Electronic Supplementary Information

Aravind Senthil Vel,* Julian Spils, Daniel Cortés-Borda and François-Xavier Felpin*

Nantes Université, CNRS, CEISAM, UMR 6230, F-44000 Nantes, France.

∗ Corresponding authors.
e-mail: aravind.senthilvel@univ-nantes.fr; ORCID: 0009-0008-6563-5652
e-mail: fx.felpin@univ-nantes.fr; ORCID: 0000-0002-8851-246X; Website: http://felpin.univ-nantes.fr/

Table of content

1. Maximum achievable throughput in the search space of [3+3] cycloaddition reaction optimization

In this section, the maximum achievable objective is explained using the optimization problem from our earlier work: the optimization of a [3+3] cycloaddition of 1,3-cyclohexanedione with citral in a flow reactor – slug flow. The optimization involved four continuous variables: temperature (T) [$25-50\ ^oC$], residence time ($\boldsymbol{\tau}$) [$1-10\ min$], reagent equivalent ($^{n_{eq}}$) [$1-2\ eq$], and catalyst loading ($^{cat_{load}}$) [$5-20\ mol\%$], along with one categorical variable consisting of six catalyst options. The objective was to maximize throughput (g/h), which is the amount of product formed per hour.

$$Throughput = \frac{Mass\ of\ product}{\tau} \tag{1}$$

$$Mass\ of\ product = C_{lim}^{inline} \times Yield \times V_{reactor} \times MW_{product} \tag{2}$$

where C_{lim}^{inline} is the inline concentration of the limiting reagent, $^{V_{reactor}(=5\ ml)}$ is the volume of reactor, and $^{MW_{product}(=246\ g/mol)}$ is the molecular weight of the product.

For the fixed total volume ($^{V_{total}=214\ \mu l}$), the inline concentration of the limiting reagent (refer Fig. S1) depends on $^{n_{eq}}$ and $^{cat_{load}}$ as described by following equations:

$$V_{total} = V_{lim} + V_{reag} + V_{cat} \tag{3}$$

$$n_{eq} = \frac{V_{reag} C_{reag}^{stock}}{V_{lim} C_{lim}^{stock}} \tag{4}$$

$$cat_{load} = \frac{V_{cat} C_{cat}^{stock}}{V_{lim} C_{lim}^{stock}} \times 100 \tag{5}$$

where $^{V_{lim}}$, $^{V_{reag}}$, and $^{V_{cat}}$ are the volumes of the limiting reagent, excess reagent, and catalyst. The stock concentrations are $C_{lim}^{stock} = 0.5\ M$, $C_{reag}^{stock} = 0.5\ M$, and $C_{cat}^{stock} = 0.05\ M$ corresponding to the limiting reagent, excess reagent, and catalyst respectively.

$^{V_{lim}}$ can then be calculated by substituting Equations (4) and (5) in (3).

$$V_{lim} = \frac{V_{total}}{\left[1 + C_{lim}^{stock}\left(\frac{n_{eq}}{C_{reag}^{stock}} + \frac{cat_{load}/100}{C_{cat}^{stock}}\right)\right]} \tag{6}$$

C_{lim}^{inline} can be then calculated as follows:

$$C_{lim}^{inline} = \frac{V_{lim} C_{lim}^{stock}}{V_{total}} \quad (7)$$

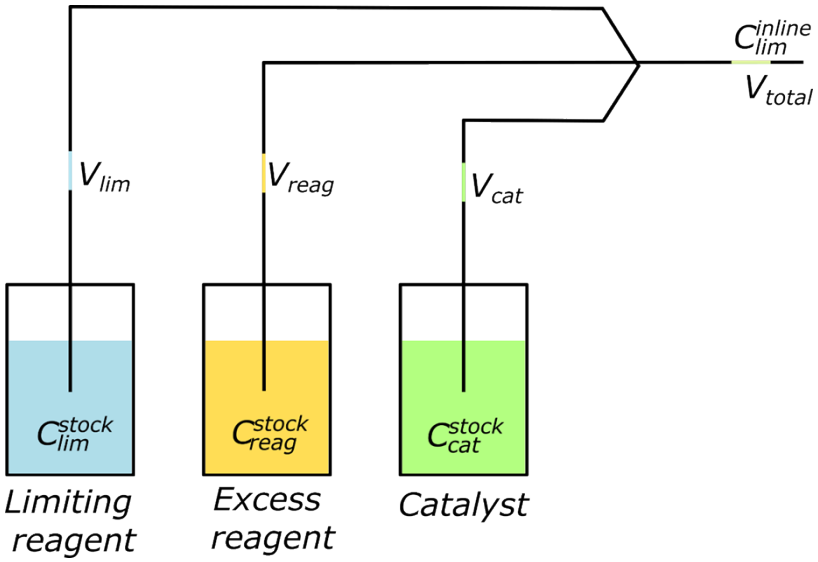


Fig. S1. Flow (slug) setup for the formal [3+3] cycloaddition reaction.

The throughput calculation incorporates the inline concentration of the limiting reagent (C_{lim}^{inline}), reactor volume ($V_{reactor}$), yield of the reaction, and the residence time ($\boldsymbol{\tau}$). While the yield depends on the nature of the reaction at specific conditions, and the reactor volume ($V_{reactor}$) is a fixed parameter, the remaining terms are determined by the chosen variables: $\boldsymbol{\tau}$ is directly specified, while C_{lim}^{inline} is based on n_{eq} and cat_{load}**.** Lower residence time ($\boldsymbol{\tau}$) favours higher throughput, as do lower equivalents of reagent (n_{eq}) and catalyst (cat_{load}). Overall, throughput depends on the yield of the reaction (which is unknown prior to conducting the reaction) and the chosen variables - residence time ($\boldsymbol{\tau}$), reagent equivalent (n_{eq}), and catalyst loading (cat_{load}) (which are known prior). For each experimental condition in the optimization space, the maximum achievable throughput can be calculated by assuming a yield of 100%.

$$Throughput = \frac{C_{lim}^{inline} \times [Yield = 100\%] \times V_{reactor} \times MW_{product}}{\tau} \quad (8)$$

Fig. S2.A depicts the maximum achievable throughput in the search space, represented solely by variables essential for calculating throughput - residence time ($\boldsymbol{\tau}$), reagent equivalent (n_{eq}), and catalyst loading (cat_{load}) - is represented in Fig. S2.A using residence time ($\boldsymbol{\tau}$) and mass of the product. The relationship between the mass of the product and the variables - reagent equivalent (n_{eq}), and catalyst loading (cat_{load}) - is represented Fig. S2.B.

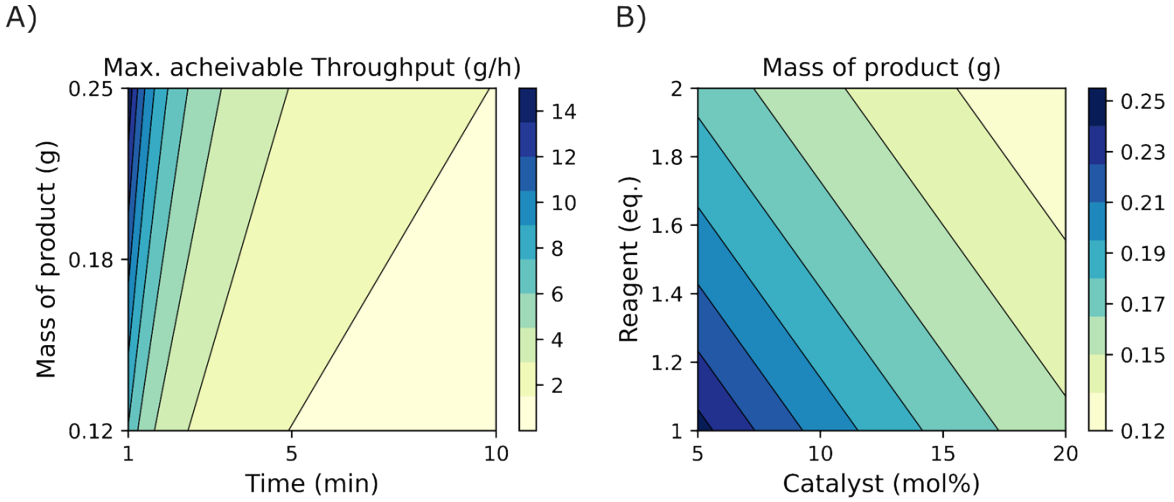


Fig. S2 A) Maximum achievable throughput for [3+3] – cycloadditions reaction - calculated assuming 100% yield within the search space. B) Relationship between mass of the product and the variables - reagent equivalent ($^{n_{eq}}$), and catalyst loading ($^{cat_{load}}$)

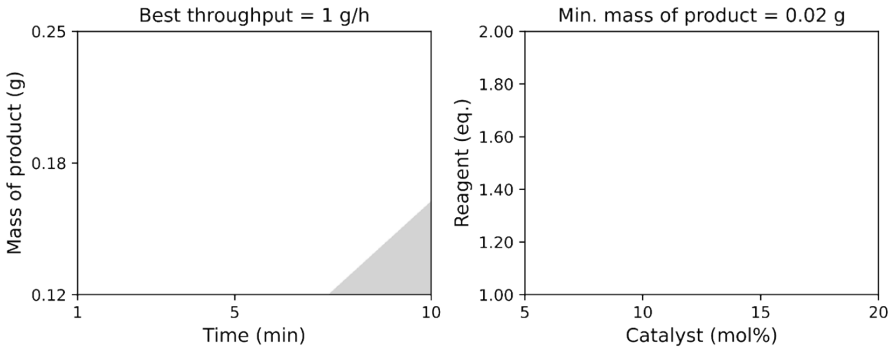
The plot clearly demonstrates the influence of variables on the objective function (throughput) calculation. Here, the maximum achievable throughput at higher residence times and low limiting reagent concentration (higher reagent equivalent ($^{n_{eq}}$) and catalyst loading ($^{cat_{load}}$)) is relatively lower compared to maximum achievable throughput at lower residence times ($\boldsymbol{\tau}$) and high limiting reagent concentration (lower reagent equivalent ($^{n_{eq}}$) and catalyst loading ($^{cat_{load}}$)). It is important not to assume that the optimal solution necessarily lies at low residence time ($\boldsymbol{\tau}$), reagent equivalent ($^{n_{eq}}$), and catalyst loading ($^{cat_{load}}$) since the maximum achievable throughput is calculated under the assumption that the yield is set at 100%. In reality, the reaction may not proceed efficiently under conditions that favour high maximum achievable throughput, so the actual global solution could be located anywhere within the search space.

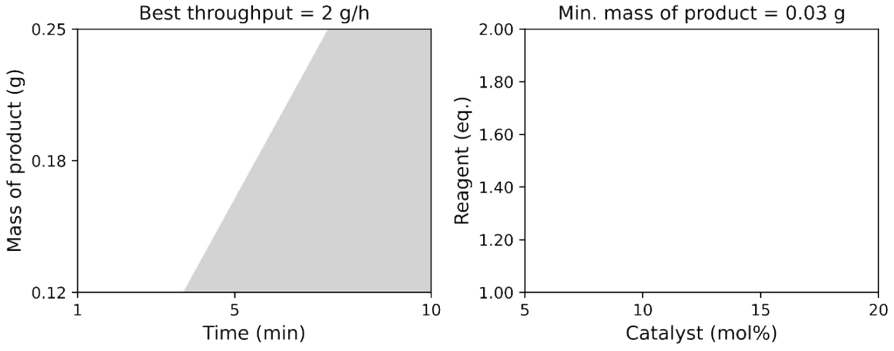
2. Futile region in the search space of [3+3] cycloaddition reaction optimization

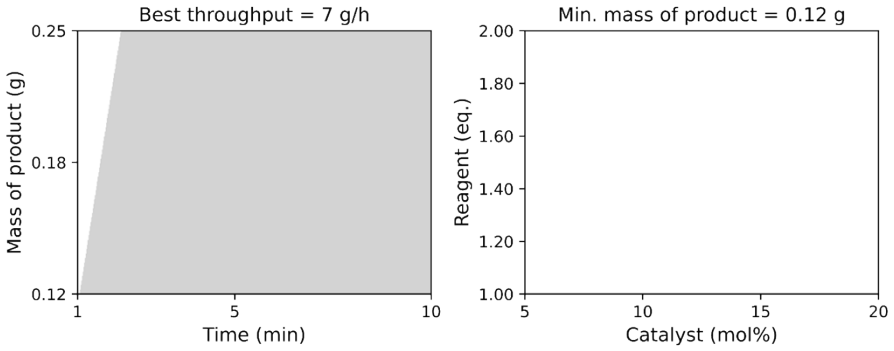
In the process of optimization, as the value of the best objective increases, the futile region within the search space also expands. This futile region represents the area where the maximum achievable objective is less than the current best objective value. The evolution of the futile space within the search space of the reaction optimization problem (mentioned earlier) is illustrated in Fig. S3, where the futile space is shown in grey. The remaining white space represents the promising search space where improvement in the objective function is theoretically possible.

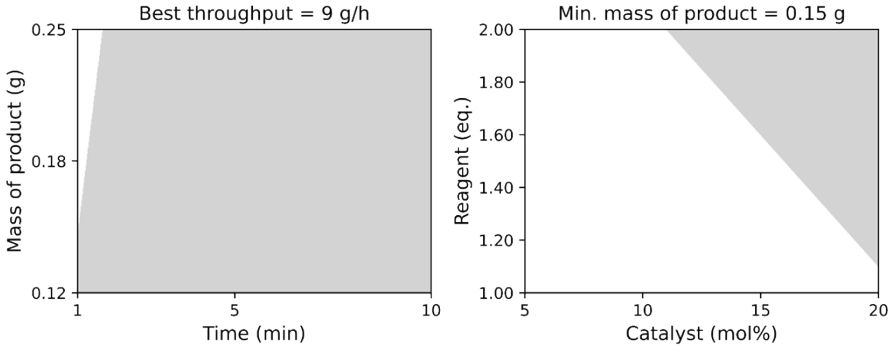
Table S1. Boundary limits of variables (involved in objective calculation) for different values of the best throughput

Best Throughput f^*	**Boundary Limit that ensures $f^{\max\, acheivable} > f^*$**		
	τ	n_{eq}	cat_{load}
(g/h)	(min)	(eq.)	(mol%)
1	14.8	35.39	348.9
2	7.4	16.95	164.5
3	4.9	10.80	103.0
4	3.7	7.72	72.2
5	3.0	5.88	53.8
6	2.5	4.65	41.5
7	2.1	3.77	32.7
8	1.8	3.11	26.1
9	1.6	2.60	21.0
10	1.5	2.19	16.9


Best throughput = 1 g/h
Min. mass of product = 0.02 g
Mass of product (g)
Time (min)
Reagent (eq.)
Catalyst (mol%)


Best throughput = 2 g/h
Min. mass of product = 0.03 g
Mass of product (g)
Time (min)
Reagent (eq.)
Catalyst (mol%)


Best throughput = 7 g/h
Min. mass of product = 0.12 g
Mass of product (g)
Time (min)
Reagent (eq.)
Catalyst (mol%)


Best throughput = 9 g/h
Min. mass of product = 0.15 g
Mass of product (g)
Time (min)
Reagent (eq.)
Catalyst (mol%)

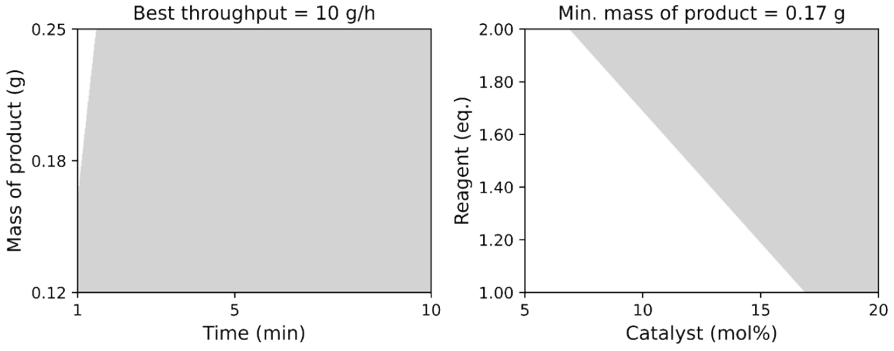

Best throughput = 10 g/h
Min. mass of product = 0.17 g
Mass of product (g)
Time (min)
Reagent (eq.)
Catalyst (mol%)

Fig. S3. Evolution of futile space within the search space for different values of the best throughput. The futile space is shown in grey, while the remaining white space represents the promising area where theoretical improvement from the best throughput value is possible. (Left) The search space is represented by the mass of the product and residence time (**τ**). (Right) The mass of the product is shown as a function of other variables - reagent equivalent (n_{eq}) and catalyst loading (cat_{load}).

It is worth noting that some of the boundaries of the variables (involved in objective calculation) become redundant as the optimization progresses. Removing such redundant boundaries from the optimization problem can reduce its complexity, thereby facilitating the optimization algorithm.

The boundary limit of a variable is determined based on the best objective value and the specified boundaries of the other variables that favour the objective value. For instance, boundary limit for residence time (τ^{lim}) is calculated (Equation 9) at the best throughput ($Throughput^*$) by assuming 100% yield and considering the lower limit of reagent equivalent (n_{eq}^{lb}) and the lower limit of catalyst loading (cat_{load}^{lb}). Similarly, the boundary limit for reagent equivalent (n_{eq}^{lim}) is calculated (Equation 10) by considering lower limit of residence time (τ^{lb}) and cat_{load}^{lb}, and the boundary limit for catalyst loading is calculated (Equation 11) using τ^{lb} and n_{eq}^{lb}.

$$\tau^{lim} = \left(\frac{1}{\left[\frac{1}{C_{lim}^{stock}} + \left(\frac{n_{eq}^{lb}}{C_{reag}^{stock}} + \frac{cat_{load}^{lb}/100}{C_{cat}^{stock}}\right)\right]}\right)\frac{[Yield = 100\%] \times V_{reactor} \times MW_{product}}{Throughput^*} \quad (9)$$

$$n_{eq}^{lim} = C_{reag}^{stock}\left(\frac{[Yield = 100\%] \times V_{reactor} \times MW_{product}}{Throughput^* \times \tau^{lb}} - \frac{1}{C_{lim}^{stock}} - \frac{cat_{load}^{lb}/100}{C_{cat}^{stock}}\right) \quad (10)$$

$$cat_{load}^{lim} =$$

$$C_{cat}^{stock} \times 100 \times \left(\frac{[Yield = 100\%] \times V_{reactor} \times MW_{product}}{Throughput^{*} \times \tau^{lb}} - \frac{1}{C_{lim}^{stock}} - \frac{n_{eq}^{lb}}{C_{reag}^{stock}} \right) \quad (11)$$

The boundary limits for the variables change with different best throughput value, is provided in Table S1. The boundary of the problem is updated only if the new boundary limit leads to a reduction in the search space. It can be observed that for lower values of the best throughput, the boundary limit exceeds the initially specified limit.

3. Handling constraint using Bayesian Optimization

When discussing the use of constraints in expensive black-box problems like reaction optimization, they can generally be classified into two types:[1]

3.1. Known constraints (constraints involving variables)

These constraints are pre-defined and can be assessed for compliance before conducting the experiment. For example, a known constraint may require that the volume of a two-reagent mixture should not exceed a specified limit. In this case, the volumes of the reagent mixtures are the variables, and it can be determined prior to the experiment whether this constraint will be violated.

3.2. Unknown constraints (constraints involving objectives)

These constraints relate to outcomes that cannot be confirmed until after the experiment is conducted. An example of this is optimizing the throughput of a reaction where the yield must exceed a certain threshold, but the actual yield is not known until the reaction is complete.

Although Bayesian optimization can address both types of constraints, handling unknown constraints is more challenging than dealing with known constraints. The constraint in our case comes under known constraints, as the maximum achievable objective value is calculated using the variables assuming 100% yield, and their satisfaction can be checked without performing experiment.

Bayesian optimization operates through two main steps: generating a surrogate model and utilizing an acquisition function to predict the optimum point. The surrogate model is a probabilistic model that serves as a cost-effective substitute for the actual model. Through the

acquisition function, each point in the variable space is scored, with the point receiving the highest score identified as the most promising for subsequent experimentation.

Different acquisition functions prioritize and score points in variable ways. For instance, the Expected Improvement function considers both the prediction and the uncertainty value. It strikes a balance between exploitation (searching near the best-known points) and exploration (searching in untested regions). The trade-off between these two can be adjusted by setting the exploration factor; a higher value leads to more exploration.

Identifying the point with the highest acquisition score presents another optimization challenge. Ideally, one could evaluate the score for all points in the search space and select the best one, but this is computationally intensive. Therefore, this selection is usually treated as an optimization problem and solved using efficient algorithms like gradient descent. Since this algorithm is a local search method, it is typically executed multiple times from different starting points to avoid local minima. Here, known constraints are incorporated into the acquisition function optimization step. The goal is to identify the point with the optimal score that simultaneously meets these constraints. Removing redundant boundaries helps in reducing the complexity of this acquisition function optimization problem.

4. Solvers

4.1. Bayesopt (MATLAB)

Bayesopt is the MATLAB's build-in function, by default, minimizes black-box problems. It utilizes a Gaussian Process (GP) as the surrogate model and employs the ARD Matern 5/2 kernel as the covariance function. It can handle continuous, integer, and categorical variables. Bayesopt offers six acquisition function options:

- Expected Improvement per Second Plus
- Expected Improvement
- Expected Improvement Plus
- Expected Improvement per Second
- Lower Confidence Bound
- Probability of Improvement

For the "Expected Improvement per Second Plus" and "Expected Improvement Plus" functions, the exploration factor can be specified (default value: 0.5). For acquisition function

optimization, the best points among the feasible points (those that satisfy constraints) are optimized using a local search method.

For simulation studies, "Expected Improvement per Second Plus" and "Expected Improvement per Second" were excluded because they incorporate the evaluation time of the objective function in the acquisition process which is not in the interest of our problem. Apart from the selection of the acquisition function and the exploration factor, the solver was run using its default settings. More details on the algorithm workflow, surrogate model, and acquisition functions, and settings can be found at: https://in.mathworks.com/help/stats/bayesian-optimization-algorithm.html

The code for initializing and running Bayesopt, both in standard BO and ABC-BO modes, is available on GitHub (https://github.com/Aravind-vel/ABC_BO). This code is particularly useful when the constraints involve discrete numeric variables. Bayesopt cannot handle discrete numeric variables directly; they need to be treated as integer variables. However, in the provided code, variables can be initiated as discrete numeric directly, simplifying the process. Furthermore, the usage of the code is demonstrated with examples provided for solving both *in silico* and practical experimental optimization problems.

Probability of Improvement

The classical version of this acquisition function chooses the point that has the highest probability of improvement over the current best objective value.

$$PoI(x) = \Phi\left(\frac{\mu(x) - \mu(x^*)}{\sigma(x)}\right)$$

where, $\Phi(.)$ is the normal cumulative distribution function, $\mu(x)$ is the mean prediction of the surrogate model of point x, $\sigma(x)$ is the standard deviation associated with the prediction model at point x, $\mu(x^*)$ is the mean prediction at the point correspond to best objective f^*.

This version has the disadvantage of over exploitation (stuck in local maxima). To counteract this, an exploration factor (ξ) can be added.

$$PoI(x) = \Phi\left(\frac{\mu(x) - \mu(x^*) - \xi}{\sigma(x)}\right)$$

In Matlab's Bayesopt, the exploration factor (ξ) by default takes the value of the estimated noise deviation (σ)

Expected Improvement

This acquisition function calculates the expected value of the improvement which takes in to account the magnitude of improvement along with probability of improvement.

$$EI(x) = \left(\mu(x) - \mu(x^*) - \xi\right)\Phi\left(\frac{\mu(x) - \mu(x^*) - \xi}{\sigma(x)}\right) + \sigma(x)\varphi\left(\frac{\mu(x) - \mu(x^*) - \xi}{\sigma(x)}\right)$$

where, $\varphi(.)$ is the normal probability density function.

In Matlab's Bayesopt, by default, the exploration factor (ξ) takes the value 0.

Expected Improvement plus

This acquisition function is the extended version of expected improvement. When the acquisition function finds the point that is overexploiting ($\frac{\sigma(x)}{\sigma} < \xi$), the algorithm adjusts the parameters of the kernel function to increase the variance between points. This adjustment encourages exploration, and the acquisition function continues searching for new points until it finds one that is not overexploiting.

The exploration factor ξ can be specified. In this study, we tested two values (0.5, 1).

Lower Confidence Bound (Upper Confidence Bound)

This acquisition function represents a curve that is n standard deviations away from the mean prediction. For maximization problems, this curve is above the mean prediction and is called the Upper Confidence Bound (UCB). For minimization problems, the curve is below the mean prediction, hence referred to as the Lower Confidence Bound (LCB).

Since MATLAB's Bayesopt minimizes the objective function by default, this acquisition function is referred to as the lower confidence bound. However, as all of the objectives in this study involve maximization, we refer to it as the upper confidence bound.

$$UCB(x) = \mu(x) + n\sigma(x)$$

n value by default takes 2 (for Bayesopt).

The optimistic approach of selecting the next point could be a reason for its relatively lower number of futile experiments compared to other acquisition functions in standard BO.

4.2. Dragonfly (python)

Dragonfly[2] is a Bayesian Optimization (BO) package (https://github.com/dragonfly/dragonfly) capable of handling a wide range of variable types, including continuous (with specified

boundaries), discrete numeric, categorical, integer, and discrete Euclidean. It employs a Gaussian Process (GP) as the surrogate model and uses the Matern-5/2 kernel for continuous and discrete numeric variables and the Hamming kernel for categorical variables. Unlike the general approach of selecting a specific acquisition function, Dragonfly uses an adaptive strategy where the acquisition function is randomly chosen from the following options: Upper Confidence Bound, Expected Improvement, Thompson Sampling, Top-two Expected Improvement[3].

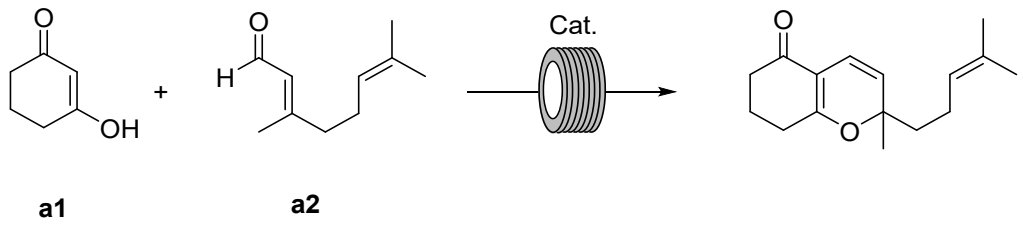
By default, the algorithm favours the acquisition function that provides improvement. However, in our case, due to the difficulty of running the optimization in a single stretch and the unavailability of an option to save information on the favoured acquisition function, we treated each experiment as a new one. Consequently, the choice of acquisition function was fully random. The solver was run using its default settings.

For purely continuous search spaces without constraints, dragonfly uses either the DiRect or PDOO algorithms for acquisition function optimization. In all other cases - such as when discrete numeric, categorical variables, or constraints are involved - dragonfly employs an evolutionary algorithm.

5. *In silico* reactions

5.1. ISR-1

This data-based *in silico* reaction involves the [3+3] cycloaddition of 1,3-cyclohexanedione **a1** with citral **a2** (**Scheme** S1), as mentioned earlier. It is derived from practical experimental work featured in our prior publications.[4,5] For prediction, we utilized a Gaussian Process (GP) model that incorporates data from both our published and unpublished studies. We have also used this model in our earlier work comparing multi-objective optimization solvers, where it was referred to as Cycloadditions-1 and Cycloadditions-2, with yield and throughput as the objectives.[6] In this work, since we are testing single-objective solvers, we optimized throughput. In ISR-1, we consider only continuous variables: temperature (T) [$25-50\ ^{o}C$], residence time ($\boldsymbol{\tau}$) [$1-10\ min$], reagent **a2** equivalent (n_{eq}) [$1-2\ eq.$], and catalyst loading (cat_{load}) [$5-20\ mol\%$]. The other details essential for calculating the objective function are as follows: the injection volume is 214 µl; **a1** has a stock concentration of 0.5 M; **a2** has a stock concentration of 0.5 M; the catalyst, ethanolamine, has a stock concentration of 0.05 M; the reactor volume is 5 ml; and the molecular weight of the product is 246 g/mol.



Scheme S1. 1,3-cyclohexanone reacts with citral.

5.2. ISR-2

ISR-2 is an extension of ISR-1, where, in addition to the four continuous variables, a categorical variable - catalyst with five choices (ethanolamine, pyrrolidine, ethylenediamine, butylamine, and piperidine) - is included in the optimization problem.

5.3. ISR-3

ISR-3 is a kinetic model based in silico reaction (Fig. S4), introduced by Reizman[7] encompassing five distinct case studies. The optimization involves the maximization of Turnover Number (TON), defined as the ratio between the product concentration and the catalyst concentration, and includes three continuous variables: residence time $\boldsymbol{\tau}$ [1-10 min], temperature $\boldsymbol{T}$ [30-110 °C], catalyst loading cat_{load} [0.5 -2.5 mol%], and one categorical variable catalyst [1,2,3,4,5,6,7,8]. This model has served as a benchmark in previous studies.[6,8,9]

$$A + B \xrightarrow{k_R} R \qquad k_R = C_{cat}^{\frac{1}{2}} A_R e^{\frac{-(E_{a_R}+E_{a_i})}{RT}}$$

$$B \xrightarrow{k_{S_1}} S_1 \qquad k_{S_1} = A_{S_1} e^{\frac{-E_{a_{S_1}}}{RT}}$$

$$B + R \xrightarrow{k_{S_2}} S_2 \qquad k_{S_2} = A_{S_2} e^{\frac{-E_{a_{S_2}}}{RT}}$$

where $A_R = 3.1 \times 10^7 L^{0.5} mol^{-1.5} s^{-1}, E_{a_R} = 55\ KJ\ mol^{-1}$

E_{a_i} values of the catalyst for specific case study

Catalyst (i)	Case 1	Case 2	Case 3/4	Case 5
1 (T<80°C)	0	0	0	-5.0
1 (T>80°C)	0	0	0	-5.0+0.3(T-80)
2	0.3	0	0.3	0.7
3	0.3	0.3	0.3	0.7
4	0.7	0.7	0.7	0.7
5	0.7	0.7	0.7	0.7
6	2.2	2.2	2.2	2.2
7	3.8	3.8	3.8	3.8
8	7.3	7.3	7.3	7.3

Fig S4. Reaction scheme and kinetic details of the *in silico* reaction ISR-3.

In our study, we considered only Case 1 out of the five case studies. The kinetic parameters associated with the different case studies are provided in Table S2.

Table S2. Overview kinetic parameters for different case studies

Case	Catalyst effect	k_{S_1}	k_{S_2}
1	$E_{A_1} > E_{A_{2-8}}$	= 0	= 0
2	$E_{A_1} = E_{A_2} > E_{A_{3-8}}$	= 0	= 0
3	$E_{A_1} > E_{A_{2-8}}$	> 0[a]	= 0
4	$E_{A_1} > E_{A_{2-8}}$	= 0	> 0[b]
5	$E_{A_1} > E_{A_{2-8}}$	= 0	= 0

a: $A_{S_1} = 1\ x\ 10^{12} s^{-1}, E_{a_{S_1}} = 100\ KJ\ mol^{-1}$

b: $A_R = 3.1\ x\ 10^5 L^{0.5} mol^{-1.5} s^{-1}, E_{a_{S_2}} = 50\ KJ\ mol^{-1}$

5.4. Global solutions

For reference, global solution (Table S3) for the *in silico* reactions (ISR-1,2 and 3) are identified by extensive grid search.

Table S3. Grid search parameters and corresponding global solutions identified for the *in silico* reactions.

in silico reaction	Grid search parameters	Global solution
ISR-1	**Continuous variables:** Temperature = [25:1:50] ^oC Residence time = [1:0.1:10] min Reagent equivalent = [1:0.01:2] $eq.$ Catalyst loading = [5:0.1:20] $mol\%$	Temperature = 30 ^oC Residence time = 1 min Reagent equivalent = 1 $eq.$ Catalyst loading = 5 $mol\%$ Throughput = 10.14 g/h
ISR-2	**Continuous variables:** Temperature = [25:1:50] ^oC Residence time = [1:0.1:10] min Reagent equivalent = [1:0.01:2] $eq.$ Catalyst loading = [5:0.1:20] $mol\%$ **Categorical variables:** Catalyst = [1, 2, 3, 4, 5]	Temperature = 30 ^oC Residence time = 1 min Reagent equivalent = 1 $eq.$ Catalyst loading = 5 $mol\%$ Catalyst = 3 Throughput = 12.52 g/h
ISR-3	**Continuous variables:** Temperature = [30:1:110] ^oC Residence time = [1:0.1:10] min	Temperature = 110 ^oC Residence time = 10 min Catalyst loading = 0.5 $mol\%$

	Catalyst loading = [5:0.01:2.5] $mol\%$	Catalyst = 1
	Categorical variables:	TON = 180.54
	Catalyst = [1, 2, 3, 4, 5,6,7,8]	

6. *In silico* simulation results

The *in silico* problems were tested using five different acquisition functions in Bayesopt and Dragonfly. Each problem was run for 21 iterations with a total budget of 30 experiments. This budget was chosen to balance the practical computational time with having enough samples to allow for statistical evaluation. The initialization phase involves center point sampling: for ISR-1, this is a single experiment, and for ISR-2, it involves 5 experiments (center point in each level of the categorical variable).

6.1. ISR-1 and ISR-2

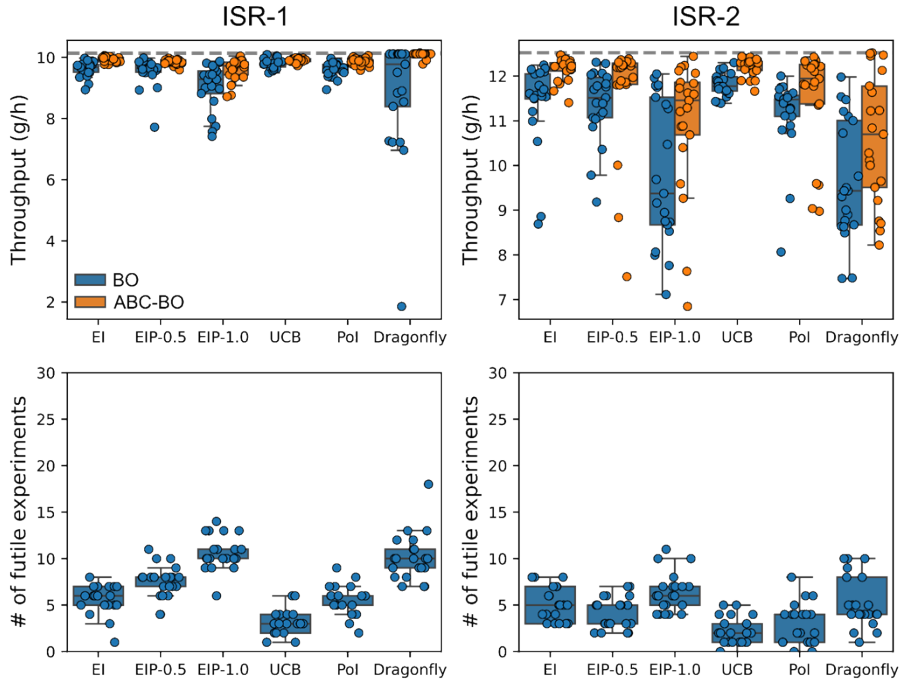


Fig. S5. The top row displays the best objective values at the end of the 20th experiment for ISR-1(left) and ISR-2 (right), with ABC-BO results in orange and BO results in blue. The grey dotted line represents the global solution for the corresponding *in silico* reaction. The bottom row shows the number of futile experiments at the end of the 20th experiment using BO.

In ABC-BO, Figs. S6 and S7 show how variable boundaries are updated after each iteration based on the best results obtained. The upper limit of the 'time' variable decreases from an initial 10 minutes to nearly 1 minute, faster than other variables such as reagent eq. and catalyst

loading. This highlights the critical role of residence time in optimizing throughput. Furthermore, the boundary adjustments are tied to the best objective values achieved; for instance, while the reagent equivalent limit stays the same in ISR-1 with a best throughput near 10 g/h, it decreases in ISR-2 where the throughput reaches 12 g/h.

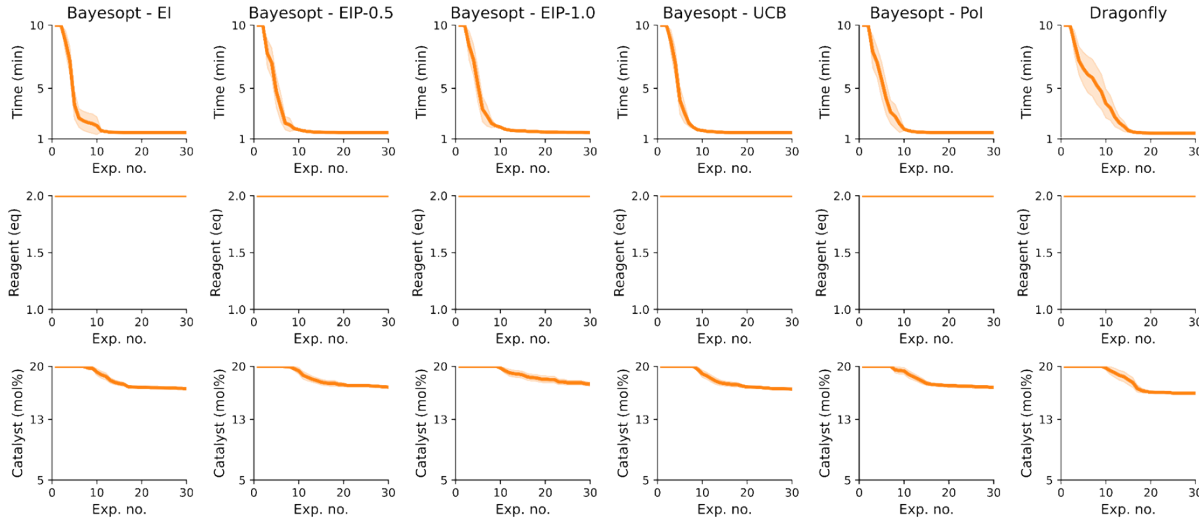


Fig. S6. Boundary reduction of influencing variables (variables involved in objective calculation) for ISR-1 during optimization using ABC-BO. The orange line represents the average, while the orange shaded area indicates the 95% confidence interval.

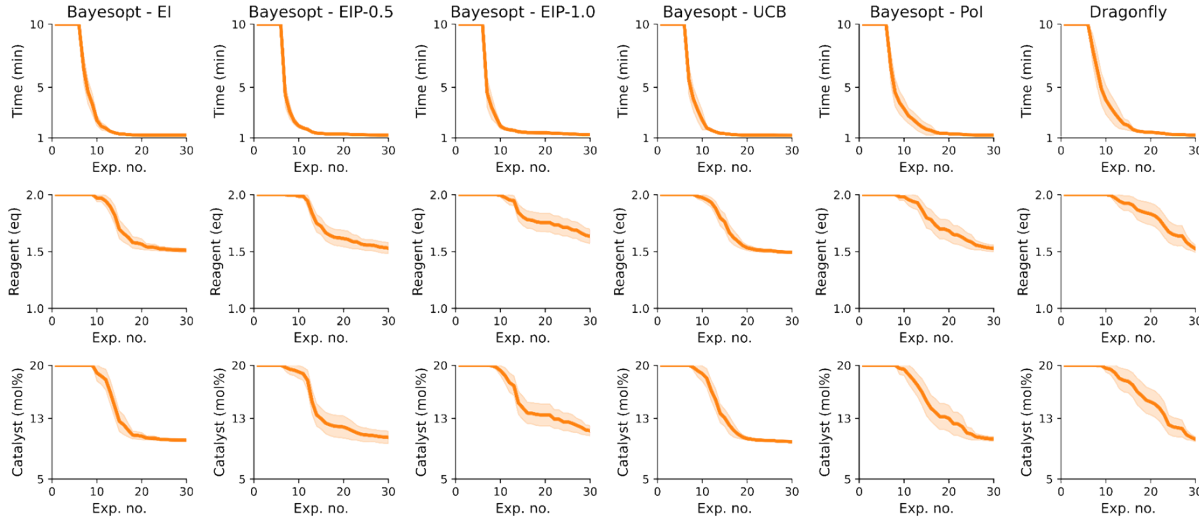


Fig. S7. Boundary reduction of influencing variables (variables involved in objective calculation) for ISR-2 during optimization using ABC-BO. The orange line represents the average, while the orange shaded area indicates the 95% confidence interval.

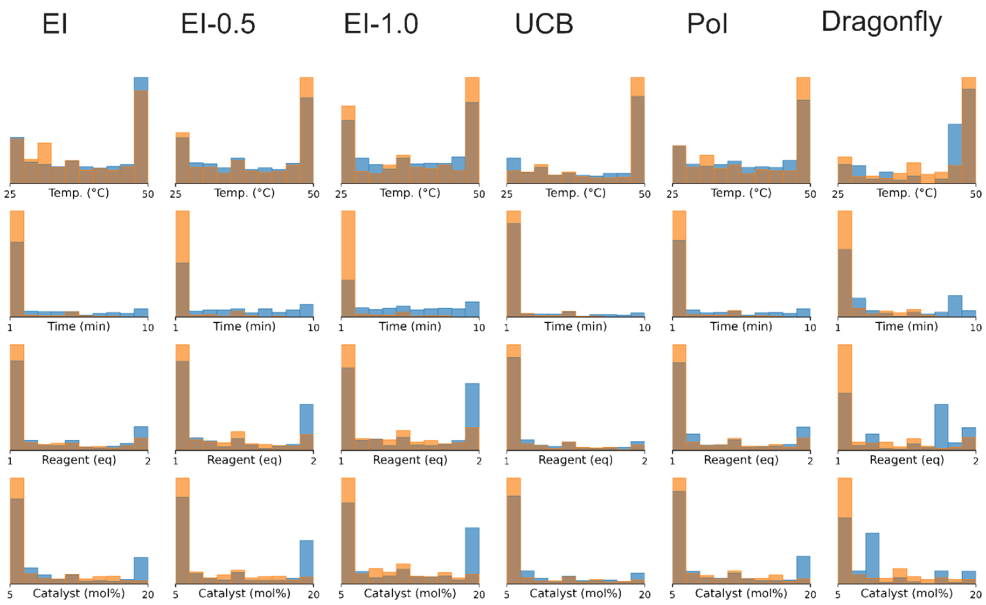


Fig. S8. Distribution of variable selection (histogram plots) in optimization for ABC-BO (orange) and BO (blue) in ISR-1

6.2. ISR-3

In the ISR-3 problem, the variable influencing TON calculation is catalyst loading. After each experiment, the boundary for this variable is updated, effectively making the constraint redundant. In other words, once the boundary is narrowed, all values of catalyst loading will naturally satisfy the constraint (refer Fig. S9). As a result, for this problem, we focused solely on reducing the boundary of catalyst loading and performed BO.

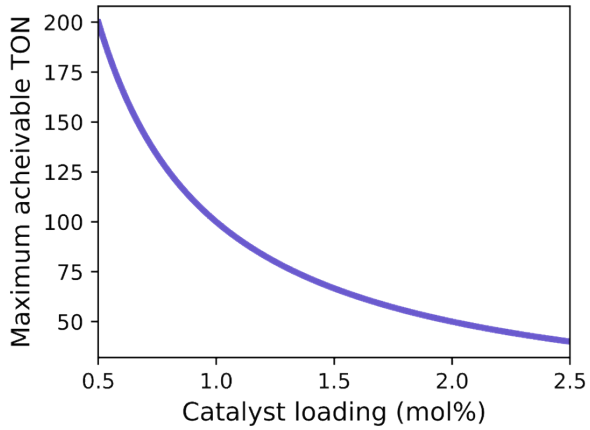


Fig. S9. Maximum achievable TON (calculated assuming 100% yield) for ISR-3, represented as a function of the influencing variable, catalyst loading.

From the simulation results (Fig. S10), it can be observed that ISR-3 exhibits a relatively lower number of futile evaluations compared to ISR-1 and ISR-2. This is likely because, in ISR-3,

only one variable (catalyst loading) influences the objective function calculation, leading to a comparatively smaller futile space than in the other problems, where three variables affect the objective calculation. Consequently, the performance improvement of ABC-BO over BO in the ISR-3 problem is less significant than in ISR-1 and ISR-2.

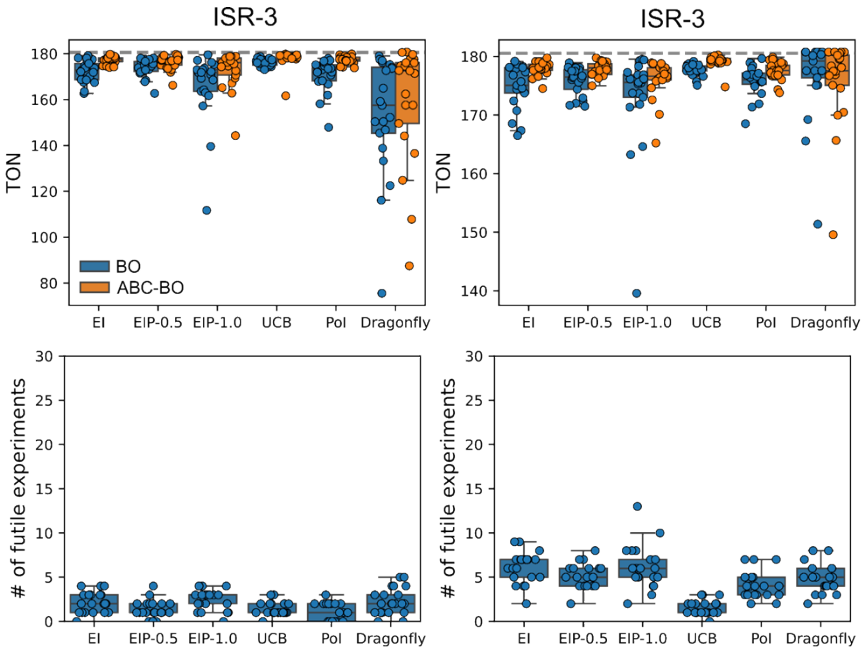


Fig. S10. The top row displays the best objective values at the end of the 20th experiment (left) and 30th experiment (right) for ISR-3, with ABC-BO results in orange and BO results in blue. The grey dotted line represents the global solution for the reaction. The bottom row shows the number of futile experiments using BO.

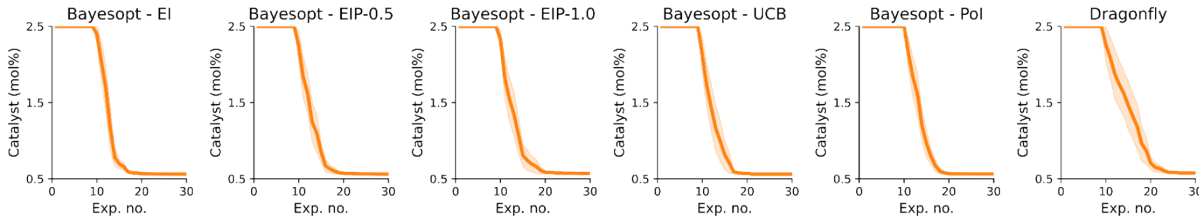


Fig. S11. Boundary reduction of the influencing variable (variable involved in objective calculation) for ISR-3 during optimization using ABC BO. The orange line represents the average, while the orange shaded area indicates the 95% confidence interval.

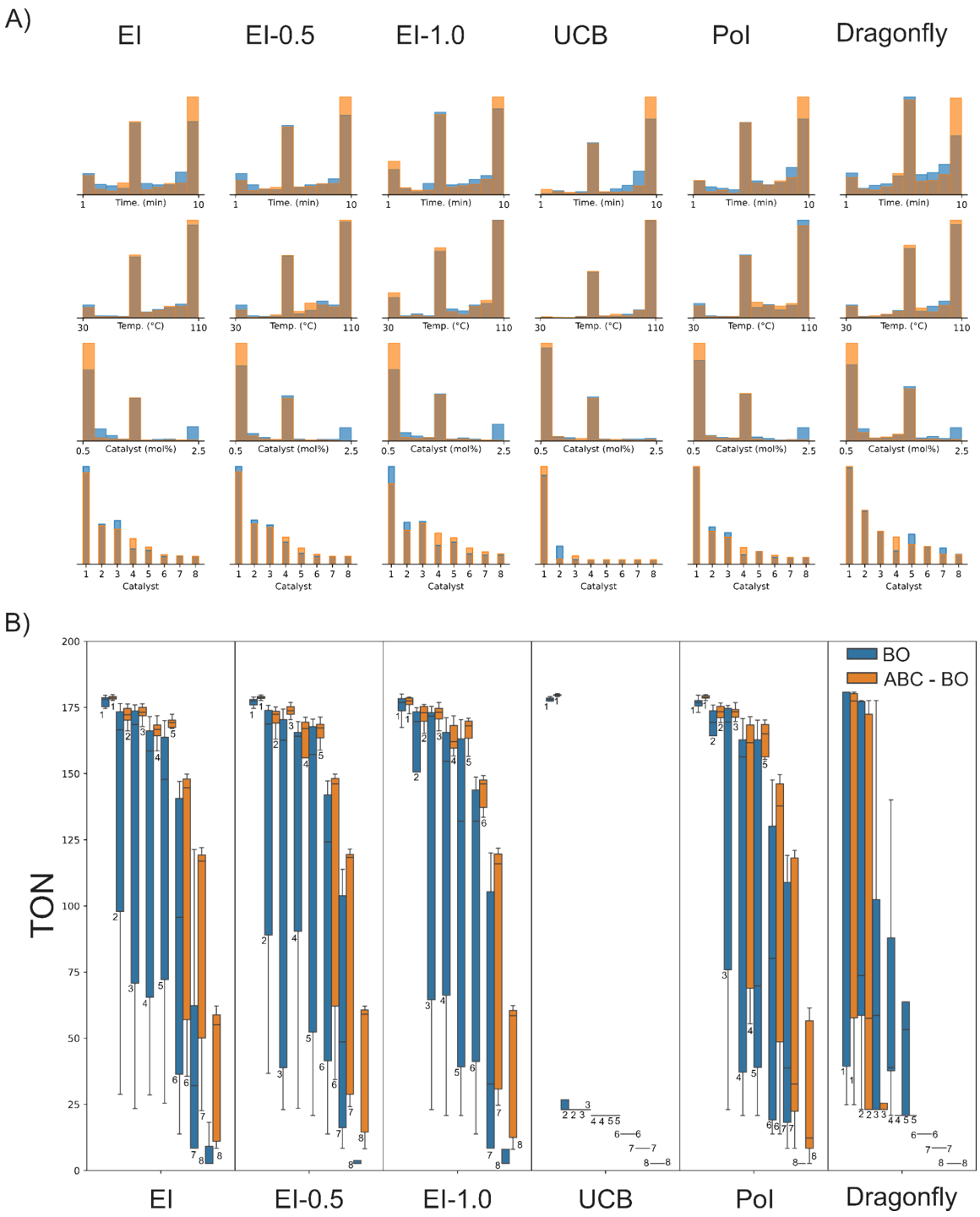


Fig. S12 A) Distribution of variable selection in optimization for ABC-BO (orange) and BO (blue) in ISR-3; histogram plots for continuous variables and bar plots for categorical variables. B) Best objective value (TON) for each catalyst in ISR-3

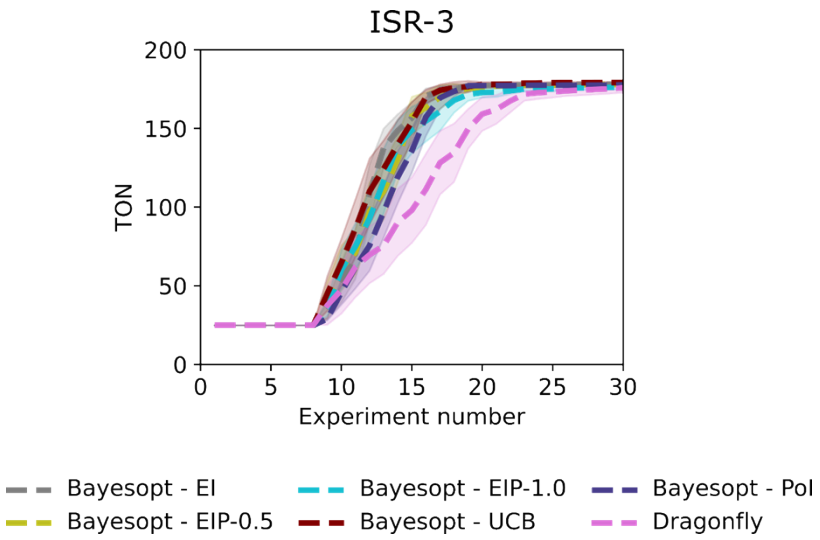
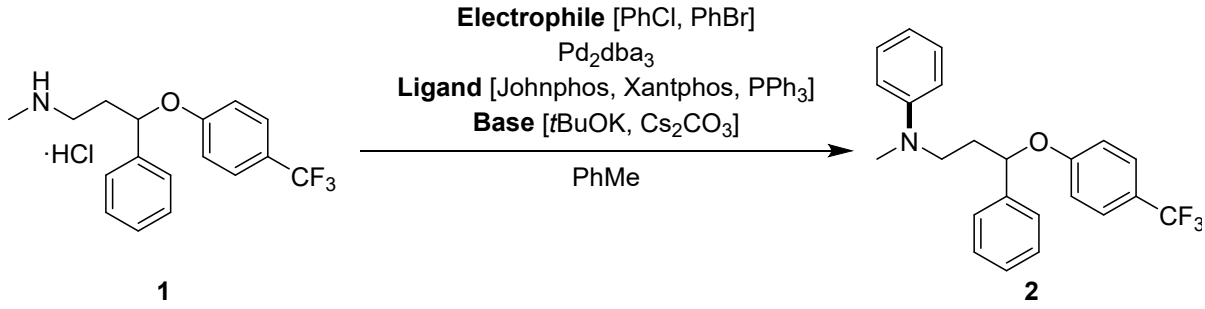


Fig. S13. Average optimization trends of ABC-BO approaches across different solvers for ISR-3. The lines represent the mean performance across multiple runs, and the shaded regions indicate the 95% confidence intervals.

7. Practical experiment

7.1. Reaction procedure



Scheme S2. Arylation of fluoxetine through Buchwald-Hartwig coupling

General procedure for initialization

To a 4 ml screw-cap vial were added Fluoxetine hydrochloride (1, 0.500 mmol, 173 mg), the base (3 eq.) and Pd2dba3 and the ligand (4x [Pd2dba3]) dissolved in dried and degassed toluene (2 ml). Afterwards, the electrophile and anisole (0.500 mmol, 54.3 µl, 1 eq.) were added, the vial was flushed with Argon, capped and stirred for the indicated time and temperature.

For 1H NMR analysis a sample (0.2 ml) was taken, diluted with CDCl3 (0.5 ml) and filtered through a syringe filter (0.20 µm). The yield of 2 was determined by the ratio of the signals of anisole (PhOMe, 3.74 ppm, s, 3H) and the product (-NPhMe, 2.86 ppm, s, 3H).

Best result – Conditions of experiment 20, ABC-BO

Following the general procedure Fluoxetine hydrochloride (1, 0.500 mmol, 173 mg), tBuOK (168 mg, 1.50 mmol, 3 eq.), Pd2dba3 (6.73 mg, 7.35 µmol, 1.47 mol%), JohnPhos (8.77 mg, 29.4 µmol 5.88 mol%) and PhBr (99.4 µl, 0.950 mmol, 1.9 eq.) were mixed and the mixture was stirred for 0.5 h at 110 °C. Afterwards, the sample was filtered over celite with EtOAc (20 ml), concentrated under reduced pressure and subjected to column chromatography (silica, EtOAc/Cy 1:40) to obtain the product (2, 151 mg, 0.391 mmol, 78%) as a colourless solid.

1H NMR (300 MHz, CDCl3) δ 7.43–7.35 (m, 2H), 7.31–7.17 (m, 5H), 7.17–7.10 (m, 2H), 6.90–6.81 (m, 2H), 6.69–6.58 (m, 3H), 5.16 (dd, 1H, J = 8.1, 4.7 Hz), 3.62–3.39 (m, 2H), 2.86 (s, 3H), 2.25–2.01 (m, 2H). 13C NMR (75 MHz, CDCl3) δ 160.5, 149.2, 140.9, 129.3, 128.9, 128.0, 126.9 (q, J = 3.8 Hz), 125.7, 124.4 (q, J = 271.2 Hz), 122.9 (q, J = 32.6 Hz), 116.5, 115.8, 112.4, 78.1, 49.1, 38.6, 36.1. 19F NMR (376 MHz, CDCl3) δ -61.5. FTIR (ATR, neat) υ 2966, 2917, 2855, 2224, 1670, 1604, 1573, 1506, 1446, 1379, 1300, 1251, 1235, 1170, 1111, 982, 832, 546 cm-1. mp = 39–41 °C. HRMS (ESI+, MeOH/H2O) m/z [M+H]+ calc. for C23H23NOF3 386.1732; Found 386.1719.

7.2. Optimization problem

The variables involved in the optimization include catalyst loading ($0.5\ to\ 5\ mol\%$), reaction time ($0.5, 1.0, \ldots, 7.5, 8.0\ h$), temperature ($50, 60, \ldots, 100, 110\ °C$), electrophile equivalent ($1.0, 1.1, \ldots, 1.9, 2.0\ eq$), electrophile ($PhCl, PhBr$), base ($KOtBu, Cs_2CO_3$), and ligand ($Johnphos, Xanphos, PPh_3$). The objective was to maximize the ratio of productivity (throughput) to cost. Productivity corresponds to the amount of product formed per unit time, while cost corresponds to the combined cost of the electrophile, catalyst, base, and ligand. The costs of the reagents were sourced from Sigma-Aldrich, dated 03 December 2024 (Table S4). Since the costs of other reagents (e.g., fluoxetine **1**, solvent) are constant across all reactions, they were not accounted for in the optimization. From the representation of the maximum achievable objective across the design space (Fig. S14, left), it can be observed that the objective corresponding to low time and cost differs significantly from the rest of the space. To enhance the performance of the surrogate model, the objective values were transformed using the natural logarithm (Fig. S14, right). This transformation helps to scale the values, particularly when there are significant differences across the design space, making the model more robust to variations.[10,11]

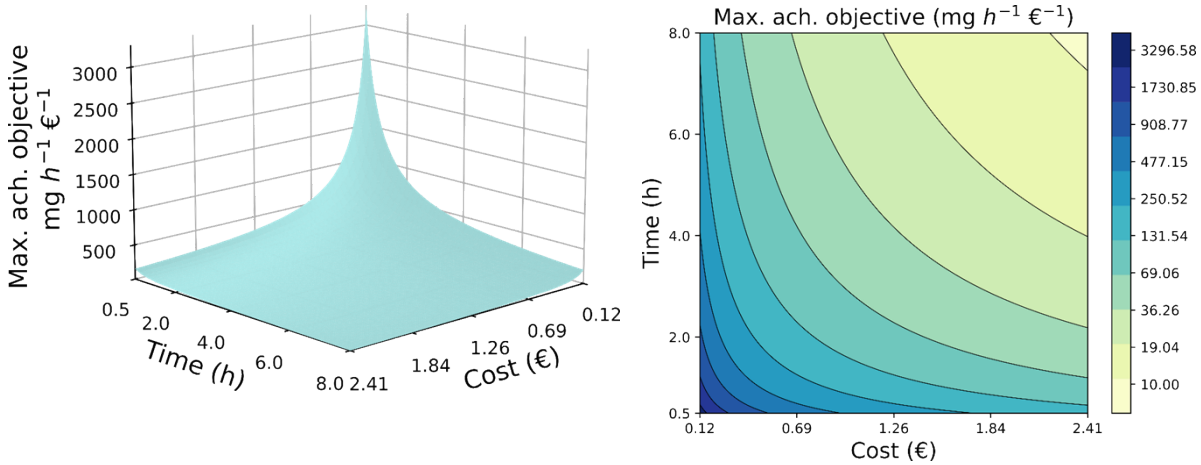


Fig. S14. Maximum achievable objective (assuming yield = 100%) within the search space, represented in terms of time and cost. (left) 3D surface plot, (right) contour plot in logarithmic scale.

The reaction was optimized using MATLAB's built-in function, Bayesopt, in the ABC-BO framework. The acquisition function used was expected improvement plus, with the default exploration factor of 0.5. The discrete numeric variables - reaction time, electrophile equivalent, and temperature - were treated as integer variables since Bayesopt handles integer variables but not discrete numeric ones directly. This approach is valid because discrete numeric and integer variables representation in surrogate model is same as continuous variables. The main difference is that, during acquisition function optimization, the search is limited to the predefined integer or discrete numeric values rather than exploring the entire continuous domain.

Table S4. Reagents cost used for evaluating the objective function

Reagents	**Cost**
PhBr	134 €/l
PhCl	18.76 €/l
Pd_2dba_3	37 €/g
Johnphos **L1**	11.48 €/g
Xantphos **L2**	22.24 €/g
PPh_3 **L3**	0.06 €/g
*t*BuOK	0.19 €/g
Cs_2CO_3	0.54 €/g

The constraint used for the ABC-BO is:

$f^{max.\,acheivable} > f_n^*$

where, $f^{max.\,acheivable}$ is the objective value assuming a 100% yield, and f_n^* represents the existing best objective value adjusted for noise (5%)

$f_n^* = f(x^*, yield^* - noise)$

Here, x^* represents the condition, and $yield^*$ is the reaction yield corresponding to best objective value f^*.

The influencing variables whose boundaries or levels (in case of categorical variables) that are eligible to update during the optimization to eliminate futile space include reaction time, electrophile equivalent, catalyst loading, electrophile, base and ligand. For the reaction time, electrophile equivalent and catalyst loading, reduction in these boundaries favours the objective value, and thus the upper limit for each variable was adjusted, if necessary, for each experiment in the optimization process.

7.3. Results

Table S5. Initialization results; center point for each combination of the categorical variables.

Exp. No.	Temp	Time	Cat. Load.	Elec. Load.	Ligand	Base	Electrophile	Yield	Productivity	Cost	Objective= $\frac{productivity}{cost}$	Max. ach. Obj.
	^oC	h	$mol\%$	$eq.$				%	mg/h	€	$mg\,h^{-1}\,€^{-1}$	$mg\,h^{-1}\,€^{-1}$
1	80	4	2.75	1.5	Johnphos	*t*BuKO	PhBr	99.4	47.9	0.70	68.82	69.23
2	80	4	2.75	1.5	Johnphos	Cs2Co3	PhBr	1	0.5	0.93	0.52	51.99
3	80	4	2.75	1.5	Xantphos	*t*BuKO	PhBr	88.5	42.6	1.22	35.09	39.64
4	80	4	2.75	1.5	Xantphos	Cs2CO3	PhBr	0	0.0	1.45	0.00	33.32
5	80	4	2.75	1.5	PPh3	*t*BuKO	PhBr	5.4	2.6	0.51	5.12	94.77
6	80	4	2.75	1.5	PPh3	Cs2Co3	PhBr	0	0.0	0.74	0.00	65.17
7	80	4	2.75	1.5	Johnphos	*t*BuKO	PhCl	94	45.3	0.69	65.94	70.15
8	80	4	2.75	1.5	Johnphos	Cs2CO3	PhCl	0	0.0	0.92	0.00	52.50
9	80	4	2.75	1.5	Xantphos	*t*BuKO	PhCl	9.4	4.5	1.21	3.75	39.94
10	80	4	2.75	1.5	Xantphos	Cs2Co3	PhCl	0	0.0	1.44	0.00	33.53
11	80	4	2.75	1.5	PPh3	*t*BuKO	PhCl	7.9	3.8	0.50	7.62	96.50
12	80	4	2.75	1.5	PPh3	Cs2CO3	PhCl	0	0.0	0.73	0.00	65.99

Table S6. Optimization results for BO

Exp. No.	Temp	Time	Cat. Load.	Elec. Load.	Ligand	Base	Electrophile	Yield	Productivity	Cost	Objective= $\frac{productivity}{cost}$	Max. ach. Obj.
	^oC	h	$mol\%$	$eq.$				%	mg/h	€	$mg\ h^{-1}\ €^{-1}$	$mg\ h^{-1}\ €^{-1}$
13	80	4.5	4.12	1.3	Johnphos	*t*BuKO	PhBr	96	41.1	1.02	40.28	41.96
14	60	1.5	2.4265	2	Johnphos	*t*BuKO	PhBr	92	118.2	0.62	190.36	206.92
15	50	8	1.416	1.7	Johnphos	*t*BuKO	PhCl	56	13.5	0.37	36.50	65.17
16	110	8	2.47	1	Johnphos	*t*BuKO	PhCl	70	16.9	0.62	27.21	38.87
17	110	8	3.0762	2	Johnphos	*t*BuKO	PhBr	93	24.1	0.78	28.83	31.00
18	90	0.5	0.5799	1.9	Johnphos	*t*BuKO	PhBr	25	96.4	0.18	528.27	2113.08
19	60	0.5	0.7252	1.2	Johnphos	*t*BuKO	PhBr	41	158	0.21	745.18	1817.52
20	110	0.5	0.5024	1.1	Johnphos	*t*BuKO	PhBr	22	84.8	0.16	535.48	2434.00

Table S7. Optimization results for ABC-BO

Exp. No.	Temp	Time	Cat. Load.	Elec. Load.	Ligand	Base	Electrophile	Yield	Productivity	Cost	Objective= $\frac{productivity}{cost}$	Max. ach. Obj.
	^oC	h	$mol\%$	$eq.$				%	mg/h	€	$mg\ h^{-1}\ €^{-1}$	$mg\ h^{-1}\ €^{-1}$
13	80	1.5	2.19	1.6	Johnphos	*t*BuKO	PhBr	95	122.1	0.56	217.07	228.49
14	100	1	1.71	2	Johnphos	*t*BuKO	PhBr	99.9	192.5	0.45	425.84	426.26
15	110	1	0.50	1.2	Johnphos	*t*BuKO	PhBr	30.5	58.8	0.16	370.27	1213.99
16	110	0.5	1.27	1.8	Johnphos	*t*BuKO	PhCl	66.4	255.9	0.33	764.27	1151.01
17	70	1	0.50	1.2	Xantphos	*t*BuKO	PhBr	7.2	13.9	0.25	54.69	759.55
18	110	0.5	1.85	1.3	Johnphos	*t*BuKO	PhCl	75.8	292.2	0.47	618.63	816.13
19	50	0.5	1.19	1.4	Johnphos	*t*BuKO	PhCl	8.7	33.5	0.31	106.57	1224.99
20	110	0.5	1.47	1.9	Johnphos	*t*BuKO	PhBr	90.2	347.7	0.39	882.94	978.87

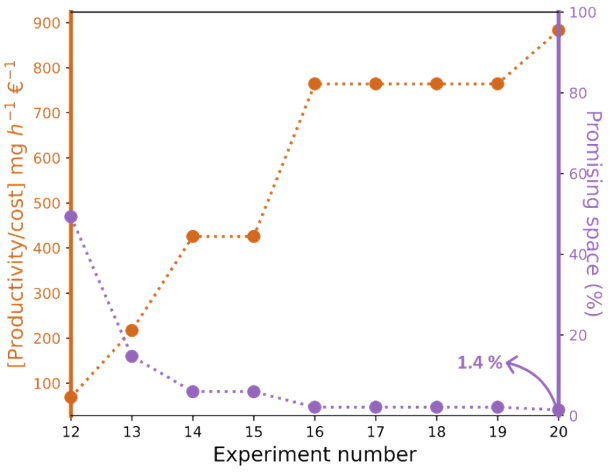


Fig. S15. Trend of the objective over the course of optimization and reduction of promising search space in ABC-BO

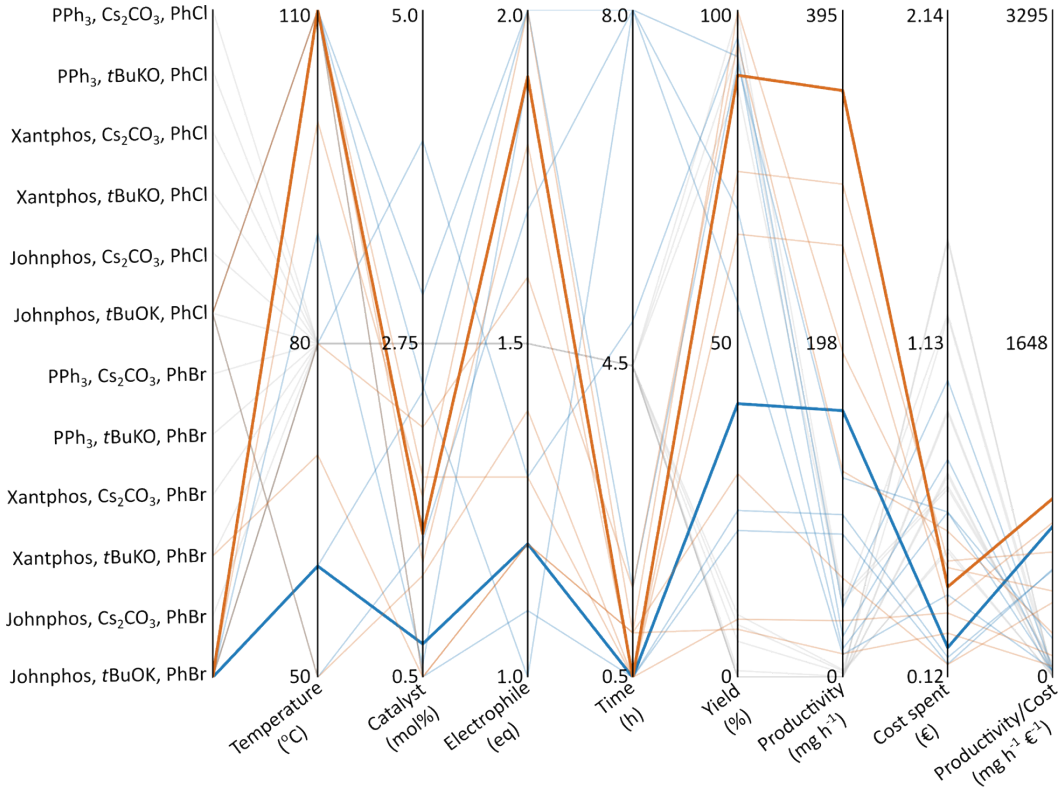


Fig. S16 Optimization results presented as a parallel coordinate plot. Each line in the plot corresponds to a different experiment, illustrating the relationship between variables, the objective (productivity/cost), and the parameters (yield, productivity, cost) used to calculate the objective. Grey lines denote initialization experiments, while blue and orange lines represent experiments conducted using BO and ABC-BO, respectively. The optimum results from BO and ABC-BO are highlighted with dark lines.

Fig. S17-S25 represent the futile space within the search space for individual combinations of the categorical variables during optimization using ABC-BO. The search space is depicted using the variables time and cost, which also account for other influencing variables such as catalyst loading and electrophile equivalent. At the end of the initialization phase (12th experiment, Fig. S17), the futile space for each combination varies due to differences in reagent costs and the incorporation of cost into the objective function. As the experiments progress and the best objective value increases, the futile space for each combination also expands. If all combinations for a particular reagent reached 100% futile space, that reagent could be removed as a level of the categorical variable. However, in our case, no single reagent combination reached 100% futile space by the 20th experiment (Fig. S25), so no reagent was eligible for

elimination. Nonetheless, the reagent combinations [Xantphos, Cs_2CO_3, PhBr] and [Xantphos, Cs_2CO_3, PhCl] reached 100% futile space.

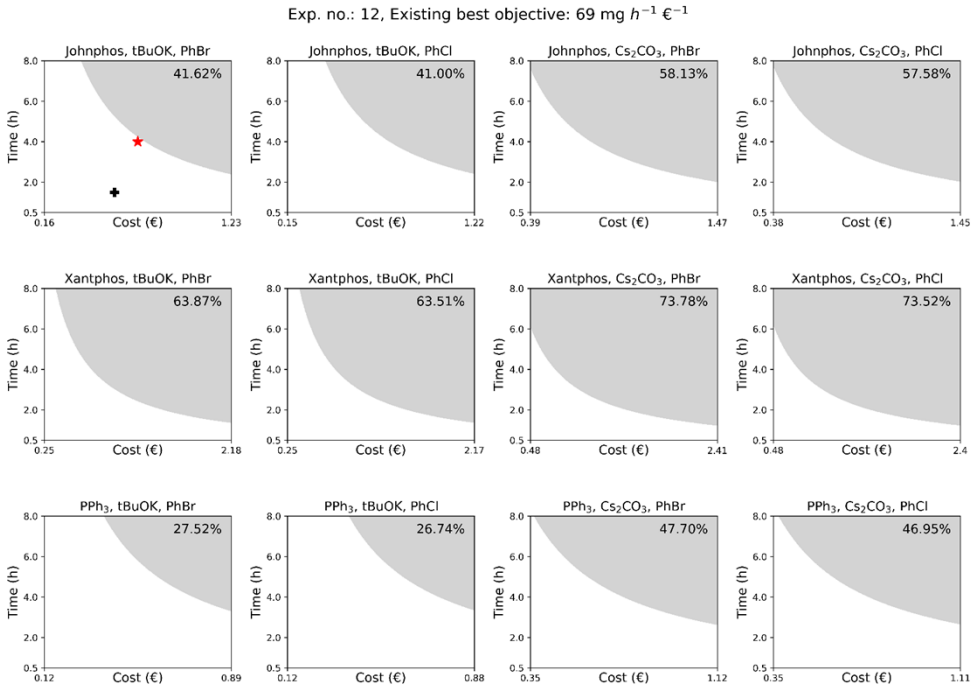


Fig. S17. Futile space (grey) within the search space corresponding to individual combinations of categorical variables at the end of experiment 12 (initialization). The red star represents the existing best objective, and the black plus sign represents the next point suggested by ABC-BO.

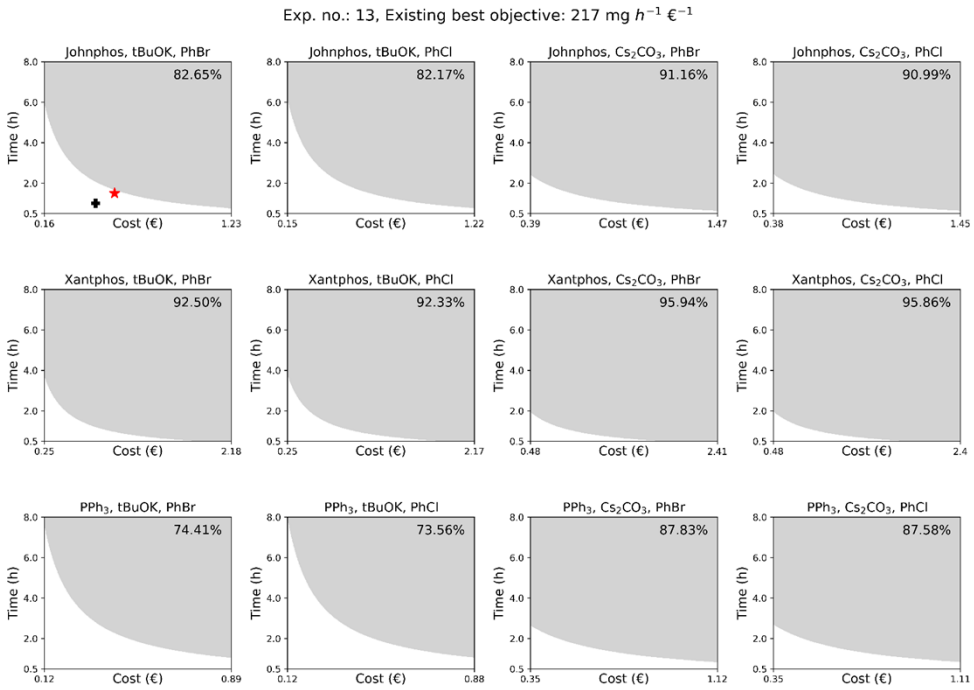


Fig. S18. Futile space (grey) within the search space corresponding to individual combinations of categorical variables at the end of experiment 13. The red star represents the existing best objective, and the black plus sign represents the next point suggested by ABC-BO.

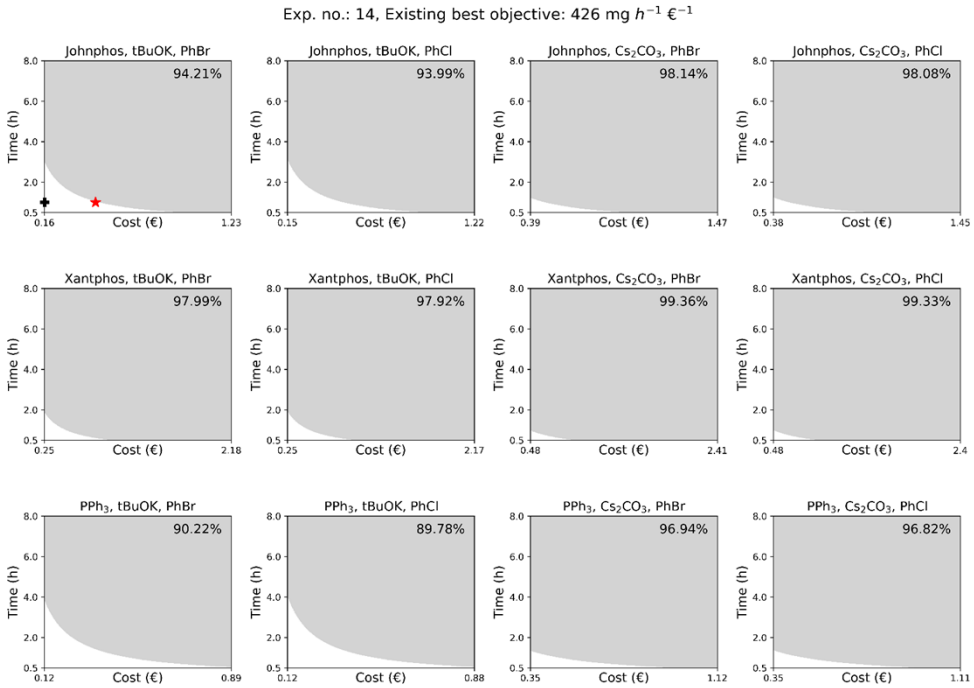


Fig. S19. Futile space (grey) within the search space corresponding to individual combinations of categorical variables at the end of experiment 14. The red star represents the existing best objective, and the black plus sign represents the next point suggested by ABC-BO.

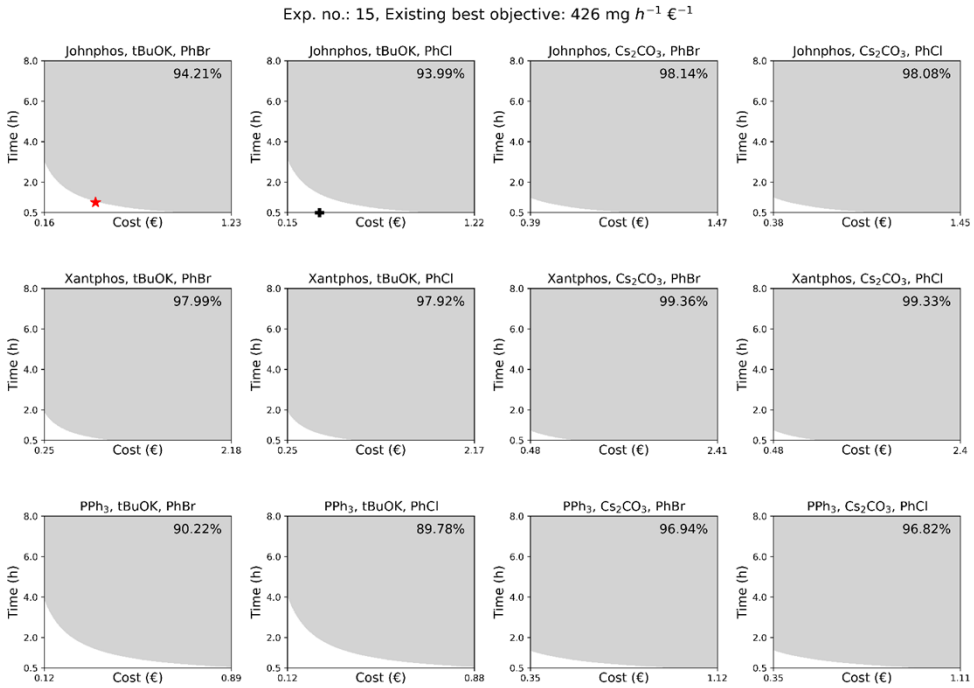


Fig. S20. Futile space (grey) within the search space corresponding to individual combinations of categorical variables at the end of experiment 15. The red star represents the existing best objective, and the black plus sign represents the next point suggested by ABC-BO.

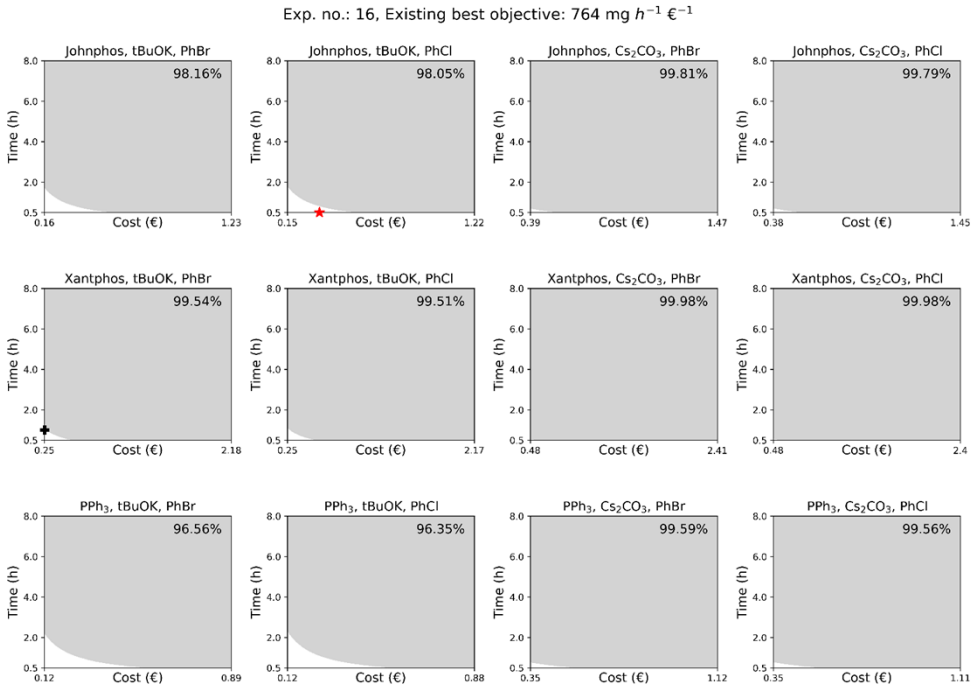


Fig. S21. Futile space (grey) within the search space corresponding to individual combinations of categorical variables at the end of experiment 16. The red star represents the existing best objective, and the black plus sign represents the next point suggested by ABC-BO.

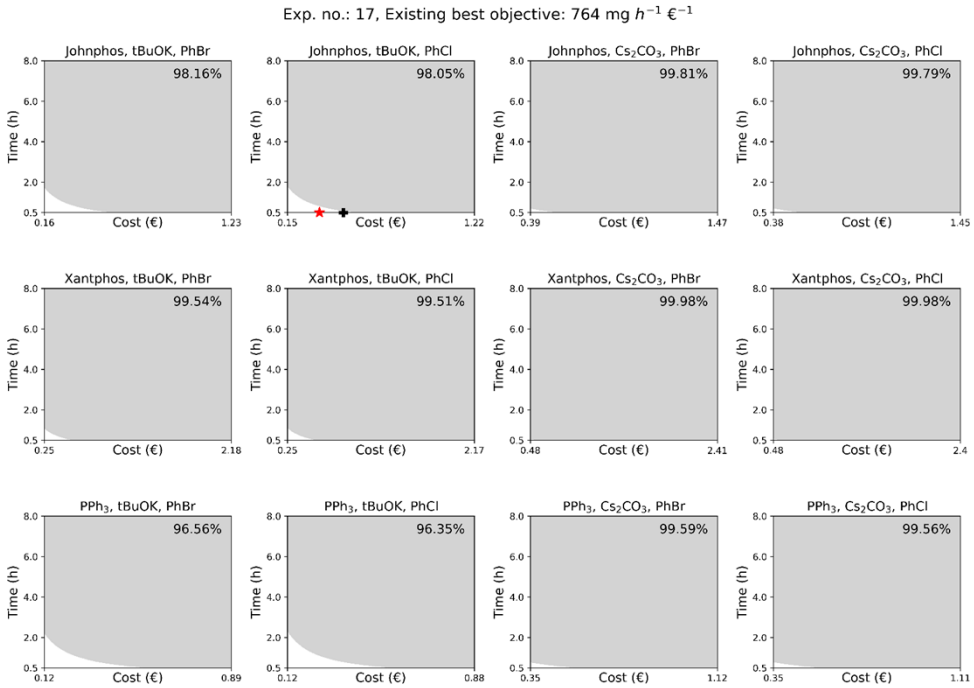


Fig. S22. Futile space (grey) within the search space corresponding to individual combinations of categorical variables at the end of experiment 17. The red star represents the existing best objective, and the black plus sign represents the next point suggested by ABC-BO.

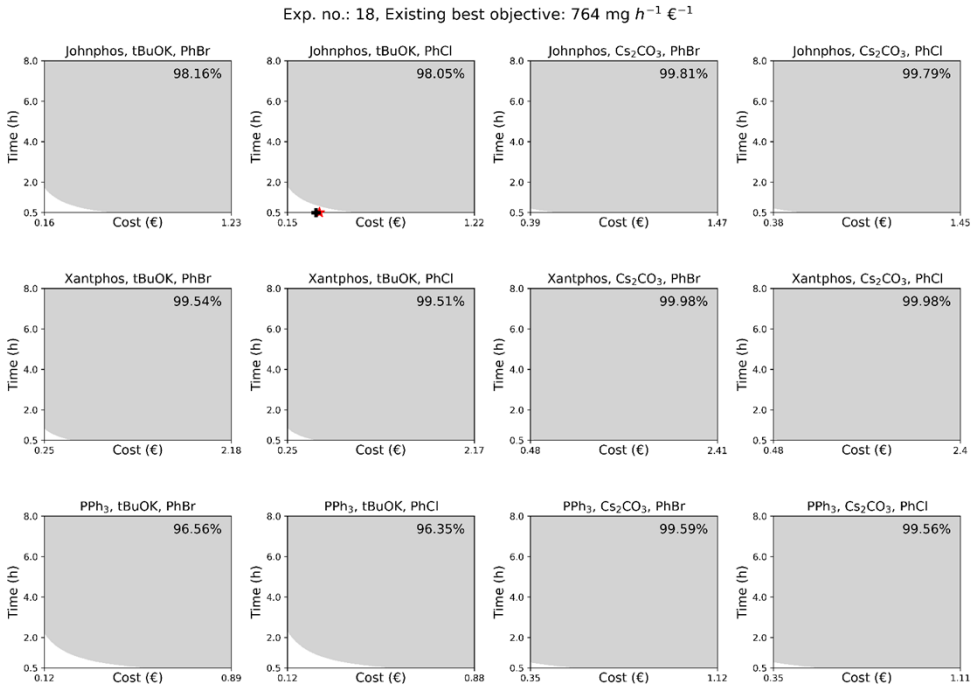


Fig. S23. Futile space (grey) within the search space corresponding to individual combinations of categorical variables at the end of experiment 18. The red star represents the existing best objective, and the black plus sign represents the next point suggested by ABC-BO.

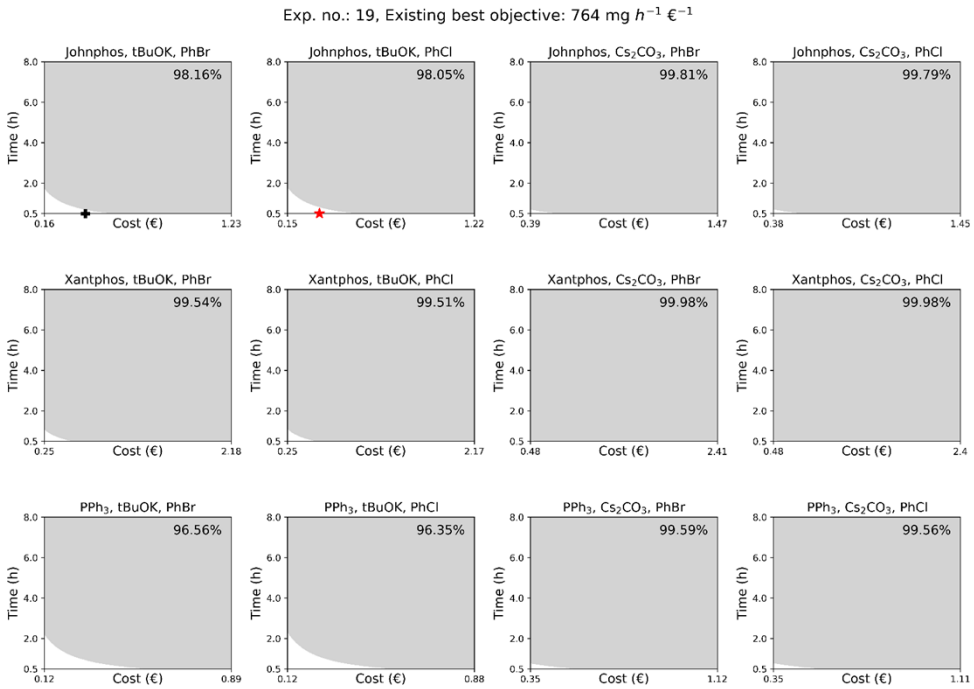


Fig. S24. Futile space (grey) within the search space corresponding to individual combinations of categorical variables at the end of experiment 19. The red star represents the existing best objective, and the black plus sign represents the next point suggested by ABC-BO.

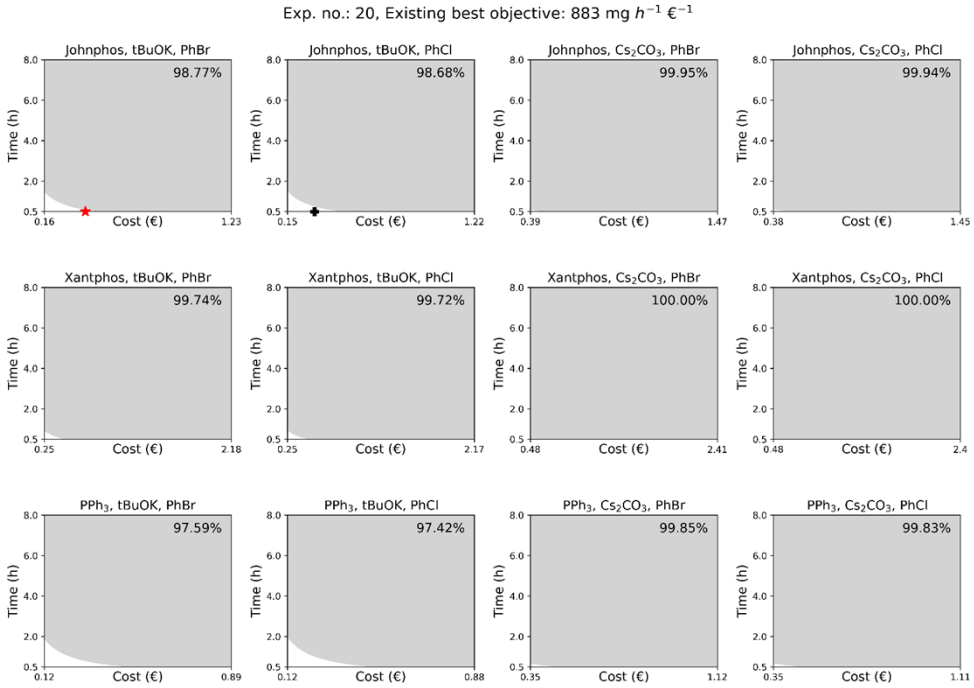


Fig. S25. Futile space (grey) within the search space corresponding to individual combinations of categorical variables at the end of experiment 20. The red star represents the existing best objective, and the black plus sign represents the next point suggested by ABC-BO.

The boundaries of the influencing variables - reaction time, electrophile equivalent, and catalyst loading - that satisfy the constraint for each combination are presented in Fig. S26-S28. It can be observed that the trends of boundary update for each variable differ, indicating their varying impact on the objective function. For the electrophile, the limits remain unchanged from the initially specified boundary.

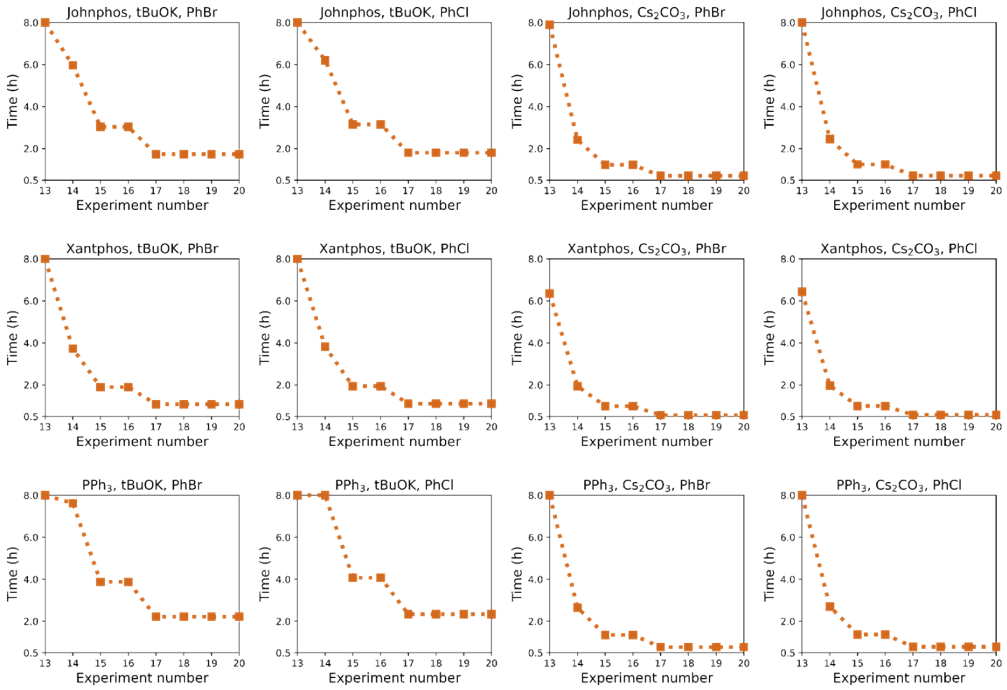


Fig. S26. Maximum time limit within the specified boundary that satisfies the constraint at the start of each experiment for all combinations of categorical variables in ABC-BO.

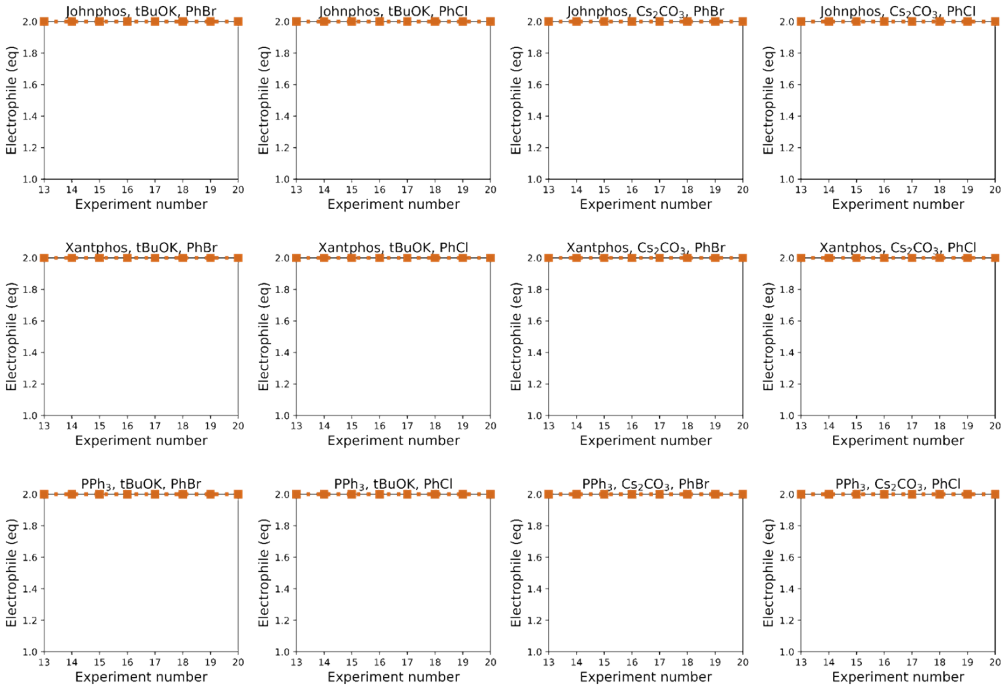


Fig. S27. Maximum electrophile limit within the specified boundary that satisfies the constraint at the start of each experiment for all combinations of categorical variables in ABC-BO.

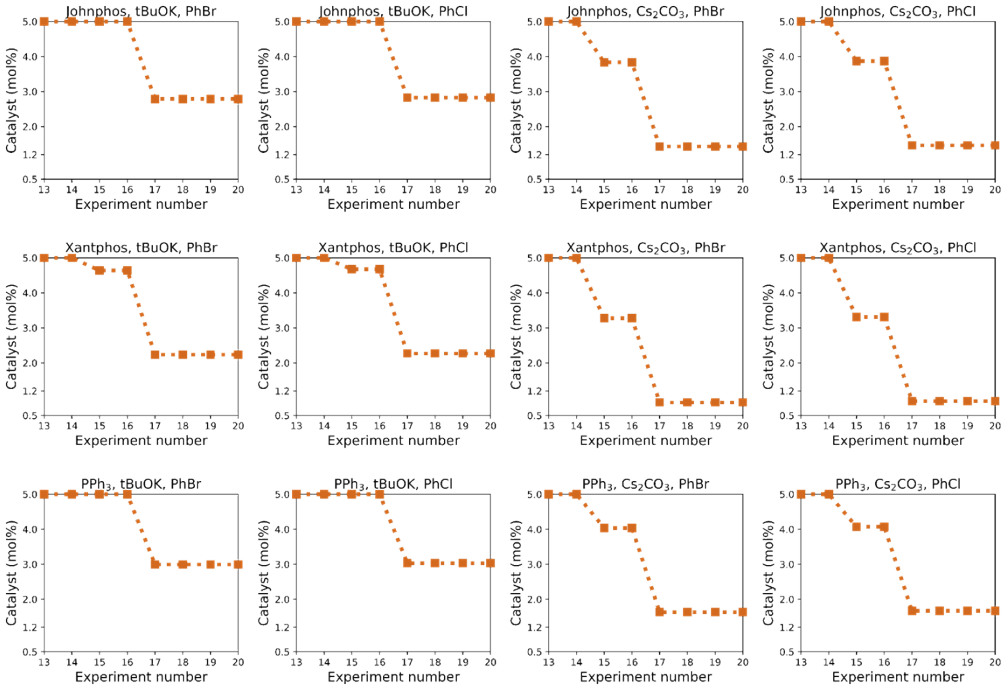


Fig. S28. Maximum catalyst loading limit within the specified boundary that satisfies the constraint at the start of each experiment for all combinations of categorical variables in ABC-BO.

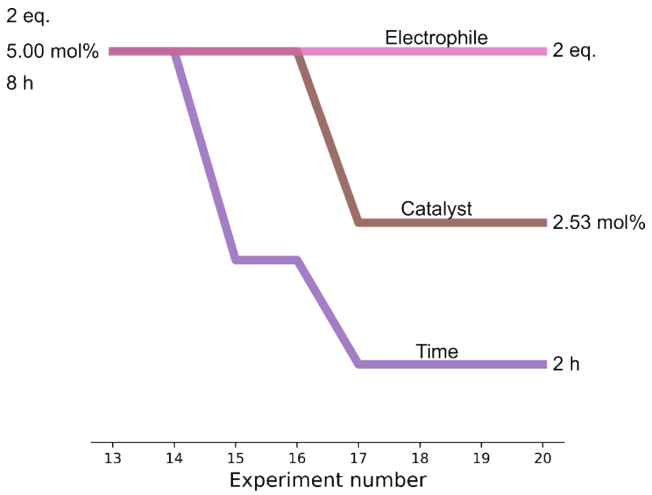


Fig. S29. Boundary reduction of the variables during optimization using ABC-BO

The constraint satisfaction score (Fig. S30) represents how much the maximum achievable objective exceeds the existing best objective (accounted for noise). A positive value indicates that the point theoretically allows improvement over the existing best objective. For ABC-BO (orange), the scores are positive for all experiments during the optimization process. In contrast, for BO, some experiments (e.g., 13, 15, 16, and 17) have negative scores, indicating futile experiments.

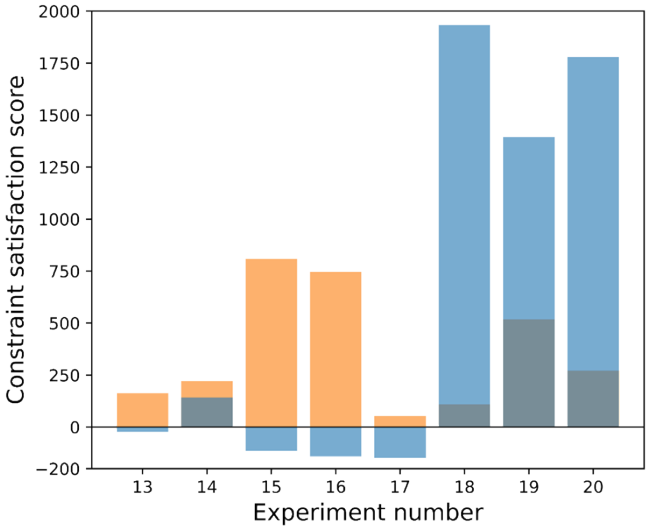


Fig. S30. Constraint score: $f^{max.\,acheivable} - f^{*}_{n}$; orange for ABC-BO, blue for BO.

8. References

1 R. J. Hickman, M. Aldeghi, F. Häse and A. Aspuru-Guzik, *Digit. Discov.*, 2022, **1**, 732–744.

2 K. Kandasamy, K. R. Vysyaraju, W. Neiswanger, B. Paria, C. R. Collins, J. Schneider, B. Poczos and E. P. Xing, *J. Mach. Learn. Res.,* 2020, **21**, 1–27.
3 C. Qin, D. Klabjan and D. Russo, in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017, **30**.
4 K. E. Konan, A. S. Vel, A. Abollé, D. Cortés-Borda and F.-X. Felpin, *React. Chem. Eng.*, 2023, **8**, 2446–2454.
5 A. S. Vel, K. E. Konan, D. Cortés-Borda and F.-X. Felpin, *Org. Process Res. Dev.*, 2024, **28**, 1597–1606.
6 A. S. Vel, D. Cortés-Borda and F.-X. Felpin, *React. Chem. Eng.*, 2024, **9**, 2882–2891.
7 B. J. Reizman, Thesis, Massachusetts Institute of Technology, 2015.
8 L. M. Baumgartner, C. W. Coley, B. J. Reizman, K. W. Gao and K. F. Jensen, *React. Chem. Eng.*, 2018, **3**, 301–311.
9 N. Aldulaijan, J. A. Marsden, J. A. Manson and A. D. Clayton, *React. Chem. Eng.*, 2024, **9**, 308–316.
10 D. R. Jones, M. Schonlau and W. J. Welch, *J. Glob. Optim.*, 1998, **13**, 455–492.
11 A. M. Schweidtmann, A. D. Clayton, N. Holmes, E. Bradford, R. A. Bourne and A. A. Lapkin, *Chem. Eng. J.*, 2018, **352**, 277–282.