

# General Reaction Conditions *via* Data-driven Optimisation

Stefan P. Schmid<sup>§\*</sup> and Kjell Jorner\*

<sup>§</sup>SCS-dsm-firmenich Award for the Best Poster Presentation in Computational Chemistry

**Abstract:** General reaction conditions are a long-standing goal in chemical synthesis, as such conditions facilitate library synthesis and a broad substrate scope. However, despite their importance, the generality of reaction conditions is mostly an afterthought when reaction conditions are optimised. Considering multiple substrates during reaction condition optimisation alleviates this problem and enables the identification of conditions that work well for multiple substrates. Inspired by data-driven optimisation techniques for one model substrate, machine learning based strategies have also been proposed to optimise reactions towards general reaction conditions. In this work, we describe recent algorithmic advances in this domain, including our state-of-the-art algorithm. This algorithm is also available as an easy-to-use website to allow experimental chemists to use it without code.

**Keywords:** Bayesian optimisation · General reaction conditions · Machine learning · Reaction development



**Stefan P. Schmid** is a doctoral researcher in the Digital Chemistry Laboratory at ETH Zurich under the guidance of Prof. Dr. Kjell Jorner. After completing his MSc in chemistry, Stefan joined the Digital Chemistry Laboratory, where his research focuses on the optimisation of chemical reactions for the discovery and development of new catalytic methods. In particular, he works on Bayesian Optimisation algorithms for

chemical laboratories to efficiently optimise reactions and accelerate catalyst development.



**Kjell Jorner** is an Assistant Professor of Digital Chemistry at ETH Zurich. He completed his PhD in computational organic chemistry at Uppsala University, studying the effect of aromaticity on photochemical reactions. He then joined AstraZeneca UK as a Postdoctoral Fellow, building machine learning models for predicting the outcome of chemical reactions, followed by a postdoctoral fellowship at the University

of Toronto, working on machine learning for molecular design with Alán Aspuru-Guzik. Since 2023, he has led a group at ETH, focusing on accelerating chemical discovery using digital tools.

## 1. Introduction

General reaction conditions are conditions that work well across many related transformations, *e.g.* different substrates of a reaction. Obtaining such general conditions is a long-standing goal in chemical synthesis<sup>[1–8]</sup> where the applicability of reactions to different substrates is evaluated *via* a substrate scope. Robust reactions are also more likely to be adopted on an industrial scale,<sup>[9]</sup> and general conditions facilitate library and high-throughput experimentation (HTE) synthesis<sup>[10]</sup> for experimental screening

efforts.<sup>[1–3,8]</sup> While identifying general reaction conditions are useful for these reasons, their identification is often not considered *a priori* within the reaction development process.

In fact, reaction development is often approached with a ‘model substrate’ approach:<sup>[1]</sup> productive reaction conditions are identified on a chosen model substrate, and subsequently optimised to maximize yield or selectivity on this model substrate. Only after such optimal conditions have been identified, their generality is evaluated by performing reactions with such conditions on a diverse array of substrates, *i.e.* the scope of the reaction is evaluated. However, when optimal conditions are identified on a model substrate, it is far from guaranteed, or even likely, that these conditions are general across a substrate scope (Fig. 1). Considering multiple substrates during the optimisation process addresses this problem and considers reaction generality already at an earlier stage.

How reaction conditions are optimised for a model substrate has also seen a shift in recent years. One-factor-at-a-time optimisation, *i.e.* only varying one component at a time, still remains a standard optimisation technique, but neglects interaction effects between components. For this reason, data-driven methods that learn more sophisticated patterns from acquired data have seen an increased adoption. Particularly popular techniques are design of experiments (DoE) and Bayesian Optimisation (BO); the latter is also becoming the standard algorithm in self-driving laboratories.<sup>[11]</sup> In a prototypical optimisation campaign, a BO algorithm needs a multi-fold input from a user: (1) the search space (*i.e.* in the context of reaction optimisation, which conditions should be considered), (2) the objective (*i.e.* what should be optimised; *e.g.* the yield, selectivity, can also be multi-objective), and (3) a set of initially performed experiments within that search space. Based on these data, a machine learning model is trained (typically a Gaussian Process), which for each point in the search space predicts its performance (yield), and the associated uncertainty of this prediction. The trained model is the basis for proposing the next measurement(s), *via* an acquisition function – this function

\*Correspondence: S. P. Schmid, E-mail: stefan.p.schmid@gmail.com; Prof. K. Jorner, E-mail: kjell.jorner@chem.ethz.ch

Institute of Chemical and Bioengineering, Department of Chemistry and Applied Biosciences, ETH Zurich, Zurich CH-8093, Switzerland & NCCR Catalysis, Switzerland

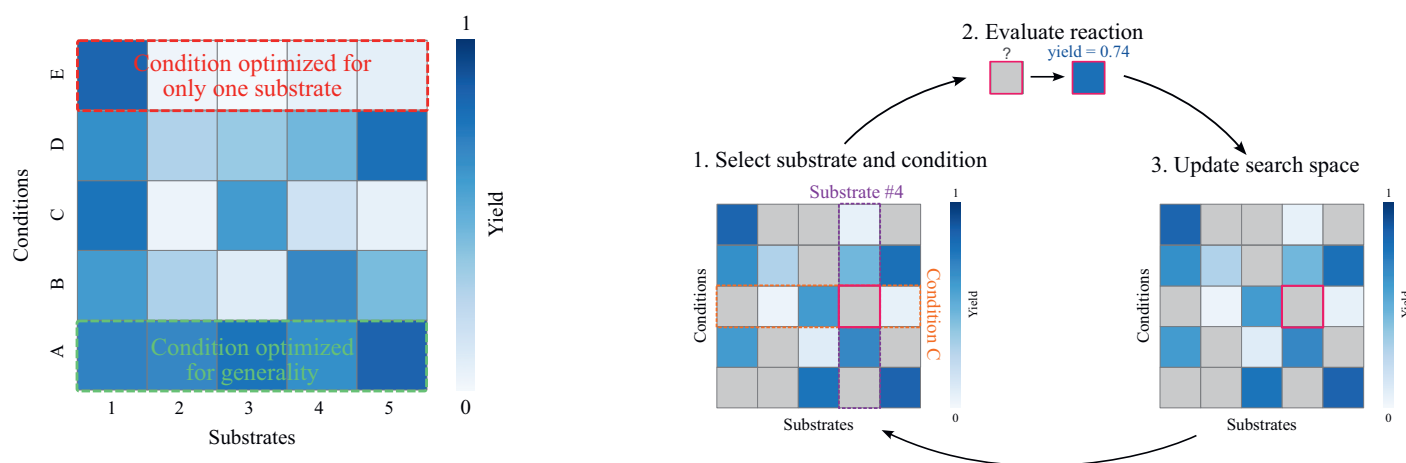


Fig. 1. Conditions optimised on one model substrate are not necessarily generally optimal. To optimise for generality, CurryBO iteratively proposes new conditions and substrates, evaluates the reaction and updates the dataset.

balances exploration of the search space (measure points with high uncertainty in underexplored regions of the search space to potentially discover new successes), and exploitation (measure points similar to well-performing ones, to improve on the current optimum). The proposed points are then measured, the dataset is updated, and the model is trained anew. This cycle is repeated until the user is content with the attained result, or a defined number of experiments has been carried out.

## 2. Optimisation Problem for General Conditions

While such a strategy has seen numerous reports of success for reaction optimisation and other domains in chemistry,<sup>[11]</sup> it is mainly used to optimise one function, *e.g.* the yield of the reaction of a model substrate as a function of the conditions (yield (conditions)). However, for general reaction conditions, multiple functions are at play, as each of the substrates has its own yield (conditions) function. For optimising general conditions, these functions need to be combined into one generality function, which can be optimised (Fig. 2). The separate functions are combined *via* a generality metric that captures the generality requirements of the chemist. Intuitive examples are maximising the average yield over all substrates, or maximising the number of substrates with a yield above a defined threshold.<sup>[4]</sup>

A centrally arising problem is that to measure the generality of certain conditions, the conditions have to be evaluated for every substrate, *e.g.* to calculate the average, you need all reac-

tion outcomes. For realistic substrate scopes, often with multiple dozens of entries, a complete measurement of the generality for each condition is thus contrary to the goal of experiment-efficient data-driven optimisation.

Instead, the generality of some conditions can be partially observed: measuring the reaction outcome of conditions on one or a few substrates provides partial insight into the generality of those conditions. Such a partial observation scenario is however not compatible with standard BO techniques and requires more complex decision making algorithms. More fundamentally, such algorithms should not only propose the next conditions, but also select a substrate on which to evaluate these conditions (Fig. 1).

## 3. Examples of Optimisation for General Conditions

Even though data-driven optimisation is becoming more adopted for model substrate optimisations, it is hardly considered for general conditions. Prominent examples have only now been recently published by Angello *et al.*<sup>[12]</sup> and Wang *et al.*<sup>[13]</sup>

The former proposes a BO approach towards this problem in a sequential acquisition strategy, at first selecting the next conditions, followed by selecting a substrate: conditions are selected by optimising a probability of improvement acquisition function (selecting the conditions with the highest probability of improving conditions); the substrate is selected by choosing the one with the highest predicted uncertainty for the selected conditions.

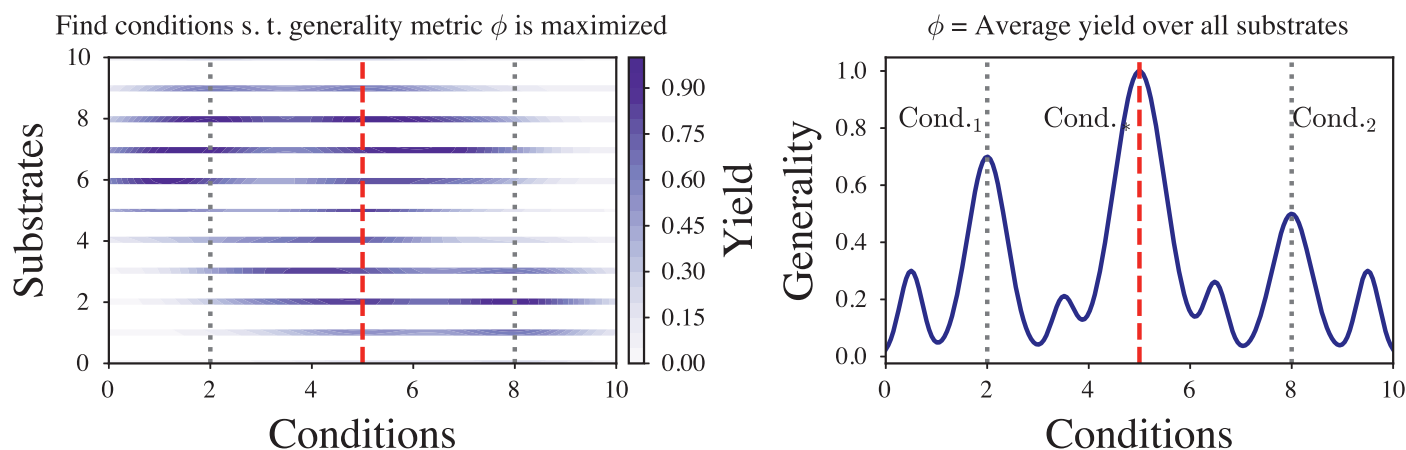


Fig. 2. An optimisation problem for general conditions consists of multiple functions, as each substrate has a yield (condition) function. For optimisation of general reaction conditions, these functions are combined to a generality function *via* a generality function  $\phi$ .

Wang *et al.*<sup>[13]</sup> report a multi-armed bandit algorithm, where each ‘arm’ corresponds to a set of conditions, and the algorithm learns which arm to ‘pull’, which corresponds to testing the most general conditions. In their sequential algorithm, the authors also first select the conditions by picking an arm to pull, and subsequently selecting the substrate randomly. However, the multi-armed bandit algorithm requires each arm (set of conditions) to be initialised, *i.e.* every condition needs to be tested once before learning by the algorithm can take place. In a hypothetical example, if a search space consists of 7 bases, 5 solvents and 5 ligands, 175 experiments have to be carried out before the algorithm starts to learn. Thus, their algorithm is only usable for small, well-defined search spaces.

While the above examples are important advances towards data-driven optimisation of general conditions, ultimately they only explore a variety of algorithmic possibilities: both methods do a sequential acquisition of conditions and substrates, but these could also be selected in the joint space, overcoming the assumption of decoupling their selection.<sup>[14]</sup> Furthermore, in their works, the authors only consider the average over all substrates as a generality metric, but others are also practically relevant, as outlined above. With these shortcomings, significant improvements in developing a more experiment-efficient algorithm for general conditions optimisation remain to be explored. As these algorithms need to be efficient to be useful to experimental chemists, such developments are paramount for the adoption in chemical laboratories.

For these reasons, we recently developed *CurryBO*, which is a flexible framework for optimisation towards general reaction conditions. While the full results will be shared in a future publication, a pre-print describing our work has recently been made available.<sup>[15]</sup> In brief, our framework enables exploration of a manifold of algorithmic designs, including selecting conditions and substrates sequentially (as previous work), and jointly, and allows for testing these algorithms on multiple generality metrics. We tested out multiple algorithms on four high-throughput datasets<sup>[16–19]</sup> consisting of multiple conditions evaluated on multiple substrates (thousands of experiments for each dataset) to identify which factors facilitate the optimisation. Through this insight, we are then able to identify an algorithm that significantly outperforms previous state-of-the-art work, including optimisation on only one model substrate.

#### 4. Conclusion and Outlook

In this contribution, we have outlined the problem of the optimisation of general reaction conditions with data-driven methods. While such conditions are a long-standing goal in chemical synthesis, the generality is often not considered *a priori* in reaction optimisation. Data-driven approaches that iteratively propose new conditions and substrates to evaluate experimentally show the potential of facilitating optimisations towards general conditions. Previously proposed algorithms only considered a narrow set of potential algorithms and generality metrics. In contrast, our recently pre-printed work<sup>[15]</sup> explores a vast algorithmic space to identify new state-of-the-art algorithms with sufficient efficiency for application.

While some experimental challenges remain, *e.g.* setting up analytics for all substrates, the described work is a stepping stone towards algorithmic usage in chemical laboratories. To this end, we also made our algorithm available as an easy-to-use website at <https://www.currybo.ethz.ch>, where chemists can use the developed algorithm for optimisation towards general conditions without a single line of code. Studies on the application to library synthesis are currently ongoing, but we invite interested readers to try out the algorithm and contact the authors for collaborations.

#### Acknowledgements

This publication was created as part of the NCCR Catalysis (grant numbers 180544 and 225147), a National Centre of Competence in Research funded by the Swiss National Science Foundation.

Additionally, the authors are grateful to dsm-firmenich and the Swiss Chemical Society for the Best Poster Presentation Award.

#### Author Contributions

S. P. S. wrote the original draft of the manuscript. S. P. S. and K. J. revised the manuscript.

Received: 12.02.2025

- [1] C. C. Wagen, S. E. McMinn, E. E. Kwan, E. N. Jacobsen, *Nature* **2022**, *610*, 680, <https://doi.org/10.1038/s41586-022-05263-2>.
- [2] C. N. Prieto Kullmer, J. A. Kautzky, S. W. Krska, T. Nowak, S. D. Dreher, D. W. C. MacMillan, *Science* **2022**, *376*, 532, <https://doi.org/10.1126/science.abn1885>.
- [3] J. Rein, S. D. Rozema, O. C. Langner, S. B. Zacate, M. A. Hardy, J. C. Siu, B. Q. Mercado, M. S. Sigman, S. J. Miller, S. Lin, *Science* **2023**, *380*, 706, <https://doi.org/10.1126/science.adf6177>.
- [4] I. O. Betinol, J. Lai, S. Thakur, J. P. Reid, *J. Am. Chem. Soc.* **2023**, *145*, 12870, <https://doi.org/10.1021/jacs.3c03989>.
- [5] D. Rana, P. M. Pflüger, N. P. Höfler, G. Tan, F. Glorius, *ACS Cent. Sci.* **2024**, *10*, 899, <https://doi.org/10.1021/acscentsci.3c01638>.
- [6] S. P. Schmid, L. Schlosser, F. Glorius, K. Jorner, *J. Org. Chem.* **2024**, *20*, 2280, <https://doi.org/10.1021/acs.joc.20.196>.
- [7] S. Gallarati, P. van Gerwen, R. Laplaza, L. Brey, A. Makaveev, C. Corminboeuf, *Chem. Sci.* **2024**, *15*, 3640, <https://doi.org/10.1039/D3SC06208B>.
- [8] S. B. Zacate, J. A. Dantas, S. Lin, A. G. Doyle, M. S. Sigman, *Angew. Chem. Int. Ed.* **2025**, *64*, e202511091, <https://doi.org/10.1002/anie.202511091>.
- [9] D. G. Brown, J. Boström, *J. Med. Chem.* **2016**, *59*, 4443, <https://doi.org/10.1021/acs.jmedchem.5b01409>.
- [10] A. W. Dombrowski, A. L. Aguirre, A. Shrestha, K. A. Sarris, Y. Wang, *J. Org. Chem.* **2022**, *87*, 1880, <https://doi.org/10.1021/acs.joc.1c01427>.
- [11] G. Tom, S. P. Schmid, S. G. Baird, Y. Cao, K. Darvish, H. Hao, S. Lo, S. Pablo-García, E. M. Rajaonson, M. Skreta, N. Yoshikawa, S. Corapi, G. D. Akkoc, F. Strieth-Kalthoff, M. Seifrid, A. Aspuru-Guzik, *Chem. Rev.* **2024**, *124*, 9633, <https://doi.org/10.1021/acs.chemrev.4c00055>.
- [12] 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, <https://doi.org/10.1126/science.adc8743>.
- [13] J. Y. Wang, J. M. Stevens, S. K. Kariofillis, M.-J. Tom, D. L. Golden, J. Li, J. E. Tabora, M. Parasram, B. J. Shields, D. N. Primer, B. Hao, D. Del Valle, S. DiSomma, A. Furman, G. G. Zipp, S. Melnikov, J. Paulson, A. G. Doyle, *Nature* **2024**, *626*, 1025, <https://doi.org/10.1038/s41586-024-07021-y>.
- [14] S. Toscano-Palmerin, P. I. Frazier, *SIAM J. Optim.* **2022**, *32*, 417, <https://doi.org/10.1137/19M1303125>.
- [15] S. P. Schmid, E. M. Rajaonson, C. T. Ser, M. Haddadnia, S. X. Leong, A. Aspuru-Guzik, A. Kristiadi, K. Jorner, F. Strieth-Kalthoff, *arXiv* **2025**, <https://doi.org/10.48550/arXiv.2502.18966>.
- [16] A. Buitrago Santanilla, E. L. Regalado, T. Pereira, M. Shevlin, K. Bateman, L.-C. Campeau, J. Schneeweis, S. Berritt, Z.-C. Shi, P. Nantermet, Y. Liu, R. Helmy, C. J. Welch, P. Vachal, I. W. Davies, T. Cernak, S. D. Dreher, *Science* **2015**, *347*, 49, <https://doi.org/10.1126/science.1259203>.
- [17] A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow, S. E. Denmark, *Science* **2019**, *363*, eaau5631, <https://doi.org/10.1126/science.aau5631>.
- [18] J. M. Stevens, J. Li, E. M. Simmons, S. R. Wisniewski, S. DiSomma, K. J. Fraunhoffer, P. Geng, B. Hao, E. W. Jackson, *Organometallics* **2022**, *41*, 1847, <https://doi.org/10.1021/acs.organomet.2c00089>.
- [19] M. K. Nielsen, D. T. Ahneman, O. Riera, A. G. Doyle, *J. Am. Chem. Soc.* **2018**, *140*, 5004, <https://doi.org/10.1021/jacs.8b01523>.

#### License and Terms



This is an Open Access article under the terms of the Creative Commons Attribution License CC BY 4.0. The material may not be used for commercial purposes.

The license is subject to the CHIMIA terms and conditions: (<https://chimia.ch/chimia/about>).

The definitive version of this article is the electronic one that can be found at <https://doi.org/10.2533/chimia.2026.242>