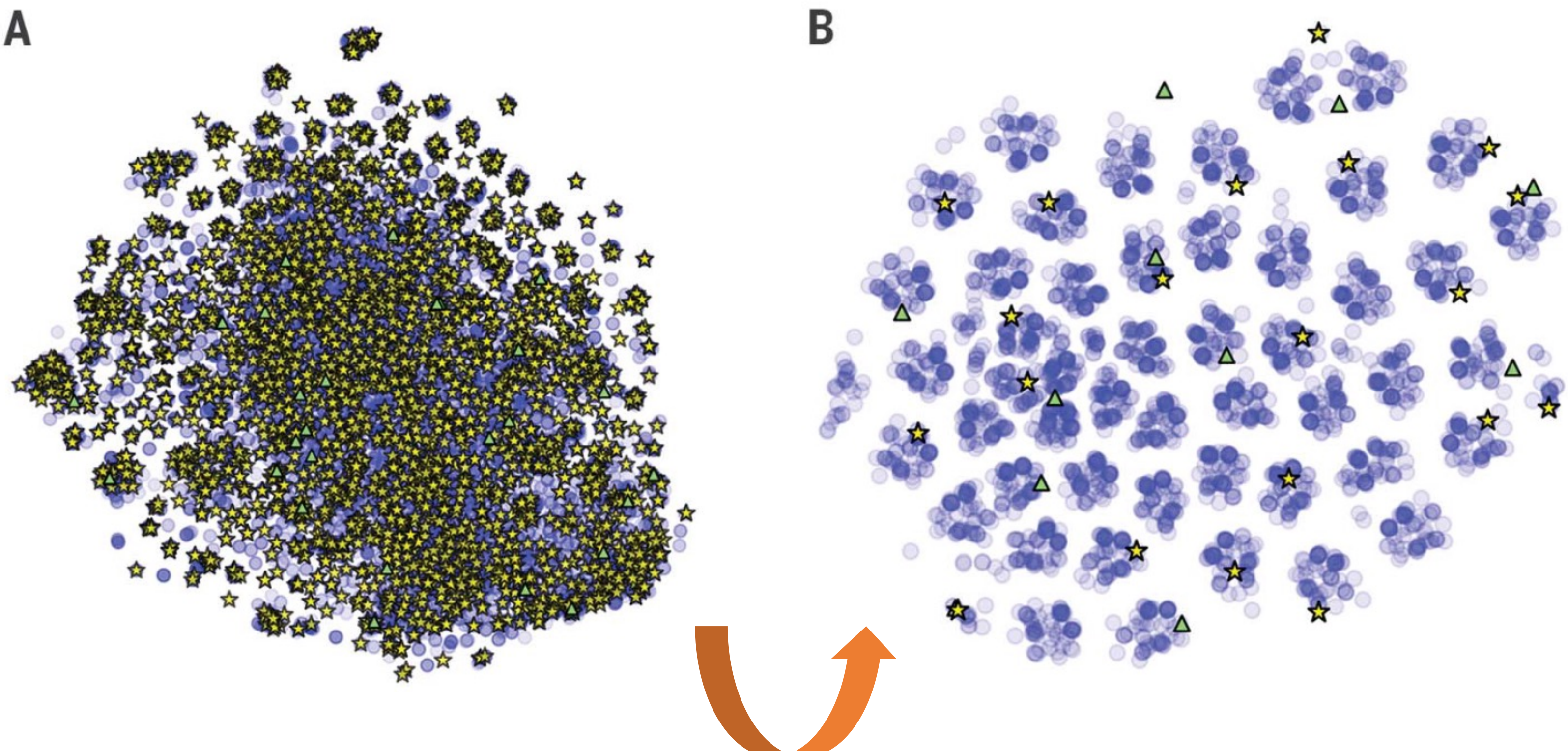
1. **Data-guided matrix down-selection** to render the vast search space tractable
2. **Modified BO** to efficiently drive prediction optimization
3. **Robotic experimentation** to increase throughput, precision, and reproducibility

➢ **Succeeds in identifying general reaction conditions!**

B. A. Grzybowski, M. D. Burke, *et al. Science* **2022**, *378*, 399–405.

Table of contents

Data-guided down-selection of substrates

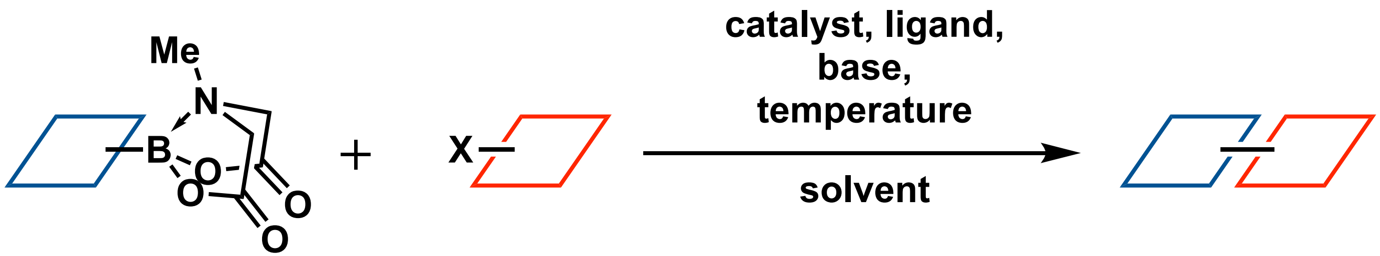


Potential product space compared to all previously reported heteroaryl products

Substrate pairs that maximize mutual dissimilarity of the resulting products

B. A. Grzybowski, M. D. Burke, *et al. Science* **2022**, *378*, 399–405.

Data-guided down-selection of conditions

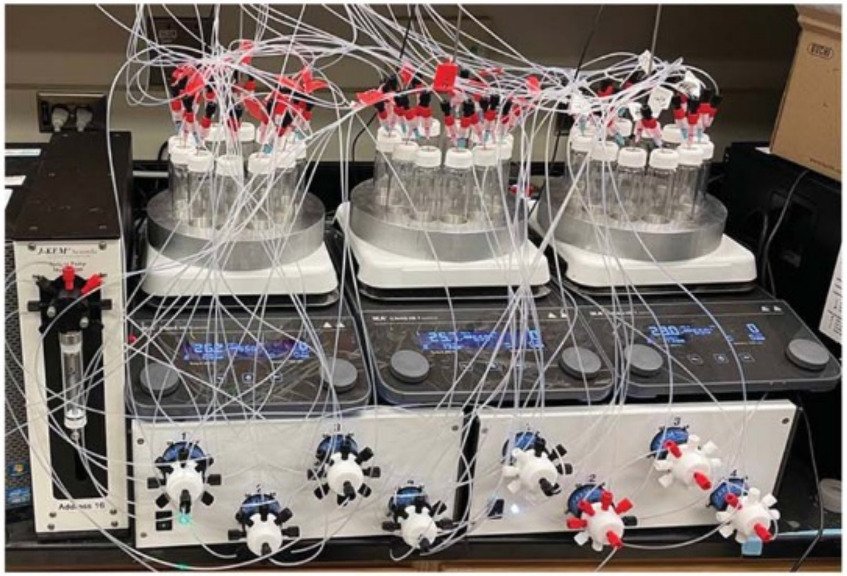


- extent of prior use
- structural diversity
- functional diversity

	selected conditions
catalysts	Pd SPhos G_4, $Pd(PPh_3)_4$,PdXPhosG_4, Pd $P(tBu)_3$ G_4, Pd PCy_3 G_4, $Pd_2(dba)_3$, and $Pd(dppf)Cl_2$
bases	sodium carbonate and potassium phosphate
temperatures	60° C and 100°C
solvents	dioxane, toluene, and dimethylformamide, all used in 5:1 mixture with water

B. A. Grzybowski, M. D. Burke, *et al. Science* **2022**, *378*, 399–405.

Robotic system for reaction performance

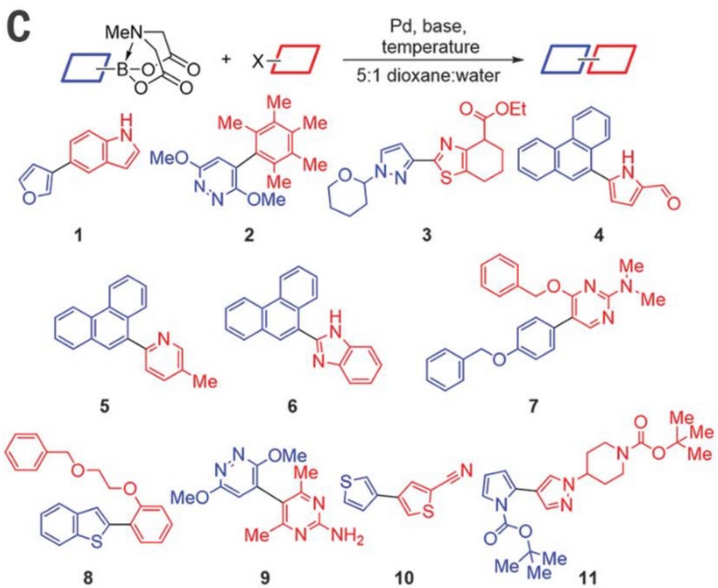


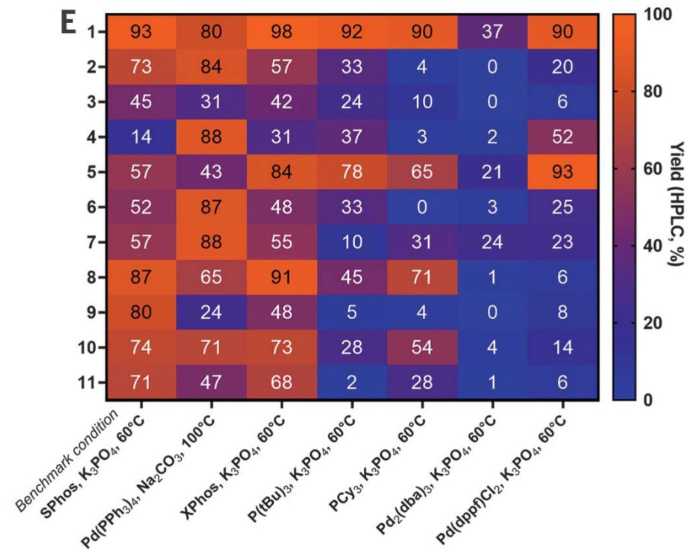
When each reaction was repeated twice, the yields exhibited only ±2% deviation.

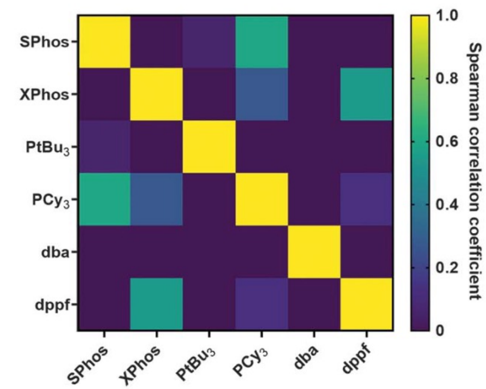
- **One of the key advantages of automated experimentation**

B. A. Grzybowski, M. D. Burke, *et al. Science* **2022**, *378*, 399–405.

Seeding experiments







Catalysts with

- **similar functions**
- **poor performance**

were eliminated.

528 reactions in total

B. A. Grzybowski, M. D. Burke, *et al.* *Science* **2022**, *378*, 399–405.

Uncertainty-minimizing ML for generality

- **Aim: to maximize the objective function f(c)**

$$f(c) = \frac{1}{|S|} \sum_{s \in S} y(s, c)$$

C={c}: the set of possible reaction conditions

S = {s}: a set of substrate pairs

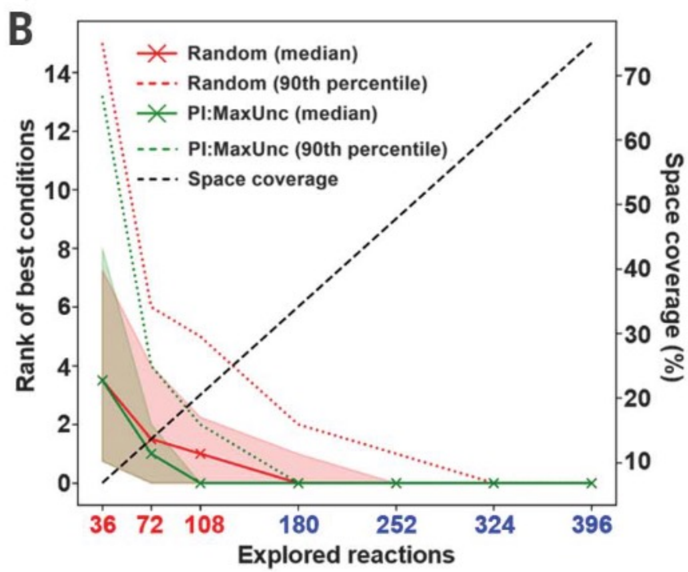
y(s,c): reaction yield

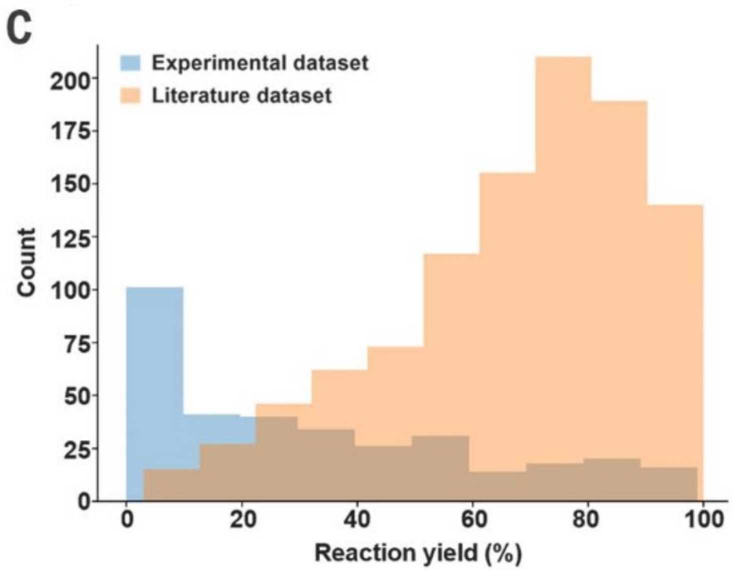
- **BO: each experiment performed immediately provides information about the objective function**
- **In this case: determination of f(c) for given conditions requires experiments with every pair of substrates in the S set**

B. A. Grzybowski, M. D. Burke, *et al. Science* **2022**, *378*, 399–405.

Table of contents

Advantages found for model-guided research

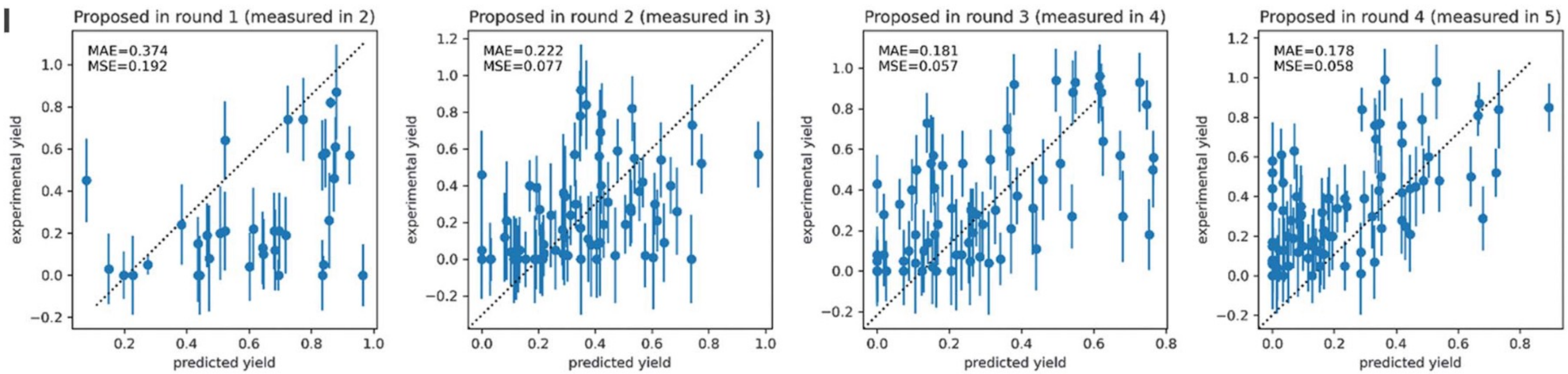




- ✓ **Apparent efficiency compared to random sampling**
- ✓ **Uniformly distributed yields over the range of possible values**

B. A. Grzybowski, M. D. Burke, *et al. Science* **2022**, *378*, 399–405.

Yields of reactions the model requested



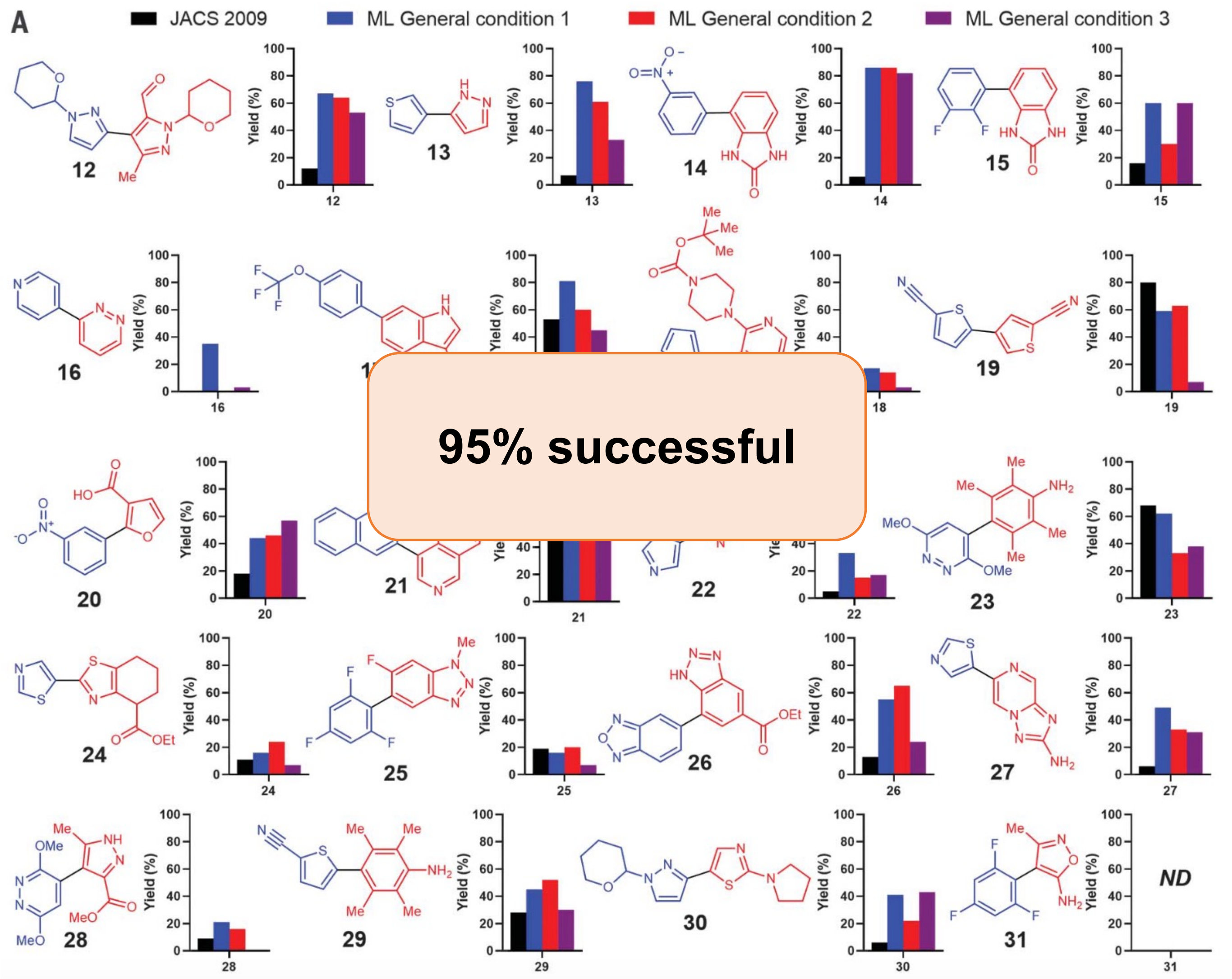
Exploration of good reactions in the second iteration

➢ **Attention shifted toward the “negative examples”**

Findings through the overall analysis:
(i) Relatively good candidate solutions were identified early
(ii) The model initially tried to look for better-yielding reactions
(iii) More and more attention was dedicated to decreasing the uncertainty of its estimates as the “loop” progressed

B. A. Grzybowski, M. D. Burke, *et al. Science* **2022**, *378*, 399–405.

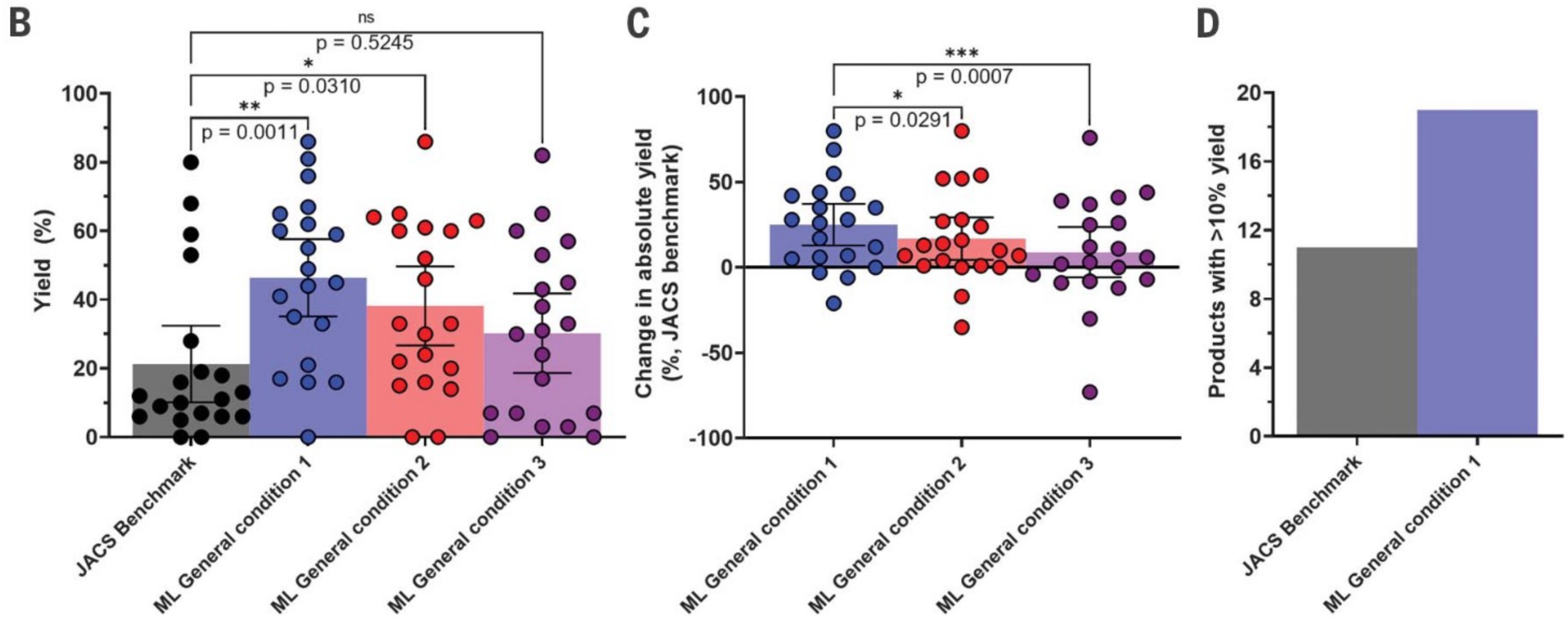
Quantifying generality



95% successful

B. A. Grzybowski, M. D. Burke, *et al. Science* **2022**, *378*, 399–405.

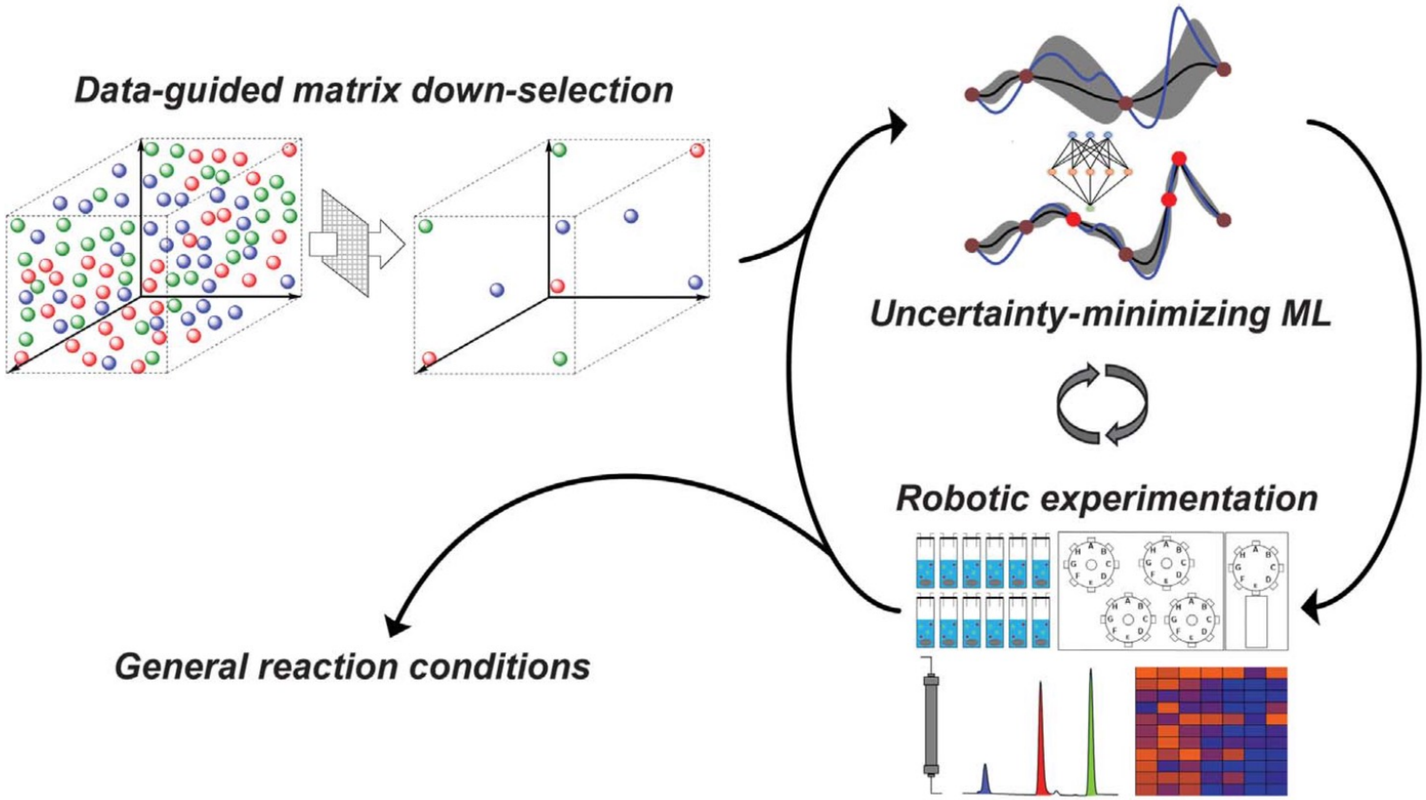
Comparison with the reported conditions



ML-discovered general reaction conditions performed substantially better.

B. A. Grzybowski, M. D. Burke, *et al.* *Science* **2022**, *378*, 399–405.

Short summary

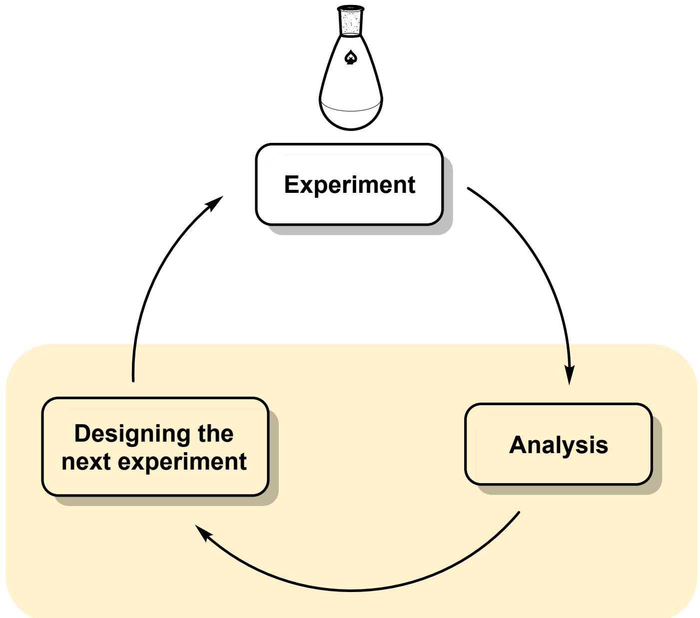


A nearly impossible challenge of finding general conditions was overcome by ML-assisted approach.

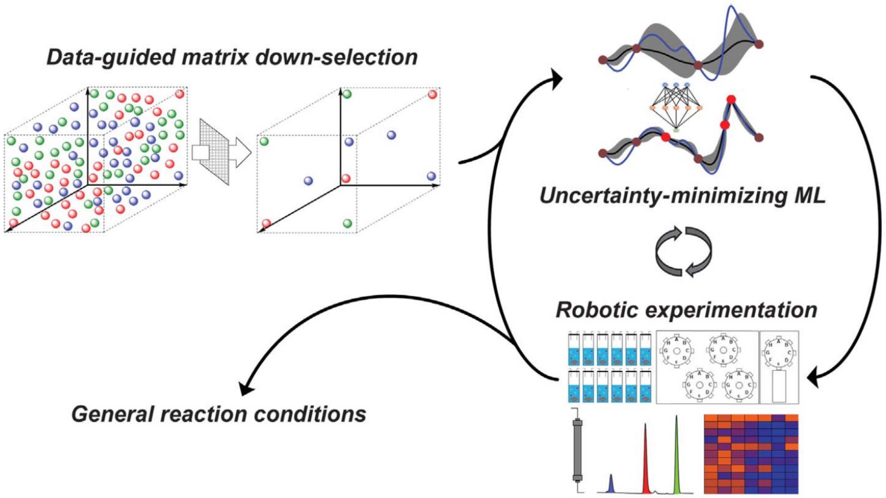
B. A. Grzybowski, M. D. Burke, *et al. Science* **2022**, *378*, 399–405.

Table of contents

Summary



Bayesian optimization (BO)



✓ **Optimization of the reaction conditions using certain substrates utilizing BO**

✓ **Further modified ML-assisted method to find general conditions applicable to various substrates**