BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

Adaptive experimentation and optimization in organic chemistry

Edited by Philippe Schwaller and Artur M. Schweidtmann

Generated on 07 February 2026, 21:41

Imprint

Beilstein Journal of Organic Chemistry
www.bjoc.org
ISSN 1860-5397
Email: journals-support@beilstein-institut.de

The *Beilstein Journal of Organic Chemistry* is published by the Beilstein-Institut zur Förderung der Chemischen Wissenschaften.

Beilstein-Institut zur Förderung der
Chemischen Wissenschaften
Trakehner Straße 7–9
60487 Frankfurt am Main
Germany
www.beilstein-institut.de

BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

Adaptive experimentation and optimization in organic chemistry

Artur M. Schweidtmann*[1] and Philippe Schwaller*[2,3]

Editorial

Address:
[1]Process Intelligence Research Group, Department of Chemical Engineering, Delft University of Technology, Delft, Netherlands, [2]Laboratory of Artificial Chemical Intelligence (LIAC), Institute of Chemical Sciences and Engineering, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland and [3]National Centre of Competence in Research (NCCR) Catalysis, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Email:
Artur M. Schweidtmann* - a.schweidtmann@tudelft.nl;
Philippe Schwaller* - philippe.schwaller@epfl.ch

* Corresponding author

Beilstein J. Org. Chem. **2025,** *21,* 2367–2368.
https://doi.org/10.3762/bjoc.21.180

Received: 16 October 2025
Accepted: 22 October 2025
Published: 03 November 2025

This article is part of the thematic issue "Adaptive experimentation and optimization in organic chemistry".

Guest Editors: A. M. Schweidtmann and P. Schwaller

The field of organic chemistry is undergoing a remarkable transformation. The convergence of laboratory automation and artificial intelligence is creating unprecedented opportunities for accelerating chemical discovery and optimization [1,2]. This thematic issue explores how adaptive experimentation, automation, and human–AI synergy are reshaping organic chemistry research.

Several key technological advances have enabled this transformation. High-throughput experimentation platforms can now rapidly test large numbers of reaction conditions [3]. Machine learning algorithms can process complex chemical data to identify promising directions [4]. Closed-loop systems can autonomously design, execute, and analyze experiments using machine learning optimization algorithms [5,6]. Together, these capabilities are dramatically increasing the speed and efficiency of chemical optimization with respect to economic and environmental objectives [7].

The contributions in this thematic issue showcase innovative approaches across multiple areas. Quijano Velasco et al. review recent advances in high-throughput automated chemical reaction platforms and machine learning algorithms for reaction optimization, showing how these approaches reduce experimentation time and human intervention [8]. They also discuss current limitations and outline future opportunities for this emerging field. Fralish and Reker demonstrate how active learning on molecular pairs can improve the identification of potent drug candidates [9]. Their ”ActiveDelta” method outperforms standard approaches while maintaining chemical diversity. Schmid et al. provide a comprehensive review of machine learning applications in enantioselective organocatalysis, highlighting both achievements and remaining challenges [10]. Guo et al. present an automated flow chemistry system for nitration reactions, combining kinetic modeling with experimental optimization [11].

However, the contributions in the thematic issue also reveal an important insight: while automation and AI are powerful tools, human chemical intuition remains invaluable. The work of Borup et al. on pK_a prediction illustrates how machine learning can complement rather than replace expert knowledge [12].

Their quantum chemistry-based workflow benefits from chemical understanding in selecting appropriate descriptors and validating predictions. Similarly, Chen and Li review how machine learning-guided optimization strategies are most effective when incorporating chemists' expertise [13].

This theme of human–AI synergy emerges repeatedly. The computational design of asymmetric catalysts by Ferrer et al. demonstrates how AI can accelerate discovery while relying on chemical principles to guide the search space [14]. The most successful approaches combine the rapid exploration capabilities of AI with the deep understanding of experienced chemists.

Looking ahead, several key challenges and opportunities are apparent. Integrating prior knowledge and transfer learning between chemical domains remains difficult but promising. Improved methods for uncertainty quantification could help identify when human oversight is most needed. The development of more interpretable AI models would facilitate collaboration between human and machine intelligence.

The future likely lies not in fully autonomous systems but in thoughtfully designed frameworks that leverage both human and artificial intelligence. As the contributions in this thematic issue demonstrate, combining these complementary strengths can accelerate discovery while maintaining chemical insight and understanding.

We are grateful to all authors who have contributed to this thematic issue. Their work illustrates the tremendous progress in this field and the exciting opportunities ahead. As methods for adaptive experimentation continue to advance, maintaining focus on effective human–AI collaboration will be crucial for realizing the full potential of these technologies in organic chemistry.

This integration of automation, machine learning, and human expertise represents a new paradigm in chemical research. We hope this thematic issue provides valuable perspectives on current capabilities and future directions in this rapidly evolving field.

Artur M. Schweidtmann and Philippe Schwaller

Delft and Lausanne, October 2025

Data Availability Statement

Data sharing is not applicable as no new data was generated or analyzed in this study.

References

1. Houben, C.; Lapkin, A. A. *Curr. Opin. Chem. Eng.* **2015,** *9,* 1–7. doi:10.1016/j.coche.2015.07.001
2. de Almeida, A. F.; Moreira, R.; Rodrigues, T. *Nat. Rev. Chem.* **2019,** *3,* 589–604. doi:10.1038/s41570-019-0124-0
3. Biyani, S. A.; Moriuchi, Y. W.; Thompson, D. H. *Chem.:Methods* **2021,** *1,* 323–339. doi:10.1002/cmtd.202100023
4. Schwaller, P.; Vaucher, A. C.; Laplaza, R.; Bunne, C.; Krause, A.; Corminboeuf, C.; Laino, T. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2022,** *12,* e1604. doi:10.1002/wcms.1604
5. Clayton, A. D.; Manson, J. A.; Taylor, C. J.; Chamberlain, T. W.; Taylor, B. A.; Clemens, G.; Bourne, R. A. *React. Chem. Eng.* **2019,** *4,* 1545–1554. doi:10.1039/c9re00209j
6. Felton, K. C.; Rittig, J. G.; Lapkin, A. A. *Chem.:Methods* **2021,** *1,* 116–122. doi:10.1002/cmtd.202000051
7. Schweidtmann, A. M.; Clayton, A. D.; Holmes, N.; Bradford, E.; Bourne, R. A.; Lapkin, A. A. *Chem. Eng. J.* **2018,** *352,* 277–282. doi:10.1016/j.cej.2018.07.031
8. Velasco, P. Q.; Hippalgaonkar, K.; Ramalingam, B. *Beilstein J. Org. Chem.* **2025,** *21,* 10–38. doi:10.3762/bjoc.21.3
9. Fralish, Z.; Reker, D. *Beilstein J. Org. Chem.* **2024,** *20,* 2152–2162. doi:10.3762/bjoc.20.185
10. Schmid, S. P.; Schlosser, L.; Glorius, F.; Jorner, K. *Beilstein J. Org. Chem.* **2024,** *20,* 2280–2304. doi:10.3762/bjoc.20.196
11. Guo, J.; Su, W.; Su, A. *Beilstein J. Org. Chem.* **2024,** *20,* 2408–2420. doi:10.3762/bjoc.20.205
12. Borup, R. M.; Ree, N.; Jensen, J. H. *Beilstein J. Org. Chem.* **2024,** *20,* 1614–1622. doi:10.3762/bjoc.20.144
13. Chen, L.-Y.; Li, Y.-P. *Beilstein J. Org. Chem.* **2024,** *20,* 2476–2492. doi:10.3762/bjoc.20.212
14. Ferrer, M.; Iribarren, I.; Renningholtz, T.; Alkorta, I.; Trujillo, C. *Beilstein J. Org. Chem.* **2024,** *20,* 2668–2681. doi:10.3762/bjoc.20.224

License and Terms

The definitive version of this article is the electronic one which can be found at:
https://doi.org/10.3762/bjoc.21.180

BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

pKalculator: A pK_a predictor for C–H bonds

Rasmus M. Borup, Nicolai Ree and Jan H. Jensen*

Full Research Paper

Open Access

Address:
Department of Chemistry, University of Copenhagen, Copenhagen,
DK-2100, Denmark

Email:
Jan H. Jensen* - jhjensen@chem.ku.dk

* Corresponding author

Keywords:
C–H pK_a values; p*K*a predictor

Beilstein J. Org. Chem. **2024,** *20,* 1614–1622.
https://doi.org/10.3762/bjoc.20.144

Received: 18 March 2024
Accepted: 02 July 2024
Published: 16 July 2024

This article is part of the thematic issue "Adaptive experimentation and optimization in organic chemistry".

Guest Editor: A. Schweidtmann

Abstract

Determining the pK_a values of various C–H sites in organic molecules offers valuable insights for synthetic chemists in predicting reaction sites. As molecular complexity increases, this task becomes more challenging. This paper introduces pKalculator, a quantum chemistry (QM)-based workflow for automatic computations of C–H pK_a values, which is used to generate a training dataset for a machine learning (ML) model. The QM workflow is benchmarked against 695 experimentally determined C–H pK_a values in DMSO. The ML model is trained on a diverse dataset of 775 molecules with 3910 C–H sites. Our ML model predicts C–H pK_a values with a mean absolute error (MAE) and a root mean squared error (RMSE) of 1.24 and 2.15 pK_a units, respectively. Furthermore, we employ our model on 1043 pK_a-dependent reactions (aldol, Claisen, and Michael) and successfully indicate the reaction sites with a Matthew's correlation coefficient (MCC) of 0.82.

Introduction

Over the years, the ability to selectively break a C–H bond to create new connections has attracted increasing interest [1]. While past methods allowed for C–H transformations in simple molecules, recent synthetic protocols [2] enable selective C–H activation and diversification in larger molecules. This has, for example, attracted the pharmaceutical industry to implement such C–H transformations to diversify different types of molecules ranging from small drug-like molecules to intermediates and lead compounds. Especially late-stage functionalization is a promising emerging field that allows chemists to efficiently explore the chemical space in complex molecules by exchanging a C–H bond with different functional groups to modify the biological activity of drugs [2]. However, pinpointing which C–H bond is reacting can be challenging.

Grzybowski and co-workers recently addressed this gap by predicting pK_a values for C–H bonds in dimethyl sulfoxide (DMSO) using a graph convolutional neural network (GCNN) [3]. Using a mix of experimental and computed pK_a data, they achieved a mean absolute error (MAE) of 2.1 pK_a units. Lee and co-workers also addressed this problem by creating a general machine learning (ML) model using either a neural network or XGBoost. They trained on experimental pK_a values in 39 solvents from the "internet Bond-energy Databank" (iBonD). Thus, they could predict the lowest pK_a value for a wide range of molecules that contain bonds such as N–H, O–H, C–H, S–H, and P–H. However, they reported a scarcity of nonaqueous pK_a values and achieved a MAE of 1.5 pK_a units for the solvent DMSO using XGBoost [4,5]. Unfortunately, neither the Grzybowski group nor the Lee group have made their models generally available to other users.

Inspired by the efforts of the Grzybowski group and the Lee group, we have developed pKalculator, a quantum chemistry (QM)-based workflow for the automatic computation of C–H pK_a values in DMSO. The computed C–H pK_a values are then used to generate training data for an ML model using LightGBM [6]. The QM-based workflow and the ML model are freely available under the MIT license.

Methods

Datasets

We compile a dataset of 732 experimental pK_a values in DMSO from two different sources, Bordwell [7] and iBonD [4]. The Bordwell dataset contains experimental C–H pK_a values in DMSO from 419 molecules. For the iBonD database, we select experimental C–H pK_a values in DMSO for 313 molecules. As the iBonD database only contains an image of each molecule, we employ the "Deep Learning for Chemical Image Recognition" software (DECIMER v. 2.0), developed by Rajan and co-workers [8-10]. While DECIMER converts molecular images into SMILES, manual intervention is required to ensure the SMILES string correctly represents the molecule. Finally, to mirror the dataset by Roszak et al. [3], we also incorporate 43 heterocycles without experimental pK_a values from Shen et al., leaving us with a dataset of 775 compounds [11]. This dataset will be used to calculate QM pK_a values using our QM workflow described in the next section.

We also create a dataset from Reaxys that contains 1043 pK_a-controlled reactions. These reactions include 584 aldol, 408 Claisen, and 51 Michael reactions. This dataset is used as an out-of-sample dataset to see how well our ML model predicts the reaction site. Additionally, we use six pharmaceutical intermediates that undergo selective borylation to compare our QM workflow and ML model with experimentally determined reaction sites.

The quantum chemistry-based workflow

Following work by Ree et al. [12-15], we present a fully automated QM-based workflow for computing C–H pK_a values. A given SMILES string undergoes modifications to produce a list of SMILES for each deprotonated C–H bond. We generate min(1 + 3n_{rot}, 20) conformers for each SMILES using RDKit (v.2022.09.4) [16,17], where (n_{rot}) represents the number of rotatable bonds. Each conformer undergoes optimization in dimethyl sulfoxide (DMSO, ε = 47.2) using GFN-FF-xTB [18] and the analytical linearized Poisson–Boltzmann (ALPB) equation [19] as the implicit solvation model. We then remove conformers with relative energies above 3 kcal/mol and select unique conformers by taking the centroids of a Butina clustering using pairwise heavy-atom root mean square deviation (RMSD) with a threshold of 0.5 Å [16,20]. For more information, refer to Supporting Information File 1, section "Selecting unique conformers".

Subsequently, we re-optimize the remaining conformers in DMSO with GFN2-xTB [21] and the ALPB implicit solvation model to identify the lowest-energy conformer. We then conduct re-optimization in ORCA (v. 5.0.4) [22,23], using the dispersion D4-corrected DFT functional CAM-B3LYP [24,25], the Karlsruhe [26,27] triple-ζ basis set, def2-TZVPPD, and the conductor-like polarizable continuum model (CPCM) [28] as the implicit solvation models. CAM-B3LYP is chosen as the optimal functional based on a benchmark study that evaluates the accuracy of different levels of theory, ranging from semiempirical methods (xTB) [21] over composite electronic structure methods (r^{2}SCAN-3c) [29] to DFT methods (CAM-B3LYP) [24,25]. All these methods are evaluated as single-point calculations or optimization and frequency calculations. For comprehensive details, refer to Supporting Information File 1, section "Benchmark study - computational methods". Hereafter, we check the geometries for imaginary frequencies and use the total thermal energy at 298.15 K. Following the approach of the Grzybowski group [3], we compute the heterolytic dissociation energy through the direct deprotonation reaction, $\mathrm{AH_{solv}} \rightleftharpoons \mathrm{A^-_{solv}}$; see Equation 1.

$$\Delta G^\circ = E\left(\mathrm{A^-_{solv}}\right) - E\left(\mathrm{AH_{solv}}\right). \qquad (1)$$

For each set of deprotonated C–H sites in a molecule, we determine the minimum heterolytic dissociation energy ($\Delta G^\circ_{\mathrm{min}}$). Hereafter, we assume a linear relationship between the experimental pK_a values and $\Delta G^\circ_{\mathrm{min}}$ as this assumption allows us to derive the empirical constants a and b and correct any systematic errors; see Equation 2, where ΔG° is replaced by $\Delta G^\circ_{\mathrm{min}}$. After retrieving the empirical constants a and b, we can determine the QM-computed pK_a values for all deprotonated C–H sites using Equation 2:

$$\mathrm{p}K_\mathrm{a} = a \cdot \Delta G^\circ + b. \qquad (2)$$

Machine learning

The feature descriptor

Recent research shows that the atomic descriptors introduced by Finkelmann et al. [30,31], using charge model 5 (CM5) atomic charges [32], are a great representation of atoms in molecules that can be used in combination with an ML model to predict a variety of properties. These properties encompass the site of metabolism [31,33], the strengths of hydrogen bond donors and acceptors [34-36], and the regioselectivity of electrophilic aromatic substitution reactions [14]. Building on the methodology from Finkelmann et al. [30,31] and Ree et al. [14], we utilize the automated approach to compute CM5 atomic charges from semiempirical tight-binding (GFN1-xTB [37]) calculations. We modify the workflow to enhance the accuracy of the computed CM5 atomic charges. Instead of generating a single random conformer, we produce 20 random conformers from a SMILES string and optimize the structure with molecular mechanics force fields [38] using RDKit [16]. The CM5 atomic charges of the lowest-energy conformer are then used to generate atomic descriptors based on sorting the CM5 charges for a given atom of the input SMILES string. Furthermore, we adjust the shell radius from 5 to 6, improving the performance of the ML model to predict pK_a values as detailed in Supporting Information File 1, section “The descriptor”.

Data preparation and hyperparameter optimization

Building on the procedure outlined by Ree et al. [14], we employ the Optuna framework (v. 3.3.0) [39] to identify optimal hyperparameters for LigthGBM regression and classification models [6]. Specifically, the Bayesian optimization technique utilizing the tree-structured Parzen estimator is applied for hyperparameter space exploration. For the regression task, the target value are the QM-computed pK_a values. For the binary classification task, which aims to predict the site with the lowest QM-computed pK_a value, labels are assigned in the following manner: ‘1’ for the lowest QM-computed pK_a value (true site) and ‘0’ for all other QM-computed pK_a values. As there is sometimes a slight variation between the pK_a value and the other pK_a values, we also introduce a tolerance where a pK_a value within +1 pK_a units or +2 pK_a units of the lowest pK_a value is accepted as ‘1’ to account for these variations, see Supporting Information File 1, section “Machine learning models” for more information. Further, given the significant imbalance between the two classes (with ‘0’s far outnumbering ‘1’s), the hyperparameter *scale_pos_weight* is invoked during hyperparameter optimization. Finally, we establish a “null model” for the classification task, wherein all sites are predicted as ‘0’.

The dataset with QM-computed pK_a values (775 compounds; 3910 pK_a values) is initially split randomly by compound into a training set (80%; 620 compounds; 3121 pK_a values) and a held-out test set (20%; 155 compounds; 789 pK_a values). For each ML model, we carry out a fivefold randomly shuffled cross-validation. Within each fold, the original training set is further split randomly into a new training set (90% of the original training set) and a validation set (10% of the original training set). This allows us to evaluate different models and estimate their performance. Hereafter, each ML model is trained on our original training set and tested against the held-out test set. Finally, we select the best-performing ML model.

Results and Discussion

Computing pK_a values

From section “The quantum chemistry-based workflow” above, we can determine the empirical values *a* and *b* in Equation 2. For each set of deprotonated sites in a molecule, we extract the computed $\Delta G^{\circ}_{\mathrm{min}}$ value and fit it against the experimental pK_a values. Hereafter, we convert the computed $\Delta G^{\circ}_{\mathrm{min}}$ to QM-computed pK_a values using Equation 2. We then inspect outliers that exceed an absolute pK_a unit difference of 5 pK_a units between the experimental pK_a value and the QM-computed pK_a value. We choose an absolute pK_a unit difference of 5 pK_a units to ensure that the QM-computed pK_a is well above the error that is to be expected on the level of theory we are using (CAM-B3LYP). The observed outliers typically result from one of the following reasons: (i) calculation errors concerning the expected minimum pK_a site, (ii) discrepancies between literature structures and database structures, (iii) mislabeled experimental pK_a values, or (iv) extrapolated pK_a values. Notably, the extrapolated pK_a values correspond to compounds beyond the scale measurable in DMSO (pK_a ≥ 35) because of the autoprotolysis of DMSO (pK_a(DMSO) = 35) [40,41]. For more information regarding finding and removing outliers, see Supporting Information File 1, section “Finding outliers”. After multiple iterations, we identified 695 molecules to have reliable experimental pK_a values and computed $\Delta G^{\circ}_{\mathrm{min}}$ values. The values for the computed $\Delta G^{\circ}_{\mathrm{min}}$ are then fitted against the experimental pK_a values, leaving us with empirical constants *a* and *b*; see Figure 1. We now use the derived linear regression to convert all computed ΔG° values into QM-computed pK_a values for our whole dataset (775 compounds). These values are used as target values for the ML part.

Machine learning models for predicting C–H pK_a values

To learn and predict C–H pK_a values, we train a LightGBM regression model with our generated dataset containing QM-computed pK_a values (775 compounds; 3910 pK_a values). Hereafter, we correlate and compare the ML-predicted pK_a values and the QM-computed pK_a values and achieve a MAE and a RMSE of 1.24 and 2.15 pK_a units, respectively, for the held-out test set (155 compounds; 789 pK_a values), as illus-

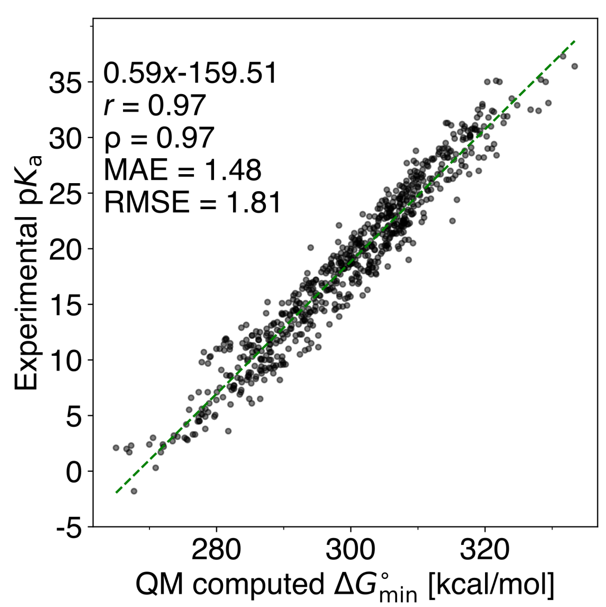


Figure 1: Correlating computed $\Delta G^{\circ}_{\text{min}}$ values and experimental pK_a values for 695 compounds. *r*: Pearson correlation coefficient; ρ: Spearman's rank correlation coefficient; MAE: mean absolute error; RMSE: root mean squared error. QM calculations were carried out at the CAM-B3LYP/def2-TZVPPD CPCM(DMSO)//GFN2-xTB ALPB(DMSO) level of theory.

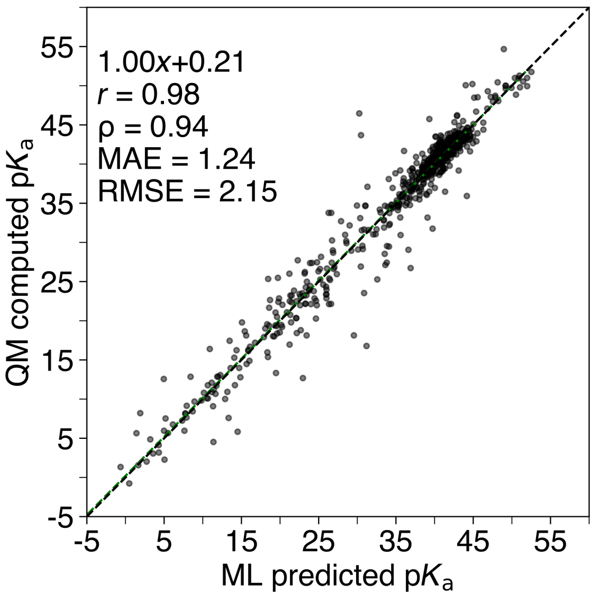


Figure 2: ML-predicted pK_a values vs QM-computed pK_a values of the held-out test set (155 compounds; 789 pK_a values). *r*: Pearson correlation coefficient; ρ: Spearman's rank correlation coefficient; MAE: mean absolute error; RMSE: root mean squared error. All predictions were made using the best ligthGBM regressor. All calculations were carried out at the CAM-B3LYP/def2-TZVPPD CPCM(DMSO)//GFN2-xTB ALPB(DMSO) level of theory.

trated in Figure 2. When zooming in on the ML-predicted pK_a values that are not correlating well with the QM-computed pK_a values, we find C–H sites that are either bridgeheads or where the negative charge is stabilized by resonance. This may be due to the nature of the chosen descriptor vector based on sorted CM5 atomic charges as it may not take into account, for example, steric strain and charge delocalisation. We discuss this further in Supporting Information File 1, section "Outliers for the test set".

We then compare our ML model with previously reported ML models for predicting pK_a values, namely, the GCNN C–H pK_a predictor by Roszak et al. [3] and the XGBoost pK_a predictor by Yang et al. [5]. Roszak et al. [3] used a mix of experimental data (414 compounds) [7], manually curated DFT data (212 compounds), and previously reported DFT data (194 C–H sites) [11]; they obtained a MAE of 2.18 pK_a units for their test set. Yang et al. [5] used filtered entries from the iBonD dataset, comprising 15338 compounds and 19397 pK_a values across 39 solvents [5]. As they not only predict C–H pK_a values, we cannot compare our result with their best ML model. However, they also report a holistic six-solvent (HM-6S) XGBoost model in DMSO (9.3% of the data), which most likely contains the majority of C–H pK_a values. For this XGBoost model, they achieved MAE and RMSE values of 1.53 and 2.35 pK_a units, respectively. A comparison between our ML model, the GCNN model of Roszak et al., and the model of Yang et al. is shown in Table 1. While a direct comparison with these studies is not feasible because of differing datasets, our model surpasses Roszak et al.'s GCNN model by a MAE of 0.94 pK_a units and outperforms Yang et al.'s HM-6S model by a MAE of 0.29 pK_a units.

Table 1: Comparing different ML models for predicting pK_a values. Mean absolute error (MAE) and root mean squared error (RMSE) are provided in pK_a units.

Method	MAE	RMSE
LGBM (this work)	**1.24**	**2.15**
GCNN [3]	2.18	—
XGBoost HM-6S (DMSO)[a] [5]	1.53	2.35

[a] HM-6S: Table 7 in their paper.

Predicting the lowest C–H pK_a value

Now that we can fairly accurately predict pK_a values with our LightGBM regressor, another use case is to be able to identify the C–H site with the lowest pK_a value to predict the site of reaction. For this purpose, we treat the task as a binary classification and train both a LightGBM classifier and a LightGMB regressor. As described earlier in section "Data preparation and hyperparameter optimization", the QM-computed pK_a values are translated into binary values, with ‘1’ representing the lowest QM-computed pK_a value and ‘0’ representing other QM-computed pK_a values. The performance metrics for the test set demonstrate that the regression model (MCC of 0.97) outperforms the classification model (MCC of 0.92) when used as a binary classifier, as seen in Table 2.

Now we train a LightGBM classifier and a LightGMB regressor for the entire dataset (775 compounds; 3910 pK_a values) of QM-computed pK_a values to assess the generalization capability of our ML models. We use an out-of-sample dataset of 1043 pK_a-dependent reactions from Reaxys, containing 584 aldol, 408 Claisen, and 51 Michael reactions. These reactions are chosen because they all involve a deprotonation step, and the C–H site with the lowest pK_a value is most likely the site of the reaction. We also use these reactions for comparison with Roszak et al. [3], who evaluated their GCNN model against 12873 pK_a-controlled reactions, including aldol, Claisen and Michael reactions, and correctly predicted the reacting site with an accuracy of 90.5%. Our out-of-sample set is also used to see how well our ML models predict the site of reaction using the lowest ML-predicted pK_a value.

To understand the result for the out-of-sample set, we show three different reactions in Scheme 1. The first step of the reaction shown in Scheme 1a is an aldol reaction where the deprotonation occurs at the least substituted C–H site next to the ketone (black arrow). Our ML model predicts a pK_a value of 24.7 for the experimental site of reaction. Also, our ML model predicts that the reaction site should be at the highlighted circle. For this site, the ML model predicts a pK_a value of 16.4. It is generally accepted that the most substituted C–H site next to a ketone will form the more stable carbanion (thermodynamic anion), whereas the least substituted carbanion will be the least stable carbanion (kinetic anion). This can generally be controlled by the type of base used. For the reaction in Scheme 1a, *n*-BuLi is commly used, which is known to lead to the kinetic anion. Because our ML model relies on the principle of lowest energy, it predicts the site with the lowest pK_a value as the site of reaction (thermodynamic carbanion) and does not account for the type of base used.

Going to Scheme 1b, we look at a Claisen reaction where the experimental site of reaction occurs at the least substituted ketone. Our ML model predicts the pK_a value here to be 20.5; however, the lowest ML-predicted pK_a value is 4.2. Again, the ML model correctly predicts the most stable carbanion (lowest pK_a value), but other factors come into play when synthesizing compounds.

Last, we have an example of the Michael reaction in Scheme 1c. Here, both the experimental site of reaction and the ML-predicted site of reaction match. Our ML model predicts the lowest pK_a value to be 12.5, whereas the second lowest ML-predicted pK_a value is 21.9 (the least substituted C–H next to a ketone). For more information, see Supporting Information File 1, section “Outliers for Reaxys”.

When we evaluate our ML models on the whole out-of-sample set, we again find that the regression model (MCC of 0.82) outperforms the classification model (MCC of 0.70) when used as a binary classifier as seen in Table 2. While a direct comparison cannot be made between Roszal et al.’s results [3] and ours, we find our result to outperform theirs with an accuracy of 0.96. In general, it is surprising that the LightGBM regressor outperforms our LightGBM classifier as Ree et al. [14] have shown the opposite to be true for electrophilic aromatic substitutions. However, our regression model serves a dual function, that is, it accurately predicts pK_a values and identifies the reaction site.

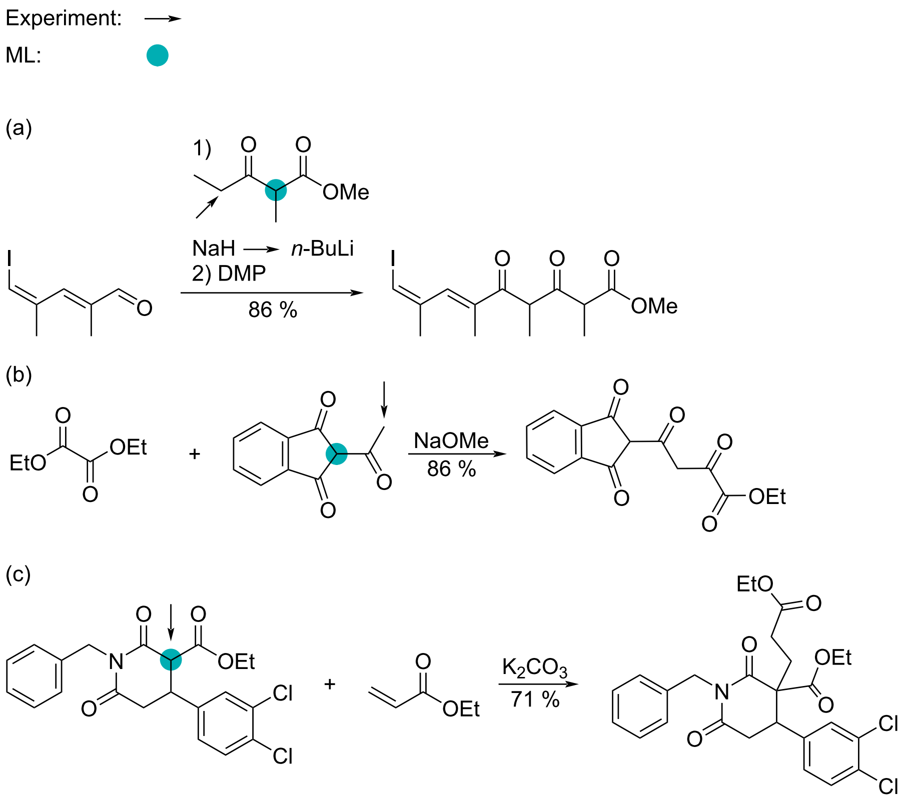
Prediction of aryl C–H borylation sites

In the previous section, we showed that our ML model is able to predict the reaction site for pK_a-dependent reactions. Now, we test the ML model on a more complex reaction type, namely, borylation reactions. Caldeweyher et al. [45] presented a workflow to predict the iridium-catalyzed borylation site of aryl C–H

Table 2: Test set performance metrics: comparison between a LightGBM classifier and a LightGBM regressor for binary classification of the lowest pK_a site. Reaxys performance metrics: comparison between a LightGBM classifier and a LightGBM regressor for binary classification of the reaction site in Reaxys. The best model is marked in bold.[a]

	Test set performance metrics						Reaxys performance metrics					
method	ACC	MCC	PPV	TPR	TNR	NPV	ACC	MCC	PPV	TPR	TNR	NPV
null model[b]	0.80	0	0	0	1.00	0.80	0.87	0	0	0	1.00	0.87
classifier	0.97	0.92	0.97	0.90	0.99	0.98	0.92	0.70	0.64	0.85	0.93	0.98
regressor	**0.99**	**0.97**	**0.97**	**0.98**	**0.99**	**1.00**	**0.96**	**0.82**	**0.84**	**0.84**	**0.98**	**0.98**

[a] ACC: accuracy; MCC: Matthew's correlation coefficient; PPV: precision/positive predictive value; TPR: recall/true-positive rate; TNR: specificity/true-negative rate; NPV: negative predictive value. [b] All predicted pK_a values are “0” to highlight the imbalance of the dataset.



Scheme 1: Predicting the reaction site for three different reactions from the out-of-sample dataset from Reaxys. (a) Aldol reaction, Reaxys reaction ID: 9947221 [42]; (b) Claisen reaction, Reaxys reaction ID: 3402137 [43]; (c) Michael reaction, Reaxys reaction ID: 29819768 [44]. Arrow: experimental site; teal filled circle: ML-predicted lowest pK_a.

bonds (SoBo) [45] and experimentally validated their approach using six pharmaceutical intermediates from medicinal chemistry programs. In the article, they state that *”Iridium catalysts ligated by bipyridine ligands catalyze the borylation of the aryl C–H bonds that are most acidic and least sterically hindered...”*[45]. For this reason, we tested both our QM workflow and the ML model to see how well they identify the reaction site when only considering the lowest aromatic C–H pK_a value; see Figure 3. For both methods, we identify the possible site of reaction if the pK_a value is within 1.5 pK_a units of the lowest pK_a value. This is slightly different from our previous approach. However, because of the higher complexity of the reaction and the similarity of aromatic C–H sites, we purposely allow the QM workflow and the ML model to assess more sites as ‘1’ or true site. When the pK_a value is within 1.5 pK_a units, we also ensure that we are within the range or the uncertainty of the QM-computed pK_a values, which have a MAE of 1.48, as discussed in section “Computing pK_a values”.

For compound **1**, the ML model predicts two low-pK_a sites, indicated by filled circles, of which none corresponds to the experimentally observed site of borylation, indicated by the arrow. However, the QM workflow predicts the correct site as the black ring indicates. Overall, the QM workflow accurately predicts four of the six borylation sites, although, in the case of compounds **2** and **6**, there are additional sites with nearly identical pK_a values. In the case of compound **3**, most chemists would expect the pK_a of pyrazole C–H sites to be considerably lower than those on the benzene ring, suggesting that factors other than pK_a determine the site of borylation for this compound. In the case of compound **5**, the most likely explanation is that the site with the lowest QM-computed pK_a value is sterically hindered compared to the experimentally observed site of borylation. The ML model predicts three borylation sites correctly, but, in the case of compound **5**, there are two additional sites with low pK_a values. One failure is for compound **3**, where the QM workflow also fails; however, for compounds **1** and **4**, the ML model fails, while the QM workflow accurately predicts the site of borylation. This indicates that these compounds are not well represented in the training set.

Conclusion

We introduce pKalculator, an automated QM-based workflow that computes C–H pK_a values with a MAE of 1.48 and a

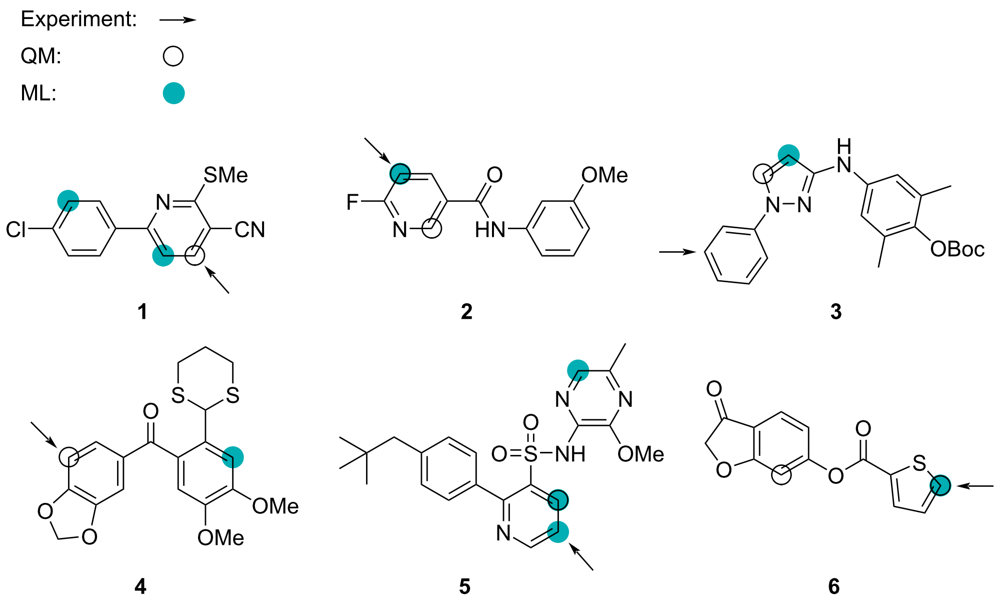


Figure 3: Predicting the site of borylation for a set of six experimentally reported borylation reactions [45]. Arrow: major experimental site/prediction by SoBo; black ring: QM-computed lowest pK_a + 1.5; teal filled circle: ML-predicted lowest pK_a + 1.5.

RMSE of 1.81 when correlating with experimental pK_a values. We use this method to generate training data for an atom-based regression model that delivers fast and relatively precise predictions with MAE and RMSE values of 1.24 and 2.15, respectively, when correlating with QM-computed pK_a values. Both methods are freely available under the MIT license. Our workflow can function as a filtering tool for computer-aided synthesis planning for the synthesis of various pK_a-dependent reactions (aldol, Michael, and Claisen), evidenced by its accurate predictions of reaction sites for 1043 reactions (MCC of 0.82). Looking ahead, we aim to explore more reactions that depend on C–H pK_a values, further enhancing the utility of pKalculator for synthetic chemists. Future iterations will consider factors such as a more extensive and diverse training set, as well as steric hindrance and base reactivity, ensuring even more precise predictions for reaction sites.

Supporting Information

Supporting Information File 1

Additional methods data.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-20-144-S1.pdf]

Funding

This work was funded by the Independent Research Foundation Denmark (DFF; grant number 1032-00129B).

Conflict of Interest

The authors declare that there are no competing interests.

Author Contributions

Rasmus M. Borup: data curation; formal analysis; investigation; methodology; software; visualization; writing – original draft; writing – review & editing. Nicolai Ree: software; supervision; validation; writing – review & editing. Jan H. Jensen: conceptualization; funding acquisition; project administration; supervision; writing – review & editing.

ORCID® iDs

Rasmus M. Borup - https://orcid.org/0000-0002-0878-1345
Nicolai Ree - https://orcid.org/0000-0001-9900-5730
Jan H. Jensen - https://orcid.org/0000-0002-1465-1010

Data Availability Statement

All data that supports the findings of this study is available in the published article and/or the supporting information to this article. The code for the automated workflow and results of the analyzed data are available at https://github.com/jensengroup/pKalculator. Aditional data is available at https://sid.erda.dk/sharelink/EyuyjllJdp. The internet Bond-energy Databank (iBonD) is accessible for non-profit academic use. Due to licensing restrictions for Reaxys, the Reaxys data cannot be shared. We have provided a list of reaction IDs together with our predictions.

Preprint

A non-peer-reviewed version of this article has been previously published as a preprint: https://doi.org/10.26434/chemrxiv-2024-56h5h

References

1. Bergman, R. G. *Nature* **2007,** *446,* 391–393. doi:10.1038/446391a
2. Guillemard, L.; Kaplaneris, N.; Ackermann, L.; Johansson, M. J. *Nat. Rev. Chem.* **2021,** *5,* 522–545. doi:10.1038/s41570-021-00300-6
3. Roszak, R.; Beker, W.; Molga, K.; Grzybowski, B. A. *J. Am. Chem. Soc.* **2019,** *141,* 17142–17149. doi:10.1021/jacs.9b05895
4. iBonD. http://ibond.nankai.edu.cn/ (accessed Oct 27, 2023).
5. Yang, Q.; Li, Y.; Yang, J.-D.; Liu, Y.; Zhang, L.; Luo, S.; Cheng, J.-P. *Angew. Chem., Int. Ed.* **2020,** *59,* 19282–19291. doi:10.1002/anie.202008528
6. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems 30 (NIPS 2017),* Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; Garnett, R., Eds.; Curran Associates, Inc., 2017.
7. Bordwell pKa Table. https://organicchemistrydata.org/hansreich/resources/pka/ (accessed Jan 11, 2023).
8. Rajan, K.; Zielesny, A.; Steinbeck, C. *J. Cheminf.* **2020,** *12,* 65. doi:10.1186/s13321-020-00469-w
9. Rajan, K.; Zielesny, A.; Steinbeck, C. *J. Cheminf.* **2021,** *13,* 61. doi:10.1186/s13321-021-00538-8
10. Rajan, K.; Brinkhaus, H. O.; Agea, M. I.; Zielesny, A.; Steinbeck, C. *Nat. Commun.* **2023,** *14,* 5045. doi:10.1038/s41467-023-40782-0
11. Shen, K.; Fu, Y.; Li, J.-N.; Liu, L.; Guo, Q.-X. *Tetrahedron* **2007,** *63,* 1568–1576. doi:10.1016/j.tet.2006.12.032
12. Ree, N.; Göller, A. H.; Jensen, J. H. *J. Cheminf.* **2021,** *13,* 10. doi:10.1186/s13321-021-00490-7
13. Ree, N.; Göller, A. H.; Jensen, J. H. *ACS Omega* **2022,** *7,* 45617–45623. doi:10.1021/acsomega.2c06378
14. Ree, N.; Göller, A. H.; Jensen, J. H. *Digital Discovery* **2022,** *1,* 108–114. doi:10.1039/d1dd00032b
15. Ree, N.; Göller, A. H.; Jensen, J. H. *Digital Discovery* **2024,** *3,* 347–354. doi:10.1039/d3dd00224a
16. *RDKit 2022_09_4 (Q3 2022) Release;* Zenodo, 2023. doi:10.5281/zenodo.7541264
17. Riniker, S.; Landrum, G. A. *J. Chem. Inf. Model.* **2015,** *55,* 2562–2574. doi:10.1021/acs.jcim.5b00654
18. Spicher, S.; Grimme, S. *Angew. Chem., Int. Ed.* **2020,** *59,* 15665–15673. doi:10.1002/anie.202004239
19. Sigalov, G.; Fenley, A.; Onufriev, A. *J. Chem. Phys.* **2006,** *124,* 124902. doi:10.1063/1.2177251
20. Butina, D. *J. Chem. Inf. Comput. Sci.* **1999,** *39,* 747–750. doi:10.1021/ci9803381
21. Bannwarth, C.; Ehlert, S.; Grimme, S. *J. Chem. Theory Comput.* **2019,** *15,* 1652–1671. doi:10.1021/acs.jctc.8b01176
22. Neese, F.; Wennmohs, F.; Becker, U.; Riplinger, C. *J. Chem. Phys.* **2020,** *152,* 224108. doi:10.1063/5.0004608
23. Neese, F. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012,** *2,* 73–78. doi:10.1002/wcms.81
24. Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004,** *393,* 51–57. doi:10.1016/j.cplett.2004.06.011
25. Caldeweyher, E.; Ehlert, S.; Hansen, A.; Neugebauer, H.; Spicher, S.; Bannwarth, C.; Grimme, S. *J. Chem. Phys.* **2019,** *150,* 154122. doi:10.1063/1.5090222
26. Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005,** *7,* 3297. doi:10.1039/b508541a
27. Rappoport, D.; Furche, F. *J. Chem. Phys.* **2010,** *133,* 134105. doi:10.1063/1.3484283
28. Barone, V.; Cossi, M. *J. Phys. Chem. A* **1998,** *102,* 1995–2001. doi:10.1021/jp9716997
29. Grimme, S.; Hansen, A.; Ehlert, S.; Mewes, J.-M. *J. Chem. Phys.* **2021,** *154,* 064103. doi:10.1063/5.0040021
30. Finkelmann, A. R.; Göller, A. H.; Schneider, G. *Chem. Commun.* **2016,** *52,* 681–684. doi:10.1039/c5cc07887c
31. Finkelmann, A. R.; Göller, A. H.; Schneider, G. *ChemMedChem* **2017,** *12,* 606–612. doi:10.1002/cmdc.201700097
32. Marenich, A. V.; Jerome, S. V.; Cramer, C. J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2012,** *8,* 527–541. doi:10.1021/ct200866d
33. Finkelmann, A. R.; Goldmann, D.; Schneider, G.; Göller, A. H. *ChemMedChem* **2018,** *13,* 2281–2289. doi:10.1002/cmdc.201800309
34. Bauer, C. A.; Schneider, G.; GÃller, A. H. *Mol. Inf.* **2019,** *38,* 1800115. doi:10.1002/minf.201800115
35. Bauer, C. A.; Schneider, G.; GÃller, A. H. *J. Cheminf.* **2019,** *11,* 59. doi:10.1186/s13321-019-0381-4
36. Kuhnke, L.; ter Laak, A.; Göller, A. H. *J. Chem. Inf. Model.* **2019,** *59,* 668–672. doi:10.1021/acs.jcim.8b00758
37. Grimme, S.; Bannwarth, C.; Shushkov, P. *J. Chem. Theory Comput.* **2017,** *13,* 1989–2009. doi:10.1021/acs.jctc.7b00118
38. Tosco, P.; Stiefl, N.; Landrum, G. *J. Cheminf.* **2014,** *6,* 37. doi:10.1186/s13321-014-0037-3
39. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. In *KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining,* 2019; pp 2623–2631. doi:10.1145/3292500.3330701
40. Matthews, W. S.; Bares, J. E.; Bartmess, J. E.; Bordwell, F. G.; Cornforth, F. J.; Drucker, G. E.; Margolin, Z.; McCallum, R. J.; McCollum, G. J.; Vanier, N. R. *J. Am. Chem. Soc.* **1975,** *97,* 7006–7014. doi:10.1021/ja00857a010
41. Koppel, I. A.; Koppel, J.; Pihl, V.; Leito, I.; Mishima, M.; Vlasov, V. M.; Yagupolskii, L. M.; Taft, R. W. *J. Chem. Soc., Perkin Trans. 2* **2000,** 1125–1133. doi:10.1039/b001792m
42. Barbarow, J. E.; Miller, A. K.; Trauner, D. *Org. Lett.* **2005,** *7,* 2901–2903. doi:10.1021/ol050831f
43. Hamama, W. S.; Hammouda, M.; Afsah, E. M. *Z. Naturforsch., B: J. Chem. Sci.* **1988,** *43,* 897–900. doi:10.1515/znb-1988-0716
44. Bettati, M.; Cavanni, P.; Di Fabio, R.; Oliosi, B.; Perini, O.; Scheid, G.; Tedesco, G.; Zonzini, L.; Micheli, F. *ChemMedChem* **2010,** *5,* 361–366. doi:10.1002/cmdc.200900482
45. Caldeweyher, E.; Elkin, M.; Gheibi, G.; Johansson, M.; Sköld, C.; Norrby, P.-O.; Hartwig, J. F. *J. Am. Chem. Soc.* **2023,** *145,* 17367–17376. doi:10.1021/jacs.3c04986

License and Terms

The definitive version of this article is the electronic one which can be found at:
https://doi.org/10.3762/bjoc.20.144

BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

Finding the most potent compounds using active learning on molecular pairs

Zachary Fralish and Daniel Reker*

Full Research Paper

Open Access

Address:
Department of Biomedical Engineering, Duke University, Durham, NC 27708, USA

Email:
Daniel Reker* - daniel.reker@duke.edu

* Corresponding author

Keywords:
active learning; drug design; machine learning; molecular optimization; potency predictions

Beilstein J. Org. Chem. **2024,** *20,* 2152–2162.
https://doi.org/10.3762/bjoc.20.185

Received: 08 April 2024
Accepted: 02 August 2024
Published: 27 August 2024

This article is part of the thematic issue "Adaptive experimentation and optimization in organic chemistry".

Guest Editor: A. Schweidtmann

Abstract

Active learning allows algorithms to steer iterative experimentation to accelerate and de-risk molecular optimizations, but actively trained models might still exhibit poor performance during early project stages where the training data is limited and model exploitation might lead to analog identification with limited scaffold diversity. Here, we present ActiveDelta, an adaptive approach that leverages paired molecular representations to predict improvements from the current best training compound to prioritize further data acquisition. We apply the ActiveDelta concept to both graph-based deep (Chemprop) and tree-based (XGBoost) models during exploitative active learning for 99 K_i benchmarking datasets. We show that both ActiveDelta implementations excel at identifying more potent inhibitors compared to the standard exploitative active learning implementations of Chemprop, XGBoost, and Random Forest. The ActiveDelta approach is also able to identify more chemically diverse inhibitors in terms of their Murcko scaffolds. Finally, deep models such as Chemprop trained on data selected through ActiveDelta approaches can more accurately identify inhibitors in test data created through simulated time-splits. Overall, this study highlights the large potential for molecular pairing approaches to further improve popular active learning strategies in low data regimes by enabling faster and more accurate identification of more diverse molecular hits against critical drug targets.

Introduction

Active learning is a powerful concept in molecular machine learning that allows algorithms to guide iterative experiments to improve model performance and identify the most optimal molecular solutions [1]. Many prominent studies have shown the potential for active learning to accelerate and de-risk the identification of optimal chemical reaction conditions [2-4] and steer molecular optimization for drug discovery [5-8]. Active learning is particularly powerful during early project stages. However, one major downside is that, at these early project stages, only a very small amount of training data is available to learn from [9] which can be insufficient to support the accurate training of data-hungry machine learning models [10,11] and thereby leading to potentially sub-optimal experimental design due to an incomplete understanding of the underlying structure–activity relationship and poor calibration of predictive uncertainty. Additionally, model exploitation can lead to analog identification, which can limit the acquired knowledge and the scaffold diversity of selected hits [1].

We previously showed that leveraging pairwise molecular representations as training data can support molecular optimization by directly training on and predicting property differences between molecules [12]. Compared to classic molecular machine learning algorithms, which are trained to predict absolute property values, such paired approaches are more well-equipped to guide molecular optimization by directly learning from and predicting molecular property differences [12-15] and by cancelling systematic assay errors [12,15]. Beyond superior performance in anticipating property improvements between molecules, the molecular pairing approach shows particularly strong performance on very small datasets by benefiting from combinatorial data expansion through the pairing of molecules [12,13]. Based on these findings, we hypothesized that we could implement exploitative active learning campaigns based on a molecular pairing approach (‘ActiveDelta’) to support rapid identification of the most potent inhibitors across a wide range of benchmark drug targets.

Active learning allows algorithms to guide iterative molecular design by identifying the most valuable next experiment [1]. This can be done by selecting the compounds the model is most uncertain of to improve model performance (‘explorative’) [16,17], retrieving compounds with desired properties (‘exploitative’) [18], or a combination of both (‘balanced’) [8]. Explorative active learning provides diverse chemical structures to support model learning while exploitative approaches instead bias towards rapid identification of favorable compounds. As such, explorative strategies may not propose as many structures with desired characteristics and exploitative strategies may not add much new knowledge for the model [1]. In pursuit of quickly finding potent leads with limited data, we selected to pursue an exploitative active learning approach for this study.

Classically during exploitative active learning, the machine learning model is trained on the available training data and the next compound to be added to the training dataset is selected based on which compound from the learning set has the highest predicted value [19] (Figure 1A). For ActiveDelta learning, training data is paired to learn property differences between molecules [12]. Then, the next compound is selected based on which compound has the greatest predicted improvement from the most promising compound currently in the training dataset (Figure 1B).

For the first time, we here present the ActiveDelta concept and evaluate the Chemprop-based [20] and XGBoost-based [21] implementations of this learning strategy against standard exploitative active learning [19] implementations of Chemprop [20], XGBoost [21], and Random Forest [22] across 99 K_i datasets with simulated time splits [23]. Across these benchmarks, the ActiveDelta approach quickly outcompeted standard active learning implementations, possibly by benefiting from the combinatorial expansion of data during pairing which enables the more accurate training of machine learning algorithms. The ActiveDelta implementations also enabled the discovery of more diverse molecules based on their Murcko scaffolds, possibly due to the ability to learn property differences rather than exploiting analog identification. Finally, the acquired data enabled the Chemprop algorithm to predict the

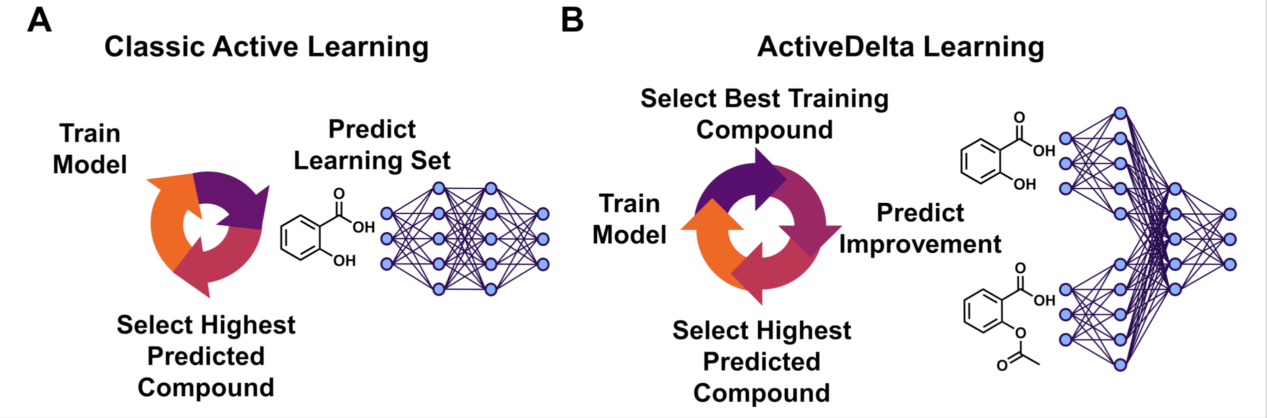


Figure 1: Comparison of active learning approaches. (A) Classic exploitative active learning uses individual molecular representations to predict absolute property values to select the most promising molecule from the learning set to add into the training set. (B) ActiveDelta learning uses paired molecular representations to predict molecular property improvements from the currently best training compound to select the best molecule to add to the training set that is predicted to improve the most compared to the currently best solution.

most promising compounds more accurately in challenging time-split test datasets. Taken together, we believe that the ActiveDelta concept and extensions thereof hold large potential to further improve popular active learning campaigns by more directly training machine learning algorithms to guide molecular optimization and by combinatorically expanding small datasets to improve learning.

Methods

Datasets

Datasets were obtained from Landrum et al. [23] which utilized their simulated medicinal chemistry project data (SIMPD) algorithm to curate and split 99 ChEMBL [24] K_i datasets with consistent values for target id, assay organism, assay category, and BioAssay Ontology (BAO) format into training and testing sets to simulate time-based splits. Datasets were split into training and test sets at an 80:20 ratio. Duplicate molecules were removed. For initial active learning training dataset formation, two random datapoints were selected from each original training dataset and the remaining training datapoints were kept in the learning datasets (Supporting Information File 1, Figure S1). The learning dataset is the pool of molecules that models will select from during active learning [25]. Exploitative active learning was repeated three times with unique starting datapoint pairs. Test sets were not used during active learning but were used only in the test set evaluation of all algorithms.

Model architecture and implementation

To evaluate ActiveDelta with a deep machine learning model, we used the previously established, two-molecule version of the directed Message Passing Neural Network (D-MPNN) Chemprop [20]. For our evaluation with tree-based models, we selected XGBoost [21] with readily available GPU acceleration [26]. Standard, single-molecule machine learning models were implemented using the single-molecule mode of Chemprop [12,27], XGBoost from the XGBoost library [22], and Random Forest models as implemented in scikit-learn [28]. To improve readability, we refer to our predictive pipeline consisting of our molecular pair pre-processing approach and the established two-molecule version of Chemprop as “ActiveDelta Chemprop” (AD-CP) and the standard active learning implementation of single-molecule Chemprop as “Chemprop”. Similarly, we refer to our pairing approach applied to XGBoost as “ActiveDelta XGBoost” (AD-XGB) and the standard single-molecule active learning implementation of XGBoost as “XGBoost”.

The Chemprop-based models were implemented for regression with default parameters and aggregation = ‘sum’ using the PyTorch deep learning framework. For the single-molecule Chemprop implementation, number_of_molecules = 1 while for the ActiveDelta implementation number_of_molecules = 2 to allow for processing of multiple inputs as described previously [29]. We previously optimized the number of epochs for single and paired implementations of Chemprop [12] and observed convergence of performance by 5 epochs for the paired implementation and convergence by 50 epochs for the single-molecule implementation. Based on these results, we set epochs = 5 for the ActiveDelta implementation and epochs = 50 for the single-molecule active learning implementation of Chemprop. XGBoost and Random Forest regression machine learning models were implemented with default parameters and molecules were described using radial chemical fingerprints (Morgan Fingerprint, radius 2, 2048 bits, rdkit.org) when used as inputs for these models. For the ActiveDelta implementation of XGBoost, we used default parameters and concatenated the fingerprints of each molecule in the molecular pairs to create paired molecular representations.

During active learning, standard approaches were trained on the active learning training set, consisting of two datapoints during the first iteration and increasing by 1 datapoint each subsequent iteration of active learning (Supporting Information File 1, Figure S1), and were then used to predict the absolute K_i value of each molecule in the learning dataset. As such, each molecule was processed individually, and predictions were made solely upon the representation of a single molecule. The datapoint with highest predicted potency was then added to the training set for the next iteration of active learning (Figure 1A). Conversely, during ActiveDelta learning, training was performed on the cross-merged training dataset to learn potency differences between molecular pairs as described previously [12]. Then, the single most potent molecule in the training set was paired with every molecule in the learning set to create new pairs for predictions on the learning data (Figure 1B). The second molecule from the molecular pair with highest predicted potency improvement was added to the training set for the next iteration of active learning, resulting in one molecule being added to the active learning training dataset at each iteration which as is commonly done in active learning except when project constraints require batch selection [1]. This datapoint would subsequently be cross-merged with all other training data compounds for ActiveDelta model retraining. For all active learning runs, analysis was repeated three times, each with a random pair of starting molecules for statistical analysis.

Evaluation of model performance and t-SNE analysis

To measure model performance during exploitative active learning, we analyzed the models’ ability to correctly identify the compounds within the top ten percentile of most potent compounds in the learning set. For evaluations on external data, we evaluated model performance after training each model on the 100 molecules this specific model selected during exploitative active learning. The models were evaluated specifically on their ability to correctly identify the top ten percentile of the most potent compounds in the test sets and evaluations were repeated three times with three distinct initial training datasets to investigate the impact of distinct starting points.

The non-parametric Wilcoxon signed-rank test was performed for all statistical comparisons following three repeats of active learning. When presenting the number of the most potent compounds identified by each approach across 3 repeats of the 99 datasets, averages and standard deviations are presented in the text while averages and standard error of the mean are presented in the plots. For plotting of chemical space, molecules were represented by radial chemical fingerprints (Morgan Fingerprint, radius 2, 2048 bits, rdkit.org). Principal component analysis (PCA) was first performed to reduce the 2048 input dimensions to 50 dimensions before t-distributed Stochastic Neighbor Embedding (t-SNE) was applied to further reduce these 50 dimensions to 2 dimensions. PCA and t-SNE were performed with scikit-learn and plotted with matplotlib. Bar plots were created in GraphPad Prism 10.2.0. Source code and datasets used in this work can be downloaded from https://github.com/RekerLab/ActiveDelta.

Results and Discussion

Identifying the most potent leads using active learning on pairs

First, we evaluated how directly learning from and predicting potency differences of molecular pairs affects adaptive learning by directly comparing the performance of specific machine learning algorithms when either applied to molecular pairs or in a classic single-molecule mode. Specifically, we evaluated the ability of the D-MPNN Chemprop and the gradient boosting tree model XGBoost to adaptively learn on molecular pairs using the ActiveDelta approach compared to their standard active learning implementations in single-molecule mode (Figure 1A). As our measure of success, we analyzed all the models' ability to identify the most potent compounds (top ten percentile) during exploitative active learning. We cold-started active learning by selecting only two random datapoints as initial training data and allowed the models to iteratively select the next molecule from the learning set that they predicted as the most potent compound to add to their training data.

When comparing the deep machine learning implementations, we observed interesting patterns. AD-CP initially underperformed compared to the single-molecule implementation of Chemprop, potentially due to the increased complexity of learning and predicting potency improvements between molecular pairs compared to simply identifying analogs of the most promising compound identified so far. However, AD-CP quickly caught up and rapidly (after 35 active learning iterations) outcompeted the single-molecule active learning implementation of Chemprop. We statistically compared the performance differences of the models at 100 and 200 active learning iterations to assess their differences. We noted that AD-CP identified a statistically significantly larger fraction of the top ten percentile of most potent compounds compared to single-molecule Chemprop after 100 iterations of active learning (61% vs 45%, +6.3 leads per dataset on average, $p = 2e - 33$, Figure 2A and Supporting Information File 1, Table S1). This improved performance extended out to 200 iterations where AD-CP had identified almost 90% of the most potent inhibitors (88% vs 79%, +4.3 leads per dataset on average, $p = 4e - 19$, Supporting Information File 1, Table S1). This data overall suggests that, while the learning from and predicting of molecular pairs might be more challenging with very limited data (<35 datapoints), the pairing rapidly enables combinatorial training data expansion that allows the more effective usage of deep neural networks for the identification of the most potent compounds from limited training data until almost all hits in the learning set are selected.

A slightly different pattern emerged when comparing the tree-based implementations. AD-XGB and XGBoost initially selected similar numbers of the most potent molecules, potentially attesting to the more robust training of tree-based models on very small datasets irrespective of whether using single molecule or paired tasks. After 13 iterations, AD-XGB started consistently outperforming XGBoost. We again compared performance statistically after 100 and 200 iterations. We noted that AD-XGB was selecting a significantly larger fraction of the most potent molecules at 100 iterations (62% vs 59%, +1.0 leads per dataset on average, $p = 0.001$, Figure 2B and Supporting Information File 1, Table S1) and at 200 iterations (88% vs. 86%, +0.8 leads per dataset on average, $p = 0.02$, Supporting Information File 1, Table S1). While this difference was not nearly as stark as for the deep neural networks, the identification of an additional lead per project might still provide tangible benefits in risky real-world drug development applications where each additional lead might provide an alternative pathway to mitigate toxicities or other compound liabilities. This further attests to the power of our pairing approach and shows that tree-based machine learning models can also benefit from the pairing to identify the most potent inhibitors in adaptive learning campaigns.

When comparing the performance of the tree-based and the deep neural network-based ActiveDelta approaches, we observed that AD-CP and AD-XGB showed no statistically signif-

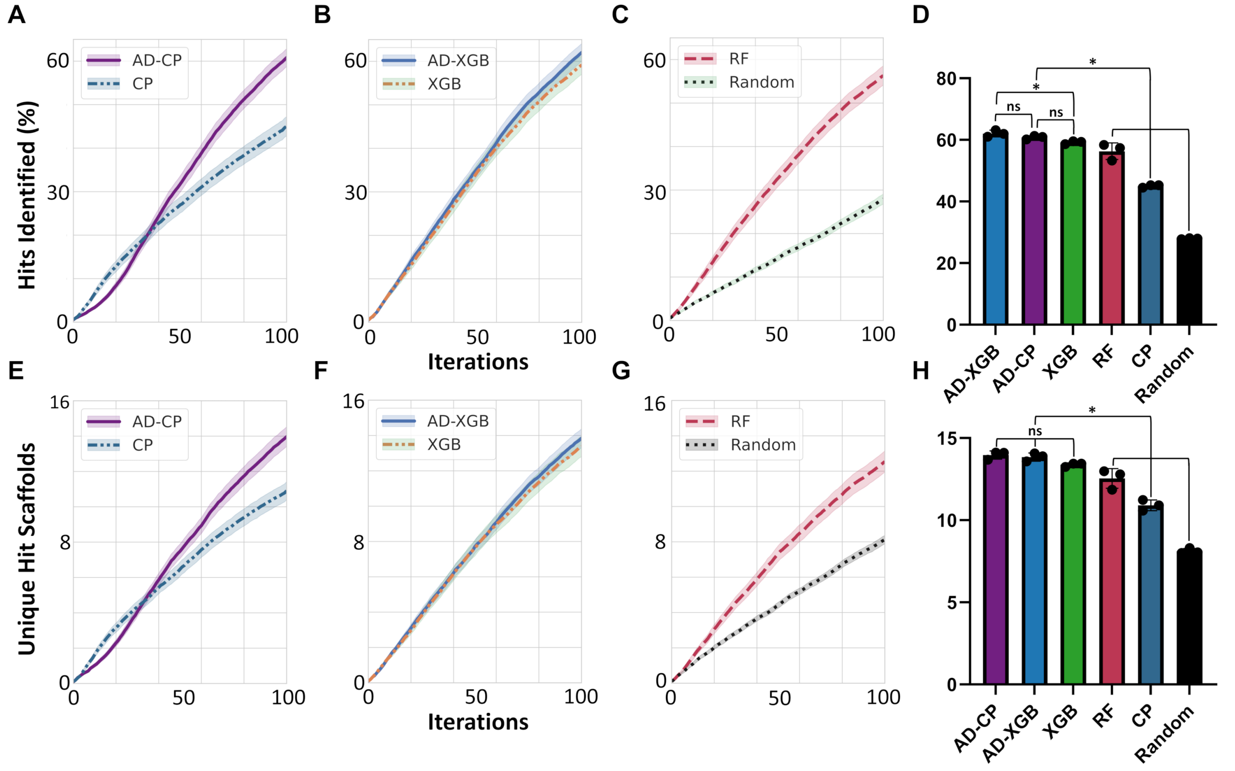


Figure 2: The ActiveDelta approach improves exploitative active learning performance. (A–C) The percentage of the top ten percentile most potent molecules ('hits') in the learning set identified over 100 iterations of active learning by (A) AD-CP and Chemprop (CP), (B) AD-XGB and XGBoost (XGB), and (C) Random Forest (RF) and random selection (Random). (D) Bar charts of the average number of identified hits of each approach at 100 iterations across all the 99 benchmarking tasks. (E–G) The number of unique scaffolds in hits selected over 100 iterations of active learning. (H) Bar charts of the average number of unique scaffolds identified by each approach at 100 iterations across all 99 benchmarking datasets. Average and standard error of the mean for three replicates across 99 K_i datasets after starting with two random datapoints is presented.

icant difference at 100 iterations (p = 0.2, Figure 2A,B, and Supporting Information File 1, Table S1) or 200 iterations (p = 0.7, Supporting Information File 1, Table S1). This suggests that the improved performance of the active learning campaigns is largely driven by the pairing and can be implemented with various underlying, established machine learning algorithms.

We next evaluated how the paired approaches were performing overall compared to standard, single-molecule active learning implementations. AD-CP outcompeted all standard implementations at 100 iterations ($p < 0.002$, Figure 2A–D and Supporting Information File 1, Table S1) except for XGBoost over which it showed a statistically nonsignificant improvement (p = 0.3, Figure 2A–D and Supporting Information File 1, Table S1) while AD-XGB outcompeted all standard implementations at 100 iterations ($p < 0.001$, Figure 2A–D and Supporting Information File 1, Table S1). By 200 iterations, both models using the ActiveDelta approach selected more of the most potent leads than any standard single-molecule active learning approach ($p < 0.04$, Supporting Information File 1, Table S1). These results highlight how a paired approach can allow models to rapidly learn in low data regimes to outcompete standard active learning implementations in identifying the most potent compounds. It also suggests that the Chemprop-based implementation requires more data than the tree-based implementation to outcompete some tree-based standard approaches, potentially hinting at the larger data requirements for deep neural networks even when combinatorially expanding datasets through pairing.

Chemical diversity in molecular selection

Beyond their ability to identify the most potent inhibitors, we sought to determine how these approaches sampled chemical space. When analyzing the scaffold diversity of hits (i.e., the number of unique Murcko scaffolds in the set of molecules selected by the different approaches whose K_i values are within the top ten percentile of the most potent compounds in the complete learning set), AD-CP selected more distinct hit scaffolds than Chemprop (Figure 2E, $p = 5e-25$ at 100 iterations) but AD-XGB's increase in distinct hit scaffolds selected was not statistically significant compared to XGBoost (Figure 2F, p = 0.1 at 100 iterations). In absolute numbers (Figure 2E–H), AD-CP selected 14.0 ± 5.6 (average and standard deviation) distinct scaffolds (59.3% of all scaffolds within the hits), AD-XGB selected 13.8 ± 5.4 (59.2%), XGBoost selected 13.4 ± 5.9 (56.6%), Random Forest selected 12.5 ± 6.1 (53.1%), Chemprop selected 10.9 ± 5.2 (47.0%), and random selection selected 8.1 ± 2.4 (36.0%). AD-CP, AD-XGB, and XGBoost showed no statistically significant differences, but all three approaches outperformed all other approaches at 100 iterations.

When analyzing the scaffold diversity of all selected compounds to understand the chemical diversity of the complete training data and not just the hits, random selection had the highest scaffold diversity of all selection strategies, while AD-CP had the most diverse scaffold selection of all active learning approaches, followed by Chemprop, Random Forest, AD-XGB, and XGBoost ($p < 0.0001$ at 100 iterations, Supporting Information File 1, Figure S2). As such, AD-CP not only finds the most chemically diverse hits, with potential to create multiple lead series to enable further development of distinct scaffolds, but this approach also enriches the scaffold diversity of “negative” training data to improve future compound selection. Although the deep learning-based ActiveDelta models were not able to identify a larger number of hit compounds than the tree-based ActiveDelta implementations here, a deep learning approach appears to be more advantageous to identify more diverse hits by selecting a greater number of distinct scaffolds during exploitative active learning.

Analyzing chemical trajectories

We next investigated how these models traversed chemical space using t-SNE analysis based on radial chemical fingerprints of molecules selected during active learning. For this analysis, we selected the most representative dataset based on similar hit retrieval rates for each algorithm on this dataset compared to the average performance of each algorithm (CHEMBL232-1, Alpha-1b adrenergic receptor). Admittedly, chemical selection trends across datasets are variable, and, as such, the following discussion is not universal but instead is a representation of the overall expected behavior of the algorithms. In the first learning iterations, AD-CP traversed chemical space broadly and jumped between clusters (Figure 3A). During 16–30 iterations, AD-CP showed a balanced behavior with equal numbers of jumps and staying within a cluster. After 30 iterations, AD-CP had identified all the relevant clusters of active compounds and largely stayed within these clusters to rapidly identify potent inhibitors. In contrast, Chemprop was more targeted at the beginning and exploited the one cluster where it could find potent inhibitors (Figure 3B). After that, Chemprop traveled more broadly and was not able to identify all clusters of potent inhibitors even after 45 iterations of learning. As expected, random selection thoroughly sampled chemical space since it is not constrained, consistently jumping between clusters (Figure 3C).

Similar to AD-CP, AD-XGB exhibited broader initial search by jumping between clusters during the first learning iterations and identified a relevant cluster of potent compounds (Figure 4A). During 16–30 iterations, AD-XGB stayed within this relevant cluster until after 30 iterations where it sampled more widely again to quickly identify another relevant cluster that it stayed within to rapidly identify additional potent inhibitors. XGBoost initially showed more targeted behavior where it exploited one cluster and then broadly searched during 16–30 iterations to discover another relevant cluster (Figure 4B). Random Forest immediately exploited the one cluster where it could find potent inhibitors, but after searching more widely it did not identify any other clusters of potent inhibitors by 45 iterations of learning and instead focused on a cluster that did not contain any of the most potent molecules (Figure 4C). Altogether, these results highlight how the ActiveDelta approach can guide models to navigate diverse clusters of distinct chemistries (Figure 2E–H) by learning effectively from the initial phases of wide investigations over chemical space instead of focusing on analog identification to effectively traverse chemical space (Figure 3 and Figure 4) to identify the most potent leads (Figure 2A–D).

For an additional global analysis across all datasets instead of focused on the representative dataset, we calculated the average Tanimoto similarities of the top molecule selected by each model compared to its respective nearest neighbor in the training data using three different molecular representations (Morgan Fingerprints, MACCS Keys, and Atom Pair Fingerprints) during the initial iterations of active learning (1–15, 16–30, and 31–45) across all 99 benchmarking datasets with three repeats (Supporting Information File 1, Table S2). Random selection consistently selected the least similar molecules of all approaches ($p < 0.005$) as expected. Of all active learning approaches, AD-CP consistently selected the least similar molecules ($p < 0.005$). Conversely, Random Forest consistently selected the most similar molecules of all approaches ($p < 0.005$). AD-XGB consistently selected less similar molecules than XGBoost ($p < 0.005$) and initially selected more similar molecules than Chemprop ($p < 0.005$), but later selected less similar molecules compared to Chemprop ($p < 0.005$). Both MPNN-based models (AD-CP and Chemprop) somewhat trended towards selecting compounds with higher similarities with increasing iterations while Random Forest somewhat trended towards less similar compounds. Random selection, XGBoost, and AD-XGB exhibited no consistent trends as iterations advanced. Ultimately, AD-CP and AD-XGB consistently selected more diverse compounds than their base models

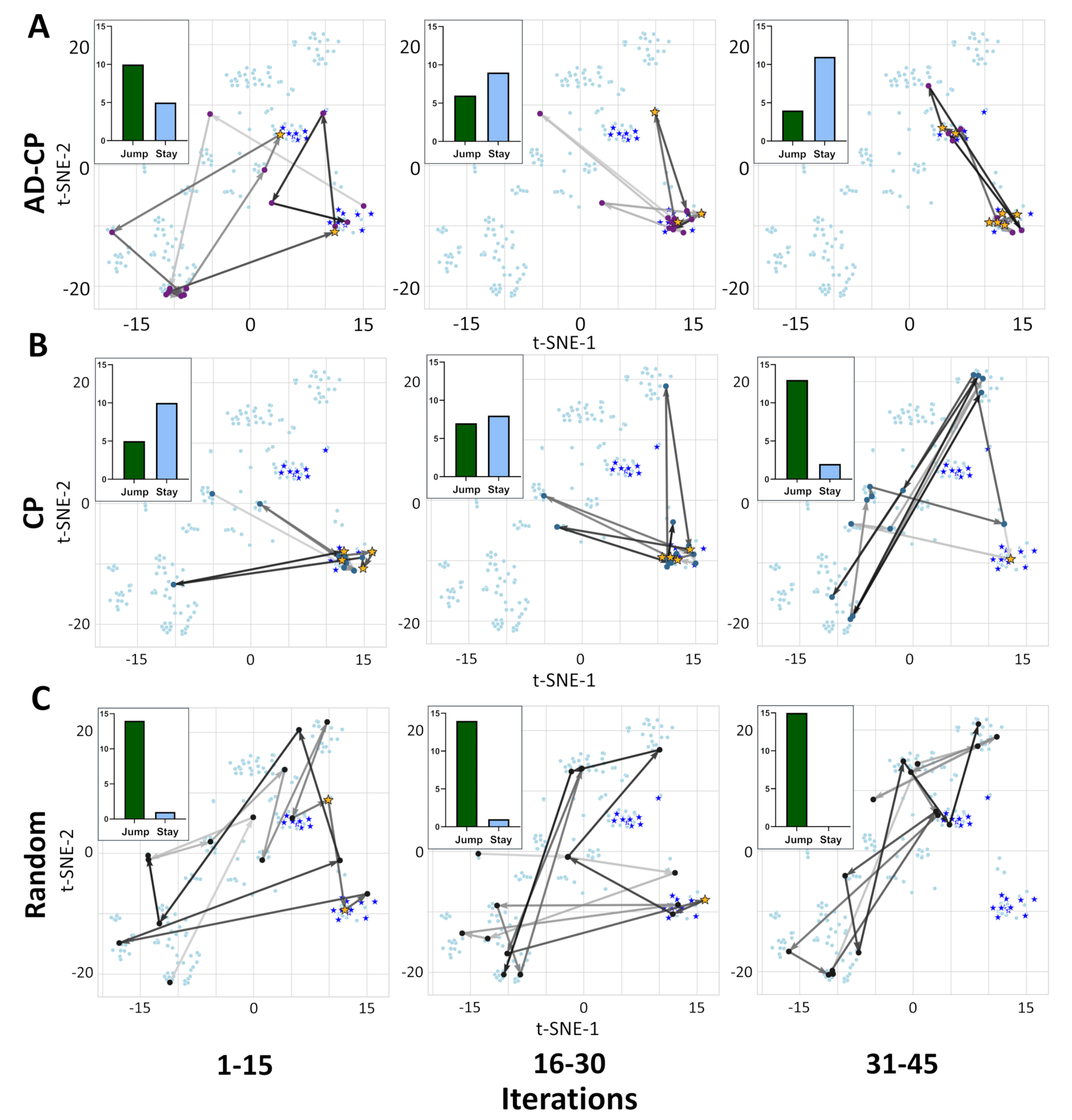


Figure 3: D-MPNN-based model navigation of chemical space. T-SNE of a representative dataset (CHEMBL232-1, Alpha-1b adrenergic receptor) highlighting molecules identified in the first 45 iterations for (A) AD-CP, (B) Chemprop (CP), and (C) random selection (Random). Top ten percentile most potent compounds are shown as stars and identified compounds are highlighted in yellow. The number of times a model 'jumps' from one cluster to another is shown in the inlet with a green bar while the times it 'stays' in the same cluster is shown with a light blue bar. Arrow gradient towards darker grey indicates increasing iteration number.

(Chemprop and XGBoost, respectively, Supporting Information File 1, Table S2) while also identifying more of most potent compounds (Figure 2D) during active learning – further highlighting how the ActiveDelta approach can guide models to rapidly identify more chemically diverse hits while also collecting more diverse training data to augment model knowledge for future compound selection.

Extrapolation to external data

Motivated by the strong ability of ActiveDelta models to effectively navigate the learning spaces, we next sought to see how readily models trained on the selected molecules by active learning could generalize to new data. We used splits that were generated to mimic real-world medicinal chemistry project data sets [23] such that the external data simulates learning from

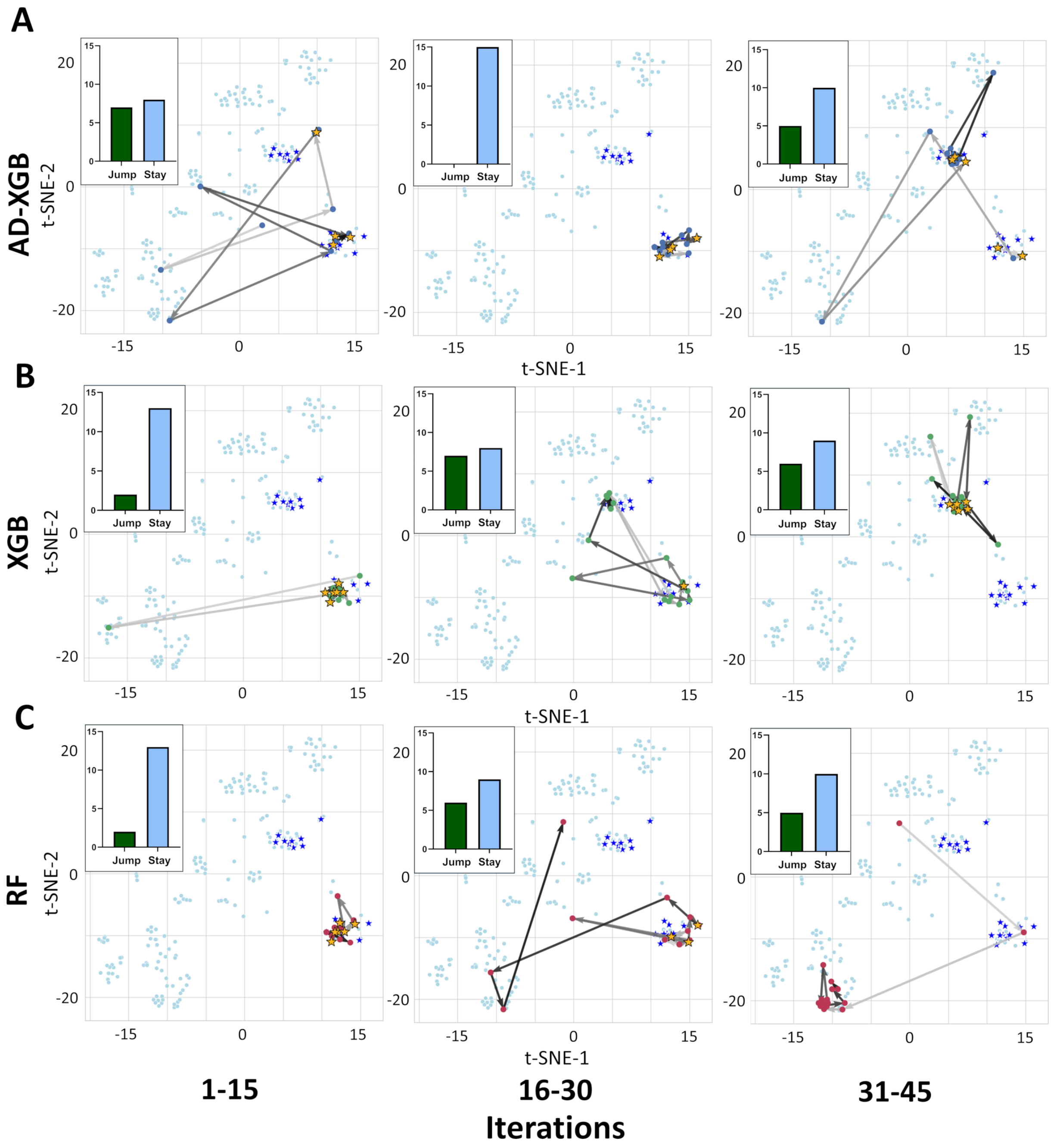


Figure 4: Tree-based model navigation of chemical space. T-SNE of a representative dataset (CHEMBL232-1, Alpha-1b adrenergic receptor) highlighting molecules identified in the first 45 iterations for (A) AD-XGB, (B) XGBoost (XGB), and (C) Random Forest (RF). Top ten percentile most potent compounds are shown as stars and identified compounds are highlighted in yellow. The number of times a model ‘jumps’ from one cluster to another is shown in the inlet with a green bar while the times it ‘stays’ in the same cluster is shown with a light blue bar. Arrow gradient towards darker grey indicates increasing iteration number.

historic data to predict undiscovered “future” compounds instead of simply being selected from a separate cluster based on chemical similarity (Supporting Information File 1, Figure S4). We evaluated all the models’ performances after training on the 100 molecules they each selected from the learning set during exploitative active learning on the task of identifying novel hits (i.e., correctly predicting the top ten percentile of the most potent compounds in the test sets). Across three repeats, AD-CP correctly identified 41.3% ± 18.5 novel hit compounds in the test set on average, AD-XGB identified 40.0% ± 18.9, XGBoost identified 40.0% ± 20.4, Random Forest identified 37.9% ± 20.4, and single-molecule Chemprop identified 27.9% ± 18.7. AD-CP showed a significant improvement over Chemprop (p = 2e − 21), but AD-XGB showed no statistically significant difference compared to XGBoost (p = 0.9), possibly driven by the good performance of XGBoost alone. AD-CP was the only approach to correctly identify 100% of the hits within a test dataset while Random Forest peaked at 89%, AD-XGB and XGBoost peaked 88%, and Chemprop peaked at 83% of correctly identified hits.

In terms of chemical diversity of the novel hits identified in the test set, AD-CP identified 3.3 ± 1.7 (42.5%) of the distinct scaffolds of the novel hit compounds, XGBoost identified 3.2 ± 1.7 (41.4%), AD-XGB identified 3.1 ± 1.6 (40.6%), Random Forest identified 2.9 ± 1.7 (37.9%), and Chemprop identified 2.2 ± 1.5 (28.5%). Similar to hit identification, AD-CP showed a significant improvement over Chemprop (p = 8e − 24) but AD-XGB showed no statistically significant difference compared to XGBoost (p = 0.7). To further evaluate the ability of the algorithms to select diverse hits, we evaluated the Tanimoto similarity of their top selected hits compared to their nearest neighbors in the training data. AD-CP selected the molecules least similar to the training set (0.83 ± 0.16, p = 0.0003, Supporting Information File 1, Table S3), followed by Chemprop (0.85 ± 0.15, p = 1e − 10, Supporting Information File 1, Table S3), XGBoost (0.89 ± 0.11, p = 0.01, Table S3), and then Random Forest (0.90 ± 0.10, Supporting Information File 1, Table S3) and AD-XGB (0.90 ± 0.11, Supporting Information File 1, Table S3). Random Forest and AD-XGB exhibited no statistically significant difference from each other (p = 0.2, Supporting Information File 1, Table S3). The increased diversity in selection from the deep models, that was heightened for our paired approach, highlights how methods that allow for appropriate application of complex models in low data regimes may expand the breadth of molecular predictions based on limited knowledge. Taken together, this data suggests that the Chemprop-based AD-CP is particularly powerful at building models that can generalize to new datasets and thereby will provide medicinal chemists with options to change utilized chemistries later in the project while utilizing knowledge generated from other molecules. Its ability to identify the most diverse scaffolds in hits will also make it a most useful tool to provide medicinal chemists with various lead series for further optimization.

Discussion

Coinciding with increased enthusiasm for machine learning methods to support drug discovery [30,31], expanded use of adaptable laboratory automation [16,32,33] will help support adaptive learning methods like active machine learning to become a cornerstone technology to guide molecular optimizations and discovery [20,34,35]. The ActiveDelta approach for active learning may efficiently guide optimization pursuits by prioritizing the most promising candidates for subsequent evaluation and could be directly integrated into robotic chemical systems to generate more potent leads through iterative design. Beyond pharmaceutical design, we expect these methods to be easily deployable for other chemical endeavors to support material design and prioritization.

Although pairwise methods like ActiveDelta exhibit increased computational costs during active learning given the combinatorial expansion of training data (Supporting Information File 1, Figure S3), these extra datapoints benefit the deep models' abilities to learn the underlying structure–activity relationships more accurately and readily identify the most potent compounds of interest with novel scaffolds. In addition, as active learning is typically conducted for smaller datasets and in early project stages, we foresee that this combinatorial data expansion will be feasible for most active learning pipelines. Furthermore, as real-world experimentation often provides a larger bottleneck than computation, the use of more complex computational architectures with improved hit retrieval rates in place of faster, but less effective, architectures should continue to be a good choice for most real-world projects. In the future, subsampling techniques may be employed to reduce computational costs and even potentially improve performance for paired approaches. For example, it has been shown that similarity-based pairing during training compound generation for Siamese neural networks can significantly improve model efficiency [36]. Additionally, active learning-based subsampling is an autonomous and adaptive approach that has been shown to improve model performance for classification tasks [37]. As the current implementation relies on exhaustive pairing of molecules, it is optimally suited for smaller datasets but allows for data-hungry deep learning models to more adequately learn from limited data amounts. Future work should evaluate the potential of non-exhaustive pairing and subsampling strategies to allow for more efficient application of this method to larger datasets, compare against standard active learning implementations of existing methods that contrast molecules, such as Siamese neural networks [36,38-43], and apply the ActiveDelta approach to these models. Additionally, an adaptive approach that begins with an exhaustive pairing approach in low data regimes and incorporates increasing rates of subsampling as dataset size increases would be worth investigating.

Given the general notion of tree-based models' robustness to training on smaller datasets [44], AD-CP's ability to outcompete standard implementations of tree-based models by only

100 iterations shows particular promise for the application of deep models for low data active learning that are typically particularly troublesome for data-hungry deep learning models [9,10]. This improved performance was maintained when extrapolating to external datasets that were generated to mimic the differences between early and late compounds from true pharmaceutical optimization projects [23], indicating the generalizability of this approach.

Conclusion

Applied to exploitative active learning, the ActiveDelta approach leverages paired molecular representations to predict molecular improvements from the best current training compound to prioritize molecules for training set expansion. Here, we have shown that this approach allows both tree-based and deep learning-based models to rapidly learn from pairwise data augmentation in low data regimes to outcompete standard active learning implementations of state-of-the-art methods in identifying the most potent compounds during exploitative active learning (Figure 2A–D) while selecting more diverse compounds (Figure 2E–H). Our t-SNE analysis suggests that ActiveDelta models will be initially forced to traverse chemical space more broadly to learn property differences between molecules rather than simply identifying analogs of promising hits (Figure 3 and Figure 4) by learning on a pairwise transformation of chemical space. The deep models using this approach also more accurately identified hits in external test sets generated through simulated temporal splits, indicating the ActiveDelta approach's applicability and generalizability to novel chemical structures that would likely be encountered during medicinal chemistry projects. We believe that ActiveDelta and other pairwise approaches show particular promise for adaptive machine learning when training data hungry neural networks on limited data and can serve as accurate platforms to guide lead optimization and prioritization during drug development.

Supporting Information

Supporting Information File 1

Supplementary figures and tables.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-20-185-S1.pdf]

Acknowledgements

We would like to thank the Chemprop, the XGBoost, and the Scikit-learn developers for making their machine learning algorithms publicly available. All computations were run on the Duke Compute Cluster.

Funding

Z.F. is supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate (NDSEG) Fellowship Program. This research was supported by the Duke Science & Technology Initiative and by the NIH NIGMS grant R35GM151255.

Conflict of Interest

D.R. acts as a consultant to the pharmaceutical and biotechnology industry, as a mentor for Start2, and on the scientific advisory board of Areteia Therapeutics.

Author Contributions

Zachary Fralish: conceptualization; data curation; formal analysis; investigation; methodology; software; validation; visualization; writing – original draft; writing – review & editing. Daniel Reker: conceptualization; funding acquisition; methodology; project administration; resources; software; supervision; validation; visualization; writing – review & editing.

ORCID® iDs

Zachary Fralish - https://orcid.org/0000-0001-6293-1730
Daniel Reker - https://orcid.org/0000-0003-4789-7380

Data Availability Statement

Source code and datasets used in this work can be downloaded from https://github.com/RekerLab/ActiveDelta.

References

1. Reker, D. *Drug Discovery Today: Technol.* **2019,** *32–33,* 73–79. doi:10.1016/j.ddtec.2020.06.001
2. Häse, F.; Roch, L. M.; Kreisbeck, C.; Aspuru-Guzik, A. *ACS Cent. Sci.* **2018,** *4,* 1134–1145. doi:10.1021/acscentsci.8b00307
3. Gong, Y.; Xue, D.; Chuai, G.; Yu, J.; Liu, Q. *Chem. Sci.* **2021,** *12,* 14459–14472. doi:10.1039/d1sc02087k
4. Reker, D.; Hoyt, E. A.; Bernardes, G. J. L.; Rodrigues, T. *Cell Rep. Phys. Sci.* **2020,** *1,* 100247. doi:10.1016/j.xcrp.2020.100247
5. Zhang, Y.; Lee, A. A. *Chem. Sci.* **2019,** *10,* 8154–8163. doi:10.1039/c9sc00616h
6. Soleimany, A. P.; Amini, A.; Goldman, S.; Rus, D.; Bhatia, S. N.; Coley, C. W. *ACS Cent. Sci.* **2021,** *7,* 1356–1367. doi:10.1021/acscentsci.1c00546
7. van Tilborg, D.; Grisoni, F. *ChemRxiv* **2023**. doi:10.26434/chemrxiv-2023-wgl32-v2
8. Reker, D.; Schneider, P.; Schneider, G. *Chem. Sci.* **2016,** *7,* 3919–3927. doi:10.1039/c5sc04272k
9. Hinton, G. E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. R. *arXiv* **2012,** 1207.0580. doi:10.48550/arxiv.1207.0580
10. Frey, N. C.; Soklaski, R.; Axelrod, S.; Samsi, S.; Gómez-Bombarelli, R.; Coley, C. W.; Gadepally, V. *Nat. Mach. Intell.* **2023,** *5,* 1297–1305. doi:10.1038/s42256-023-00740-3
11. van Tilborg, D.; Brinkmann, H.; Criscuolo, E.; Rossen, L.; Özçelik, R.; Grisoni, F. *ChemRxiv* **2024**. doi:10.26434/chemrxiv-2024-w0wvl

12. Fralish, Z.; Chen, A.; Skaluba, P.; Reker, D. *J. Cheminf.* **2023,** *15,* 101. doi:10.1186/s13321-023-00769-x
13. Fralish, Z.; Skaluba, P.; Reker, D. *RSC Med. Chem.* **2024,** *15,* 2474–2482. doi:10.1039/d4md00325j
14. Matsumoto, K.; Miyao, T.; Funatsu, K. *ACS Omega* **2021,** *6,* 11964–11973. doi:10.1021/acsomega.1c00463
15. Tynes, M.; Gao, W.; Burrill, D. J.; Batista, E. R.; Perez, D.; Yang, P.; Lubbers, N. *J. Chem. Inf. Model.* **2021,** *61,* 3846–3857. doi:10.1021/acs.jcim.1c00670
16. Desai, B.; Dixon, K.; Farrant, E.; Feng, Q.; Gibson, K. R.; van Hoorn, W. P.; Mills, J.; Morgan, T.; Parry, D. M.; Ramjee, M. K.; Selway, C. N.; Tarver, G. J.; Whitlock, G.; Wright, A. G. *J. Med. Chem.* **2013,** *56,* 3033–3047. doi:10.1021/jm400099d
17. Naik, A. W.; Kangas, J. D.; Sullivan, D. P.; Murphy, R. F. *eLife* **2016,** *5,* e10047. doi:10.7554/elife.10047
18. Besnard, J.; Ruda, G. F.; Setola, V.; Abecassis, K.; Rodriguiz, R. M.; Huang, X.-P.; Norval, S.; Sassano, M. F.; Shin, A. I.; Webster, L. A.; Simeons, F. R. C.; Stojanovski, L.; Prat, A.; Seidah, N. G.; Constam, D. B.; Bickerton, G. R.; Read, K. D.; Wetsel, W. C.; Gilbert, I. H.; Roth, B. L.; Hopkins, A. L. *Nature* **2012,** *492,* 215–220. doi:10.1038/nature11691
19. Reker, D.; Schneider, G. *Drug Discovery Today* **2015,** *20,* 458–465. doi:10.1016/j.drudis.2014.12.004
20. Heid, E.; Greenman, K. P.; Chung, Y.; Li, S.-C.; Graff, D. E.; Vermeire, F. H.; Wu, H.; Green, W. H.; McGill, C. J. *J. Chem. Inf. Model.* **2024,** *64,* 9–17. doi:10.1021/acs.jcim.3c01250
21. Mitchell, R.; Adinets, A.; Rao, T.; Frank, E. *arXiv* **2018,** 1806.11248. doi:10.48550/arxiv.1806.11248
22. Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. *J. Chem. Inf. Comput. Sci.* **2003,** *43,* 1947–1958. doi:10.1021/ci034160g
23. Landrum, G. A.; Beckers, M.; Lanini, J.; Schneider, N.; Stiefl, N.; Riniker, S. *J. Cheminf.* **2023,** *15,* 119. doi:10.1186/s13321-023-00787-9
24. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. *Nucleic Acids Res.* **2012,** *40,* D1100–D1107. doi:10.1093/nar/gkr777
25. Reker, D.; Brown, J. B. Selection of Informative Examples in Chemogenomic Datasets. *Computational Chemogenomics;* Methods in Molecular Biology, Vol. 1825; Humana Press: New York, NY, USA, 2018; pp 369–410. doi:10.1007/978-1-4939-8639-2_13
26. Mitchell, R.; Frank, E. *PeerJ. Comput. Sci.* **2017,** *3,* e127. doi:10.7717/peerj-cs.127
27. Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. *J. Chem. Inf. Model.* **2019,** *59,* 3370–3388. doi:10.1021/acs.jcim.9b00237
28. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. *J. Mach. Learn. Res.* **2011,** *12,* 2825–2830.
29. Vermeire, F. H.; Green, W. H. *Chem. Eng. J.* **2021,** *418,* 129307. doi:10.1016/j.cej.2021.129307
30. Markey, C.; Croset, S.; Woolley, O. R.; Buldun, C. M.; Koch, C.; Koller, D.; Reker, D. *Nat. Comput. Sci.* **2024,** *4,* 96–103. doi:10.1038/s43588-024-00594-8
31. Brown, N.; Ertl, P.; Lewis, R.; Luksch, T.; Reker, D.; Schneider, N. *J. Comput.-Aided Mol. Des.* **2020,** *34,* 709–715. doi:10.1007/s10822-020-00317-x
32. Schneider, G. *Nat. Rev. Drug Discovery* **2018,** *17,* 97–113. doi:10.1038/nrd.2017.232
33. Granda, J. M.; Donina, L.; Dragone, V.; Long, D.-L.; Cronin, L. *Nature* **2018,** *559,* 377–381. doi:10.1038/s41586-018-0307-8
34. Eisenstein, M. *Nat. Biotechnol.* **2020,** *38,* 512–514. doi:10.1038/s41587-020-0521-4
35. Bustillo, L.; Laino, T.; Rodrigues, T. *Chem. Sci.* **2023,** *14,* 10378–10384. doi:10.1039/d3sc03367h
36. Zhang, Y.; Menke, J.; He, J.; Nittinger, E.; Tyrchan, C.; Koch, O.; Zhao, H. *J. Cheminf.* **2023,** *15,* 75. doi:10.1186/s13321-023-00744-6
37. Wen, Y.; Li, Z.; Xiang, Y.; Reker, D. *Digital Discovery* **2023,** *2,* 1134–1142. doi:10.1039/d3dd00037k
38. Altalib, M. K.; Salim, N. *ACS Omega* **2022,** *7,* 4769–4786. doi:10.1021/acsomega.1c04587
39. Fernández-Llaneza, D.; Ulander, S.; Gogishvili, D.; Nittinger, E.; Zhao, H.; Tyrchan, C. *ACS Omega* **2021,** *6,* 11086–11094. doi:10.1021/acsomega.1c01266
40. Altae-Tran, H.; Ramsundar, B.; Pappu, A. S.; Pande, V. *ACS Cent. Sci.* **2017,** *3,* 283–293. doi:10.1021/acscentsci.6b00367
41. Schwarz, K.; Allam, A.; Perez Gonzalez, N. A.; Krauthammer, M. *BMC Bioinf.* **2021,** *22,* 412. doi:10.1186/s12859-021-04325-y
42. McNutt, A. T.; Koes, D. R. *J. Chem. Inf. Model.* **2022,** *62,* 1819–1829. doi:10.1021/acs.jcim.1c01497
43. Jeon, M.; Park, D.; Lee, J.; Jeon, H.; Ko, M.; Kim, S.; Choi, Y.; Tan, A.-C.; Kang, J. *Bioinformatics* **2019,** *35,* 5249–5256. doi:10.1093/bioinformatics/btz411
44. Grinsztajn, L.; Oyallon, E.; Varoquaux, G. *Adv. Neural. Inf. Process. Syst.* **2022,** *35,* 507–520.

License and Terms

The definitive version of this article is the electronic one which can be found at:
https://doi.org/10.3762/bjoc.20.185

BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

Catalysing (organo-)catalysis: Trends in the application of machine learning to enantioselective organocatalysis

Stefan P. Schmid‡1, Leon Schlosser‡2, Frank Glorius*2 and Kjell Jorner*1,3

Review

Open Access

Address:
[1]Institute of Chemical and Bioengineering, Department of Chemistry and Applied Biosciences, ETH Zurich, Zurich CH-8093, Switzerland, [2]Organisch-Chemisches Institut, Universität Münster, 48149 Münster, Germany and [3]National Centre of Competence in Research (NCCR) Catalysis, ETH Zurich, Zurich CH-8093, Switzerland

Email:
Frank Glorius* - glorius@uni-muenster.de; Kjell Jorner* - kjell.jorner@chem.ethz.ch

* Corresponding author ‡ Equal contributors

Keywords:
catalyst design; machine learning; modelling; organocatalysis; selectivity prediction

Beilstein J. Org. Chem. **2024,** *20,* 2280–2304.
https://doi.org/10.3762/bjoc.20.196

Received: 09 May 2024
Accepted: 09 August 2024
Published: 10 September 2024

This article is part of the thematic issue "Adaptive experimentation and optimization in organic chemistry".

Guest Editor: A. Schweidtmann

Abstract

Organocatalysis has established itself as a third pillar of homogeneous catalysis, besides transition metal catalysis and biocatalysis, as its use for enantioselective reactions has gathered significant interest over the last decades. Concurrent to this development, machine learning (ML) has been increasingly applied in the chemical domain to efficiently uncover hidden patterns in data and accelerate scientific discovery. While the uptake of ML in organocatalysis has been comparably slow, the last two decades have showed an increased interest from the community. This review gives an overview of the work in the field of ML in organocatalysis. The review starts by giving a short primer on ML for experimental chemists, before discussing its application for predicting the selectivity of organocatalytic transformations. Subsequently, we review ML employed for privileged catalysts, before focusing on its application for catalyst and reaction design. Concluding, we give our view on current challenges and future directions for this field, drawing inspiration from the application of ML to other scientific domains.

Introduction

Since the beginning of the 21st century, organocatalysts [1] have established themselves as a third group of homogeneous catalysts, next to biocatalysts [2] (enzymes) and transition metal-based catalysts [3]. In particular, enantioselective organocatalysis has shown an impressive rise in the last decades, owing to the tunability of catalysts and different modes of activation, enabling a manifold of different transformations [4,5]. The development of the field, driven by many researchers, led to the award of the Nobel Prize to List and MacMillan in 2021 ‘for the development of asymmetric organocatalysis’. Organocatalytic transformations have also seen the transition to industrial processes for the production of a variety of pesticides and medicinal compounds, as recently reviewed [6-9].

Despite the prominence of organocatalytic reactions, catalyst development has so far mostly been conducted guided by intuition of skilled organic chemists. Given that organocatalytic reactions are governed by different competing interactions, the influence of a change in molecular structure is often non-trivial, even for highly experienced experts. Thus, intuition-guided catalyst development is regarded as suboptimally efficient and furthermore highly subjective to the experience of the chemists carrying out the study [10-15]. Considering the demand of organocatalysts, their accelerated and reliable development is highly desirable [16]. In the spirit of accelerated discovery, the development of organocatalysts has been augmented with computational catalyst design [17,18]. Multiple programs for automated catalyst simulation have been developed in the last decade. Notable examples include the development of ACE (Asymmetric Catalyst Evaluation) [19,20], AARON (Automated Reaction Optimiser for New Catalysts) [21] or CatVS (Catalyst Virtual Screening) [22]. Such tools have been extensively reviewed in the past years [23-25]. Based on a known mechanism, the tools calculate the energies of relevant species either via force field or quantum chemical methods to assess the properties of a reaction such as activation energies or selectivity. Irrespective of the degree of automation, in silico calculations are often less time-sensitive than wet-lab experiments and can be used to reduce the number of required experiments. As such, these methods contribute to the acceleration of catalyst discovery, for example through high-throughput virtual screening.

Predating these computational techniques is the desire to understand and explain experimental outcomes in organic chemistry with physicochemical descriptors. A prominent early example are Hammett parameters, developed in 1937 [26,27], that relate substituent parameters to the equilibrium constant of the deprotonation of a substituted benzoic acid. The derived substituent parameters are used to gain insight into the mechanism of reactions by observing the influence of substituents on a reaction outcome. However, Hammett parameters have shown to not fully describe observed trends. Therefore, complementary representations capturing other properties of a molecule have been derived (vide infra) [28].

While traditional linear free energy relationships such as those using Hammett parameters used linear models, the emergence of ML has led to the development of more complex algorithms, better suited for extracting hidden patterns in data. The ability of ML to efficiently capture complex relationships allows to extract influences on catalyst properties and thus makes it suited towards the accelerated design of chemicals and materials, including organocatalysts [29]. Due to this potential, an increasing number of research groups have used ML to predict and develop new organocatalytic reactions.

This review aims to provide a critical overview of developments in ML specifically for organocatalysis over the last decade, with a focus on its applications. We aim to provide a starting point to catalysis researchers who are interested in ML as well as an assessment of critical challenges to more experienced ML users. We will first give a primer on ML, equipping experimentalists with the knowledge necessary to follow the developments in the field. The rest of the review is divided into three parts: (1) ML for reactivity and selectivity prediction, (2) ML for the design of privileged organocatalysts and (3) ML for catalyst and reaction design. Ultimately, the review will give an outlook on the authors' expectation of the future of the field.

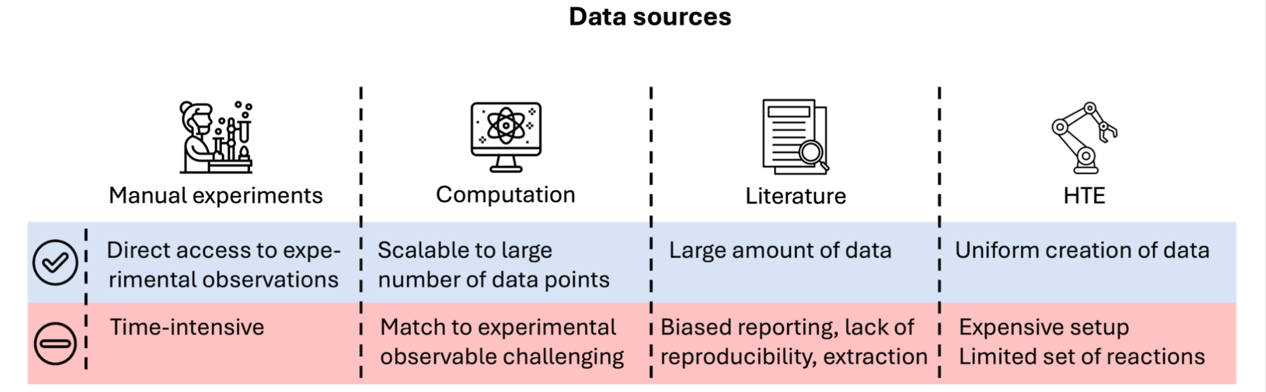
Review

1. Primer on ML

1.1 Data

The foundation for any predictive model is the underlying data. It represents the source from which the model extracts relevant patterns and relations. Therefore, the size and quality of the underlying dataset will determine the model's predictive capabilities. To obtain high predictive accuracy for a broad range of problems, a data set is sought which covers the problem space comprehensively. This does not only encompass the chemical diversity of the included molecules, but also the range of results, e.g., reactions with low, medium and high selectivity [30]. Predictions for data points outside of the applicability domain, e.g., the region which is not sufficiently covered by the provided training data, are less reliable, which is why an appropriate choice of training data is paramount for predictive modelling. Depending on the problem at hand, different sources of data are available (Figure 1).

Apart from experimental data, the creation of large amounts of in silico data is possible with sufficient computational resources [31,32]. While this approach is useful in cases where the experimental determination is challenging, some experimental properties, like the reaction yield, remain elusive to be reliably computed due to the myriad of factors (side-reactions, impurities, solvation effects, interface effects,...) that influence this observable [33,34]. Another pitfall regarding computational data is its accuracy with respect to the ground truth, in particular for multiple factors relevant throughout catalysis, such as noncovalent interactions (NCIs) for organocatalysis or spin properties for transition metal catalysis [35,36]. While most quantities can in principle be computed with the highest accuracy using



Data sources

	Manual experiments	Computation	Literature	HTE
✓	Direct access to experimental observations	Scalable to large number of data points	Large amount of data	Uniform creation of data
⊖	Time-intensive	Match to experimental observable challenging	Biased reporting, lack of reproducibility, extraction	Expensive setup Limited set of reactions

Figure 1: Schematic depiction of available data sources for predictive modelling, each with its advantages and disadvantages. Icon 'Manual experiments' made by Eucalyp from flaticon.com. This content is not subject to CC BY 4.0. Icon 'Computation' made by Wichai.wi from flaticon.com. This content is not subject to CC BY 4.0. Icon 'Literature' made by Muhammad Atif from flaticon.com. This content is not subject to CC BY 4.0. Icon 'HTE' made by Nuricon from flaticon.com. This content is not subject to CC BY 4.0. Icon 'Pros' made by Aldo Cervantes from flaticon.com. This content is not subject to CC BY 4.0. Icon 'Cons' made by Yogi Aprelliyanto from flaticon.com. This content is not subject to CC BY 4.0.

advanced tools, the associated computational cost needs to be considered [18,24].

Therefore, the use of experimental data is advantageous as less assumptions have to be made and the quantity of interest is directly represented. The results of a great number of experiments can be found in literature, as well as patents. Manual curation of this data is possible, but for larger amounts of data it is usually impractical. Therefore, automated extraction tools have been reported yielding the data in a structured format suitable for ML [37-41]. While some important efforts have been made to establish uniform data reporting standards [42,43], they are getting picked up by the community rather slowly. With data from experiments conducted by different scientists under varying conditions and adhering to various standards, reproducibility remains a major challenge in organic chemistry and restricts the applicability of literature data for statistical modelling [30]. Despite emerging high-throughput experimentation (HTE) pipelines [44,45], large datasets of high-quality are still scarce. While multiple large datasets are available for transition metal catalysis [46-48] and biocatalysis [49-51], they are however not common for organocatalysis. Therefore, much research has been devoted to develop models that perform well on the available small data sets [52,53].

1.2 Representation

In order to be processed by any ML model, the data needs to be provided in a machine-readable way. Unlike chemists who typically use drawings of Lewis structures to represent molecules, computers require a numerical representation of the molecular structure. Since the information that describes the input directly influences what relationships a model can learn from the presented data, different representations might be suitable depending on the task.

Besides the most commonly used string-based representations, such as the Simplified Molecular Input Line Entry Specification (SMILES) [54] and fingerprints like the extended connectivity fingerprint (ECFP) [55], molecules can be directly represented as graphs (Figure 2).

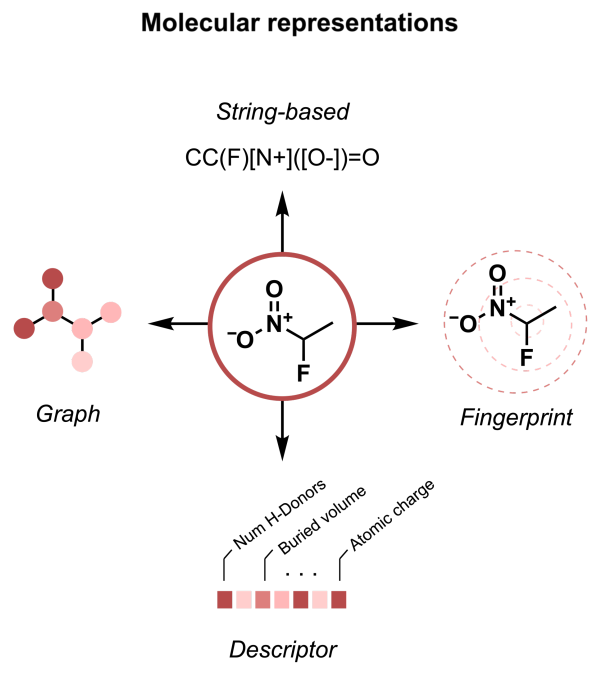


Figure 2: Schematic depiction of different kinds of molecular representations for fluoronitroethane. Among the most common representations are string-based notations, such as SMILES, structural fingerprints, like the ECFP, or molecular graphs. Another way of encoding a molecule is through descriptors that often contain steric or electronic properties.

In graphs, the atoms and bonds are represented as nodes, and edges, respectively [56]. While these kind of representations are well suited for the description of most organocatalysts with distinct bonds, they have limitations when describing coordination compounds as commonly found in transition metal catalysis for example [57].

Another kind of representation that has found considerable application for ML in organocatalysis, is the use of descriptors. These are sets of numerical or categorical values to encode a molecule. A plethora of descriptors with varying degree of computational effort for their calculation are available. Among the most commonly employed descriptors in organocatalysis are steric and electronic descriptors. Section 2.1 provides a detailed overview of examples where different kind of descriptors have been successfully applied for predictive modelling in organocatalysis. In contrast to the representations through graphs, or SMILES, which can be directly obtained from the molecular structure, the selection of appropriate descriptors is problem-specific and requires knowledge about the fundamental interactions governing the reaction outcome. Hence, making the selection of input features a key step for successful modelling [58-63].

1.3 Modelling

The third important requirement for building a predictive model is the model architecture. Generally, ML algorithms can be divided into reinforced, unsupervised and supervised learning. In reinforcement learning, an agent is trained to make decisions by interacting with an environment, receiving feedback in the form of rewards or penalties, and adjusting its behaviour to maximise cumulative rewards over time [64].

While reinforcement learning has not yet found widespread application in organocatalysis, supervised and unsupervised learning are widely employed techniques. The latter uses unlabelled data (e.g., data without a label or numerical value), to identify patterns and relationships within the provided data. Popular tools are Principal Component Analysis (PCA), Uniform Manifold Approximation and Projection (UMAP) [65], or t-distributed Stochastic Neighbour Embedding (t-SNE) [66], which have found application in organocatalysis to reduce the dimension of the respective reaction space, e.g., for visualization purposes. Another widely applied unsupervised ML technique is clustering, which aims to group similar data points together and thus enables a diverse selection by uniformly sampling from the created space [67,68]. Supervised learning requires labelled data and aims at identifying correlations between the target values and the corresponding input features. In the context of addressing chemical problems, this can be used to correlate reaction specific features with the reaction outcome, such as the yield or selectivity. A plethora of different supervised learning algorithms are available and a priori knowledge which architecture works best is challenging. Some of the most widely used algorithms include multivariate linear regression (MLR) [69] in which the target is linearly modelled by multiple independent variables. Other notable architectures include decision trees [70], support vector machines [67] and deep neural networks [71,72]. While the accuracy of the model is paramount, interpretability is also highly desirable. In this regard, MLR bears the advantage that it yields a directly interpretable function which can be used for mechanistic inference. However, it is important to note that the caveat of correlation and causality must be considered. Also, for other kind of models, e.g., random forests, it is common practice to consider the importance of individual features for the model's prediction to gain mechanistic insight. Careful attention must be paid to the collinearity of features [73], such that they are not too strongly related to each other, which complicates any quantitative interpretation of feature importance. Thus, thorough analysis and special strategies to address collinearity, such as hierarchical clustering [74] or threshold-based pre-selection [75] have to be considered to ensure reliable interpretability [69].

It is worth mentioning that all the above-mentioned techniques are not limited to applications in organocatalysis but are used for a wide variety of chemical problems.

2 ML for selectivity predictions

In the context of organocatalysis, for a majority of published work, the reaction property of interest is the selectivity (either enantio- or diastereoselectivity), which is predicted as the difference in energies between the selectivity-governing transition states $\Delta\Delta G^{\ddagger}$ (Figure 3).

Whereas the application of the above described representations and models to such problems is rather modern, the interest to describe the influence of substrate or catalyst structures on the rate or selectivity of a reaction is well-established and led among others to the introduction of Hammett parameters to relate chemical structures to both kinetic and thermodynamic reaction properties [28] (Figure 4).

As Hammett parameters account only for the electronic effect of substituents, much research has been devoted to develop physical-organic descriptors, which consider steric effects and separate the electronic effect into contributions from resonance and induction, among others [27,77-81].

In this chapter, we first discuss the evolution of physical-organic descriptors for the representation of organocatalysts [82]. Later, we examine the effects of increasing data availability towards the application of ML in this field.

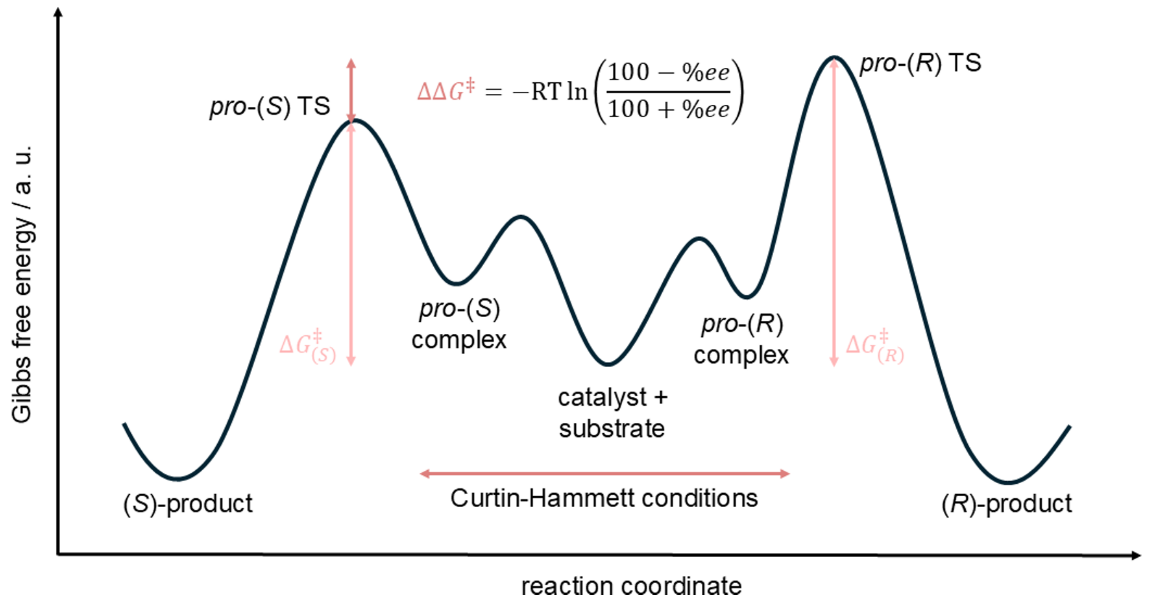
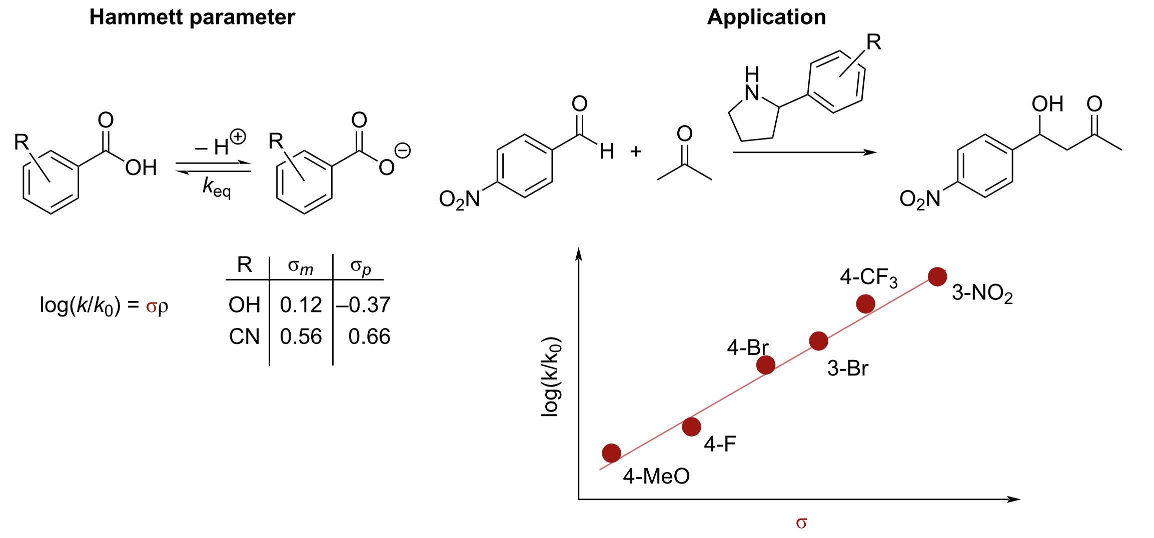

Gibbs free energy / a. u.
reaction coordinate
$\Delta\Delta G^{\ddagger} = -\mathrm{RT}\ln\left(\frac{100-\%ee}{100+\%ee}\right)$
pro-(S) TS
pro-(R) TS
$\Delta G^{\ddagger}_{(S)}$
$\Delta G^{\ddagger}_{(R)}$
pro-(S) complex
pro-(R) complex
catalyst + substrate
(S)-product
Curtin-Hammett conditions
(R)-product

Figure 3: Depiction of the energy diagram of a generic enantioselective reaction. In the centre, catalyst and substrate are separated. They associate with each other to either the pro-(*R*) or pro-(*S*) complex, with all these reactions taking place in a fast equilibrium (Curtin–Hammett conditions). From these complexes, the products are formed via separate transition states. The energy difference between these two transition states is termed $\Delta\Delta G^{\ddagger}$ and determines the selectivity.



R	σ_m	σ_p
OH	0.12	–0.37
CN	0.56	0.66

Figure 4: Hammett parameters are derived from the equilibrium constant of substituted benzoic acids (example from Rogers et al. [76] to correlate Hammett parameters of the arylpyrrolidine catalysts to the reaction kinetics of the aldol reaction).

2.1 Evolution of physical-organic descriptors in organocatalysis

Drawing inspiration from linear free energy relationships, MLR models, pioneered by Norrby and co-workers [83] and later further developed by Sigman and co-workers [69,82], are commonly used for the prediction of enantioselectivity. In such models, the substrates, catalysts, and other relevant reaction species are encoded via a suitable representation of expert-chosen descriptors. Subsequently, the target property of interest, commonly $\Delta\Delta G^{\ddagger}$, is fitted to the representation via a linear fit of the form $y = m_1x_1 + m_2x_2 + \ldots + m_nx_n + k$, where y is the target property, m_1, ... , m_n are the regression coefficients, k is the offset and x_1, ..., x_n are the molecular descriptors. The regression coefficients are also indicative of the importance of the respective molecular parameter. Thus, MLR models provide the capability to directly interpret the prediction results and form mechanistic hypotheses based on the importance of distinct descriptors.

Given the importance of the chosen representation, the search for descriptive parameters has always been a cornerstone in this field. While Taft [77] and Charton [81] describe steric properties as singular substituent values, Harper et al. [60] showed that a singular value is insufficient to represent steric substituent properties. Instead, the authors used Sterimol parameters [84] as steric descriptors (Figure 5), showing superior correlations towards the enantioselectivity for a multitude of organocatalytic reactions.

Sterimol parameters are calculated from a given 3D structure and consist of three parameters, describing the minimum and maximum (rotational) width as well as the depth of a substituent. Nowadays, Sterimol parameters are established as standard parameters to describe steric residue properties. Since Sterimol parameters are calculated from a 3D structure, it is important to include information from relevant conformers. To avoid losing important information from discarding conformers, Paton and co-workers [85] introduced wSterimol, which takes into account structures from the entire conformer ensemble via Boltzmann-weighting. The authors used their descriptors for the prediction of the enantioselectivity for several previously reported reactions, showing improved prediction performance compared to non-Boltzmann-weighted Sterimol parameters. Apart from considering parameters of the entire conformer ensemble, it has been shown that informative models can be developed by considering active structures. This was demonstrated by Crawford et al. [86] in their investigation of a peptide-catalysed atroposelective bromination (Figure 5). The authors found that the peptidic catalysts can broadly be defined in two categories of β-turns: a type I' pre-helical and type II' β-hairpin. Even though the latter was consistently lower in ground state energy (up to 6 kcal/mol for some catalysts), predictive models for enantioselectivity were found for both catalyst conformers in separate MLR models. For organophosphorous ligands of transition metal complexes, the minimum buried volume in a conformer ensemble was identified to determine the ligation state towards a metal centre as either mono- or bis-ligated and thus providing a threshold for catalytically active ligands [87]. All of these examples demonstrate that not only the type of descriptor is important, but also the structure for which the descriptors are considered. This can either be ensured by expert-knowledge of preselecting relevant structures, for example based on a known mechanism, or by considering information from the entire conformer ensemble.

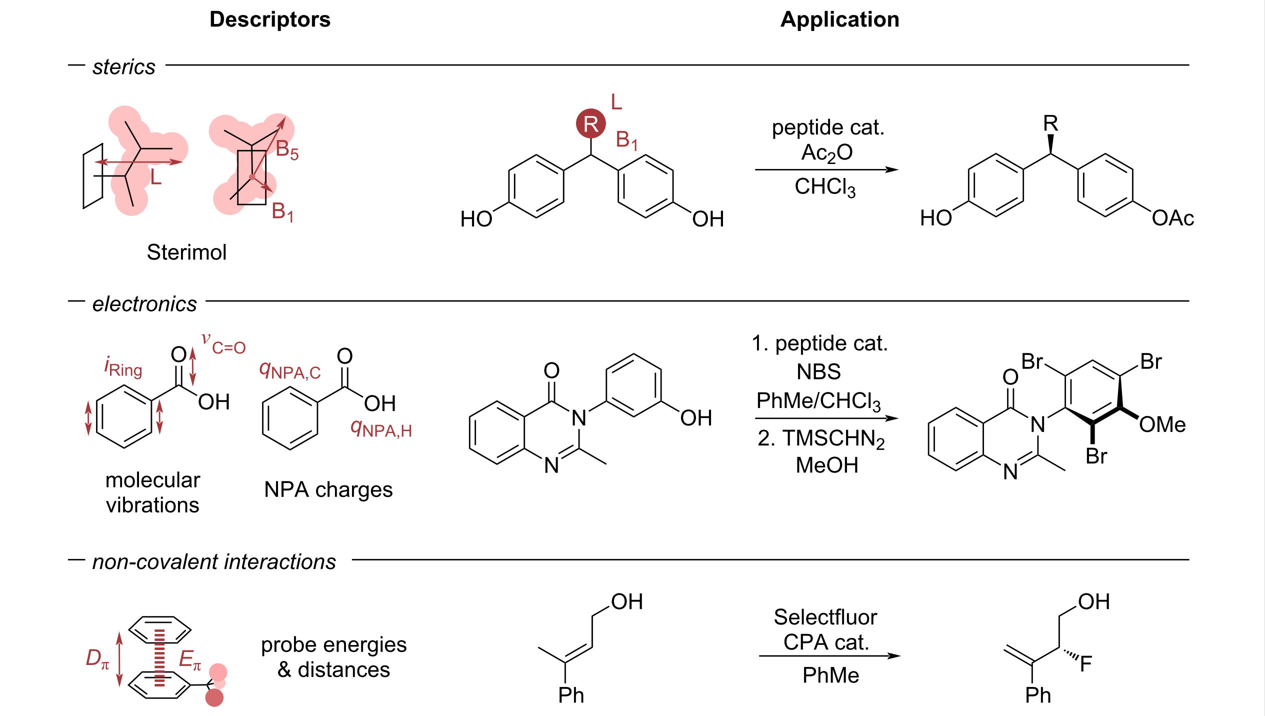


Figure 5: Selected examples of popular descriptors applied to model organocatalytic reactions. Descriptors encompass steric features modelled via Sterimol parameters [84] (example from Harper et al. [60] correlating the Sterimol B1 and L parameters of the bisphenols to the enantioselectivity of the peptide catalysed desymmetrisation), electronic features modelled via vibrations or NPA charges (example from Crawford et al. [86]) and NCIs, modelled via interaction distances and energies with a defined probe (example from Orlandi et al. [61]).

Parallel to the evolution in modelling steric effects, the representation of electronic effects has also been further developed. Milo et al. [58] introduced the intensity and frequency of manually selected molecular vibrations as descriptors (Figure 5). For the selection of relevant vibrations, a mechanistic proposal is required a priori, commonly based on a manual analysis of the probed substrates. The inclusion of electronic parameters led to a considerable improvement in predicting the enantioselectivity of a peptide-catalysed bisphenol desymmetrisation compared to their omission, showcasing the importance of capturing relevant molecular properties via descriptors. Apart from molecular vibrations, electronic influences are commonly modelled via global properties of a molecule (such as HOMO/LUMO energies) or local properties (such as natural population analysis (NPA) charges/NMR shifts), as shown in Figure 5 [69,72,88,89].

With respect to organocatalysis, NCIs are often a major factor in determining selectivities, which are hard to describe via standard molecular descriptors. Therefore, Orlandi et al. [61] introduced computed NCI distances and energies between benzene and a probe residue as descriptors for NCIs (Figure 5).

Notably, the NCI energies are inspired by previous work from Wheeler and Houk [90,91] and are defined as the computed energetic difference between the complex of the benzene ring and the probe residue and the separated species. Orlandi et al. used the NCI parameters in combination with other descriptors to model the enantioselectivities of a kinetic resolution of benzyl alcohols and an enantiodivergent fluorination of allylic alcohols, observing good correlations for both reactions. Since then, the proposed NCI descriptors have been successfully applied to multiple different reactions, such as an allenoate Claisen rearrangement [92] and a phase-transfer catalysed oxidative amination reaction [93]. In the latter, NCI descriptors were both used to simplify previously existing MLR models and also led to a hypothesis of key NCIs in the transition state. Whereas these descriptors require the selection of a suitable probe model, Chen and Pollice proposed P_{int} as a descriptor of the London dispersion potential that is universal and can be calculated without a probe system [94]. Although P_{int} has not been utilised for organocatalysis, the authors applied it to a Pd-metal-catalysed enantioselective 1,1-diarylation of benzyl acrylates [95] and found a similar performance compared to NCI probe descriptors.

Despite the success of this approach, it is important to remember that descriptors do not have to be parameters of one molecule and that intermolecular terms can be used to derive mechanistic hypotheses. Toste and co-workers [96] investigated a bromocyclization catalysed by a chiral phosphoric acid (CPA) and a DABConium brominating reagent (Figure 6). The authors calculated transition state conformer ensembles for several flexible DABConium systems and performed energy decomposition analysis to separate the interactions between catalyst, substrate and the DABConium moiety. Subsequently, a random forest model was used to predict *exo*/*endo*- and regioselectivity of the reaction. Using random forest as an interpretable machine learning model allowed to extract the important features of the model, which indicated that the dispersion interaction between the DABConium system and the CPA is governing the *exo*-selectivity.

For the application of the ML techniques discussed above, it is assumed that all studied reactions follow the same mechanism. If that is not the case, models cannot be reliably fit to the data

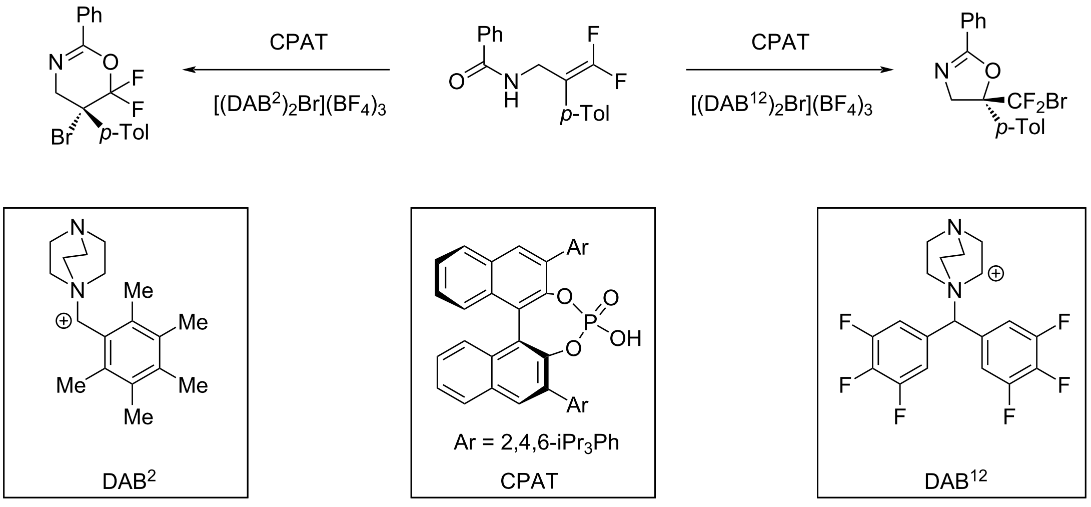


Figure 6: Example bromocyclization reaction from Toste and co-workers using a DABConium catalyst system and CPA phase transfer catalyst [96].

points, similar to mechanistic breaks in Hammett plots. However, deliberate data set design to systematically cover the relevant chemical space can aid in detecting outliers and aid in creating more relevant models, as demonstrated by Neel et al. for an enantiodivergent fluorination of allylic alcohols, catalysed by a CPA as phase transfer catalyst and an arylboronic acid [97] (Figure 7).

After a systematic data set design involving eight phosphoric acids and eight boronic acids, the authors observed breaks in linearity of the model of enantioinduction for some catalyst combinations. Further experiments, such as non-linear effect studies and isotopic substitution experiments revealed multiple different mechanisms of enantioinduction for the respective combinations. To rationalise relevant interactions, MLR models were trained on subsets of the data set. For each different mechanism of enantioinduction previously elucidated, the authors developed a separate model to gain a sufficiently interpretable model, finding that some parameters remain important throughout the different subsets. This example demonstrates both the strength of careful data analysis and the intricacies of dealing with chemical reactivity data.

The above outlined examples demonstrate the relevance of efficient representations, to which the development of advanced descriptors contributed. However, the usage of descriptors also restricts the generalizability of models, as they have to be expert derived. Interestingly, descriptor-based MLR models have also been used to predict the Mayr–Patz nucleophilicity parameter N, which estimates the nucleophilicity of a nucleophile based on experimentally measured kinetic data. The MLR models are used to predict N for more than 1200 nucleophiles, enabling the prediction of N for further nucleophiles [98-101]. While this complicates the usage of descriptors for a multitude of different reactions, it also enables an efficient representation by representing chemical hypotheses. Even though descriptors have been proposed for a number of different interactions, others are not easily represented via descriptors but remain highly important towards enantioselectivity, e.g., solvent-solute interactions.

When interpreting the importance of descriptors, effects such as overfitting and collinearity of features must be accounted for. Particularly in the low-data regime, the importance of selected features can vary based on the reactions that are contained in the training and test set. While descriptors can help in gaining mechanistic insight, it is important to not overinterpret the significance of single features to form a mechanistic hypothesis.

Ideally, to overcome issues such as a high dataset dependence, larger reaction datasets are available. In terms of data set sizes, the presented studies all worked in the low to medium data set size, with up to few hundred experiments [102,103], where careful considerations must be paid towards the applicability domain, overfitting and interpretability. With HTE platforms established and due to their importance to ML campaigns, the past few years have seen a trend in creating larger experimental chemical reactivity datasets, in particular for transition metal catalysis [47,48].

2.2 Increasing data availability in ML for organocatalysis

While, to the best of the authors' knowledge, no HTE dataset has found widespread application in ML for organocatalysis, Denmark and co-workers published a data set comprising more than 1,000 organocatalytic transformations [67]. In their work, the authors demonstrated a data-driven workflow to study the

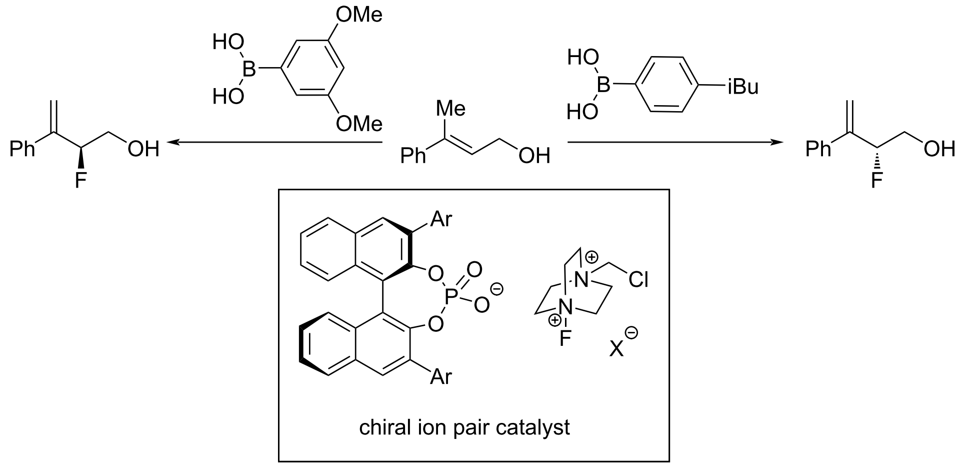


Figure 7: Example from Neel et al. using a chiral ion pair catalyst for the selective fluorination of allylic alcohols [97].

enantioselective formation of N,S-acetals catalysed by CPAs. To represent the catalysts, the authors developed the average steric occupancy (ASO) descriptors, a representation inspired by CoMFA [104-106], which recently also was applied in the selectivity prediction of aldehydes to nitroalkenes [68]. In ASO, all catalysts are aligned on a 3D-grid and the descriptor is calculated as the average occupancy of voxels on the 3D grid, where a voxel is occupied if it is within the van der Waals radius of an atom. The steric descriptors were combined with electronic descriptors called Average Electronic Indicator Field (AEIF), which are calculated for each CPA substituent (R) by observing the electrostatic potential of a quarternary ammonium ion with the substituent of interest (NMe_3R^+). The authors performed unsupervised clustering on an in silico library to select a 'Universal Training Set' (UTS) consisting of 24 catalysts, aiming to effectively represent the chemical space of CPAs. This UTS was selected by first reducing the dimension of the combined descriptor space using PCA and subsequent uniform sampling of the catalysts using a clustering algorithm (see Section 1.3), which ensures a broad coverage of CPA chemical space. Notably, this data-driven technique is not restricted to the reaction chosen by the authors. The UTS, combined with 19 'test set' catalysts, 5 nucleophiles and 5 electrophiles, constitutes a dataset of 1,075 reactions with associated enantioselectivity values (Figure 8).

The size of the data set allowed the authors to perform various ML experiments: a random (600:475) split on the data set, a substrate test set where $\Delta\Delta G^{\ddagger}$ of known catalysts with new substrate combinations were predicted, a catalyst test set where the substrates were known but the catalysts not and a test set were both components were not known beforehand. Even in the most challenging case, predictions were highly accurate with a mean absolute deviation of 0.24 kcal/mol. Further, the authors performed a split where the models were only trained on reactions with an ee < 80% (718:357 split), still showing good extrapolation performance with an error of only 0.33 kcal/mol on the test set with higher enantioselectivity.

The open availability of larger, high-quality datasets also inspires other researchers to develop and apply ML algorithms and molecular representations. The previously described dataset from Denmark and co-workers has been adopted by other groups to develop and/or benchmark descriptors [107,108], models that use architectures designed to deal with multiple conformers [109-111] (see Figure 9A and also Section 2.1) or models that are based on multiple fingerprints [112].

In addition, such larger data sets also lead to an increased interest in the application of deep learning tools, such as graph-based neural networks, to organocatalysis. One particular example was published by Hong and co-workers [113], who developed a chemistry-informed graph model for the prediction of enantioselectivities (Figure 9B). In their model, molecules were represented as graphs, where local steric and electronic information was added to each node (atom). Additionally, the used graph neural network contains a molecular interaction module that allows the model to learn synergistic effects between molecules, crucial for reactivity prediction tasks. While reaching state-of-the art performance in predicting $\Delta\Delta G^{\ddagger}$ on the data set from Denmark and co-workers, the designed neural network also enables to interpret the effects leading to the observed enantioselectivity by eliminating the atom features and observing the change in predictive performance. Using this method, the authors observed that the main contribution towards enantioinduction by CPAs is through steric effects, in line with previous literature.

Besides the establishment of experimental data sets, the number of ML data sets based on quantum mechanical calculations is also increasing, such as a data set that considers propargylation reactions catalysed by bipyridine *N*,*N*'-dioxide-derived scaffolds, created by Wheeler and co-workers using their AARON toolkit [21,114-116]. Similar to experimental data, computational data sets also lead to the development of ML innovation [117,118]. One example is the development of a new reaction representation based on the geometry of reactants and

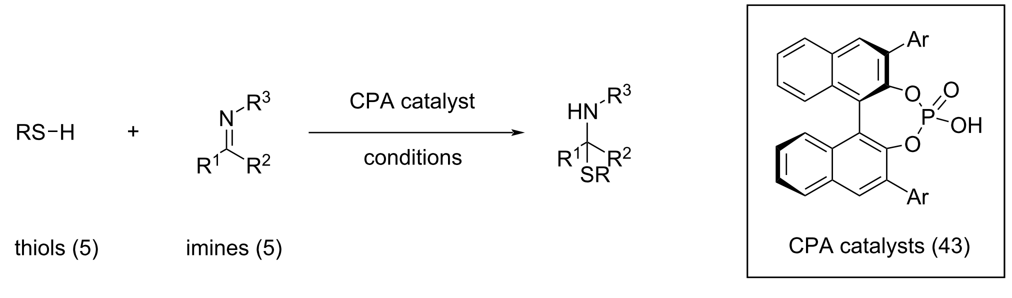


Figure 8: Data set created by Denmark and co-workers for the CPA-catalysed thiol addition to *N*-acylimines [67]. The combinatorial data set encompasses the enantioselectivities from 5 thiols and 5 imines in combination with 43 CPA catalysts for a total of 1,075 data points.

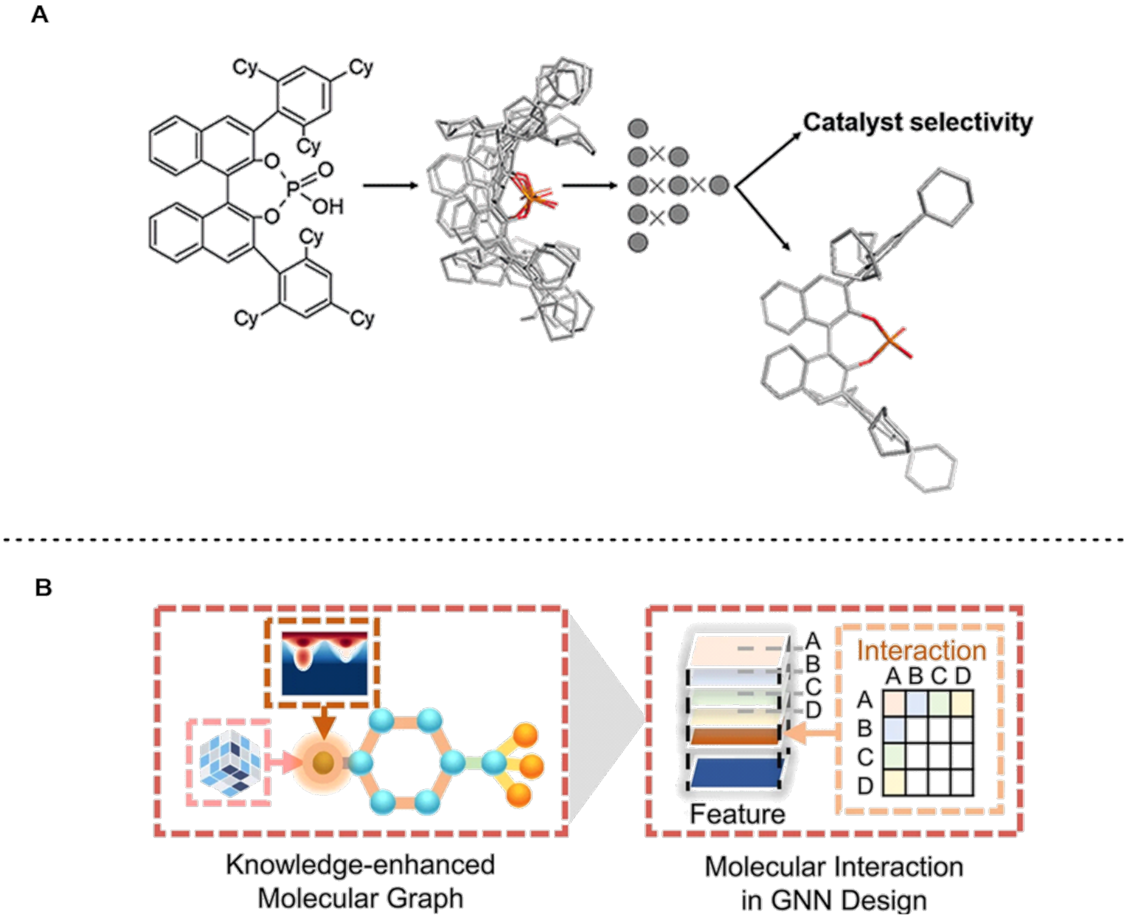


Figure 9: Selected examples of ML developments that used the dataset from Denmark and co-workers [67]. (A) Varnek and co-workers used ML models designed to deal with multiple catalyst conformers for the prediction of catalyst selectivity. Reproduced with permission from reference [109], © 2021 Georg Thieme Verlag KG. This content is not subject to CC BY 4.0. (B): Hong and co-workers utilised a molecular graph based on knowledge about the local steric and electronic information, coupled with a graph neural network equipped with a module designed to capture molecular interactions. Figure adapted from reference [113] (© 2023 S.-W. Li et al., published by Springer Nature, distributed under the terms of the Creative Commons Attribution 4.0 International License, https://creativecommons.org/licenses/by/4.0).

products [89]. Unlike expert-chosen descriptors, this representation is generalisable to other systems. Although not concerned with selectivity, Corminboeuf and co-workers reported OSCAR, a computational repository of 4,000 organocatalyst structures mined from the literature and Cambridge Structural Database (CSD) [31].

In addition, the authors utilised the combinatorial nature of organocatalysts to create data bases comprising more than 8,000 NHC-type catalysts and more than one million double hydrogen bond donor catalysts. While this repository does not provide any reactivity data, it still comprises a valuable map of organocatalyst chemical space to aid in catalyst design.

The creation of these larger datasets, both experimental and in silico, has enabled the interest of the ML in chemistry community towards enantioselective organocatalysis. With these datasets, it is now possible to test different algorithms and benchmark varying chemical representations. Despite these advances, the existence of few large datasets in enantioselective organocatalysis might lead to a bias in developed algorithms and representations. Since few datasets are available, advances are benchmarked on these datasets and commonly only published if they provide state-of-the-art performance. Thus, a bias towards representations and algorithms that capture relevant effects of the existing datasets are conceivable, while other important effects that govern selectivities remain underexplored by the community. Therefore, it is highly relevant to extend the available chemical space to underexplored regions and to acquire large datasets for such cases to allow for more holistic investigations of algorithms and chemical representations.

To summarise, the last decade has seen a steady refinement in the representation of chemical species, considering sterics, electronic properties and non-covalent interactions. Since these interactions are governing any reactivity, accurate description is relevant for a successful ML campaign. Most of the work in organocatalysis using expert-derived descriptors has been conducted in the low to middle data-regime. Only recently, the focus has shifted towards bigger data sets of more than 1,000 reactions, the first one of which has already inspired a manifold of other groups to develop new ML techniques, including graph neural networks. With the continued rise of high-throughput experimentation in organocatalysis [40], we expect ML to be applied to more data sets in this domain to aid in answering a wider variety of research questions. For the prediction of selectivities, we expect more advanced techniques to be adopted, establishing ML as a powerful tool for the evaluation of organocatalysts.

3 ML for the design of privileged organocatalysts

Throughout the development of organocatalysis, privileged catalysts, i.e., catalysts which catalyse a wide variety of different reactions through the same mechanism of enantioinduction, have emerged in multiple organocatalytic transformations [119]. The examples discussed in Section 2 all have seen the application of ML techniques to predict the selectivity of a reaction of interest. However, since the mechanism of enantioinduction is similar for multiple reactions catalysed by a privileged catalyst class, these 'related' reactions can in principle be modelled together. The reactions are assumed to be mechanistically transferable.

The similarity of multiple reactions led to two different applications of ML to organocatalysis: (**1**) prediction of reaction properties (e.g., selectivity) for multiple mechanistically transferable reactions, and (**2**) employing ML in the search to predict the generality of a catalyst. This chapter will discuss prominent examples in both applications.

3.1 ML for transferable reactions

The key to modelling transferable reactions together is to find a representation that can describe all relevant reacting species. While such representations commonly exist in chemistry, e.g., SMILES and graphs, the most common representation for transferable reactions is via expert-chosen descriptors. As such, the space of relevant reactions has to be carefully studied, e.g., with respect to the different reactant or catalyst classes. Once this space is defined, the descriptors have to be chosen such that they are specific enough to provide information to the ML model while also general enough to cover the space of interest.

One pioneering study in the field of mechanistic transferability for enantioselectivity prediction was published by Reid and Sigman [120] in 2019. The authors manually combined 367 different published reactions of BINOL-phosphoric acid catalysed nucleophilic additions to imines, comprising alcohols, thiols, phosphonates, diazoacetamides, peroxides, benzothiazolines and more as nucleophiles. Apart from reactant classes, the reactions also vary in additives, and solvent among others. Since these reactions all adhere to the same mechanism of enantioinduction, the authors chose to consider them in the same ML campaign, even though the nucleophiles vary significantly. As descriptors, the authors used the overlapping features of nucleophiles, imines and catalysts to derive steric and electronic parameters as well as topological descriptors for solvents, where less structural overlap is present [121].

For every reaction, the imine is categorised as either an *E*- or *Z*-imine, based on the sign of the recorded enantiomeric excess. Further, molecular descriptors, either physicochemical properties or topological, are calculated for all reaction partners. This data is used to develop a comprehensive model, finding that imine parameters govern the defining transition state and hence the preferred enantiomer. In a focused modelling, two separate models are constructed, one for all *E*- and *Z*-imines, respectively, finding substrate–catalyst matching is important for *E*- and *Z*-imines. The focused correlations enabled the authors to identify subtle mechanistic differences between reactions of *E*- and *Z*-imines, such as the role of steric and electronic properties of the imine for *E*- and *Z*-imines, respectively. The two-stage workflow, using the comprehensive model to distinguish the imine-type and subsequently using the focused model for detailed predictions, proved successful for out-of-sample reaction predictions with new nucleophiles, such as enecarbamates. Further, the authors also tested their models on the dataset published by Denmark and co-workers [67] (see Figure 10), showcasing the importance of high-quality datasets for ML applications.

Due to their prominence in organocatalysis, CPAs have been a common catalyst class when considering mechanistically transferable reactions for modelling. Further work on CPA catalysed reactions was performed by Shoja et al. [122], considering a multitude of different reaction types, ranging from hydrogenations to epoxidations and dearomatization reactions. In a further study, the generalisation of the obtained model to reactions involving more complex substrates was demonstrated [123]. For the comparison of different reaction descriptors, Asahara and Miyao [108] considered different CPA-catalysed nucleophilic additions to imines, comprising aza-Mannich reactions and Friedel–Crafts reactions among others. Different reactions were also combined by Liles et al. [124]. For a transfer hydrogena-

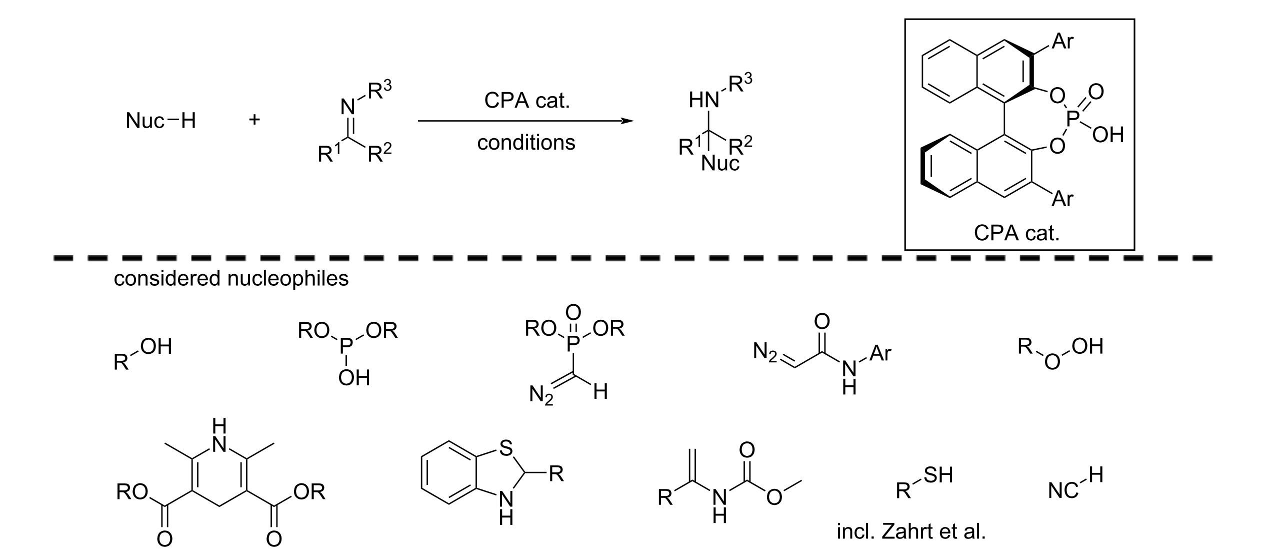


Figure 10: Study from Reid and Sigman developing statistical models for CPA-catalysed nucleophilic addition reactions to imines for different classes of nucleophiles [120].

tion reaction, the authors used a workflow consisting of training set design, classification, MLR and extrapolation to predict a new class of CPA catalysts with enhanced enantioselectivity. Subsequently, the new catalyst class was tested for cyclodehydration and oxetane desymmetrisation reactions, where a comprehensive model was developed for the three different reactions (Figure 11A).

Mechanistic model transferability for CPA-catalysed Minisci reactions [125] was utilised for the derivatization of quinolines and pyridines. Models trained on these compound classes show good generalisation towards other nitrogen-containing heteroaromatics including pyrimidines and pyrazines.

The importance of mechanistic understanding for model building was underlined by Kuang et al. [126], where the authors considered multi-catalyst enantioselective reactions, where one catalyst was an organocatalyst, either CPA or an amine. The co-catalyst was included in the ML model by being considered as a nucleophile or electrophile, depending on the reaction mechanism. Descriptors allowed for the inclusion of a variety of co-catalysts, ranging from Fe-piano stool complexes to copper complexes. The consideration of co-catalysis into model development further expands the considerable reaction space in organocatalysis.

The discussed principle of mechanistic transferability has also been employed outside of CPA catalysis, with a focus on amine-based hydrogen-bond donors, for example imidodiphosphorimidate-type catalysts for the construction of THF and THP rings [107] (Figure 11B). Werth and Sigman [127] investigated multiple nucleophilic additions to nitroalkenes, catalysed by bifunctional hydrogen bond donors, observing good correlations to new bi-functional donors, new nucleophiles, new electrophiles and even similar cascade-type reactions.

In the authors' perspective, the exploitation of the concept of mechanistic transferability is a promising avenue for the application of ML in enantioselective organocatalysis, as combining data from multiple reactions enlarges datasets. As such, it is an important stepping stone towards the development of more generally applicable models. However, when applying these models, potential mechanistic breaks as well as utility of the chosen representations (descriptors) across the entire dataset have to be considered. Currently, the work mainly focuses on CPAs for which a vast number of reactions are reported. While this underlines the importance of CPAs as enantioselective organocatalysts, work exploring the mechanistic transferability of other catalyst classes should not be neglected in order to fulfill the potential that the application of ML in organocatalysis holds.

3.2 ML for general organocatalysts

While it is important to consider catalysts achieving high enantiomeric excess (ee) on relevant reactions, the deployment of general catalysts that provide a reasonable ee for a variety of reactions has gained more attention over the last years [128-130]. Catalysts that fulfil such demands are coined 'general catalysts'.

While the concept of generality was recently explored in a closed-loop fashion for Suzuki–Miyaura cross couplings to find

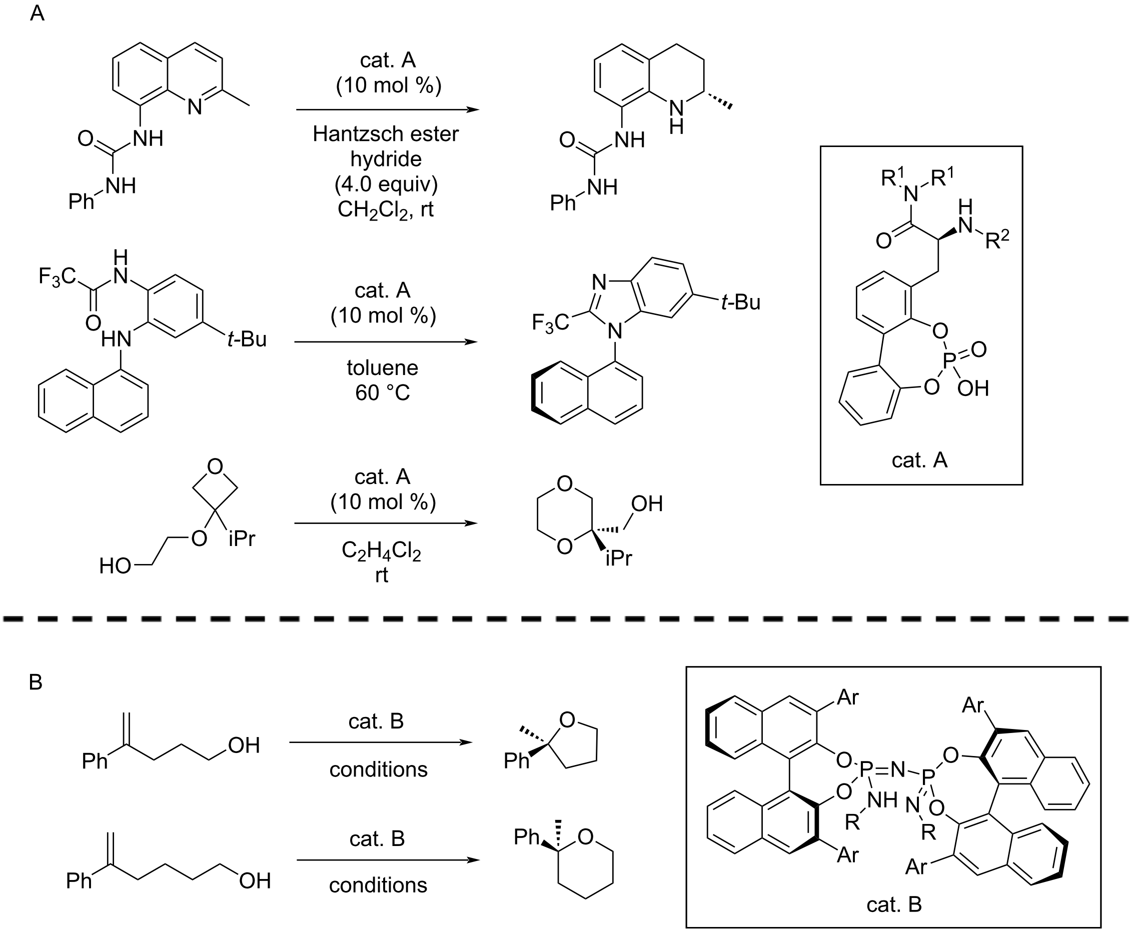


Figure 11: Selected examples of studies where mechanistic transferability was exploited to model multiple reactions together. (A): Liles et al. used univariate classification and MLR to develop a new CPA catalyst achieving high enantioselectivity for transfer hydrogenations. A comprehensive model for multiple reactions was developed under the assumption of mechanistic transferability [124]. (B): List and co-workers employed Support Vector Machines trained on data of different cyclisations to find an optimal catalyst for tetrahydropyran synthesis [107]. R = general residue, Ar = aromatic residue.

the most general catalyst and reaction conditions [131], the application of this concept in the context of ML has found comparatively less attention in organocatalysis, despite the prominence of privileged catalysts.

Despite the intuitive explanation of generality to chemists, a clear mathematical definition of chemical generality remains elusive, exacerbating the integration of the generality concept towards machine learning algorithms. As such, different implementations were chosen to tackle this problem.

In 2022, Denmark and co-workers [132] (Figure 12) investigated a disulfonimide-catalysed atroposelective iodination with the intention of finding a general reaction procedure. After constructing an in silico library consisting of 1,478 catalysts, a universal training set was constructed consisting of 18 catalysts. Subsequently, the enantioselectivity of each catalyst with 13 model substrates was experimentally evaluated. 13 different models, one for each substrate, were developed. To find a general catalyst, a technique termed ’catalyst selection by committee’ (CSC) was employed: for each substrate, all in silico catalysts were evaluated and catalysts in the most enantioselective 1% of catalysts considered received one ’vote’. After this process was performed for each of the 13 model substrates, catalysts with more votes were termed as being more general, balancing high enantioselectivity with a broader substrate scope. CSC enabled the identification of two well-performing, general catalysts.

A different generality metric was proposed by Betinol et al. [133] (Figure 13). The authors performed clustering on the reaction space of interest representing the molecule either by topological or quantum mechanical descriptors. The generality of a catalyst was then assigned by considering the fraction of clus-

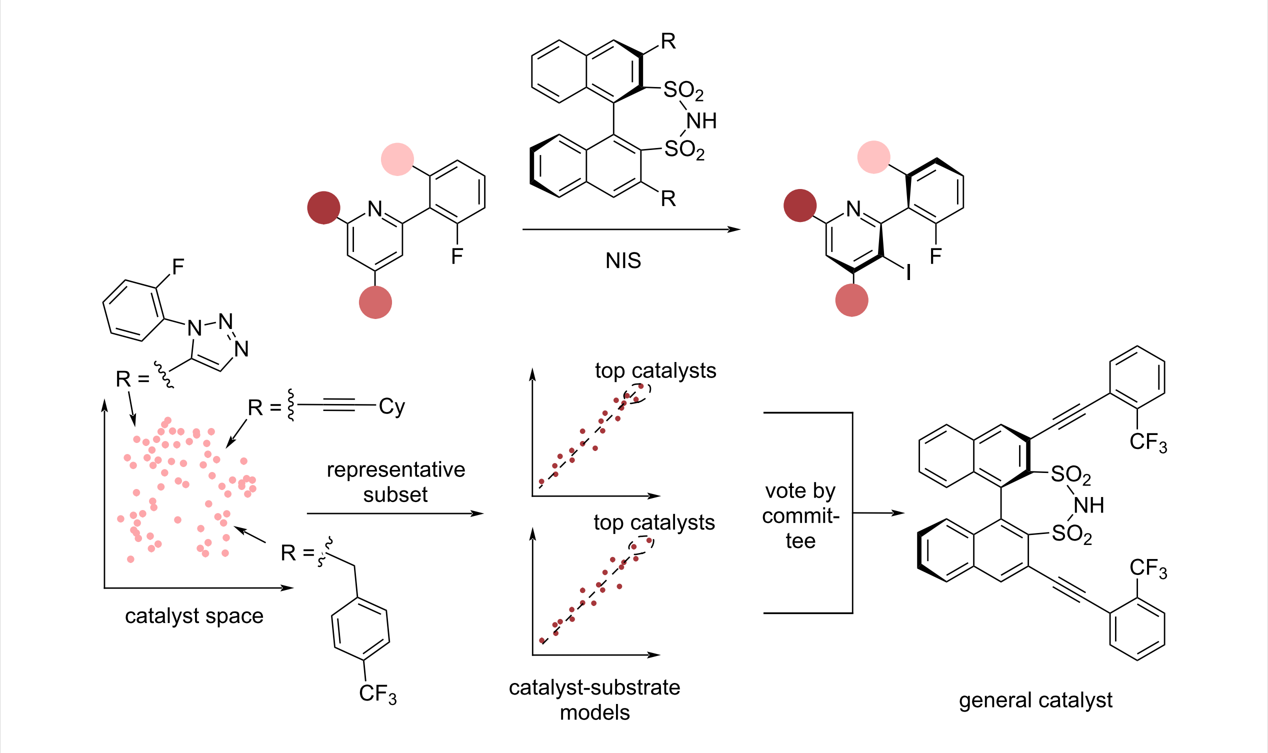


Figure 12: Generality approach by Denmark and co-workers [132] for the iodination of arylpyridines. From the relevant chemical space, a representative subset of 18 catalysts is selected. For each of the 13 model substrates, a catalyst-substrate model is trained. Catalysts that are top performers for multiple substrates are considered general catalysts.

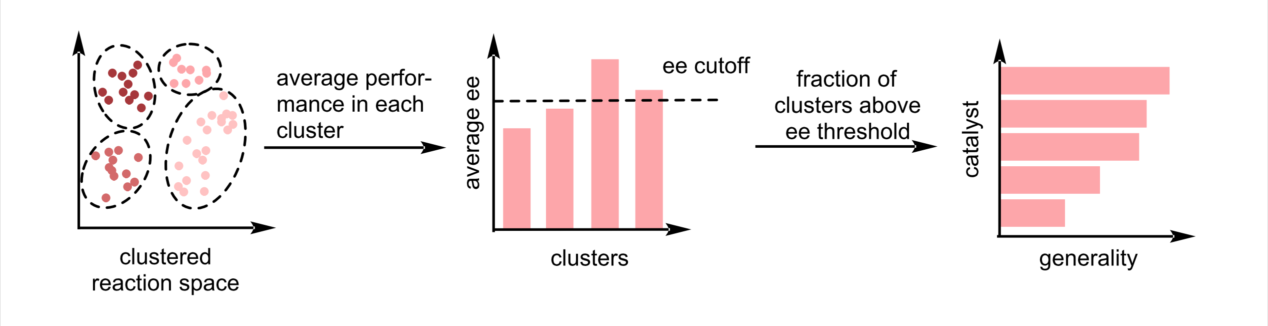


Figure 13: Betinol et al. [133] clustered the relevant chemical space and then evaluated the average ee for every cluster. Catalysts that perform above a user-defined ee cutoff for many clusters are considered general catalysts.

ters for which the average cluster enantioselectivity of a catalyst exceeds a user-defined threshold. This threshold can be used to balance the need for a wide substrate scope and enantioselectivity requirements, while accounting for the specifics of a reaction and the requirements of the user. The authors applied their method on 3,003 literature-mined Mannich reactions from 106 publications to find that urea-based catalysts are the most general organocatalysts for this reaction class (ee threshold 80%), even though amine-based catalysts demonstrate a higher average ee. Notably, this strategy is not restricted to literature-extracted examples and can also be applied to enantioselectivities calculated via quantum chemical calculations or predictions from an ML model. The latter was used by the authors as an augmentation technique towards an imbalanced dataset for CPA-catalysed nucleophilic additions. Further, the authors also propose an order of generality for CPAs catalysing nucleophilic additions to imines, with TRIP being the highest ranked (ee threshold 60%). Thus, the authors recommend that for developing a new reaction, their metric can be used to decide which catalyst should be tested first based on the expected success. This generality-based guiding principle of experimental design showcases a further possibility for data-driven methods to complement and augment experimental chemistry.

In addition to these methods, Corminboeuf and co-workers [134] proposed a genetic algorithm for the de novo design of general catalysts (Figure 14). Considering the Pictet–Spengler cyclization of tryptamine derivatives catalysed by hydrogen-bond donors, the authors considered a general catalyst to display both high enantioselectivity and turn-over frequency for a broad substrate scope. The substrate scope, termed generality probing set (GPS), was selected based on farthest point sampling of a literature mined reaction space to cover a wide chemical space. To assess the enantioselectivity and turn-over frequency for reactions with a new catalyst, which is required for de novo design, the authors used different strategies. To predict enantioselectivity of a previously unseen reaction, the authors used the reported enantioselectivities in their initial literature-mined reaction database to train an XGBoost model. The turn-over frequency of a reaction was determined using a volcano plot based on reaction energies [135-137], where the latter were again predicted using an XGBoost model based on the literature-mined dataset. Using fragments derived from their OSCAR [31] database, the authors used the NaviCatGA genetic algorithm [118] to find the most general catalysts. The fitness function comprised multiple objectives, including the median of the enantioselectivity and activity across the generality probing set. The usage of a multi-objective optimization algorithm allowed them to discover multiple trade-off optima, enabling a scientist to select the ideal catalyst based on the specific requirements of catalytic activity and selectivity, while still accounting for catalyst generality through design of the objectives. Noticeably, data analysis allowed to identify regions in the chemical space where highly ranked catalysts underperform as well as less sensitive areas in chemical space, further providing mechanistic insight into the mechanism of stereoinduction and activity trends.

With the concept of privileged catalysts deeply rooted in organocatalysis, we expect a steady increase in studies aiming to bridge the gaps between different reactions that are mechanistically transferable via ML. Using this strategy, it is possible to both increase the available data (since more reactions are considered), as well as investigate more general mechanisms. However, careful consideration has to be paid towards combining different reactions, as mechanistic transferability has to be ensured. Furthermore, the usage of ML to identify general catalysts demonstrate that the application of modern ML tools is not limited to predicting selective catalysts.

4 ML for catalyst and reaction design

The design of chemical reactions encompasses various aspects, from the choice of the employed catalyst to the selection of ideal reaction conditions. While traditionally, all of this has been performed by chemical knowledge, intuition and rational design, the last years have witnessed a surge in data-driven approaches to improve the design of reactions, e.g., by inferring mechanistic features through statistical modelling, the generation of catalyst structures with increased catalytic activity, or optimising the reaction conditions to maximise the yield or selectivity. In contrast to the direct approach as seen in many examples discussed so far, where starting from a molecular structure and a set of conditions, the reaction outcome is predicted, optimising the design of a reaction can be framed as an inverse design approach [138]. Given a target, e.g., fast conversion or high selectivity, the task is to find a catalyst structure or a set of conditions to satisfy the requirement. The following chapter will give an overview of recent advances in the design of organocatalytic reactions.

4.1 Mechanistic understanding

The design of a catalyst requires detailed understanding of the key catalytic steps [23,139-142] and commonly uses calculated or measured physical parameters of reaction components to make decisions in a design effort. In line with the early developments of statistical modelling through Hammett parameters to correlate substrate properties to kinetic properties of the reaction (Section 2), advanced ML tools can help to unravel key mechanistic features in higher dimensions and with stronger interactions, which can be used to tailor a reaction to match

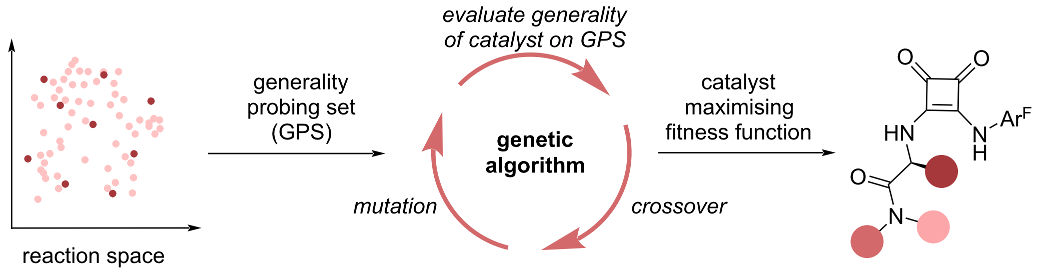

reaction space
generality probing set (GPS)
evaluate generality of catalyst on GPS
genetic algorithm
mutation
crossover
catalyst maximising fitness function
O
O
HN
N
H
Ar^F
O
N

Figure 14: Corminboeuf and co-workers [134] chose a representative subset of the reaction space (indicated by dark red points) and used it to evaluate the generality of catalyst structures optimised through their genetic algorithm.

desired properties. Sigman and co-workers demonstrated this by complementing knowledge from physical organic chemistry with data-driven analysis techniques, in particular MLR, to gain a greater understanding of the enantioselectivity-determining steps for a C–N coupling catalysed by CPA derivatives (Figure 15A) [143]. Based on their findings that π–π interactions between the catalyst's triazole substituent and the substrate is key for stereoinduction, they designed new catalyst structures maximising the predicted selectivity. The predictions were experimentally validated confirming that their model can be used to guide the design of highly selective catalysts (Figure 15B).

4.2 High-throughput virtual screening

Although such approaches showcase the ability of ML models to unravel structure–activity relationships and thereby guide the development of catalysts, the design of new structures remains influenced by the prevailing design principles of chemists. In this regard, approaches to explore uncharted regions of chemical space in a more unbiased way can help to identify previously unknown structures that exhibit desired properties. The advent of statistical models that can predict key catalytic properties has enabled pipelines to assess a great number of candidates in high-throughput virtual screening approaches [107,144-147]. Thereby, experimental efforts can be focused on the most promising candidates predicted by the model. Denmark and co-workers utilised such an approach to design highly selective catalysts for a peptide-catalysed annulation reaction [68]. Using conformer-dependent steric and electronic descriptors, they built a universal training set (UTS) consisting of 161 tripeptide catalysts. Based on models trained on the UTS they were able to identify highly selective tripeptide catalysts from a virtual library containing more than 30,000 structures. Remarkably, the predicted peptide catalysts did not follow the prevailing design principles of experimentally optimised peptide catalysts, demonstrating how ML can help to explore novel classes of catalysts. While high-throughput screening campaigns can be powerful tools for the discovery of novel structures with desired properties, their scope can be limited due to the effort associated with computing the descriptors for each individual molecule. Corminboeuf and co-workers utilised a fragment-based approach exploiting the modularity of commonly used organocatalysts. By considering individual contributions of catalyst fragments, they were able to build a combinatorial library of cata-

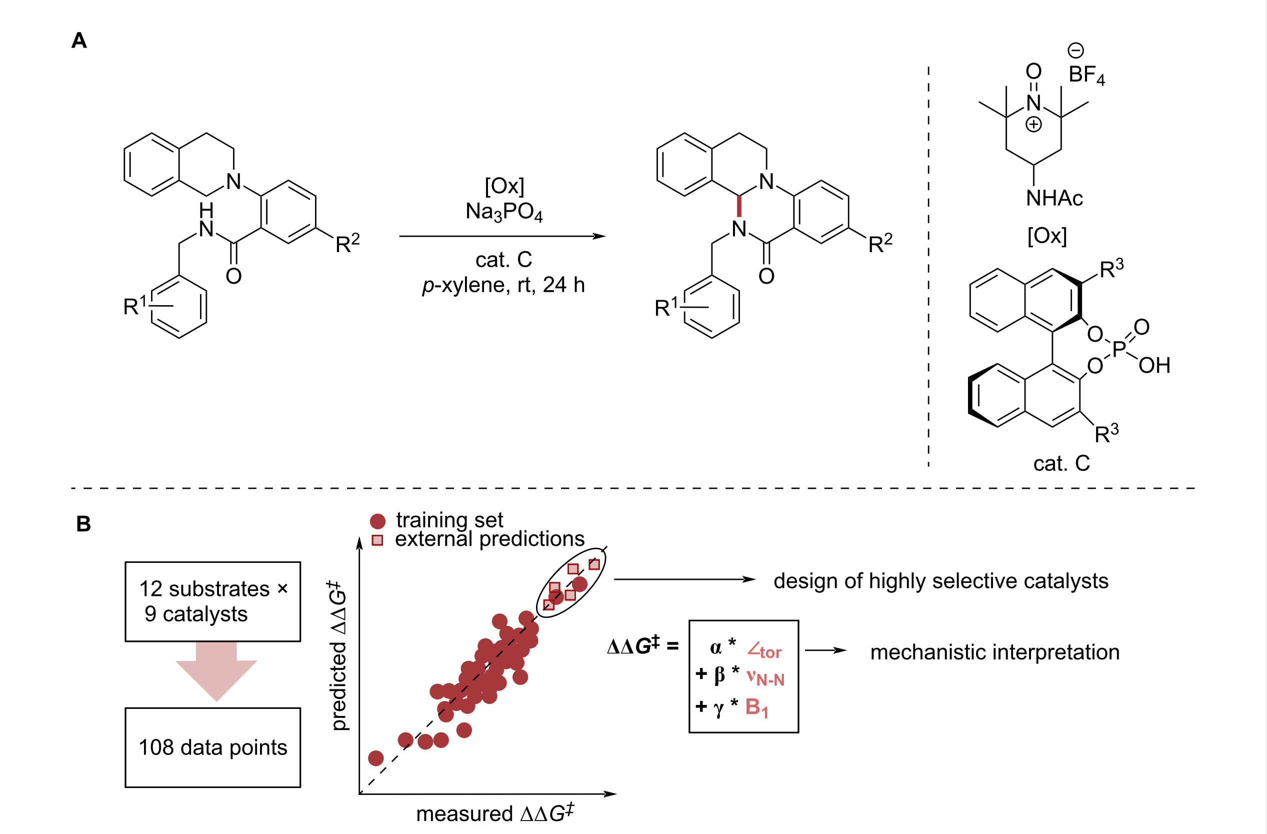


Figure 15: Example for data-driven modelling to improve substrate and catalyst design. (A) C–N coupling catalysed by CPA derivatives studied by Sigman and co-workers [143]. (B) The library for the study consisted of 12 substrates and 9 triazolyl catalysts. This data was used to train an MLR model and infer key mechanistic features as well as the design of highly selective catalysts which were experimentally verified after prediction.

lysts and predicted novel catalysts with increased reactivity for an organocatalysed Diels–Alder reaction [148].

4.3 Genetic algorithms

An alternative approach for chemical space exploration is the use of genetic algorithms (GAs) [149]. Inspired by biological evolution, they aim to maximise a fitness function using biology-inspired operations such as mutation and crossovers. Jensen and co-workers demonstrated the utility of GAs by optimising the structure of a tertiary amine catalyst for the Morita–Baylis–Hillman reaction [150] (Figure 16).

First, the rate determining step was identified (within the proposed reaction mechanism). Then, the organocatalyst's structure was optimised to decrease the barrier of this step. After identification of the most potent structures by the GA, they verified experimentally that the identified structure increases the reaction rate by a factor of 7.8 compared to the commonly used DABCO catalyst. While this clearly demonstrates the capabilities of the GA to accelerate the discovery of organocatalysts, the authors note that the success of their approach is dependent on the detailed knowledge of the underlying mechanism. Therefore, the discovery of catalysts for novel reaction mechanism is still an ongoing challenge [151-153]. In order to make GAs for catalyst discovery more generally available, the Corminboeuf group developed the software suit 'NaviCatGA' [118] which is designed for the optimisation of catalysts with desired catalytic properties. The tool provides the user with considerable flexibility, e.g., the definition of the employed fitness function or the genetic operations to be applied. Further, it supports the multi-objective optimisation based on multiple target properties, which is of particular importance as an ideal catalyst combines a number of properties that need to be taken into account, e.g., solubility, stability and synthesisability. The authors exemplify this by optimising simultaneously for catalytic activity and selectivity using two individual MLR expressions in their fitness function. Doing so, their algorithm is able to tailor the structure of the employed base for a Lewis-base catalysed enantioselective propargylation of benzaldehyde in this multi-objective optimisation task [118].

Importantly, molecules designed by generative models need to be tested experimentally. This allows one to verify the assumptions made during modelling and validate the model’s ability to propose molecules tailored to a given application. In this regard, the synthesisability of the generated molecules plays a decisive role and remains a major bottleneck which currently restricts the effective use of generative models [154].

4.4 ML-driven experimental design

Besides the design of employed catalysts, reaction design involves the identification of optimal reaction conditions, which poses a formidable challenge due to the high dimensionality of the reaction space. In the simplest approach, ideal reaction conditions are identified by changing one parameter at a time based on the chemist’s intuition. While this shows the influ-

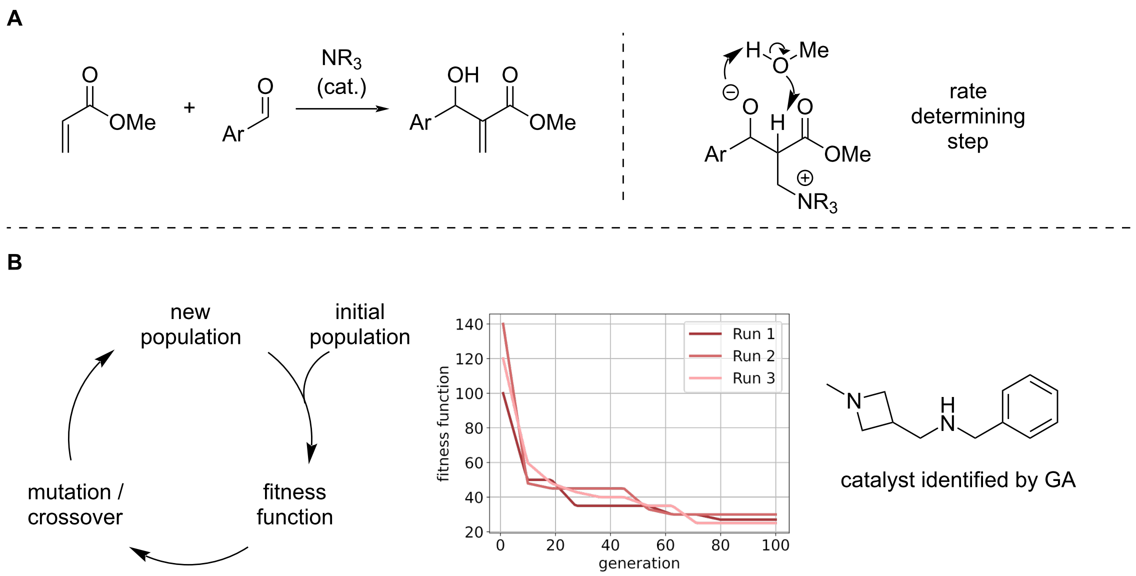


Figure 16: Example for utilising a genetic algorithm for catalyst design. (A) Morita–Baylis–Hillman reaction studied by Jensen and co-workers [150] (B) Left: A genetic algorithm performs mutation and crossover operations on a set of catalysts with the goal to optimise the fitness function. Middle: Schematic depiction how the fitness is iteratively optimised across multiple runs ("generations") of the genetic cycle. Right: Identified catalyst structure with increased catalytic activity.

ence of the varied parameter on the observable, interaction effects between the parameters are significantly harder to capture with this approach. Design of experiments (DoE) is a more systematic approach where parameters are varied simultaneously to unravel their effect on the outcome [155,156]. Although multiple variants of DoE are available, the number of required experiments can quickly exceed what is feasible for most applications. Driven by optimisation problems in other fields, like ML model parameters, more efficient optimisation strategies have therefore been explored recently. Particularly Bayesian optimisation is widely used for optimisation problems where the quantity of interest is expensive to obtain, such as quantifying the yield of a reaction. Therefore, it has found application for the optimisation of chemical problems [157-167] and demonstrated its effectiveness by outperforming human optimisation strategies [168]. However, even with efficient optimisation algorithms, conducting experiments and analysing the reaction outcome remains a major bottleneck. Performing chemistry in flow provides several advantages in this regard, as reaction parameters can be varied on-the-fly [169]. In combination with ML optimisation strategies, this can lead to efficient optimisation of reaction conditions as demonstrated by Kondo et al. where they utilised Gaussian Process Regression (GPR) [170] to optimise the flow rate, the temperature as well as the stoichiometry of the reactant for the organocatalysed synthesis of spirooxindole analogues [171] (Figure 17).

In a later study the same group expanded the search space for a Brønsted acid-catalysed cross-coupling for the synthesis of biaryl compounds [172]. They utilised Bayesian optimisation to explore a total of six numerical and categorical parameters. With as little as 15 data points they were able to find optimal conditions which yielded the desired product in 96% yield. This showcases the application of ML-driven optimisation strategies for efficient multi-parameter screening problems, however, manual action is still required for experimental setup and analysis. Automating these operations would significantly increase productivity and reproducibility and is a research area of high interest termed self-driving laboratories [173,174]. Cooper and co-workers exemplified the opportunities of a self-driving laboratory by utilising a free-roaming robot that autonomously conducted and analysed 688 experiments selected by a Bayesian optimisation algorithm [175]. Within eight days it discovered a set of parameters that yielded a six-fold increase of activity for the photocatalytic hydrogen evolution from water compared to the baseline formulation. These examples show the possibilities that ML offers for optimising experimental design in organocatalysis. However, the use of data-driven methods to optimise reactions is still far from routine. It is expected that the recent surge of Large Language Models (LLMs) will support this development and further improve accessibility and the interaction between humans and ML-based models [176-178]. While the works presented give a glimpse of what is possible with automated experimentation pipelines in combination with ML, the wide adoption of such methods is limited by the high acquisition costs of the setup, the expertise and time required to implement and maintain the hardware in the research environment and the limited versatility of the methods to a broad range of problems [179].

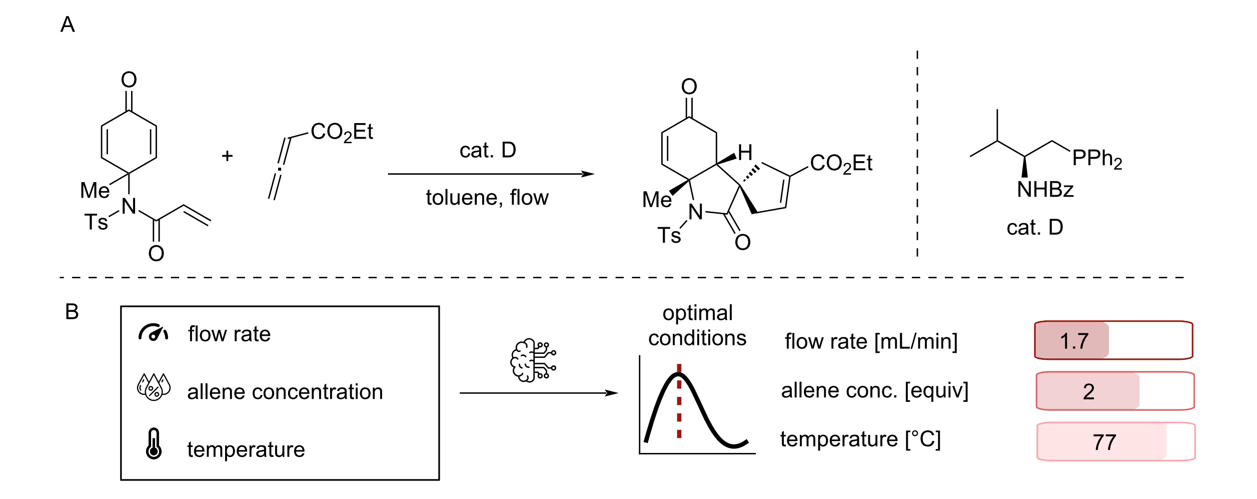


Figure 17: Organocatalysed synthesis of spirooxindole analogues by Kondo et al. [171] (A) Reaction scheme of dienones with allenoates to form chiral spirooxindole analogues using a chiral amine organocatalyst. (B) Schematic depiction of the employed optimisation to identify optimal conditions for flow reaction using Gaussian Process Regression. Icon 'Flow rate' made by Gregor Cresnar from flaticon.com. This content is not subject to CC BY 4.0. Icon 'Allene concentration' made by Nadiinko from flaticon.com. This content is not subject to CC BY 4.0. Icon 'Temperature' made by hirschwolf from flaticon.com. This content is not subject to CC BY 4.0. Icon 'ML model' made by VectorPortal from flaticon.com. This content is not subject to CC BY 4.0.

Conclusion

The tremendous potential of utilising ML tools to support organocatalysis is clearly demonstrated in the above presented works. Nevertheless, it remains to be seen whether these examples provide general solutions and are applicable to a wide range of problems. In this regard, the domain of applicability needs to be carefully analysed in order to obtain reliable and robust predictions [180,181]. While some works exemplified the ability of data-driven models to provide interpretable results, their validity is far from being universally applicable. It should be remembered that correlations in statistical models don’t equal causation, and that hypotheses made from feature importances need to be followed up by mechanistic studies to avoid potentially misleading conclusions.

One common bottleneck for further improvements and the wider application of statistical tools is the generation and availability of high-quality data [182] (Figure 18). As the bottlenecks are prevalent throughout the sub-disciplines of homogeneous catalysis, we expect that developments for the application of ML in one area will have a strong influence across the whole domain.

The utilisation of electronic lab notebooks [183-185] and the adoption of standardised formats for collecting and sharing data such as the Open Reaction Database (ORD) scheme could significantly improve the broadness of available data sets [42,43,186-188]. Moreover, standardised protocols for performing experiments, for example for probing the robustness or the sensitivity of a reaction [189-191], as well as the selection of the substrate scope can help to provide valuable information in a reproducible fashion [192,193].

Further, this also requires a paradigm shift towards keeping track of and publishing all conducted experiments, regardless of whether the expected outcome was achieved or not. While HTE campaigns typically yield a broader distribution of reaction outcomes [67], unsuccessful reactivity from traditional "benchtop" chemistry is only rarely reported. Nevertheless, authors are beginning to include a selection of "unsuccessful

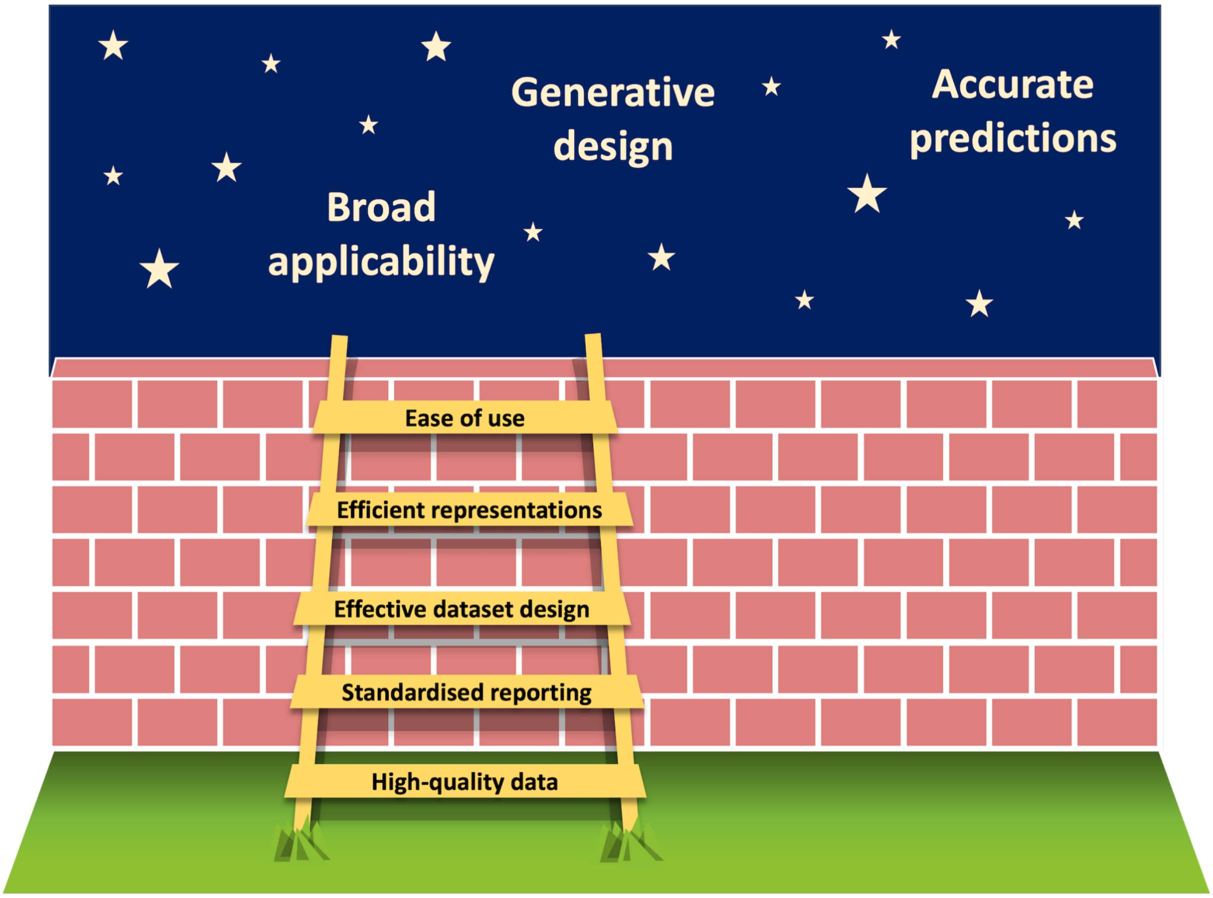


Figure 18: Schematic depiction of required developments in order to overcome current limitations of ML for organocatalysis.

substrates" in the supporting information [194-198]. In this context, it is necessary to highlight the importance of publishing data in accordance with the FAIR (Findable, Accessible, Interoperable, Reusable) principles to allow for wide usage by the community. Importantly, this does not only apply to experimental work, but also all results from data-driven modelling.

In terms of data set design, Bayesian optimisation bears the potential to maximise the information gained by ML algorithms without the need for extensive experimental effort. In combination with closed-loop high-throughput experimentation, this would allow for fast access to data that cover the problem space adequately and thereby enable optimal modelling. Current challenges for automation pipelines include the purification and analysis of the reaction outcome [199], which is particularly challenging in asymmetric organocatalysis. Due to its relevance for industrial processes however, we expect an increased interest in HTE platforms specifically tailored to organocatalysis, especially (organo-)photocatalysis [200]. In this context, flow chemistry could provide a promising platform to enable closed-loop, multi-objective optimisations and facile scale-up of reactions [201]. With ML tools becoming increasingly accessible for non-experts through easy-to-use interfaces [202,203], their application is expected to gain greater popularity and be integrated into existing routines [204]. This could involve ML-guided catalyst screening, obtaining entries for the substrate scope through unsupervised learning or ML-based reaction condition optimisation. This development will be supported through the advent of LLMs and their incorporation into chemical workflows [176,178] which increase the accessibility of ML tools for synthetic chemistry. While a low entry barrier does not make the knowledge of statistics and coding (primarily in Python) redundant, the abundance of online tutorials and courses on ML allows also non-experts to acquire fundamental skills and to apply such techniques to their own problems. As statistical and coding competencies are becoming more relevant to scientists, courses focused on these fundamentals are being continuously integrated in chemistry curricula at universities.

The last decade has shown the pace at which data-driven tools can be utilised in organocatalysis and led to powerful tools that can augment synthetic chemists. Most works have focused on enantioselectivity as the quantity of interest. Recently, many works have also applied ML for investigating privileged organocatalytic systems. However, there are other objectives that are worth considering when developing a reaction, for example sustainability, complexity, or cost aspects. In this regard future work might involve multi-objective optimisation schemes and generative modelling to account for the plethora of requirements in reaction and process development. Moreover, recent trends in organocatalysis, such as photocatalysis, halogen-bonding, or cooperative catalysis [205], provide new synthetic opportunities, whose advancements are expected to be supported through data-driven modelling.

Acknowledgements

The authors thank Lauriane Jacot-Descombes for assistance in design of Figure 18. Further, we acknowledge different artists for icons used in the Graphical Abstract: Icon ‘Database’ made by The Chohans Brand from flaticon.com. This content is not subject to CC BY 4.0. Icon ‘Molecule’ made by Freepik from flaticon.com. This content is not subject to CC BY 4.0. Icon ‘Brain’ made by Freepik from flaticon.com. This content is not subject to CC BY 4.0.

Funding

This study was created as part of NCCR Catalysis (grant number 180544), a National Centre of Competence in Research, funded by the Swiss National Science Foundation. The authors thank the Deutsche Forschungsgemeinschaft (SPP2363 – Utilisation and Development of Machine Learning for Molecular Applications – Molecular Machine Learning, L.S.).

ORCID® iDs

Stefan P. Schmid - https://orcid.org/0000-0002-0965-0208
Leon Schlosser - https://orcid.org/0009-0007-6764-6497
Frank Glorius - https://orcid.org/0000-0002-0648-956X
Kjell Jorner - https://orcid.org/0000-0002-4191-6790

Data Availability Statement

Data sharing is not applicable as no new data was generated or analyzed in this study.

Preprint

A non-peer-reviewed version of this article has been previously published as a preprint: https://doi.org/10.26434/chemrxiv-2024-xfdn8

References

1. Benaglia, M., Ed. *Organocatalysis. Stereoselective Reactions and Applications in Organic Synthesis;* De Gruyter: Berlin, Germany, 2021. doi:10.1515/9783110590050
2. Bell, E. L.; Finnigan, W.; France, S. P.; Green, A. P.; Hayes, M. A.; Hepworth, L. J.; Lovelock, S. L.; Niikura, H.; Osuna, S.; Romero, E.; Ryan, K. S.; Turner, N. J.; Flitsch, S. L. *Nat. Rev. Methods Primers* **2021,** *1,* 46. doi:10.1038/s43586-021-00044-z
3. Twilton, J.; Le, C.; Zhang, P.; Shaw, M. H.; Evans, R. W.; MacMillan, D. W. C. *Nat. Rev. Chem.* **2017,** *1,* 52. doi:10.1038/s41570-017-0052
4. Kerru, N.; Katari, N. K.; Jonnalagadda, S. B. *Phys. Sci. Rev.* **2022,** *7,* 325–344. doi:10.1515/psr-2021-0022
5. Xiang, S.-H.; Tan, B. *Nat. Commun.* **2020,** *11,* 3786. doi:10.1038/s41467-020-17580-z

6. Bernardi, L.; Carlone, A.; Fini, F. Industrial Relevance of Asymmetric Organocatalysis in the Preparation of Chiral Amine Derivatives. In *Methodologies in Amine Synthesis;* Ricci, A.; Bernardi, L., Eds.; Wiley-VCH: Weinheim, Germany, 2021; pp 187–241. doi:10.1002/9783527826186.ch6
7. Bulger, P. G. Industrial Applications of Organocatalysis. In *Comprehensive Chirality;* Carreira, E. M.; Yamamoto, H., Eds.; Elsevier: Amsterdam, Netherlands, 2012; pp 228–252. doi:10.1016/b978-0-08-095167-6.00911-3
8. Han, B.; He, X.-H.; Liu, Y.-Q.; He, G.; Peng, C.; Li, J.-L. *Chem. Soc. Rev.* **2021,** *50,* 1522–1586. doi:10.1039/d0cs00196a
9. Hughes, D. L. *Org. Process Res. Dev.* **2018,** *22,* 574–584. doi:10.1021/acs.oprd.8b00096
10. Keith, J. A.; Vassilev-Galindo, V.; Cheng, B.; Chmiela, S.; Gastegger, M.; Müller, K.-R.; Tkatchenko, A. *Chem. Rev.* **2021,** *121,* 9816–9872. doi:10.1021/acs.chemrev.1c00107
11. Toyao, T.; Maeno, Z.; Takakusagi, S.; Kamachi, T.; Takigawa, I.; Shimizu, K.-i. *ACS Catal.* **2020,** *10,* 2260–2297. doi:10.1021/acscatal.9b04186
12. Kitchin, J. R. *Nat. Catal.* **2018,** *1,* 230–232. doi:10.1038/s41929-018-0056-y
13. Li, Z.; Wang, S.; Xin, H. *Nat. Catal.* **2018,** *1,* 641–642. doi:10.1038/s41929-018-0150-1
14. Yang, W.; Fidelis, T. T.; Sun, W.-H. *ACS Omega* **2020,** *5,* 83–88. doi:10.1021/acsomega.9b03673
15. Esterhuizen, J. A.; Goldsmith, B. R.; Linic, S. *Nat. Catal.* **2022,** *5,* 175–184. doi:10.1038/s41929-022-00744-z
16. Gomollón-Bel, F. *Chem. Int.* **2019,** *41* (2), 12–17. doi:10.1515/ci-2019-0203
17. Houk, K. N.; Cheong, P. H.-Y. *Nature* **2008,** *455,* 309–313. doi:10.1038/nature07368
18. Sterling, A. J.; Zavitsanou, S.; Ford, J.; Duarte, F. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2021,** *11,* e1518. doi:10.1002/wcms.1518
19. Corbeil, C. R.; Thielges, S.; Schwartzentruber, J. A.; Moitessier, N. *Angew. Chem., Int. Ed.* **2008,** *47,* 2635–2638. doi:10.1002/anie.200704774
20. Weill, N.; Corbeil, C. R.; De Schutter, J. W.; Moitessier, N. *J. Comput. Chem.* **2011,** *32,* 2878–2889. doi:10.1002/jcc.21869
21. Guan, Y.; Ingman, V. M.; Rooks, B. J.; Wheeler, S. E. *J. Chem. Theory Comput.* **2018,** *14,* 5249–5261. doi:10.1021/acs.jctc.8b00578
22. Rosales, A. R.; Wahlers, J.; Limé, E.; Meadows, R. E.; Leslie, K. W.; Savin, R.; Bell, F.; Hansen, E.; Helquist, P.; Munday, R. H.; Wiest, O.; Norrby, P.-O. *Nat. Catal.* **2019,** *2,* 41–45. doi:10.1038/s41929-018-0193-3
23. Iribarren, I.; Garcia, M. R.; Trujillo, C. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2022,** *12,* e1616. doi:10.1002/wcms.1616
24. Melnyk, N.; Iribarren, I.; Mates-Torres, E.; Trujillo, C. *Chem. – Eur. J.* **2022,** *28,* e202201570. doi:10.1002/chem.202201570
25. Melnyk, N.; Garcia, M. R.; Iribarren, I.; Trujillo, C. *Tetrahedron Chem* **2023,** *5,* 100035. doi:10.1016/j.tchem.2023.100035
26. Hammett, L. P. *J. Am. Chem. Soc.* **1937,** *59,* 96–103. doi:10.1021/ja01280a022
27. Williams, W. L.; Zeng, L.; Gensch, T.; Sigman, M. S.; Doyle, A. G.; Anslyn, E. V. *ACS Cent. Sci.* **2021,** *7,* 1622–1637. doi:10.1021/acscentsci.1c00535
28. Hansch, C.; Leo, A.; Taft, R. W. *Chem. Rev.* **1991,** *91,* 165–195. doi:10.1021/cr00002a004
29. Suvarna, M.; Pérez-Ramírez, J. *Nat. Catal.* **2024,** *7,* 624–635. doi:10.1038/s41929-024-01150-3
30. Strieth-Kalthoff, F.; Sandfort, F.; Kühnemund, M.; Schäfer, F. R.; Kuchen, H.; Glorius, F. *Angew. Chem., Int. Ed.* **2022,** *61,* e202204647. doi:10.1002/anie.202204647
31. Gallarati, S.; van Gerwen, P.; Laplaza, R.; Vela, S.; Fabrizio, A.; Corminboeuf, C. *Chem. Sci.* **2022,** *13,* 13782–13794. doi:10.1039/d2sc04251g
32. Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. *Sci. Data* **2014,** *1,* 140022. doi:10.1038/sdata.2014.22
33. Strieth-Kalthoff, F.; Sandfort, F.; Segler, M. H. S.; Glorius, F. *Chem. Soc. Rev.* **2020,** *49,* 6154–6168. doi:10.1039/c9cs00786e
34. Haghighatlari, M.; Li, J.; Heidar-Zadeh, F.; Liu, Y.; Guan, X.; Head-Gordon, T. *Chem* **2020,** *6,* 1527–1542. doi:10.1016/j.chempr.2020.05.014
35. Wappett, D. A.; Goerigk, L. *J. Chem. Theory Comput.* **2023,** *19,* 8365–8383. doi:10.1021/acs.jctc.3c00558
36. Taylor, M. G.; Yang, T.; Lin, S.; Nandy, A.; Janet, J. P.; Duan, C.; Kulik, H. J. *J. Phys. Chem. A* **2020,** *124,* 3286–3299. doi:10.1021/acs.jpca.0c01458
37. Swain, M. C.; Cole, J. M. *J. Chem. Inf. Model.* **2016,** *56,* 1894–1904. doi:10.1021/acs.jcim.6b00207
38. Vaucher, A. C.; Zipoli, F.; Geluykens, J.; Nair, V. H.; Schwaller, P.; Laino, T. *Nat. Commun.* **2020,** *11,* 3601. doi:10.1038/s41467-020-17266-6
39. Zheng, Z.; Zhang, O.; Borgs, C.; Chayes, J. T.; Yaghi, O. M. *J. Am. Chem. Soc.* **2023,** *145,* 18048–18062. doi:10.1021/jacs.3c05819
40. Fan, V.; Qian, Y.; Wang, A.; Wang, A.; Coley, C. W.; Barzilay, R. *J. Chem. Inf. Model.* **2024,** *64,* 5521–5534. doi:10.1021/acs.jcim.4c00572
41. Ai, Q.; Meng, F.; Shi, J.; Pelkie, B.; Coley, C. W. *Digital Discovery* **2024,** in press. doi:10.1039/d4dd00091a
42. Kearnes, S. M.; Maser, M. R.; Wleklinski, M.; Kast, A.; Doyle, A. G.; Dreher, S. D.; Hawkins, J. M.; Jensen, K. F.; Coley, C. W. *J. Am. Chem. Soc.* **2021,** *143,* 18820–18826. doi:10.1021/jacs.1c09820
43. Nippa, D. F.; Müller, A. T.; Atz, K.; Konrad, D. B.; Grether, U.; Martin, R. E.; Schneider, G. *ChemRxiv* **2023**. doi:10.26434/chemrxiv-2023-nfq7h
44. Nie, W.; Wan, Q.; Sun, J.; Chen, M.; Gao, M.; Chen, S. *Nat. Commun.* **2023,** *14,* 6671. doi:10.1038/s41467-023-42446-5
45. Krska, S. W.; DiRocco, D. A.; Dreher, S. D.; Shevlin, M. *Acc. Chem. Res.* **2017,** *50,* 2976–2985. doi:10.1021/acs.accounts.7b00428
46. Buitrago Santanilla, A.; Regalado, E. L.; Pereira, T.; Shevlin, M.; Bateman, K.; Campeau, L.-C.; Schneeweis, J.; Berritt, S.; Shi, Z.-C.; Nantermet, P.; Liu, Y.; Helmy, R.; Welch, C. J.; Vachal, P.; Davies, I. W.; Cernak, T.; Dreher, S. D. *Science* **2015,** *347,* 49–53. doi:10.1126/science.1259203
47. Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. *Science* **2018,** *360,* 186–190. doi:10.1126/science.aar5169
48. Perera, D.; Tucker, J. W.; Brahmbhatt, S.; Helal, C. J.; Chong, A.; Farrell, W.; Richardson, P.; Sach, N. W. *Science* **2018,** *359,* 429–434. doi:10.1126/science.aap9112
49. Heid, E.; Probst, D.; Green, W. H.; Madsen, G. K. H. *Chem. Sci.* **2023,** *14,* 14229–14242. doi:10.1039/d3sc02048g

50. Morgat, A.; Axelsen, K. B.; Lombardot, T.; Alcántara, R.; Aimo, L.; Zerara, M.; Niknejad, A.; Belda, E.; Hyka-Nouspikel, N.; Coudert, E.; Redaschi, N.; Bougueleret, L.; Steinbeck, C.; Xenarios, I.; Bridge, A. *Nucleic Acids Res.* **2015,** *43,* D459–D464. doi:10.1093/nar/gku961
51. Jeske, L.; Placzek, S.; Schomburg, I.; Chang, A.; Schomburg, D. *Nucleic Acids Res.* **2019,** *47,* D542–D549. doi:10.1093/nar/gky1048
52. Shalit Peleg, H.; Milo, A. *Angew. Chem., Int. Ed.* **2023,** *62,* e202219070. doi:10.1002/anie.202219070
53. Davies, I. W. *Nature* **2019,** *570,* 175–181. doi:10.1038/s41586-019-1288-y
54. Weininger, D. *J. Chem. Inf. Comput. Sci.* **1988,** *28,* 31–36. doi:10.1021/ci00057a005
55. Rogers, D.; Hahn, M. *J. Chem. Inf. Model.* **2010,** *50,* 742–754. doi:10.1021/ci100050t
56. Wigh, D. S.; Goodman, J. M.; Lapkin, A. A. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2022,** *12,* e1603. doi:10.1002/wcms.1603
57. David, L.; Thakkar, A.; Mercado, R.; Engkvist, O. *J. Cheminf.* **2020,** *12,* 56. doi:10.1186/s13321-020-00460-5
58. Milo, A.; Bess, E. N.; Sigman, M. S. *Nature* **2014,** *507,* 210–214. doi:10.1038/nature13019
59. Gallegos, L. C.; Luchini, G.; St. John, P. C.; Kim, S.; Paton, R. S. *Acc. Chem. Res.* **2021,** *54,* 827–836. doi:10.1021/acs.accounts.0c00745
60. Harper, K. C.; Bess, E. N.; Sigman, M. S. *Nat. Chem.* **2012,** *4,* 366–374. doi:10.1038/nchem.1297
61. Orlandi, M.; Coelho, J. A. S.; Hilton, M. J.; Toste, F. D.; Sigman, M. S. *J. Am. Chem. Soc.* **2017,** *139,* 6803–6806. doi:10.1021/jacs.7b02311
62. Hickey, D. P.; Schiedler, D. A.; Matanovic, I.; Doan, P. V.; Atanassov, P.; Minteer, S. D.; Sigman, M. S. *J. Am. Chem. Soc.* **2015,** *137,* 16179–16186. doi:10.1021/jacs.5b11252
63. Dhayalan, V.; Gadekar, S. C.; Alassad, Z.; Milo, A. *Nat. Chem.* **2019,** *11,* 543–551. doi:10.1038/s41557-019-0258-1
64. Gow, S.; Niranjan, M.; Kanza, S.; Frey, J. G. *Digital Discovery* **2022,** *1,* 551–567. doi:10.1039/d2dd00047d
65. McInnes, L.; Healy, J.; Melville, J. *arXiv* **2018,** 1802.03426. doi:10.48550/arxiv.1802.03426
66. van der Maaten, L.; Hinton, G. E. *J. Mach. Learn. Res.* **2008,** *9,* 2579–2605.
67. Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science* **2019,** *363,* eaau5631. doi:10.1126/science.aau5631
68. Schnitzer, T.; Schnurr, M.; Zahrt, A. F.; Sakhaee, N.; Denmark, S. E.; Wennemers, H. *ACS Cent. Sci.* **2024,** *10,* 367–373. doi:10.1021/acscentsci.3c01284
69. Santiago, C. B.; Guo, J.-Y.; Sigman, M. S. *Chem. Sci.* **2018,** *9,* 2398–2412. doi:10.1039/c7sc04679k
70. Noto, N.; Yada, A.; Yanai, T.; Saito, S. *Angew. Chem., Int. Ed.* **2023,** *62,* e202219107. doi:10.1002/anie.202219107
71. Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. *ACS Cent. Sci.* **2017,** *3,* 434–443. doi:10.1021/acscentsci.7b00064
72. Banerjee, S.; Sreenithya, A.; Sunoj, R. B. *Phys. Chem. Chem. Phys.* **2018,** *20,* 18311–18318. doi:10.1039/c8cp03141j
73. Dormann, C. F.; Elith, J.; Bacher, S.; Buchmann, C.; Carl, G.; Carré, G.; Marquéz, J. R. G.; Gruber, B.; Lafourcade, B.; Leitão, P. J.; Münkemüller, T.; McClean, C.; Osborne, P. E.; Reineking, B.; Schröder, B.; Skidmore, A. K.; Zurell, D.; Lautenbach, S. *Ecography* **2013,** *36,* 27–46. doi:10.1111/j.1600-0587.2012.07348.x
74. Harrell, F. E., Jr. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis;* Springer: New York, NY, USA, 2001. doi:10.1007/978-1-4757-3462-1
75. Murray, K.; Conner, M. M. *Ecology* **2009,** *90,* 348–355. doi:10.1890/07-1929.1
76. Rogers, C. J.; Dickerson, T. J.; Brogan, A. P.; Janda, K. D. *J. Org. Chem.* **2005,** *70,* 3705–3708. doi:10.1021/jo050161r
77. Taft, R. W., Jr. *J. Am. Chem. Soc.* **1952,** *74,* 2729–2732. doi:10.1021/ja01131a010
78. Taft, R. W., Jr. *J. Am. Chem. Soc.* **1952,** *74,* 3120–3128. doi:10.1021/ja01132a049
79. Taft, R. W., Jr. *J. Am. Chem. Soc.* **1953,** *75,* 4538–4539. doi:10.1021/ja01114a044
80. Taft, R. W.; Topsom, R. D. The Nature and Analysis of Substitutent Electronic Effects. *Progress in Physical Organic Chemistry;* John Wiley & Sons: New York, NY, USA, 1987; Vol. 16, pp 1–83. doi:10.1002/9780470171950.ch1
81. Charton, M. *J. Am. Chem. Soc.* **1975,** *97,* 1552–1556. doi:10.1021/ja00839a047
82. Crawford, J. M.; Kingston, C.; Toste, F. D.; Sigman, M. S. *Acc. Chem. Res.* **2021,** *54,* 3136–3148. doi:10.1021/acs.accounts.1c00285
83. Oslob, J. D.; Åkermark, B.; Helquist, P.; Norrby, P.-O. *Organometallics* **1997,** *16,* 3015–3021. doi:10.1021/om9700371
84. Verloop, A.; Hoogenstraaten, W.; Tipker, J. Development and Application of New Steric Substituent Parameters in Drug Design. In *Medicinal Chemistry: A Series of Monographs;* Ariëns, E. J., Ed.; Academic Press: Amsterdam, Netherlands, 1976; pp 165–207. doi:10.1016/b978-0-12-060307-7.50010-9
85. Brethomé, A. V.; Fletcher, S. P.; Paton, R. S. *ACS Catal.* **2019,** *9,* 2313–2323. doi:10.1021/acscatal.8b04043
86. Crawford, J. M.; Stone, E. A.; Metrano, A. J.; Miller, S. J.; Sigman, M. S. *J. Am. Chem. Soc.* **2018,** *140,* 868–871. doi:10.1021/jacs.7b11303
87. Newman-Stonebraker, S. H.; Smith, S. R.; Borowski, J. E.; Peters, E.; Gensch, T.; Johnson, H. C.; Sigman, M. S.; Doyle, A. G. *Science* **2021,** *374,* 301–308. doi:10.1126/science.abj4213
88. Durand, D. J.; Fey, N. *Chem. Rev.* **2019,** *119,* 6561–6594. doi:10.1021/acs.chemrev.8b00588
89. Gallarati, S.; Fabregat, R.; Laplaza, R.; Bhattacharjee, S.; Wodrich, M. D.; Corminboeuf, C. *Chem. Sci.* **2021,** *12,* 6879–6889. doi:10.1039/d1sc00482d
90. Wheeler, S. E.; Houk, K. N. *J. Am. Chem. Soc.* **2008,** *130,* 10854–10855. doi:10.1021/ja802849j
91. Wheeler, S. E. *Acc. Chem. Res.* **2013,** *46,* 1029–1038. doi:10.1021/ar300109n
92. Miró, J.; Gensch, T.; Ellwart, M.; Han, S.-J.; Lin, H.-H.; Sigman, M. S.; Toste, F. D. *J. Am. Chem. Soc.* **2020,** *142,* 6390–6399. doi:10.1021/jacs.0c01637
93. Orlandi, M.; Toste, F. D.; Sigman, M. S. *Angew. Chem., Int. Ed.* **2017,** *56,* 14080–14084. doi:10.1002/anie.201707644
94. Pollice, R.; Chen, P. *Angew. Chem., Int. Ed.* **2019,** *58,* 9758–9769. doi:10.1002/anie.201905439
95. Orlandi, M.; Hilton, M. J.; Yamamoto, E.; Toste, F. D.; Sigman, M. S. *J. Am. Chem. Soc.* **2017,** *139,* 12688–12695. doi:10.1021/jacs.7b06917
96. Miller, E.; Mai, B. K.; Read, J. A.; Bell, W. C.; Derrick, J. S.; Liu, P.; Toste, F. D. *ACS Catal.* **2022,** *12,* 12369–12385. doi:10.1021/acscatal.2c03077

97. Neel, A. J.; Milo, A.; Sigman, M. S.; Toste, F. D. *J. Am. Chem. Soc.* **2016,** *138,* 3863–3875. doi:10.1021/jacs.6b00356
98. Mayr, H.; Kempf, B.; Ofial, A. R. *Acc. Chem. Res.* **2003,** *36,* 66–77. doi:10.1021/ar020094c
99. Mayr, H.; Ofial, A. R. *J. Phys. Org. Chem.* **2008,** *21,* 584–595. doi:10.1002/poc.1325
100. Mayr, H.; Patz, M. *Angew. Chem., Int. Ed. Engl.* **1994,** *33,* 938–957. doi:10.1002/anie.199409381
101. Orlandi, M.; Escudero-Casao, M.; Licini, G. *J. Org. Chem.* **2021,** *86,* 3555–3564. doi:10.1021/acs.joc.0c02952
102. Jorner, K. *Chimia* **2023,** *77,* 22–30. doi:10.2533/chimia.2023.22
103. Heid, E.; McGill, C. J.; Vermeire, F. H.; Green, W. H. *J. Chem. Inf. Model.* **2023,** *63,* 4012–4029. doi:10.1021/acs.jcim.3c00373
104. Yamaguchi, S. *Org. Biomol. Chem.* **2022,** *20,* 6057–6071. doi:10.1039/d2ob00228k
105. Lipkowitz, K. B.; Pradhan, M. *J. Org. Chem.* **2003,** *68,* 4648–4656. doi:10.1021/jo0267697
106. Melville, J. L.; Andrews, B. I.; Lygo, B.; Hirst, J. D. *Chem. Commun.* **2004,** 1410–1411. doi:10.1039/b402378a
107. Tsuji, N.; Sidorov, P.; Zhu, C.; Nagata, Y.; Gimadiev, T.; Varnek, A.; List, B. *Angew. Chem., Int. Ed.* **2023,** *62,* e202218659. doi:10.1002/anie.202218659
108. Asahara, R.; Miyao, T. *ACS Omega* **2022,** *7,* 26952–26964. doi:10.1021/acsomega.2c03812
109. Zankov, D.; Polishchuk, P.; Madzhidov, T.; Varnek, A. *Synlett* **2021,** *32,* 1833–1836. doi:10.1055/a-1553-0427
110. Zankov, D.; Madzhidov, T.; Polishchuk, P.; Sidorov, P.; Varnek, A. *J. Chem. Inf. Model.* **2023,** *63,* 6629–6641. doi:10.1021/acs.jcim.3c00393
111. Zankov, D.; Madzhidov, T.; Varnek, A.; Polishchuk, P. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2024,** *14,* e1698. doi:10.1002/wcms.1698
112. Sandfort, F.; Strieth-Kalthoff, F.; Kühnemund, M.; Beecks, C.; Glorius, F. *Chem* **2020,** *6,* 1379–1390. doi:10.1016/j.chempr.2020.02.017
113. Li, S.-W.; Xu, L.-C.; Zhang, C.; Zhang, S.-Q.; Hong, X. *Nat. Commun.* **2023,** *14,* 3569. doi:10.1038/s41467-023-39283-x
114. Lu, T.; Zhu, R.; An, Y.; Wheeler, S. E. *J. Am. Chem. Soc.* **2012,** *134,* 3095–3102. doi:10.1021/ja209241n
115. Sepúlveda, D.; Lu, T.; Wheeler, S. E. *Org. Biomol. Chem.* **2014,** *12,* 8346–8353. doi:10.1039/c4ob01719f
116. Doney, A. C.; Rooks, B. J.; Lu, T.; Wheeler, S. E. *ACS Catal.* **2016,** *6,* 7948–7955. doi:10.1021/acscatal.6b02366
117. van Gerwen, P.; Fabrizio, A.; Wodrich, M. D.; Corminboeuf, C. *Mach. Learn.: Sci. Technol.* **2022,** *3,* 045005. doi:10.1088/2632-2153/ac8f1a
118. Laplaza, R.; Gallarati, S.; Corminboeuf, C. *Chem.: Methods* **2022,** *2,* e202100107. doi:10.1002/cmtd.202100107
119. Yoon, T. P.; Jacobsen, E. N. *Science* **2003,** *299,* 1691–1693. doi:10.1126/science.1083622
120. Reid, J. P.; Sigman, M. S. *Nature* **2019,** *571,* 343–348. doi:10.1038/s41586-019-1384-z
121. Roy, K.; Das, R. N. *Curr. Drug Metab.* **2014,** *15,* 346–379. doi:10.2174/1389200215666140908102230
122. Shoja, A.; Zhai, J.; Reid, J. P. *ACS Catal.* **2021,** *11,* 11897–11905. doi:10.1021/acscatal.1c03520
123. Betinol, I. O.; Kuang, Y.; Reid, J. P. *Org. Lett.* **2022,** *24,* 1429–1433. doi:10.1021/acs.orglett.1c04134
124. Liles, J. P.; Rouget-Virbel, C.; Wahlman, J. L. H.; Rahimoff, R.; Crawford, J. M.; Medlin, A.; O'Connor, V. S.; Li, J.; Roytman, V. A.; Toste, F. D.; Sigman, M. S. *Chem* **2023,** *9,* 1518–1537. doi:10.1016/j.chempr.2023.02.020
125. Reid, J. P.; Proctor, R. S. J.; Sigman, M. S.; Phipps, R. J. *J. Am. Chem. Soc.* **2019,** *141,* 19178–19185. doi:10.1021/jacs.9b11658
126. Kuang, Y.; Lai, J.; Reid, J. P. *Chem. Sci.* **2023,** *14,* 1885–1895. doi:10.1039/d2sc05974f
127. Werth, J.; Sigman, M. S. *J. Am. Chem. Soc.* **2020,** *142,* 16382–16391. doi:10.1021/jacs.0c06905
128. Wagen, C. C.; McMinn, S. E.; Kwan, E. E.; Jacobsen, E. N. *Nature* **2022,** *610,* 680–686. doi:10.1038/s41586-022-05263-2
129. Strassfeld, D. A.; Algera, R. F.; Wickens, Z. K.; Jacobsen, E. N. *J. Am. Chem. Soc.* **2021,** *143,* 9585–9594. doi:10.1021/jacs.1c03992
130. Wang, J. Y.; Stevens, J. M.; Kariofillis, S. K.; Tom, M.-J.; Golden, D. L.; Li, J.; Tabora, J. E.; Parasram, M.; Shields, B. J.; Primer, D. N.; Hao, B.; Del Valle, D.; DiSomma, S.; Furman, A.; Zipp, G. G.; Melnikov, S.; Paulson, J.; Doyle, A. G. *Nature* **2024,** *626,* 1025–1033. doi:10.1038/s41586-024-07021-y
131. Angello, N. H.; Rathore, V.; Beker, W.; Wołos, A.; Jira, E. R.; Roszak, R.; Wu, T. C.; Schroeder, C. M.; Aspuru-Guzik, A.; Grzybowski, B. A.; Burke, M. D. *Science* **2022,** *378,* 399–405. doi:10.1126/science.adc8743
132. Rose, B. T.; Timmerman, J. C.; Bawel, S. A.; Chin, S.; Zhang, H.; Denmark, S. E. *J. Am. Chem. Soc.* **2022,** *144,* 22950–22964. doi:10.1021/jacs.2c08820
133. Betinol, I. O.; Lai, J.; Thakur, S.; Reid, J. P. *J. Am. Chem. Soc.* **2023,** *145,* 12870–12883. doi:10.1021/jacs.3c03989
134. Gallarati, S.; van Gerwen, P.; Laplaza, R.; Brey, L.; Makaveev, A.; Corminboeuf, C. *Chem. Sci.* **2024,** *15,* 3640–3660. doi:10.1039/d3sc06208b
135. Nørskov, J. K.; Bligaard, T.; Hvolbæk, B.; Abild-Pedersen, F.; Chorkendorff, I.; Christensen, C. H. *Chem. Soc. Rev.* **2008,** *37,* 2163–2171. doi:10.1039/b800260f
136. Kulkarni, A.; Siahrostami, S.; Patel, A.; Nørskov, J. K. *Chem. Rev.* **2018,** *118,* 2302–2312. doi:10.1021/acs.chemrev.7b00488
137. Wodrich, M. D.; Sawatlon, B.; Busch, M.; Corminboeuf, C. *Acc. Chem. Res.* **2021,** *54,* 1107–1117. doi:10.1021/acs.accounts.0c00857
138. Sanchez-Lengeling, B.; Aspuru-Guzik, A. *Science* **2018,** *361,* 360–365. doi:10.1126/science.aat2663
139. Liu, S.-J.; Chen, Z.-H.; Chen, J.-Y.; Ni, S.-F.; Zhang, Y.-C.; Shi, F. *Angew. Chem., Int. Ed.* **2022,** *61,* e202112226. doi:10.1002/anie.202112226
140. Gerosa, G. G.; Marcarino, M. O.; Spanevello, R. A.; Suárez, A. G.; Sarotti, A. M. *J. Org. Chem.* **2020,** *85,* 9969–9978. doi:10.1021/acs.joc.0c01256
141. Handoko; Satishkumar, S.; Panigrahi, N. R.; Arora, P. S. *J. Am. Chem. Soc.* **2019,** *141,* 15977–15985. doi:10.1021/jacs.9b07742
142. Iribarren, I.; Trujillo, C. *Phys. Chem. Chem. Phys.* **2020,** *22,* 21015–21021. doi:10.1039/d0cp02012e
143. Milo, A.; Neel, A. J.; Toste, F. D.; Sigman, M. S. *Science* **2015,** *347,* 737–743. doi:10.1126/science.1261043
144. Wan, Y.; Ramirez, F.; Zhang, X.; Nguyen, T.-Q.; Bazan, G. C.; Lu, G. *npj Comput. Mater.* **2021,** *7,* 69. doi:10.1038/s41524-021-00541-5
145. Nandy, A.; Duan, C.; Goffinet, C.; Kulik, H. J. *JACS Au* **2022,** *2,* 1200–1213. doi:10.1021/jacsau.2c00176

146. Bai, Y.; Wilbraham, L.; Slater, B. J.; Zwijnenburg, M. A.; Sprick, R. S.; Cooper, A. I. *J. Am. Chem. Soc.* **2019,** *141,* 9063–9071. doi:10.1021/jacs.9b03591
147. Meyer, B.; Sawatlon, B.; Heinen, S.; von Lilienfeld, O. A.; Corminboeuf, C. *Chem. Sci.* **2018,** *9,* 7069–7077. doi:10.1039/c8sc01949e
148. Gallarati, S.; Laplaza, R.; Corminboeuf, C. *Org. Chem. Front.* **2022,** *9,* 4041–4051. doi:10.1039/d2qo00550f
149. Anstine, D. M.; Isayev, O. *J. Am. Chem. Soc.* **2023,** *145,* 8736–8750. doi:10.1021/jacs.2c13467
150. Seumer, J.; Kirschner Solberg Hansen, J.; Brøndsted Nielsen, M.; Jensen, J. H. *Angew. Chem., Int. Ed.* **2023,** *62,* e202218565. doi:10.1002/anie.202218565
151. Rasmussen, M. H.; Jensen, J. H. *PeerJ Phys. Chem.* **2022,** *4,* e22. doi:10.7717/peerj-pchem.22
152. Habershon, S. *J. Chem. Theory Comput.* **2016,** *12,* 1786–1798. doi:10.1021/acs.jctc.6b00005
153. Bensberg, M.; Reiher, M. *Isr. J. Chem.* **2023,** *63,* e202200123. doi:10.1002/ijch.202200123
154. Gao, W.; Coley, C. W. *J. Chem. Inf. Model.* **2020,** *60,* 5714–5723. doi:10.1021/acs.jcim.0c00174
155. Weissman, S. A.; Anderson, N. G. *Org. Process Res. Dev.* **2015,** *19,* 1605–1633. doi:10.1021/op500169m
156. Nori, V.; Sinibaldi, A.; Giorgianni, G.; Pesciaioli, F.; Di Donato, F.; Cocco, E.; Biancolillo, A.; Landa, A.; Carlone, A. *Chem. – Eur. J.* **2022,** *28,* e202104524. doi:10.1002/chem.202104524
157. Häse, F.; Roch, L. M.; Kreisbeck, C.; Aspuru-Guzik, A. *ACS Cent. Sci.* **2018,** *4,* 1134–1145. doi:10.1021/acscentsci.8b00307
158. Reker, D.; Hoyt, E. A.; Bernardes, G. J. L.; Rodrigues, T. *Cell Rep. Phys. Sci.* **2020,** *1,* 100247. doi:10.1016/j.xcrp.2020.100247
159. Taylor, C. J.; Pomberger, A.; Felton, K. C.; Grainger, R.; Barecka, M.; Chamberlain, T. W.; Bourne, R. A.; Johnson, C. N.; Lapkin, A. A. *Chem. Rev.* **2023,** *123,* 3089–3126. doi:10.1021/acs.chemrev.2c00798
160. Clayton, A. D.; Manson, J. A.; Taylor, C. J.; Chamberlain, T. W.; Taylor, B. A.; Clemens, G.; Bourne, R. A. *React. Chem. Eng.* **2019,** *4,* 1545–1554. doi:10.1039/c9re00209j
161. Mateos, C.; Nieves-Remacha, M. J.; Rincón, J. A. *React. Chem. Eng.* **2019,** *4,* 1536–1544. doi:10.1039/c9re00116f
162. Sans, V.; Cronin, L. *Chem. Soc. Rev.* **2016,** *45,* 2032–2043. doi:10.1039/c5cs00793c
163. Reizman, B. J.; Jensen, K. F. *Acc. Chem. Res.* **2016,** *49,* 1786–1796. doi:10.1021/acs.accounts.6b00261
164. Fabry, D. C.; Sugiono, E.; Rueping, M. *Isr. J. Chem.* **2014,** *54,* 341–350. doi:10.1002/ijch.201300080
165. Fabry, D. C.; Sugiono, E.; Rueping, M. *React. Chem. Eng.* **2016,** *1,* 129–133. doi:10.1039/c5re00038f
166. James, D. M.; Lindsey, J. S. *JALA (1998-2010)* **2004,** *9,* 364–374. doi:10.1016/j.jala.2004.08.004
167. Houben, C.; Lapkin, A. A. *Curr. Opin. Chem. Eng.* **2015,** *9,* 1–7. doi:10.1016/j.coche.2015.07.001
168. Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021,** *590,* 89–96. doi:10.1038/s41586-021-03213-y
169. Plutschack, M. B.; Pieber, B.; Gilmore, K.; Seeberger, P. H. *Chem. Rev.* **2017,** *117,* 11796–11893. doi:10.1021/acs.chemrev.7b00183
170. Deringer, V. L.; Bartók, A. P.; Bernstein, N.; Wilkins, D. M.; Ceriotti, M.; Csányi, G. *Chem. Rev.* **2021,** *121,* 10073–10141. doi:10.1021/acs.chemrev.1c00022
171. Kondo, M.; Wathsala, H. D. P.; Sako, M.; Hanatani, Y.; Ishikawa, K.; Hara, S.; Takaai, T.; Washio, T.; Takizawa, S.; Sasai, H. *Chem. Commun.* **2020,** *56,* 1259–1262. doi:10.1039/c9cc08526b
172. Kondo, M.; Wathsala, H. D. P.; Salem, M. S. H.; Ishikawa, K.; Hara, S.; Takaai, T.; Washio, T.; Sasai, H.; Takizawa, S. *Commun. Chem.* **2022,** *5,* 148. doi:10.1038/s42004-022-00764-7
173. Tom, G.; Schmid, S. P.; Baird, S. G.; Cao, Y.; Darvish, K.; Hao, H.; Lo, S.; Pablo-García, S.; Rajaonson, E. M.; Skreta, M.; Yoshikawa, N.; Corapi, S.; Akkoc, G. D.; Strieth-Kalthoff, F.; Seifrid, M.; Aspuru-Guzik, A. *Chem. Rev.* **2024,** *124,* 9633–9732. doi:10.1021/acs.chemrev.4c00055
174. Abolhasani, M.; Kumacheva, E. *Nat. Synth.* **2023,** *2,* 483–492. doi:10.1038/s44160-022-00231-0
175. Burger, B.; Maffettone, P. M.; Gusev, V. V.; Aitchison, C. M.; Bai, Y.; Wang, X.; Li, X.; Alston, B. M.; Li, B.; Clowes, R.; Rankin, N.; Harris, B.; Sprick, R. S.; Cooper, A. I. *Nature* **2020,** *583,* 237–241. doi:10.1038/s41586-020-2442-2
176. Boiko, D. A.; MacKnight, R.; Kline, B.; Gomes, G. *Nature* **2023,** *624,* 570–578. doi:10.1038/s41586-023-06792-0
177. Jablonka, K. M.; Schwaller, P.; Ortega-Guerrero, A.; Smit, B. *Nat. Mach. Intell.* **2024,** *6,* 161–169. doi:10.1038/s42256-023-00788-1
178. Bran, A. M.; Cox, S.; Schilter, O.; Baldassari, C.; White, A. D.; Schwaller, P. *Nat. Mach. Intell.* **2024,** *6,* 525–535. doi:10.1038/s42256-024-00832-8
179. Pablo-García, S.; García, Á.; Deniz Akkoc, G.; Sim, M.; Cao, Y.; Somers, M.; Hattrick, C.; Yoshikawa, N.; Dworschak, D.; Hao, H.; Aspuru-Guzik, A. *ChemRxiv* **2024**. doi:10.26434/chemrxiv-2024-cwnwc
180. Tetko, I. V.; Sushko, I.; Pandey, A. K.; Zhu, H.; Tropsha, A.; Papa, E.; Öberg, T.; Todeschini, R.; Fourches, D.; Varnek, A. *J. Chem. Inf. Model.* **2008,** *48,* 1733–1746. doi:10.1021/ci800151m
181. Sushko, I.; Novotarskyi, S.; Körner, R.; Pandey, A. K.; Kovalishyn, V. V.; Prokopenko, V. V.; Tetko, I. V. *J. Chemom.* **2010,** *24,* 202–208. doi:10.1002/cem.1296
182. Schrader, M. L.; Schäfer, F. R.; Schäfers, F.; Glorius, F. *Nat. Chem.* **2024,** *16,* 491–498. doi:10.1038/s41557-024-01470-8
183. Tremouilhac, P.; Nguyen, A.; Huang, Y.-C.; Kotov, S.; Lütjohann, D. S.; Hübsch, F.; Jung, N.; Bräse, S. *J. Cheminf.* **2017,** *9,* 54. doi:10.1186/s13321-017-0240-0
184. Scroggie, K. R.; Burrell-Sander, K. J.; Rutledge, P. J.; Motion, A. *Digital Discovery* **2023,** *2,* 1188–1196. doi:10.1039/d3dd00032j
185. Boobier, S.; Davies, J. C.; Derbenev, I. N.; Handley, C. M.; Hirst, J. D. *J. Chem. Inf. Model.* **2023,** *63,* 2895–2901. doi:10.1021/acs.jcim.3c00306
186. Pistoia Alliance, UDM. https://github.com/PistoiaAlliance/UDM.
187. Jablonka, K. M.; Patiny, L.; Smit, B. *Nat. Chem.* **2022,** *14,* 365–376. doi:10.1038/s41557-022-00910-7
188. Wigh, D. S.; Arrowsmith, J.; Pomberger, A.; Felton, K. C.; Lapkin, A. A. *J. Chem. Inf. Model.* **2024,** *64,* 3790–3798. doi:10.1021/acs.jcim.4c00292
189. Collins, K. D.; Glorius, F. *Nat. Chem.* **2013,** *5,* 597–601. doi:10.1038/nchem.1669
190. Gensch, T.; Teders, M.; Glorius, F. *J. Org. Chem.* **2017,** *82,* 9154–9159. doi:10.1021/acs.joc.7b01139
191. Pitzer, L.; Schäfers, F.; Glorius, F. *Angew. Chem., Int. Ed.* **2019,** *58,* 8572–8576. doi:10.1002/anie.201901935
192. Kariofillis, S. K.; Jiang, S.; Żurański, A. M.; Gandhi, S. S.; Martinez Alvarado, J. I.; Doyle, A. G. *J. Am. Chem. Soc.* **2022,** *144,* 1045–1055. doi:10.1021/jacs.1c12203

193. Rana, D.; Pflüger, P. M.; Hölter, N. P.; Tan, G.; Glorius, F. *ACS Cent. Sci.* **2024,** *10,* 899–906. doi:10.1021/acscentsci.3c01638
194. Saebi, M.; Nan, B.; Herr, J. E.; Wahlers, J.; Guo, Z.; Zurański, A. M.; Kogej, T.; Norrby, P.-O.; Doyle, A. G.; Chawla, N. V.; Wiest, O. *Chem. Sci.* **2023,** *14,* 4997–5005. doi:10.1039/d2sc06041h
195. Kleinmans, R.; Pinkert, T.; Dutta, S.; Paulisch, T. O.; Keum, H.; Daniliuc, C. G.; Glorius, F. *Nature* **2022,** *605,* 477–482. doi:10.1038/s41586-022-04636-x
196. Formica, M.; Rogova, T.; Shi, H.; Sahara, N.; Ferko, B.; Farley, A. J. M.; Christensen, K. E.; Duarte, F.; Yamazaki, K.; Dixon, D. J. *Nat. Chem.* **2023,** *15,* 714–721. doi:10.1038/s41557-023-01165-6
197. Huang, C.; Xiao, P.; Ye, Z.-M.; Wang, C.-L.; Kang, C.; Tang, S.; Wei, Z.; Cai, H. *Org. Lett.* **2024,** *26,* 304–309. doi:10.1021/acs.orglett.3c03980
198. Ji, D.-S.; Zhang, R.; Han, X.-Y.; Hu, X.-Q.; Xu, P.-F. *Org. Lett.* **2024,** *26,* 315–320. doi:10.1021/acs.orglett.3c03861
199. Welch, C. J. *React. Chem. Eng.* **2019,** *4,* 1895–1911. doi:10.1039/c9re00234k
200. Mennen, S. M.; Alhambra, C.; Allen, C. L.; Barberis, M.; Berritt, S.; Brandt, T. A.; Campbell, A. D.; Castañón, J.; Cherney, A. H.; Christensen, M.; Damon, D. B.; Eugenio de Diego, J.; García-Cerrada, S.; García-Losada, P.; Haro, R.; Janey, J.; Leitch, D. C.; Li, L.; Liu, F.; Lobben, P. C.; MacMillan, D. W. C.; Magano, J.; McInturff, E.; Monfette, S.; Post, R. J.; Schultz, D.; Sitter, B. J.; Stevens, J. M.; Strambeanu, I. I.; Twilton, J.; Wang, K.; Zajac, M. A. *Org. Process Res. Dev.* **2019,** *23,* 1213–1242. doi:10.1021/acs.oprd.9b00140
201. Slattery, A.; Wen, Z.; Tenblad, P.; Sanjosé-Orduna, J.; Pintossi, D.; den Hartog, T.; Noël, T. *Science* **2024,** *383,* eadj1817. doi:10.1126/science.adj1817
202. Falivene, L.; Credendino, R.; Poater, A.; Petta, A.; Serra, L.; Oliva, R.; Scarano, V.; Cavallo, L. *Organometallics* **2016,** *35,* 2286–2293. doi:10.1021/acs.organomet.6b00371
203. Ertl, P. *Chem.: Methods* **2022,** *2,* e202200041. doi:10.1002/cmtd.202200041
204. Strieth-Kalthoff, F.; Szymkuć, S.; Molga, K.; Aspuru-Guzik, A.; Glorius, F.; Grzybowski, B. A. *J. Am. Chem. Soc.* **2024,** *146,* 11005–11017. doi:10.1021/jacs.4c00338
205. García Mancheño, O.; Waser, M. *Eur. J. Org. Chem.* **2023,** *26,* e202200950. doi:10.1002/ejoc.202200950

License and Terms

BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

Homogeneous continuous flow nitration of *O*-methylisouronium sulfate and its optimization by kinetic modeling

Jiapeng Guo[1], Weike Su[1] and An Su[*1,2]

Full Research Paper

Open Access

Address:
[1]Key Laboratory of Pharmaceutical Engineering of Zhejiang Province, Key Laboratory for Green Pharmaceutical Technologies and Related Equipment of Ministry of Education, Collaborative Innovation Center of Yangtze River Delta Region Green Pharmaceuticals, Zhejiang University of Technology, Hangzhou, 310014, P. R. China and [2]State Key Laboratory Breeding Base of Green Chemistry-Synthesis Technology, Key Laboratory of Green Chemistry-Synthesis Technology of Zhejiang Province, College of Chemical Engineering, Zhejiang University of Technology, Hangzhou, Zhejiang 310014, China

Email:
An Su[*] - ansu@zjut.edu.cn

* Corresponding author

Keywords:
continuous flow; kinetic modeling; nitration; reaction optimization; static mixer

Beilstein J. Org. Chem. **2024,** *20,* 2408–2420.
https://doi.org/10.3762/bjoc.20.205

Received: 11 June 2024
Accepted: 10 September 2024
Published: 24 September 2024

This article is part of the thematic issue "Adaptive experimentation and optimization in organic chemistry".

Guest Editor: A. Schweidtmann

Abstract

Nitration of *O*-methylisouronium sulfate under mixed acid conditions gives *O*-methyl-*N*-nitroisourea, a key intermediate of neonicotinoid insecticides with high application value. The reaction is a fast and highly exothermic process with a high mass transfer resistance, making its control difficult and risky. In this paper, a homogeneous continuous flow microreactor system was developed for the nitration of *O*-methylisouronium sulfate under high concentrations of mixed acids, with a homemade static mixer eliminating the mass transfer resistance. In addition, the kinetic modeling of this reaction was performed based on the theory of NO_2^+ attack, with the activation energy and pre-exponential factor determined. Finally, based on the response surface generated by the kinetic model, the reaction was optimized with a conversion of 87.4% under a sulfuric acid mass fraction of 94%, initial reactant concentration of 0.5 mol/L, reaction temperature of 40 °C, molar ratio of reactants at 4.4:1, and a residence time of 12.36 minutes.

Introduction

The demand for high-quality insecticides is increasing as the world's food crisis intensifies due to the changes in the natural environment and ongoing geopolitical crises [1]. *O*-Methyl-*N*-nitroisourea (NIO) is a pivotal pesticide intermediate in the preparation of nitroguanidine derivatives, which are the raw material for highly effective and non-toxic neonicotinoid insecticides, such as dinotefuran and clothianidin [2-4]. Currently, the industrial production of *O*-methyl-*N*-nitroisourea usually involves the nitration of *O*-methylisouronium sulfate (IO) with a mixture of sulfuric acid (H_2SO_4) and nitric acid (HNO_3) in a batch reactor [3]. The reaction is a typical aliphatic nitration, which is fast and highly exothermic, requiring low reaction temperatures. In addition, the safety hazard of this reaction is increased by using concentrated nitric and sulfuric acids. Therefore, it is necessary to modify the nitrification reaction process of *O*-methylisouronium sulfate to improve the reaction efficiency and intrinsic safety.

In recent years, continuous flow microreactors have been recognized due to their excellent mass and heat transfer performance, precise control over reaction parameters, and intrinsic safety [5-8]. Guo et al. constructed a continuous flow microsystem for *o*-xylene nitrification and proved the process safety of by the adiabatic temperature rise of the nitrification reaction and the characteristic heat transfer time of the microreactor [9]. The residence time of the microreactor was reduced by an order of magnitude and the volumetric mass transfer coefficient was increased by several orders of magnitude compared with that of a conventional stirred-tank reactor. Jin et al. developed a continuous flow microreactor system for the non-homogeneous nitrification of nitrobenzene using mixed acids [10]. The reaction time and temperature were reduced from >2 h and 80 °C in industrial operation to 10 min and 65 °C in the microreactor with high conversion and selectivity. Since *O*-methylisouronium sulfate can be dissolved in high concentrations of sulfuric acid, it is expected to construct a homogeneous continuous flow nitrification system, leading to better elimination of the effects of mass and heat transfer [11].

Kinetic modeling is a classical approach to chemical reaction optimization, where the effects of various reaction parameters on the results are effectively quantified by mathematical formulas, thus providing an efficient guide to optimize reaction conditions [12]. Taylor et al. [13] and Bures et al. [14] have performed kinetic modeling with data collected from continuous flow systems with automated platforms. Yao et al. constructed a kinetic model on thermal dissociation and oligomerization of dicyclopentadiene (DCPD) in a continuous flow microreactor [15]. Where cyclopentadiene was the target intermediate formed by the thermal dissociation of dicyclopentadiene, cascade oligomerization was a side reaction to be avoided. Based on the deep understanding of the kinetic differences between thermal dissociation and oligomerization, the residence time and temperature were designed rationally to improve the yield of cyclopentadiene. Since NO_2^+ is the actual substance that plays a role in the nitrification process [16], kinetic modeling based on the concentration of NO_2^+ is essential for the understanding of the nitrification mechanism and optimization of the reaction. Luo et al. have carried out extensive research on this topic and obtained accurate kinetic data for the nitration of chlorobenzene [17], *o*-nitrotoluene [18], and *p*-nitrotoluene [19] by constructing a homogeneous continuous flow reaction system. Therefore, it is feasible to model homogeneous nitrification and optimize the reaction in a continuous flow system based on NO_2^+.

An important prerequisite for kinetic modeling is the elimination of issues related to mass and heat transfer. The effect of mass transfer resistance is greater for highly viscous reaction systems, especially at higher reactant concentrations. It is still difficult to eliminate the mass transfer effect using conventional microreactors, leading to errors in the determination of nitration kinetics. Therefore, more efficient mixers are needed to overcome the effects of mass and heat transfer. According to the mixing principle, there are active mixers and passive mixers. Passive mixers do not require overly complex equipment and external energy inputs and are extensively used in continuous flow reactions [20,21]. Passive mixers enhance the passive mixing of the liquid–liquid two-phase mass transfer process on a microscopic scale, mainly by optimizing the microchannel geometry [22], addition of in-channel obstacles, etc. [23-25]. Santana et al. designed an efficient fluid mixer "Elis" consisting of internal walls and circular obstacles. This static mixer achieves efficient mixing in a wide range of Reynolds numbers at the micro- and milliscale. However, many static mixer designs are structurally complex and require the use of 3D printing technology to aid in their manufacture, which is more expensive to use. Kilcher et al. investigated in detail the efficient mixing of organic phases (cyclopentadiene, 1,2-dichloroethane, and $MeBu_3NCl$) and aqueous phases (30% NaOH) and optimized it by the use of a simple homemade “PTFE Raschig ring static mixer” (RRSM). The RRSM is simple in structure, easy to fabricate, inexpensive for many flow reaction systems, and has a promising application.

In this work, we constructed a continuous flow microreactor system to determine the kinetic parameters of IO nitration, which allows precise control of temperature and residence time (Figure 1). Due to the high viscosity of the reaction system, a simple and practical static mixer was designed to eliminate the effect of mass transfer on the kinetic measurements and validated experimentally. We developed a kinetic model for the nitration of *O*-methylisouronium sulfate and optimized the reaction conditions for conversion rates, which is crucial for theoretical significance and practical value for process optimization.

Results and Discussion

In this section, we perform kinetic modeling for the continuous flow synthesis of NIO from IO and mixed acid (Scheme 1). The reaction was then optimized by kinetic modeling.

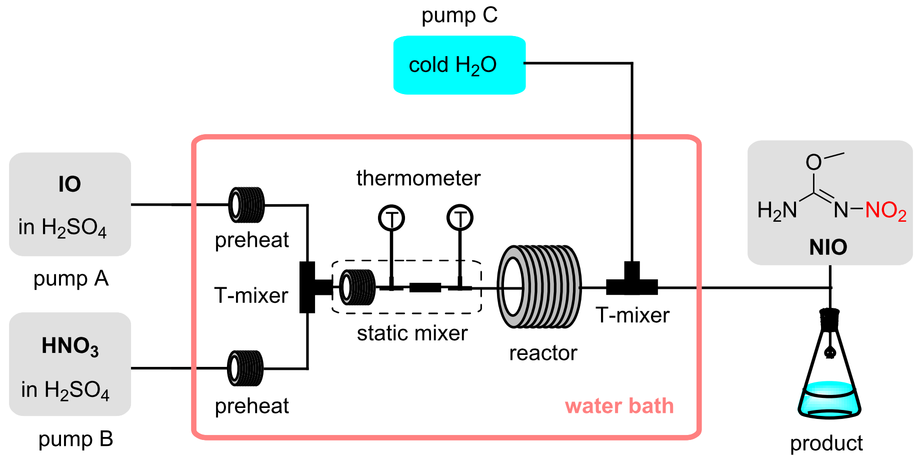
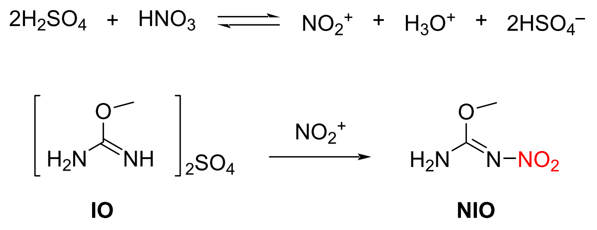


Figure 1: The schematic diagram of the continuous flow microreactor system.



Scheme 1: Nitration of IO with mixed acid.

Prescreening experiments

The solubility of IO in H_2SO_4 is critical in ensuring the smooth progression of the nitration reaction within a homogeneous system. Given the strongly exothermic nature of this reaction, an excessively high concentration of IO can lead to an overproduction of heat, thereby elevating the associated risks. In contrast, a concentration that is too low may fall beneath the detection threshold, compromising the reliability of the experimental data. To strike a balance, the initial concentration of IO was set to 0.5 mol/L in the reaction mixture. In addition, the effect of the molar ratio between the two reactants was examined. As shown in Figure S1 in Supporting Information File 1, the conversion of IO gradually increased as the molar ratio of HNO_3 elevated. The molar ratio of HNO_3 was established at 4.4 equiv, a value chosen to optimize both conversion and atom efficiency.

Effect of two types of mixing equipment

Upon achieving homogeneous nitration conditions, our next objective was to eliminate the influence of mass transfer. We assessed the impact of flow rate on the reaction conversion under two distinct mixing scenarios (Figure 2a and 2c). The assessments were performed with reaction temperatures at 30–40 °C to eliminate the impact of the high viscosity of sulfuric acid [26]. Figure 2a illustrates the scenario employing solely a T-mixer and Figure 2c shows the effect of flow rate on the conversion under this setup. Even when the flow rate was escalated to 14 mL/min, the conversion failed to stabilize at a plateau, suggesting that mass transfer limitations had not been fully addressed. Conversely, with the addition of our home-made static mixer which consists of a 1/16-inch mixing coil and a SiO_2 beads-filled column (Figure 2b), the conversion rate plateaued once the total flow rate surpassed 8 mL/min (Figure 2d), suggesting the elimination of mass transfer limitations. The improved mixing efficiency can be attributed to the mixer's design features, such as its double reverse rotating vortex [27,28], large specific surface area [29], and the incorporation of obstacles within the flow channel [30,31].

Determining reaction orders

The reaction orders for IO and HNO_3 were determined in the continuous flow microreactor system, and the results are shown in Figure S2 of Supporting Information File 1. The initial concentration of HNO_3 was set at a level significantly higher (14 times greater) than that of IO. This approach allowed for the assumption that the concentration of HNO_3 remained constant throughout the reaction, enabling the conversion of the rate constant to K_β (Equation 1). The relationship between reaction time and the conversion of IO was analyzed according to the first-order (Equation 2) and second-order (Equation 3) reaction kinetics, where x_{IO} represented the conversion of IO and t denoted the reaction time. The outcome of these fittings is presented in Figure 3a for first-order and Figure 3b for second-order. Notably, the higher R^2 observed in Figure 3a compared to Figure 3b suggests that the reaction of IO follows first-order kinetics.

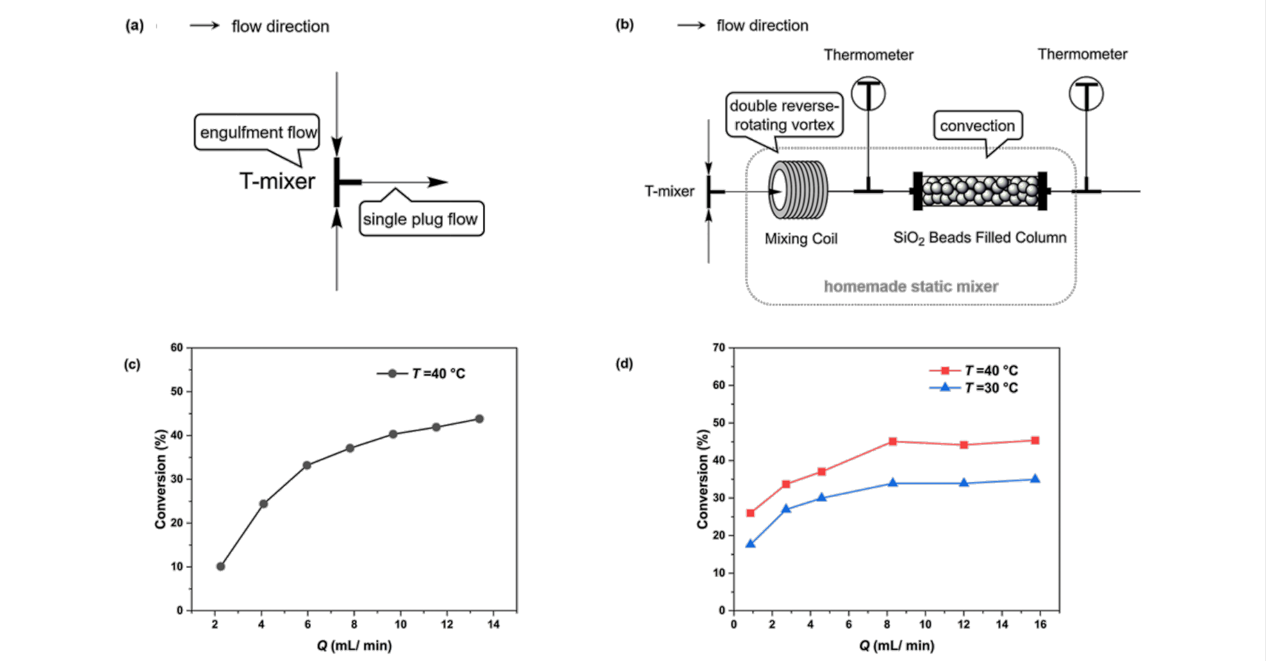


Figure 2: Two mixing setups: (a) a T-mixer and (b) a T-mixer combined with a homemade static mixer, and the effect of the two mixing setups on the mixing process; (c) the T-mixer and (d) the T-mixer plus the homemade static mixer effect of flow rate on conversion. Reaction conditions: H_2SO_4 mass fractions = 98%, reaction temperature T = 40 °C, residence time t = 2 min, initial concentration of reactants c_{IO} = 1 mol/L, c_{HNO3} = 4.4 mol/L.

$$-\frac{dc_{IO}}{dt} = kc_{IO}{}^{\alpha}c_{HNO_3}{}^{\beta} \approx K_{\beta}c_{IO}{}^{\alpha} \quad (1)$$

$$\ln(1-x_{IO}) = -K_{\beta}t \quad (2)$$

$$\frac{1}{1-x_{IO}} = 1-K_{\beta}t \quad (3)$$

Given that the reaction order of IO was determined to be 1, Equation 1 was subsequently transformed into Equation 4. As nitration reactions are predominantly second-order, we explored the potential for the reaction order of HNO_3 (β) to be either 0 or 1 by fitting the reaction data to Equation 5 and Equation 6, respectively.

$$-\frac{dc_{IO}}{dt} = kc_{IO}c_{HNO_3}{}^{\beta} \quad (4)$$

$$\ln(1-x_{IO}) = -Kt \quad (5)$$

$$\frac{1}{c_{HNO_3\,0} - c_{IO\,0}}\ln\left(\frac{1-x_{HNO_3}}{1-x_{IO}}\right) = -Kt \quad (6)$$

The fitting results, as depicted in Figure 3c for β = 0 and Figure 3d for β = 1, revealed that R^2 for the latter scenario (R^2 = 0.993) was higher than that for the former (R^2 = 0.986). This outcome indicates that the reaction order of HNO_3 is also 1, which transforms Equation 4 into Equation 7.

$$-\frac{dc_{IO}}{dt} = kc_{IO}c_{HNO_3} \quad (7)$$

Also, with M=cHNO30cIO0, Equation 6 can be rewritten to Equation 8.

$$\ln\left[\frac{M-x_{IO}}{M(1-x_{IO})}\right] = (M-1)c_{IO0}kt \quad (8)$$

After the reaction order being determined, the rest of the experiments were conducted in the continuous flow reactor and t in Equation 8 refers to the residence time.

Determining the apparent reaction kinetics

The variation in the conversion of IO (x_{IO}) as the function of time (t) at different temperatures (30 °C, 35 °C, 40 °C) and H_2SO_4 mass fractions (88%, 90%, 92%, 94%, 96%, and 98%) is depicted in Figure S3 in Supporting Information File 1 and subsequently modeled using Equation 8. The fitting results shown in Figure 4 exhibit robust linear correlations ($R^2 > 0.99$), facilitating the calculation of rate constants based on the slopes of these lines across the varied temperatures and H_2SO_4 concentrations. Table 1 indicates that the reaction rate constants

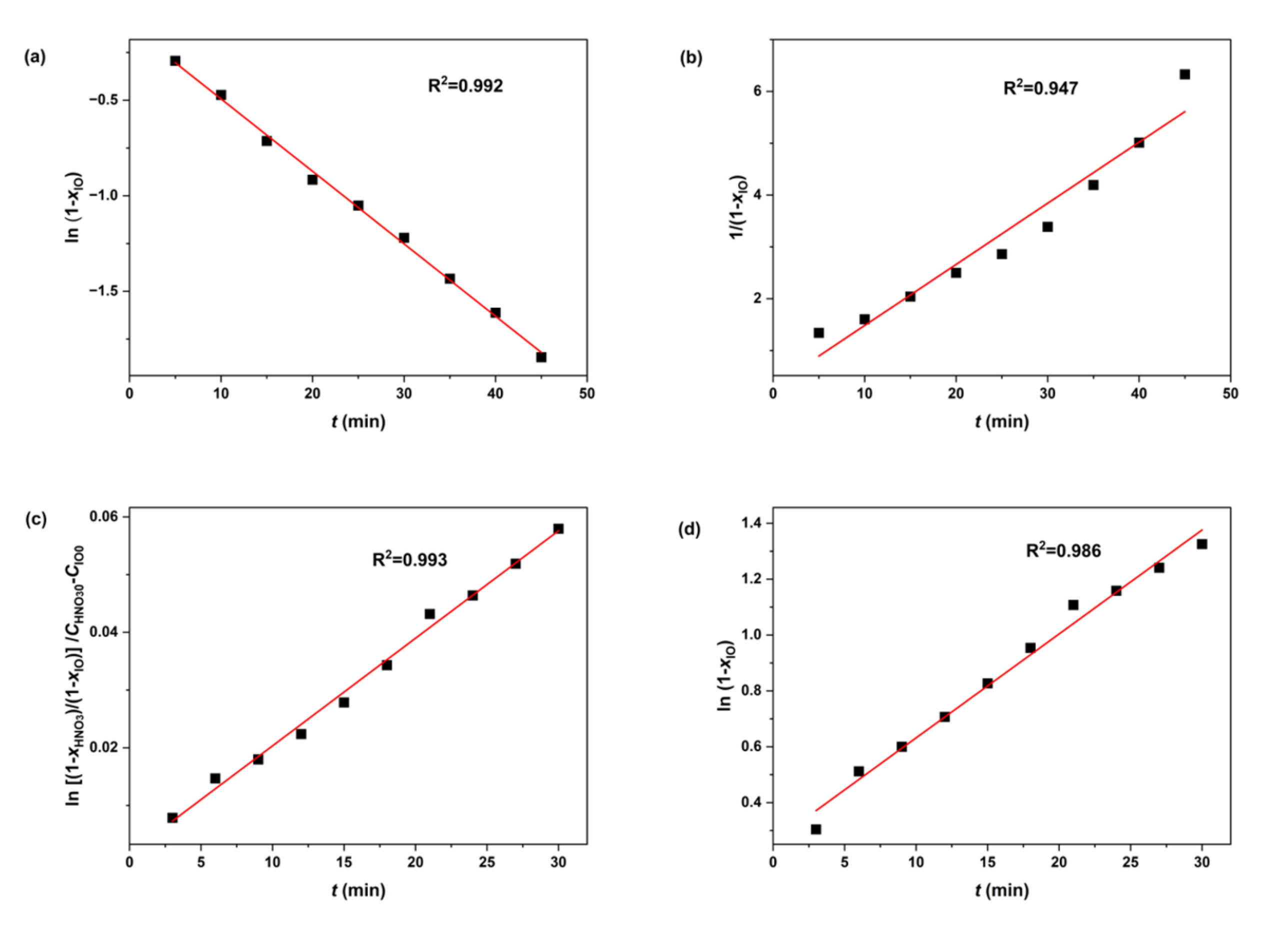


Figure 3: Determination of the number of reaction orders. a) ln(1−x_{IO}) versus *t*; b) $\frac{1}{1-x_{IO}}$ versus *t*; c) ln(1−x_{IO}) versus *t*; d) $\frac{1}{c_{HNO_30}-c_{IO0}}\ln\left(\frac{1-x_{HNO_3}}{1-x_{IO}}\right)$ versus *t*. Reaction conditions for determining IO's reaction order: H_2SO_4 mass fractions = 98%, reaction temperature, *T* = 0 °C; initial concentration of reactants in the reaction mixture: c_{IO0} = 1 mol/L, c_{HNO30} = 15 mol/L. Reaction conditions for determining HNO_3's reaction order: reaction temperature, *T* = 0 °C; initial concentration of reactants in the reaction mixture: c_{NIO0} = 1 mol/L, c_{HNO30} = 4.4 mol/L.

escalate with increasing H_2SO_4 mass fraction, which aligns with the findings from previous studies on mixed acid-catalyzed nitration reactions [32,33]. However, the data also reveal a decline in rate constants when the H_2SO_4 mass fraction exceeds 94%, suggesting a complex interaction at higher acid concentrations.

Table 1: Values of *k* for different H_2SO_4 mass fractions and at different temperatures.

Mass fraction of H_2SO_4 (wt %)	k × 10^2 (L/mol/s)		
	30 °C	35 °C	40 °C
88	2.26	2.98	4.31
90	5.51	6.91	8.40
92	7.48	10.2	12.6
94	10.3	12.3	15.1
96	9.56	10.8	11.4
98	8.13	8.70	9.37

Determining the intrinsic reaction kinetics

Given the strong correlation between the observed HNO_3-based reaction rate constant and the H_2SO_4 mass fraction, intrinsic reaction constants independent of H_2SO_4 concentrations were determined to study the intrinsic kinetics of the reaction. Previous research has established that the relationship between the apparent and intrinsic kinetics of nitrification can be described by Equation 9 [17,19].

$$\lg k = \lg\left(\frac{c_{NO_2^+}}{c_{HNO_3}}\right) + nM_c + \lg k_0 \qquad (9)$$

where k_0 is the intrinsic rate constant only based on NO_2^+ and independent of sulfuric acid concentration [34], *n* is a thermo-

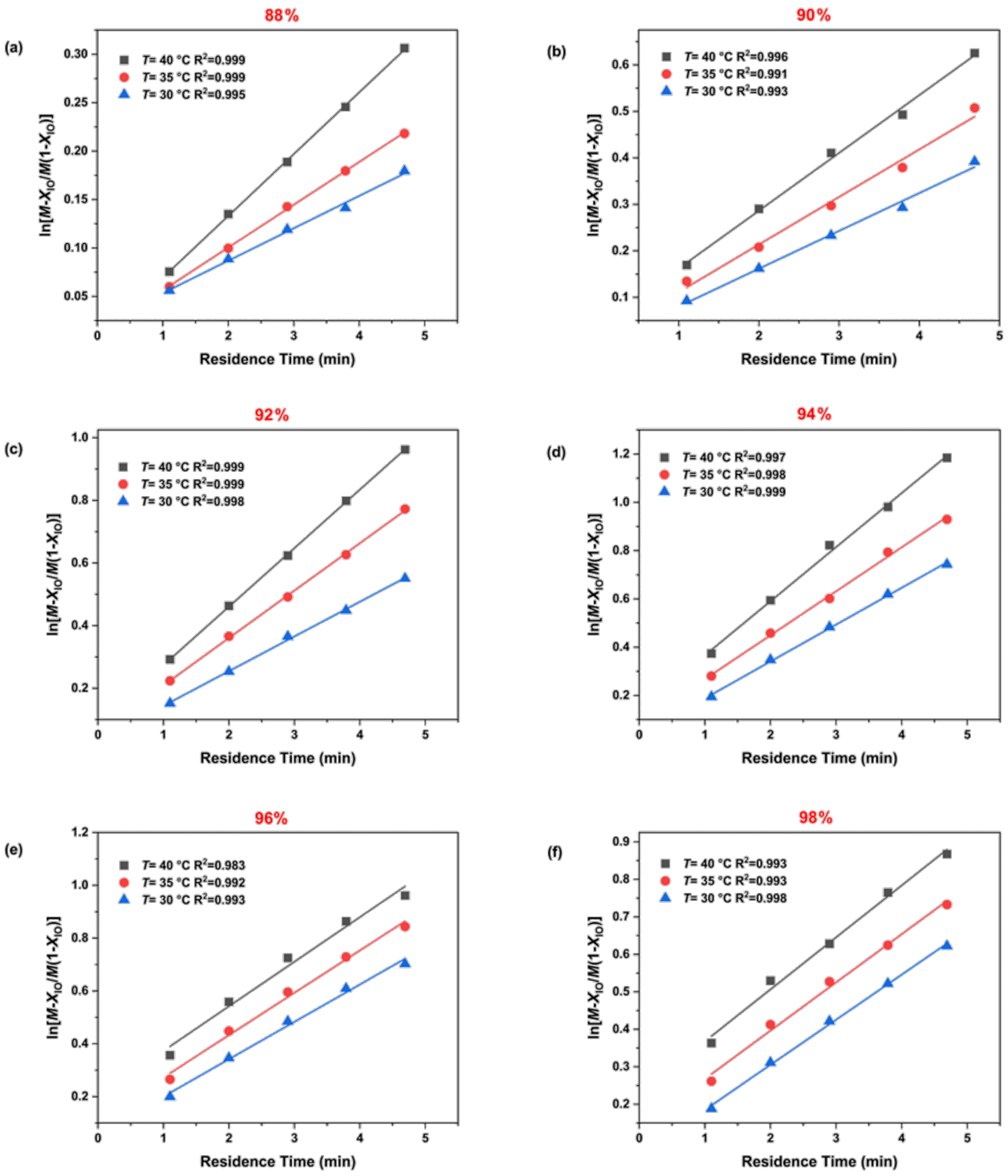


Figure 4: Determination of (M−1)$c_{IO0}k$ at different temperatures and H_2SO_4 mass fractions. (a) 88% H_2SO_4, (b) 90% H_2SO_4, (c) 92% H_2SO_4, (d) 94% H_2SO_4, (e) 96% H_2SO_4, and (f) 98% H_2SO_4.

dynamic parameter related to the type of compound, and M_c is the activity coefficient function introduced in the next section.

By shifting the terms in Equation 9, Equation 10 can be obtained as:

$$\lg k - \lg\left(\frac{c_{NO_2^+}}{c_{HNO_3}}\right) = nM_c + \lg k_0 \qquad (10)$$

Therefore, by plotting

$$\lg k - \lg\left(\frac{c_{NO_2^+}}{c_{HNO_3}}\right)$$

as the vertical coordinate and M_c as the horizontal coordinate, the values of n and k_0 can be obtained from the slope and intercept of the resulting fitting line. Since the values of M_c and

$$\lg\left(\frac{c_{NO_2^+}}{c_{HNO_3}}\right)$$

change with the change in temperature and sulfuric acid mass fraction, we determined the values of M_c and

$$\lg\left(\frac{c_{NO_2^+}}{c_{HNO_3}}\right)$$

according to the method proposed by Luo et al. [17,19]. As the ranges of sulfuric acid concentrations and temperature in our study were different from Luo et al.'s study, recalculations were required to obtain the values of M_c and

$$\lg\left(\frac{c_{NO_2^+}}{c_{HNO_3}}\right).$$

Determination of M_c values

The value of M_c can be calculated using Equation 11 and Equation 12. Equation 11 [35] was employed to predict M_c at various H_2SO_4 concentrations at 298 K, specifically when the H_2SO_4 concentrations were between 15.2 and 18.4 mol/L. By fitting the predicted data, M_c as a function of the H_2SO_4 concentration at a given temperature was determined (Figure S4 in Supporting Information File 1). In addition, the values of M_c for different sulfuric acid concentrations at a given temperature can be obtained by substituting the corresponding temperature into Equation 12, as first introduced by Marziano et al.

$$\begin{aligned} -M_c(298\text{ K}) = 2.16\times10^{-4}c_{H_2SO_4}{}^5 - 1.27\times10^{-2}c_{H_2SO_4}{}^4 \\ +0.28c_{H_2SO_4}{}^3 - 2.73c_{H_2SO_4}{}^2 + 10.6c_{H_2SO_4} \end{aligned} \qquad (11)$$

$$M_c(T) = M_c(298\text{ K})\left[\frac{200}{T} + 0.3292\right] \qquad (12)$$

Determination of lg(c_{NO2+}/c_{HNO3}) values

Since NO_2^+ is the actual reactive species in the nitration reaction, an accurate estimation of its concentration is essential for the study of intrinsic kinetics. Based on the values of $\frac{c_{NO_2^+}}{c_{HNO_3}}$, reported in previous studies for different temperatures and sulfuric acid concentrations [36-38], the mass fraction of sulfuric acid was plotted against

$$\lg\left(\frac{c_{NO_2^+}}{c_{HNO_3}}\right).$$

The fitting results shown in Figure 5a exhibit robust linear correlations, enabling the calculation of

$$\lg\left(\frac{c_{NO_2^+}}{c_{HNO_3}}\right)$$

at temperatures of 23 °C, 40 °C, and 60 °C. Subsequently, by plotting

$$\lg\left(\frac{c_{NO_2^+}}{c_{HNO_3}}\right)$$

versus $1/T$, a series of fitted curves for the studied range of sulfuric acid concentrations (88–98 wt %) can be obtained, as shown in Figure 5b. Thus, the values of

$$\lg\left(\frac{c_{NO_2^+}}{c_{HNO_3}}\right)$$

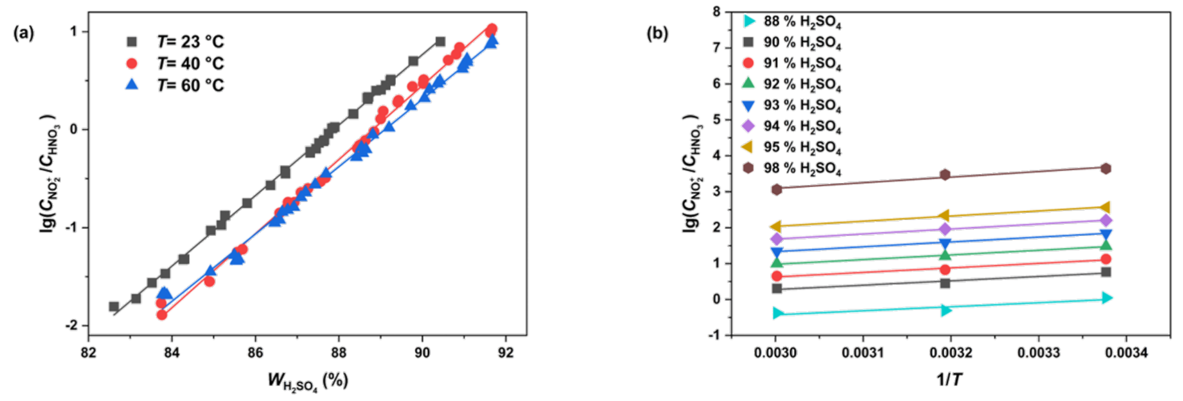


Figure 5: Variations and fitting of as a function of a) the mass fraction of H_2SO_4 at 23 °C, 40 °C, and 60 °C and b) 1/*T* at different H_2SO_4 concentrations $\lg\frac{c_{NO_2^+}}{c_{HNO_3}}$.

at different sulfuric acid concentrations and temperatures can be determined.

Determination of intrinsic kinetic parameters

With

$$\lg\left(\frac{c_{NO_2^+}}{c_{HNO_3}}\right)$$

and M_c at different conditions determined in Figure 5,

$$\lg k - \lg\left(\frac{c_{NO_2^+}}{c_{HNO_3}}\right)$$

was plotted against M_c at different temperatures (Figure 6a–c), and fitting these data into Equation 10 leads to ($R^2 > 0.99$). The values of k_0 and n at different temperatures are shown in Table 2. The value of k_0 increases with increasing temperature and the value of n remains almost constant with temperature, which is consistent with the results reported in previous studies for other mixed acid-catalyzed nitration reactions [17,39].

Table 2: Values of *n* and lgk_0 at different temperatures.

Temperature (°C)	*n*	lgk_0
30	1.0764	11.3749
35	1.1127	11.8556
40	1.1577	12.4352

According to the values of k_0 at different temperatures, the activation energy for the electrophilic attack of NO_2^+ on the IO can be calculated by the Arrhenius equation:

$$\ln k_0 = -\frac{E_a}{RT} + \ln A \quad (13)$$

where R is the molar gas constant and T denotes the temperature in Kelvin, and E_a and A are the activation energy and pre-exponential factors for the IO nitration.

By fitting lnk_0 versus 1/T into Equation 13 (Figure 6d), the values of E_a and ln$A$ were determined (Table 3).

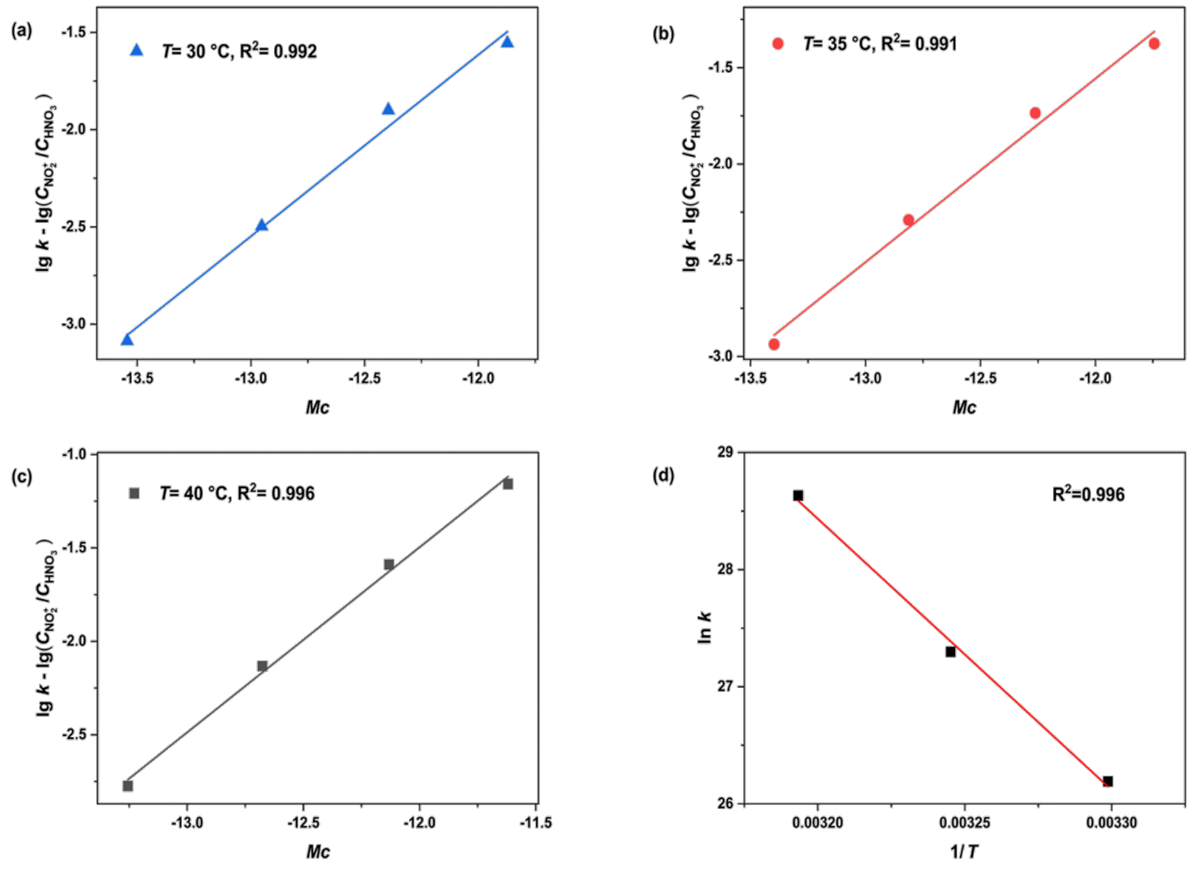


Figure 6: Determination of thermodynamic parameters *n* and k_0 and determination of the activation energy and pre-exponential factors.

Table 3: Values of the pre-exponential factor and activation energy.

Factors	E_a (kJ/mol)	ln*A*
values	192.57	102.55

The synergic effect of temperature and sulfuric acid concentration on the apparent kinetics

As discussed above, the apparent rate constant is determined by three components,

$$\lg\left(\frac{c_{\mathrm{NO_2^+}}}{c_{\mathrm{HNO_3}}}\right),$$

nM_c, and $\lg k_0$. First, the intrinsic rate constant k_0 is only temperature-dependent and is not affected by the concentration of sulfuric acid (Equation 9). In addition, Figure 5b shows that

$$\lg\left(\frac{c_{\mathrm{NO_2^+}}}{c_{\mathrm{HNO_3}}}\right)$$

increases with the increase in sulfuric acid concentration when the temperature is fixed. In contrast, nM_c is a negative value that decreases with higher sulfuric acid concentration (Table S1 in Supporting Information File 1). As the concentration of sulfuric acid increases, the decrease in nM_c gradually surpassed the increase in

$$\lg\left(\frac{c_{\mathrm{NO_2^+}}}{c_{\mathrm{HNO_3}}}\right)$$

when the sulfuric acid concentration exceeded 94%, resulting in an overall decrease of k (Figure 7a). Similar trends were reported in the nitration of nitrobenzene [40] and *o*-nitrotoluene [18], suggesting that the phenomenon observed in our study is not isolated.

Validation, extrapolation, and optimization

To validate the kinetic model and assess its ability to extrapolate, we conducted 18 experiments varying three residence times, three reaction temperatures, and two sulfuric acid concentrations. We then compared the theoretical and experimental values of conversion rates under these conditions (Figure 7b and Table S2 in Supporting Information File 1). Notably, 16 of these experiments were performed with a residence time exceeding the upper limit of the model construction, 4.7 min. The results revealed a strong alignment between the predicted and experimental conversion rates, with an average discrepancy of less than 2%. The smallest error was observed with a 98% sulfuric acid concentration at 35 °C and a residence time of 8.0 min, where the theoretical and experimental values nearly matched. Conversely, the largest error was at 94% sulfuric acid concentration, 40 °C, and a residence time of 9.3 min, with theoretical and experimental values of 90% and 86%, respectively. Increasing the residence time to 12.36 min amplified the error to approximately 8% (Figure 7c). A similar increase in error with prolonged residence time was noted in Kappe et al.'s kinetic modeling of the Buchwald–Hartwig amination reaction [41], where the theoretical and experimental values diverged by 4.1% when the residence time increased from 0.5 min to 4.2 min.

Building on the model's demonstrated ability to extrapolate at prolonged residence times, we performed additional experiments with the reaction temperature increased to 45 °C (Figure 7c and Table S3 in Supporting Information File 1). This temperature exceeds the highest temperature used during the initial development of the kinetic model, which was 40 °C. This further extrapolation led to a 10% error at a residence time of 13.7 min, inferring that it would be prudent to avoid increasing the temperature to 45 °C if the aim is to maintain the discrepancy between the model predicted and experimental conversion rates below 10%.

Based on the observations above, the optimized reaction conditions were obtained: the sulfuric acid mass fraction was 94%, the initial concentration of IO was 0.5 mol/L, the reaction temperature was 40 °C, the molar ratio was 4.4:1, and the reaction time was 12.36 min. Under these conditions, the experimentally measured conversion was 87.4%.

This study marks the first time that the intrinsic kinetics of this reaction have been reported and utilized to optimize the process of nitration of *O*-methylisouronium sulfate within a continuous flow device. The highly exothermic nature of nitration makes the conversion from batch to continuous flow significantly safer. Additionally, the optimization model demonstrates excellent scalability and can accurately predict reaction conversions, with errors not exceeding 4%, for residence times beyond the modeling range (extending from the initial 1–5 minutes to 5–12 minutes in validation experiments). Compared to the original patent [2], the reaction time has been significantly reduced from tens of minutes to hours to less than 20 minutes while maintaining a lower sulfuric acid mass fraction and achieving higher conversion rates. Furthermore, the process does not require low temperatures, thereby reducing energy consumption and simplifying operation.

Conclusion

In this work, a homogeneous nitration system for the synthesis of *O*-methyl-*N*-nitroisourea was constructed. To eliminate the

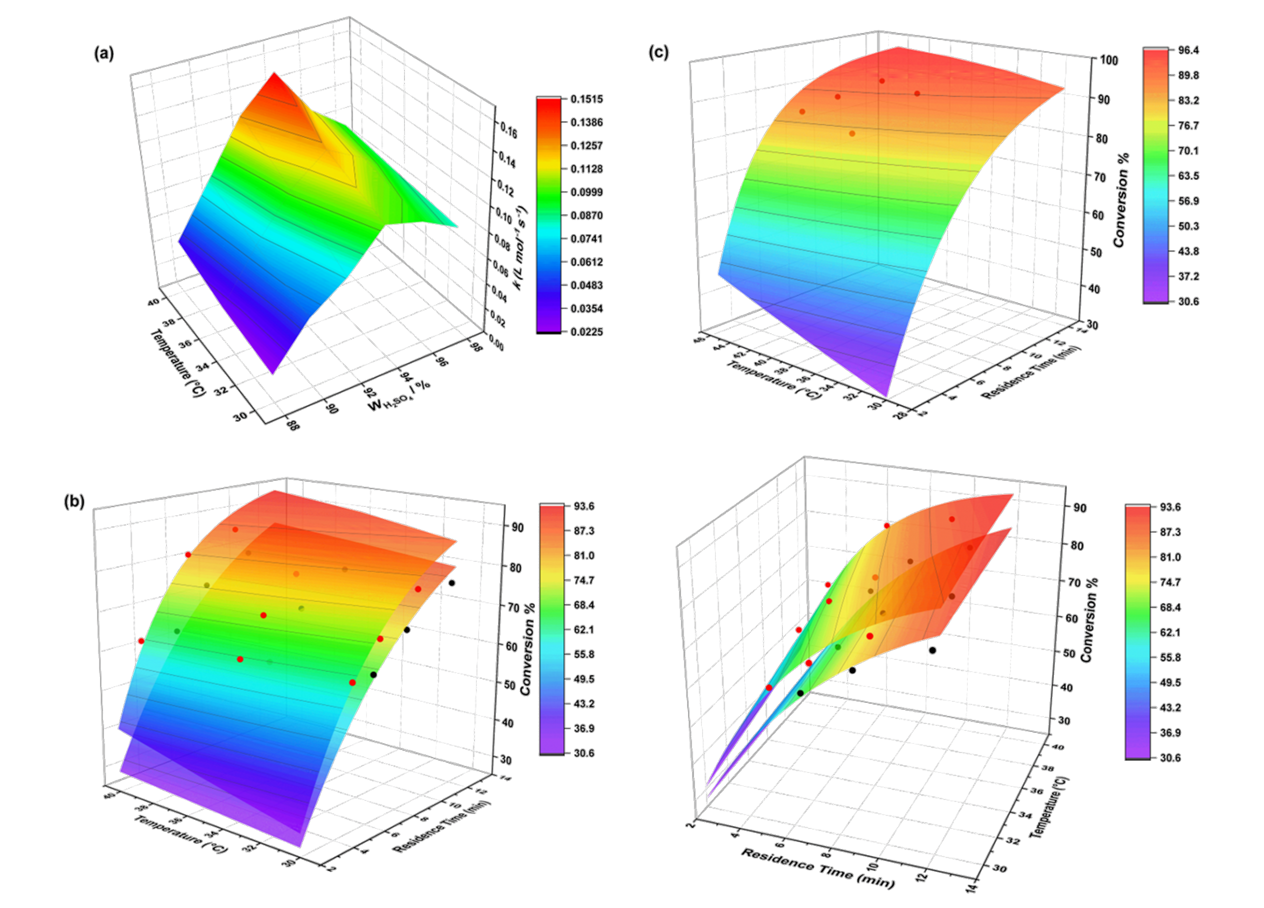


Figure 7: a) The value of apparent rate constant *k* at various H_2SO_4 mass fractions and different temperatures. b) Validation of reaction kinetic models with different mass fractions of sulfuric acid (from top to bottom: the theoretical response surfaces and experimental values are shown for sulfuric acid mass fractions of 94% and 98%, respectively). c) Response surface of the kinetic model under optimized conditions and experimental results. The red dots show the experimental values for a sulfuric acid mass fraction of 94% and the black dots show the experimental values for a sulfuric acid mass fraction of 98%.

mass transfer resistance between the two liquid phases during the reaction, a homemade simple and effective static mixer was used which rapidly achieved thorough mixing of the two phases with little temperature fluctuation. The effects of temperature, residence time, and sulfuric acid mass fraction on the reaction were investigated as well as the apparent and intrinsic rate constants based on nitric acid and NO_2^+ observations were obtained, respectively. The apparent rate constants observed based on nitric acid are highly correlated with the mass fraction of sulfuric acid, increasing and then decreasing as the mass fraction of sulfuric acid increases, with 94% sulfuric acid being the turning point. This is the result of a combination of the intrinsic rate constant, the sulfuric acid activity coefficient function, and the NO_2^+ concentration. Thus, the effect of different sulfuric acid mass fractions and temperatures on the apparent rate constants can be understood. In addition, a complete kinetic model of IO nitration based on NO_2^+ was developed to describe the reaction process, the activation energy of the IO nitration was calculated to be 192.57 kJ/mol. Furthermore, the accuracy of the kinetic model was verified by comparing the predicted data with the experimental data. Finally, the reaction was optimized by kinetic modeling and 87.4% conversion of IO was achieved under optimum conditions. This kinetic model can be used to understand the nitration process of IO and optimize the reactor design, which can serve as guidance for industrial production.

Experimental

Materials and methods

Chemicals

O-Methylisouronium sulfate (IO, 95%) was purchased from Shanghai Yien Chemical Technology Co., Ltd; fuming nitric acid (HNO_3, 98.0%) was purchased from Sinopharm Chemical Reagent Co., Ltd.; sulfuric acid (H_2SO_4, 98.0%) was purchased from Sinopharm Chemical Reagent Co., Ltd.; pure water from AR, Hangzhou Wahaha Group Co., Ltd.; all reagents were used without further purification. Sulfuric acid solutions of different mass fractions were prepared with pure water and 98% concentrated sulfuric acid in an ice bath with stirring.

Solution A (IO): IO (0.1 mol, 24.64 g) was dissolved in H_2SO_4 (100 mL) under stirring conditions in an ice bath, solution volume V_A = 118 mL.

Solution B (H_2SO_4 + HNO_3): HNO_3 (0.44 mol, 18.49 mL) was dissolved in H_2SO_4 (100 mL) under stirring conditions in an ice bath, solution volume V_B = 112 mL.

Continuous flow microreactor system

The continuous flow microreactor system is shown in Figure 1. Solutions A and B were stored in two glass vials (500 mL) with lids and were preheated by two high-pressure PTFE pumps (pump A, pump B, JJRZ-10004F, Hangzhou JingJin Technology Co., Ltd.) and pumped into coiled stainless steel capillary tubes (SS316L, 1/16-inch diameter) that were sufficiently long (1 m). After being preheated to reaction temperature, the material was first initially mixed in a T-mixer (SS316L, 1/16-inch diameter), followed by a homemade static mixer at the outlet of the T-mixer to fully mix the material. The reaction coil (SS316L, 1/8-inch diameter) was connected directly to the outlet of the homemade static mixer, nitration took place in the reaction coil. The residence time was precisely controlled by changing the flow rate of the reaction mixture or the length of the reaction coil. All preheat tubes, mixers, and reaction coils were immersed in the same water bath to maintain a constant temperature. Finally, after controlling the residence time, the reaction was terminated by pumping excess pure ice water through a high-pressure PTFE pump (Pump C, JJRZ-10004F, Hangzhou Jingjin Technology Co., Ltd.) into the second T-mixer.

The homemade static mixer consisted of two different mixing units as shown in Figure 2b (total internal volume: 1.3154 mL). The first mixing unit consists of a section of stainless steel coil (SS316L, 1/16-inch diameter, Beijing Xiongchuan Technology Co. Ltd.) and an electronic thermometer (Beijing Xiongchuan Technology Co. Ltd.). The second mixing unit consisted of a section of PTFE piping filled with SiO_2 beads (SiO_2 beads, 3 mm diameter; piping, 1/4-inch diameter,10 cm length, Wuxi Hongxin Special Material Technology Co.) and an electronic thermometer connected to the outlet.

Sample analysis

When the continuous flow system was operated at steady state (after 2–3 times the residence time), the reaction solution was quenched and diluted by a large amount of ice water at the outlet of the reaction system. The quenched and diluted reaction solution was collected and analyzed by high-performance liquid chromatography (HPLC, ThermoFisher Ulcel3000), and the conversion of the samples was derived from the external standard method based on the regression equation of the HPLC standard curve. HPLC detection conditions: C18 column (10 μm, 4.6 × 250 mm, Welch Materials Shanghai, China), the mobile phase was 80% MeOH and 20% ultrapure water at a flow rate of 1 mL/min, and the detection wavelength was 195 nm. The conversion of IO was calculated by the following equation:

$$x_{\mathrm{IO}} = \left(1 - \frac{c_{\mathrm{IO}}}{c_{\mathrm{IO}} + c_{\mathrm{NIO}}}\right) \qquad (14)$$

The residence time was calculated as follows:

$$t = \frac{V}{Q_{\mathrm{IO}} + Q_{\mathrm{HNO_3}}} \qquad (15)$$

where t is the reaction residence time and V is the volume of the microchannel. Q_{IO} and Q_{HNO3} are the volume flow rates of the raw material aqueous solution, respectively. Samples were tested three times under the same conditions and averaged to minimize errors.

Kinetic modeling optimization process

The classical integral method was employed to determine the reaction order [42]. Various integral forms of kinetic equations corresponding to different reaction orders were fitted against the experimental data. The reaction orders yielding the highest R^2 were selected as the best fit. Subsequently, the least squares method was used to fit the kinetic data obtained under different reaction conditions, allowing for the determination of the pre-exponential factors and activation energies. Finally, the accuracy of the resulting kinetic model was validated through experimental testing.

Supporting Information

Supporting Information File 1

Additional information.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-20-205-S1.pdf]

Funding

This research was supported by the Joint Funds of the Zhejiang Provincial Natural Science Foundation of China under Grant No. LHDMZ23B060001, Zhejiang Province Science and Technology Plan Project under Grant No. 2022C01179, and the National Natural Science Foundation of China under Grant No. 22108252.

Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Contributions

Jiapeng Guo: data curation; formal analysis; investigation; methodology; validation; visualization; writing – original draft. Weike Su: conceptualization; project administration; resources; supervision; writing – review & editing. An Su: conceptualization; funding acquisition; project administration; resources; supervision; writing – review & editing.

ORCID® iDs

An Su - https://orcid.org/0000-0002-6544-3959

Data Availability Statement

All data that supports the findings of this study is available in the published article and/or the supporting information to this article.

Preprint

A non-peer-reviewed version of this article has been previously published as a preprint: https://doi.org/10.3762/bxiv.2024.39.v1

References

1. Elbert, A.; Haas, M.; Springer, B.; Thielert, W.; Nauen, R. *Pest Manage. Sci.* **2008,** *64,* 1099–1105. doi:10.1002/ps.1616
2. Uneme, H.; Konobe, M.; Ishizuka, H.; Kamiya, Y. Preparation of heteroarylmethylisoureas and related compounds. WO Patent WO9700867, Jan 9, 1997.
3. Uneme, H.; Kamiya, Y.; Konobe, M.; Yamada, J. Manufacture of *N*-(heterocyclylmethyl)-*N'*-nitroisoureas. WO Patent WO9933809, July 8, 1999.
4. Brady, J. F.; Simmons, D. P.; Wilson, T. E. Immunoassay for neonicotinoid insecticides. WO Patent WO2001042787, June 14, 2001.
5. Köckinger, M.; Wyler, B.; Aellig, C.; Roberge, D. M.; Hone, C. A.; Kappe, C. O. *Org. Process Res. Dev.* **2020,** *24,* 2217–2227. doi:10.1021/acs.oprd.0c00254
6. Magosso, M.; van den Berg, M.; van der Schaaf, J. *React. Chem. Eng.* **2021,** *6,* 1574–1590. doi:10.1039/d1re00141h
7. Sheng, L.; Chen, Y.; Wang, K.; Deng, J.; Luo, G. *Chem. Eng. Sci.* **2021,** *239,* 116621. doi:10.1016/j.ces.2021.116621
8. Sheng, L.; Ma, L.; Chen, Y.; Deng, J.; Luo, G. *Chem. Eng. J.* **2022,** *427,* 132067. doi:10.1016/j.cej.2021.132067
9. Guo, S.; Zhan, L.-w.; Li, B.-d. *Chem. Eng. J.* **2023,** *468,* 143468. doi:10.1016/j.cej.2023.143468
10. Jin, N.; Song, Y.; Yue, J.; Wang, Q.; Lu, P.; Li, Y.; Zhao, Y. *Chem. Eng. Sci.* **2023,** *281,* 119198. doi:10.1016/j.ces.2023.119198
11. Rahaman, M.; Mandal, B. P.; Ghosh, P. *AIChE J.* **2007,** *53,* 2476–2480. doi:10.1002/aic.11222
12. Taylor, C. J.; Pomberger, A.; Felton, K. C.; Grainger, R.; Barecka, M.; Chamberlain, T. W.; Bourne, R. A.; Johnson, C. N.; Lapkin, A. A. *Chem. Rev.* **2023,** *123,* 3089–3126. doi:10.1021/acs.chemrev.2c00798
13. Taylor, C. J.; Booth, M.; Manson, J. A.; Willis, M. J.; Clemens, G.; Taylor, B. A.; Chamberlain, T. W.; Bourne, R. A. *Chem. Eng. J.* **2021,** *413,* 127017. doi:10.1016/j.cej.2020.127017
14. Burés, J.; Larrosa, I. *Nature* **2023,** *613,* 689–695. doi:10.1038/s41586-022-05639-4
15. Yao, Z.; Xu, X.; Dong, Y.; Liu, X.; Yuan, B.; Wang, K.; Cao, K.; Luo, G. *Chem. Eng. Sci.* **2020,** *228,* 115892. doi:10.1016/j.ces.2020.115892
16. Hughes, E. D.; Ingold, C. K.; Reed, R. I. *Nature* **1946,** *158,* 448–449. doi:10.1038/158448c0
17. Cui, Y.; Song, J.; Du, C.; Deng, J.; Luo, G. *AIChE J.* **2022,** *68,* e17564. doi:10.1002/aic.17564
18. Song, J.; Cui, Y.; Luo, G.; Deng, J.; Wang, Y. *React. Chem. Eng.* **2022,** *7,* 111–122. doi:10.1039/d1re00362c
19. Song, J.; Cui, Y.; Sheng, L.; Wang, Y.; Du, C.; Deng, J.; Luo, G. *Chem. Eng. Sci.* **2022,** *247,* 117041. doi:10.1016/j.ces.2021.117041
20. Zhang, S.; Zhu, C.; Feng, H.; Fu, T.; Ma, Y. *Chem. Eng. Sci.* **2021,** *229,* 116040. doi:10.1016/j.ces.2020.116040
21. Holvey, C. P.; Roberge, D. M.; Gottsponer, M.; Kockmann, N.; Macchi, A. *Chem. Eng. Process.* **2011,** *50,* 1069–1075. doi:10.1016/j.cep.2011.05.016
22. Tajik Ghanbari, T.; Rahimi, M.; Ranjbar, A. A.; Pahamli, Y.; Torbatinejad, A. *Phys. Fluids* **2023,** *35,* 122013. doi:10.1063/5.0177344
23. Al-Azzawi, M.; Mjalli, F. S.; Husain, A.; Al-Dahhan, M. *Ind. Eng. Chem. Res.* **2021,** *60,* 5049–5075. doi:10.1021/acs.iecr.0c05858
24. Bhagat, A. A. S.; Peterson, E. T. K.; Papautsky, I. *J. Micromech. Microeng.* **2007,** *17,* 1017–1024. doi:10.1088/0960-1317/17/5/023
25. Su, Y.; Chen, G.; Yuan, Q. *Chem. Eng. Sci.* **2011,** *66,* 2912–2919. doi:10.1016/j.ces.2011.03.024
26. Dunstan, A. E. *Proc. Chem. Soc., London* **1914,** *30,* 104–105.
27. Dean, W. R.; Hurst, J. M. *Mathematika* **1959,** *6,* 77–85. doi:10.1112/s0025579300001947
28. Jonas Bolinder, C.; Sundén, B. *Exp. Therm. Fluid Sci.* **1995,** *11,* 348–363. doi:10.1016/0894-1777(95)00040-2
29. Lü, Y.; Zhu, S.; Wang, K.; Luo, G. *Chin. J. Chem. Eng.* **2016,** *24,* 711–718. doi:10.1016/j.cjche.2016.01.011
30. Tata Rao, L.; Goel, S.; Kumar Dubey, S.; Javed, A. *J. Phys.: Conf. Ser.* **2019,** *1276,* 012003. doi:10.1088/1742-6596/1276/1/012003
31. Bazargan-Lari, Y.; Movahed, S.; Mashhoodi, M. *J. Mech.* **2017,** *33,* 387–394. doi:10.1017/jmech.2016.81
32. Li, S.; Zhang, X.; Ji, D.; Wang, Q.; Jin, N.; Zhao, Y. *Chem. Eng. Sci.* **2022,** *255,* 117657. doi:10.1016/j.ces.2022.117657
33. Yang, M.; Liao, C.; Tang, C.; Zhang, P.; Huang, Z.; Li, J. *Phys. Chem. Chem. Phys.* **2021,** *23,* 4658–4668. doi:10.1039/d0cp05935h
34. Olah, G. A.; Malhotra, R.; Narang, S .C. NITRATION: Methods and Mechanisms. *Across Conventional Lines;* World Scientific Publishing, 1989; pp 975–979.
35. Marziano, N. C.; Tomasin, A.; Traverso, P. G. *J. Chem. Soc., Perkin Trans. 2* **1981,** 1070–1075. doi:10.1039/p29810001070
36. Deno, N. C.; Peterson, H. J.; Sacher, E. *J. Phys. Chem.* **1961,** *65,* 199–201. doi:10.1021/j100820a002
37. Marziano, N. C.; Tomasin, A.; Tortato, C.; Zaldivar, J. M. *J. Chem. Soc., Perkin Trans. 2* **1998,** 1973–1982. doi:10.1039/a802521e
38. Ross, D. S.; Kuhlmann, K. F.; Malhotra, R. *J. Am. Chem. Soc.* **1983,** *105,* 4299–4302. doi:10.1021/ja00351a030

39. Wen, Z.; Yang, M.; Zhao, S.; Zhou, F.; Chen, G. *React. Chem. Eng.* **2018,** *3,* 379–387. doi:10.1039/c7re00182g
40. Rahaman, M.; Mandal, B.; Ghosh, P. *AIChE J.* **2010,** *56,* 737–748. doi:10.1002/aic.11989
41. Wagner, F.; Sagmeister, P.; Jusner, C. E.; Tampone, T. G.; Manee, V.; Buono, F. G.; Williams, J. D.; Kappe, C. O. *Adv. Sci.* **2024,** *11,* 2308034. doi:10.1002/advs.202308034
42. Xu, Q.; Fan, H.; Yao, H.; Wang, D.; Yu, H.; Chen, B.; Yu, Z.; Su, W. *Chem. Eng. J.* **2020,** *398,* 125584. doi:10.1016/j.cej.2020.125584

License and Terms

The definitive version of this article is the electronic one which can be found at:
https://doi.org/10.3762/bjoc.20.205

BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

Machine learning-guided strategies for reaction conditions design and optimization

Lung-Yi Chen[1] and Yi-Pei Li[*1,2]

Address:
[1]Department of Chemical Engineering, National Taiwan University, No. 1, Sec. 4, Roosevelt Road, Taipei 10617, Taiwan and [2]Taiwan International Graduate Program on Sustainable Chemical Science and Technology (TIGP-SCST), No. 128, Sec. 2, Academia Road, Taipei 11529, Taiwan

Email:
Yi-Pei Li[*] - yipeili@ntu.edu.tw

* Corresponding author

Review

Open Access

Beilstein J. Org. Chem. **2024,** *20,* 2476–2492.
https://doi.org/10.3762/bjoc.20.212

Received: 28 June 2024
Accepted: 19 September 2024
Published: 04 October 2024

This article is part of the thematic issue "Adaptive experimentation and optimization in organic chemistry".

Guest Editor: A. Schweidtmann

Keywords:
data preprocessing; reaction conditions prediction; reaction data mining; reaction optimization; reaction representation

Abstract

This review surveys the recent advances and challenges in predicting and optimizing reaction conditions using machine learning techniques. The paper emphasizes the importance of acquiring and processing large and diverse datasets of chemical reactions, and the use of both global and local models to guide the design of synthetic processes. Global models exploit the information from comprehensive databases to suggest general reaction conditions for new reactions, while local models fine-tune the specific parameters for a given reaction family to improve yield and selectivity. The paper also identifies the current limitations and opportunities in this field, such as the data quality and availability, and the integration of high-throughput experimentation. The paper demonstrates how the combination of chemical engineering, data science, and ML algorithms can enhance the efficiency and effectiveness of reaction conditions design, and enable novel discoveries in synthetic chemistry.

Introduction

Machine learning (ML) techniques have been widely applied to various chemical-related tasks, such as computer-aided synthesis planning (CASP) [1-4], which can recommend possible synthetic routes for a target molecule and potentially improve the efficiency of developing new synthetic pathways. Many studies have shown that ML-based retrosynthesis models can reproduce patent-derived pathways for known compounds, and even suggest more diverse and efficient alternatives [5-8]. Building upon the retrosynthesis, the reaction conditions prediction models can help in identifying appropriate conditions for each step, ensuring compatibility with the platform and addressing safety concerns. On the other aspect, forward reaction prediction normally plays the role of validating the feasibility of a reaction pathway predicted by retrosynthetic models and to further enhance reaction yields by optimizing reaction parameters such as temperature, pressure, and solvent choice, thus it polishes and trims the suggested routes. As a result, CASP tools have attracted commercial interest and stimulated the development of integrated robotic platforms for automated flow synthesis [9-11].

However, as Coley et al. [12] pointed out, there are still challenges to achieve a fully automated and self-driving synthesis process. One of the key challenges is to automatically select appropriate reaction conditions for each synthesis step without human intervention. Conventionally, the common strategy to determine suitable reaction conditions is to adopt the previously reported conditions for the same or similar reaction types and conduct several experimental trials to evaluate the resulting reaction yields. However, this empirical approach is unlikely to find the optimal conditions, since the reaction outcome depends on a large and complex combination of factors, such as catalysts, solvents, substrate concentrations, and temperature. In academia, especially, the "one factor at a time" (OFAT) approach, which involves changing one factor while keeping the others constant, is frequently used to examine the effect of individual reaction parameters [13]. However, the OFAT method is simplistic and may fail to identify the optimal reaction conditions, since it ignores the possible interactions among the experimental factors.

With the rapid development of high-throughput experimentation (HTE) techniques and ML, it has become more feasible to collect large volumes of data and accelerate the prediction of optimal reaction condition combinations. It has been widely demonstrated that ML algorithms can be used for various chemistry-related tasks, such as yield prediction [14,15], site-selectivity prediction [16,17], reaction conditions recommendation [18], and reaction conditions optimization [13]. These techniques have also been integrated with robotic platforms to speed up the discovery and synthesis of new materials and drug candidates, showcasing the potential and promising benefits of self-driving chemistry labs [19].

Raghavan et al. [20] compared two types of reaction condition models based on their scope of applicability and dataset size: global and local models. The global models cover a wide range of reaction types and typically predict the experimental conditions based on a predefined list derived from literature data. However, this method requires sufficient and diverse reaction data for training, so that the models can have broader applicability and usefulness for CASP in autonomous robotic platforms [12,21]. On the other hand, the local models focus on a single reaction type. Generally, more fine-grained levels of experimental conditions, such as substrate concentrations, bases, and additives, are considered in local models. The development of these models usually involves using HTE [22-24] for efficient data collection, coupled with Bayesian optimization (BO) [25] for searching the best reaction conditions to achieve the desired reaction outcomes.

In this review, we delve into the various methodologies used for predicting and optimizing reaction conditions, and illustrate their diverse applications across different chemical domains. Given the importance of data collection for building data-driven models, we review different aspects of the dataset features and data preprocessing methods. Moreover, we introduce common algorithms and representative studies for developing both global and local models. We highlight representative studies that demonstrate the effectiveness and applicability of these algorithms in real-world chemical scenarios. Finally, we summarize the progress in this field and underline the remaining challenges in the area of reaction condition design.

Review

Reaction data collection and preprocessing

One of the major challenges in building ML models for global reaction conditions prediction is the data scarcity and diversity, as they need to cover a vast reaction space [26,27]. However, collecting data relevant to chemical reactions represents a significant challenge. While specific molecular properties can be precisely computed using existing simulation methods like quantum chemical calculations – allowing for the generation of extensive data through large-scale simulations – chemical reactions pose a much greater difficulty for accurate simulation. The development of systematic theoretical calculations to model correlations between reaction yields and various substrates and catalysts requires extensive effort. This involves complex parameter optimization, meticulous validation against experimental data, and careful consideration of diverse reaction conditions and possible reaction mechanisms [28,29]. Although some studies employ transition-state (TS) theory to simulate activation energies and compute reaction enthalpy for particular types of reactions [30], this approach often demands significant computational resources to determine accurate TSs and activation energies. The complexity increases further when considering the impacts of solvents and catalysts, which means that large-scale theoretical calculations are typically restricted to gas-phase reactions [31]. Despite these challenges, recent advances in quantum chemical methods have shown that theoretical calculations can provide practical guidance for validating experimental results [32]. Thus, we posit that the role of theoretical calculations in generating data for ML applications will grow increasingly critical. At present, employing theoretical calculations systematically to construct accurate, large-scale databases of reaction conditions remains highly challenging for complex reaction systems, leading to a primary reliance on experimental data for building ML models.

Overview of data sources for chemical reaction modeling

Table 1 summarizes some of the commonly used chemical reaction databases and their characteristics. These databases differ in the types and sources of reactions they contain [33], as well as in the formats used for data recording. Predominantly, these databases rely on experimental chemical data; however, most are proprietary and require subscription-based access. This restricts the availability and comparability of data essential for developing global reaction conditions prediction models and often leads to duplicated efforts in data collection. For instance, Gao et al. [18] trained a reaction conditions recommender on about 10 million reactions from Reaxys [34], but subsequent studies could not access or use the same data for model evaluation or improvement [35]. To address this issue, Coley et al. proposed the Open Reaction Database (ORD) [36], an open-source initiative to collect and standardize chemical synthesis data from various literature sources. The ORD allows chemists to upload reaction data associated with their publications, and aims to serve as a benchmark for ML development. However, the ORD is still in its infancy and contains mostly literature-extracted USPTO data [37], with only a small fraction of manually curated data. Therefore, there is a need for more community involvement and data contribution to make the ORD a comprehensive and reliable resource for global reaction modeling.

Local reaction datasets, on the other hand, usually focus on a specific reaction family and record reactions with relatively less structural variation in reactants and products. Various combinations of reaction conditions are tested to investigate the output yields in these reaction-specific datasets, which are typically obtained from HTE [41]. Some representative datasets are summarized in Table 2 and can be retrieved from the original papers or ORD. Local reaction datasets have several advantages over global datasets, despite containing less than 10k reactions. For instance, HTE data include failed experiments with zero yields, which are often omitted in large-scale commercial databases that only extract the most successful conditions per reference, as discussed by Chen et al. [42]. This selection bias can lead to overestimation of reaction yields by ML models and limit their generalization capabilities [43]. Therefore, many studies have called for more comprehensive documentation of all experimental results and submission of data in machine-readable formats [44-46]. Another potential issue with data from various sources is the discrepancy in yield definition, as pointed out by Mercado et al. [47]. Literature-extracted yields can be derived from different methods, such as crude yield, isolated yield, quantitative NMR, and liquid chromatography area percentage, and they can also vary in precision due to human bias or equipment quality. HTE data for specific reactions, however, are usually measured using more standardized procedures and are less affected by this issue. In summary, while global models have the appealing feature of wider applicability, local models offer a more practical fit for optimizing real chemical reaction conditions [20]. The choice of datasets depends on the application scenario, whether it is to establish a comprehensive CASP system or to focus on specific reaction types.

Besides the existing datasets, alternative approaches for constructing reaction data through automatic literature mining have also been proposed. These approaches leverage the rapid advancement of natural language processing (NLP) techniques to extract experimental data from unstructured text. For example, Vaucher et al. [69] combined rule-based models and deep-learning techniques to convert experimental procedures into standardized synthetic steps. They further used this data extraction technique to construct a dataset of ≈693k reactions with detailed procedures and developed a sequence-to-sequence model to predict synthetic steps that are actionable and compatible with robotic platforms [70]. Guo et al. [71] conducted a continual pretraining scheme on the BERT model [72] to obtain a domain-adaptive encoder, ChemBERT, which was pretrained on an unlabeled corpus of ≈200k chemical journal articles. They then finetuned ChemBERT on a small annotated dataset for reaction role labeling, resulting in ChemRxnBERT, which can identify the reaction transformation and distinguish reactants,

Table 1: Summary of large-scale chemical reaction databases.

Database	Reference	No. of the reactions	Availability
Reaxys	[34]	≈65 millions	proprietary
ORD	[36]	≈1.7 million reactions from USPTO [37] and ≈91k reactions from the chemical community	open access
Scifinder^n	[38]	≈150 millions	proprietary
Pistachio	[39]	≈13 millions	proprietary
Spresi	[40]	≈4.6 millions	proprietary

Table 2: Summary of chemical reaction yield datasets obtained from HTE.

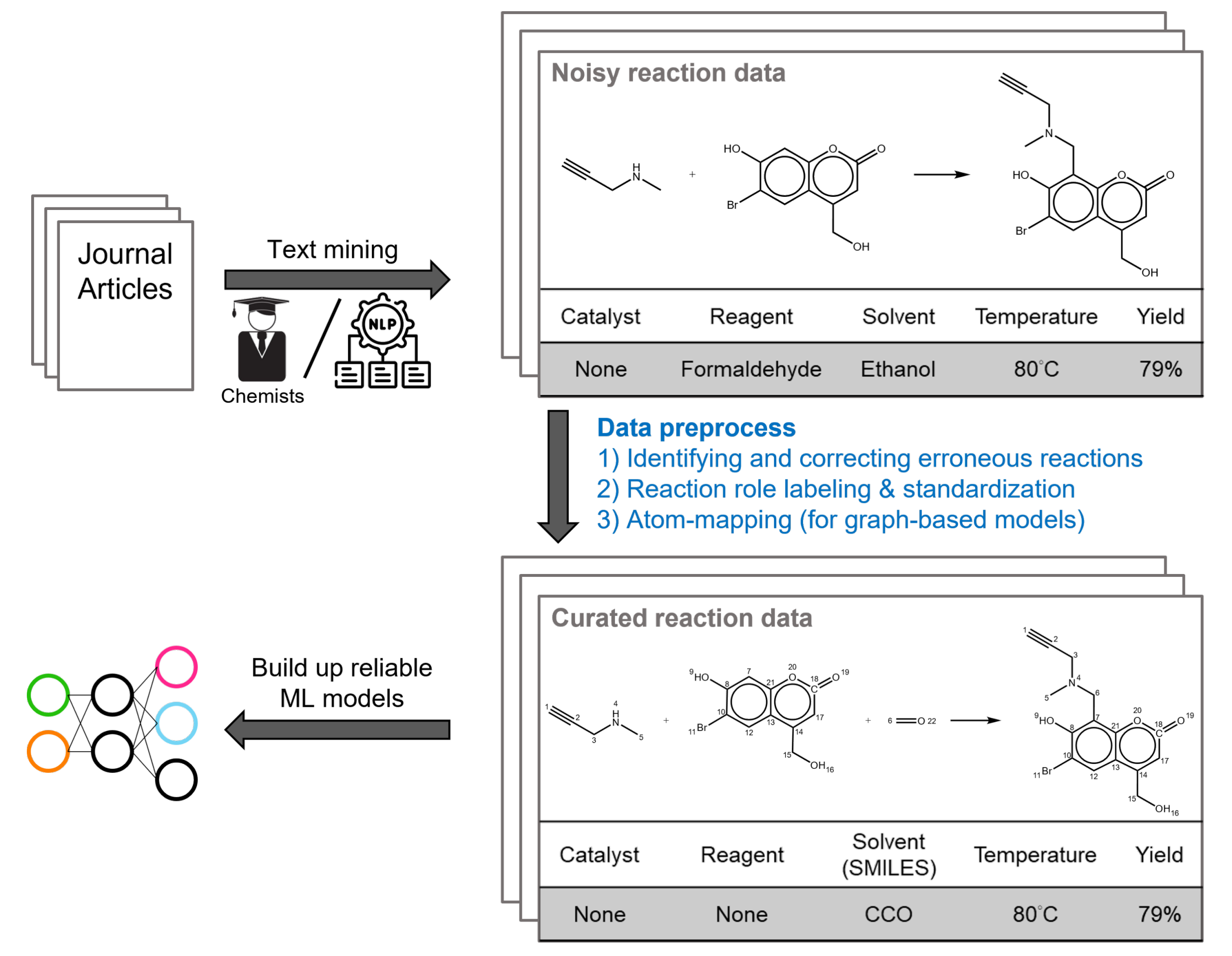
Dataset	Reference	No. of reactions
Buchwald–Hartwig (1)	[48]	4,608
Buchwald–Hartwig (2)	[49]	288
Buchwald–Hartwig (3)	[50]	750
Pd-catalyzed cross-coupling	[49]	1,536
Suzuki–Miyaura coupling (1)	[51]	5,760
Suzuki–Miyaura coupling (2)	[52]	384
Suzuki–Miyaura coupling (3)	[53]	534
electroreductive coupling of alkenyl and benzyl halides	[54]	27
Mizoroki–Heck reaction	[55]	384
coupling of α-carboxyl sp3-carbons with aryl halides	[56]	24
Biginelli condensation	[57]	48
deoxyfluorination	[58]	80
coupling reactions	[59]	264
synthesis of sulfonamide	[60]	39
Ni-catalyzed Suzuki–Miyaura	[61]	450
Mitsunobu reaction	[62]	40
Ni-catalyzed borylation	[63]	1,296
amide coupling (1)	[64]	1,280
amide coupling (2)	[65]	960
Pd-catalysed C–H arylation	[65]	1,536
Ni-catalyzed C–O coupling	[66]	2,003
Ir(I)-catalyzed O–H bond insertion	[67]	653
Pd-catalyzed C–N coupling	[68]	767

catalysts, solvents, and reagents from chemistry passages. However, many chemical literature records depict reactions using diagrams, which can have various formats such as single-line, multiple-line, tree, and graph representations. Extracting data from reaction diagrams requires the use of image recognition to parse molecular structures and convert them into textual representations. Qian et al. [73,74] demonstrated that this task of optical chemical structure recognition (OCSR) [75] can be handled with a model that combines an image encoder and a molecular graph decoder. Despite the promising machine-learning solutions for reaction diagram parsing [76,77], there are still some limitations. For instance, sometimes the reaction conditions are listed in tables, and certain functional groups in images are represented by abbreviations (e.g., R-groups). To achieve more complete data extraction, future efforts will need to employ multi-modal modeling approaches [78-80] that can collect information from different sources and provide robust results. Recently, Fan et al. developed the OpenChemIE toolkit [81], which integrates extraction methods from text, images, and tables, automating the capture of experimental records of chemical reactions from chemical synthesis papers. This development demonstrates significant advancements in streamlining the data extraction process for chemical research.

Implicit data issues and data preprocessing tools

The quality of training data is a crucial factor for the robustness of ML models in chemistry. However, chemical reaction data may contain errors or incompleteness, which can adversely affect the model performance and reliability. The common errors in reaction data can be roughly categorized into two types: (1) erroneous reactions, such as those with mislabeled, missing, or extra atoms in reactants or products, and (2) incomplete reactions, such as those with missing reactants, which are often due to insufficient documentation of the involved species. Erroneous reactions usually require the removal of the corresponding entries from the dataset, as it is hard to determine whether the recorded reactants or products are correct and consistent. Incomplete reactions could be mitigated by using heuristic methods to complete the missing species. In this section, we explain the details of data collection and preprocessing, and we present a schematic representation of the workflow in Figure 1.

One approach to remove erroneous reactions is based on the concept of "catastrophic forgetting", which refers to the model's tendency to forget previously learned events during the training process. Toniato et al. [83] proposed to use this idea as


Journal Articles
Text mining
Chemists
NLP
Noisy reaction data
HO
Br
O
O
OH
N
Data preprocess
1) Identifying and correcting erroneous reactions
2) Reaction role labeling & standardization
3) Atom-mapping (for graph-based models)
Curated reaction data
Build up reliable ML models

Catalyst	Reagent	Solvent	Temperature	Yield
None	Formaldehyde	Ethanol	80°C	79%

Catalyst	Reagent	Solvent (SMILES)	Temperature	Yield
None	None	CCO	80°C	79%

Figure 1: Schematic diagram illustrating the data mining and preprocessing steps for chemical reaction datasets. This process includes data collection, filtering, completion, and atom mapping. In the depicted example, formaldehyde, which contributes a carbon atom to the product, is classified as a reactant due to its active involvement in the reaction. It is also noteworthy that while reaction conditions are ideally encoded in the simplified molecular-input line-entry system (SMILES) to aid in text-based model predictions [82], certain reagents and catalysts, especially those comprising specific mixtures, cannot be represented in SMILES and are instead described using textual labels.

a criterion to filter out the reactions that are more difficult for the model to learn, assuming that they are more likely to contain errors. However, this protocol depends on the choice of the model and does not require any chemistry-informed knowledge for preprocessing.

For dealing with incomplete reactions, the first step is to identify the missing component, which can be facilitated by atom-mapping packages [84-87] that assign a unique label to each atom in the reactants and products. With the atom-mapping information, one can apply the rule-based method, CGRTools [88], to add small molecules (e.g., H_2O and HCl) in reactions, but this method is limited by the availability and coverage of predefined reaction rules. Alternatively, language models have been developed to predict the missing part of molecules given a partial reaction equation, as reported in the work of Zipoli et al. [89] and Zhang et al. [90]. These ML-based approaches can balance reactions without exhaustive rule definition, but they may not be able to recover complex molecules. A promising data preprocessing strategy that addresses this issue is proposed by Phan et al. [91], who formulated the omission of molecules as a maximum common subgraph (MCS) problem and aligned reactants and products to identify non-overlapping segments, thereby generating the missing compounds. Another novel method is AutoTemplate [92], which extracts generic reaction templates from the reactions being preprocessed and recursively applies them on the products of the dataset to validate and correct reaction data. This approach can not only fill in missing reactants, but also fix atom-mapping errors and remove incorrect data entries, thus improving the quality of chemical reaction datasets.

Although many data preprocessing tools have been proposed, we believe more research in this direction can be beneficial to the performance and reliability of machine learning models. Ideally, a unified standard data processing workflow should be established in the future to benefit various reaction prediction and synthesis tasks.

Reaction representations for reaction modeling

The choice of featurization strategy for chemical reactions is crucial for building predictive models for reaction conditions. Compared to the extensive research on molecular representation learning, the development of reaction encoding methods is relatively less explored [93]. Most of the existing methods were originally designed for predicting reaction properties (such as activation energy, reaction enthalpy, etc.) or classifying reactions, but they can be potentially adapted for reaction conditions prediction by modifying the output layer of the model. Both global reaction conditions prediction and local reaction optimization, which use the structures of reactants and products as inputs to predict their corresponding targets, require suitable choices of reaction featurization. The common methods can be categorized into three types: (1) descriptor-based, (2) graph-based, and (3) text-based featurization, as shown in Figure 2. Descriptor-based methods are often used for datasets with limited samples, since they incorporate chemistry- or physics-informed features that can enhance the model's ability to fit the data. Graph-based and text-based methods rely on deep-learning architectures that can learn latent patterns from the reactants and products, but they require sufficient data to train both the feature extractor and the neural network. These methods also reduce the need for manual feature selection by chemists.

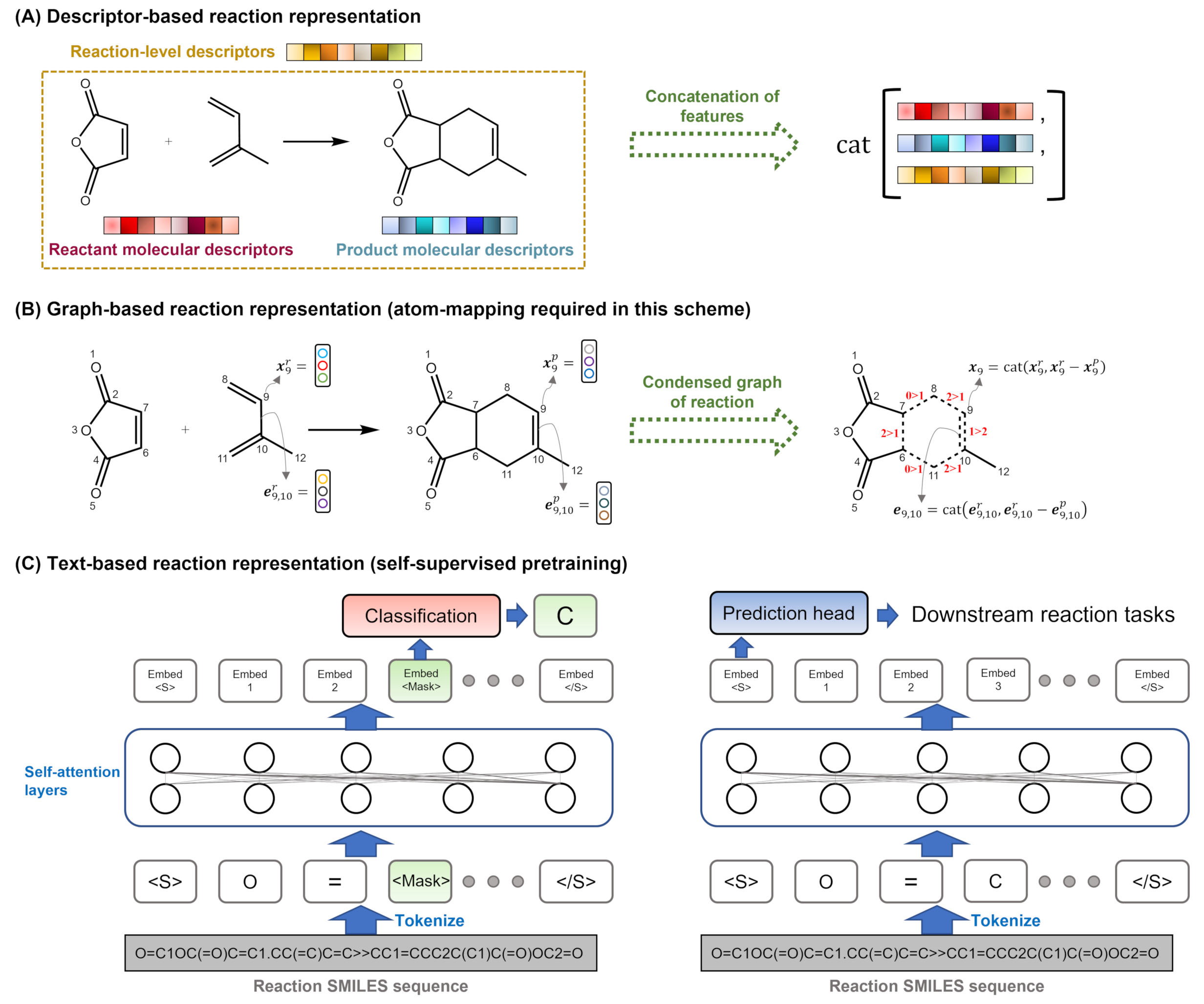

(A) Descriptor-based reaction representation
Reaction-level descriptors
Reactant molecular descriptors
Product molecular descriptors
Concatenation of features
cat
(B) Graph-based reaction representation (atom-mapping required in this scheme)
$x_9^r =$
$e_{9,10}^r =$
$x_9^p =$
$e_{9,10}^p =$
Condensed graph of reaction
0>1
2>1
1>2
$x_9 = \mathrm{cat}(x_9^r, x_9^r - x_9^p)$
$e_{9,10} = \mathrm{cat}(e_{9,10}^r, e_{9,10}^r - e_{9,10}^p)$
(C) Text-based reaction representation (self-supervised pretraining)
Classification
C
Embed <S>
Embed 1
Embed 2
Embed <Mask>
Embed </S>
Self-attention layers
<S>
O
=
<Mask>
</S>
Tokenize
O=C1OC(=O)C=C1.CC(=C)C=C>>CC1=CCC2C(C1)C(=O)OC2=O
Reaction SMILES sequence
Prediction head
Downstream reaction tasks
Embed 3
C

Figure 2: A comparison of three types of reaction embedding methods: (A) descriptor-based, which use predefined features from reactants and products, (B) graph-based, which use neural networks to learn features from molecular graphs, and (C) text-based, which use natural language processing to learn features from reaction SMILES. These methods vary in their computational efficiency, data requirements, and feature interpretability.

Descriptor-based representation

Descriptor-based methods are often used for datasets with limited samples, since they incorporate features that are informed by chemistry or physics and that can enhance the model's ability to fit the data [94]. Molecular-level descriptors of reactants and products are concatenated to obtain reaction-level descriptors, which can be computed by various methods [95]. These include substructure keys-based [96-100], circular [101-103], physicochemical [104-107], and quantum mechanical (QM) features [108-112]. The choice of descriptors depends on the size and scope of the dataset. For large-scale global models, descriptors with longer feature lengths and higher computational efficiency, such as the first four methods, are preferred. However, for small-scale local models, QM features can offer more compact and accurate information, but they require sampling and optimizing the 3D conformers of molecules using density functional theory (DFT) calculations, which are computationally expensive and time-consuming [62]. To overcome this challenge, some studies have proposed to pre-generate QM properties datasets and train ML models to serve as fast feature generators for new molecules [16]. However, this approach requires careful validation of the training data coverage and the extrapolation ability of the surrogate models.

Reaction-level descriptors based on DFT calculations of the TS structures of chemical reactions can provide valuable insights for predicting rate constants [113-117], regioselectivity, and site-selectivity [16,17,118-120]. However, this approach is also computationally demanding and requires a good initial guess of the TS structure. Moreover, it may face difficulties in simulating some classes of reactions and large-size molecules [121], and the solvent effects may complicate the results [122]. Therefore, reaction-level DFT-based descriptors are not widely used for reaction featurization. A more popular alternative is the differential reaction fingerprint (DRFP) developed by Probst et al. [123], which converts a reaction SMILES sequence into a binary fingerprint by comparing the symmetric difference of two sets of circular molecular substructures. The DRFP fingerprint can be seen as the reaction version of the ECFP molecular fingerprint [103]. Due to its fast computation and compatibility with conventional ML models, it has been widely used or benchmarked in various reaction-related tasks [124-128], and has become one of the mainstream reaction-level featurization techniques.

Graph-based representation

Graph neural networks (GNNs) have been widely applied to various chemical tasks, such as predicting molecular properties [129-133], reaction product prediction [134-136], and inverse materials design [137-139]. Chemical molecules can be naturally represented as undirected graphs, where nodes and edges encode atomic and bond information, respectively. GNNs update and aggregate the hidden features of nodes and edges through recursive message passing and a readout function, resulting in a molecular representation. There are many variants of GNN models [140-143], most of which are based on the message passing neural network (MPNN) framework proposed by Gilmer et al. [144].

Encoding reactions as graph representations is more challenging than encoding molecular structures, as reactions involve multiple disconnected molecular graphs and complex interactions. Graph-based reaction representations can be divided into two categories: AAM-exempted and AAM-required methods. Atom-to-atom mapping (AAM) is a process that establishes the correspondence between atoms before and after a reaction, reflecting the reaction mechanism.

AAM-exempted methods [145-150] apply graph convolutions to each reactant and product molecule separately, and then use a pooling function or attention layers to obtain a reaction fingerprint. These methods are scalable and compatible with conventional GNN models, requiring minimal modifications. AAM-required methods [151-153] assign labels to each atom and adapt the algorithms accordingly. Grambow et al. [151] and Yarish et al. [153] both subtract the hidden node vectors of the reactants from those of the products, and use the resulting differential atomic fingerprints to generate reaction representations. Heid et al. [152] developed a more general AAM-required reaction encoding method that operates graph convolutions on the condensed graph of reaction (CGR) [154,155]. The CGR is the superposition of reactant and product graphs, where nodes and edges can incorporate features from both sides of the reaction, as shown in Figure 2B. This method can also handle imbalanced reactions by imputing or zeroing the missing nodes.

The AAM procedure can provide valuable chemical insights into graph-based reaction encoding, as it reveals how the reaction center atoms influence the bond breaking and formation. However, obtaining accurate AAM for reactions can be difficult and depends on the complexity of the reaction types, as shown by Lin et al. [156]. Moreover, it is unclear whether AAM significantly improves the accuracy of reaction modeling. The AAM-required methods are usually tested on specific reaction types, where the reaction transformations and AAM are clear and correct. However, most large-scale reaction datasets do not have AAM information, and thus require the use of high-accuracy and automated AAM tools [84-87]. These tools may still introduce errors and affect the prediction of new reactions. Therefore, although GNN models are popular and successful for tasks at the molecule level, their effectiveness in reaction-level applications can still be enhanced.

Text-based representation

Recent years have witnessed the emergence of large language models (LLMs) [157-159], such as ChatGPT, that learn the statistical and semantic patterns of language through extensive self-supervised training. These models have broad applicability and robust learning capabilities, and thus have attracted the interest of the chemistry domain to tackle relevant problems. One common way to represent chemical molecular structures in chemical databases is the SMILES notation [160], which is a text-based expression with specific grammar rules and can be tokenized as input for language models.

Many studies have adopted the BERT model architecture and the masked language modeling (MLM) method to pretrain on millions of molecular SMILES and finetune on small-sample molecular property datasets [161-164]. For reaction-level prediction tasks, the textual input for pretraining can be changed to reaction SMILES, as shown in Figure 2C. Schwaller et al. [165] first demonstrated this idea and showed that pretraining in this way significantly improved reaction classification accuracy and could automatically generate AAM for reactants and products by analyzing the attention weights of each token in the reaction sequence.

The key to effective language modeling and its powerful reasoning abilities is the size of the pretraining data [166]. However, unlike molecular SMILES, which can be generated from existing databases (e.g., GDB-13 [167]) or by methods that produce reasonable structures [168], reaction SMILES data are often limited by the availability of experimental databases. Therefore, various data augmentation methods [169-171] have been proposed to increase the data size. These methods mainly involve changing the order of SMILES without affecting their molecular structures or modifying specific functional groups in coupling reactions with chemistry-informed reaction templates. Despite the need for large amounts of data to train base models, the main advantage of text-based reaction representation is that it can be easily applied to different downstream tasks by finetuning on small-sample data [172,173], without the need for tedious chemistry-informed feature generation and selection beforehand.

Reaction conditions design

In this section, we discuss the practical applications of different methods for featurizing reactions in predicting and optimizing reaction conditions. The design of reaction conditions depends on the availability of data and the specific application scenario. For example, if the aim is to predict the reaction conditions for each step in a synthesis pathway as part of an ML-aided CASP system, global models that can handle diverse reactions need to be built using large-scale reaction datasets. These models can then provide a range of general reaction conditions for chemists to select from. Alternatively, if the aim is to optimize the yield and selectivity of a specific reaction, more fine-grained variations of reaction conditions need to be explored. For this purpose, local models that are tailored for specific reaction families need to be trained to provide more focused guidance.

Global models for direct reaction conditions predictions

A common approach for chemists to develop novel reactions is to reference similar chemical reactions using reaction similarity search [174,175] and adopt the reaction conditions used in the literature. ML can leverage the large-scale reaction databases to build global models that can predict reaction conditions for diverse and novel chemical reactions, providing initial guidance for chemists.

Most of the existing research on global reaction conditions models involves predicting the reagents used in the dataset as labels, along with the reaction temperatures, using multi-class or multi-label classification methods [176]. This is a convenient way to represent the prediction targets, as some additives, such as molecular sieves and zeolites, cannot be represented by SMILES notation. However, the labels in the datasets may have some inconsistencies, such as different names for the same chemical, which may affect the learning and performance of the models. Therefore, a preprocessing step to standardize the labels and reduce redundancy is also essential.

Gao et al. [18] developed a large-scale model for predicting reaction conditions, using a deep learning approach trained on the Reaxys database. Their model could sequentially predict the catalysts, solvents, and reagents for a given reaction. This approach demonstrated the model's ability to handle complex and diverse datasets. However, the model assumed that each reaction had a single optimal set of conditions, ignoring the fact that some reactions might have multiple viable alternatives. This limitation reduced the diversity of options available for experimentalists. Subsequent studies have attempted to overcome this challenge by proposing different solutions. Kwon et al. [145] used a variational autoencoder (VAE) architecture to sample different reaction conditions, while Chen et al. [42] designed a two-stage recommendation system that predicted and ranked various reaction conditions based on the reaction yields. These methods enabled the prediction of a range of reaction conditions, allowing experimentalists to choose their preferred ones. However, building such a model is difficult, as most reaction databases, such as Reaxys, only record the highest-yield reaction conditions from a single publication. Therefore, the data might lack diversity in reaction conditions for a given reaction, unless the same reaction appears in multiple publications with different conditions.

A variety of ML approaches have been applied to the prediction of reaction conditions, including descriptor-, graph-, and text-based methods, as summarized in Table 3. However, these studies use different reaction datasets to evaluate their models, making it difficult to compare their accuracy objectively. A more standardized and open-source way of storing and accessing chemical reaction data, such as the ORD [36,177] or the curated USPTO dataset [35], would facilitate the benchmarking of models in predicting reaction conditions. Moreover, ML models may not always learn to predict meaningful reaction conditions; they may simply memorize the most frequently reported solvents and reagents in the literature. Beker et al. [178] showed that some machine learning models could not outperform simple statistical analyses based on the popularity of reported conditions in the literature, using the Suzuki–Miyaura coupling as an example. Therefore, to assess the predictive capabilities of models more rigorously, popularity-based baselines should be used as a reference.

The choice of reaction conditions is crucial for CASP applications, as it affects the cost, yield, and environmental impact of the synthetic route [4,182]. Moreover, predicting reaction conditions can help optimize the synthetic route [183] by providing the necessary information for each synthetic step. Coley et al. [12] integrated ASKCOS [184], an automated CASP software, with the self-driving lab [185] and demonstrated the synthesis of 15 small molecules. Guo et al. [186] used a synthesis strategy that combines Monte Carlo Tree Search (MCTS) with reinforcement learning to model the retrosynthesis game, aiming to identify high-value synthetic pathways. Recently, Koscher et al. [21] have shown the simultaneous design and synthesis of dye molecules through design–make–test–analyze (DMTA) cycles [187]. Given the limited experimental throughput, it is important to prioritize the molecular properties that are predicted to be superior, along with their synthesis costs, during the chemical experiments. The reaction conditions prediction model plays a vital role in this context; it filters out inaccessible and incompatible conditions, such as high-temperature reactions, high-reactive gases, insoluble solid reagents, and environmentally unfriendly reagents.

The examples above illustrate the usefulness of global reaction conditions prediction models, which use historical literature on similar chemical contexts to suggest suitable reaction conditions for synthetic steps. However, the predictive output often

Table 3: Representative works on predicting globally reaction conditions. The references are sorted chronologically.

Reference	Data	Model type	Description
[18]	≈10 million general reactions from Reaxys	ECFP + DNN	the model has the most access to proprietary training data
[179]	4 types of totally ≈191k reactions from Reaxys	descriptors + GBM and GCNs	the output labels were systematically categorized with chemical insights
[70]	≈693k reactions from Pistachio	nearest-neighbor, transformer and BART	the work demonstrates the first utilization of NLP models to generate the step-by-step experimental procedures
[180]	≈6k Buchwald–Hartwig coupling reactions from in-house lab notebooks	ECFP + DNN	it showed that multi-label predictions are more advantageous than single-label predictions
[145]	4 types of totally ≈191k reactions from Reaxys	GNN + VAE	the models provide multiple reaction conditions by repeatedly sampling from the VAE space
[82]	480k USPTO-MIT dataset [134]	reaction SMILES + transformer	this work directly predicts SMILES representation of the combination of reaction conditions
[35]	curated USPTO-condition dataset with ≈680k reactions and Reaxys-TotalSyn-Condition dataset with ≈180k reactions	reaction SMILES + transformer	this work demonstrates the benefits of MLM pretraining for the downstream reaction conditions prediction task
[42]	10 types of totally ≈74k reactions from Reaxys	ECFP + DNN	it models the reaction conditions prediction problem as recommendation system and artificially generate fake reaction conditions for data augmentation
[181]	curated USPTO-Condition dataset with ≈680k reactions	SMILES-to-text retriever and text-augmented predictor	the two-stage model first uses multimodal retrieval to obtain related chemistry literature and then combines it with reaction input to predict reaction conditions

lacks fine-grained details such as reaction time, pressure, and pH values. These details depend on the problem formulation specific to each individual synthetic step. To further improve yields, it is necessary to perform local reaction optimization, which is discussed below.

Local reaction optimization

ML-guided local reaction optimization, or self-optimization, is an automated and generalizable approach that can accelerate the discovery of optimal reaction conditions, as illustrated in Figure 3. The first step is problem formulation, which involves defining the reaction parameters to be optimized and the target objectives, such as yield and selectivity. The reaction parameters include categorical variables, such as catalysts, solvents, and acid–base salts, and continuous variables, such as temperature, pressure, substrate concentration, and residence time. Regression prediction models are then built for these reaction parameters and target objectives by collecting experimental data and conducting statistical analysis.

Many reaction optimization platforms have been developed [188-191], which integrate software optimization algorithms with hardware automation for experiments, enabling large-scale experimentation and data collection. Among these, BO [192] is the most classic and widely used algorithm, which leverages kernel density estimators to efficiently explore parameter space. This method updates prior probability distributions with new experimental results and optimizes the reaction conditions by focusing on regions of the parameter space predicted to improve objectives. The power of BO lies in its ability to balance exploration and exploitation, making it highly effective for complex, multidimensional optimization tasks in chemical processes. BO has also demonstrated robust performance in many benchmark tasks [193-195], and numerous chemical reaction optimization packages have been developed to support this algorithm [196-200].

A typical example is the work by Shields et al. [62], who used different featurization strategies, such as DFT [108], cheminformatics [107], and binary one-hot-encoded, in conjunction with the BO algorithm to optimize reaction conditions. Their experimental results showed that DFT features could train probabilistic surrogate models more effectively and that the optimization efficiency was superior to manual adjustments made by professional chemists. They also applied this approach to the Mitsunobu reaction and deoxyfluorination reaction, rapidly identifying medium to high-yield results from approximately 100,000 experimental conditions using fewer than 100 experiments.

Moving from individual synthetic steps to CASP, Nambiar et al. [201] investigated the impact of integrating a global reaction conditions prediction model with local reaction optimization on enhancing the overall chemical synthesis pathway. They demonstrated the predictive pathway for sonidegib synthesis, but it still required chemical insights to verify the compatibility of the solvents predicted by the global model with the reactants. Moreover, in a multistep synthesis route, the interdependencies between different reaction sequences, such as additional separation and purification steps, could reduce the overall yield [202].

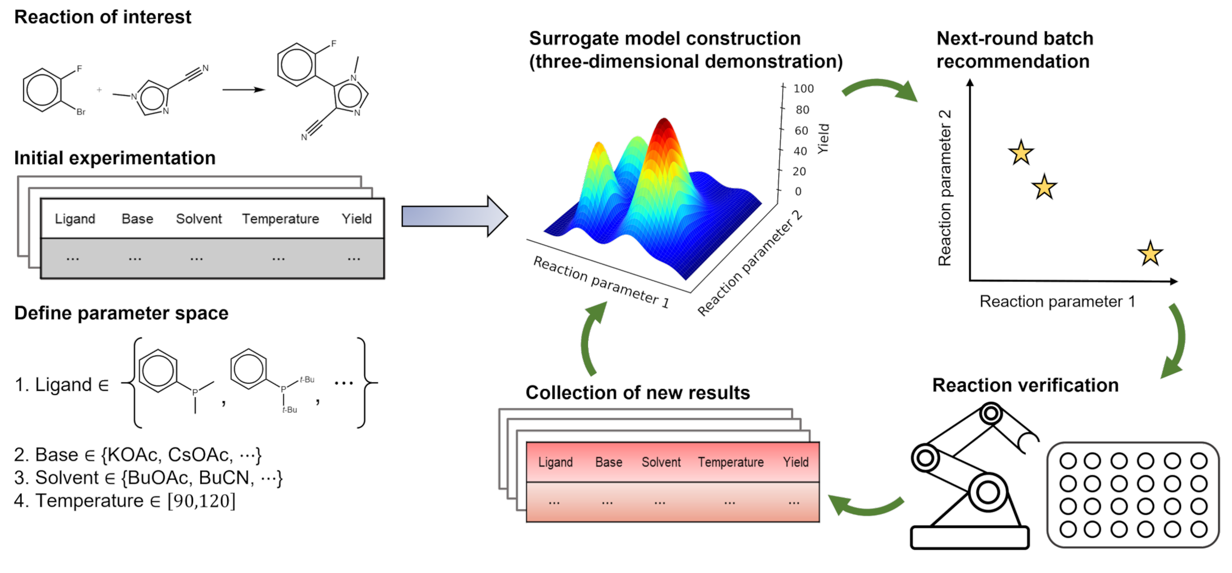


Figure 3: A schematic diagram of how ML algorithms can be combined with HTE platforms to optimize reaction conditions for CASP.

This indicates that the suboptimal combination of each reaction does not necessarily represent the global optimum for multistep synthesis [203-205]. In contrast, telescoped flow sequences [206-208] or one-pot batch synthesis [209] emphasize the use of chemically compatible reagents and solvents in each reaction step to minimize intermediate purification steps. Volk et al. [210] developed AlphaFlow, which utilizes reinforcement learning as an optimization algorithm for the shell growth of core-shell semiconductor nanoparticles. This involves various unit operations such as phase separation, washing, and continuous in-situ spectral monitoring. Although the process conditions for this reaction system do not have as extensive a literature base for training data, this study was still able to identify better solutions than conventional designs through reinforcement learning in multistep processes.

Besides maximizing the reaction yield for a given reaction with given substrates, another goal of reaction optimization is to discover general reaction conditions that are applicable to various substrates within the same reaction type [211-215]. For instance, the generality of chiral catalysts for asymmetric or enantioselective catalysis has been a longstanding interest in synthetic chemistry [216]. Angello et al. [53] applied uncertainty-minimizing ML and automated robotic experimentation to accelerate the exploration of general reaction conditions for heteroaryl Suzuki–Miyaura cross-coupling. They achieved an average yield that was twice as high as that of previous human-guided experiments. Recently, Wang et al. [65] formulated the optimization of general reaction conditions as a multi-armed bandit problem, where each set of reaction conditions is a slot machine, and each experiment is a round of playing on one of these machines. The challenge is to find the slot machine with the highest win rate using a limited number of rounds. For chemical experiments, this entails a strategic balance between exploring new reaction conditions (or 'slot machines') and exploiting known conditions that deliver high yields. Therefore, they proposed a more efficient sampling strategy based on reinforcement learning to dynamically adjust the selection process, thereby optimizing the exploration–exploitation trade-off.

The preceding examples demonstrate how the combination of HTE chemistry tools and optimization algorithms has significantly advanced the field of reaction optimization. However, this protocol also has some limitations, especially regarding the suitability of the chemical system under investigation. First, in terms of hardware implementation, setting up an HTE platform with robotic technologies entails high financial costs and specialized knowledge for installation, which may not be accessible for smaller-scale or less-funded research entities [217]. Moreover, to enable experimentation with various reaction conditions, a large chemical storage capacity is necessary. Otherwise, the scope of research would be confined to only a few types of chemical reactions [21]. Additionally, to ensure experimental safety, chemists must rigorously verify the compatibility of each solvent and reagent combination used in reactions and eliminate any potential hazards [218]. Second, in terms of algorithmic approaches, the widely used BO requires initial data to build a probabilistic surrogate model. Although the data might be sourced from related literature, caution is advised as experimental apparatus from different sources could introduce systematic errors in reported yields [46]. Furthermore, BO cannot generalize well from past reactions to unseen reaction transformations, which inherently requires gathering new relevant data for new chemical reactions [219]. Regarding general reaction conditions, the typically limited experimental budgets in laboratories restrict the ability to explore a diverse range of reaction conditions [65]. Thus, initial filtering by chemists, which removes known impractical conditions, is essential. Despite these existing challenges, reaction optimization continues to play a vital role in both academia and industry in the age of big data [23].

Outlook and Perspectives

As we explore the future of ML in designing and optimizing reaction conditions, several promising avenues and challenges are poised to shape this interdisciplinary field. The integration of HTE with ML is revolutionizing how chemists approach reaction conditions. Future efforts should aim to enhance these technologies to enable faster and more comprehensive data collection, potentially leading to the automation of HTE and ML integration into real-time adaptive systems that learn from each experiment.

The discussions in this review about global and local models underline the critical need for large, comprehensive, and coherent datasets. Advancements in data processing and model training methodologies, such as transfer learning and reinforcement learning, are essential to boost the predictive power and efficiency of these models. Platforms like the ORD are crucial in meeting the demand for accessible and standardized chemical data. The expansion of such platforms and fostering wider community involvement will be key to advancing data-driven approaches in chemistry. A community dedicated to openly sharing data and findings will likely accelerate innovation and enhance the robustness of ML tools.

Moreover, computational models that integrate theoretical chemistry and ML could unlock deeper insights into poorly understood or complex reaction mechanisms. These models are particularly valuable in areas where experimental data are sparse or challenging to obtain, thereby extending the range of ML applications in chemistry.

Educating the next generation of chemists, engineers, and data scientists in both ML and chemical synthesis is critical. Interdisciplinary programs can develop a workforce skilled in applying AI to complex chemical issues, fostering more innovative and efficient solutions. Enhancing international cooperation can standardize data collection and sharing practices, simplifying the process of building and validating models across various laboratories and contexts. Such global collaboration is instrumental in addressing widespread challenges like climate change and sustainability through smarter chemical processes.

By focusing on these directions, we anticipate a future where ML not only supports but significantly propels the field of synthetic chemistry forward, making it more innovative, efficient, and sustainable. The ongoing development of ML in reaction conditions design and optimization holds the promise of unlocking new capabilities and achieving transformative breakthroughs in the field.

Acknowledgements

We are grateful to the National Center for High-performance Computing (NCHC) and the Computer and Information Networking Center at NTU for the support of computing facilities. During the preparation of this work, the authors used AI tools in order to correct grammatical mistakes and enhance the fluency of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding

Y.P.L. is supported by Taiwan NSTC 2030 Cross-Generation Young Scholars Program (113-2628-E-002-017-MY3) and the Higher Education Sprout Project by the Ministry of Education in Taiwan (113L891305).

ORCID® iDs

Lung-Yi Chen - https://orcid.org/0000-0002-9411-6404
Yi-Pei Li - https://orcid.org/0000-0002-1314-3276

Data Availability Statement

Data sharing is not applicable as no new data was generated or analyzed in this study.

Preprint

A non-peer-reviewed version of this article has been previously published as a preprint: doi:10.26434/chemrxiv-2024-wt75q

References

1. Chen, S.; Jung, Y. *JACS Au* **2021,** *1,* 1612–1620. doi:10.1021/jacsau.1c00246
2. Finnigan, W.; Hepworth, L. J.; Flitsch, S. L.; Turner, N. J. *Nat. Catal.* **2021,** *4,* 98–104. doi:10.1038/s41929-020-00556-z
3. Fortunato, M. E.; Coley, C. W.; Barnes, B. C.; Jensen, K. F. *J. Chem. Inf. Model.* **2020,** *60,* 3398–3407. doi:10.1021/acs.jcim.0c00403
4. Thakkar, A.; Johansson, S.; Jorner, K.; Buttar, D.; Reymond, J.-L.; Engkvist, O. *React. Chem. Eng.* **2021,** *6,* 27–51. doi:10.1039/d0re00340a
5. Klucznik, T.; Mikulak-Klucznik, B.; McCormack, M. P.; Lima, H.; Szymkuć, S.; Bhowmick, M.; Molga, K.; Zhou, Y.; Rickershauser, L.; Gajewska, E. P.; Toutchkine, A.; Dittwald, P.; Startek, M. P.; Kirkovits, G. J.; Roszak, R.; Adamski, A.; Sieredzińska, B.; Mrksich, M.; Trice, S. L. J.; Grzybowski, B. A. *Chem* **2018,** *4,* 522–532. doi:10.1016/j.chempr.2018.02.002
6. Mo, Y.; Guan, Y.; Verma, P.; Guo, J.; Fortunato, M. E.; Lu, Z.; Coley, C. W.; Jensen, K. F. *Chem. Sci.* **2021,** *12,* 1469–1478. doi:10.1039/d0sc05078d
7. Molga, K.; Dittwald, P.; Grzybowski, B. A. *Chem* **2019,** *5,* 460–473. doi:10.1016/j.chempr.2018.12.004
8. Sankaranarayanan, K.; Jensen, K. F. *Chem. Sci.* **2023,** *14,* 6467–6475. doi:10.1039/d3sc01355c
9. Ha, T.; Lee, D.; Kwon, Y.; Park, M. S.; Lee, S.; Jang, J.; Choi, B.; Jeon, H.; Kim, J.; Choi, H.; Seo, H.-T.; Choi, W.; Hong, W.; Park, Y. J.; Jang, J.; Cho, J.; Kim, B.; Kwon, H.; Kim, G.; Oh, W. S.; Kim, J. W.; Choi, J.; Min, M.; Jeon, A.; Jung, Y.; Kim, E.; Lee, H.; Choi, Y.-S. *Sci. Adv.* **2023,** *9,* eadj0461. doi:10.1126/sciadv.adj0461
10. Hardwick, T.; Ahmed, N. *Chem. Sci.* **2020,** *11,* 11973–11988. doi:10.1039/d0sc04250a
11. Shen, Y.; Borowski, J. E.; Hardy, M. A.; Sarpong, R.; Doyle, A. G.; Cernak, T. *Nat. Rev. Methods Primers* **2021,** *1,* 23. doi:10.1038/s43586-021-00022-5
12. Coley, C. W.; Thomas, D. A., III; Lummiss, J. A. M.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H.; Hicklin, R. W.; Plehiers, P. P.; Byington, J.; Piotti, J. S.; Green, W. H.; Hart, A. J.; Jamison, T. F.; Jensen, K. F. *Science* **2019,** *365,* eaax1566. doi:10.1126/science.aax1566
13. Taylor, C. J.; Pomberger, A.; Felton, K. C.; Grainger, R.; Barecka, M.; Chamberlain, T. W.; Bourne, R. A.; Johnson, C. N.; Lapkin, A. A. *Chem. Rev.* **2023,** *123,* 3089–3126. doi:10.1021/acs.chemrev.2c00798
14. Voinarovska, V.; Kabeshov, M.; Dudenko, D.; Genheden, S.; Tetko, I. V. *J. Chem. Inf. Model.* **2024,** *64,* 42–56. doi:10.1021/acs.jcim.3c01524
15. Żurański, A. M.; Martinez Alvarado, J. I.; Shields, B. J.; Doyle, A. G. *Acc. Chem. Res.* **2021,** *54,* 1856–1865. doi:10.1021/acs.accounts.0c00770
16. Guan, Y.; Coley, C. W.; Wu, H.; Ranasinghe, D.; Heid, E.; Struble, T. J.; Pattanaik, L.; Green, W. H.; Jensen, K. F. *Chem. Sci.* **2021,** *12,* 2198–2208. doi:10.1039/d0sc04823b
17. Struble, T. J.; Coley, C. W.; Jensen, K. F. *React. Chem. Eng.* **2020,** *5,* 896–902. doi:10.1039/d0re00071j
18. Gao, H.; Struble, T. J.; Coley, C. W.; Wang, Y.; Green, W. H.; Jensen, K. F. *ACS Cent. Sci.* **2018,** *4,* 1465–1476. doi:10.1021/acscentsci.8b00357
19. Abolhasani, M.; Kumacheva, E. *Nat. Synth.* **2023,** *2,* 483–492. doi:10.1038/s44160-022-00231-0
20. Raghavan, P.; Haas, B. C.; Ruos, M. E.; Schleinitz, J.; Doyle, A. G.; Reisman, S. E.; Sigman, M. S.; Coley, C. W. *ACS Cent. Sci.* **2023,** *9,* 2196–2204. doi:10.1021/acscentsci.3c01163

21. Koscher, B. A.; Canty, R. B.; McDonald, M. A.; Greenman, K. P.; McGill, C. J.; Bilodeau, C. L.; Jin, W.; Wu, H.; Vermeire, F. H.; Jin, B.; Hart, T.; Kulesza, T.; Li, S.-C.; Jaakkola, T. S.; Barzilay, R.; Gómez-Bombarelli, R.; Green, W. H.; Jensen, K. F. *Science* **2023,** *382,* eadi1407. doi:10.1126/science.adi1407
22. Eyke, N. S.; Koscher, B. A.; Jensen, K. F. *Trends Chem.* **2021,** *3,* 120–132. doi:10.1016/j.trechm.2020.12.001
23. Krska, S. W.; DiRocco, D. A.; Dreher, S. D.; Shevlin, M. *Acc. Chem. Res.* **2017,** *50,* 2976–2985. doi:10.1021/acs.accounts.7b00428
24. Shevlin, M. *ACS Med. Chem. Lett.* **2017,** *8,* 601–607. doi:10.1021/acsmedchemlett.7b00165
25. Snoek, J.; Larochelle, H.; Adams, R. P. Practical Bayesian Optimization of Machine Learning Algorithms. In *Advances in Neural Information Processing Systems 25,* 26th annual conference on neural information processing systems 2012, Lake Tahoe, NV, USA, Dec 3–6, 2012; Pereira, F.; Burges, C. J.; Bottou, L.; Weinberger, K. Q., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2013.
26. Andronov, M.; Fedorov, M. V.; Sosnin, S. *ACS Omega* **2021,** *6,* 30743–30751. doi:10.1021/acsomega.1c04778
27. Dobson, C. M. *Nature* **2004,** *432,* 824–828. doi:10.1038/nature03192
28. Cheng, G.-J.; Zhang, X.; Chung, L. W.; Xu, L.; Wu, Y.-D. *J. Am. Chem. Soc.* **2015,** *137,* 1706–1725. doi:10.1021/ja5112749
29. Vogiatzis, K. D.; Polynski, M. V.; Kirkland, J. K.; Townsend, J.; Hashemi, A.; Liu, C.; Pidko, E. A. *Chem. Rev.* **2019,** *119,* 2453–2523. doi:10.1021/acs.chemrev.8b00361
30. Jorner, K.; Tomberg, A.; Bauer, C.; Sköld, C.; Norrby, P.-O. *Nat. Rev. Chem.* **2021,** *5,* 240–255. doi:10.1038/s41570-021-00260-x
31. Plata, R. E.; Singleton, D. A. *J. Am. Chem. Soc.* **2015,** *137,* 3811–3826. doi:10.1021/ja5111392
32. Mata, R. A.; Suhm, M. A. *Angew. Chem., Int. Ed.* **2017,** *56,* 11011–11018. doi:10.1002/anie.201611308
33. Thakkar, A.; Kogej, T.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. J. *Chem. Sci.* **2020,** *11,* 154–168. doi:10.1039/c9sc04944d
34. Reaxys. https://www.reaxys.com (accessed March 25, 2024).
35. Wang, X.; Hsieh, C.-Y.; Yin, X.; Wang, J.; Li, Y.; Deng, Y.; Jiang, D.; Wu, Z.; Du, H.; Chen, H.; Li, Y.; Liu, H.; Wang, Y.; Luo, P.; Hou, T.; Yao, X. *Research (Washington, DC, U. S.)* **2023,** *6,* 0231. doi:10.34133/research.0231
36. Kearnes, S. M.; Maser, M. R.; Wleklinski, M.; Kast, A.; Doyle, A. G.; Dreher, S. D.; Hawkins, J. M.; Jensen, K. F.; Coley, C. W. *J. Am. Chem. Soc.* **2021,** *143,* 18820–18826. doi:10.1021/jacs.1c09820
37. Lowe, D. Chemical reactions from US patents (1976-Sep2016). https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873 (accessed March 25, 2024).
38. CAS, SciFinder. https://scifinder-n.cas.org (accessed March 25, 2024).
39. Nextmove Software Pistachio. https://www.nextmovesoftware.com/pistachio.html (accessed March 25, 2024).
40. Roth, D. L. *J. Chem. Inf. Model.* **2005,** *45,* 1470–1473. doi:10.1021/ci050274b
41. Mahjour, B.; Shen, Y.; Cernak, T. *Acc. Chem. Res.* **2021,** *54,* 2337–2346. doi:10.1021/acs.accounts.1c00119
42. Chen, L.-Y.; Li, Y.-P. *J. Cheminf.* **2024,** *16,* 11. doi:10.1186/s13321-024-00805-4
43. StriethKalthoff, F.; Sandfort, F.; Kühnemund, M.; Schäfer, F. R.; Kuchen, H.; Glorius, F. *Angew. Chem., Int. Ed.* **2022,** *61,* e202204647. doi:10.1002/anie.202204647
44. Herres-Pawlis, S.; Bach, F.; Bruno, I. J.; Chalk, S. J.; Jung, N.; Liermann, J. C.; McEwen, L. R.; Neumann, S.; Steinbeck, C.; Razum, M.; Koepler, O. *Angew. Chem., Int. Ed.* **2022,** *61,* e202203038. doi:10.1002/anie.202203038
45. Hunter, A. M.; Carreira, E. M.; Miller, S. J. *J. Org. Chem.* **2020,** *85,* 1773–1774. doi:10.1021/acs.joc.0c00248
46. Maloney, M. P.; Coley, C. W.; Genheden, S.; Carson, N.; Helquist, P.; Norrby, P.-O.; Wiest, O. *Org. Lett.* **2023,** *25,* 2945–2947. doi:10.1021/acs.orglett.3c01282
47. Mercado, R.; Kearnes, S. M.; Coley, C. W. *J. Chem. Inf. Model.* **2023,** *63,* 4253–4265. doi:10.1021/acs.jcim.3c00607
48. Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. *Science* **2018,** *360,* 186–190. doi:10.1126/science.aar5169
49. Buitrago Santanilla, A.; Regalado, E. L.; Pereira, T.; Shevlin, M.; Bateman, K.; Campeau, L.-C.; Schneeweis, J.; Berritt, S.; Shi, Z.-C.; Nantermet, P.; Liu, Y.; Helmy, R.; Welch, C. J.; Vachal, P.; Davies, I. W.; Cernak, T.; Dreher, S. D. *Science* **2015,** *347,* 49–53. doi:10.1126/science.1259203
50. Saebi, M.; Nan, B.; Herr, J. E.; Wahlers, J.; Guo, Z.; Zurański, A. M.; Kogej, T.; Norrby, P.-O.; Doyle, A. G.; Chawla, N. V.; Wiest, O. *Chem. Sci.* **2023,** *14,* 4997–5005. doi:10.1039/d2sc06041h
51. Perera, D.; Tucker, J. W.; Brahmbhatt, S.; Helal, C. J.; Chong, A.; Farrell, W.; Richardson, P.; Sach, N. W. *Science* **2018,** *359,* 429–434. doi:10.1126/science.aap9112
52. Reizman, B. J.; Wang, Y.-M.; Buchwald, S. L.; Jensen, K. F. *React. Chem. Eng.* **2016,** *1,* 658–666. doi:10.1039/c6re00153j
53. Angello, N. H.; Rathore, V.; Beker, W.; Wołos, A.; Jira, E. R.; Roszak, R.; Wu, T. C.; Schroeder, C. M.; Aspuru-Guzik, A.; Grzybowski, B. A.; Burke, M. D. *Science* **2022,** *378,* 399–405. doi:10.1126/science.adc8743
54. DeLano, T. J.; Reisman, S. E. *ACS Catal.* **2019,** *9,* 6751–6754. doi:10.1021/acscatal.9b01785
55. Isbrandt, E. S.; Chapple, D. E.; Tu, N. T. P.; Dimakos, V.; Beardall, A. M. M.; Boyle, P. D.; Rowley, C. N.; Blacquiere, J. M.; Newman, S. G. *J. Am. Chem. Soc.* **2024,** *146,* 5650–5660. doi:10.1021/jacs.3c14612
56. Zuo, Z.; Ahneman, D. T.; Chu, L.; Terrett, J. A.; Doyle, A. G.; MacMillan, D. W. C. *Science* **2014,** *345,* 437–440. doi:10.1126/science.1255525
57. Stadler, A.; Kappe, C. O. *J. Comb. Chem.* **2001,** *3,* 624–630. doi:10.1021/cc010044j
58. Nielsen, M. K.; Ahneman, D. T.; Riera, O.; Doyle, A. G. *J. Am. Chem. Soc.* **2018,** *140,* 5004–5008. doi:10.1021/jacs.8b01523
59. Kutchukian, P. S.; Dropinski, J. F.; Dykstra, K. D.; Li, B.; DiRocco, D. A.; Streckfuss, E. C.; Campeau, L.-C.; Cernak, T.; Vachal, P.; Davies, I. W.; Krska, S. W.; Dreher, S. D. *Chem. Sci.* **2016,** *7,* 2604–2613. doi:10.1039/c5sc04751j
60. Gioiello, A.; Rosatelli, E.; Teofrasti, M.; Filipponi, P.; Pellicciari, R. *ACS Comb. Sci.* **2013,** *15,* 235–239. doi:10.1021/co400012m
61. Newman-Stonebraker, S. H.; Smith, S. R.; Borowski, J. E.; Peters, E.; Gensch, T.; Johnson, H. C.; Sigman, M. S.; Doyle, A. G. *Science* **2021,** *374,* 301–308. doi:10.1126/science.abj4213
62. Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021,** *590,* 89–96. doi:10.1038/s41586-021-03213-y

63. Stevens, J. M.; Li, J.; Simmons, E. M.; Wisniewski, S. R.; DiSomma, S.; Fraunhoffer, K. J.; Geng, P.; Hao, B.; Jackson, E. W. *Organometallics* **2022,** *41,* 1847–1864. doi:10.1021/acs.organomet.2c00089
64. Mahjour, B.; Zhang, R.; Shen, Y.; McGrath, A.; Zhao, R.; Mohamed, O. G.; Lin, Y.; Zhang, Z.; Douthwaite, J. L.; Tripathi, A.; Cernak, T. *Nat. Commun.* **2023,** *14,* 3924. doi:10.1038/s41467-023-39531-0
65. Wang, J. Y.; Stevens, J. M.; Kariofillis, S. K.; Tom, M.-J.; Golden, D. L.; Li, J.; Tabora, J. E.; Parasram, M.; Shields, B. J.; Primer, D. N.; Hao, B.; Del Valle, D.; DiSomma, S.; Furman, A.; Zipp, G. G.; Melnikov, S.; Paulson, J.; Doyle, A. G. *Nature* **2024,** *626,* 1025–1033. doi:10.1038/s41586-024-07021-y
66. Schleinitz, J.; Langevin, M.; Smail, Y.; Wehnert, B.; Grimaud, L.; Vuilleumier, R. *J. Am. Chem. Soc.* **2022,** *144,* 14722–14730. doi:10.1021/jacs.2c05302
67. Xu, Y.; Ren, F.; Su, L.; Xiong, Z.; Zhu, X.; Lin, X.; Qiao, N.; Tian, H.; Tian, C.; Liao, K. *Org. Chem. Front.* **2023,** *10,* 1153–1159. doi:10.1039/d2qo01954j
68. Fitzner, M.; Wuitschik, G.; Koller, R.; Adam, J.-M.; Schindler, T. *ACS Omega* **2023,** *8,* 3017–3025. doi:10.1021/acsomega.2c05546
69. Vaucher, A. C.; Zipoli, F.; Geluykens, J.; Nair, V. H.; Schwaller, P.; Laino, T. *Nat. Commun.* **2020,** *11,* 3601. doi:10.1038/s41467-020-17266-6
70. Vaucher, A. C.; Schwaller, P.; Geluykens, J.; Nair, V. H.; Iuliano, A.; Laino, T. *Nat. Commun.* **2021,** *12,* 2573. doi:10.1038/s41467-021-22951-1
71. Guo, J.; Ibanez-Lopez, A. S.; Gao, H.; Quach, V.; Coley, C. W.; Jensen, K. F.; Barzilay, R. *J. Chem. Inf. Model.* **2022,** *62,* 2035–2045. doi:10.1021/acs.jcim.1c00284
72. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. *arXiv* **2018,** 1810.04805. doi:10.48550/arxiv.1810.04805
73. Qian, Y.; Guo, J.; Tu, Z.; Coley, C. W.; Barzilay, R. *J. Chem. Inf. Model.* **2023,** *63,* 4030–4041. doi:10.1021/acs.jcim.3c00439
74. Qian, Y.; Guo, J.; Tu, Z.; Li, Z.; Coley, C. W.; Barzilay, R. *J. Chem. Inf. Model.* **2023,** *63,* 1925–1934. doi:10.1021/acs.jcim.2c01480
75. McDaniel, J. R.; Balmuth, J. R. *J. Chem. Inf. Comput. Sci.* **1992,** *32,* 373–378. doi:10.1021/ci00008a018
76. Beard, E. J.; Cole, J. M. *J. Chem. Inf. Model.* **2020,** *60,* 2059–2072. doi:10.1021/acs.jcim.0c00042
77. Wilary, D. M.; Cole, J. M. *J. Chem. Inf. Model.* **2021,** *61,* 4962–4974. doi:10.1021/acs.jcim.1c01017
78. Huang, H.; Zheng, O.; Wang, D.; Yin, J.; Wang, Z.; Ding, S.; Yin, H.; Xu, C.; Yang, R.; Zheng, Q.; Shi, B. *Int. J. Oral Sci.* **2023,** *15,* 29. doi:10.1038/s41368-023-00239-y
79. Song, B.; Zhou, R.; Ahmed, F. *J. Comput. Inf. Sci. Eng.* **2024,** *24,* 010801. doi:10.1115/1.4063954
80. Wang, X.; Chen, G.; Qian, G.; Gao, P.; Wei, X.-Y.; Wang, Y.; Tian, Y.; Gao, W. *Mach. Intell. Res.* **2023,** *20,* 447–482. doi:10.1007/s11633-022-1410-8
81. Fan, V.; Qian, Y.; Wang, A.; Wang, A.; Coley, C. W.; Barzilay, R. *J. Chem. Inf. Model.* **2024,** *64,* 5521–5534. doi:10.1021/acs.jcim.4c00572
82. Andronov, M.; Voinarovska, V.; Andronova, N.; Wand, M.; Clevert, D.-A.; Schmidhuber, J. *Chem. Sci.* **2023,** *14,* 3235–3246. doi:10.1039/d2sc06798f
83. Toniato, A.; Schwaller, P.; Cardinale, A.; Geluykens, J.; Laino, T. *Nat. Mach. Intell.* **2021,** *3,* 485–494. doi:10.1038/s42256-021-00319-w
84. Chen, S.; An, S.; Babazade, R.; Jung, Y. *Nat. Commun.* **2024,** *15,* 2250. doi:10.1038/s41467-024-46364-y
85. Chen, W. L.; Chen, D. Z.; Taylor, K. T. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2013,** *3,* 560–593. doi:10.1002/wcms.1140
86. Nugmanov, R.; Dyubankova, N.; Gedich, A.; Wegner, J. K. *J. Chem. Inf. Model.* **2022,** *62,* 3307–3315. doi:10.1021/acs.jcim.2c00344
87. Schwaller, P.; Hoover, B.; Reymond, J.-L.; Strobelt, H.; Laino, T. *Sci. Adv.* **2021,** *7,* eabe4166. doi:10.1126/sciadv.abe4166
88. Nugmanov, R. I.; Mukhametgaleev, R. N.; Akhmetshin, T.; Gimadiev, T. R.; Afonina, V. A.; Madzhidov, T. I.; Varnek, A. *J. Chem. Inf. Model.* **2019,** *59,* 2516–2521. doi:10.1021/acs.jcim.9b00102
89. Zipoli, F.; Ayadi, Z.; Schwaller, P.; Laino, T.; Vaucher, A. C. *Mach. Learn.: Sci. Technol.* **2024,** *5,* 025071. doi:10.1088/2632-2153/ad5413
90. Zhang, C.; Arun, A.; Lapkin, A. A. *ACS Omega* **2024,** *9,* 18385–18399. doi:10.1021/acsomega.4c00262
91. Phan, T.-L.; Weinbauer, K.; Gärtner, T.; Merkle, D.; Andersen, J. L.; Fagerberg, R.; Stadler, P. F. *ChemRxiv* **2024**. doi:10.26434/chemrxiv-2024-hltm9
92. Chen, L.-Y.; Li, Y.-P. *J. Cheminf.* **2024,** *16,* 74. doi:10.1186/s13321-024-00869-2
93. Ding, Y.; Qiang, B.; Chen, Q.; Liu, Y.; Zhang, L.; Liu, Z. *J. Chem. Inf. Model.* **2024,** *64,* 2955–2970. doi:10.1021/acs.jcim.4c00004
94. Singh, A.; Thakur, N.; Sharma, A. A review of supervised machine learning algorithms. In *2016 3rd international conference on computing for sustainable global development (INDIACom),* New Delhi, India, March 16–18, 2016; IEEE: NJ, USA, 2016; pp 1310–1315.
95. Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. *Methods* **2015,** *71,* 58–63. doi:10.1016/j.ymeth.2014.08.005
96. Barnard, J. M.; Downs, G. M. *J. Chem. Inf. Comput. Sci.* **1997,** *37,* 141–142. doi:10.1021/ci960090k
97. Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. *Annu. Rep. Comput. Chem.* **2008,** *4,* 217–241. doi:10.1016/s1574-1400(08)00012-1
98. Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. *J. Chem. Inf. Comput. Sci.* **2002,** *42,* 1273–1280. doi:10.1021/ci010132r
99. Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. *J. Chem. Inf. Comput. Sci.* **1996,** *36,* 128–136. doi:10.1021/ci950275b
100. Tovar, A.; Eckert, H.; Bajorath, J. *ChemMedChem* **2007,** *2,* 208–217. doi:10.1002/cmdc.200600225
101. Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. *J. Chem. Inf. Comput. Sci.* **2004,** *44,* 170–178. doi:10.1021/ci034207y
102. Probst, D.; Reymond, J.-L. *J. Cheminf.* **2018,** *10,* 66. doi:10.1186/s13321-018-0321-8
103. Rogers, D.; Hahn, M. *J. Chem. Inf. Model.* **2010,** *50,* 742–754. doi:10.1021/ci100050t
104. Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. *J. Chem. Inf. Comput. Sci.* **1996,** *36,* 118–127. doi:10.1021/ci950274j
105. Raevsky, O. A. *Mini-Rev. Med. Chem.* **2004,** *4,* 1041–1052. doi:10.2174/1389557043402964
106. Zhang, Q.-Y.; Aires-de-Sousa, J. *J. Chem. Inf. Model.* **2007,** *47,* 1–8. doi:10.1021/ci050520j

107. Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. *J. Cheminf.* **2018,** *10,* 4. doi:10.1186/s13321-018-0258-y
108. Żurański, A. M.; Wang, J. Y.; Shields, B. J.; Doyle, A. G. *React. Chem. Eng.* **2022,** *7,* 1276–1284. doi:10.1039/d2re00030j
109. Li, S.-C.; Wu, H.; Menon, A.; Spiekermann, K. A.; Li, Y.-P.; Green, W. H. *J. Am. Chem. Soc.* **2024,** *146,* 23103–23120. doi:10.1021/jacs.4c04670
110. Low, K.; Coote, M. L.; Izgorodina, E. I. *J. Chem. Theory Comput.* **2023,** *19,* 1466–1475. doi:10.1021/acs.jctc.2c00984
111. Neeser, R. M.; Isert, C.; Stuyver, T.; Schneider, G.; Coley, C. W. *Chem. Data Collect.* **2023,** *46,* 101040. doi:10.1016/j.cdc.2023.101040
112. Low, K.; Coote, M. L.; Izgorodina, E. I. *J. Chem. Theory Comput.* **2022,** *18,* 1607–1618. doi:10.1021/acs.jctc.1c01264
113. Al Ibrahim, E.; Farooq, A. *J. Phys. Chem. A* **2022,** *126,* 4617–4629. doi:10.1021/acs.jpca.2c00713
114. Komp, E.; Janulaitis, N.; Valleau, S. *Phys. Chem. Chem. Phys.* **2022,** *24,* 2692–2705. doi:10.1039/d1cp04422b
115. Sanches-Neto, F. O.; Dias-Silva, J. R.; Keng Queiroz Junior, L. H.; Carvalho-Silva, V. H. *Environ. Sci. Technol.* **2021,** *55,* 12437–12448. doi:10.1021/acs.est.1c04326
116. Zhang, Y.; Yu, J.; Song, H.; Yang, M. *J. Chem. Inf. Model.* **2023,** *63,* 5097–5106. doi:10.1021/acs.jcim.3c00892
117. Johnson, M. S.; Green, W. H. *React. Chem. Eng.* **2024,** *9,* 1364–1380. doi:10.1039/d3re00684k
118. Beker, W.; Gajewska, E. P.; Badowski, T.; Grzybowski, B. A. *Angew. Chem., Int. Ed.* **2019,** *58,* 4515–4519. doi:10.1002/anie.201806920
119. Li, X.; Zhang, S.-Q.; Xu, L.-C.; Hong, X. *Angew. Chem., Int. Ed.* **2020,** *59,* 13253–13259. doi:10.1002/anie.202000959
120. Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science* **2019,** *363,* eaau5631. doi:10.1126/science.aau5631
121. Jacobson, L. D.; Bochevarov, A. D.; Watson, M. A.; Hughes, T. F.; Rinaldo, D.; Ehrlich, S.; Steinbrecher, T. B.; Vaitheeswaran, S.; Philipp, D. M.; Halls, M. D.; Friesner, R. A. *J. Chem. Theory Comput.* **2017,** *13,* 5780–5797. doi:10.1021/acs.jctc.7b00764
122. Liu, S.-C.; Zhu, X.-R.; Liu, D.-Y.; Fang, D.-C. *Phys. Chem. Chem. Phys.* **2023,** *25,* 913–931. doi:10.1039/d2cp04720a
123. Probst, D.; Schwaller, P.; Reymond, J.-L. *Digital Discovery* **2022,** *1,* 91–97. doi:10.1039/d1dd00006c
124. Chen, K.; Chen, G.; Li, J.; Huang, Y.; Wang, E.; Hou, T.; Heng, P.-A. *J. Cheminf.* **2023,** *15,* 43. doi:10.1186/s13321-023-00715-x
125. Kroll, A.; Rousset, Y.; Hu, X.-P.; Liebrand, N. A.; Lercher, M. J. *Nat. Commun.* **2023,** *14,* 4139. doi:10.1038/s41467-023-39840-4
126. Neves, P.; McClure, K.; Verhoeven, J.; Dyubankova, N.; Nugmanov, R.; Gedich, A.; Menon, S.; Shi, Z.; Wegner, J. K. *J. Cheminf.* **2023,** *15,* 20. doi:10.1186/s13321-023-00685-0
127. Ranković, B.; Griffiths, R.-R.; Moss, H. B.; Schwaller, P. *Digital Discovery* **2024,** *3,* 654–666. doi:10.1039/d3dd00096f
128. Wen, M.; Blau, S. M.; Xie, X.; Dwaraknath, S.; Persson, K. A. *Chem. Sci.* **2022,** *13,* 1446–1458. doi:10.1039/d1sc06515g
129. Chen, L.-Y.; Hsu, T.-W.; Hsiung, T.-C.; Li, Y.-P. *J. Phys. Chem. A* **2022,** *126,* 7548–7556. doi:10.1021/acs.jpca.2c04848
130. Li, Y.-P.; Han, K.; Grambow, C. A.; Green, W. H. *J. Phys. Chem. A* **2019,** *123,* 2142–2152. doi:10.1021/acs.jpca.8b10789
131. Lin, Y.-H.; Liang, H.-H.; Lin, S.-T.; Li, Y.-P. *ChemRxiv* **2024**. doi:10.26434/chemrxiv-2024-nmnlk
132. Muthiah, B.; Li, S.-C.; Li, Y.-P. *J. Taiwan Inst. Chem. Eng.* **2023,** *151,* 105123. doi:10.1016/j.jtice.2023.105123
133. Yang, C.-I.; Li, Y.-P. *J. Cheminf.* **2023,** *15,* 13. doi:10.1186/s13321-023-00682-3
134. Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. *Chem. Sci.* **2019,** *10,* 370–377. doi:10.1039/c8sc04228d
135. Keto, A.; Guo, T.; Underdue, M.; Stuyver, T.; Coley, C. W.; Zhang, X.; Krenske, E. H.; Wiest, O. *J. Am. Chem. Soc.* **2024,** *146,* 16052–16061. doi:10.1021/jacs.4c03131
136. Wu, Y.; Zhang, C.; Wang, L.; Duan, H. *Chem. Commun.* **2021,** *57,* 4114–4117. doi:10.1039/d1cc00586c
137. Dold, D.; Aranguren van Egmond, D. *Cell Rep. Phys. Sci.* **2023,** *4,* 101586. doi:10.1016/j.xcrp.2023.101586
138. Wang, Q.; Zhang, L. *Nat. Commun.* **2021,** *12,* 5359. doi:10.1038/s41467-021-25490-x
139. Xiong, J.; Xiong, Z.; Chen, K.; Jiang, H.; Zheng, M. *Drug Discovery Today* **2021,** *26,* 1382–1393. doi:10.1016/j.drudis.2021.02.011
140. Wang, Y.; Wang, J.; Cao, Z.; Barati Farimani, A. *Nat. Mach. Intell.* **2022,** *4,* 279–287. doi:10.1038/s42256-022-00447-x
141. Wieder, O.; Kohlbacher, S.; Kuenemann, M.; Garon, A.; Ducrot, P.; Seidel, T.; Langer, T. *Drug Discovery Today: Technol.* **2020,** *37,* 1–12. doi:10.1016/j.ddtec.2020.11.009
142. Zang, X.; Zhao, X.; Tang, B. *Commun. Chem.* **2023,** *6,* 34. doi:10.1038/s42004-023-00825-5
143. Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; Sun, M. *AI Open* **2020,** *1,* 57–81. doi:10.1016/j.aiopen.2021.01.001
144. Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. *Proc. Mach. Learn. Res.* **2017,** *70,* 1263–1272.
145. Kwon, Y.; Kim, S.; Choi, Y.-S.; Kang, S. *J. Chem. Inf. Model.* **2022,** *62,* 5952–5960. doi:10.1021/acs.jcim.2c01085
146. Li, B.; Su, S.; Zhu, C.; Lin, J.; Hu, X.; Su, L.; Yu, Z.; Liao, K.; Chen, H. *J. Cheminf.* **2023,** *15,* 72. doi:10.1186/s13321-023-00732-w
147. Kwon, Y.; Lee, D.; Choi, Y.-S.; Kang, S. *J. Cheminf.* **2022,** *14,* 2. doi:10.1186/s13321-021-00579-z
148. Kwon, Y.; Lee, D.; Kim, J. W.; Choi, Y.-S.; Kim, S. *ACS Omega* **2022,** *7,* 44939–44950. doi:10.1021/acsomega.2c05165
149. Li, S.-W.; Xu, L.-C.; Zhang, C.; Zhang, S.-Q.; Hong, X. *Nat. Commun.* **2023,** *14,* 3569. doi:10.1038/s41467-023-39283-x
150. Tavakoli, M.; Shmakov, A.; Ceccarelli, F.; Baldi, P. *arXiv* **2022,** 2201.01196. doi:10.48550/arxiv.2201.01196
151. Grambow, C. A.; Pattanaik, L.; Green, W. H. *J. Phys. Chem. Lett.* **2020,** *11,* 2992–2997. doi:10.1021/acs.jpclett.0c00500
152. Heid, E.; Green, W. H. *J. Chem. Inf. Model.* **2022,** *62,* 2101–2110. doi:10.1021/acs.jcim.1c00975
153. Yarish, D.; Garkot, S.; Grygorenko, O. O.; Radchenko, D. S.; Moroz, Y. S.; Gurbych, O. *J. Comput. Chem.* **2023,** *44,* 76–92. doi:10.1002/jcc.27016
154. Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. P. *J. Comput.-Aided Mol. Des.* **2005,** *19,* 693–703. doi:10.1007/s10822-005-9008-0
155. Gimadiev, T.; Nugmanov, R.; Khakimova, A.; Fatykhova, A.; Madzhidov, T.; Sidorov, P.; Varnek, A. *J. Chem. Inf. Model.* **2022,** *62,* 2015–2020. doi:10.1021/acs.jcim.1c01105
156. Lin, A.; Dyubankova, N.; Madzhidov, T. I.; Nugmanov, R. I.; Verhoeven, J.; Gimadiev, T. R.; Afonina, V. A.; Ibragimova, Z.; Rakhimbekova, A.; Sidorov, P.; Gedich, A.; Suleymanov, R.; Mukhametgaleev, R.; Wegner, J.; Ceulemans, H.; Varnek, A. *Mol. Inf.* **2022,** *41,* 2100138. doi:10.1002/minf.202100138

157. Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; Ye, W.; Zhang, Y.; Chang, Y.; Yu, P. S.; Yang, Q.; Xie, X. *ACM Trans. Intell. Syst. Technol.* **2024,** *15,* 1–45. doi:10.1145/3641289
158. Kasneci, E.; Seßler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; Krusche, S.; Kutyniok, G.; Michaeli, T.; Nerdel, C.; Pfeffer, J.; Poquet, O.; Sailer, M.; Schmidt, A.; Seidel, T.; Stadler, M.; Weller, J.; Kuhn, J.; Kasneci, G. *Learn. Individ. Diff.* **2023,** *103,* 102274. doi:10.35542/osf.io/5er8f
159. Thirunavukarasu, A. J.; Ting, D. S. J.; Elangovan, K.; Gutierrez, L.; Tan, T. F.; Ting, D. S. W. *Nat. Med.* **2023,** *29,* 1930–1940. doi:10.1038/s41591-023-02448-8
160. Weininger, D.; Weininger, A.; Weininger, J. L. *J. Chem. Inf. Comput. Sci.* **1989,** *29,* 97–101. doi:10.1021/ci00062a008
161. Chithrananda, S.; Grand, G.; Ramsundar, B. *arXiv* **2020,** 2010.09885. doi:10.48550/arxiv.2010.09885
162. Li, J.; Jiang, X. *Wirel. Commun. Mob. Com.* **2021,** 7181815. doi:10.1155/2021/7181815
163. Wang, S.; Guo, Y.; Wang, Y.; Sun, H.; Huang, J. Smiles-bert: large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics,* Niagara Falls, NY, USA, Sept 7–10, 2019; pp 429–436. doi:10.1145/3307339.3342186
164. Wu, Z.; Jiang, D.; Wang, J.; Zhang, X.; Du, H.; Pan, L.; Hsieh, C.-Y.; Cao, D.; Hou, T. *Briefings Bioinf.* **2022,** *23,* bbac131. doi:10.1093/bib/bbac131
165. Schwaller, P.; Probst, D.; Vaucher, A. C.; Nair, V. H.; Kreutter, D.; Laino, T.; Reymond, J.-L. *Nat. Mach. Intell.* **2021,** *3,* 144–152. doi:10.1038/s42256-020-00284-w
166. Frey, N. C.; Soklaski, R.; Axelrod, S.; Samsi, S.; Gómez-Bombarelli, R.; Coley, C. W.; Gadepally, V. *Nat. Mach. Intell.* **2023,** *5,* 1297–1305. doi:10.1038/s42256-023-00740-3
167. Blum, L. C.; Reymond, J.-L. *J. Am. Chem. Soc.* **2009,** *131,* 8732–8733. doi:10.1021/ja902302h
168. Ma, R.; Luo, T. *J. Chem. Inf. Model.* **2020,** *60,* 4684–4690. doi:10.1021/acs.jcim.0c00726
169. Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. *Mach. Learn.: Sci. Technol.* **2021,** *2,* 015016. doi:10.1088/2632-2153/abc81d
170. Wu, X.; Zhang, Y.; Yu, J.; Zhang, C.; Qiao, H.; Wu, Y.; Wang, X.; Wu, Z.; Duan, H. *Sci. Rep.* **2022,** *12,* 17098. doi:10.1038/s41598-022-21524-6
171. Zhong, Z.; Song, J.; Feng, Z.; Liu, T.; Jia, L.; Yao, S.; Wu, M.; Hou, T.; Song, M. *Chem. Sci.* **2022,** *13,* 9023–9034. doi:10.1039/d2sc02763a
172. Jaume-Santero, F.; Bornet, A.; Valery, A.; Naderi, N.; Vicente Alvarez, D.; Proios, D.; Yazdani, A.; Bournez, C.; Fessard, T.; Teodoro, D. *J. Chem. Inf. Model.* **2023,** *63,* 1914–1924. doi:10.1021/acs.jcim.2c01407
173. Lu, J.; Zhang, Y. *J. Chem. Inf. Model.* **2022,** *62,* 1376–1387. doi:10.1021/acs.jcim.1c01467
174. Dobbelaere, M. R.; Lengyel, I.; Stevens, C. V.; Van Geem, K. M. *J. Cheminf.* **2024,** *16,* 37. doi:10.1186/s13321-024-00834-z
175. Hu, Q.-N.; Deng, Z.; Hu, H.; Cao, D.-S.; Liang, Y.-Z. *Bioinformatics* **2011,** *27,* 2465–2467. doi:10.1093/bioinformatics/btr413
176. Zhang, M.-L.; Zhou, Z.-H. *IEEE Trans. Knowl. Data Eng.* **2014,** *26,* 1819–1837. doi:10.1109/tkde.2013.39
177. Wigh, D. S.; Arrowsmith, J.; Pomberger, A.; Felton, K. C.; Lapkin, A. A. *J. Chem. Inf. Model.* **2024,** *64,* 3790–3798. doi:10.1021/acs.jcim.4c00292
178. Beker, W.; Roszak, R.; Wołos, A.; Angello, N. H.; Rathore, V.; Burke, M. D.; Grzybowski, B. A. *J. Am. Chem. Soc.* **2022,** *144,* 4819–4827. doi:10.1021/jacs.1c12005
179. Maser, M. R.; Cui, A. Y.; Ryou, S.; DeLano, T. J.; Yue, Y.; Reisman, S. E. *J. Chem. Inf. Model.* **2021,** *61,* 156–166. doi:10.1021/acs.jcim.0c01234
180. Genheden, S.; Mårdh, A.; Lahti, G.; Engkvist, O.; Olsson, S.; Kogej, T. *Mol. Inf.* **2022,** *41,* 2100294. doi:10.1002/minf.202100294
181. Qian, Y.; Li, Z.; Tu, Z.; Coley, C. W.; Barzilay, R. *arXiv* **2023,** 2312.04881. doi:10.48550/arxiv.2312.04881
182. Wang, W.; Liu, Y.; Wang, Z.; Hao, G.; Song, B. *Chem. Sci.* **2022,** *13,* 12604–12615. doi:10.1039/d2sc04419f
183. Griffin, D. J.; Coley, C. W.; Frank, S. A.; Hawkins, J. M.; Jensen, K. F. *Org. Process Res. Dev.* **2023,** *27,* 1868–1879. doi:10.1021/acs.oprd.3c00229
184. ASKCOS; Computer-aided tools for Organic Synthesis. https://askcos.mit.edu (accessed March 27, 2024).
185. Seifrid, M.; Pollice, R.; Aguilar-Granda, A.; Morgan Chan, Z.; Hotta, K.; Ser, C. T.; Vestfrid, J.; Wu, T. C.; Aspuru-Guzik, A. *Acc. Chem. Res.* **2022,** *55,* 2454–2466. doi:10.1021/acs.accounts.2c00220
186. Guo, J.; Yu, C.; Li, K.; Zhang, Y.; Wang, G.; Li, S.; Dong, H. *J. Chem. Theory Comput.* **2024,** *20,* 4921–4938. doi:10.1021/acs.jctc.4c00071
187. Janet, J. P.; Mervin, L.; Engkvist, O. *Curr. Opin. Struct. Biol.* **2023,** *80,* 102575. doi:10.1016/j.sbi.2023.102575
188. Sim, M.; Vakili, M. G.; Strieth-Kalthoff, F.; Hao, H.; Hickman, R. J.; Miret, S.; Pablo-García, S.; Aspuru-Guzik, A. *Matter* **2024,** *7,* 2959–2977. doi:10.1016/j.matt.2024.04.022
189. Hammer, A. J. S.; Leonov, A. I.; Bell, N. L.; Cronin, L. *JACS Au* **2021,** *1,* 1572–1587. doi:10.1021/jacsau.1c00303
190. Torres, J. A. G.; Lau, S. H.; Anchuri, P.; Stevens, J. M.; Tabora, J. E.; Li, J.; Borovika, A.; Adams, R. P.; Doyle, A. G. *J. Am. Chem. Soc.* **2022,** *144,* 19999–20007. doi:10.1021/jacs.2c08592
191. Ruan, Y.; Lin, S.; Mo, Y. *J. Chem. Inf. Model.* **2023,** *63,* 770–781. doi:10.1021/acs.jcim.2c01168
192. Frazier, P. I. *arXiv* **2018,** 1807.02811. doi:10.48550/arxiv.1807.02811
193. Häse, F.; Aldeghi, M.; Hickman, R. J.; Roch, L. M.; Christensen, M.; Liles, E.; Hein, J. E.; Aspuru-Guzik, A. *Mach. Learn.: Sci. Technol.* **2021,** *2,* 035021. doi:10.1088/2632-2153/abedc8
194. Hickman, R.; Parakh, P.; Cheng, A.; Ai, Q.; Schrier, J.; Aldeghi, M.; Aspuru-Guzik, A. *ChemRxiv* **2023**. doi:10.26434/chemrxiv-2023-8nrxx
195. Kang, Y.; Yin, H.; Berger, C. *IEEE Trans. Intell. Veh.* **2019,** *4,* 171–185. doi:10.1109/tiv.2018.2886678
196. Balandat, M.; Karrer, B.; Jiang, D.; Daulton, S.; Letham, B.; Wilson, A. G.; Bakshy, E. *Adv. Neural Inf. Process. Syst.* **2020,** *33,* 21524–21538.
197. Griffiths, R.-R.; Klarner, L.; Moss, H. B.; Ravuri, A.; Truong, S.; Stanton, S.; Tom, G.; Rankovic, B.; Du, Y.; Jamasb, A.; Deshwal, A.; Schwartz, J.; Tripp, A.; Kell, G.; Frieder, S.; Bourached, A.; Chan, A.; Moss, J.; Guo, C.; Durholt, J.; Chaurasia, S.; Strieth-Kalthoff, F.; Lee, A. A.; Cheng, B.; Aspuru-Guzik, A.; Schwaller, P.; Tang, J. *arXiv* **2023,** 2212.04450. doi:10.48550/arxiv.2212.04450
198. Häse, F.; Aldeghi, M.; Hickman, R. J.; Roch, L. M.; Aspuru-Guzik, A. *Appl. Phys. Rev.* **2021,** *8,* 031406. doi:10.1063/5.0048164
199. Kandasamy, K.; Vysyaraju, K. R.; Neiswanger, W.; Paria, B.; Collins, C. R.; Schneider, J.; Poczos, B.; Xing, E. P. *J. Mach. Learn. Res.* **2020,** *21,* 1–27.

200. Paria, B.; Kandasamy, K.; Póczos, B. A flexible framework for multi-objective bayesian optimization using random scalarizations. In *Uncertainty in Artificial Intelligence,* 2020; pp 766–776.
201. Nambiar, A. M. K.; Breen, C. P.; Hart, T.; Kulesza, T.; Jamison, T. F.; Jensen, K. F. *ACS Cent. Sci.* **2022,** *8,* 825–836. doi:10.1021/acscentsci.2c00207
202. Wang, G.; Ang, H. T.; Dubbaka, S. R.; O'Neill, P.; Wu, J. *Trends Chem.* **2023,** *5,* 432–445. doi:10.1016/j.trechm.2023.03.008
203. Clayton, A. D. *Chem.: Methods* **2023,** *3,* e202300021. doi:10.1002/cmtd.202300021
204. Dietz, T.; Klamroth, K.; Kraus, K.; Ruzika, S.; Schäfer, L. E.; Schulze, B.; Stiglmayr, M.; Wiecek, M. M. *Eur. J. Oper. Res.* **2020,** *280,* 581–596. doi:10.1016/j.ejor.2019.07.027
205. Papoulias, S. A.; Grossmann, I. E. *Comput. Chem. Eng.* **1983,** *7,* 723–734. doi:10.1016/0098-1354(83)85024-8
206. Clayton, A. D.; Pyzer-Knapp, E. O.; Purdie, M.; Jones, M. F.; Barthelme, A.; Pavey, J.; Kapur, N.; Chamberlain, T. W.; Blacker, A. J.; Bourne, R. A. *Angew. Chem., Int. Ed.* **2023,** *62,* e202214511. doi:10.1002/anie.202214511
207. Kearney, A. M.; Collins, S. G.; Maguire, A. R. *React. Chem. Eng.* **2024,** *9,* 990–1013. doi:10.1039/d3re00678f
208. Nolan, L. J.; King, S. J.; Wharry, S.; Moody, T. S.; Smyth, M. *Curr. Opin. Green Sustainable Chem.* **2024,** *46,* 100886. doi:10.1016/j.cogsc.2024.100886
209. Climent, M. J.; Corma, A.; Iborra, S. *Chem. Rev.* **2011,** *111,* 1072–1133. doi:10.1021/cr1002084
210. Volk, A. A.; Epps, R. W.; Yonemoto, D. T.; Masters, B. S.; Castellano, F. N.; Reyes, K. G.; Abolhasani, M. *Nat. Commun.* **2023,** *14,* 1403. doi:10.1038/s41467-023-37139-y
211. Betinol, I. O.; Lai, J.; Thakur, S.; Reid, J. P. *J. Am. Chem. Soc.* **2023,** *145,* 12870–12883. doi:10.1021/jacs.3c03989
212. Kim, H.; Gerosa, G.; Aronow, J.; Kasaplar, P.; Ouyang, J.; Lingnau, J. B.; Guerry, P.; Farès, C.; List, B. *Nat. Commun.* **2019,** *10,* 770. doi:10.1038/s41467-019-08374-z
213. Rein, J.; Rozema, S. D.; Langner, O. C.; Zacate, S. B.; Hardy, M. A.; Siu, J. C.; Mercado, B. Q.; Sigman, M. S.; Miller, S. J.; Lin, S. *Science* **2023,** *380,* 706–712. doi:10.1126/science.adf6177
214. Rinehart, N. I.; Saunthwal, R. K.; Wellauer, J.; Zahrt, A. F.; Schlemper, L.; Shved, A. S.; Bigler, R.; Fantasia, S.; Denmark, S. E. *Science* **2023,** *381,* 965–972. doi:10.1126/science.adg2114
215. Wagen, C. C.; McMinn, S. E.; Kwan, E. E.; Jacobsen, E. N. *Nature* **2022,** *610,* 680–686. doi:10.1038/s41586-022-05263-2
216. Strassfeld, D. A.; Algera, R. F.; Wickens, Z. K.; Jacobsen, E. N. *J. Am. Chem. Soc.* **2021,** *143,* 9585–9594. doi:10.1021/jacs.1c03992
217. Strambeanu, I. I.; Diccianni, J. B. High-Throughput Experimentation in Discovery Chemistry: A Perspective on HTE Uses and Laboratory Setup. *The Power of High-Throughput Experimentation: General Topics and Enabling Technologies for Synthesis and Catalysis (Volume 1);* ACS Symposium Series, Vol. 1419; 2022; pp 11–22. doi:10.1021/bk-2022-1419.ch002
218. Buglioni, L.; Raymenants, F.; Slattery, A.; Zondag, S. D. A.; Noël, T. *Chem. Rev.* **2022,** *122,* 2752–2906. doi:10.1021/acs.chemrev.1c00332
219. Taylor, C. J.; Felton, K. C.; Wigh, D.; Jeraal, M. I.; Grainger, R.; Chessari, G.; Johnson, C. N.; Lapkin, A. A. *ACS Cent. Sci.* **2023,** *9,* 957–968. doi:10.1021/acscentsci.3c00050

License and Terms

BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

Computational design for enantioselective CO_2 capture: asymmetric frustrated Lewis pairs in epoxide transformations

Maxime Ferrer*[1], Iñigo Iribarren[2], Tim Renningholtz[3], Ibon Alkorta[1] and Cristina Trujillo*[3,4]

Full Research Paper

Open Access

Address:
[1]Instituto de Química Médica (CSIC), Juan de la Cierva, 3, 28006 Madrid, Spain, [2]Technische Universität München (TUM), School of Computation, Information and Technology, D-85748 Garching, Germany, [3]Department of Chemistry, The University of Manchester, Oxford Road, Manchester, M13 9PL, UK and [4]Trinity Biomedical Sciences Institute, School of Chemistry, The University of Dublin, Trinity College, D02 R590 Dublin 2, Ireland

Email:
Maxime Ferrer* - maxime.ferrer@iqm.csic.es; Cristina Trujillo* - cristina.trujillodelvalle@manchester.ac.uk

* Corresponding author

Keywords:
asymmetric catalysis; carbon dioxide; CO_2; epoxide; frustrated Lewis pairs (FLPs); volcano plot

Beilstein J. Org. Chem. **2024,** *20,* 2668–2681.
https://doi.org/10.3762/bjoc.20.224

Received: 09 July 2024
Accepted: 13 September 2024
Published: 22 October 2024

This article is part of the thematic issue "Adaptive experimentation and optimization in organic chemistry".

Guest Editor: A. Schweidtmann

Abstract

Carbon capture and utilisation (CCU) technologies offer a compelling strategy to mitigate rising atmospheric carbon dioxide levels. Despite extensive research on the CO_2 insertion into epoxides to form cyclic carbonates, the stereochemical implications of this reaction have been largely overlooked, despite the prevalence of racemic epoxide solutions. This study introduces an in silico approach to design asymmetric frustrated Lewis pairs (FLPs) aimed at controlling reaction stereochemistry. Four FLP scaffolds, incorporating diverse Lewis acids (LA), Lewis bases (LB), and substituents, were assessed via volcano plot analysis to identify the most promising catalysts. By strategically modifying LB substituents to induce asymmetry, a stereoselective catalytic scaffold was developed, favouring one enantiomer from both epoxide enantiomers. This work advances the in silico design of FLPs, highlighting their potential as asymmetric CCU catalysts with implications for optimising catalyst efficiency and selectivity in sustainable chemistry applications.

Introduction

The field of frustrated Lewis pairs (FLPs) has flourished since their seminal discovery in 2006 by Stephan and colleagues [1]. These compounds, which feature a Lewis acid (LA) and a Lewis base (LB), whose interaction is hindered by bulky substituents or chain strain, have garnered significant attention. Initially explored for their ability to trap small molecules [2,3], such as H_2 [4], CO_2 [5-7], N_2O [8,9], and alkenes [10,11], they have since found applications in catalysis [12,13].

Among the first catalytic uses of FLPs were the hydrogenation of unsaturated compounds [12,14] and the reduction of CO_2 using H_2 as a reductant [7,15-17]. FLPs have become an attractive avenue for the reduction of CO_2, particularly given the increasing levels of CO_2 in the atmosphere. However, challenges persist in understanding and optimising the reactivity of these systems.

One significant obstacle is the tendency for CO_2 to react preferentially with FLPs over H_2. As such, the design of FLPs that prioritise the capture of H_2 over CO_2 becomes crucial for effective CO_2 reduction [7]. Additionally, the strength of the interaction between the catalyst and the resulting system after hydride transfer presents a limitation. The formation of a robust LA–oxygen interaction may impede proton transfer to the basic oxygen atom. These limitations suggest that a more viable approach to employing FLPs as catalysts for CO_2-related reactions could involve their use in CO_2 activation [7,18,19]. In particular, the capture of CO_2 by FLPs enhances the electrophilicity of the CO_2 carbon atom and the nucleophilicity of one of the CO_2 oxygen atoms [6,7].

Carbon capture and utilisation (CCU) technologies involve the extraction of CO_2 from the atmosphere of the Earth to generate value-added chemicals, which can serve as platform chemicals in other chemical processes [20,21]. This is achieved by inserting CO_2 as a C1 building block into readily available substrates such as epoxides, resulting in the formation of polycarbonates or monomeric cyclic carbonates [22]. Depending on the substitution pattern in the epoxide, a chiral centre is present in the product.

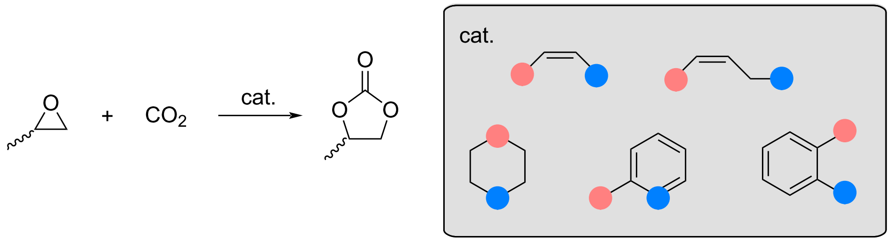
The insertion of CO_2 into epoxides has been the subject of numerous studies, but the stereochemical aspects of this reaction, particularly through the use of FLP catalysts, have been largely overlooked, despite the prevalence of racemic epoxide solutions. Only one study has addressed the asymmetric insertion of CO_2 into propylene oxide (PO) using a transition-metal catalyst [23-25]. Therefore, the stereochemical aspects of CO_2 insertion into PO enabled by FLP catalysts should be investigated.

To the best of our knowledge, only one paper has proposed an asymmetric approach to this reaction using a metal-based catalyst [23]. However, our approach differs significantly and seeks to explore new possibilities in this area. Herein, the present study focusses on the asymmetric insertion of CO_2 into PO using asymmetric FLPs as catalysts. Initially, five FLP scaffolds with different substituents, LA and LB, were tested, resulting in a total of 53 potential catalysts (Scheme 1). The most promising catalyst scaffolds for the reaction under study were identified by volcano plot analysis [26,27]. Inspired by the asymmetric oxazoline synthesised by Gao et al. [28], and guided by the volcano plot results, modifications to these FLP scaffolds facilitated the development of an asymmetric FLP and consequently an asymmetric catalyst. The subsequent study explores the asymmetric insertion of CO_2 into chiral PO catalysed by the proposed in silico designed catalyst.

Computational Details

During the benchmark to choose the best catalyst, the reported geometries were optimised with the Gaussian16 quantum chemical software package [29], using the B3LYP density functional [30,31] along with the Grimme dispersion correction including Becke and Johnson damping D3(BJ) [32-34] and the def2-TZVP basis set [35]. Harmonic frequencies were computed at the optimisation level to confirm that the relaxed structures correspond to local minima (no imaginary frequencies) or transition states (one imaginary frequency). The reaction simulations were run in chloroform using the "Solvation Model based on Density" (SMD) [36] at 273.0 K to reproduce the most commonly used experimental conditions [37-39].

When considering asymmetry, it was necessary to include large substituents on the catalyst to induce steric hindrance. These modifications increase the size of the asymmetric catalysts.



Scheme 1: Reaction between propylene oxide (PO) and CO_2 and the five catalyst scaffolds under study. The position of the LB along with an appropriate number of substituents is indicated by blue dots and that of the LA by pink dots.

Thus, the calculations presented in subsection "Asymmetric catalysis" were optimised at the B3LYP-D3(BJ)/def2-SVP computational level. Single point energy calculations on the optimised structures were run at the B3LYP-D3(BJ)/def2-TZVP level to obtain more accurate electronic energies. The reported free energies in this section correspond to the sum of the triple-zeta electronic energy and the free energy correction at double-zeta.

The kinetics of some reactions were calculated, applying the transition state theory [40]. Within this theory, the rate constant of an elementary reaction with the free energy barrier $\Delta G^{\ddagger}$ is given by Equation 1,

$$k = \frac{k_{\mathrm{B}}T}{h} e^{\frac{-\Delta G^{\ddagger}}{RT}}, \tag{1}$$

where k is the rate constant in s^{-1}, k_{B} is the Boltzmann constant, T is the temperature in Kelvin, h is the Planck constant, and R is the gas constant.

The enantiomeric excess (%ee) was calculated using Equation 2 [41]. k_{fav} stands for the kinetic rate constant of the most favourable process, and k_{defav} stands for the rate constant of the less favourable process.

$$\%\mathrm{ee} = \frac{k_{\mathrm{fav}} - k_{\mathrm{defav}}}{k_{\mathrm{fav}} + k_{\mathrm{defav}}}. \tag{2}$$

During the study, it will be observed that several transition states (TSs) can lead to the same product. As there is no possible interconversion between the reactant states, the different reactions will be considered independent, and it will be necessary to use an effective rate constant (k_{eff}). The definition given by Williams will be used (Equation 3, [42]):

$$k_{\mathrm{eff}} = \sum_{j}^{N_{\mathrm{TS}}} e^{-\Delta^{\ddagger}G_j / RT}. \tag{3}$$

The proton affinity (PA) [43] of the LB and the fluoride ion affinity (FIA) [44] of the LA of a given FLP are generally used to rationalise the FLP reactivity observed [45,46]. Thus, PA and FIA of the different scaffolds considered were calculated using Equation 4 and Equation 5, respectively, where H(A) stands for the enthalpy of the FLP, H(H$^+$) for the enthalpy of the proton, H(F$^-$) for the enthalpy of the fluoride ion, and H([A-H$^+$]) and H([A-F$^-$]) for the enthalpies of the complexes formed between the FLP and a proton and a fluoride ion, respectively.

$$\mathrm{PA} = \mathrm{H}(\mathrm{A}) + \mathrm{H}\left(\mathrm{H}^{+}\right) - \mathrm{H}\left(\left[\text{A-H}\right]^{+}\right), \tag{4}$$

$$\mathrm{FIA} = \mathrm{H}(\mathrm{A}) + \mathrm{H}\left(\mathrm{F}^{-}\right) - \mathrm{H}\left(\left[\text{A-F}\right]^{-}\right). \tag{5}$$

Volcanic 1.3.3, a Python package for the NaviCat platform, was used to generate 3D volcano plots, facilitating the identification of the most appropriate catalyst for the coupling reaction being considered [27].

Volcano plots

Volcano plots are a visualisation of the Sabatier principle [47], a qualitative concept originating in heterogeneous catalysis for assessing the performance of different catalysts. According to this principle, an ideal catalyst interacts with reactants neither strongly nor weakly. This idea is visualised in volcano plots, where a metric of the catalyst performance, for example, the reaction rate, is displayed as a function of the catalyst–substrate interaction, for example, the adsorption energy when considering a heterogeneous catalyst (Figure 1A). The resulting plot exhibits a volcano-like shape consisting of at least two slopes with the best-performing catalysts located at the top. The top of the volcano plot corresponds to the scenario where the catalyst exhibits high catalytic activity, as it achieves an optimal balance in binding to the reactants, neither too strongly nor too weakly (Figure 1A, purple points). The catalysts with binding energies lower than the catalysts at the top show lower catalytic activity due to insufficient activation of the substrate (Figure 1A, green points). Conversely, catalysts that bind too strongly impede the detachment of the catalyst–reactant complex, thereby reducing the catalyst turnover (Figure 1A, pink points) [26,27].

The previous plots are effective for metal-based catalysts and relatively simple catalytic reactions; however, they fall short when reactions involve multiple steps and independent activation barriers. In this paper, instead of focusing solely on the activation energy, the energy span of the catalytic reaction (δE) is considered. King et al. [48] introduced the concept of the energy span of a simulated catalytic cycle by defining it as the difference between the highest and lowest free energy stationary points [49,50]. More precisely, the energy span can be defined using Equation 6, where T_i is the energy of the rate-limiting TS, I_j the energy of the most populated intermediate, and $\delta G_{i,j}$ a correction that accounts for the cyclic nature of the catalytic cycle [26]:

$$\delta E = \max_{i,j}\left(T_i - I_j\right) + \delta G_{i,j}. \tag{6}$$

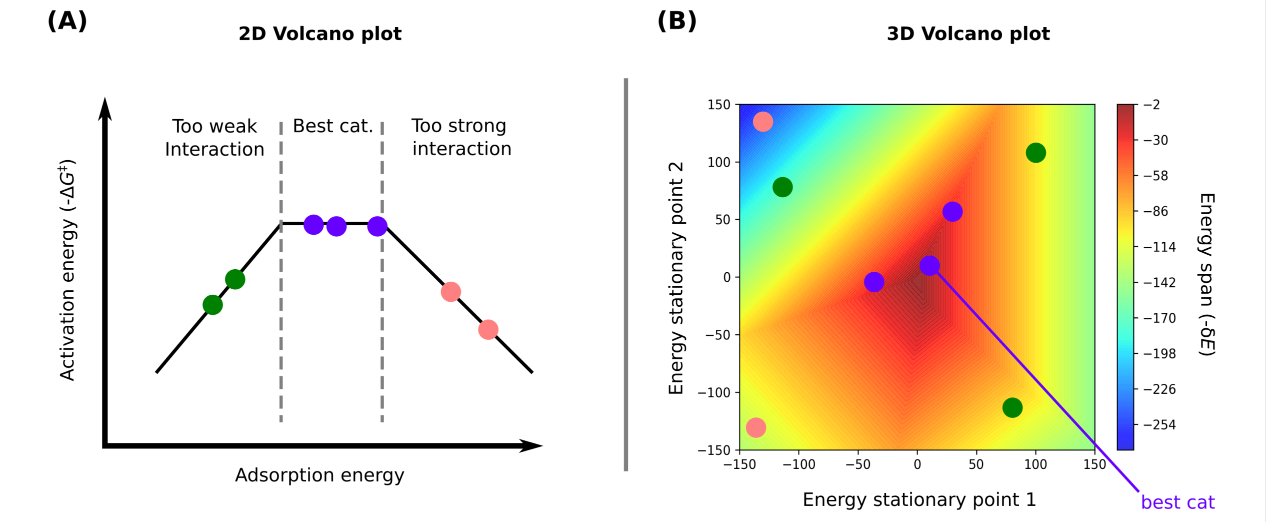


Figure 1: Schematic representation of an (A) 2D and a (B) 3D volcano plot. The abbreviation “cat.” stands for catalyst.

The energy span is a crucial parameter as it directly correlates with the turnover frequency (TOF) of the catalytic reaction (Equation 7). A flatter energy profile, indicated by a δ*E* value closer to zero, signifies more efficient catalysis:

$$\mathrm{TOF} = \frac{k_\mathrm{B}T}{h} e^{\frac{-\delta E}{RT}} . \qquad (7)$$

In this work, to achieve better correlations between the energy span and the system energies, two energies were used (Figure 1B). The volcanic program employs a multivariate linear regression process. Considering a reaction with six stationary points (REACTANTS, E1, TS2, E2, TS3, and PRODUCTS), which can be catalysed by *n* potential catalysts, the program calculates the correlation between the energy span and all possible pairs of stationary points for the *n* catalysts. For instance, it determines a function such as E1 = *f*(E2, TS3). The quality of these correlations is assessed through the square of the Pearson coefficient. The pair of stationary points has to correlate with the energies of the six stationary points previously presented. The quality of the pair considered to describe well the energies of the catalytic reaction is, thus, obtained by taking the mean value of R^2 of the six correlations. The pair with the largest mean R^2, is considered as the optimal pair of stationary points; it is then used to predict the energy span, resulting in a 2D contour plot (Figure 1B). The *x* axis represents the free energy of the first stationary point of the selected pair and the *y* axis represents the free energy of the other stationary point of the pair. According to the volcano plots, the best theoretically predicted catalysts are those nearest to the lowest predicted δ*E* values, depicted by the purple points in Figure 1B.

Results and Discussion

The following nomenclature will be used during the volcano plot analysis: FX_LBLA_S1_S2 where X is the label of the family (1, 2, 3, 5, or 6), LB is the Lewis base considered (N or P), LA is the Lewis acid (in this particular study only B), S1 is the substituent on the LB, and S2 is the substituent on the LA.

Capture of CO_2 and PO by an FLP

Chemoselectivity

Our investigations began by examining the uncatalysed coupling reaction between CO_2 and PO (Scheme 1), which exhibits a calculated activation barrier ($\Delta G^{\ddagger}$) greater than 55 kcal·mol^{-1} (Figure S6, Supporting Information File 1). Therefore, in order to observe the coupling between these two moieties under standard conditions, the presence of a catalyst is necessary. In the literature, metal-based and organocatalysts have been reported as efficient catalysts for this reaction [24,25]. As noted previously [51,52], the reaction depicted in Scheme 1 can proceed via two distinct mechanisms.

In the first mechanism, the catalyst initiates epoxide opening, followed by CO_2 insertion. The second mechanism suggests that CO_2 activation by the catalyst precedes its transfer to the epoxide. To determine the more feasible mechanism, a comprehensive investigation of both possibilities was conducted. To determine the most probable mechanism within our system, the capture of CO_2 and a symmetric epoxide (E) using the FLP proposed by Stephan et al. [37] was evaluated (Scheme 2). A symmetric epoxide was chosen to avoid addressing asymmetry concerns at this stage. The capture exhibiting the lowest activation barrier was considered the first step of the coupling reaction for the remainder of the study. The free-energy profiles of both capture processes are depicted in Figure S1 (Supporting

Information File 1). Notably, the CO_2 capture exhibits a lower activation barrier compared to the capture of epoxide (+10.0 kcal·mol^{-1} vs +30.0 kcal·mol^{-1}). Using transition state theory [40] as expressed in Equation 1, the rate constants were calculated for binding to either molecule at 273.0 K, resulting in $k_1 = 5.47{\cdot}10^4$ s^{-1} for capturing CO_2 and $k_2 = 4.85{\cdot}10^{-12}$ s^{-1} for capturing the epoxide. Despite the FLP–CO_2 adduct being less thermodynamically stable than the FLP–epoxide adduct (−10.1 kcal·mol^{-1} vs −44.8 kcal·mol^{-1}), the lower activation barrier for the capture of CO_2 and the temperature considered (273.0 K) suggest a kinetically controlled reaction. To further shift the chemical equilibrium toward CO_2 capture, increasing steric hindrance at the epoxide was explored by introducing bulky substituents into the scaffold. This resulted in an increase in activation barriers for adduct formation. Including a methyl group, for instance, increased the barrier by 1.4 kcal·mol^{-1}, a phenyl group by 1.7 kcal·mol^{-1}, and a *tert*-butyl group by more than 2 kcal·mol^{-1} (Table S1, Supporting Information File 1). This observation is consistent with reports in the literature [24,53-56]. Based on this initial study, it can be concluded that the mechanism for our system proceeds according to mechanism two. The following simulations were performed on this conclusion.

Regioselectivity

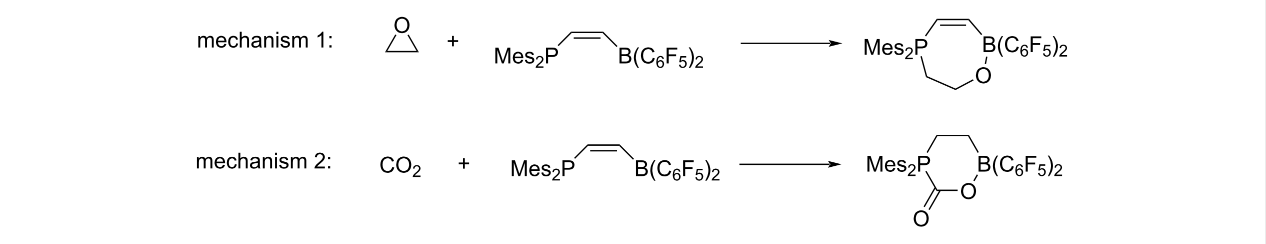
PO exhibits two distinct electrophilic sites, which can be subject to nucleophilic attack (Figure 2B). Thus, the regioselectivity of the CO_2 insertion into PO must be addressed as part of the full mechanistic investigation. The compound 3-boryl-2-propen-1-amine is now considered as the catalyst (Figure 2B). As observed in Figure 2A, the bond length and electron density at the bond critical point (BCP) difference are minimal and do not conclusively suggest that one bond will be broken more easily than the other. Therefore, both scenarios will be explored to see if the coupling reaction could proceed more easily by breaking the O–CH(CH_3) bond rather than the O–CH_2 bond.

Based on our investigations, the opening of PO with activated CO_2 was found to proceed through two transition states. The calculations showed that the breaking of the O–CH(CH_3) bond was more kinetically favourable, with a TS 7.6 kcal·mol^{-1} lower in free energy than the corresponding TS for breaking the O–CH_2 bond. The electron-donating nature of the methyl group facilitates a greater stabilisation of the intermediary positive charge at the central carbon compared to the hydrogen after bond-breaking at the terminal carbon, thereby reducing the activation barrier.

Henceforth, in this paper, the optimised TSs will consistently represent the breaking of the O–CH(CH_3) bond. Additionally, the (*S*)-epoxide enantiomer was employed consistently.

Symmetric FLP scaffolds – achiral environment

Following the initial exploration and preliminary results, our attention shifted toward the identification of a suitable catalyst. Drawing inspiration from the literature, fourteen FLP scaffolds have been evaluated (Figure 3), focussing specifically on N/B



Scheme 2: Capture reactions of CO_2 or an epoxide by FLP.

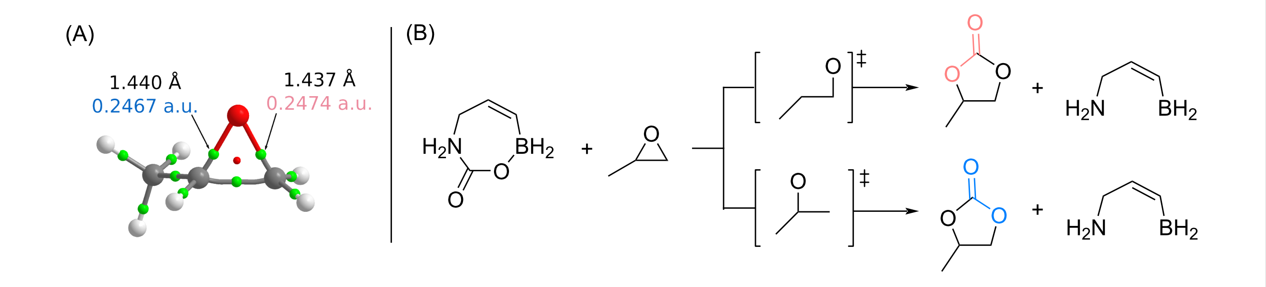


Figure 2: (A) Structure of PO annotated with the C–O bond distances and electron densities at the BCPs. BCPs are indicated by green spheres and the ring critical point by a red sphere. (B) Schematic representation of the two possible ring-opening reactions of PO in the presence of activated CO_2.

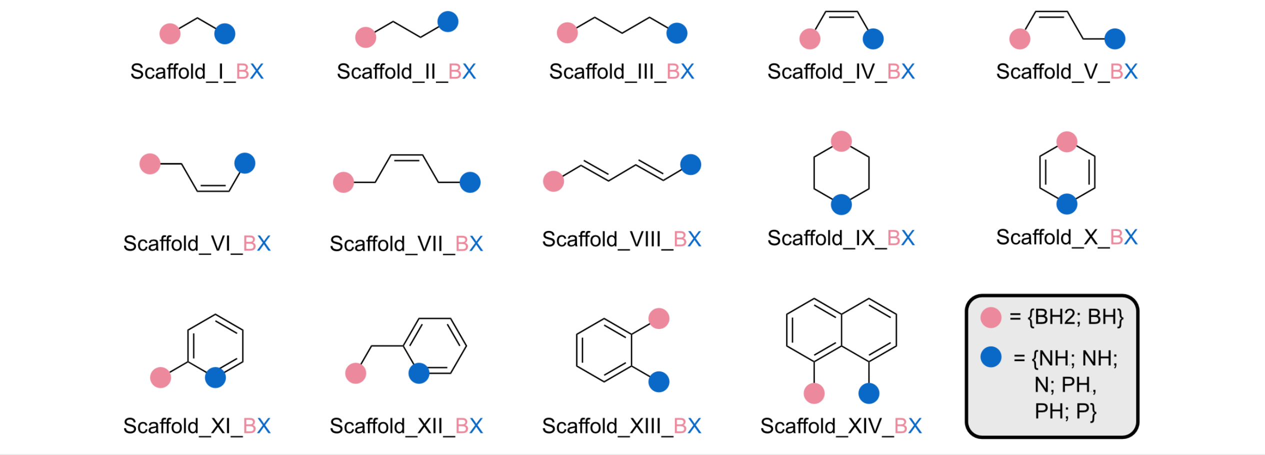


Figure 3: Symmetric FLP scaffolds considered in the first study. X denotes N or P.

and P/B FLPs because of their widespread application in this field, especially considering the initial step involving CO_2 capture [6,12,21].

Selection of the scaffolds and substituents

Volcano plots were introduced to find the most efficient catalyst for a given reaction [26,57]. They are a valuable tool for the in silico design of catalysts [27,58]. Volcano plot analysis requires a set of reactions that follow the same mechanism but whose stationary points possess different energies. Generally, the larger the differences in energy between the stationary points, the better the exploration of the catalytic space.

To determine the set of scaffolds to be used for volcano plot analysis, the CO_2–FLP adduct of each of the fourteen scaffolds was optimised (Figure 3). Based on the stability of the optimised adducts, families can be selected to cover a wide energy range. The obtained free energies of formation are presented in Figure S2 (Supporting Information File 1). The stabilities of the N/B adducts range from −9 to +48 kcal·mol^{-1}, while the P/B adducts vary from +10 to +36 kcal·mol^{-1}. The reliability of a volcano plot is based on an extensive exploration of the energetic space. Thus, because of the larger variation in the energy of the CO_2–FLP adducts employing N/B FLPs, it was decided to choose systems based on FLPs with N/B. It is expected that the large energy range of the CO_2–FLP adducts will propagate along the reaction path, enabling appropriate energy exploration. The scaffolds V_BX (family 1, ΔG(adduct) = −0.4 kcal·mol^{-1}), IX_BX (family 2, ΔG(adduct) = −2.1 kcal·mol^{-1}), XIII_BX (family 3, ΔG(adduct) = +3.0 kcal·mol^{-1}), IV_BX (family 4, ΔG(adduct) = +15.4 kcal·mol^{-1}), and XI_BX (family 5, ΔG(adduct) = −19.5 kcal·mol^{-1}) were selected for further investigation (Figure 4). This selection allows us to obtain free energy differences of 35 kcal·mol^{-1} already in the adduct stationary point.

After selecting the scaffolds to work with, the next step is to choose substituents for placement on the LA and LB positions.

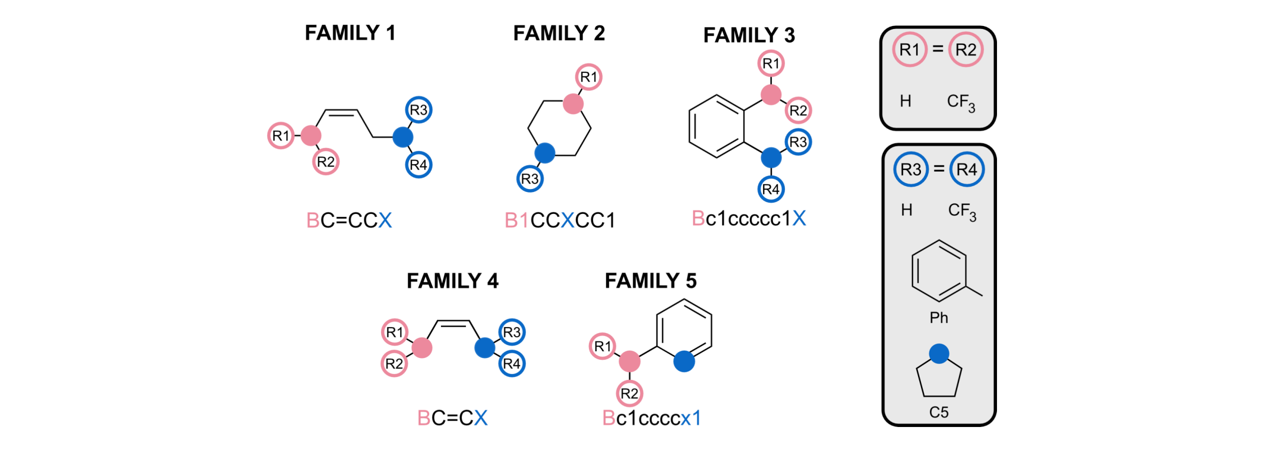


Figure 4: Subset of FLP scaffolds considered in the catalyst optimisation study. Substituents and labels are indicated.

These substituents will have two main effects on the FLP. First, they will alter the Lewis acidity and basicity of the LA and LB centres, respectively; second, they may induce steric hindrance. The first effect is perhaps the most intriguing to consider, as the acidity and basicity of the LA/LB centres are indicative of the FLP's reactivity [45,46]. Thus, substituents must be selected to ensure a broad spectrum of acidity and basicity of the LA and LB. Different methods for determining these properties have been described in the literature. Because of their easy computation, the proton affinity [43] and fluoride ion affinity [44] were selected to compute the basicity and acidity of the systems considered. By selecting the substituents presented in Figure 4, FIAs spanning a range of 60 kcal·mol^{-1} and PAs spanning a range of 48 kcal·mol^{-1} were obtained (Figure S3, Supporting Information File 1). All the structures studied exhibit the classical FLP characteristics, except for some systems that can be considered as "masked FLPs" (Table S4, Supporting Information File 1).

Possible mechanisms

As established in the previous section. the general mechanism of the coupling reaction proceeds through three steps: (1) capture of CO_2, (2) opening of PO and addition of the activated CO_2, and (3) liberation of the product (Figure 5).

During the study of the selected catalysts (Figure 4), it was observed that, depending on the catalyst considered, the reaction followed a different mechanism, that is, mechanism 1, 2, or 3 (Figure 5). Mechanism 1 (Figure 5A), comprises three steps. First, CO_2 is captured by FLP (TS01), and subsequently activated CO_2 is inserted into the epoxide (TS12). TS12 corresponds to the concerted opening of PO and the insertion of CO_2. The product is later released via TS23. This mechanism is followed by 40% of the catalysed reactions studied. Mechanism 2 (Figure 5B) contains an additional step. In this mechanism, the epoxide is first isomerised through TS14, resulting in the formation of the aldehyde (Min4). It can be observed that the opening of the epoxide is catalysed by the presence of the CO_2 adduct. In the gas phase and isolated, the isomerisation of the epoxide exhibits a barrier of 52.6 kcal·mol^{-1}. In the case of F2_NB_H_H, the barrier is reduced to 37.0 kcal·mol^{-1}. CO_2 later reacts with the aldehyde, forming the insertion product already observed in mechanism 1 (Min2). Passing through TS23, the product is released. Similar to mechanism 2, mechanism 3 contains eleven stationary points (Figure 5C). After the capture of CO_2 by the FLP, the opening of the epoxide takes place along with the insertion reaction. The main difference from the previous two mechanisms is that a new intermediate (Min5) is stabilised, in which the oxygen of CO_2 has attacked the electrophilic carbon of PO, and the oxygen atom of PO interacts with the LB. This mechanism is exclusive to phosphorus-containing FLPs, as nitrogen does not support this type of reactivity. Subsequently, the intermediate undergoes reorganization, leading to Min2.

Surprisingly, family 5, having phosphorus as the Lewis base, presents a different reactivity from the other families (Figure S3, Supporting Information File 1). Compounds F5_PB_H and

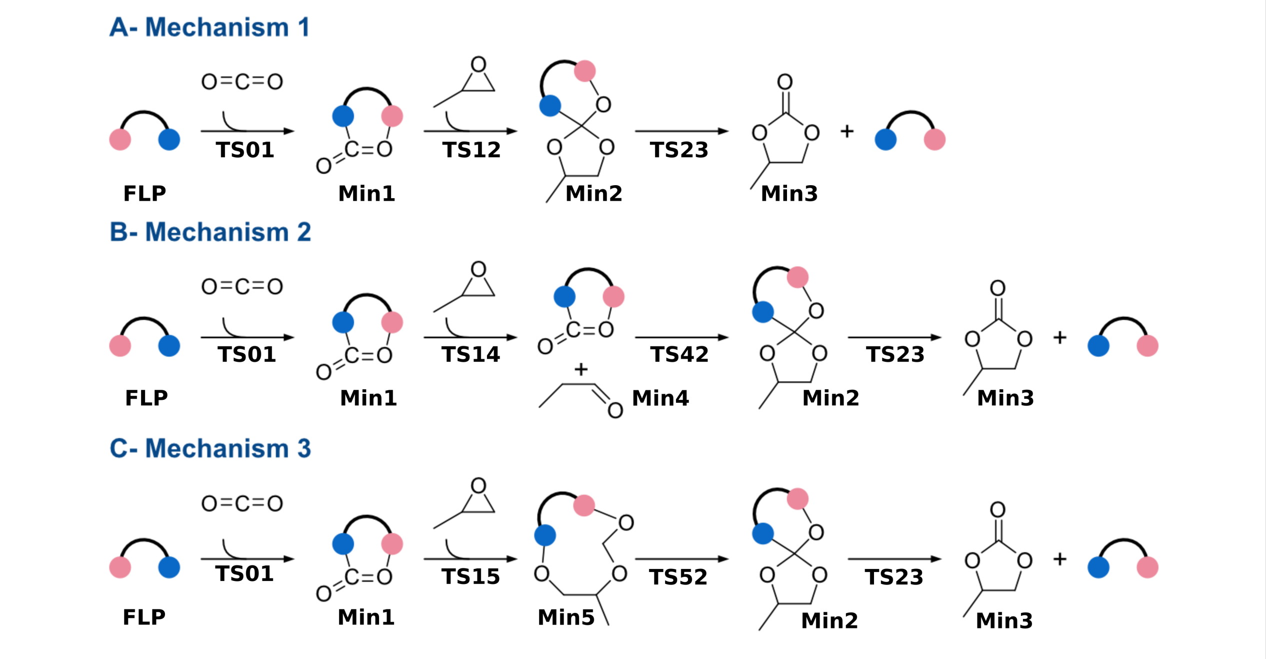


Figure 5: Coupling reaction between PO and CO_2. Depending on the catalyst considered, the reaction follows mechanism 1, 2, or 3.

F5_PB_CF3 react following mechanism 3 (Figure 5C), but the reaction proceeds directly from Min5 to Min3, with no Min2 observed. These two cases were then removed from the volcano plot analysis. The remaining two catalysts from family 5, namely, F5_NB_H and F5_NB_CF3, react according to mechanism 1 (Figure 5A, Figure S4, Supporting Information File 1). The energy matrix obtained can be found in Table S2 and Table S3 (Supporting Information File 1). It is interesting to observe that of the remaining 47 catalysts, 12 are not catalytically active, having their largest activation barrier greater than the 55.0 kcal·mol^{-1} previously reported for the uncatalysed reaction. Most of these belong to family 1.

Optimising catalyst selection

The optimal scaffolds and substituents for the studied reaction were identified through analysis using a volcano plot. This analysis requires that all catalytic cycles present the same number of steps. Because of varying numbers of steps between mechanism 1 and mechanisms 2 and 3, they had to be treated separately. The first group comprises catalysts that yield reactions following mechanism 1 (Figure 5A), characterised by nine stationary points. The second group consists of reactions with eleven stationary points, indicating that FLPs catalyse reactions following mechanisms 2 or 3 (Figure 5B,C). For each group, an analysis was performed using two volcano plots. The first plot aids in identifying the best families, which are then exclusively considered for the second volcano plot. The second plot helps to determine the most appropriate substituents to consider, thereby highlighting the optimal catalyst.

The first group of reactions, those following mechanism 1 (Figure 5A), comprises a total of twenty-two FLPs, accounting for 40% of the 55 catalysts considered. This group 1 includes FLPs from families 1, 3, 4, and 5. Sixteen compounds are based on an N/B pair, while the remaining six are P/B FLPs. Given the relative complexity of the mechanism studied, it was necessary to employ a 3D volcano plot using the energy span (δE) and two energies of the system. Analysis of the correlations revealed that the most suitable combination of energies to consider involved the energy of pre-TS01 assembly, which is the non-covalent complex formed between the FLP and the CO_2 molecule, and the energy of the intermediate Min2 (Figure 5). Correlating these parameters with the energy span yields an R^2 value of 0.79, a mean absolute error (MAE) of 2.59, and a standard mean absolute percentage error (MAPE) of 0.35.

As depicted in Figure 6A, families 3, 4, and 5 emerge as catalysts that catalyse the reaction most effectively. This aligns with previous findings that family 1 is not suitable for catalysing the reaction. However, compounds F1_PB_Ph_H and F1_PB_Ph_CF3 from family 1 are exceptions as they exhibit acceptable catalytic activity.

To identify the most suitable substituents, compounds of family 1 were excluded (except F1_PB_Ph_H and F1_PB_Ph_CF3), and a new volcano plot (Figure 6B) was generated. This plot employs the same axes as before (ΔG(pre-TS01 assembly) and ΔG(Min2)) and identifies a catalyst worthy of special consideration, that is, F3_NB_C5_CF3.

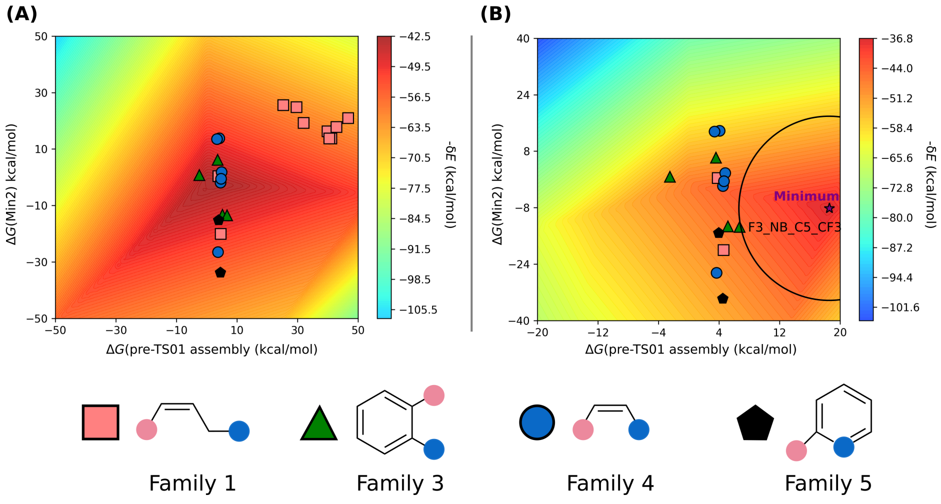


Figure 6: VOLCANO plot group 1. The free energies of pre-TS01 assembly and Min2 are considered for the correlation. (A) On the left, the compounds of families 1, 3, 4, and 5 were used for the plot. (B) On the right, only families 3, 4, and 5 are considered along with compounds F1_PB_Ph_H and F1_PB_Ph_CF3. In (B), a purple star was used to locate the minimum of the surface, along with a circle centred at the minimum to locate the closest systems. Also in (B), the names of the most effective catalysts are indicated in black.

Additionally, it is observed that an efficient catalyst for this reaction should have an unstable pre-TS assembly, pre-TS01 assembly, (E1 > 0), and an intermediate Min2 with an energy close to 0 kcal·mol^{-1}. Remarkably, among the most efficient catalysts within this group of FLPs those with a nitrogen LB stand out. This phenomenon could be attributed to the exceptional stability of the covalent adduct formed between phosphorus-based FLPs and CO_2.

The second group comprises compounds that undergo reactions following mechanisms 2 or 3 (Figure 5). This group represents 60% of the 55 catalysts considered. This time, the set is richer in FLPs based on phosphorus, comprising 21 out of 29 compounds. It includes compounds from families 1, 2, 3, and 4. Similarly to the previous group, a 3D volcano plot was utilised. The same variables (energy of pre-TS01 assembly and Min2) were considered, which yielded a correlation with a R^2 value of 0.71.

As depicted in Figure 7A, it is clear that the best family for this mechanism is family 2, followed by families 3 and 4. Family 1, similar to the previous group, exhibits the lowest catalytic activity. The low reactivity could be attributed to the masked character of this family. FLP monomers belonging to family 1 can be considered as masked FLPs [59,60], requiring breaking the LA–LB bond to achieve reactivity. Consequently, the pre-TS assembly formed between CO_2 and the FLPs from family 1 are less stable than the pre-TS assembly between CO_2 and the other FLP families, because of the absence of possible interactions between CO_2 and LA or L). Furthermore, TS01, corresponding to the capture of CO_2, is higher in energy because of the need for breaking the LA–LB bond.

By exclusively considering family 2 and using the energy of pre-TS01 assembly and product P4 for the second volcano plot (Figure 7B), it is observed that the best candidates are F2_PB_H_CF3 and F2_PB_Ph_CF3. Then, it can be concluded that the catalytic activity of the FLP is more efficient if the boron bears CF_3 substituents. Thus, an acidic boron atom seems to increase the reactivity of the considered system. Concerning the LB, it appears that, as opposed to the first group of compounds, a phosphorus atom is more active than a nitrogen atom.

Asymmetric catalysis

After examining the volcano results, we looked into the literature to explore examples of asymmetric FLPs. This exploration revealed three main types, namely, intramolecular chiral FLPs, intermolecular FLPs composed of a chiral acid and an achiral base, and intermolecular FLPs comprising an achiral acid and a chiral base [61,62]. One study reported a reaction involving the asymmetric reduction of ketones using an achiral borane, denoted as $B(p\text{-}HC_6F_4)_3$, paired with a chiral oxazoline, as depicted in Scheme 3 [28]. Remarkably, in this study, these FLPs demonstrated the capability to achieve high conversion rates and enantiomeric excess.

Following the volcano plot analysis presented in the previous section, F3_NB_C5_CF3 emerged as one of the top FLP catalysts under study. This catalyst, adhering to mechanism 1, incorporates a CF_3 group on the boron atom, serving as a simpli-

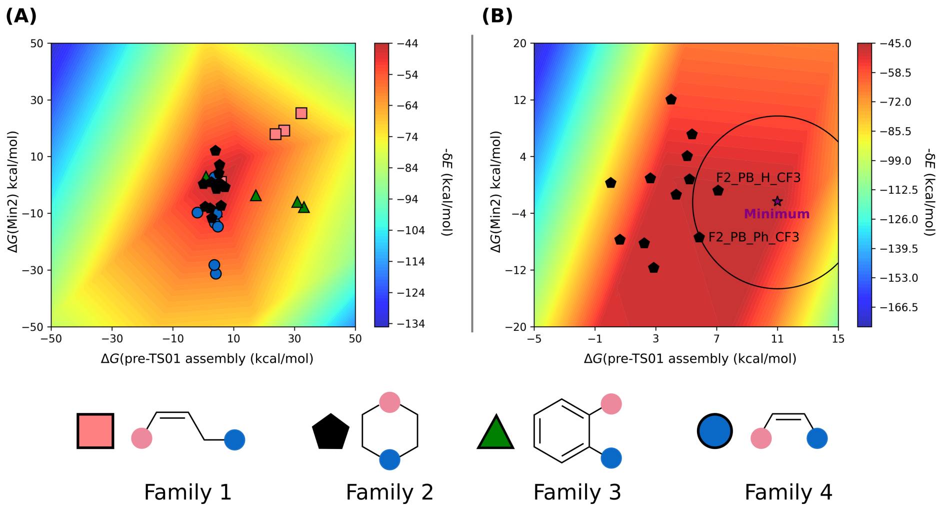
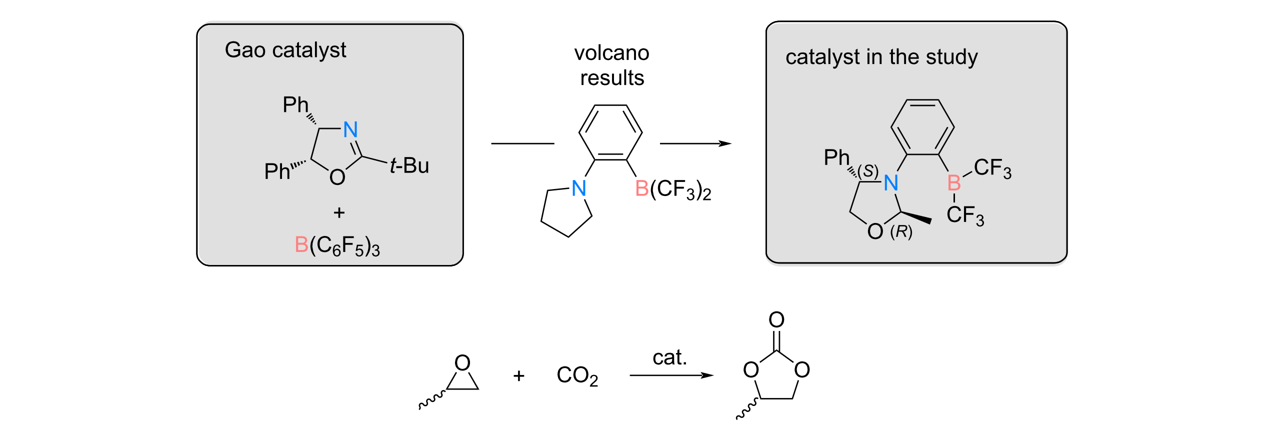


Figure 7: VOLCANO plot group 2. The free energies of pre-TS01 assembly and Min2 are considered for the correlation. (A) On the left, the compounds of families 1, 2, 3 and 4 are depicted. (B) On the right, only family 2 is considered. In (B) a purple star was used to locate the minimum of the surface, along with a circle centred at the minimum to locate the closest systems. Also in (B), the names of the most effective catalysts are indicated in black.



Scheme 3: Asymmetric catalysis studied. On the left, the catalyst proposed by Gao et al. for the asymmetric hydrogenation of a ketone is shown [28]. On the right, the catalyst design inspired by the Gao catalyst and the volcano plot results is shown. At the bottom the reaction under study is presented.

fied version of the B(p-HC_6F_4)$_3$. Notably, the nitrogen in this FLP is situated within a five-membered ring. Using this structural insight, an asymmetric catalyst was subsequently designed by strategically modifying the pyrrolidine substituent (C5 in Figure 4) based on the most efficient FLP.

The coupling reaction proposed in Scheme 3 was studied. In order to minimise the computational costs associated with the study, the asymmetric catalyst was obtained by removing a phenyl group and exchanging the *t*-Bu group with a methyl group in the catalyst of Gao [28]. It appears that the capture of CO_2 by the catalyst is barrierless and results in the formation of an adduct with a relative free energy of 0.7 kcal·mol^{-1}. Thus, the evaluation of the stereoselectivity of the designed catalyst was conducted by only studying the steps after the capture of CO_2 by the catalyst.

The reaction occurs in two steps (Table 1). Initially, a pre-TS assembly, with the PO compound positioned 2.67 Å from the CO_2 carbon is formed. Overcoming a TS, an intermediate is generated. In this intermediate, the distance between PO and the CO_2 carbon decreases to 1.61 Å from the initial 2.67 Å, and the interaction between nitrogen and the CO_2 carbon weakens. The intermediate is highly energetic and closely positioned to the TS. In the case of the (*R*) mechanism, the intermediate is slightly higher in energy than the TS, potentially because of methodological error. The intermediate further reacts with the activated CO_2 to generate the corresponding product. As can be observed in Table 1, the mechanism leading to the (*S*) product presents an asymmetric TS, 1.2 kcal·mol^{-1} higher in energy than that of the (*R*) mechanism. Thus, the asymmetric catalyst enables to generate an enantiomeric excess of 95% with the (*R*) product being the most abundant product.

Surprisingly, a transition state connecting the (*S*) epoxide with the (*R*) product was identified (Figure 8). Even more intriguingly, this new TS (TS_S_R in), verified by the IRC calculation (Figure S5, Supporting Information File 1) is the most stable TS located (Figure 9).

In this TS, the epoxy ring opens (Figure 9). Because of a shorter C–C distance between the CH3 group in the catalyst and the epoxy carbon atom (3.45 Å vs 3.75 Å in TS_S_S), a steric clash between the two methyl groups occurs (Figure 9). This results in an inversion of stereochemistry via rotation of the epoxy C–C bond, leading to the formation of the (*R*) product. As two TSs now yield the same product, it is necessary to recalculate the %ee, but this time using an effective rate constant k_{eff} (Equation 3). In doing so, a small increase in enantioselectivity is observed, with now a (*R*) enantiomeric excess of 96%ee. The de-

Table 1: Free energy reaction profile of the asymmetric coupling between propylene oxide and CO_2 catalysed by the catalyst depicted in Scheme 3. The energies are reported in kcal·mol^{-1}, and the 0.0 energy was set to be the FLP–CO_2 adduct and the isolated propylene oxide.

Enantiomer	Pre-TS	TS1	Int.	Asymmetric TS	Product
R	9.4	27.6	27.9	30.9	–14.2
S	8.9	25.0	24.5	32.1	–13.4

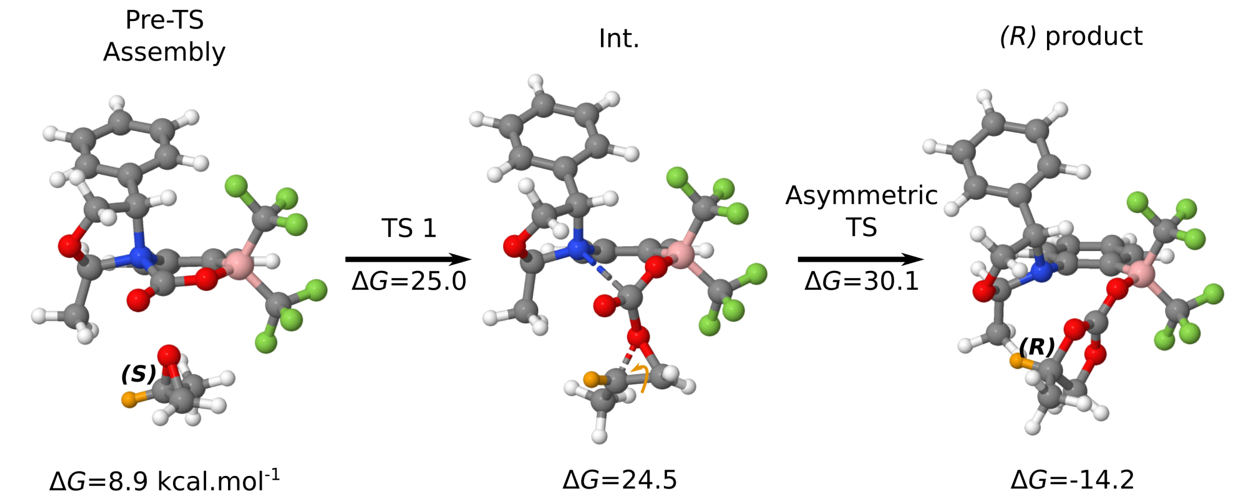


Figure 8: Catalysed reaction between the (S)-enantiomer of propylene oxide and CO_2 resulting in the formation of the (*R*)-product. In orange, the hydrogen atom that illustrates the stereochemistry inversion. The free energies are given in kcal·mol^{-1}

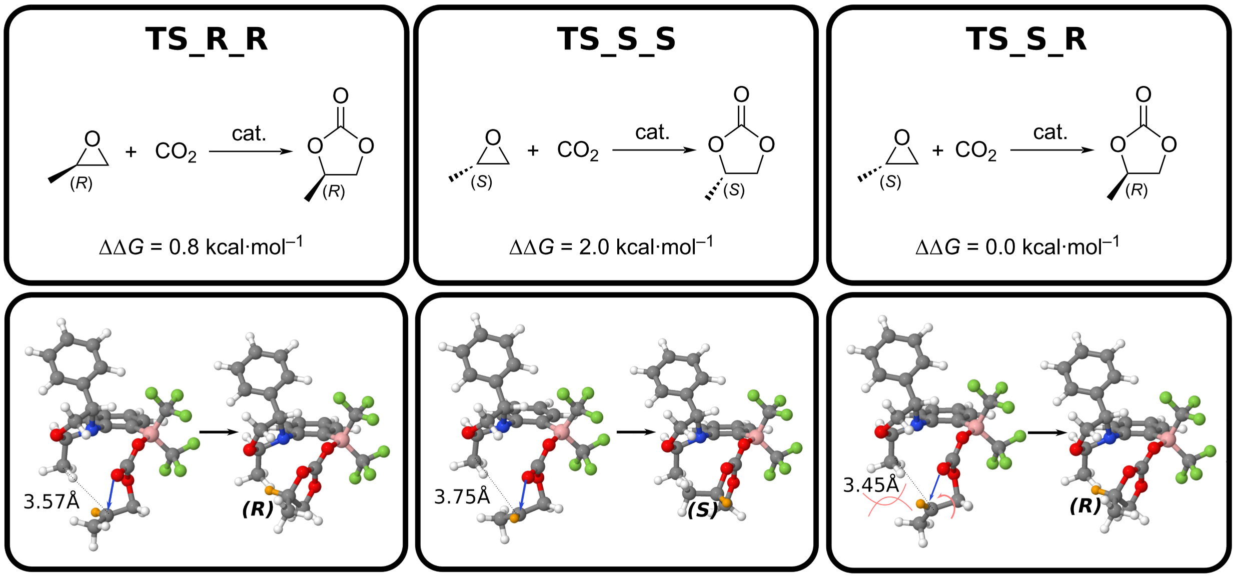


Figure 9: Schemes of the different asymmetric reactions observed. Hydrogen capable of rotation is marked in orange, influencing the stereochemistry at the TS.

signed catalyst enables the generation of an almost enantiomerically pure product from a racemic mixture.

Conclusion

Carbon capture and utilisation technologies represent a promising avenue for addressing increasing atmospheric carbon dioxide levels. The reaction involving the insertion of CO_2 into epoxides to form cyclic carbonates is a key focus within this domain. Despite extensive exploration, the stereochemical aspects of this reaction have been surprisingly underexplored, especially in the context of racemic epoxide mixtures commonly encountered in practice.

This study introduces an innovative in silico design strategy for asymmetric frustrated Lewis pairs tailored specifically to control the stereochemistry of the CO_2 insertion reaction. Computational evaluations of four distinct FLP scaffolds, incorporating various Lewis acids, Lewis bases, and substituents, identify the most promising catalyst candidates through volcano plot analysis. The volcano plot analysis reveals that the best candidate is F3_NB_C5_CF3, which is the catalyst based on the 2-borylbenzenamine scaffold, with a pyrrolidine substituent on the nitrogen atom and CF_3 substituents on the boron.

Through strategic modification of the Lewis base substituents, a stereoselective catalyst was engineered to produce a single enantiomer preferentially from both enantiomers of the epoxide substrate. An enantiomeric excess of 95%ee was initially achieved, with the predominant (*R*) enantiomer. Enhanced selectivity was subsequently observed through additional transition states, resulting in a remarkable 96%ee yielded by the catalyst.

Supporting Information

Supporting information features geometries of the different stationary points optimised as well as figures and tables mentioned in the main text. The outputs of the calculations presented can be found at the following link: https://doi.org/10.5281/zenodo.

Supporting Information File 1

Supporting figures and tables.
[https://www.beilstein-journals.org/bjoc/content/supplementary/1860-5397-20-224-S1.pdf]

Acknowledgements

The authors acknowledge the assistance provided by Research IT and the use of the Computational Shared Facility at The University of Manchester, HPC-Europe3, the Centro de Supercomputacion de Galicia (CESGA), and the Irish Centre for High-End Computing (ICHEC) for their continued computational support. Additionally, the authors express their gratitude to Dr. Rubén LaPlaza for his valuable scientific discussions and insights.

Funding

This publication has emanated from research supported by Science Foundation Ireland (SFI 18/SIRG/5517) and the Ministerio de Ciencia, Innovación y Universidades (PID2021-125207NB-C3).

Author Contributions

Maxime Ferrer: conceptualization; data curation; investigation; writing – original draft; writing – review & editing. Iñigo Iribarren: conceptualization; data curation; formal analysis; writing – review & editing. Tim Renningholtz: formal analysis; investigation; writing – review & editing. Ibon Alkorta: conceptualization; data curation; investigation; supervision; writing – review & editing. Cristina Trujillo: conceptualization; formal analysis; funding acquisition; investigation; project administration; writing – review & editing.

ORCID® iDs

Maxime Ferrer - https://orcid.org/0000-0001-7838-9974
Iñigo Iribarren - https://orcid.org/0000-0003-0373-8687
Tim Renningholtz - https://orcid.org/0000-0001-8793-4057
Ibon Alkorta - https://orcid.org/0000-0001-6876-6211
Cristina Trujillo - https://orcid.org/0000-0001-9178-5146

Data Availability Statement

The data generated and analyzed during this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.12633864.

Preprint

A non-peer-reviewed version of this article has been previously published as a preprint: https://doi.org/10.3762/bxiv.2024.47.v1

References

1. Welch, G. C.; Juan, R. R. S.; Masuda, J. D.; Stephan, D. W. *Science* **2006,** *314,* 1124–1126. doi:10.1126/science.1134230
2. Ghara, M.; Mondal, H.; Pal, R.; Chattaraj, P. K. *J. Phys. Chem. A* **2023,** *127,* 4561–4582. doi:10.1021/acs.jpca.3c02141
3. Mondal, H.; Patra, S. G.; Chattaraj, P. K. *Struct. Chem.* **2022,** *33,* 1853–1865. doi:10.1007/s11224-022-01934-y
4. Stephan, D. W. *J. Am. Chem. Soc.* **2021,** *143,* 20002–20014. doi:10.1021/jacs.1c10845
5. Stephan, D. W.; Erker, G. *Chem. Sci.* **2014,** *5,* 2625–2641. doi:10.1039/c4sc00395k
6. Pérez-Jiménez, M.; Corona, H.; de la Cruz-Martínez, F.; Campos, J. *Chem. – Eur. J.* **2023,** *29,* e202301428. doi:10.1002/chem.202301428
7. Khan, M. N.; van Ingen, Y.; Boruah, T.; McLauchlan, A.; Wirth, T.; Melen, R. L. *Chem. Sci.* **2023,** *14,* 13661–13695. doi:10.1039/d3sc03907b
8. Otten, E.; Neu, R. C.; Stephan, D. W. *J. Am. Chem. Soc.* **2009,** *131,* 9918–9919. doi:10.1021/ja904377v
9. Neu, R. C.; Otten, E.; Lough, A.; Stephan, D. W. *Chem. Sci.* **2011,** *2,* 170–176. doi:10.1039/c0sc00398k
10. Ullrich, M.; Seto, K. S.-H.; Lough, A. J.; Stephan, D. W. *Chem. Commun.* **2009,** 2335–2337. doi:10.1039/b901212e
11. Jie, X.; Sun, Q.; Daniliuc, C. G.; Knitsch, R.; Hansen, M. R.; Eckert, H.; Kehr, G.; Erker, G. *Chem. – Eur. J.* **2020,** *26,* 1269–1273. doi:10.1002/chem.201905171
12. Stephan, D. W. *Acc. Chem. Res.* **2015,** *48,* 306–316. doi:10.1021/ar500375j
13. Paradies, J. *Eur. J. Org. Chem.* **2019,** 283–294. doi:10.1002/ejoc.201800944
14. Scott, D. J.; Fuchter, M. J.; Ashley, A. E. *Chem. Soc. Rev.* **2017,** *46,* 5689–5700. doi:10.1039/c7cs00154a
15. Liu, L.; Qiu, M.; Liu, S.; Ma, H.; Huang, Z.-Q.; Chang, C.-R. *J. Phys. Chem. C* **2023,** *127,* 6714–6722. doi:10.1021/acs.jpcc.3c00396
16. Das, S.; Laplaza, R.; Blaskovits, J. T.; Corminboeuf, C. *J. Am. Chem. Soc.* **2024,** *146,* 15806–15814. doi:10.1021/jacs.4c01890

17. Pimbaotham, P.; Injongkol, Y.; Jungsuttiwong, S.; Yodsin, N. *J. Catal.* **2024,** *436,* 115571. doi:10.1016/j.jcat.2024.115571
18. Mondal, H.; Chattaraj, P. K. *J. Comput. Chem.* **2024,** *45,* 1098–1111. doi:10.1002/jcc.27285
19. Mondal, H.; Patra, S. G.; Chattaraj, P. K. *J. Chem. Sci.* **2022,** *134,* 119. doi:10.1007/s12039-022-02119-0
20. Song, Q.-W.; Zhou, Z.-H.; He, L.-N. *Green Chem.* **2017,** *19,* 3707–3728. doi:10.1039/c7gc00199a
21. Fu, H.-C.; You, F.; Li, H.-R.; He, L.-N. *Front. Chem. (Lausanne, Switz.)* **2019,** *7,* 525. doi:10.3389/fchem.2019.00525
22. Song, X.; Wang, J.; Yang, L.; Pan, H.; Zheng, B. *Inorg. Chem. Commun.* **2020,** *121,* 108197. doi:10.1016/j.inoche.2020.108197
23. Berkessel, A.; Brandenburg, M. *Org. Lett.* **2006,** *8,* 4401–4404. doi:10.1021/ol061501d
24. Andrea, K. A.; Kerton, F. M. *ACS Catal.* **2019,** *9,* 1799–1809. doi:10.1021/acscatal.8b04282
25. Horton, T. A. R.; Wang, M.; Shaver, M. P. *Chem. Sci.* **2022,** *13,* 3845–3850. doi:10.1039/d2sc00894g
26. Wodrich, M. D.; Sawatlon, B.; Busch, M.; Corminboeuf, C. *Acc. Chem. Res.* **2021,** *54,* 1107–1117. doi:10.1021/acs.accounts.0c00857
27. Laplaza, R.; Das, S.; Wodrich, M. D.; Corminboeuf, C. *Nat. Protoc.* **2022,** *17,* 2550–2569. doi:10.1038/s41596-022-00726-2
28. Gao, B.; Feng, X.; Meng, W.; Du, H. *Angew. Chem., Int. Ed.* **2020,** *59,* 4498–4504. doi:10.1002/anie.201914568
29. *Gaussian 16,* Revision C.01; Gaussian, Inc.: Wallingford, CT, 2016.
30. Becke, A. D. *J. Chem. Phys.* **1992,** *96,* 2155–2160. doi:10.1063/1.462066
31. Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988,** *37,* 785–789. doi:10.1103/physrevb.37.785
32. Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *J. Chem. Phys.* **2010,** *132,* 154104. doi:10.1063/1.3382344
33. Becke, A. D.; Johnson, E. R. *J. Chem. Phys.* **2005,** *123,* 154101. doi:10.1063/1.2065267
34. Grimme, S.; Ehrlich, S.; Goerigk, L. *J. Comput. Chem.* **2011,** *32,* 1456–1465. doi:10.1002/jcc.21759
35. Weigend, F. *Phys. Chem. Chem. Phys.* **2006,** *8,* 1057–1065. doi:10.1039/b515623h
36. Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2009,** *113,* 6378–6396. doi:10.1021/jp810292n
37. Mömming, C. M.; Otten, E.; Kehr, G.; Fröhlich, R.; Grimme, S.; Stephan, D. W.; Erker, G. *Angew. Chem., Int. Ed.* **2009,** *48,* 6643–6646. doi:10.1002/anie.200901636
38. Krachko, T.; Nicolas, E.; Ehlers, A. W.; Nieger, M.; Slootweg, J. C. *Chem. – Eur. J.* **2018,** *24,* 12669–12677. doi:10.1002/chem.201801909
39. Dalpozzo, R.; Della Ca', N.; Gabriele, B.; Mancuso, R. *Catalysts* **2019,** *9,* 511. doi:10.3390/catal9060511
40. Eyring, H. *J. Chem. Phys.* **1935,** *3,* 107–115. doi:10.1063/1.1749604
41. Schneebeli, S. T.; Hall, M. L.; Breslow, R.; Friesner, R. *J. Am. Chem. Soc.* **2009,** *131,* 3965–3973. doi:10.1021/ja806951r
42. Williams, I. H. *J. Phys. Org. Chem.* **2022,** *35,* e4312. doi:10.1002/poc.4312
43. Kolboe, S. *J. Chem. Theory Comput.* **2014,** *10,* 3123–3128. doi:10.1021/ct500315c
44. Erdmann, P.; Leitner, J.; Schwarz, J.; Greb, L. *ChemPhysChem* **2020,** *21,* 987–994. doi:10.1002/cphc.202000244
45. Ferrer, M.; Alkorta, I.; Elguero, J.; Oliva-Enrich, J. M. *ChemPhysChem* **2022,** *23,* e202200204. doi:10.1002/cphc.202200204
46. Ferrer, M.; Alkorta, I.; Elguero, J.; Oliva-Enrich, J. M. *Phys. Chem. Chem. Phys.* **2024,** *26,* 12433–12443. doi:10.1039/d4cp00496e
47. Sabatier, P. In *La catalyse en chimie organique;* Béranger, C., Ed.; Librairie polytechnique: Paris, 1920; Vol. 3.
48. King, E. L.; Altman, C. *J. Phys. Chem.* **1956,** *60,* 1375–1378. doi:10.1021/j150544a010
49. Kozuch, S.; Shaik, S. *J. Am. Chem. Soc.* **2006,** *128,* 3355–3365. doi:10.1021/ja0559146
50. Kozuch, S.; Shaik, S. *Acc. Chem. Res.* **2011,** *44,* 101–110. doi:10.1021/ar1000956
51. Fiorani, G.; Guo, W.; Kleij, A. W. *Green Chem.* **2015,** *17,* 1375–1389. doi:10.1039/c4gc01959h
52. Shaikh, R. R.; Pornpraprom, S.; D'Elia, V. *ACS Catal.* **2018,** *8,* 419–450. doi:10.1021/acscatal.7b03580
53. Kayaki, Y.; Yamamoto, M.; Ikariya, T. *Angew. Chem., Int. Ed.* **2009,** *48,* 4194–4197. doi:10.1002/anie.200901399
54. Maya, E. M.; Rangel-Rangel, E.; Díaz, U.; Iglesias, M. *J. CO2 Util.* **2018,** *25,* 170–179. doi:10.1016/j.jcou.2018.04.001
55. Tsutsumi, Y.; Yamakawa, K.; Yoshida, M.; Ema, T.; Sakai, T. *Org. Lett.* **2010,** *12,* 5728–5731. doi:10.1021/ol102539x
56. Chatelet, B.; Joucla, L.; Dutasta, J.-P.; Martinez, A.; Szeto, K. C.; Dufaud, V. *J. Am. Chem. Soc.* **2013,** *135,* 5348–5351. doi:10.1021/ja402053d
57. Stratton, S. M.; Zhang, S.; Montemore, M. M. *Surf. Sci. Rep.* **2023,** *78,* 100597. doi:10.1016/j.surfrep.2023.100597
58. Wodrich, M. D.; Busch, M.; Corminboeuf, C. *Chem. Sci.* **2016,** *7,* 5723–5735. doi:10.1039/c6sc01660j
59. Fontaine, F.-G.; Stephan, D. W. *Philos. Trans. R. Soc., A* **2017,** *375,* 20170004. doi:10.1098/rsta.2017.0004
60. Ferrer, M.; Alkorta, I.; Elguero, J.; Oliva-Enrich, J. M. *ChemPhysChem* **2024,** *25,* e202300750. doi:10.1002/cphc.202300750
61. Feng, X.; Meng, W.; Du, H. Frustrated Lewis Pair Catalyzed Asymmetric Reactions. In *Frustrated Lewis Pairs;* Slootweg, J. C.; Jupp, A. R., Eds.; Springer: Cham, Switzerland, 2021; pp 29–86. doi:10.1007/978-3-030-58888-5_2
62. Kótai, B.; Laczkó, G.; Hamza, A.; Pápai, I. *Chem. – Eur. J.* **2024,** *30,* e202400241. doi:10.1002/chem.202400241

License and Terms

The definitive version of this article is the electronic one which can be found at:
https://doi.org/10.3762/bjoc.20.224

BEILSTEIN JOURNAL OF ORGANIC CHEMISTRY

Emerging trends in the optimization of organic synthesis through high-throughput tools and machine learning

Pablo Quijano Velasco[1], Kedar Hippalgaonkar[*1,2,3] and Balamurugan Ramalingam[*1,4]

Review

Open Access

Address:
[1]Institute of Materials Research and Engineering (IMRE), Agency for Science Technology and Research (A*STAR), 2 Fusionopolis Way, Singapore 138634, Republic of Singapore, [2]Department of Materials Science and Engineering, Nanyang Technological University, Singapore 639798, Republic of Singapore, [3]Institute for Functional Intelligent Materials, National University of Singapore, 4 Science Drive 2, Singapore 117544, Republic of Singapore and [4]Institute of Sustainability for Chemicals, Energy and Environment (ISCE2), Agency for Science Technology and Research (A*STAR), 1 Pesek Road, Jurong Island, Singapore 627833, Republic of Singapore

Email:
Kedar Hippalgaonkar* - kedar@ntu.edu.sg;
Balamurugan Ramalingam* -
balamurugan_ramalingam@imre.a-star.edu.sg

* Corresponding author

Keywords:
autonomous reactors; data processing; high-throughput experimentation; machine learning; reaction optimization

Beilstein J. Org. Chem. **2025,** *21,* 10–38.
https://doi.org/10.3762/bjoc.21.3

Received: 04 July 2024
Accepted: 26 November 2024
Published: 06 January 2025

This article is part of the thematic issue "Adaptive experimentation and optimization in organic chemistry".

Guest Editor: A. Schweidtmann

Abstract

The discovery of the optimal conditions for chemical reactions is a labor-intensive, time-consuming task that requires exploring a high-dimensional parametric space. Historically, the optimization of chemical reactions has been performed by manual experimentation guided by human intuition and through the design of experiments where reaction variables are modified one at a time to find the optimal conditions for a specific reaction outcome. Recently, a paradigm change in chemical reaction optimization has been enabled by advances in lab automation and the introduction of machine learning algorithms. Therein, multiple reaction variables can be synchronously optimized to obtain the optimal reaction conditions, requiring a shorter experimentation time and minimal human intervention. Herein, we review the currently used state-of-the-art high-throughput automated chemical reaction platforms and machine learning algorithms that drive the optimization of chemical reactions, highlighting the limitations and future opportunities of this new field of research.

Introduction

Organic synthesis plays a crucial role in drug discovery, polymer synthesis, materials science, agrochemicals, and specialty chemicals. Their synthesis and process optimization require substantial resources and are labor-intensive, often exploring only a single variable in search of the optimal conditions while disregarding the intricate interactions among competing variables within the synthesis process. The complexity of the problem requires consideration that process optimization often demands solutions that meet multiple targets, such as yield, selectivity, purity, cost, environmental impact, etc. In recent years, the advancement of artificial intelligence (AI), machine learning (ML), and automation has produced a paradigm shift for chemical synthesis optimization techniques. By leveraging on ML models to predict reaction outcomes and ML optimization algorithms, this new approach has demonstrated the ability to navigate the complex relationships between reaction variables and finding the global optimal conditions within a fewer number of experiments than with traditional methods [1,2]. In addition, machine-guided optimization has emerged as a promising framework to obtain reaction conditions that perform optimally for single- or multiple-target objectives, enabling researchers to explore diverse solution spaces and uncover the optimal conditions that strike a balance between consonant and/or conflicting targets. In addition, the incorporation of lab robotics into chemical synthesis has enabled the development of closed-loop optimization platforms capable of executing optimization campaigns rapidly with minimal human intervention, relieving experimenters from labor-intensive tasks and reducing the overall process development lead time [3,4].

A standard workflow and general methodology for organic reaction optimization through ML methods is shown in Figure 1. The workflow comprises (i) careful design of experiments (DOE); (ii) reaction execution with commercial high-throughput systems or in-house designed reaction modules; (iii) data collection by in-line/offline analytical tools; (iv) mapping the collected data points with the target objectives; (v) prediction of the next set of reaction conditions towards attaining optimal solutions; and (vi) experimental validation of suggested optimization results. Through an examination of methodologies, algorithms, and various case studies, this article offers our perspective on the state-of-the-art techniques for optimizing the synthesis of organic molecules, highlighting both challenges and prospects. The structure of this review follows the steps presented in Figure 1. In the following section, we review the high-throughput platforms currently used to perform chemical reaction optimization. Thereafter, we discuss the development and use of analytical tools and data processing algorithms. After that, we discuss the latest trends in the selection of optimization algorithms for chemical synthesis. Finally, we highlight the future directions and opportunities in the field. For an in-depth overview on the topic of chemical reaction optimization, the readers are referred to prominent reviews by Taylor et al. [5], Griffin et al. [6], and Sagmeister et al. [7]. The first two offer valuable perspectives on chemical reaction optimization, particularly focusing on process scale-up, while the latter discusses the potential of flow platforms for self-optimization reactions. Additionally, we refer the readers to the following literature in other areas relevant to the application of ML to chemical synthesis that are not covered by our review, such as small molecule discovery [8], drug discovery [9,10], retrosynthesis [11,12], and catalyst selection and design [13,14].

Review

HTE platforms

HTE platforms were designed to accelerate the discovery and development of organic molecules by the rapid screening and analysis of large numbers of experimental conditions simultaneously. For the purpose of this article, we define HTE as a technique that leverages a combination of automation, parallelization of experiments, advanced analytics, and data processing methods to streamline repetitive experimental tasks, reduce manual intervention, and increase the rate of experimental execution in comparison to traditional manual experimentation. In conventional chemical synthesis, several sequential steps are typically undertaken, involving the setup of the reaction, mixing of reactants, reaction workup, product analysis, and product purification. To perform all these basic chemistry tasks effec-

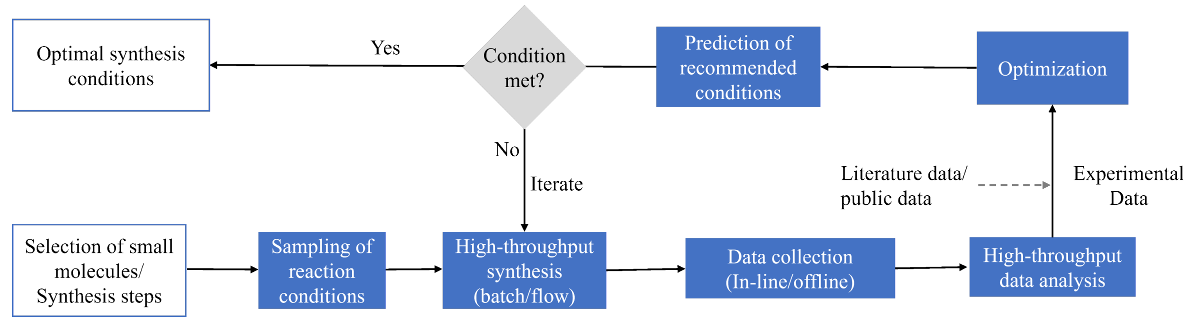

Optimal synthesis conditions
Yes
Condition met?
Prediction of recommended conditions
Optimization
No
Iterate
Literature data/ public data
Experimental Data
Selection of small molecules/ Synthesis steps
Sampling of reaction conditions
High-throughput synthesis (batch/flow)
Data collection (In-line/offline)
High-throughput data analysis

Figure 1: A high-level representation of the workflow and framework used for the optimization of organic reactions through optimization algorithms and high-throughput experimentation (HTE) tools.

tively, customizable HTE platforms are available from various laboratory instrument manufacturers or can be assembled using a mix of commercial and in-house developed equipment. Normally, HTE for organic chemistry will include a liquid transfer module, a reactor stage, and analytical tools for product characterization. When the full experimental process is automated and coupled with a centralized control system performing ML optimization, the HTE can function as a self-driving platform where the next iteration of experiments is automatically selected by algorithm without human intervention. This section will highlight the key features of various HTE platforms, including benefits, limitations, and applications to organic molecule synthesis.

HTE using batch modules

Batch reactions occur without flow of the reagents/products into or out of the reaction vessel until a target conversion has been achieved. HTE batch platforms leverage on parallelization of experiments to perform several reactions under different conditions simultaneously to increase the experimental throughput. Commonly, batch platforms include a liquid handling system for setting up reactions based on a plunger pump (e.g., syringe, pipette), a reactor capable of heating and mixing, and in-line/online analytical tools. HTE in batch excels in the control of categorical and continuous variables, in particular for the stoichiometry and chemical formulation of reaction mixtures. Many HTE batch experiments have been performed in self-contained automated platforms developed by various instrument manufacturers (Chemspeed, Zinsser Analytic, Mettler Toledo, Tecan, Unchained Labs, etc.). In these HTE platforms, microtiter well plates (MTP) and reaction blocks containing 96/48/24-well plates are widely used as reaction and characterization vessels [5]. UltraHTE configurations typically incorporate 1536-well plates, enabling the exploration of lager spaces of reaction parameters. While ultraHTE was initially tailored to biological assays, the versatility of these modules has been extended to optimizing chemistry-related processes [5]. The Chemspeed SWING robotic system, equipped with two fluoropolymers and PFA-mat-sealed 96-well metal blocks, was used for the exploration of stereoselective Suzuki–Miyaura couplings, offering precise control over both categorical and continuous variables (Figure 2a) [15]. The integrated robotic system containing a four-needle dispense head facilitated the delivery of reagents in low volume and slurries, ensuring the accuracy and throughput of the process. The entire experimental workflow was further optimized through parallelization, dividing reactions into eight loops, enabling them to complete 192 reactions within 24 loops, achieving a significant throughput in within four days. Other reports for various reactions include the Buchwald–Hartwig aminations [16-19], Suzuki couplings [16,17,20], N-alkylations [21], hydroxylations [22], and photochemical reactions [23-29]. The versatility to handle multiple reagents and the widespread availability of 96-well plates have facilitated the extensive adoption of HTE under batch conditions for optimizing chemical synthesis. However, several challenges arise when MTP are used as reaction vessels. First, the independent control of variables such as reaction time, temperature, and pressure within individual wells is not possible due to the inherent design constraints of parallel reactors that share the same MTP. In addition, challenges arise when standard MTP-based reaction vessels are used at a temperature near the solvent's boiling point, as this labware is not enclosed or able to cool the top of the reaction vessel to facilitate reflux conditions. Although some research groups have developed custom tools to enable high-temperature reactions, these reactors are currently not commercially available. For an in-depth discussion on the limitations of batch reactors for HTE, we refer the readers to a review by Taylor et al. on chemical reaction optimization [5].

In recent years, research laboratories have deviated from traditional commercial tools to HTE systems custom-built to the chemists' requirements and demands. Burger et al. [30] have creatively developed a mobile robot equipped with sample-handling arms, tailored for the precise execution of photocatalytic reactions for water molecule cleavage to produce hydrogen. The mobile robot (Figure 2b) acted as a substitute of a human experimenter by executing tasks and linking eight separate experimental stations, including solid and liquid dispensing, sonication, several characterization equipments, and stations for consumables and sample storage. Remarkably, through a tedious ten-dimensional parameter search spanning eight days, the robot achieved an impressive hydrogen evolution rate of approximately 21.05 $\mu mol \cdot h^{-1}$. Despite the initial investment and two-year development timeline, the versatility of this robotic system promises remarkable applications in materials, polymers, and chemical synthesis. Most automated synthesis platforms are based on expensive scientific equipment, have a large equipment footprint, and need extensive reconfiguration to adapt to new synthetic protocols. To address this issue, Manzano et al. [31] have developed a small-footprint portable chemical synthesis platform able to perform liquid and solid phase organic reactions (Figure 2c). The platform utilizes 3D-printed reactors that can be generated on demand based on the targeted reaction and features liquid handling, stirring, heating, and cooling modules for enhanced versatility. In addition, the platform is capable of performing under inert and low-pressure atmospheres, handling separation steps, and pressure sensing for reaction monitoring. Its efficacy and robustness were confirmed through the successful synthesis of five small organic molecules, four oligopeptides, and four oligonucleotides in high purity and impressive yield. Although, in the current configuration, the platform lacks characterization modules

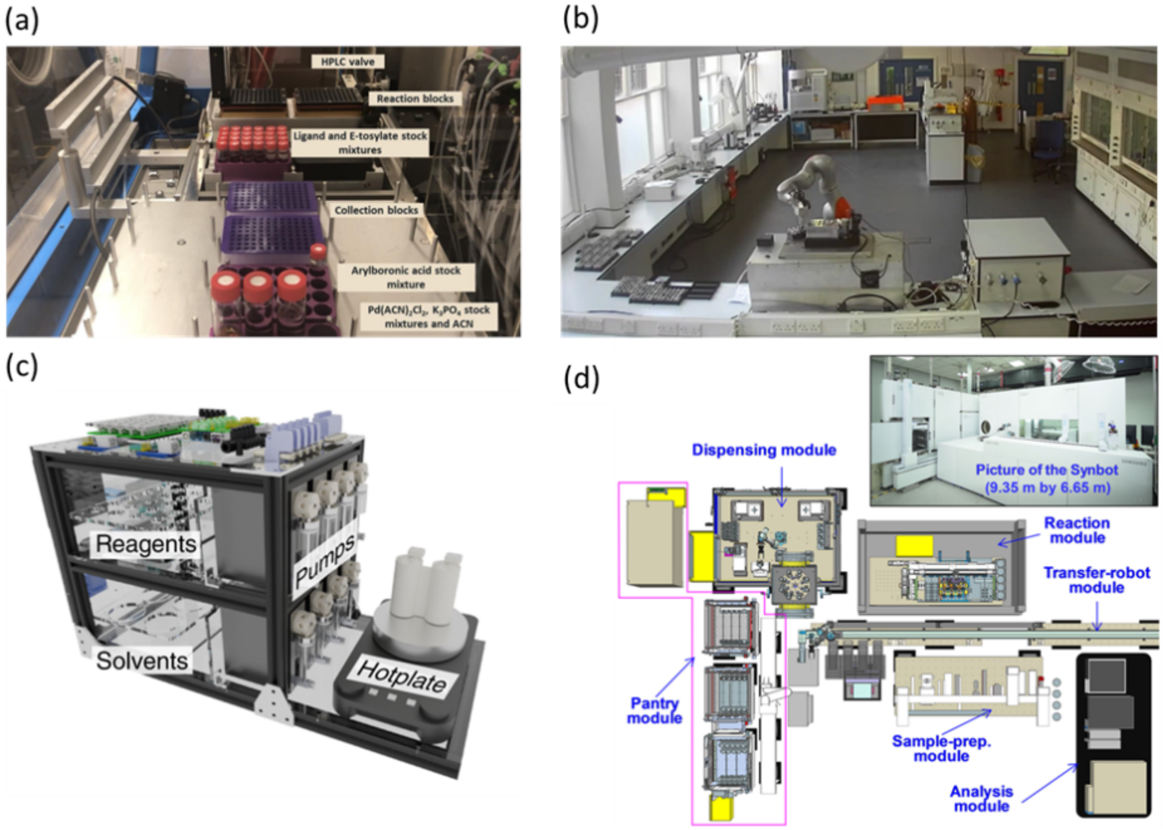


Figure 2: (a) Photograph showing a Chemspeed HTE platform using 96-well reaction blocks. (b) Mobile robot equipment performing tasks normally executed by human experimenters for the photocatalytic conversion of water to hydrogen. (c) Small-footprint portable chemical synthesis platform. (d) Schematic of the SynBot platform developed by Samsung researchers, showing each module used for chemical synthesis. Figure 2a was reproduced from [15] (© 2021 M. Christensen et al., published by Springer Nature, distributed under the terms of the Creative Commons Attribution 4.0 International License, https://creativecommons.org/licenses/by/4.0). Figure 2b is from [30] and was reprinted by permission from Springer Nature from the journal Nature ("A mobile robotic chemist" by B. Burger; P. M. Maffettone; V. V. Gusev; C. M. Aitchison; Y. Bai; X. Wang; X. Li; B. M. Alston; B. Li; R. Clowes; N. Rankin; B. Harris; R. S. Sprick; A. I. Cooper), Copyright © 2020 The Author(s), under exclusive licence to Springer Nature Limited. This content is not subject to CC BY 4.0. Figure 2c is from [31] and was reprinted by permission from Springer Nature from the journal Nature Chemistry ("An autonomous portable platform for universal chemical synthesis" by J. S. Manzano; W. Hou; S. S. Zalesskiy; P. Frei; H. Wang; P. J. Kitson; L. Cronin), Copyright © 2022 The Author(s), under exclusive licence to Springer Nature Limited. This content is not subject to CC BY 4.0. Figure 2d was adapted from [32] (© 2023 T. Ha et al., published by American Association for the Advancement of Science, distributed under the terms of the Creative Commons Attribution 4.0 International License, https://creativecommons.org/licenses/by/4.0).

and has a lower throughput in comparison to other automated platforms, it does offer a low-cost alternative that can be adapted to perform chemical reaction optimization.

In addition to academia, industry is increasingly recognizing the value of investing in custom-built HTE setups to automate their synthesis workflows for enhanced productivity. A fully integrated, cloud-accessible, automated synthesis laboratory (ASL) was designed and built by Eli Lilly [33]. This state-of-the-art facility allowed for heating, cryogenic conditions, microwaving, high-pressure reactions, evaporation, and workup, empowering researchers to conduct an extensive array of chemical reactions. The ASL comprises of three bench spaces dedicated to either high temperature reactions, cryogenic/microwave reactions, or reaction workup. On each bench, a translational combination of robotic arms performs the specific experiments using the modular platforms, while consumables and samples are transferred between benches through a conveyor belt, linking them together. According to the report, the ASL has facilitated over 16,350 gram-scale reactions across various case studies, showcasing the widespread capability. Researchers at Samsung have pioneered the development of SynBot, an innovative autonomous synthesis robot that uses AI and robotic technology to establish optimal synthetic procedures [32]. Similar to ASL, SynBot consists of five modules connected through a conveyor belt backbone, with a robot arm in charge of transferring the samples between them. The modules include a pantry for chemical storage and chemical selection, a dispensing module for solids and liquids, a reaction module capable of heating and stirring, a sample preparation module, and a LC–MS characterization module (Figure 2d). The efficiency of the system has been demonstrated in three reactions types, namely Suzuki–Miyaura coupling, Buchwald–Hartwig amination, and Ullmann coupling. These experiments showcased a conversion rate that outperformed existing reference systems and provided at least six times the efficiency in the experimentation, besides synthesis planning, optimization, and downstream workup tasks. The throughput of SynBot is estimated to be an average of 12 reactions within 24 hours depending on the reaction time. IBM has developed RoboRXN, a remotely accessible autonomous chemical laboratory that enables notable acceleration in chemical synthesis by leveraging cloud computing, AI, and automation [34]. The technology relies on an AI model that recommends the sequence of operations needed to perform the corresponding chemical reactions, including the order of reagent addition [35]. This model facilitates that the synthesis tasks are sent to a robotic research lab from anywhere in the world, allowing the robot to execute the recommended retrosynthesis provided by the user. The enhancements in RoboRXN assist chemists in predicting the environmental impact of chemical processes, and the new AI model also helps identify more environmentally friendly enzymes for chemical reactions. Although RoboRXN has demonstrated the ability to perform most tasks in chemical and material synthesis, the hardware currently cannot perform product purification and multistep synthesis continuously.

HTE using flow platforms

Flow reactions are characterized by a constant flow of reagents and products into and out of the reaction vessel. A flow platform consists of a fluid delivery system, mixing tools, reactors, quenching units, pressure regulation units, and collection vessels. The fluid delivery is normally executed using either high-pressure liquid chromatography (HPLC), a syringe, or peristaltic pumps. A passive mixing stage where the reagents are introduced to the system through a Y- or T-connection is the most common approach observed for most flow reactions, while more specialized mixing tools can be incorporated depending on the reaction prerequisites. The most common reactors used are either microfluidic chip- or coil-based reactors for solution chemistry. Packed bed reactors are used when solid heterogeneous catalysts and reagents (e.g., inorganic bases) are handled. Specialized reactors for electro- [36,37] and photochemical [38-40] experiments have also been developed. Depending on the flow of the reaction mixture, flow reactions can be continuous or segmented (also known as slug). Segmented flow reactions present an efficient means to gather diverse data points by creating segmented or droplet flow within microfluidic reactors. Each droplet is carefully separated by either an antisolvent or an inert gas, thus providing every droplet with the functionality of an individual reactor. This segmentation ensures precise control over reactions and prevents interference between different reaction environments. Moreover, the ratio of reagents within these droplets is easily modulated using syringe pumps, providing users with a convenient means to collect data efficiently and coherently. This approach streamlines experimentation processes, enhances reproducibility, and facilitates the exploration of complex reaction spaces with unprecedented accuracy.

Droplet microfluidics has emerged as a powerful tool across diverse scientific disciplines, with dedicated literature offering concepts behind droplet formation [41,42]. An example of a segmented flow droplet system was employed to screen a range of organic solvents to obtain optimal conditions for the monoalkylation of *trans*-1,2-diaminocyclohexane [43]. The HTE methodology in combination with feedback DOE facilitated the rapid identification of appropriate solvents. Notably, the use of DMSO, DMF, and pyridine led to an enhanced yield of the monoalkylated product. An experimental setup was developed for single-droplet studies of visible-light photoredox catalysis using an oscillatory flow strategy [44,45]. In an oscillatory reactor, an alternating pressure gradient is applied within the reactor, causing a back-and-forth oscillation of the reaction slugs, which leads to higher control in mixing and an extended residence time of the reaction mixture. About 150 reaction conditions were explored, using a total volume of 4.5 mL reaction mixture, and the screening results can be readily translated to continuous flow synthesis. The application of segmented flow or microslug reactors was demonstrated in the decarboxylative arylation cross-coupling reaction promoted by catalysts and light [40]. The design allows the screening to be more material- and time-efficient in the optimization of both continuous variables (e.g., temperature and residence time) and discrete variables (e.g., catalyst, base). Pieber et al. [46] reported the application of a segmental flow reactor for heterogeneous solid–liquid reactions. In their report, they described the reaction slugs as serial microbatch reactors (SMBRs) separated through gas segments that incorporated liquid reagents and solid photocatalysts in a continuous flow. The slugs were generated by establishing a stable gas–liquid segmented-flow pattern using a Y-shaped mixer, followed by the suspension of the catalyst via a T-mixer. This technology was utilized to develop selective and efficient decarboxylative fluorination reactions. Recently, a slug flow platform was developed (Figure 3a) by injecting segments of gas as a separating medium for enhancing the optimization of the Buchwald–Hartwig amination intermediate, which is crucial for synthesizing the drug olanzapine [47]. The reactor setup was integrated with spectroscopic and chromatographic in-line analytical tools, enabling real-time monitoring of products and reaction intermediates. A detailed discussion on the optimization strategy is described in the section Machine-learning-driven optimization of chemical reactions.

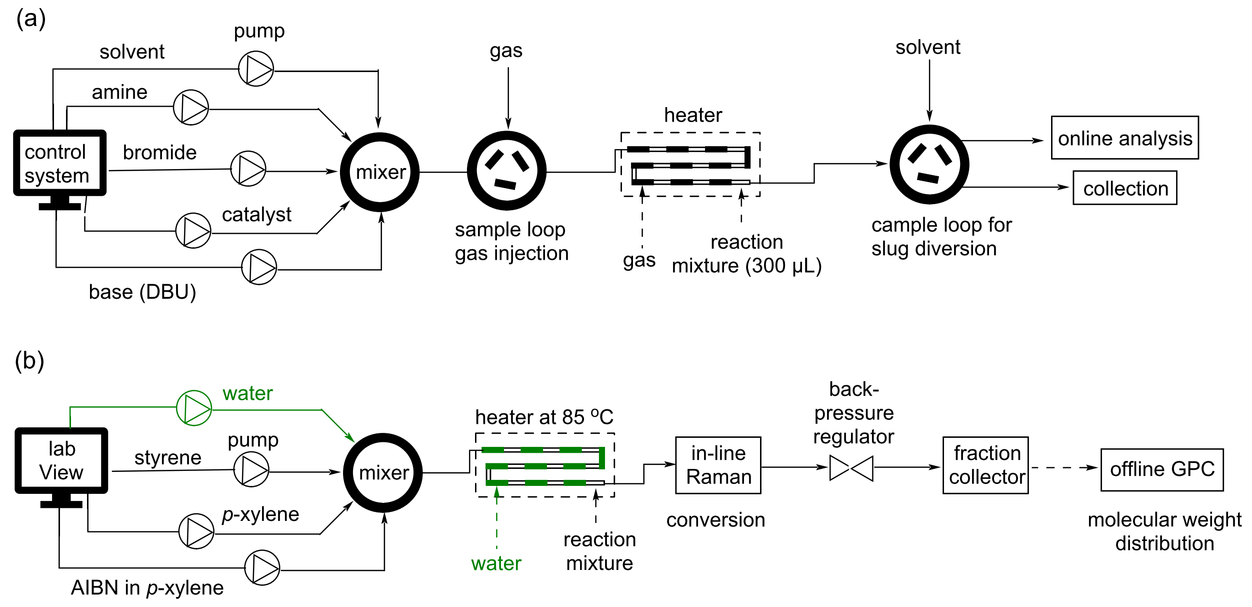


Figure 3: (a) Description of a slug flow platform developed using segments of gas as separation medium for high-throughput data collection in a Buchwald–Hartwig amination. A six-way mixer was used to mix the solvents and reagents. (b) Schematic representation of a computer-controlled segmented flow pattern developed using degassed water as an antisolvent for the high-throughput polymerization of styrene in *p*-xylene. A staggered infusion of organic and aqueous phases resulted in the exploration of a wider parameter space.

Robochem, a HTE platform, was designed to streamline the screening of photochemical reactions, facilitating the rapid generation of diverse reaction mixtures, each comprising 650 μL within a slug flow reactor [48]. This innovative system features precise monitoring of the reaction slug through a dedicated array of phase sensors and an algorithm designed for detecting its passage. As a result, the workflow delivers a notable boost in productivity, surpassing traditional batch reactions by over a 500-fold and outperforming flow reactions with a five-fold improvement. A fully integrated automated multistep chemical synthesizer (AutoSyn) was reported to be able to autonomously synthesize milligram- to gram-scale amounts of any organic or drug-like molecule [49]. The system comprised of a flow chemistry synthesis platform, a reagent delivery system, a packed bed reactor, process-analytical tools, and an integrated software control system that automates end-to-end process operations and monitoring. The system has been used to demonstrate the synthesis of at least ten drug molecules autonomously, and it does not include a closed-loop optimization framework. The Pfizer research team developed a custom-designed flow system for rapid reaction screening of the Suzuki–Miyaura coupling reaction on a nanomolar scale [50]. The platform included a modified HPLC system that supplied a flowing stream of 12 selectable solvents, an autosampler that injected microliter amounts of preselected reaction mixtures, and an LC–MS device for product characterization. Approximately 5,760 reactions were screened across a selection of 11 ligands, seven bases, and four solvents, along with appropriate control experiments being performed. The nanomolar droplet system enabled a very high throughput, exceeding 1,500 reactions every 24 hours. This extensive and intelligent screening approach identified optimal conditions for scaling up selected reactions to 50–200 mg under batch and flow conditions.

In addition to organic synthesis, the slug flow methodology has found application in polymer synthesis. A flow platform capable of polymerizing 397 unique copolymer compositions was developed by Reis et al. [51] using a droplet flow reactor. The methodology and high-fidelity data enabled them to discover more than ten copolymer compositions of promising ^{19}F MRI agents that outperformed state-of-the-art materials. A rapid generation of copolymer libraries was achieved by forming a droplet flow in an automated HTE flow setup [52]. This approach not only assists in overcoming viscosity challenges in conventional photopolymerization reactions but also helps to identify structure–property relationships for copolymer libraries. We have generated a segmented flow pattern (Figure 3b) by alternating the infusion of organic components and degassed water to create nine different compositions [53]. The organic components consisting of styrene, α,α′-azobisisobutyronitrile (AIBN), and *p*-xylene were infused using a computer-controlled segmented-flow platform. These approaches allow the compartmentalization of reaction mixtures without cross-contamination and enhance experimental throughput significantly. The concept of parallel flow reactors, where several distinct reactions conditions are tested simultaneously, has been proposed as a pathway to increase the throughput of flow reactions. Ahn et al. [54] designed and fabricated a complete prototype equipped with a unique built-in flow distributor (Figure 4) and 16 microreactors capable of executing diverse conditions in parallel, including photochemistry. The temperature of the capillary reactors can be controlled independently, providing flexibility in experimentation. The reservoir-type distributor, featuring a baffle structure, not only ensures uniform flow of reagents even when one or more reactors experience clogging but also allows for variation of the residence time of individual capillary reactors. The authors demonstrated the capabilities of their platform by executing 12 distinct reactions, which encompassed six different types of chemical transformations based on diazonium chemistry, in parallel (Figure 4). A total of 96 reaction conditions were tested, leading to optimized reaction parameters in less than an hour.

Chatterjee et al. [55] introduced the concept of radial synthesis to perform multiple single-step chemical reactions or to decouple multistep reactions into parallel processes. Individually accessible reactors are arranged around a central switching station that enables the delivery of independent reaction mixtures or reagents. Each reactor loop functions as an independent unit to carry out thermal or photochemical reactions under different conditions. This parallel reactor setup was successfully utilized for the multistep synthesis of 18 compounds of an anticonvulsant drug, employing various reaction pathways to perform photoredox carbon–nitrogen cross-coupling reactions. A parallel droplet flow system was developed by Eyke et al. [56] to significantly increase the throughput of reaction screening. A closed-loop Bayesian optimization (BO) framework was integrated to optimize reactions involving both continuous and categorical variables. The team upgraded the oscillatory droplet reactor platform to a high-throughput version consisting of multiple independent parallel reactors. This paral-

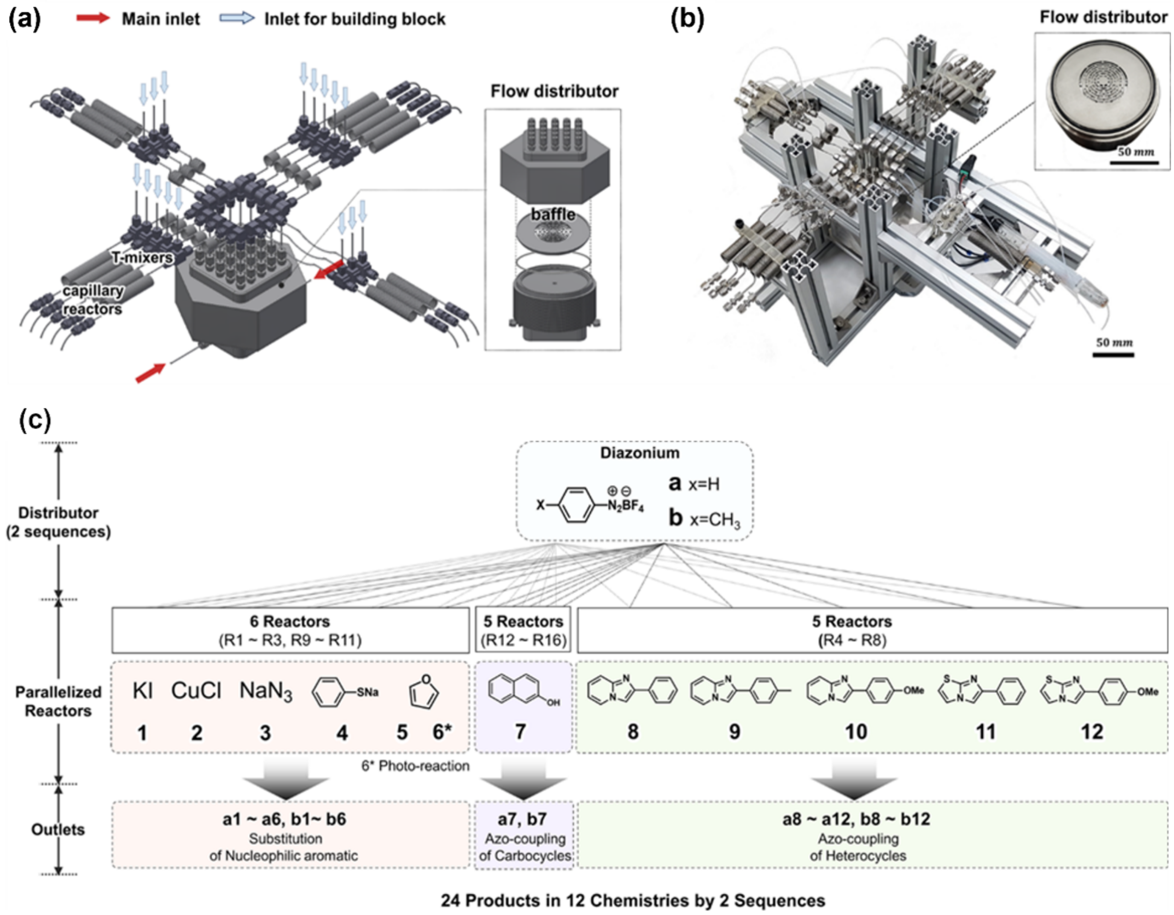


Figure 4: Schematic representation (a) and photograph (b) of the flow parallel synthesizer intelligently designed for screening optimal conditions for diazonium chemistry (c) for 24 products within two sequences. Figure 4 was reproduced from [54] (© 2021, G.-N. Anh et al., published by Springer Nature, distributed under the terms of the Creative Commons Attribution 4.0 International License, https://creativecommons.org/licenses/by/4.0).

lelization enables the collection of high-fidelity data for reaction kinetics and optimization for at least six different chemical reactions. The major bottleneck in HTE synthesis lies in the challenge of isolating and purifying reaction products once experiments are performed. Despite this bottleneck, the landscape is evolving, with various practical tools emerging to streamline purification processes. From prepacked silica gel tubes to the precision semipreparative liquid chromatography and the versatile capabilities of various scavenger resins, laboratories are witnessing a surge in options for efficient high-throughput purification, particularly in chemical synthesis on a modest scale. A change in thinking beyond conventional purification methods presents an opportunity to revolutionize HTE flow platforms. A completely novel design, differing from established isolation and separation techniques, holds the promise of not only enhancing the efficiency of HTE flow synthesis but also paving the way for more sustainable growth in this research area.

Autonomous self-optimizing flow reactors

Autonomous self-optimizing flow reactors (ASFRs) represent a promising advancement in the process optimization of chemical reactions. ASFR combines principles of automation, AI, in-line analytics, and robotics to streamline and accelerate the process optimization workflow. ASFRs enhance the yield and throughput of synthesis by minimizing waste. Engaging in-line/online analytics and integrating them with flow systems is relatively straightforward. The real-time processing of analytical data allows for immediate adjustments to the reaction parameters, enabling the attainment of optimal solutions rapidly. Consequently, the process can lead to lower energy consumption and reduced use of hazardous materials, contributing to more sustainable chemical processes. Integrating ML algorithms to simultaneously optimize multiple parameters such as yield, purity, and cost within a closed-loop represents a significant advancement in process design. Furthermore, automation in ASFRs reduces the need for constant human oversight, lowering operational costs and minimizing the risk of human errors. A schematic representation of ASFR is provided in Figure 5a.

A self-optimizing microreactor system has been devised specifically for closed-loop optimization of the Heck reaction, employing a "black-box" optimization strategy directed by Nelder–Mead simplex method algorithm [57]. In-line HPLC analysis was performed to determine the product yield in real time and give feedback to the control system to direct the input conditions to achieve the optimum product yield in 19 automated experiments. The optimum conditions for the formation of the monoarylated product **1** (Figure 5b) identified in a

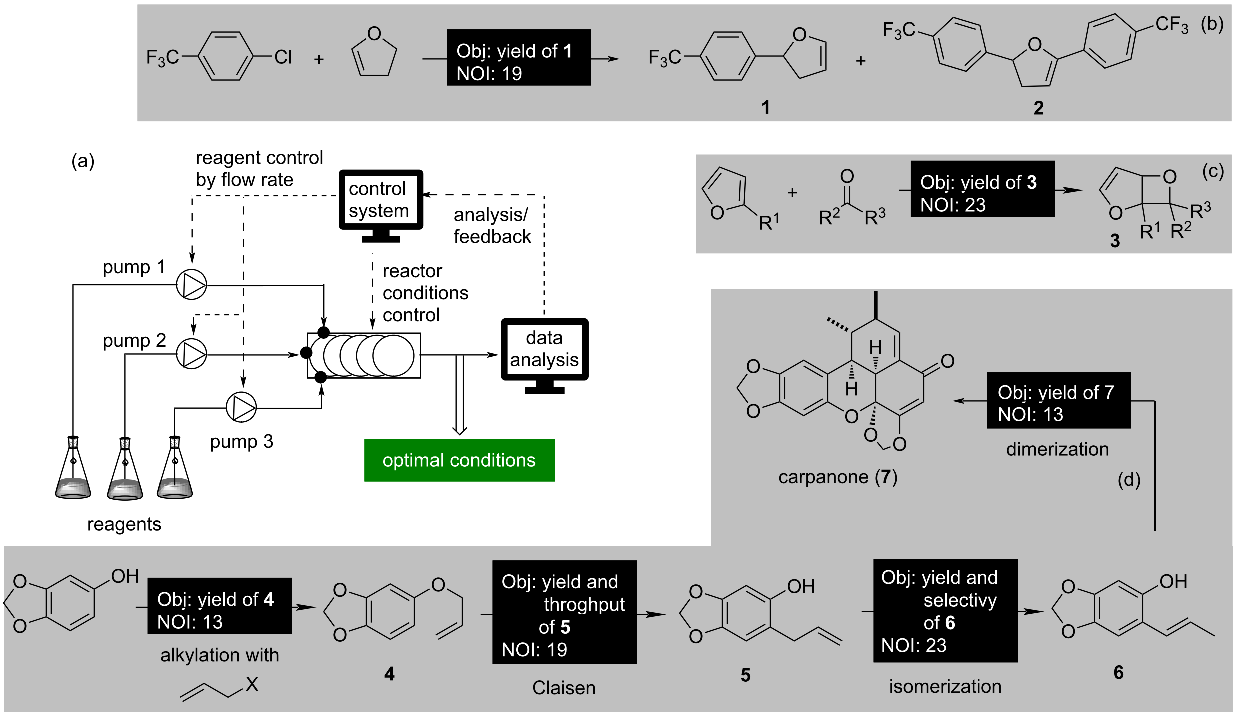


Figure 5: (a) Schematic representation of an ASFR for obtaining an optimal solution with minimal human intervention. Selected case studies (b–d) with closed-loop optimization are provided. Abbreviations “Obj” and “NOI” represent “objective functions” and “number of interactions”.

microfluidics system were successfully translated to a mesofluidics system on a 50-fold scale to afford 26.9 g of the product **1**. LeyLab, a modular software system developed by Fitzpatrick et al. [58], allows researchers to oversee chemical reactions online. The hydration of 3-cyanopyridine to an amide was monitored by online MS, offering real-time conversion insights. Through 30 experiments within ten hours, five key reaction parameters were finely tuned for optimal conditions.

Photochemical reactions require uniform light penetration of the reaction mixture, and flow setups with uniform path lengths would be ideal for such reactions. A self-optimizing continuous-flow reactor was designed by Poscharny et al. [59] for [2 + 2]-cycloaddition reactions promoted by light. The optimization (modified simplex) algorithm elaborated the optimal conditions within 25 iterative experiments to afford compound **3** (Figure 5c) in good yield. A modular autonomous flow reactor controlled via MATLAB was designed for the synthesis of carpanone (**7**, Figure 5d) using a modified Nelder−Mead algorithm [60]. The four-step process involves allylation, Claisen rearrangement, isomerization, and oxidative dimerization. Each reaction step was optimized independently by using either online HPLC or in-line benchtop NMR spectroscopy to afford an overall yield of 67% in 66 iterative experiments over four linear reaction steps. Nandiwale et al. [61] reported the autonomous optimization of three multiphase catalytic reactions involving the handling of solid substrates, operating the photoreactor, and feeding of the slurries, catalysts, and inorganic bases in an automated flow platform comprising a continuous stirred tank reactor (CSTR) cascade. The platform allowed to showcase the autonomous optimization to find the ideal reaction conditions for Suzuki–Miyaura and photoredox-catalyzed coupling reactions.

A plug-and-play, continuous-flow chemical synthesis system (Figure 6a) was intelligently designed by Bédard et al. [62] to mitigate some of the challenges in traditional organic synthesis by the integration of hardware, software, and analytics. Comprising an array of modular components, including units for heating, cooling, LED light exposure, and packed bed reactors, it provides a flexible platform for various reaction categories. The system consists of a liquid–liquid separator and an in-line/online analytical tool to facilitate closed-loop autonomous optimization. The capability of the system was demonstrated in the optimization of C–C and C–N cross-coupling, olefination, reductive amination, photoredox-catalytic, and nucleophilic aromatic substitution reactions, as well as in the two-step synthesis of cyclobutanone. The molecules synthesized under the optimal conditions are presented in Figure 6b, employing the stable noisy optimization by branch and fit (SNOBFIT) algorithm. SNOBFIT offers a convenient methodology for global optimization, eliminating the necessity of a theoretical model. A

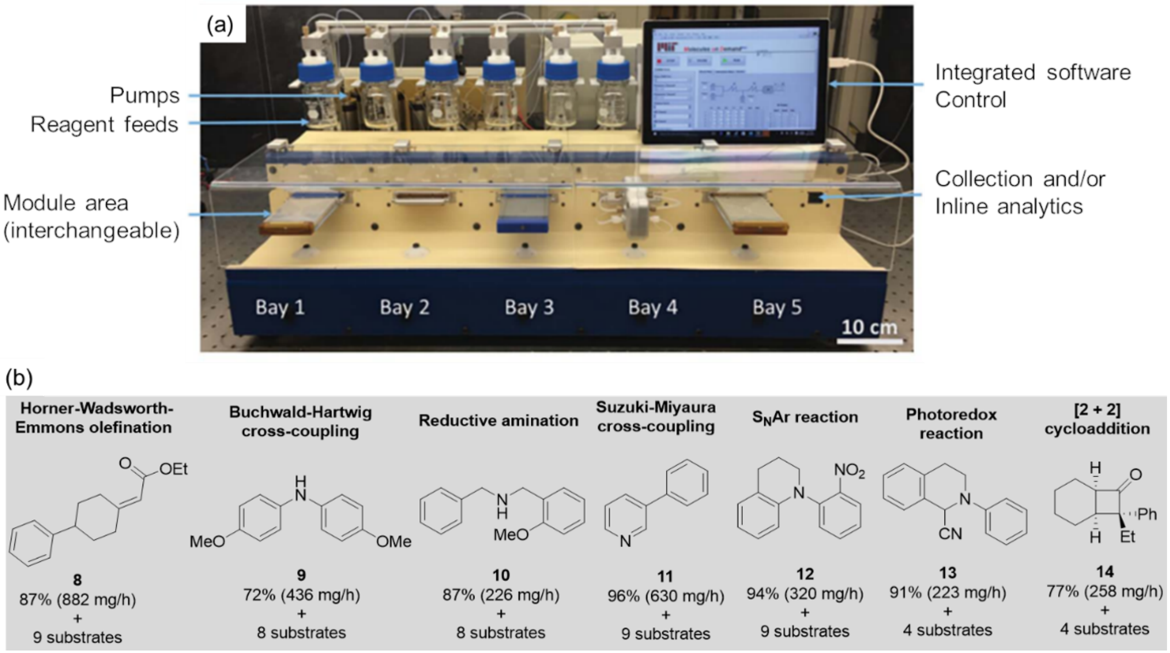

(a)
Pumps
Reagent feeds
Module area (interchangeable)
Integrated software Control
Collection and/or Inline analytics
Bay 1
Bay 2
Bay 3
Bay 4
Bay 5
10 cm
(b)
Horner-Wadsworth-Emmons olefination
Buchwald-Hartwig cross-coupling
Reductive amination
Suzuki-Miyaura cross-coupling
S_NAr reaction
Photoredox reaction
[2 + 2] cycloaddition
OEt
O
H
N
MeO
OMe
N
H
MeO
N
N
NO_2
N
CN
H
O
Ph
H
Et
8
87% (882 mg/h) + 9 substrates
9
72% (436 mg/h) + 8 substrates
10
87% (226 mg/h) + 8 substrates
11
96% (630 mg/h) + 9 substrates
12
94% (320 mg/h) + 9 substrates
13
91% (223 mg/h) + 4 substrates
14
77% (258 mg/h) + 4 substrates

Figure 6: (a) A modular flow platform developed for a wider variety of chemical syntheses. (b) Various categories of chemical reactions optimized and molecules synthesized in a continuous flow system are given. Figure 6a is from [62]. Reprinted with permission from AAAS. This content is not subject to CC BY 4.0.

reconfigurable automated flow platform integrating online HPLC monitoring was used for the cobalt-catalyzed aerobic oxidative dimerization of desmethoxycarpacine (**6**) to carpanone (**7**) in the presence of oxygen as an oxidant [63]. A gas−liquid segmented or a tube-in-tube strategy was adopted to achieve a higher yield within a shorter residence time. Substantial further developments have been made in applying ASFR in multiobjective optimizations, which will be discussed in detail below in the section "Machine-learning-driven optimization of chemical reactions".

Real-time analytics and high-throughput data processing

Real-time analytics play a critical role in the optimization of chemical reactions via high-throughput synthesis and ML algorithms. Process-analytical technology (PAT) tools empower researchers to obtain chemical insights from a large number of experiments, facilitating the precise measurement of optimization targets. The integration of real-time analysis in HTE presents a multitude of advantages over traditional, one-time final product evaluations, as outlined below:

(i) Real-time analysis facilitates rapid decision-making, enabling researchers to continuously monitor and analyze data as it is generated and allows for immediate adjustments to process parameters during experiments.

(ii) Early detection of trends or anomalies are made possible through real-time analysis, providing valuable insights that can guide subsequent experiments and inform iterative improvements and optimizations in experimental protocols.

(iii) By optimizing experimental workflows and minimizing waste through real-time analysis, researchers can allocate resources more efficiently, ensuring that resources are utilized effectively to maximize experimental outcomes.

(iv) Enhanced experimental control on the process to deliver constant product quality to meet desired specifications and standards.

(v) By providing instantaneous feedback, real-time analysis accelerates the optimization process, reducing the experimentation time and expediting the discovery of optimal reaction conditions with minimum material use.

Analytical tools are integral components of high-throughput platforms and are found in various configurations, such as in-line, online, at-line, and offline, contingent upon their placement within the experimental workflow. In Table 1, we describe the subtle disparities for clarity and reference.

Self-optimizing HTE throughput platforms require in-line and/or online characterization as well as data analysis and processing for rapid optimization of organic reactions. Chromatographic (i.e., HPLC, GC) and spectroscopic (e.g., NMR, FTIR, UV–vis, Raman) characterization methods are commonly used in real-time reaction monitoring. To quantify the products of a chemical reaction, a calibration curve is required before the optimization campaign. The following sequential steps are typically employed to refine raw data into actionable inputs for building ML models for optimization: (i) extraction and categorization of appropriate spectra; (ii) fitting of spectral peaks utilizing predefined functional models, alongside deconvolution of overlapping signals; (iii) consolidation of extracted peak information and generation of relevant data plots; and (iv) extracting the relevant information and formatting into input data for ML models. A recent review by Felpin and Rodriguez-Zubiri [64] highlighted the selection of in-line/online analytical tools that can be integrated into flow reactors for the monitoring of chemical reactions. In the current review, we focus on the high-throughput data processing that complements the HTE platforms for rapid optimization of organic reactions. Although

Table 1: Different analytical methods depending on their position within the experimental workflow.

analytical method	description
in-line	Analyzed in real time during the reaction or production process by directly integrating appropriate devices.
online	Sampling and analysis take place while the reaction or process is running. The analysis is done on a device located nearby. Online analyses can be carried out continuously or at set intervals. Autonomous sampling allows for direct online analysis with the instrument.
at-line	Like online analysis, the samples are analyzed usually within a manufacturing facility. An aliquot of the reaction mixture is taken for analysis. Human intervention is often required for this task. At-line analysis still provides relatively rapid results compared to offline methods, offering a balance between real-time monitoring and convenience.
offline	Analysis conducted outside of the process environment and separate from ongoing operations. Provides a more detailed and comprehensive analysis compared to real-time monitoring.

multivariate data analysis has frequently been adopted in analytical chemistry for rapid data processing, the availability of relevant open-source code is relatively low [65,66]. Consequently, the development of open-source code for data processing is interesting for the scientific community.

Jansen et al. [67] have developed HappyTools, a tool for the analysis of HPLC measurements, able to calibrate retention time, perform peak quantification, and use various quality criteria to curate the compiled data. For the quantification and calibration of chromatographic peaks, the user can either input a peak list containing the retention time window of the target chemicals, or the tool can use an automated peak detection algorithm, removing the need of user input. The peak detection algorithm was developed using a loop to attain the user-specified cut-off value of the highest-intensity peak. A new univariate spline is fitted for each iteration, from which the local maxima and minima are determined. Overall, HappyTools showed similar or better performance in comparison to existing commercial software. In particular, HappyTools showed an enhanced throughput, demonstrating up to a ten-fold reduction of the total processing time for biopharmaceutical samples. The authors have released the source code and an executable program in an online repository to be employed freely for research purposes.

In addition to HappyTools, there are other available open-source Python packages to analyze chromatographic and spectroscopic data. A cross-platform Python package named Aston can be used to process both UV–vis and MS data. The open-source library is written using Python, NumPy, and SciPy and is openly hosted in an online repository [68]. Similarly, for processing chromatographic data from GC–FID, HPLC–UV, or HPLC–FD, packages are also available open source. Embedding these codes into HTE and ML workflow dramatically improves the efficiency and speed of the optimization processes. Liu et al. [69] developed a custom-built Python script to study the kinetics of carbonyldiimidazole-mediated amide formation by analyzing data from online HPLC and in-line FTIR-spectroscopic measurements. Their algorithm was able to automatically detect peaks from chromatographic spectra and to automatically assign the peaks to reagents or products depending on the decrease or increase in peak intensity over time. In addition to monitoring the evolution of the reaction, the IR spectral data was processed in real time. This was to ensure the complete consumption of acid reactant and to feed this information back to the pump for immediate quenching of carbonyldiimidazole to prevent any side reactions. The entire process allows to control the acid activation and amide formation precisely to afford the desired final product in quantitative yield.

Recently, Sagmeister et al. [70] assembled four complementary PATs, including in-line NMR, UV–vis, IR, and online ultra-high-performance liquid chromatography (UHPLC) to meticulously monitor the intricate three-step linear synthesis of the drug mesalazine (**18**, Figure 7) with a 1.6 g·h^{-1} throughput. In the first step, the nitration reaction was monitored by in-line NMR. The overlapping peaks were resolved for accurate quantification by building a chemometric model. The model also allowed for flexibility to small changes in peak positions and shapes in repetitive analyses. An in-house-designed flow cell equipped with a reflectance probe was employed for real-time monitoring of hydrolysis by in-line UV–vis spectroscopy. The raw data was processed using a sophisticated neural network algorithm, yielding rapid quantification with an impressive processing time of 1.4 ms per spectrum. This streamlined approach ensured efficient and timely data analysis, facilitating seamless real-time monitoring of the hydrolysis of **16**. The final hydrogenation step was monitored by an in-line IR probe. The spectral data was processed using a partial least squares regression

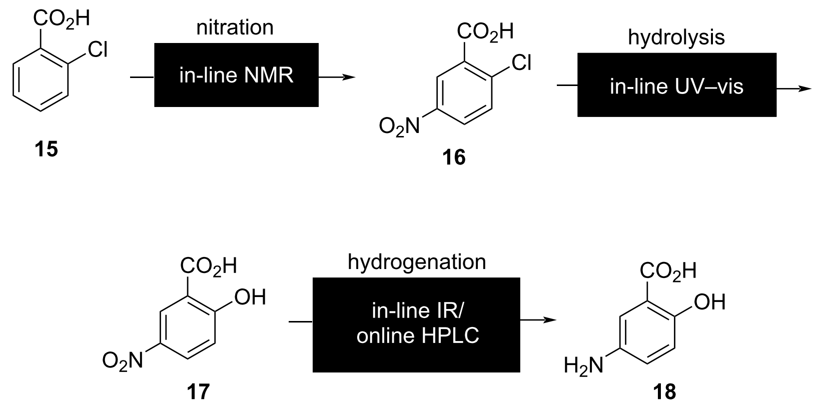


Figure 7: Implementation of four complementary PATs into the optimization process of a three-step synthesis.

model and quantified. An online UHPLC was used to analyze the final composition of the reaction mixture after three reaction steps. The integration of all PAT tools into the three-step reaction was carefully executed with an open platform communications unified architecture (OPCUA) platform for interplatform equipment communication. The adoption of the OPCUA platform ensured seamless communication between different equipment platforms for enhanced efficiency and accuracy in data analysis.

A recent study introduced a novel approach for directly processing and analyzing HPLC−DAD raw data using Python [71]. This method leverages the Multivariate Online Contextual Chromatographic Analysis (MOCCA) package, designed for integration into both automated and manual workflows. MOCCA offers a range of benefits, including automated management of internal standards for precise relative quantification, reliable peak assignments, accelerated sample processing, and efficient deconvolution of overlapping peaks. Its versatility was showcased through the successful completion of four comprehensive case studies, demonstrating its broad applicability across diverse analytical scenarios. Recently, we implemented in-line Raman spectroscopy to monitor the real-time conversion of styrene to polystyrene, utilizing a custom Python package developed in-house [53]. This approach enabled us to track the conversion process at different residence times. Specifically, we quantified the conversion by analyzing the area under the curve (AUC) of the Raman-active vibrational modes associated with the styrene–vinyl C=C stretch (≈1630 cm^{-1}), which we calibrated against signals from *p*-xylene (≈830 cm^{-1}). To resolve overlapping peaks, we employed curve-fitting techniques utilizing Lorentzian functional forms, facilitated by the lmfit Python package. This methodology (Figure 8) allowed us to accurately calculate conversion rates and to make precise predictions using ML models. Traditionally, the optimization of a chemical reaction, the development of kinetic models, and optimization of analytical characterization parameters are undertaken independently. With this approach, many overlapping tasks are performed in parallel, thus leading to long lead times and inefficient personnel allocation. To overcome these issues, Sagmeister et al. [72] developed a dual modelling approach using a single platform that seamlessly integrates the calibration of PAT, reaction optimization, kinetic modelling, and parametrizes a process model for scale-up within approximately eight hours. Their platform consisted of a flow reactor connected to an in-line FTIR spectrometer. In addition, the platform has two valves that allow a stream of reagents or target product to bypass the reactor coil directly into the in-line FTIR

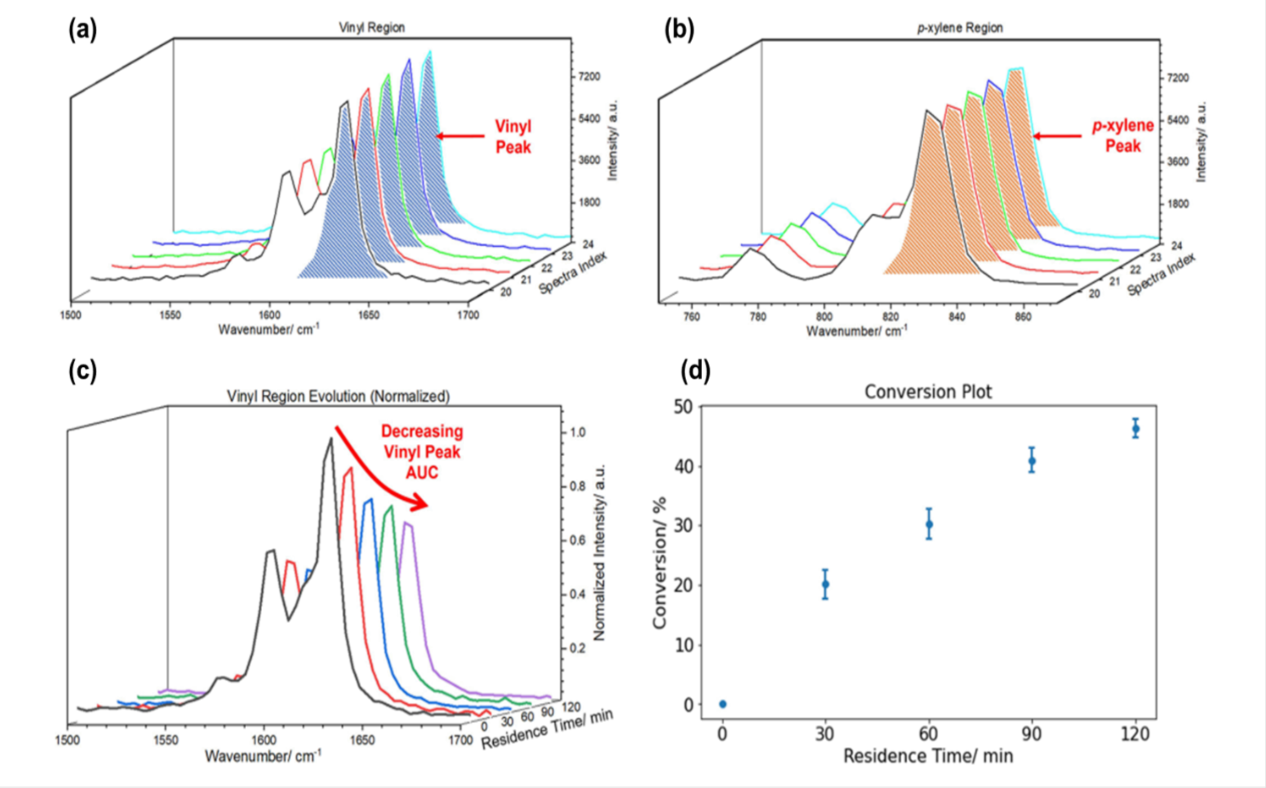


Figure 8: Overlay of several Raman spectra of a single condition featuring the styrene vinyl region (a) and the *p*-xylene region (b). (c) Waterfall plot depicting the decrease in the vinyl peak AUC over time. (d) A representative conversion plot shows an increasing conversion with residence time. Figure 8 is adapted from [53]. Figure 8 was reprinted with permission from [53], Copyright 2023 American Chemical Society.

spectrometer. Using this configuration, the platform can perform a calibration of the reagent and product concentration through a standard addition method. Once the PAT is calibrated, the platform performs dynamic experiments where the concentration of the reagents is ramped to explore the parametric space. Finally, using a scientific programming language called Julia, the collected data can be fitted to the kinetic model parameters, and in silico optimization of the reaction parameters can be performed.

Machine-learning-driven optimization of chemical reactions

Historically, optimization of chemical reactions has been performed based on DOE methodologies, with the objective of maximizing the yield of the reaction product. However, these techniques are not well suited to find the global optimal conditions and scale exponentially with the number of variables. Computational approaches that rely on optimization algorithms offer more efficient ways to obtain the optimal conditions, without requiring an exponential number of experiments per variable to be optimized. Early examples of chemical reaction conditions optimization through computational approaches focused on the application of black-box optimization algorithms, such as steepest descent, SNOBFIT, and Nelder–Mead simplex, which demonstrated positive results and the ability to perform self-optimizing automated workflows with little human intervention [57,58,60,62,73-75]. In recent years, ML optimization methods have demonstrated the ability to obtain optimal reaction conditions within a reduced number of experiments in comparison to human intuition, traditional DOE, and other black-box optimization algorithms [2,76,77]. Unlike traditional optimization algorithms, the ML approach focuses on building predictive surrogate models for objective functions. These models learned the relationships between the reaction conditions and the target optimization objectives based on experimental data. In a second step, these models are efficiently probed to identify the most promising values for optimizing the objective function. In this section, we review the latest developments in ML optimization strategies for the optimization of chemical reactions.

Figure 9a outlines the basic steps for the optimization of chemical reactions using ML methods. The workflow requires an initial set of experimental data that contains different variables for reaction conditions (i.e., temperature, time, solvent, catalyst, etc.) and the corresponding outcome values for the target optimization objectives (e.g., yield, purity, cost, etc.). The initial dataset is commonly obtained by sampling a combination of reaction variables from the parametric space, performing the synthetic experiments under the selected reaction conditions, and measuring the values for the target optimization objectives. The sampling of the initial reaction variables is often performed through near-random statistical methods, such as Latin hypercube sampling (LHS), Sobol sampling, full factorial sampling, and centerpoint sampling methods. Alternatively, the initial dataset can be obtained from values previously reported in the literature. After that, one or various predictive models are fitted to the initial dataset to predict the expected values of the optimization objectives. The number of models that are fitted depends on the number of optimization objectives, and normally one model is constructed for each optimization objective. The next step involves the application of an optimization algorithm to find the parameters that would most likely lead to optimal outcomes of the target optimization objectives. Finally, a set of the most promising suggestions is selected and tested experimentally. The dataset is then updated with the outcomes of the latest experimental parameters, and the process is repeated until the optimal conditions have been found. Depending on the number of objectives, optimization campaigns are classified as single-objective (Figure 9b) or multiobjective optimizations (Figure 9c). In single-objective optimizations, the algorithm will explore the parametric space to determine the optimal conditions by finding the variables that either maximize or minimize the target objective function. In multiobjective optimizations, the algorithms will search for optimal conditions that either maximize or minimize each objective function. On the other hand, when competing objectives are optimized, the algorithm aims to discover the set of solutions where the improvement of one objective results in the deterioration of the other. This set of solutions is called the Pareto front of the system (also known as nondominated solutions), and all other solutions that are not part of the Pareto front are not optimal for any of the objectives and are referred to as dominated solutions. Since all solutions in the Pareto front are optimal, the user is responsible for choosing the set of conditions for their specific application.

The first reports on the application of ML in the optimization of chemical reactions appeared over 20 years ago. A handful of studies used ML algorithms, such as neural networks and support vector machines, to fit models to chemical reaction data that were then optimized by genetic algorithms [78-80]. However, the use of ML for chemical reaction optimization did not become popular until the introduction of BO techniques by Lapkin and Bourne et al. [81]. BO is a global optimization method that fits a probabilistic function to model the objective function and utilizes it to search for parameters that will likely lead to optimal objective values. Commonly, BO uses a Gaussian process (GP) to create surrogate models that map the relationships between the variables and objectives (Figure 9a). Then, the surrogate model is sampled, and the output values are passed to an acquisition function that balances the surrogate

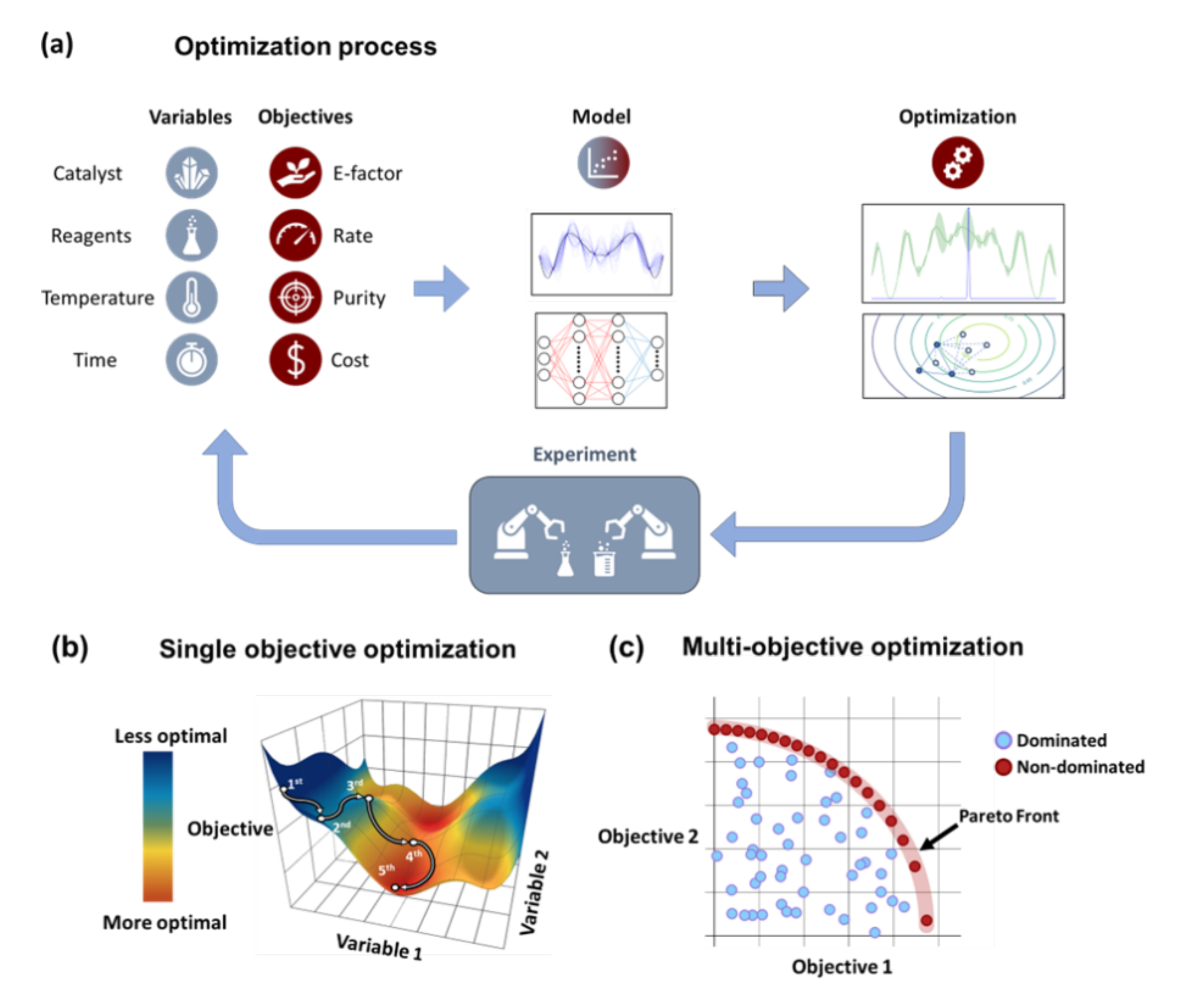


Figure 9: (a) Schematic description of the process of chemical reaction optimization through ML methods. (b) 3D representation of the objective function depending on two variables, showing the path of five optimization iterations that aim to minimize the value of the objective function. (c) Representation of the outcomes of a multiobjective optimization campaign. Each data point represents one experimental reaction condition. The Pareto front of the system, where the improvement of one objective leads to the deterioration of the other, is highlighted in red.

model predictions and uncertainties to find variable combinations that are likely to lead to optimal solutions (Figure 9a). The application of GPs and BO to optimize chemical reactions has the advantages of being able to model complex nonlinear relationships between multiple variables and of incorporating uncertainty into the predictions, making them suitable for the optimization of noisy and expensive evaluation functions.

Multiobjective optimization of chemical synthesis

Different BO algorithms can be implemented depending on the acquisition function used to evaluate the surrogate models and the strategies used to suggest the most likely optimal values for a target objective. Table 2 summarizes the use of various ML algorithms for the optimization of chemical syntheses with multiple objective functions. For chemical reaction optimization, the Thompson sampling efficient multiobjective optimization (TSEMO) algorithm has been the most widely used due to its capability to model noisy functions, efficient computation, and ability to model functions in the absence of any prior knowledge. The TSEMO algorithm utilizes a GP to model each objective function and utilizes an approach based on Thompson sampling to recommend the next set of conditions that maximizes the evaluated objective functions [82]. The use of TSEMO for the optimization of a chemical reaction was first reported by Schweidtmann et al. [81]. In this study, the multiobjective Bayesian optimization (MOBO) was used to optimize an S_NAr reaction (Table 2, entry 1) and an N-benzylation reaction (Table 2, entry 2) using an automated flow reactor. The objectives of the optimization were to maximize the space–time yield (STY) while minimizing either the E-factor of the S_NAr reac-

Table 2: Multiobjective optimization of synthetic organic case studies using ML methods and single-objective optimization of telescoped reactions.

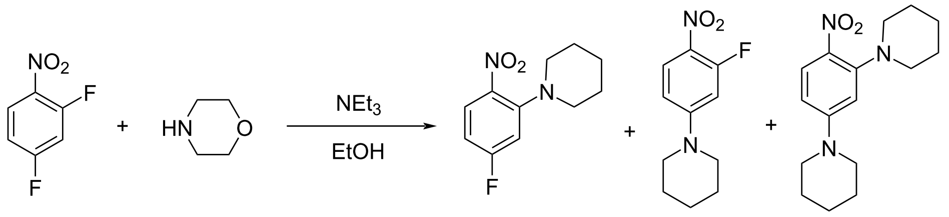
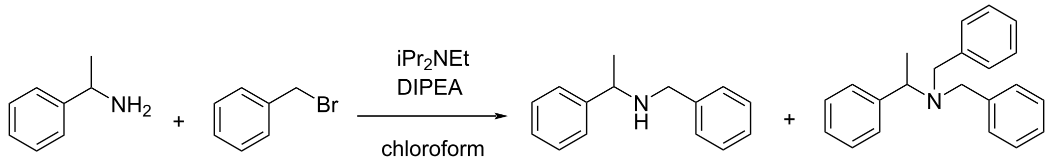
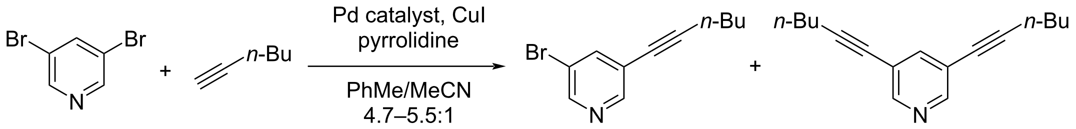
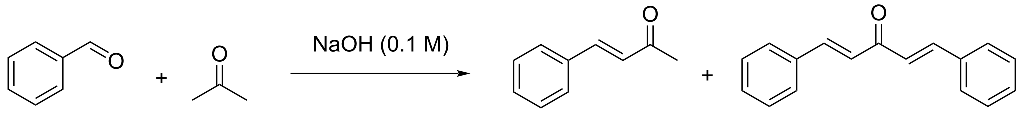
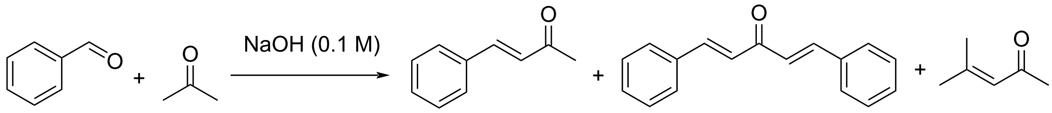
entry	platform	algorithm	variables	objectives	Refs.
1	 **19** **20** **21** **22** **23**				[81]
	flow	BO (TSEMO)	residence time equiv of **20** concn of **19** temp	↑ [a] STY of **21** ↓ [b] E-factor	
2	 **24** **25** **26** **27**				[81]
	flow	BO (TSEMO)	flow rate **24**/**25** ratio solvent temp	↑ STY of **26** ↓ yield of **27**	
3	 **28** **29** **30** **31**				[86]
	flow (CSTR)	TSEMO	residence time equiv of **29** temp	↑ STY of **30** ↓ yield of **31**	
4	 **32** **33** **34** **35**				[86]
	flow (CSTR)	TSEMO	flow of **32** equiv of **33** equiv of NaOH temp	↑ STY of **34** ↓ yield of **35** ↑ RME[c] of **32**	
5	 **32** **33** **34** **35** **36**				[87]
	flow	TSEMO	equiv **33** equiv of NaOH temp residence time	↑ yield of **34** ↓ cost ↓ E-factor	

Table 2: Multiobjective optimization of synthetic organic case studies using ML methods and single-objective optimization of telescoped reactions. (continued)

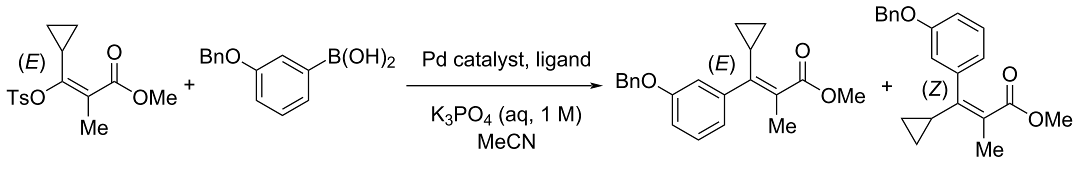
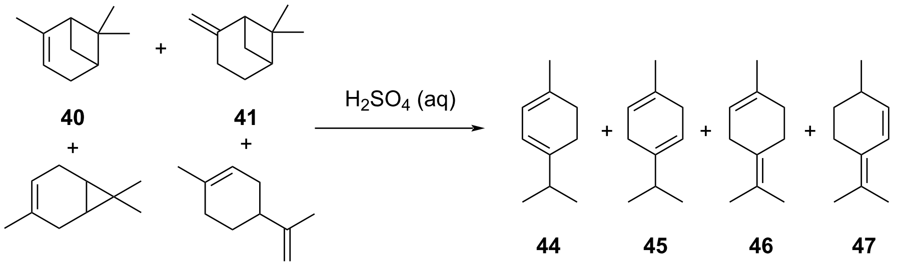
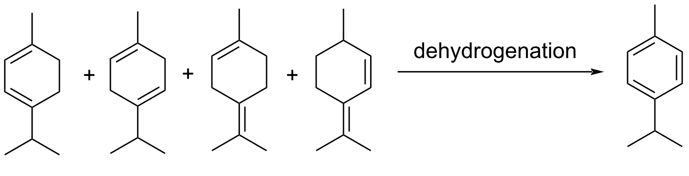
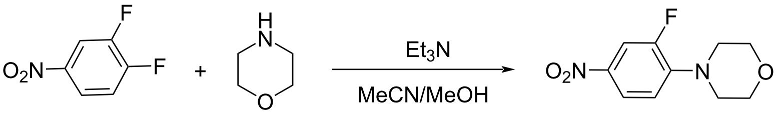
6	 (*E*) TsO OMe Me BnO B(OH)$_2$ Pd catalyst, ligand K_3PO_4 (aq, 1 M) MeCN (*E*) (*Z*) **37** **38** (*E*)-**39** (*Z*)-**39**				[15]
	batch	Phoenics Gryffin	ligand ligand/Pd ratio Pd loading equiv of **38** temp	↑ yield of (*E*)-**39** ↓ yield of (*Z*)-**39** ↓ Pd loading ↓ equiv of **38**	
7	 **40** **41** **42** **43** **44** **45** **46** **47**				[88]
	batch	TSEMO	temp concn of H_2SO_4 aqueous/organic phase ratio time equiv of **40** equiv of **41** equiv of **42** equiv of **43**	↑ conversion of **40**–**43** ↑ yield of **44**–**47**	
8	 **44** **45** **46** **47** **48**				[88]
	flow	TSEMO	temp air flow liquid flow time equiv of **44** equiv of **45** equiv of **46** equiv of **47**	↑ conversion of **44**–**47** ↑ yield of **48**	
9	 F O_2N F H N O Et_3N MeCN/MeOH F O_2N N O **19** **20** **21**				[83]
	flow	TSEMO	temp residence time concn of **19** equiv of **20** Et_3N	↑ conversion of **19** ↑ STY of **21** ↓ E-factor	

Table 2: Multiobjective optimization of synthetic organic case studies using ML methods and single-objective optimization of telescoped reactions. (continued)

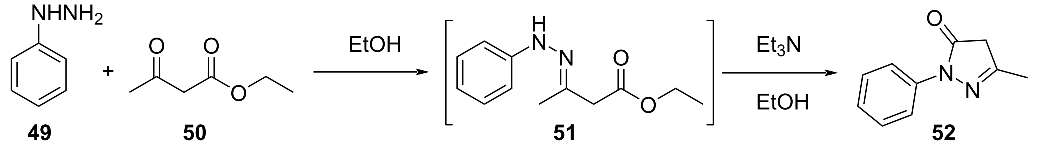
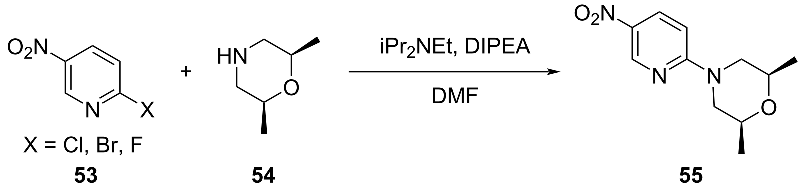
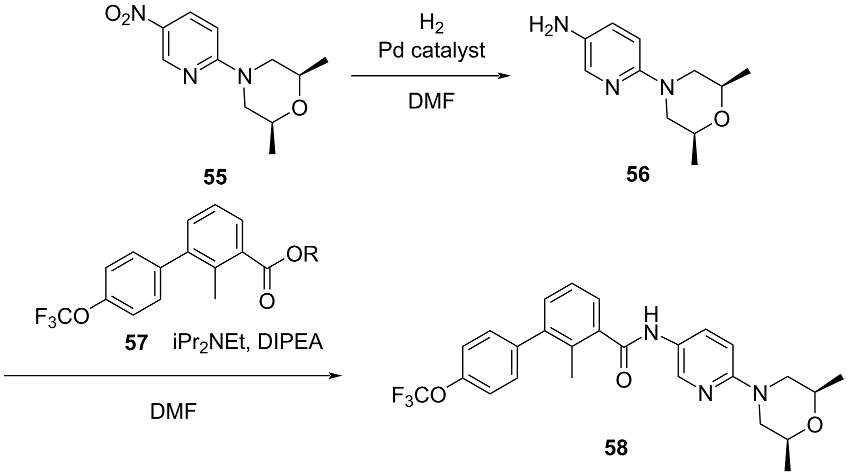
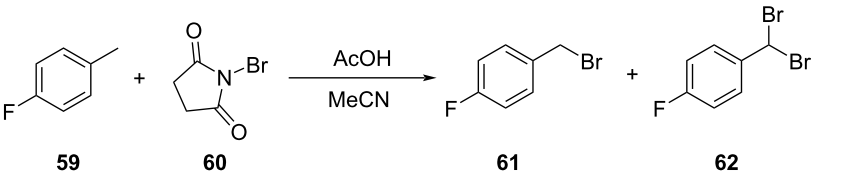
10	 NHNH$_2$ EtOH H N N O Et$_3$N EtOH O N N 49 50 51 52				[83]
	flow	TSEMO	equiv of **49** concn of **49** and **50** residence time of step 2 temp of step 1 temp of step 2 equiv of Et$_3$N	↑ yield of **51** ↑ STY of **52** ↓ equiv of **49** and Et$_3$N	
11	 O$_2$N N X X = Cl, Br, F 53 HN O 54 iPr$_2$NEt, DIPEA DMF O$_2$N N N O 55				[84]
	flow	Dragonfly	temp residence time equiv of **53** equiv of DIPEA leaving group X	↓ cost ↑ productivity of **55** ↑ yield of **55**	
12	 O$_2$N N N O 55 H$_2$ Pd catalyst DMF H$_2$N N N O 56 OR O F$_3$CO 57 iPr$_2$NEt, DIPEA DMF H N O N N O F$_3$CO 58				[84]
	flow	Dragonfly	activation time **55**/**57** equiv ratio temp of step 2 reactor vol substituent R	↑ yield of **58** ↑ productivity of **58**	
13	 F 59 O N Br O 60 AcOH MeCN Br F 61 Br Br F 62				[89]
	flow	TSEMO	equiv of **59** temp concn of **60** equiv of AcOH light intensity residence time	↑ STY of **61** ↑ conversion of **60** ↑ selectivity	

Table 2: Multiobjective optimization of synthetic organic case studies using ML methods and single-objective optimization of telescoped reactions. (continued)

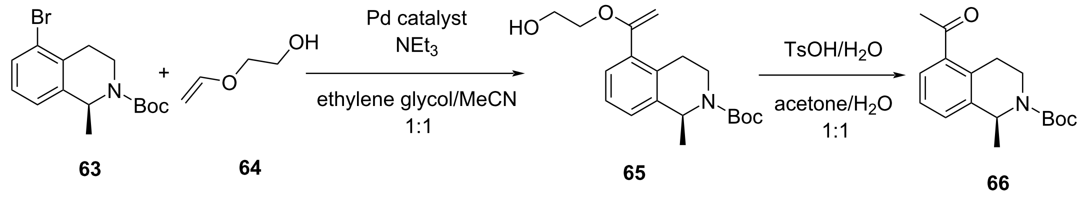
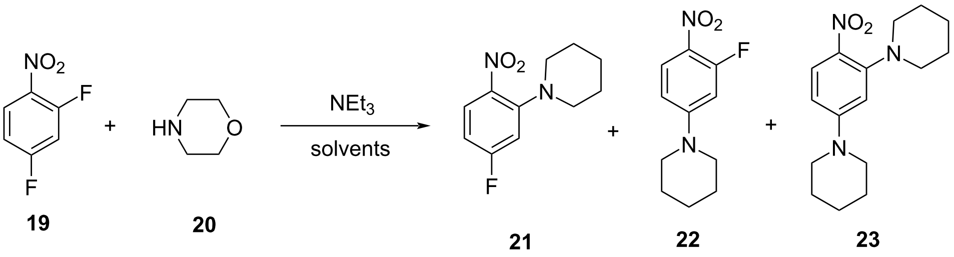
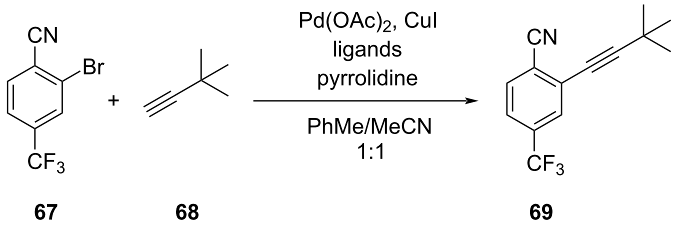
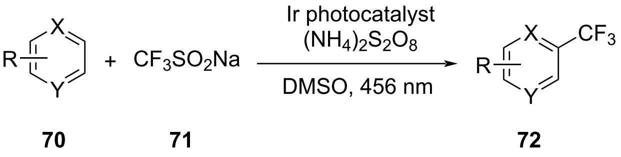
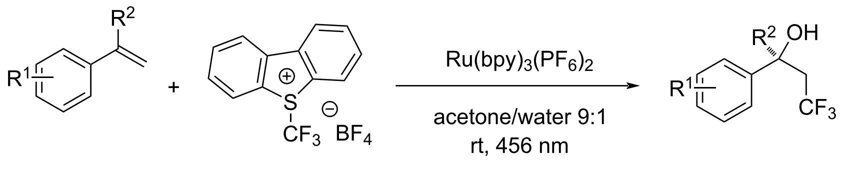
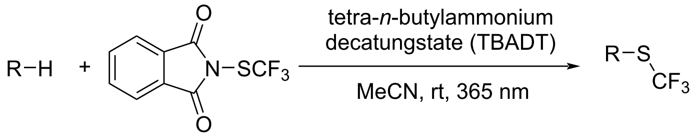
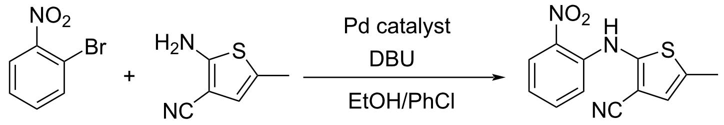
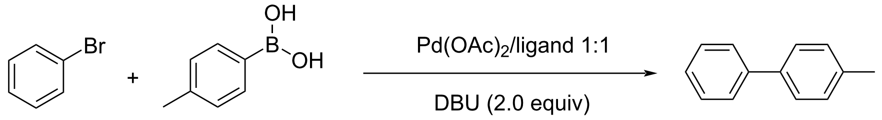
14	 Br N Boc + OH O Pd catalyst NEt₃ ethylene glycol/MeCN 1:1 HO O N Boc TsOH/H₂O acetone/H₂O 1:1 O N Boc 63 64 65 66				[85]
	flow	BOAEI[d]	residence time equiv of **64** temp equiv of TsOH	↑ yield of **66**	
15	 NO₂ F F + HN O NEt₃ solvents NO₂ N F + NO₂ F N + NO₂ N N 19 20 21 22 23				[90]
	flow	MVMOO[e]	solvents residence time concn of **19** equiv of **20** temp	↑ yield of **21** ↑ yield of **22**	
16	 CN Br CF₃ + Pd(OAc)₂, CuI ligands pyrrolidine PhMe/MeCN 1:1 CN CF₃ 67 68 69				[90]
	flow	MVMOO	ligands residence time equiv of **68** temp	↑ RME ↑ STY of **69**	
17	 R X Y + CF₃SO₂Na Ir photocatalyst (NH₄)₂S₂O₈ DMSO, 456 nm R X CF₃ Y 70 71 72				[48]
	photoflow reactor (Robochem)	BO	concn of **70** cat. loading concn of CF_3SO_2Na $(NH_4)_2S_2O_8$ loading residence time light intensity	↑ yield of **72** ↑ throughput	

Table 2: Multiobjective optimization of synthetic organic case studies using ML methods and single-objective optimization of telescoped reactions. (continued)

18	 R2 R1 CF3 BF4 Ru(bpy)3(PF6)2 acetone/water 9:1 rt, 456 nm R2 OH R1 CF3 **73** + **74** → **75** photoflow reactor (Robochem)	BO	concn of **73** concn of **74** cat. loading residence time light intensity	↑ yield of **75** ↑ throughput	[48]
19	 R−H N−SCF3 tetra-n-butylammonium decatungstate (TBADT) MeCN, rt, 365 nm R−S CF3 **76** + **77** → **78** photoflow reactor (Robochem)	BO	concn of **76** loading of **76** TBADT loading residence time light intensity	↑ yield of **78** ↑ throughput	[48]
20	 NO2 Br H2N S NC Pd catalyst DBU EtOH/PhCl NO2 H N S NC **79** + **80** → **81** slug flow reactor	TSEMO	residence time concn of **79** and **80** equiv of **80** temp equiv of DBU cat. loading	↑ yield of **81** ↑ STY of **81** ↓ cost	[47]
21	 Br OH B OH Pd(OAc)2/ligand 1:1 DBU (2.0 equiv) THF/H2O 5:1 **82** + **83** → **84** flow	ALaBO[f]	residence time cat. loading temp phosphine ligand	↑ yield of **84** ↑ turnover number	[91]

[a]Maximization. [b]Minimization. [c]Reaction mass efficiency. [d]Bayesian optimization algorithm with an adaptive expected improvement acquisition function. [e]Mixed-variable multiobjective optimization. [f]Adaptive latent Bayesian optimization.

tion or the impurity concentration of the N-benzylation reaction. For both reactions, there were four variables to optimize, including metrics for reaction time, reagent concentration, and temperature. After an initial sampling of 20 experimental conditions by LHS, the choice of reaction conditions was left to the TSEMO algorithm, optimizing the S_NAr within a total of 48 iterations and the N-benzylation reaction within a total of 58 iterations. Both optimizations resulted in the discovery of a dense Pareto front with approximately 30–50% of the total suggested conditions resulting in nondominated solutions. Since then, multiple reports have demonstrated the ability of TSEMO to optimize multiobjective optimizations for the synthesis of organic molecules (see examples in Table 2, entries 3–5 and 7–9). A particularly noteworthy development is the application of TSEMO for the optimization of synthetic routes composed of two and more successive reaction steps or telescoped reactions

[83-85]. Sagmeister et al. [83] reported the optimization of a two-step telescoped synthesis of the active pharmaceutical ingredient edaravone (**52**, Table 2, entry 10). In this study, a self-optimizing flow reactor was used to run the optimization of seven continuous variables, including three variables for the first step and four variables for the second step. The optimization had the objective of maximizing the yield of the imine intermediate **51** obtained after the first reaction, the STY of **52**, and minimizing the overall used equivalents of the reagents. After 85 iterations, a maximum yield of 95% for the synthesis of **51** and a maximum STY of 5.42 kg/h for the synthesis of edaravone (**52**) were achieved. Setting the objective to reducing the quantity of reagents led to the discovery of unexpected reaction conditions where a substoichiometric amount of triethylamine was sufficient to promote the second reaction step, decreasing the waste produced during synthesis. Although no global solution that provided the optimal reaction conditions for all three objectives was found, a distinct set of reaction conditions was identified that led to a high yield and a low overall number of equivalents of reagent.

Accelerating optimization campaigns

Shortening the optimization time is desirable, especially when manufacturing active pharmaceutical ingredients where only small amounts of materials are available in each step in the development. Currently, optimization methods require an initialization step where reaction conditions are sampled and executed to train the surrogate models used during the optimization (Figure 9a). Sagmeister et al. [83] performed a multiobjective optimization of an S_NAr reaction in an automated flow reactor platform and compared initialization sampling methods to understand how different methods affect the final number of experiments required to find optimal conditions (Table 2, entry 9). They compared LHS (20 experiments), full factorial DoE (17 experiments), and centerpoint (only one experiment) as the starting data points. They found that LHS and full factorial DoE required a smaller number of optimization iterations after the initial set of experiments was conducted due to the better predictive capability of GPs trained with larger amounts of data. However, when the total number of experiments including the initialization set was considered, the number of experiments required to obtain optimal reaction values was larger than, or equal to the situation where only one starting point was used as the only initial sample of reaction conditions. Thus, the authors concluded that it is beneficial to start the algorithm-driven optimization as soon as possible instead of performing an initial thorough exploration of the parametric space. However, they did not fully explore if there was a trade-off between a reduced number of initialization sampling and a total number of experiments to achieve the optimal reaction conditions. Further studies are required to understand this relationship.

Recently, Taylor et al. [92] introduced the concept of multitask Bayesian optimization (MTBO) for chemical reaction optimization. Analogous to transfer learning in ML models, the idea behind multitask learning is to pretrain the surrogate GP models with data that has been previously collected from similar reactions to eliminate the need of an initial sampling step and reduce the overall number of experiments required to obtain the optimal reaction conditions. In MTBO, the standard GP surrogate models are replaced with multitask GPs that use kernels able to create correlations between multiple GPs. The GP that models the experimental conditions that are being optimized is called the main task, while any other GP trained on previous data is called an auxiliary task (Figure 10a). The authors benchmarked MTBO in silico for a single objective optimization for a Suzuki–Miyaura reaction. They discovered that in most cases, pretraining the multitask GPs using a single dataset as an auxiliary task resulted in fewer iterations in comparison to standard BO in order to achieve the optimal conditions. Moreover, the authors observed that when four auxiliary tasks were used instead of 1, the number of iterations required to the obtain optimal reaction conditions was reduced from 15 to fewer than five experiments (Figure 10b). Finally, the authors tested the performance of MTBO in a series of palladium-catalyzed C–H activation reactions of chloroacetanilides in an automated flow reactor to produce the corresponding oxindoles (Figure 10c). For all reactions, three continuous and one categorical variable were optimized to maximize the reaction yield. The authors first performed a standard single-objective BO of reaction (i) in Figure 10c. The optimization was initialized with a set of 16 distinct reaction conditions sampled by LHS, reaching optimal reaction conditions within seven further BO iterations. Subsequently, reaction (ii), yielding a similar oxindole product, was optimized using MTBO, wherein the data gathered from the previous optimization was used to train the auxiliary GP, obtaining the optimal conditions within only 11 iterations, in comparison to 18 required for the first reaction. Reaction (iii), yielding another similar oxindole product, was optimized using the previous data from the first two optimization campaigns to train the auxiliary task GP. The authors found the optimal conditions within five iterations by the algorithm. Further, the authors tested the ability of MTBO to learn from previous experiments by performing the optimization of two other C–H activation reactions, where the structure of the substrate **91** was substantially different in comparison to the first three optimizations. Thus, for the fourth campaign, they tested the optimization of a reaction that produced a six-membered quinolinone ring instead of the five-membered ring present in oxindoles. The MTBO was able to find optimal reaction conditions within ten iterations, demonstrating the capability of the algorithm to handle the optimization of reactions that show small structural deviations from the auxiliary task. Finally, the limits of the

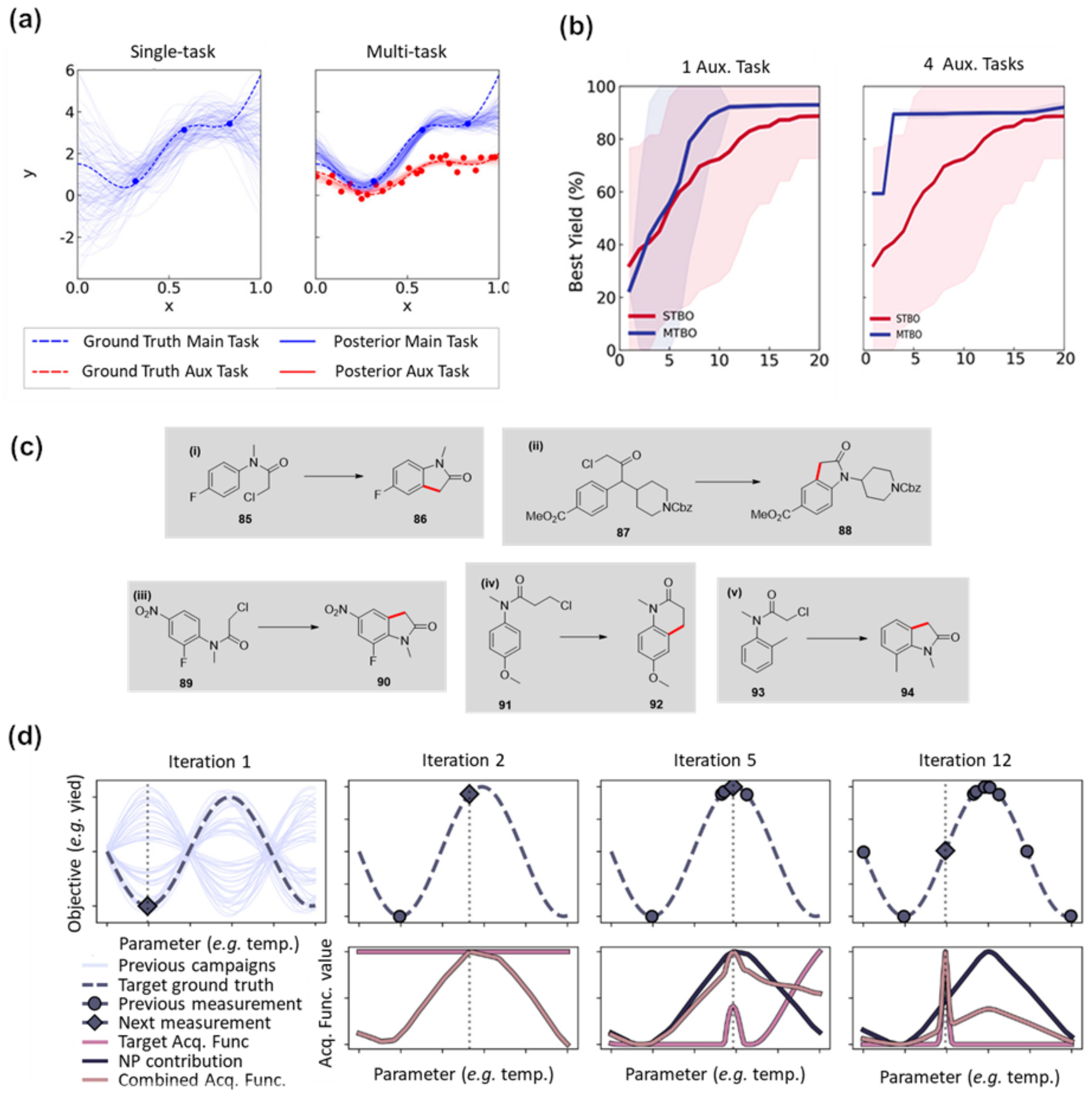


Figure 10: (a) Comparison between a standard GP (single-task) and a multitask GP. Training an auxiliary task using data collected from a similar reaction reduces the uncertainties associated with the GP predictions. (b) Comparison of reaction optimizations performed in silico for single-task and multitask BO. Multitask BO requires a reduced number of iterations to find optimal parameters that maximize the reaction yield. The performance is further improved by incorporating a larger number of auxiliary tasks. (c) Reactions used to test multitask BO under experimental conditions. Reaction (i) was performed using standard single-task BO, where each subsequent reaction incorporated the previously collected data to train auxiliary tasks. (d) Example of SeMOpt algorithm maximizing a sine function. The upper row shows the ground truth function with the sampled points and the best suggested candidate by the BO algorithm. The bottom row shows the values from the acquisition function from the surrogate of the target objective, the neural processes (NPs), and their combination. Figure 10a and 10b were reproduced from [92] (© 2023 C. J. Taylor et al., published by American Chemical Society, distributed under the terms of the Creative Commons Attribution 4.0 International License, https://creativecommons.org/licenses/by/4.0). Figure 10d was republished with permission of The Royal Society of Chemistry, from [93] ("Equipping data-driven experiment planning for Self-driving Laboratories with semantic memory: case studies of transfer learning in chemical reaction optimization" by R. J. Hickman et al., React. Chem. Eng., vol. 8, issue 9, © 2023); permission conveyed through Copyright Clearance Center, Inc. This content is not subject to CC BY 4.0.

MTBO were tested by using a chloroacetanilide **93** having an electron-rich aromatic ring. Therein, the MTBO was unable to discover satisfying reaction conditions.

Recently, researchers from Atinary Technologies reported the development of SeMOpt, a BO framework that, similarly to MTBO, aims to transfer knowledge obtained from previous optimization campaigns to accelerate chemical reaction optimization [93]. In comparison to MTBO, SeMOpt has the advantage of being an agnostic model, and thus it can be applied to any combination of surrogate model and acquisition function used during the BO campaign. In addition to the surrogate model used for BO (see Figure 9a), SeMOpt introduces a surrogate NP to model and make predictions based on previously gathered data. Then, an acquisition function is used to select likely candidates by evaluating both the surrogate model and NP predictions. SeMOpt introduces the knowledge learnt by biasing the acquisition function of the surrogate model for the target optimization with the acquisition function evaluated using the NP model (Figure 10d). In addition, the bias introduced to the acquisition function by the NP is continuously updated and decreases as the number of optimization iterations increases. In this way, the optimization surrogate will eventually disregard the bias introduced by the NP whenever it becomes uninformative. The authors benchmarked the performance of the SeMOpt framework by performing an in silico single-objective optimization of a simulated cross-coupling reaction and a Buchwald–Hartwig cross-coupling of aryl halides. For the benchmarking, the authors used several different BO algorithms and compared their performance when paired with SeMOpt. The authors observed that in all cases, the application of SeMOpt outperformed the single-task implementation of the same BO algorithm. In addition, they compared the performance of SeMOpt against other algorithms that include some knowledge transfer into the optimization workflow, including MTBO. The authors observed that SeMOpt outperformed most of the other algorithms, with MTBO closely matching the performance of SemMOpt.

Mixed-variable optimizations

A challenge in BO is to include categorical variables (i.e., noncontinuous) into the optimization procedures due to the inherent limitations of standard GPs to include discrete variables into their predictions. Categorical variables, such as choice of solvent, catalyst, ligands, additives, etc., are crucial for many chemical reactions. For this purpose, new algorithms have been developed to include categorical variables into MOBOs. Kershaw et al. [90] utilized an MVMOO algorithm developed in house, employing GP regression surrogate models tailored for predictions with discrete variable inputs. Their study employed a self-driving flow reactor to optimize the synthesis of *ortho-* and *para-*isomers **21** and **22** of an SNAr reaction, leveraging four continuous variables alongside a single discrete variable representing the solvent (Table 2, entry 15). After 99 sequential reactions (25 LHS steps and 74 optimization iterations), the researchers found 20 nondominated solutions that mapped the Pareto front from a highly dominant *ortho-*product **21** to a 50:50 split between the isomers. In addition, the researchers explored the optimization of a Sonogashira cross-coupling to optimize the STY and RME for the synthesis of **69** (Table 2, entry 16). In this case, the optimization involved three continuous variables and the selection of a ligand for the catalyst as a discrete variable. After 69 sequential experiments (25 LHS steps, 44 optimizations), the platform was able to identify 12 nondominated solutions that demonstrated the trade-off between RME and STY. In general, most Pareto solutions were obtained when triphenylphosphine was used as the catalyst ligand. Interestingly, triphenylphosphine was the least sterically hindering ligand, which is counterintuitive to expert intuition that may identify sterically demanding ligands as more favorable choices for cross-coupling reactions.

Another noteworthy approach for the optimization of both continuous and categorical variables for a Suzuki–Miyaura coupling reaction was reported by Christensen et al. [15] using BO algorithms developed in house called Phoenics and Gryffin (Table 2, entry 6). The Gryffin algorithm uses Bayesian neural networks to construct the surrogate model, circumventing the limitations of GPs to fit categorical variables. The authors chose a total of four continuous reaction variables and selected a catalyst ligand as the unique categorical variable for the optimization. The algorithm targeted the optimal reaction variables for four objectives, including the maximization of the targeted stereoisomer (*E*)-**39**, the minimization of the undesirable one (*Z*)-**39**, catalyst loading, and reagent equivalents. Twelve ligands were initially selected based on domain expert knowledge, and after 120 trials, the best conditions were found to be similar to those previously reported in the literature. To further improve the performance of the reaction, the authors used DFT simulations to compute the chemical properties of 365 commercially available phosphine ligands, and by using *k*-means clustering, they grouped the ligands into 24 distinct regions. Through the strategic selection of a representative ligand from each distinct region, the researchers identified a novel set of ligands that differed from conventional recommendations based on domain expertise. Following the optimization of the reaction conditions using these 23 new ligands, the authors observed enhanced performance, surpassing that of previous reports (Figure 11). This study showcased how data science, ML algorithms, and reaction optimization can be used to discover reaction conditions that would have otherwise been overlooked by human intuition. Another great example of a combination of ML and AI cheminformatic tools and reaction optimization was reported by Nambiar et al. [84], who presented the use of a computer-aided synthesis planning (CASP) tool to find a three-step reaction pathway for the synthesis of the active pharmaceutical ingredient sonidegib (**58**). After the generation of multiple reaction pathways by the CASP tool, the authors manually selected a highly ranked route based on synthetic feasibility. This three-step reaction comprised an S_NAr, hydrogenative reduction of a nitro group, and an amide coupling (Table 2, entries 11 and 12). Using an automated flow reactor, the researchers attempted to perform the optimization of the fully telescoped reaction. However, the optimization campaign had to

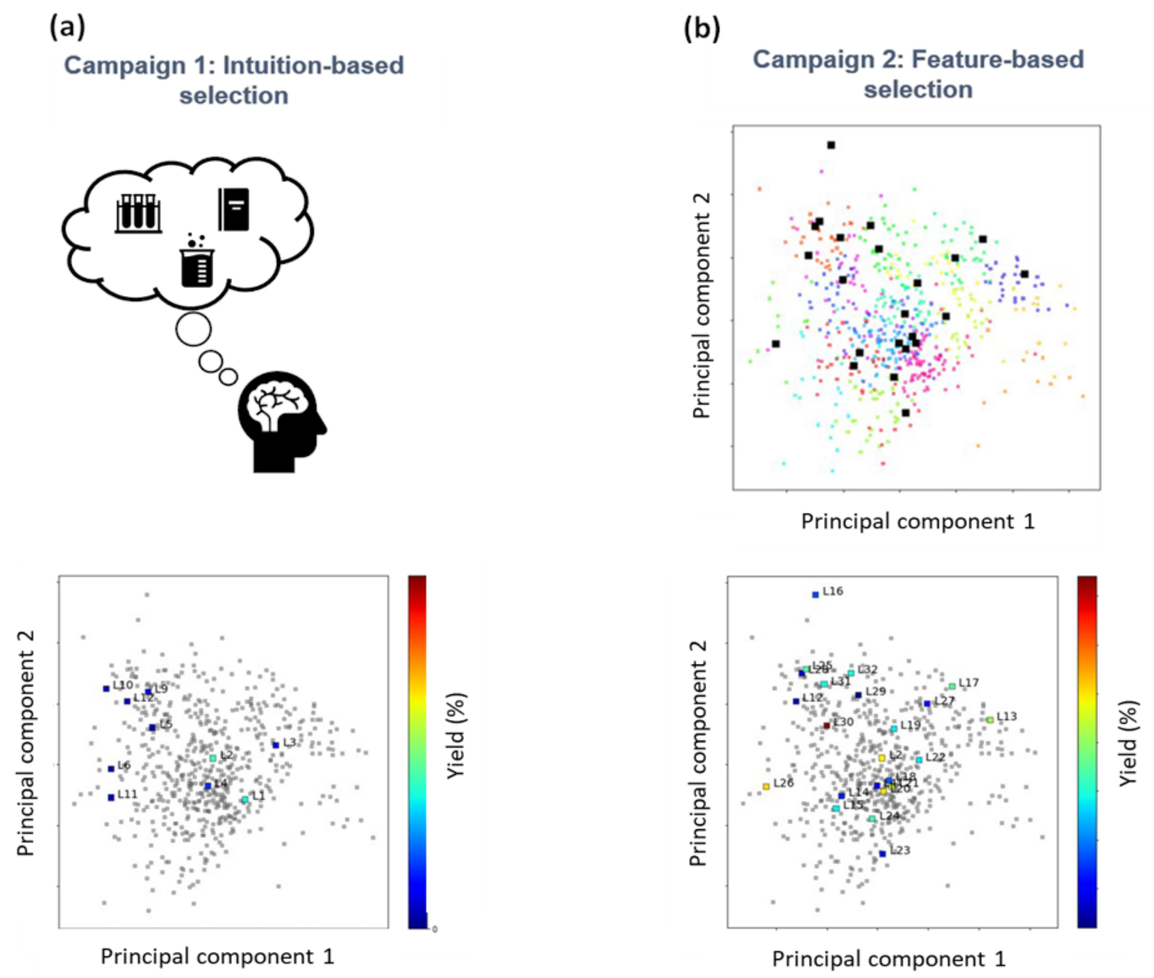


Figure 11: Comparison of the reaction yield between optimizations campaign where the catalyst ligand selection was based on (a) expert intuition and (b) sampling the distinct ligand clusters obtained from *k*-means clustering in the calculated chemical space. Figure 11 was adapted from [15] (© 2021 M. Christensen et al., published by Springer Nature, distributed under the terms of the Creative Commons Attribution 4.0 International License, https://creativecommons.org/licenses/by/4.0.

be restructured into two independent optimizations due to the side products of the S_NAr reaction poisoning the Pd catalyst used in the hydrogenation reaction. Thus, the MOBO of the S_NAr reaction was performed to maximize the yield of **55**, the productivity, and to minimize the cost of the reagents per mole of product by optimizing four continuous and one categorical variable. The second optimization campaign was performed for the telescoped reaction, which included the hydrogenation step and the amide coupling. Therein, the objectives of the optimization were to maximize the yield and productivity by optimizing two categorical and three continuous variables.

Dragonfly, an open-sourced BO package, was used to optimize both categorical and continuous reaction variables. An increase in yield and productivity was observed as the optimization progressed. The authors found that the selection of F as a leaving group led to the highest yield (98.3%) and productivity (5.97 g/h) for the synthesis of **58**. However, if Cl was selected as the leaving group, only a marginal reduction in yield and productivity was observed (93.8% and 5.70 g/h), but a 33% reduction in the cost. In the second reaction, both a high yield and productivity were achieved concurrently. Because these objectives were positively correlated, no trade-offs were observed in the optimization suggestions. Recently, Aldulaijan et al. [91] reported a novel single-objective ALaBO algorithm that can optimize continuous and categorical variables simultaneously. This algorithm first encodes the continuous and categorical variables into a 2D latent space, creating a continuous response surface for the objective function, which can be modeled by standard GPs and optimized by standard acquisition functions, such as adaptive expected improvement. Once the likelihood for the optimal variables is determined within the latent space, they can be decoded into their original continuous and categorical forms. Thus, this approach enables a "one-shot" optimization of both kinds of variables without the need for specialized GP modeling techniques. The authors evaluated the efficacy of the ALaBO through the optimization of catalytic reactions, demonstrating a faster convergence to optimal values in comparison to Dragonfly.

Benchmarking of optimization algorithms

With an increasing number of optimization algorithms, an effort to benchmark their performance is required. Felton et al. [76] have highlighted the fact that the ability of an algorithm to perform well in a specific task may not translate universally to other problems, and thus a specific algorithm for chemical reaction optimization may have different performances depending on the nature of the target variables, objectives, and chemical reaction. Also, the computational time required to execute an algorithm varies, and it should be taken into consideration in order to select the most appropriate variation for each case study. To benchmark different optimization algorithms, Felton et al. released Summit, a Python module containing several optimization algorithms and two benchmark in silico models, to compare the performance of algorithms. Initially, the benchmarking models included in Summit were a kinetic model for the S_NAr reaction of difluoronitrobenzene with pyrrolidine and a neural network forward model for the prediction of the yield of diphenylamine in a Pd-catalyzed C–N cross-coupling reaction trained on a previously published dataset containing 96 unique sets of reaction conditions. The optimization for the S_NAr reaction included four continuous variables and two optimization objectives, while the C–N cross-coupling included three continuous variables, two categorical variables, and two optimization objectives. The algorithms used during the optimization included non-ML algorithms (Nelder–Mead, SNOBFIT), BO algorithms (Gryffin, SOBO, TSEMO), and distributionally robust optimization (DRO), a pretrained reinforcement learning agent algorithm. For the optimization of S_NAr reaction, BO methods were superior to any other of the algorithms, reaching a higher hypervolume within a smaller number of iterations. When the BO algorithms were compared, TSEMO outperformed Gryffin and SOBO by a significant margin. For the C–N cross-coupling, all models had a similar hypervolume performance, including a random search of reaction conditions, due to the small parametric space for the selected categorical variable. Müller et al. [77] also conducted a benchmarking in silico study for six different chemical reactions using previously reported kinetic models. Therein, three distinct BO algorithms (TSEMO, ParEGO, EIM-EGO) and a genetic algorithm (NSGA-II) were compared. The authors demonstrated that BO methods outperformed non-BO methods such as NSGA-II, which is consistent with the earlier studies by Felton et al. [76].

Conclusion

In this article, we outlined the latest advances in ML-driven multiobjective optimization for chemical synthesis, in addition to breakthroughs in HTE and analytical techniques. The recent developments of ML algorithms, HTE tools, data processing techniques, and self-optimizing reactors has been a transformative force for chemical optimization processes. Nonetheless, there are still plenty of research opportunities to continue the transformation of the field and to accelerate the execution of chemical reaction optimization. Given the time-consuming nature inherit to organic synthesis and characterization, optimization campaigns are significantly limited by the time required to test new reaction conditions. This is importantly true for campaigns aimed to map a Pareto front, which can often require too many evaluations to be conducted experimentally. Innovative approaches such as MTBO and transfer learning have already demonstrated improvements in reducing the number of experiments to find optimal solutions. However, developing novel algorithms that address the limitations of traditional BO approaches would also yield substantial benefits. For example, existing BO algorithms are often concerned with optimizing the objective and fail to uniformly map the Pareto front [94,95]. New algorithms that integrate sampling procedures based on single-step evolutionary algorithms in conjunction with BO have demonstrated fast convergence, decreased sampling wastage, and uniform exploration of the Pareto front, which could be promising in the field of organic reaction optimization [95]. We anticipate that further advancements will lead to better-performing algorithms that require a minimal number of experiments to achieve optimal solutions.

The field has experienced substantial progress in optimizing multiple continuous variables, yet the utilization of categorical variables in chemical synthesis optimization has predominantly been confined to single-step reactions with one or two optimization objectives. The development of ML algorithms that can efficiently optimize a larger number of categorical variables will be crucial to unlocking the full potential of optimization methods. This is particularly true when objective functions that go beyond direct measurements of the reaction product outputs (e.g., yield, throughput, selectivity, etc.) are targeted. For example, optimizations that aim to minimize the environmental impact of chemical synthesis are becoming a priority in industry. The environmental impact of a reaction not only depends on the efficiency of the process (i.e., yield and throughput) but will be highly affected by the nature of the solvent, catalyst, reagents, downstream workup, etc. used in the synthesis. To obtain optimal reaction conditions that minimize the environmental impact, the exploration of a large number of different reagents may be required, which is not possible through traditional optimization methodologies. Nonetheless, ML algorithms could offer an efficient approach to navigating the parametric space and to reduce the experimentation time to find the conditions that minimize the environmental impact of a particular manufacturing process. However, the state-of-the-art optimization algorithms that incorporate mixed variables still fall short of handling the large number of categorical variables required for these studies.

Manufacturing of pharmaceutical and specialty chemicals commonly involves multiple reaction steps in order to transform the starting reagents into the final product. So far, optimization algorithms have been mostly applied to single-step reactions or step by step to each reaction of a multistep procedure. Few examples in the literature have demonstrated the ability of ML methods to optimize telescoped reactions in automated flow reactors, but the positive results should encourage further research in this field. However, situations where the telescoped reactions are not feasible due to competing chemical interactions of the reagents in the reaction mixture are bound to occur. Thus, more research should investigate optimization strategies in multistep reaction procedures in which the final objective function has input variables from multiple steps of the synthetic route.

The application of ML algorithms to aid the discovery of new chemistry knowledge is flourishing, from generative design to property prediction and reaction planning. Further work should incorporate the diverse applications of ML into chemical reaction optimization campaigns to open new avenues for research and discovery. In particular, ML tools have great potential for the planning of reaction optimization campaigns to assist the selection of categorical chemical variables (e.g., catalysts, ligands, additives, etc.). Christensen et al. [15] have already demonstrated the advantages of applying ML clustering methods to discover new ligands for catalysts that would have been missed if the selection of test ligands had only relied on human chemical intuition. Taylor et al. [92] also highlighted the use of DFT or ML alternatives to find similarities between reaction models in order to apply efficient multitask learning to chemical reaction optimization. Another potential application of ML tools is the use of CASP to discover alternative reaction routes, with the potential to improve the efficiency of current manufacturing methods. Finally, leveraging on the large quantities of data generated from self-optimizing chemical platforms and their experimental versatility, we envision the incorporation of reaction optimization methods with generative design to create full-driving laboratories. These could tackle both the discovery of new molecules and the search for optimal synthesis conditions to meet the production requirements for a chemical commodity.

Future research on the optimization of organic chemistry reactions should leverage advanced deep learning models. In particular, we highlight large language models (LLMs) as a promising technology to enable the extraction of chemical knowledge from previous literature. LLMs can be used to generate synthesis protocols for target materials through data mining of peer-reviewed literature [96,97]. Bran et al. [98] recently demonstrated an advanced LLM-powered chemistry engine called ChemCrow that is capable of planning and executing the synthesis of organic molecules. The LLM integrated 18 cheminformatic tools and performed the reasoning steps based on the information supplied by these tools to accomplish specific chemistry tasks. Along these lines, we envision that the integration of CASP tools and LLMs could accelerate the optimization of organic reactions by providing viable reaction routes with starting conditions that are close to the reaction optimum based on previous studies. LLMs could also assist researchers with limited coding experience to write the code required for automating their experimental workflows and execute their reaction optimizations. However, the use of LLMs to drive experimental campaigns is still in its early stages, making it crucial to understand their limitations and potential shortcomings in generating valuable content for chemical sciences. A recent study has shown that LLMs can generate erroneous and misleading information regarding chemical safety, which requires to be addressed to avoid accidents in autonomous platforms controlled by these models [99]. Early findings suggest that prompt engineering [100], fine-tuning [97], and retrieval-augmented generation [101] could improve the reliability of LLMs in chemistry-related tasks and enable their widespread application in the field.

Standardizing benchmarking methods for ML optimization algorithms will be crucial as the number of optimization methodologies increases. Foundational work has been laid by the Lapkin research group with the release of the Summit open-source software package [76]. Given the vast spectrum of chemical reactions, there is a necessity to develop a diverse array of reaction models to comprehensively assess the suitability of optimization methods for various scenarios. The field should leverage the ability of HTE to produce large amounts of data to create reliable forward models that can be incorporated into an online repository. Thus, researchers could access this online repository to benchmark new optimization algorithms by performing in silico optimization campaigns of the chemical reaction models.

For the continued advancement of this research, it is paramount to democratize access to proprietary autonomous platforms and algorithms and to foster collaboration to share expertise within academia. While particularly significant advances have been made in addressing immediate challenges, we are convinced that the full potential of ML and AI is yet to be realized. This highlights the importance of raising cross-functional expertise both within universities and at preuniversity levels, thereby nurturing a broader knowledge base. Such an approach will empower young researchers to tackle complex scientific challenges holistically right from the outset, thereby unlocking new possibilities for innovation and advancement.

Funding

The authors acknowledge funding from the Materials Generative Design and Testing Framework (MAT-GDT) Program at A*STAR, provided through the AME Programmatic Fund (Grant No. M24N4b0034). K. H. acknowledges the National Research Foundation's Competitive Research Programme (NRF-CRP) in Singapore (Grant No. NRF-CRP25-2020-0002). B. R. thanks the Horizontal Technology Coordinating Office of A*STAR for seed funding under project No. C231218004.

Author Contributions

Pablo Quijano Velasco: conceptualization; methodology; writing – original draft; writing – review & editing. Kedar Hippalgaonkar: conceptualization; methodology; validation; writing – review & editing. Balamurugan Ramalingam: conceptualization; investigation; methodology; writing – original draft; writing – review & editing.

ORCID® iDs

Pablo Quijano Velasco - https://orcid.org/0000-0001-6924-4642
Balamurugan Ramalingam - https://orcid.org/0000-0002-6688-1205

Data Availability Statement

Data sharing is not applicable as no new data was generated or analyzed in this study.

Preprint

A non-peer-reviewed version of this article has been previously published as a preprint: doi:10.26434/chemrxiv-2024-vbgc6

References

1. Meuwly, M. *Chem. Rev.* **2021,** *121,* 10218–10239. doi:10.1021/acs.chemrev.1c00033
2. Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G. *Nature* **2021,** *590,* 89–96. doi:10.1038/s41586-021-03213-y
3. Gao, W.; Raghavan, P.; Coley, C. W. *Nat. Commun.* **2022,** *13,* 1075. doi:10.1038/s41467-022-28736-4
4. Seifrid, M.; Pollice, R.; Aguilar-Granda, A.; Morgan Chan, Z.; Hotta, K.; Ser, C. T.; Vestfrid, J.; Wu, T. C.; Aspuru-Guzik, A. *Acc. Chem. Res.* **2022,** *55,* 2454–2466. doi:10.1021/acs.accounts.2c00220
5. Taylor, C. J.; Pomberger, A.; Felton, K. C.; Grainger, R.; Barecka, M.; Chamberlain, T. W.; Bourne, R. A.; Johnson, C. N.; Lapkin, A. A. *Chem. Rev.* **2023,** *123,* 3089–3126. doi:10.1021/acs.chemrev.2c00798
6. Griffin, D. J.; Coley, C. W.; Frank, S. A.; Hawkins, J. M.; Jensen, K. F. *Org. Process Res. Dev.* **2023,** *27,* 1868–1879. doi:10.1021/acs.oprd.3c00229
7. Sagmeister, P.; Williams, J. D.; Kappe, C. O. *Chimia* **2023,** *77,* 300. doi:10.2533/chimia.2023.300
8. Fromer, J. C.; Coley, C. W. *Patterns* **2023,** *4,* 100678. doi:10.1016/j.patter.2023.100678
9. Askr, H.; Elgeldawi, E.; Aboul Ella, H.; Elshaier, Y. A. M. M.; Gomaa, M. M.; Hassanien, A. E. *Artif. Intell. Rev.* **2023,** *56,* 5975–6037. doi:10.1007/s10462-022-10306-1
10. Vamathevan, J.; Clark, D.; Czodrowski, P.; Dunham, I.; Ferran, E.; Lee, G.; Li, B.; Madabhushi, A.; Shah, P.; Spitzer, M.; Zhao, S. *Nat. Rev. Drug Discovery* **2019,** *18,* 463–477. doi:10.1038/s41573-019-0024-5
11. Kreutter, D.; Reymond, J.-L. *Chem. Sci.* **2023,** *14,* 9959–9969. doi:10.1039/d3sc01604h
12. Mo, Y.; Guan, Y.; Verma, P.; Guo, J.; Fortunato, M. E.; Lu, Z.; Coley, C. W.; Jensen, K. F. *Chem. Sci.* **2021,** *12,* 1469–1478. doi:10.1039/d0sc05078d
13. Bennett, J. A.; Orouji, N.; Khan, M.; Sadeghi, S.; Rodgers, J.; Abolhasani, M. *Nat. Chem. Eng.* **2024,** *1,* 240–250. doi:10.1038/s44286-024-00033-5
14. Yang, W.; Fidelis, T. T.; Sun, W.-H. *ACS Omega* **2020,** *5,* 83–88. doi:10.1021/acsomega.9b03673
15. Christensen, M.; Yunker, L. P. E.; Adedeji, F.; Häse, F.; Roch, L. M.; Gensch, T.; dos Passos Gomes, G.; Zepel, T.; Sigman, M. S.; Aspuru-Guzik, A.; Hein, J. E. *Commun. Chem.* **2021,** *4,* 112. doi:10.1038/s42004-021-00550-x
16. Buitrago Santanilla, A.; Christensen, M.; Campeau, L.-C.; Davies, I. W.; Dreher, S. D. *Org. Lett.* **2015,** *17,* 3370–3373. doi:10.1021/acs.orglett.5b01648
17. Brocklehurst, C. E.; Gallou, F.; Hartwieg, J. C. D.; Palmieri, M.; Rufle, D. *Org. Process Res. Dev.* **2018,** *22,* 1453–1457. doi:10.1021/acs.oprd.8b00200
18. Grainger, R.; Heightman, T. D.; Ley, S. V.; Lima, F.; Johnson, C. N. *Chem. Sci.* **2019,** *10,* 2264–2271. doi:10.1039/c8sc04789h
19. Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. *Science* **2018,** *360,* 186–190. doi:10.1126/science.aar5169
20. Gesmundo, N.; Dykstra, K.; Douthwaite, J. L.; Kao, Y.-T.; Zhao, R.; Mahjour, B.; Ferguson, R.; Dreher, S.; Sauvagnat, B.; Saurí, J.; Cernak, T. *Nat. Synth.* **2023,** *2,* 1082–1091. doi:10.1038/s44160-023-00351-1
21. Zarate, C.; Ardolino, M.; Morriello, G. J.; Logan, K. M.; Kaplan, W. P.; Torres, L.; Li, D.; Chen, M.; Li, H.; Su, J.; Fuller, P.; Maddess, M. L.; Song, Z. J. *Org. Process Res. Dev.* **2021,** *25,* 642–647. doi:10.1021/acs.oprd.0c00446
22. Fier, P. S.; Maloney, K. M. *Org. Lett.* **2017,** *19,* 3033–3036. doi:10.1021/acs.orglett.7b01403
23. Huff, C. A.; Cohen, R. D.; Dykstra, K. D.; Streckfuss, E.; DiRocco, D. A.; Krska, S. W. *J. Org. Chem.* **2016,** *81,* 6980–6987. doi:10.1021/acs.joc.6b00811
24. DiRocco, D. A.; Dykstra, K.; Krska, S.; Vachal, P.; Conway, D. V.; Tudge, M. *Angew. Chem., Int. Ed.* **2014,** *53,* 4802–4806. doi:10.1002/anie.201402023
25. Primer, D. N.; Molander, G. A. *J. Am. Chem. Soc.* **2017,** *139,* 9847–9850. doi:10.1021/jacs.7b06288

26. Nicastri, M. C.; Lehnherr, D.; Lam, Y.-h.; DiRocco, D. A.; Rovis, T. *J. Am. Chem. Soc.* **2020,** *142,* 987–998. doi:10.1021/jacs.9b10871
27. Corcoran, E. B.; Pirnot, M. T.; Lin, S.; Dreher, S. D.; DiRocco, D. A.; Davies, I. W.; Buchwald, S. L.; MacMillan, D. W. C. *Science* **2016,** *353,* 279–283. doi:10.1126/science.aag0209
28. Lin, S.; Dikler, S.; Blincoe, W. D.; Ferguson, R. D.; Sheridan, R. P.; Peng, Z.; Conway, D. V.; Zawatzky, K.; Wang, H.; Cernak, T.; Davies, I. W.; DiRocco, D. A.; Sheng, H.; Welch, C. J.; Dreher, S. D. *Science* **2018,** *361,* eaar6236. doi:10.1126/science.aar6236
29. Lee, H.; Boyer, N. C.; Deng, Q.; Kim, H.-Y.; Sawyer, T. K.; Sciammetta, N. *Chem. Sci.* **2019,** *10,* 5073–5078. doi:10.1039/c9sc00694j
30. Burger, B.; Maffettone, P. M.; Gusev, V. V.; Aitchison, C. M.; Bai, Y.; Wang, X.; Li, X.; Alston, B. M.; Li, B.; Clowes, R.; Rankin, N.; Harris, B.; Sprick, R. S.; Cooper, A. I. *Nature* **2020,** *583,* 237–241. doi:10.1038/s41586-020-2442-2
31. Manzano, J. S.; Hou, W.; Zalesskiy, S. S.; Frei, P.; Wang, H.; Kitson, P. J.; Cronin, L. *Nat. Chem.* **2022,** *14,* 1311–1318. doi:10.1038/s41557-022-01016-w
32. Ha, T.; Lee, D.; Kwon, Y.; Park, M. S.; Lee, S.; Jang, J.; Choi, B.; Jeon, H.; Kim, J.; Choi, H.; Seo, H.-T.; Choi, W.; Hong, W.; Park, Y. J.; Jang, J.; Cho, J.; Kim, B.; Kwon, H.; Kim, G.; Oh, W. S.; Kim, J. W.; Choi, J.; Min, M.; Jeon, A.; Jung, Y.; Kim, E.; Lee, H.; Choi, Y.-S. *Sci. Adv.* **2023,** *9,* eadj0461. doi:10.1126/sciadv.adj0461
33. Godfrey, A. G.; Masquelin, T.; Hemmerle, H. *Drug Discovery Today* **2013,** *18,* 795–802. doi:10.1016/j.drudis.2013.03.001
34. IBM RXN for Chemistry. https://rxn.res.ibm.com/rxn/robo-rxn/welcome (accessed April 9, 2024).
35. Vaucher, A. C.; Zipoli, F.; Geluykens, J.; Nair, V. H.; Schwaller, P.; Laino, T. *Nat. Commun.* **2020,** *11,* 3601. doi:10.1038/s41467-020-17266-6
36. Kuleshova, J.; Hill-Cousins, J. T.; Birkin, P. R.; Brown, R. C. D.; Pletcher, D.; Underwood, T. J. *Electrochim. Acta* **2011,** *56,* 4322–4326. doi:10.1016/j.electacta.2011.01.036
37. Green, R. A.; Brown, R. C. D.; Pletcher, D.; Harji, B. *Electrochem. Commun.* **2016,** *73,* 63–66. doi:10.1016/j.elecom.2016.11.004
38. Sun, A. C.; Steyer, D. J.; Robinson, R. I.; Ginsburg-Moraff, C.; Plummer, S.; Gao, J.; Tucker, J. W.; Alpers, D.; Stephenson, C. R. J.; Kennedy, R. T. *Angew. Chem., Int. Ed.* **2023,** *62,* e202301664. doi:10.1002/anie.202301664
39. González-Esguevillas, M.; Fernández, D. F.; Rincón, J. A.; Barberis, M.; de Frutos, O.; Mateos, C.; García-Cerrada, S.; Agejas, J.; MacMillan, D. W. C. *ACS Cent. Sci.* **2021,** *7,* 1126–1134. doi:10.1021/acscentsci.1c00303
40. Hsieh, H.-W.; Coley, C. W.; Baumgartner, L. M.; Jensen, K. F.; Robinson, R. I. *Org. Process Res. Dev.* **2018,** *22,* 542–550. doi:10.1021/acs.oprd.8b00018
41. Churski, K.; Korczyk, P.; Garstecki, P. *Lab Chip* **2010,** *10,* 816. doi:10.1039/b925500a
42. Kaminski, T. S.; Jakiela, S.; Czekalska, M. A.; Postek, W.; Garstecki, P. *Lab Chip* **2012,** *12,* 3995. doi:10.1039/c2lc40540g
43. Reizman, B. J.; Jensen, K. F. *Chem. Commun.* **2015,** *51,* 13290–13293. doi:10.1039/c5cc03651h
44. Abolhasani, M.; Jensen, K. F. *Lab Chip* **2016,** *16,* 2775–2784. doi:10.1039/c6lc00728g
45. Coley, C. W.; Abolhasani, M.; Lin, H.; Jensen, K. F. *Angew. Chem., Int. Ed.* **2017,** *56,* 9847–9850. doi:10.1002/anie.201705148
46. Pieber, B.; Shalom, M.; Antonietti, M.; Seeberger, P. H.; Gilmore, K. *Angew. Chem., Int. Ed.* **2018,** *57,* 9976–9979. doi:10.1002/anie.201712568
47. Wagner, F.; Sagmeister, P.; Jusner, C. E.; Tampone, T. G.; Manee, V.; Buono, F. G.; Williams, J. D.; Kappe, C. O. *Adv. Sci.* **2024,** *11,* 2308034. doi:10.1002/advs.202308034
48. Slattery, A.; Wen, Z.; Tenblad, P.; Pintossi, D.; Sanjose-Orduna, J.; den Hartog, T.; Noel, T. *ChemRxiv* **2023**. doi:10.26434/chemrxiv-2023-r0drq
49. Collins, N.; Stout, D.; Lim, J.-P.; Malerich, J. P.; White, J. D.; Madrid, P. B.; Latendresse, M.; Krieger, D.; Szeto, J.; Vu, V.-A.; Rucker, K.; Deleo, M.; Gorfu, Y.; Krummenacker, M.; Hokama, L. A.; Karp, P.; Mallya, S. *Org. Process Res. Dev.* **2020,** *24,* 2064–2077. doi:10.1021/acs.oprd.0c00143
50. Perera, D.; Tucker, J. W.; Brahmbhatt, S.; Helal, C. J.; Chong, A.; Farrell, W.; Richardson, P.; Sach, N. W. *Science* **2018,** *359,* 429–434. doi:10.1126/science.aap9112
51. Reis, M.; Gusev, F.; Taylor, N. G.; Chung, S. H.; Verber, M. D.; Lee, Y. Z.; Isayev, O.; Leibfarth, F. A. *J. Am. Chem. Soc.* **2021,** *143,* 17677–17689. doi:10.1021/jacs.1c08181
52. Zhou, Y.; Gu, Y.; Jiang, K.; Chen, M. *Macromolecules* **2019,** *52,* 5611–5617. doi:10.1021/acs.macromol.9b00846
53. Tan, J. D.; Ramalingam, B.; Wong, S. L.; Cheng, J. J. W.; Lim, Y.-F.; Chellappan, V.; Khan, S. A.; Kumar, J.; Hippalgaonkar, K. *J. Chem. Inf. Model.* **2023,** *63,* 4560–4573. doi:10.1021/acs.jcim.3c00504
54. Ahn, G.-N.; Sharma, B. M.; Lahore, S.; Yim, S.-J.; Vidyacharan, S.; Kim, D.-P. *Commun. Chem.* **2021,** *4,* 53. doi:10.1038/s42004-021-00490-6
55. Chatterjee, S.; Guidi, M.; Seeberger, P. H.; Gilmore, K. *Nature* **2020,** *579,* 379–384. doi:10.1038/s41586-020-2083-5
56. Eyke, N. S.; Schneider, T. N.; Jin, B.; Hart, T.; Monfette, S.; Hawkins, J. M.; Morse, P. D.; Howard, R. M.; Pfisterer, D. M.; Nandiwale, K. Y.; Jensen, K. F. *Chem. Sci.* **2023,** *14,* 8798–8809. doi:10.1039/d3sc02082g
57. McMullen, J. P.; Stone, M. T.; Buchwald, S. L.; Jensen, K. F. *Angew. Chem., Int. Ed.* **2010,** *49,* 7076–7080. doi:10.1002/anie.201002590
58. Fitzpatrick, D. E.; Battilocchio, C.; Ley, S. V. *Org. Process Res. Dev.* **2016,** *20,* 386–394. doi:10.1021/acs.oprd.5b00313
59. Poscharny, K.; Fabry, D. C.; Heddrich, S.; Sugiono, E.; Liauw, M. A.; Rueping, M. *Tetrahedron* **2018,** *74,* 3171–3175. doi:10.1016/j.tet.2018.04.019
60. Cortés-Borda, D.; Wimmer, E.; Gouilleux, B.; Barré, E.; Oger, N.; Goulamaly, L.; Peault, L.; Charrier, B.; Truchet, C.; Giraudeau, P.; Rodriguez-Zubiri, M.; Le Grognec, E.; Felpin, F.-X. *J. Org. Chem.* **2018,** *83,* 14286–14299. doi:10.1021/acs.joc.8b01821
61. Nandiwale, K. Y.; Hart, T.; Zahrt, A. F.; Nambiar, A. M. K.; Mahesh, P. T.; Mo, Y.; Nieves-Remacha, M. J.; Johnson, M. D.; García-Losada, P.; Mateos, C.; Rincón, J. A.; Jensen, K. F. *React. Chem. Eng.* **2022,** *7,* 1315–1327. doi:10.1039/d2re00054g
62. Bédard, A.-C.; Adamo, A.; Aroh, K. C.; Russell, M. G.; Bedermann, A. A.; Torosian, J.; Yue, B.; Jensen, K. F.; Jamison, T. F. *Science* **2018,** *361,* 1220–1225. doi:10.1126/science.aat0650
63. Aka, E. C.; Wimmer, E.; Barré, E.; Cortés-Borda, D.; Ekou, T.; Ekou, L.; Rodriguez-Zubiri, M.; Felpin, F.-X. *Org. Process Res. Dev.* **2020,** *24,* 745–751. doi:10.1021/acs.oprd.9b00525
64. Rodriguez-Zubiri, M.; Felpin, F.-X. *Org. Process Res. Dev.* **2022,** *26,* 1766–1793. doi:10.1021/acs.oprd.2c00102

65. Roos, G.; Röseler, C.; Berger Büter, K.; Simmen, U. *Planta Med.* **2004,** *70,* 771–777. doi:10.1055/s-2004-827210
66. Schmidt, B.; Jaroszewski, J. W.; Bro, R.; Witt, M. *Anal. Chem. (Washington, DC, U. S.)* **2008,** *80,* 1978–1987. doi:10.1021/ac702064p
67. Jansen, B. C.; Hafkenscheid, L.; Bondt, A.; Gardner, R. A.; Hendel, J. L.; Wuhrer, M.; Spencer, D. I. R. *PLoS One* **2018,** *13,* e0200280. doi:10.1371/journal.pone.0200280
68. Bovee, R. https://github.com/bovee/Aston (accessed April 9, 2024).
69. Liu, J.; Sato, Y.; Yang, F.; Kukor, A. J.; Hein, J. E. *Chem.:Methods* **2022,** *2,* e202200009. doi:10.1002/cmtd.202200009
70. Sagmeister, P.; Lebl, R.; Castillo, I.; Rehrl, J.; Kruisz, J.; Sipek, M.; Horn, M.; Sacher, S.; Cantillo, D.; Williams, J. D.; Kappe, C. O. *Angew. Chem., Int. Ed.* **2021,** *60,* 8139–8148. doi:10.1002/anie.202016007
71. Haas, C. P.; Lübbesmeyer, M.; Jin, E. H.; McDonald, M. A.; Koscher, B. A.; Guimond, N.; Di Rocco, L.; Kayser, H.; Leweke, S.; Niedenführ, S.; Nicholls, R.; Greeves, E.; Barber, D. M.; Hillenbrand, J.; Volpin, G.; Jensen, K. F. *ACS Cent. Sci.* **2023,** *9,* 307–317. doi:10.1021/acscentsci.2c01042
72. Sagmeister, P.; Melnizky, L.; Williams, J.; Kappe, C. O. *ChemRxiv* **2024**. doi:10.26434/chemrxiv-2024-mj3h8
73. Holmes, N.; Akien, G. R.; Savage, R. J. D.; Stanetty, C.; Baxendale, I. R.; Blacker, A. J.; Taylor, B. A.; Woodward, R. L.; Meadows, R. E.; Bourne, R. A. *React. Chem. Eng.* **2016,** *1,* 96–100. doi:10.1039/c5re00083a
74. Parrott, A. J.; Bourne, R. A.; Akien, G. R.; Irvine, D. J.; Poliakoff, M. *Angew. Chem., Int. Ed.* **2011,** *50,* 3788–3792. doi:10.1002/anie.201100412
75. Sans, V.; Porwol, L.; Dragone, V.; Cronin, L. *Chem. Sci.* **2015,** *6,* 1258–1264. doi:10.1039/c4sc03075c
76. Felton, K. C.; Rittig, J. G.; Lapkin, A. A. *Chem.:Methods* **2021,** *1,* 116–122. doi:10.1002/cmtd.202000051
77. Müller, P.; Clayton, A. D.; Manson, J.; Riley, S.; May, O. S.; Govan, N.; Notman, S.; Ley, S. V.; Chamberlain, T. W.; Bourne, R. A. *React. Chem. Eng.* **2022,** *7,* 987–993. doi:10.1039/d1re00549a
78. Nandi, S.; Badhe, Y.; Lonari, J.; Sridevi, U.; Rao, B. S.; Tambe, S. S.; Kulkarni, B. D. *Chem. Eng. J.* **2004,** *97,* 115–129. doi:10.1016/s1385-8947(03)00150-5
79. Lahiri, S. K.; Khalfe, N. *Chem. Prod. Process Model.* **2008,** *3*. doi:10.2202/1934-2659.1261
80. Lahiri, S. K.; Khalfe, N. *Can. J. Chem. Eng.* **2009,** *87,* 118–128. doi:10.1002/cjce.20123
81. Schweidtmann, A. M.; Clayton, A. D.; Holmes, N.; Bradford, E.; Bourne, R. A.; Lapkin, A. A. *Chem. Eng. J.* **2018,** *352,* 277–282. doi:10.1016/j.cej.2018.07.031
82. Bradford, E.; Schweidtmann, A. M.; Lapkin, A. *J. Glob. Optim.* **2018,** *71,* 407–438. doi:10.1007/s10898-018-0609-2
83. Sagmeister, P.; Ort, F. F.; Jusner, C. E.; Hebrault, D.; Tampone, T.; Buono, F. G.; Williams, J. D.; Kappe, C. O. *Adv. Sci.* **2022,** *9,* 2105547. doi:10.1002/advs.202105547
84. Nambiar, A. M. K.; Breen, C. P.; Hart, T.; Kulesza, T.; Jamison, T. F.; Jensen, K. F. *ACS Cent. Sci.* **2022,** *8,* 825–836. doi:10.1021/acscentsci.2c00207
85. Clayton, A. D.; Pyzer-Knapp, E. O.; Purdie, M.; Jones, M. F.; Barthelme, A.; Pavey, J.; Kapur, N.; Chamberlain, T. W.; Blacker, A. J.; Bourne, R. A. *Angew. Chem., Int. Ed.* **2023,** *62,* e202214511. doi:10.1002/anie.202214511
86. Clayton, A. D.; Schweidtmann, A. M.; Clemens, G.; Manson, J. A.; Taylor, C. J.; Niño, C. G.; Chamberlain, T. W.; Kapur, N.; Blacker, A. J.; Lapkin, A. A.; Bourne, R. A. *Chem. Eng. J.* **2020,** *384,* 123340. doi:10.1016/j.cej.2019.123340
87. Jeraal, M. I.; Sung, S.; Lapkin, A. A. *Chem.:Methods* **2021,** *1,* 71–77. doi:10.1002/cmtd.202000044
88. Jorayev, P.; Russo, D.; Tibbetts, J. D.; Schweidtmann, A. M.; Deutsch, P.; Bull, S. D.; Lapkin, A. A. *Chem. Eng. Sci.* **2022,** *247,* 116938. doi:10.1016/j.ces.2021.116938
89. Knoll, S.; Jusner, C. E.; Sagmeister, P.; Williams, J. D.; Hone, C. A.; Horn, M.; Kappe, C. O. *React. Chem. Eng.* **2022,** *7,* 2375–2384. doi:10.1039/d2re00208f
90. Kershaw, O. J.; Clayton, A. D.; Manson, J. A.; Barthelme, A.; Pavey, J.; Peach, P.; Mustakis, J.; Howard, R. M.; Chamberlain, T. W.; Warren, N. J.; Bourne, R. A. *Chem. Eng. J.* **2023,** *451,* 138443. doi:10.1016/j.cej.2022.138443
91. Aldulaijan, N.; Marsden, J. A.; Manson, J. A.; Clayton, A. D. *React. Chem. Eng.* **2024,** *9,* 308–316. doi:10.1039/d3re00476g
92. Taylor, C. J.; Felton, K. C.; Wigh, D.; Jeraal, M. I.; Grainger, R.; Chessari, G.; Johnson, C. N.; Lapkin, A. A. *ACS Cent. Sci.* **2023,** *9,* 957–968. doi:10.1021/acscentsci.3c00050
93. Hickman, R. J.; Ruža, J.; Tribukait, H.; Roch, L. M.; García-Durán, A. *React. Chem. Eng.* **2023,** *8,* 2284–2296. doi:10.1039/d3re00008g
94. Low, A. K. Y.; Vissol-Gaudin, E.; Lim, Y.-F.; Hippalgaonkar, K. *J. Mater. Inf.* **2023,** *3,* 11. doi:10.20517/jmi.2023.02
95. Low, A. K. Y.; Mekki-Berrada, F.; Gupta, A.; Ostudin, A.; Xie, J.; Vissol-Gaudin, E.; Lim, Y.-F.; Li, Q.; Ong, Y. S.; Khan, S. A.; Hippalgaonkar, K. *npj Comput. Mater.* **2024,** *10,* 104. doi:10.1038/s41524-024-01274-x
96. Thway, M.; Low, A. K. Y.; Khetan, S.; Dai, H.; Recatala-Gomez, J.; Chen, A. P.; Hippalgaonkar, K. *Digital Discovery* **2024,** *3,* 328–336. doi:10.1039/d3dd00202k
97. Zhang, W.; Wang, Q.; Kong, X.; Xiong, J.; Ni, S.; Cao, D.; Niu, B.; Chen, M.; Li, Y.; Zhang, R.; Wang, Y.; Zhang, L.; Li, X.; Xiong, Z.; Shi, Q.; Huang, Z.; Fu, Z.; Zheng, M. *Chem. Sci.* **2024,** *15,* 10600–10611. doi:10.1039/d4sc00924j
98. Bran, A. M.; Cox, S.; Schilter, O.; Baldassari, C.; White, A. D.; Schwaller, P. *Nat. Mach. Intell.* **2024,** *6,* 525–535. doi:10.1038/s42256-024-00832-8
99. Mirza, A.; Alampara, N.; Kunchapu, S.; Emoekabu, B.; Krishnan, A.; Wilhelmi, M.; Okereke, M.; Eberhardt, J.; Elahi, A. M.; Greiner, M.; Holick, C. T.; Gupta, T.; Asgari, M.; Glaubitz, C.; Klepsch, L. C.; Köster, Y.; Meyer, J.; Miret, S.; Hoffmann, T.; Kreth, F. A.; Ringleb, M.; Roesner, N.; Schubert, U. S.; Stafast, L. M.; Wonanke, D.; Pieler, M.; Schwaller, P.; Jablonka, K. M. *arXiv* **2024,** 2404.01475. doi:10.48550/arxiv.2404.01475
100. Liu, H.; Yin, H.; Luo, Z.; Wang, X. *Synth. Syst. Biotechnol.* **2025,** *10,* 23–38. doi:10.1016/j.synbio.2024.07.004
101. Lála, J.; O'Donoghue, O.; Shtedritski, A.; Cox, S.; Rodriques, S. G.; White, A. D. *arXiv* **2023,** 2312.07559. doi:10.48550/arxiv.2312.07559

License and Terms

The definitive version of this article is the electronic one which can be found at:
https://doi.org/10.3762/bjoc.21.3