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]

Review

Open Access

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

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

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

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.

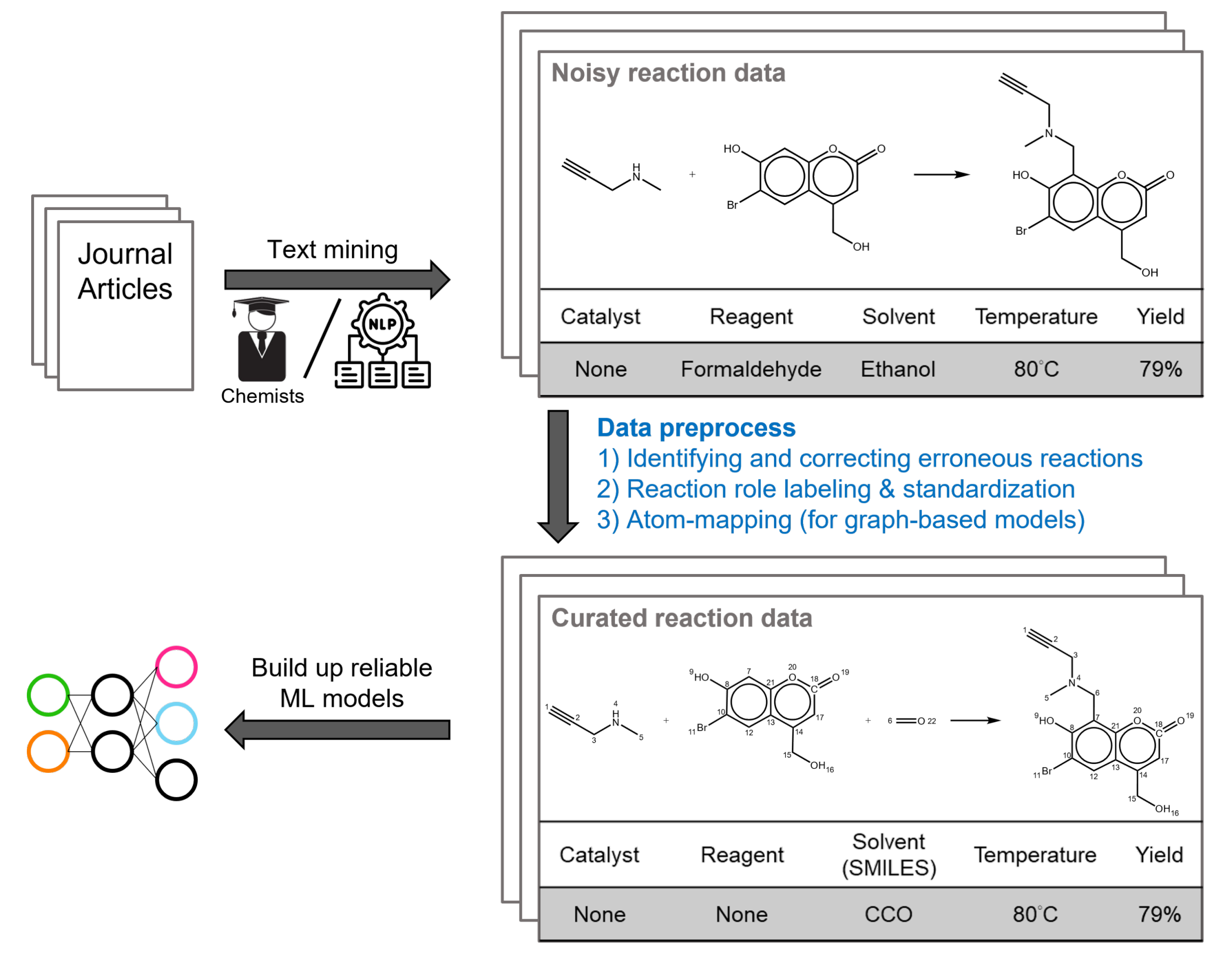
Dataset	Reference	No. of reactions
Buchwald–Hartwig (1)	[48]	4,608
Buchwald–Hartwig (2)	[49]	288
Buchwald–Hartwig (3)	[50]	750
Pd-catalyzed cross-coupling	[49]	1,536
Suzuki–Miyaura coupling (1)	[51]	5,760
Suzuki–Miyaura coupling (2)	[52]	384
Suzuki–Miyaura coupling (3)	[53]	534
electroreductive coupling 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.

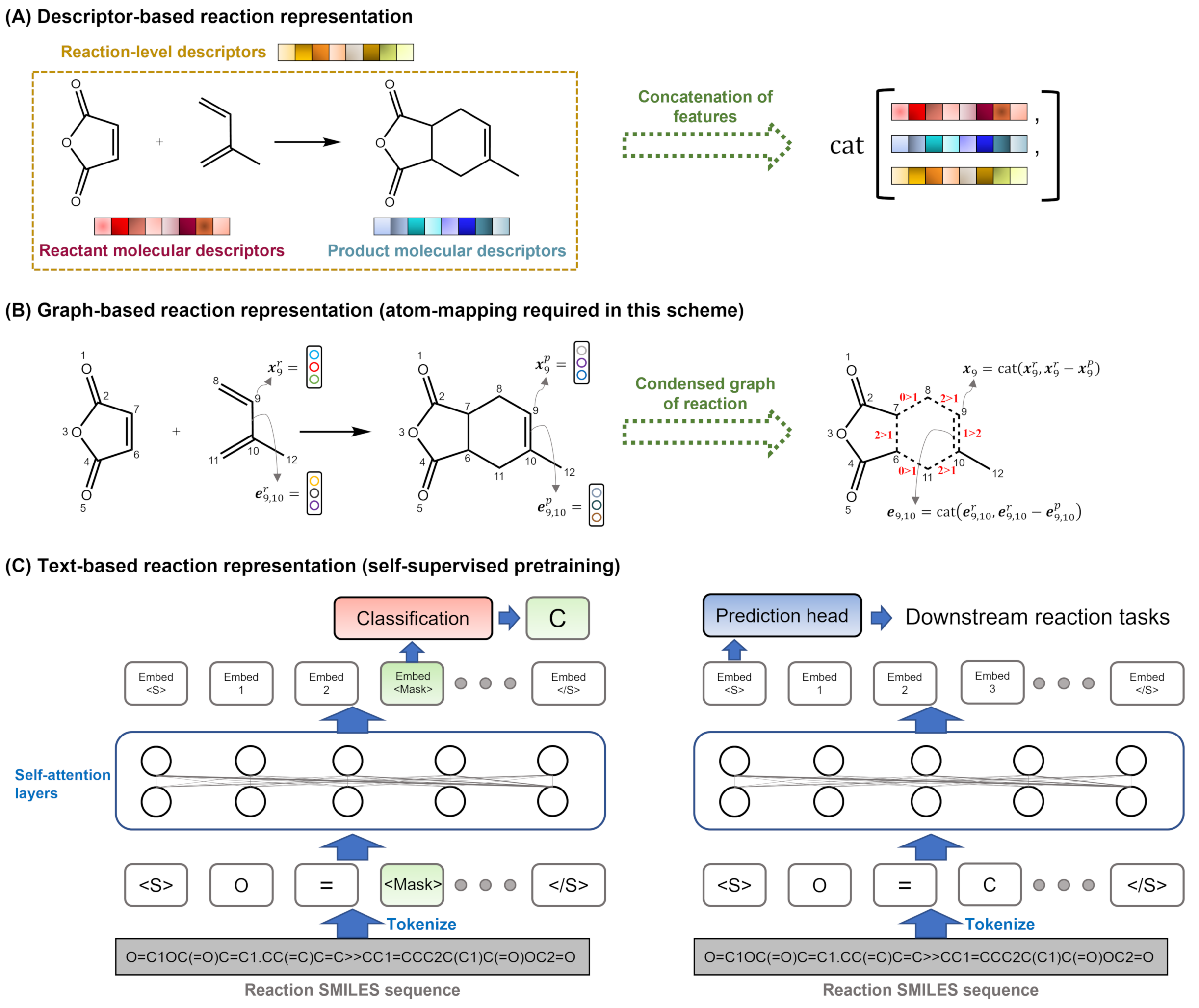

(A) Descriptor-based reaction representation
Reaction-level descriptors
Reactant molecular descriptors
Product molecular descriptors
Concatenation of features
cat
(B) Graph-based reaction representation (atom-mapping required in this scheme)
$x_9^r =$
$e_{9,10}^r =$
$x_9^p =$
$e_{9,10}^p =$
Condensed 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 fine-tuning 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].

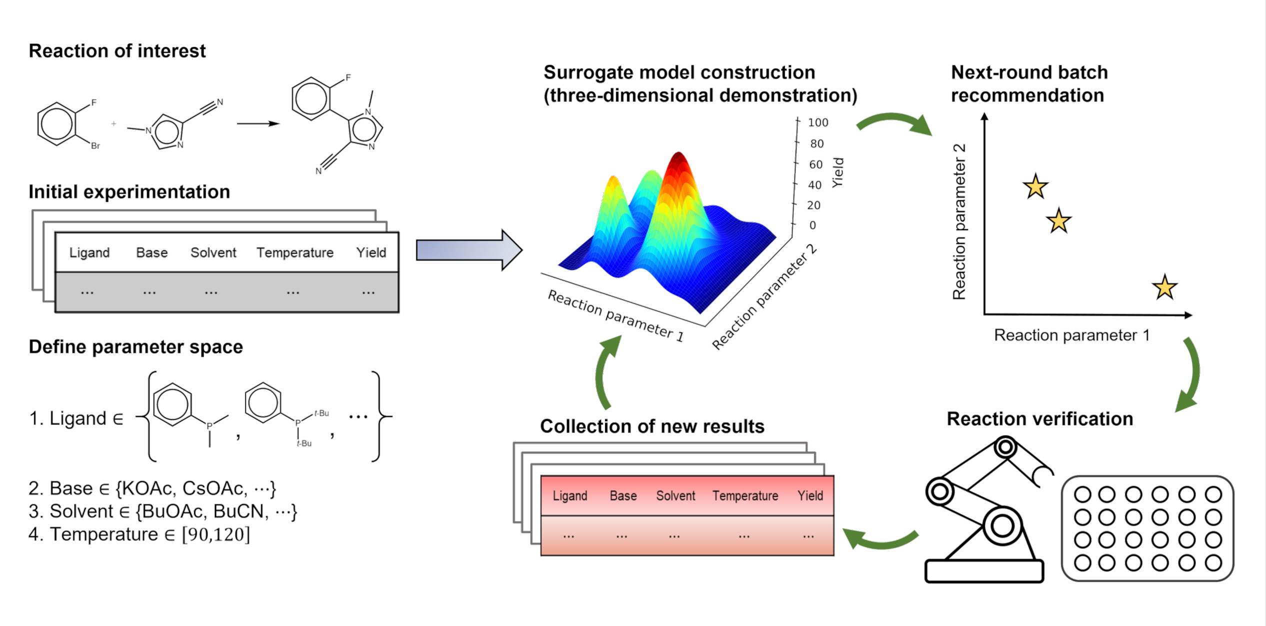


Figure 3: A schematic diagram of how ML algorithms can be combined with HTE platforms to optimize reaction conditions for CASP.

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