Machine Learning-Guided Strategies for Reaction Condition 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

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 condition design, and enable novel discoveries in synthetic chemistry.

Keywords

Reaction data mining; Data preprocessing; Reaction representation; Reaction condition prediction; Reaction optimization

1. 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]. 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 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 condition recommendation [18], and reaction condition 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 high-throughput experimentation (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.

2. Reaction data collection and preprocessing

One of the major challenges in building ML models for global reaction condition prediction is the data scarcity and diversity, as they need to cover a vast reaction space [26, 27]. Collecting experimental data for chemical reactions is not trivial, as it involves laborious and costly synthesis procedures. Some studies attempt to estimate activation energy and reaction enthalpy for specific reaction types using theoretical methods based on quantum mechanical calculations [28]. However, this approach is limited by the accuracy and applicability of the computational methods, especially for reactions in complex environments, such as those with ionic intermediates in solvents [29] or porous catalysts [30, 31]. Therefore, most ML models rely on experimental data from literature sources.

2.1 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 [32]. Most of them are proprietary and require subscription-based access, which limits the availability and comparability of data for developing global reaction condition prediction models. For example, Gao et al. [18] trained a reaction condition recommender on about 10 million reactions from Reaxys [33], but subsequent studies could not access or use the same data for model evaluation or improvement [34]. To address this issue, Coley et al. proposed the Open Reaction Database (ORD) [35], 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 [36], 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 [37]. 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 condition per reference, as discussed by Chen et al. [38]. This selection bias can lead to overestimation of reaction yields by ML models and limit their generalization capabilities [39]. Therefore, many studies have called for more comprehensive documentation of all experimental results and submission of data in machine-readable formats [40-42]. Another potential issue with data from various sources is the discrepancy in yield definition, as pointed out by Mercado et al. [43]. 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. [44] 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 [45]. Guo et al. [46] conducted a continual pretraining scheme on the BERT model [47] 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, 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. [48, 49] demonstrated that this task of optical chemical structure recognition (OCSR) [50] 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 [51, 52], 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 [53-55] that can collect information from different sources and provide robust results. Recently, Fan et al. developed the OpenChemIE toolkit [56], 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.

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

Database	Reference	No. of reactions	Availability
Reaxys	[33]	~65 millions	Proprietary
ORD	[35]	~1.7 million reactions from USPTO [36] and ~91k reactions from chemical community	Open source
Scifinder^n	[57]	~150 millions	Proprietary
Pistachio	[58]	~13 millions	Proprietary
Spresi	[59]	~4.6 millions	Proprietary

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

Dataset	Reference	No. of reactions
Buchwald-Hartwig (1)	[60]	4,608
Buchwald-Hartwig (2)	[61]	288
Buchwald-Hartwig (3)	[62]	750
Pd-catalyzed cross-coupling	[61]	1,536
Suzuki–Miyaura coupling (1)	[63]	5,760
Suzuki–Miyaura coupling (2)	[64]	384
Suzuki–Miyaura coupling (3)	[65]	534
Electroreductive coupling of alkenyl and benzyl halides	[66]	27
Mizoroki–Heck reaction	[67]	384
Coupling of α-carboxyl sp3-carbons with aryl halides	[68]	24
Biginelli condensation	[69]	48
Deoxyfluorination	[70]	80
Coupling reactions	[71]	264
Synthesis of sulfonamide	[72]	39
Ni-catalyzed Suzuki–Miyaura	[73]	450
Mitsunobu reaction	[74]	40
Ni-catalyzed borylation	[75]	1,296
Amide coupling (1)	[76]	1,280
Amide coupling (2)	[77]	960
Pd-catalysed C–H arylation	[77]	1,536
Ni-catalyzed C–O coupling	[78]	2,003

Ir(I)-catalyzed O–H bond insertion	[79]	653
Pd-catalyzed C−N coupling	[80]	767

2.2 Implicit data issues and data preprocessing tools

The quality of training data is a crucial factor for the robustness of machine learning 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. [81] proposed to use this idea as 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 [82-85] 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 [86], to add small molecules (e.g., H2O 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. [87] and Zhang et al. [88]. 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. [89], 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 [90], 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.

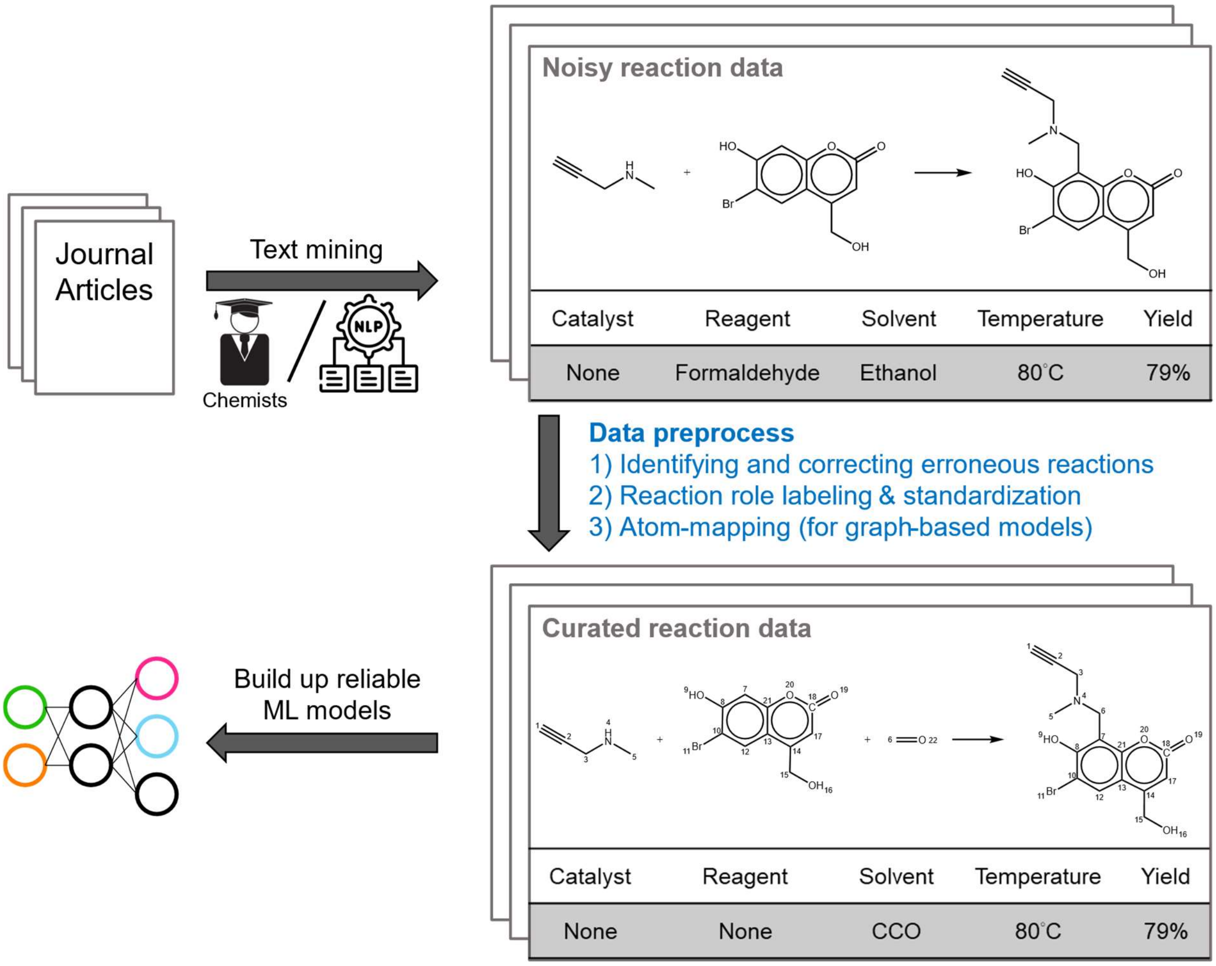


Figure 1: A schematic diagram of the data mining and preprocessing steps for chemical reaction datasets, including data collection, data filtering, data completion, and atom mapping.

3. 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 [91]. 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 condition prediction by modifying the output layer of the model. 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.

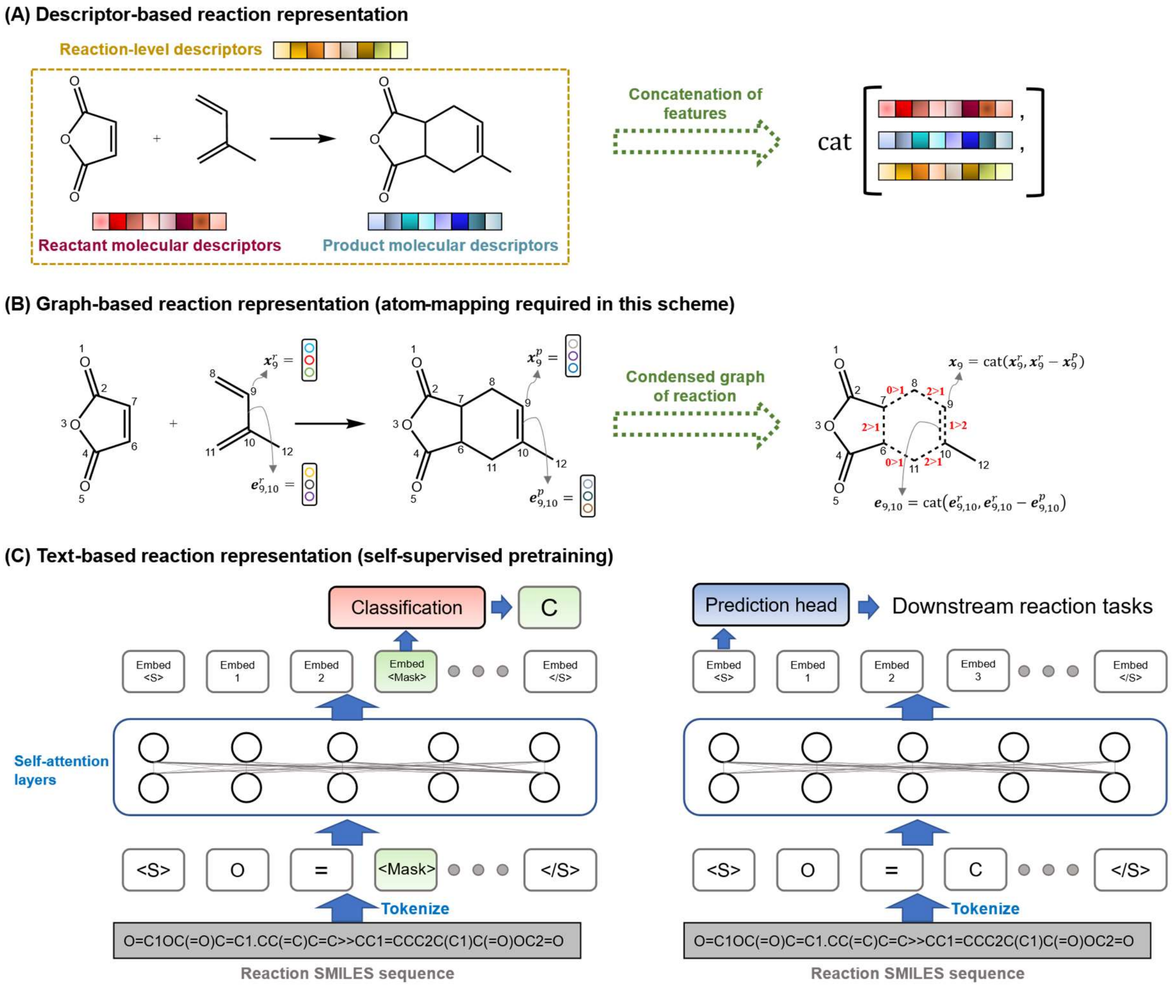

(A) Descriptor-based reaction representation
Reaction-level descriptors
Reactant molecular descriptors
Product molecular descriptors
Concatenation of features
cat
(B) Graph-based reaction representation (atom-mapping required in this scheme)
$x_9^r =$
$e_{9,10}^r =$
$x_9^p =$
$e_{9,10}^p =$
Condensed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

3.1 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 [92]. Molecular-level descriptors of reactants and products are concatenated to obtain reaction-level descriptors, which can be computed by various methods [93]. These include substructure keys-based [94-98], circular [99-101], physicochemical [102-105] and quantum chemistry (QM) features [106-110]. 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 [74]. To overcome this challenge, some studies have proposed to pre-generate QM properties datasets and train machine learning 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 transition state (TS) structures of chemical reactions can provide valuable insights for predicting rate constants [111-115], regioselectivity and site-selectivity [16, 17, 116-118]. 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 [119], and the solvent effects may complicate the results [120]. 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. [121], 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 [101]. Due to its fast computation and compatibility with conventional ML models, it has been widely used or benchmarked in various reaction-related tasks [122-126], and has become one of the mainstream reaction-level featurization techniques.

3.2 Graph-based representation

Graph neural networks (GNNs) have been widely applied to various chemical tasks, such as predicting molecular properties [127-131], reaction product prediction [132-134], and inverse materials design [135-137]. 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 [138-141], most of which are based on the message passing neural network (MPNN) framework proposed by Gilmer et al. [142].

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 [143-148] 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 [149-151] assign labels to each atom and adapt the algorithms accordingly. Grambow et al. [149] and Yarish et al. [151] 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. [150] developed a more general AAM-required reaction encoding method that operates graph convolutions on the condensed graph of reaction (CGR) [152, 153]. 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. [154]. 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 [82-85]. 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.

3.3 Text-based representation

Recent years have witnessed the emergence of large language models (LLMs) [155-157], 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 Simplified Molecular-Input Line-Entry System (SMILES) notation [158], 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 modelling (MLM) method to pretrain on millions of molecular SMILES and finetune on small-sample molecular property datasets [159-162]. For reaction-level prediction tasks, the textual input for pretraining can be changed to reaction SMILES, as shown in Figure 2C. Schwaller et al. [163] 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 modelling and its powerful reasoning abilities is the size of the pretraining data [164]. However, unlike molecular SMILES, which can be generated from existing databases (e.g., GDB-13 [165]) or by methods that produce reasonable structures [166], reaction SMILES data are often limited by the availability of experimental databases. Therefore, various data augmentation methods [167-169] 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 [170, 171], without the need for tedious chemistry-informed feature generation and selection beforehand.

4. Reaction condition 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.

4.1 Global models for direct reaction condition predictions

A common approach for chemists to synthesize novel reactions is to reference similar chemical reactions using reaction similarity search [172, 173] and adopt the reaction conditions used in the literature. Machine learning 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 condition models involves predicting the reagents used in the dataset as labels, along with the reaction temperatures, using multi-class or multi-label classification methods [174]. 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. [143] used a variational autoencoder (VAE) architecture to sample different reaction conditions, while Chen et al. [38] 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 condition 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 [35, 175] or the curated USPTO dataset [34], 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. [176] 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, 177]. Moreover, predicting reaction conditions can help optimize the synthetic route [178] by providing the necessary information for each synthetic step. Coley et al. [12] integrated ASKCOS [179], an automated CASP software, with the self-driving lab [180] and demonstrated the synthesis of 15 small molecules. Guo et al. [181] 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 [182]. 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 condition 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 condition prediction models, which use historical literature on similar chemical contexts to suggest suitable reaction conditions for synthetic steps. However, the predictive output often 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.

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.
[183]	4 types of totally ~191k reactions from Reaxys	Descriptors + GBM and GCNs	The output labels were systematically categorized with chemical insights.

[45]	~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.
[184]	~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.
[143]	4 types of totally ~191k reactions from Reaxys	GNN + VAE	The models provide multiple reaction conditions by repeatedly sampling from the VAE space.
[185]	480k USPTO-MIT dataset [132]	Reaction SMILES + Transformer	This work directly predicts SMILES representation of the combination of reaction conditions.
[34]	Curated USPTO-Condition dataset with ~680k reactions and Reaxys-TotalSyn-	Reaction SMILES + Transformer	This work demonstrates the benefits of MLM pretraining for the downstream reaction condition prediction task.

	Condition dataset with ~180k reactions		
[38]	10 types of totally ~74k reactions from Reaxys	ECFP + DNN	It models the reaction condition prediction problem as recommendation system and artificially generate fake reaction conditions for data augmentation.
[186]	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.

4.2 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.

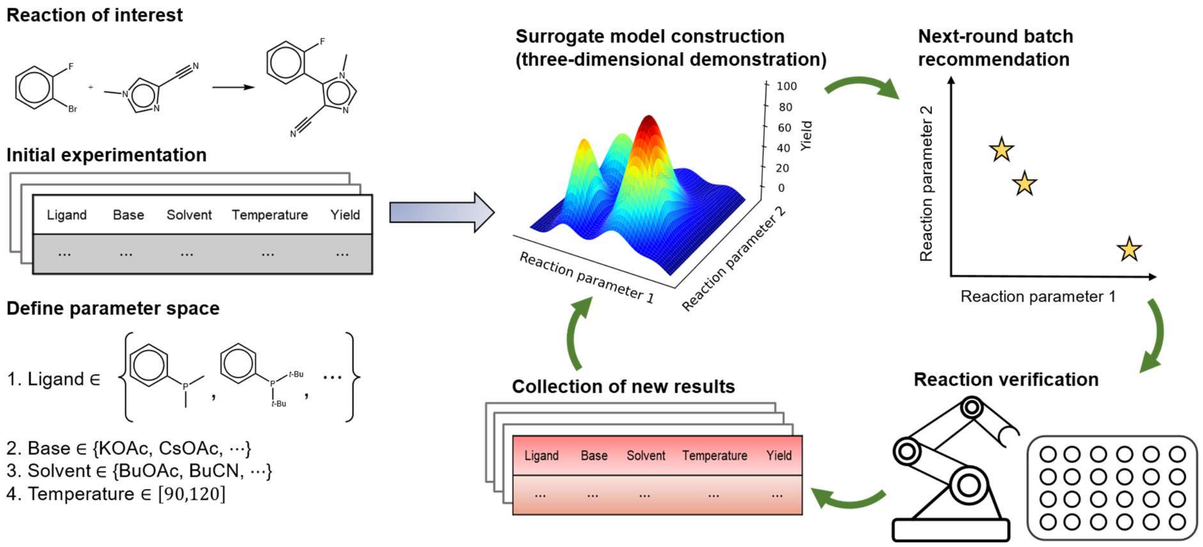
Many reaction optimization platforms have been developed [187-190], which integrate software optimization algorithms with hardware automation for experiments, enabling large-scale experimentation and data collection. Among these, Bayesian optimization (BO) [191] is the most classic and widely used algorithm, 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 Bayesian optimization lies in its ability to balance exploration and exploitation, making it highly effective for complex, multi-dimensional optimization tasks in chemical processes. BO has also demonstrated robust performance in many benchmark tasks [192-194], and numerous chemical reaction optimization packages have been developed to support this algorithm [195-199].

A typical example is the work by Shields et al. [74], who used different featurization strategies, such as DFT [106], cheminformatic [105], 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. [200] investigated the impact of integrating a global reaction condition 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 [201]. This indicates that the suboptimal combination of each reaction does not necessarily represent the global optimum for multistep synthesis [202-204]. In contrast, telescoped flow sequences [205-207] or one-pot batch synthesis [208] emphasize the use of chemically compatible reagents and solvents in each reaction step to minimize intermediate purification steps. Volk et al. [209] 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 [210-214]. For instance, the generality of chiral catalysts for asymmetric or enantioselective catalysis has been a longstanding interest in synthetic chemistry [215]. Angello et al. [65] 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. [77] 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 [216]. 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 [217]. 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 [42]. Furthermore, BO cannot generalize well from past reactions to unseen reaction transformations, which inherently requires gathering new relevant data for new chemical reactions [218]. Regarding general reaction conditions, the typically limited experimental budgets in laboratories restrict the ability to explore a diverse range of reaction conditions [77]. 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].



Ligand	Base	Solvent	Temperature	Yield
···	···	···	···	···

1. Ligand ∈ { , , ··· }
2. Base ∈ {KOAc, CsOAc, ···}
3. Solvent ∈ {BuOAc, BuCN, ···}
4. Temperature ∈ [90,120]

Ligand	Base	Solvent	Temperature	Yield
···	···	···	···	···

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

5. Outlook and perspectives

In conclusion, this review paper has demonstrated the importance of reaction conditions in CASP and the potential of ML to assist in their design. We have discussed the current state of the art in data collection, data preprocessing, model development, global prediction, and local optimization for reaction condition design using ML. We have also identified some of the challenges and limitations that need to be addressed in future research, such as the quality and availability of reaction datasets and the cost and accessibility of automated reaction optimization tools. We hope that this review paper will inspire researchers to adopt ML approaches for reaction condition design and to collaborate across disciplines of organic synthesis, process engineering, and ML algorithms. This will enable the development of more efficient, sustainable, and innovative synthetic pathways for CASP.

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 Young Scholar Fellowship Einstein Program (112-2636-E-002-005) and the Higher Education Sprout Project by the Ministry of Education in Taiwan (113L891305).

References

1. Chen, S.; Jung, Y. *JACS Au* **2021**, *1* (10), 1612-1620.
2. Finnigan, W.; Hepworth, L. J.; Flitsch, S. L.; Turner, N. J. *Nature catalysis* **2021**, *4* (2), 98-104.
3. Fortunato, M. E.; Coley, C. W.; Barnes, B. C.; Jensen, K. F. *Journal of chemical information and modeling* **2020**, *60* (7), 3398-3407.
4. Thakkar, A.; Johansson, S.; Jorner, K.; Buttar, D.; Reymond, J.-L.; Engkvist, O. *Reaction chemistry & engineering* **2021**, *6* (1), 27-51.
5. Klucznik, T.; Mikulak-Klucznik, B.; McCormack, M. P.; Lima, H.; Szymkuć, S.; Bhowmick, M.; Molga, K.; Zhou, Y.; Rickershauser, L.; Gajewska, E. P. *Chem* **2018**, *4* (3), 522-532.

6. Mo, Y.; Guan, Y.; Verma, P.; Guo, J.; Fortunato, M. E.; Lu, Z.; Coley, C. W.; Jensen, K. F. *Chemical science* **2021**, *12* (4), 1469-1478.
7. Molga, K.; Dittwald, P.; Grzybowski, B. A. *Chem* **2019**, *5* (2), 460-473.
8. Sankaranarayanan, K.; Jensen, K. F. *Chemical Science* **2023**, *14* (23), 6467-6475.
9. Ha, T.; Lee, D.; Kwon, Y.; Park, M. S.; Lee, S.; Jang, J.; Choi, B.; Jeon, H.; Kim, J.; Choi, H. *Science advances* **2023**, *9* (44), eadj0461.
10. Hardwick, T.; Ahmed, N. *Chemical Science* **2020**, *11* (44), 11973-11988.
11. Shen, Y.; Borowski, J. E.; Hardy, M. A.; Sarpong, R.; Doyle, A. G.; Cernak, T. *Nature Reviews Methods Primers* **2021**, *1* (1), 1-23.
12. Coley, C. W.; Thomas III, D. A.; Lummiss, J. A.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H. *Science* **2019**, *365* (6453), eaax1566.
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. *Chemical Reviews* **2023**, *123* (6), 3089-3126.
14. Voinarovska, V.; Kabeshov, M.; Dudenko, D.; Genheden, S.; Tetko, I. V. *Journal of Chemical Information and Modeling* **2023**, *64* (1), 42-56.
15. Zuranski, A. M.; Martinez Alvarado, J. I.; Shields, B. J.; Doyle, A. G. *Accounts of chemical research* **2021**, *54* (8), 1856-1865.
16. Guan, Y.; Coley, C. W.; Wu, H.; Ranasinghe, D.; Heid, E.; Struble, T. J.; Pattanaik, L.; Green, W. H.; Jensen, K. F. *Chemical Science* **2021**, *12* (6), 2198-2208.
17. Struble, T. J.; Coley, C. W.; Jensen, K. F. *Reaction Chemistry & Engineering* **2020**, *5* (5), 896-902.
18. Gao, H.; Struble, T. J.; Coley, C. W.; Wang, Y.; Green, W. H.; Jensen, K. F. *ACS central science* **2018**, *4* (11), 1465-1476.
19. Abolhasani, M.; Kumacheva, E. *Nature Synthesis* **2023**, *2* (6), 483-492.
20. Raghavan, P.; Haas, B. C.; Ruos, M. E.; Schleinitz, J.; Doyle, A. G.; Reisman, S. E.; Sigman, M. S.; Coley, C. W. *ACS Central Science* **2023**, *9* (12), 2196-2204.
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. *Science* **2023**, *382* (6677), eadi1407.
22. Eyke, N. S.; Koscher, B. A.; Jensen, K. F. *Trends in Chemistry* **2021**, *3* (2), 120-132.
23. Krska, S. W.; DiRocco, D. A.; Dreher, S. D.; Shevlin, M. *Accounts of chemical research* **2017**, *50* (12), 2976-2985.
24. Shevlin, M. *ACS medicinal chemistry letters* **2017**, *8* (6), 601-607.
25. Snoek, J.; Larochelle, H.; Adams, R. P. *Advances in neural information processing systems* **2012**, *25*.
26. Andronov, M.; Fedorov, M. V.; Sosnin, S. *ACS omega* **2021**, *6* (45), 30743-30751.
27. Dobson, C. M. *Nature* **2004**, *432* (7019).
28. Jorner, K.; Tomberg, A.; Bauer, C.; Sköld, C.; Norrby, P.-O. *Nature Reviews Chemistry* **2021**, *5* (4), 240-255.
29. Plata, R. E.; Singleton, D. A. *Journal of the American Chemical Society* **2015**, *137* (11), 3811-3826.
30. Li, S.-C.; Lin, Y.-C.; Li, Y.-P. *Catalysts* **2021**, *11* (9), 1114.
31. Yeh, J. Y.; Li, S. C.; Chen, C. H.; Wu, K. C. W.; Li, Y. P. *Chemistry–An Asian Journal* **2021**, *16* (9), 1049-1056.

32. Thakkar, A.; Kogej, T.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. J. *Chemical science* **2020**, *11* (1), 154-168.
33. Reaxys. https://www.reaxys.com/ (accessed March 25, 2024).
34. Wang, X.; Hsieh, C.-Y.; Yin, X.; Wang, J.; Li, Y.; Deng, Y.; Jiang, D.; Wu, Z.; Du, H.; Chen, H. *Research* **2023**, *6*, 0231.
35. 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. *Journal of the American Chemical Society* **2021**, *143* (45), 18820-18826.
36. 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).
37. Mahjour, B.; Shen, Y.; Cernak, T. *Accounts of Chemical Research* **2021**, *54* (10), 2337-2346.
38. Chen, L.-Y.; Li, Y.-P. *Journal of Cheminformatics* **2024**, *16* (1), 11.
39. Strieth-Kalthoff, F.; Sandfort, F.; Kühnemund, M.; Schäfer, F. R.; Kuchen, H.; Glorius, F. *Angewandte Chemie International Edition* **2022**, *61* (29), e202204647.
40. 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. *Angewandte Chemie International Edition* **2022**, *61* (51), e202203038.
41. Hunter, A. M.; Carreira, E. M.; Miller, S. J., Encouraging Submission of FAIR Data at The Journal of Organic Chemistry and Organic Letters. ACS Publications: 2020; Vol. 85, pp 1773-1774.
42. Maloney, M. P.; Coley, C. W.; Genheden, S.; Carson, N.; Helquist, P.; Norrby, P.-O.; Wiest, O. *Organic Letters* **2023**, *25* (17), 2945-2947.
43. Mercado, R.; Kearnes, S. M.; Coley, C. W. *Journal of Chemical Information and Modeling* **2023**, *63* (14), 4253-4265.
44. Vaucher, A. C.; Zipoli, F.; Geluykens, J.; Nair, V. H.; Schwaller, P.; Laino, T. *Nature communications* **2020**, *11* (1), 3601.
45. Vaucher, A. C.; Schwaller, P.; Geluykens, J.; Nair, V. H.; Iuliano, A.; Laino, T. *Nature communications* **2021**, *12* (1), 2573.
46. Guo, J.; Ibanez-Lopez, A. S.; Gao, H.; Quach, V.; Coley, C. W.; Jensen, K. F.; Barzilay, R. *Journal of chemical information and modeling* **2021**, *62* (9), 2035-2045.
47. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. *arXiv preprint arXiv:1810.04805* **2018**.
48. Qian, Y.; Guo, J.; Tu, Z.; Coley, C. W.; Barzilay, R. *Journal of Chemical Information and Modeling* **2023**, *63* (13), 4030-4041.
49. Qian, Y.; Guo, J.; Tu, Z.; Li, Z.; Coley, C. W.; Barzilay, R. *Journal of Chemical Information and Modeling* **2023**, *63* (7), 1925-1934.
50. McDaniel, J. R.; Balmuth, J. R. *Journal of chemical information and computer sciences* **1992**, *32* (4), 373-378.
51. Beard, E. J.; Cole, J. M. *Journal of chemical information and modeling* **2020**, *60* (4), 2059-2072.
52. Wilary, D. M.; Cole, J. M. *Journal of Chemical Information and Modeling* **2021**, *61* (10), 4962-4974.
53. Huang, H.; Zheng, O.; Wang, D.; Yin, J.; Wang, Z.; Ding, S.; Yin, H.; Xu, C.; Yang, R.; Zheng, Q. *International Journal of Oral Science* **2023**, *15* (1), 29.
54. Song, B.; Zhou, R.; Ahmed, F. *Journal of Computing and Information Science in Engineering* **2024**, *24* (1), 010801.

55. Wang, X.; Chen, G.; Qian, G.; Gao, P.; Wei, X.-Y.; Wang, Y.; Tian, Y.; Gao, W. *Machine Intelligence Research* **2023**, *20* (4), 447-482.
56. Fan, V.; Qian, Y.; Wang, A.; Wang, A.; Coley, C. W.; Barzilay, R. *arXiv preprint arXiv:2404.01462* **2024**.
57. CAS, SciFinder-n. https://scifinder-n.cas.org/ (accessed March 25, 2024).
58. Nextmove Software Pistachio. https://www.nextmovesoftware.com/pistachio.html (accessed March 25, 2024).
59. Roth, D. L., SPRESIweb 2.1, a selective chemical synthesis and reaction database. ACS Publications: 2005.
60. Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. *Science* **2018**, *360* (6385), 186-190.
61. Buitrago Santanilla, A.; Regalado, E. L.; Pereira, T.; Shevlin, M.; Bateman, K.; Campeau, L.-C.; Schneeweis, J.; Berritt, S.; Shi, Z.-C.; Nantermet, P. *Science* **2015**, *347* (6217), 49-53.
62. 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. *Chemical Science* **2023**, *14* (19), 4997-5005.
63. Perera, D.; Tucker, J. W.; Brahmbhatt, S.; Helal, C. J.; Chong, A.; Farrell, W.; Richardson, P.; Sach, N. W. *Science* **2018**, *359* (6374), 429-434.
64. Reizman, B. J.; Wang, Y.-M.; Buchwald, S. L.; Jensen, K. F. *Reaction chemistry & engineering* **2016**, *1* (6), 658-666.
65. 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. *Science* **2022**, *378* (6618), 399-405.
66. DeLano, T. J.; Reisman, S. E. *ACS catalysis* **2019**, *9* (8), 6751-6754.
67. 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. *Journal of the American Chemical Society* **2023**.
68. Zuo, Z.; Ahneman, D. T.; Chu, L.; Terrett, J. A.; Doyle, A. G.; MacMillan, D. W. *Science* **2014**, *345* (6195), 437-440.
69. Stadler, A.; Kappe, C. O. *Journal of combinatorial chemistry* **2001**, *3* (6), 624-630.
70. Nielsen, M. K.; Ahneman, D. T.; Riera, O.; Doyle, A. G. *Journal of the American Chemical Society* **2018**, *140* (15), 5004-5008.
71. 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. *Chemical science* **2016**, *7* (4), 2604-2613.
72. Gioiello, A.; Rosatelli, E.; Teofrasti, M.; Filipponi, P.; Pellicciari, R. *ACS Combinatorial Science* **2013**, *15* (5), 235-239.
73. 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* (6565), 301-308.
74. 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* (7844), 89-96.
75. 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* (14), 1847-1864.
76. Mahjour, B.; Zhang, R.; Shen, Y.; McGrath, A.; Zhao, R.; Mohamed, O. G.; Lin, Y.; Zhang, Z.; Douthwaite, J. L.; Tripathi, A. *Nature Communications* **2023**, *14* (1), 3924.

77. 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. *Nature* **2024**, *626* (8001), 1025-1033.
78. Schleinitz, J.; Langevin, M.; Smail, Y.; Wehnert, B.; Grimaud, L.; Vuilleumier, R. *Journal of the American Chemical Society* **2022**, *144* (32), 14722-14730.
79. Xu, Y.; Ren, F.; Su, L.; Xiong, Z.; Zhu, X.; Lin, X.; Qiao, N.; Tian, H.; Tian, C.; Liao, K. *Organic Chemistry Frontiers* **2023**, *10* (5), 1153-1159.
80. Fitzner, M.; Wuitschik, G.; Koller, R.; Adam, J.-M.; Schindler, T. *ACS omega* **2023**, *8* (3), 3017-3025.
81. Toniato, A.; Schwaller, P.; Cardinale, A.; Geluykens, J.; Laino, T. *Nature Machine Intelligence* **2021**, *3* (6), 485-494.
82. Chen, S.; An, S.; Babazade, R.; Jung, Y. *Nature Communications* **2024**, *15* (1), 2250.
83. Chen, W. L.; Chen, D. Z.; Taylor, K. T. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2013**, *3* (6), 560-593.
84. Nugmanov, R.; Dyubankova, N.; Gedich, A.; Wegner, J. K. *Journal of Chemical Information and Modeling* **2022**, *62* (14), 3307-3315.
85. Schwaller, P.; Hoover, B.; Reymond, J.-L.; Strobelt, H.; Laino, T. *Science Advances* **2021**, *7* (15), eabe4166.
86. Nugmanov, R. I.; Mukhametgaleev, R. N.; Akhmetshin, T.; Gimadiev, T. R.; Afonina, V. A.; Madzhidov, T. I.; Varnek, A. *Journal of chemical information and modeling* **2019**, *59* (6), 2516-2521.
87. Zipoli, F.; Ayadi, Z.; Schwaller, P.; Laino, T.; Vaucher, A. C. *Machine Learning: Science and Technology* **2024**.
88. Zhang, C.; Arun, A.; Lapkin, A. A. *ACS omega* **2024**.
89. Phan, T.-L.; Weinbauer, K.; Gärtner, T.; Merkle, D.; Andersen, J. L.; Fagerberg, R.; Stadler, P. F. *Chemrxiv* **2024**.
90. Chen, L.-Y.; Li, Y.-P. *Journal of Cheminformatics* **2024**, *16* (1), 74.
91. Ding, Y.; Qiang, B.; Chen, Q.; Liu, Y.; Zhang, L.; Liu, Z. *Journal of Chemical Information and Modeling* **2024**.
92. *2016 3rd international conference on computing for sustainable global development (INDIACom)*, Ieee: 2016.
93. Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. *Methods* **2015**, *71*, 58-63.
94. Barnard, J. M.; Downs, G. M. *Journal of chemical information and computer sciences* **1997**, *37* (1), 141-142.
95. Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. PubChem: integrated platform of small molecules and biological activities. In *Annual reports in computational chemistry*; Elsevier: 2008; Vol. 4, pp 217-241.
96. Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. *Journal of chemical information and computer sciences* **2002**, *42* (6), 1273-1280.
97. Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. *Journal of chemical information and computer sciences* **1996**, *36* (1), 128-136.
98. Tovar, A.; Eckert, H.; Bajorath, J. *ChemMedChem: Chemistry Enabling Drug Discovery* **2007**, *2* (2), 208-217.
99. Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. *Journal of chemical information and computer sciences* **2004**, *44* (1), 170-178.
100. Probst, D.; Reymond, J.-L. *Journal of cheminformatics* **2018**, *10*, 1-12.
101. Rogers, D.; Hahn, M. *Journal of chemical information and modeling* **2010**, *50* (5), 742-754.

102. Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. *Journal of Chemical Information and Computer Sciences* **1996**, *36* (1), 118-127.
103. Raevsky, O. A. *Mini reviews in medicinal chemistry* **2004**, *4* (10), 1041-1052.
104. Zhang, Q.-Y.; Aires-de-Sousa, J. *Journal of chemical information and modeling* **2007**, *47* (1), 1-8.
105. Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. *Journal of cheminformatics* **2018**, *10* (1), 1-14.
106. Żurański, A. M.; Wang, J. Y.; Shields, B. J.; Doyle, A. G. *Reaction Chemistry & Engineering* **2022**, *7* (6), 1276-1284.
107. Li, S.-C.; Wu, H.; Menon, A.; Spiekermann, K.; Li, Y.-P.; Green, W. *Chemrxiv* **2024**.
108. Low, K.; Coote, M. L.; Izgorodina, E. I. *Journal of Chemical Theory and Computation* **2023**, *19* (5), 1466-1475.
109. *Journal of Chemical Theory and Computation* **2022**, *18* (3), 1607-1618.
110. Neeser, R. M.; Isert, C.; Stuyver, T.; Schneider, G.; Coley, C. W. *Chemical Data Collections* **2023**, *46*, 101040.
111. Al Ibrahim, E.; Farooq, A. *The Journal of Physical Chemistry A* **2022**, *126* (28), 4617-4629.
112. Komp, E.; Janulaitis, N.; Valleau, S. *Physical Chemistry Chemical Physics* **2022**, *24* (5), 2692-2705.
113. Sanches-Neto, F. O.; Dias-Silva, J. R.; Keng Queiroz Junior, L. H.; Carvalho-Silva, V. H. *Environmental Science & Technology* **2021**, *55* (18), 12437-12448.
114. Zhang, Y.; Yu, J.; Song, H.; Yang, M. *Journal of Chemical Information and Modeling* **2023**, *63* (16), 5097-5106.
115. Johnson, M. S.; Green, W. H. *Reaction Chemistry & Engineering* **2024**.
116. Beker, W.; Gajewska, E. P.; Badowski, T.; Grzybowski, B. A. *Angewandte Chemie International Edition* **2019**, *58* (14), 4515-4519.
117. Li, X.; Zhang, S. Q.; Xu, L. C.; Hong, X. *Angewandte Chemie International Edition* **2020**, *59* (32), 13253-13259.
118. Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. *Science* **2019**, *363* (6424), eaau5631.
119. 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. *Journal of chemical theory and computation* **2017**, *13* (11), 5780-5797.
120. Liu, S.-C.; Zhu, X.-R.; Liu, D.-Y.; Fang, D.-C. *Physical Chemistry Chemical Physics* **2023**, *25* (2), 913-931.
121. Probst, D.; Schwaller, P.; Reymond, J.-L. *Digital discovery* **2022**, *1* (2), 91-97.
122. Chen, K.; Chen, G.; Li, J.; Huang, Y.; Wang, E.; Hou, T.; Heng, P.-A. *Journal of Cheminformatics* **2023**, *15* (1), 43.
123. Kroll, A.; Rousset, Y.; Hu, X.-P.; Liebrand, N. A.; Lercher, M. J. *Nature Communications* **2023**, *14* (1), 4139.
124. Neves, P.; McClure, K.; Verhoeven, J.; Dyubankova, N.; Nugmanov, R.; Gedich, A.; Menon, S.; Shi, Z.; Wegner, J. K. *Journal of cheminformatics* **2023**, *15* (1), 20.
125. Ranković, B.; Griffiths, R.-R.; Moss, H. B.; Schwaller, P. *Digital Discovery* **2024**.
126. Wen, M.; Blau, S. M.; Xie, X.; Dwaraknath, S.; Persson, K. A. *Chemical science* **2022**, *13* (5), 1446-1458.
127. Chen, L.-Y.; Hsu, T.-W.; Hsiung, T.-C.; Li, Y.-P. *The Journal of Physical Chemistry A* **2022**, *126* (41), 7548-7556.

128. Li, Y.-P.; Han, K.; Grambow, C. A.; Green, W. H. *The Journal of Physical Chemistry A* **2019**, *123* (10), 2142-2152.
129. Lin, Y.-H.; Liang, H.-H.; Lin, S.-T.; Li, Y.-P. **2024**.
130. Muthiah, B.; Li, S.-C.; Li, Y.-P. *Journal of the Taiwan Institute of Chemical Engineers* **2023**, *151*, 105123.
131. Yang, C.-I.; Li, Y.-P. *Journal of Cheminformatics* **2023**, *15* (1), 13.
132. Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. *Chemical science* **2019**, *10* (2), 370-377.
133. Keto, A.; Guo, T.; Underdue, M.; Stuyver, T.; Coley, C.; Zhang, X.; Krenske, E.; Wiest, O. **2024**.
134. Wu, Y.; Zhang, C.; Wang, L.; Duan, H. *Chemical Communications* **2021**, *57* (34), 4114-4117.
135. Dold, D.; van Egmond, D. A. *Cell Reports Physical Science* **2023**, *4* (10).
136. Wang, Q.; Zhang, L. *Nature communications* **2021**, *12* (1), 5359.
137. Xiong, J.; Xiong, Z.; Chen, K.; Jiang, H.; Zheng, M. *Drug discovery today* **2021**, *26* (6), 1382-1393.
138. Wang, Y.; Wang, J.; Cao, Z.; Barati Farimani, A. *Nature Machine Intelligence* **2022**, *4* (3), 279-287.
139. Wieder, O.; Kohlbacher, S.; Kuenemann, M.; Garon, A.; Ducrot, P.; Seidel, T.; Langer, T. *Drug Discovery Today: Technologies* **2020**, *37*, 1-12.
140. Zang, X.; Zhao, X.; Tang, B. *Communications Chemistry* **2023**, *6* (1), 34.
141. Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; Sun, M. *AI open* **2020**, *1*, 57-81.
142. *International conference on machine learning*, PMLR: 2017.
143. Kwon, Y.; Kim, S.; Choi, Y.-S.; Kang, S. *Journal of Chemical Information and Modeling* **2022**, *62* (23), 5952-5960.
144. Li, B.; Su, S.; Zhu, C.; Lin, J.; Hu, X.; Su, L.; Yu, Z.; Liao, K.; Chen, H. *Journal of Cheminformatics* **2023**, *15* (1), 72.
145. Kwon, Y.; Lee, D.; Choi, Y.-S.; Kang, S. *Journal of Cheminformatics* **2022**, *14*, 1-10.
146. Kwon, Y.; Lee, D.; Kim, J. W.; Choi, Y.-S.; Kim, S. *ACS omega* **2022**, *7* (49), 44939-44950.
147. Li, S.-W.; Xu, L.-C.; Zhang, C.; Zhang, S.-Q.; Hong, X. *Nature Communications* **2023**, *14* (1), 3569.
148. Tavakoli, M.; Shmakov, A.; Ceccarelli, F.; Baldi, P. *arXiv preprint arXiv:2201.01196* **2022**.
149. Grambow, C. A.; Pattanaik, L.; Green, W. H. *The journal of physical chemistry letters* **2020**, *11* (8), 2992-2997.
150. Heid, E.; Green, W. H. *Journal of Chemical Information and Modeling* **2021**, *62* (9), 2101-2110.
151. Yarish, D.; Garkot, S.; Grygorenko, O. O.; Radchenko, D. S.; Moroz, Y. S.; Gurbych, O. *Journal of Computational Chemistry* **2023**, *44* (2), 76-92.
152. Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. P. *Journal of computer-aided molecular design* **2005**, *19*, 693-703.
153. Gimadiev, T.; Nugmanov, R.; Khakimova, A.; Fatykhova, A.; Madzhidov, T.; Sidorov, P.; Varnek, A. *Journal of Chemical Information and Modeling* **2021**, *62* (9), 2015-2020.
154. Lin, A.; Dyubankova, N.; Madzhidov, T. I.; Nugmanov, R. I.; Verhoeven, J.; Gimadiev, T. R.; Afonina, V. A.; Ibragimova, Z.; Rakhimbekova, A.; Sidorov, P. *Molecular Informatics* **2022**, *41* (4), 2100138.

155. Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y. *ACM Transactions on Intelligent Systems and Technology* **2023**.

156. Kasneci, E.; Seßler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E. *Learning and individual differences* **2023**, *103*, 102274.

157. Thirunavukarasu, A. J.; Ting, D. S. J.; Elangovan, K.; Gutierrez, L.; Tan, T. F.; Ting, D. S. W. *Nature medicine* **2023**, *29* (8), 1930-1940.

158. Weininger, D.; Weininger, A.; Weininger, J. L. *Journal of chemical information and computer sciences* **1989**, *29* (2), 97-101.

159. Chithrananda, S.; Grand, G.; Ramsundar, B. *arXiv preprint arXiv:2010.09885* **2020**.

160. Li, J.; Jiang, X. *Wireless Communications and Mobile Computing* **2021**, *2021*, 1-7.

161. *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, 2019.

162. Wu, Z.; Jiang, D.; Wang, J.; Zhang, X.; Du, H.; Pan, L.; Hsieh, C.-Y.; Cao, D.; Hou, T. *Briefings in Bioinformatics* **2022**, *23* (3), bbac131.

163. Schwaller, P.; Probst, D.; Vaucher, A. C.; Nair, V. H.; Kreutter, D.; Laino, T.; Reymond, J.-L. *Nature machine intelligence* **2021**, *3* (2), 144-152.

164. Frey, N. C.; Soklaski, R.; Axelrod, S.; Samsi, S.; Gomez-Bombarelli, R.; Coley, C. W.; Gadepally, V. *Nature Machine Intelligence* **2023**, *5* (11), 1297-1305.

165. Blum, L. C.; Reymond, J.-L. *Journal of the American Chemical Society* **2009**, *131* (25), 8732-8733.

166. Ma, R.; Luo, T. *Journal of Chemical Information and Modeling* **2020**, *60* (10), 4684-4690.

167. Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. **2020**.

168. Wu, X.; Zhang, Y.; Yu, J.; Zhang, C.; Qiao, H.; Wu, Y.; Wang, X.; Wu, Z.; Duan, H. *Scientific Reports* **2022**, *12* (1), 17098.

169. Zhong, Z.; Song, J.; Feng, Z.; Liu, T.; Jia, L.; Yao, S.; Wu, M.; Hou, T.; Song, M. *Chemical Science* **2022**, *13* (31), 9023-9034.

170. Jaume-Santero, F.; Bornet, A.; Valery, A.; Naderi, N.; Vicente Alvarez, D.; Proios, D.; Yazdani, A.; Bournez, C.; Fessard, T.; Teodoro, D. *Journal of chemical information and modeling* **2023**, *63* (7), 1914-1924.

171. Lu, J.; Zhang, Y. *Journal of Chemical Information and Modeling* **2022**, *62* (6), 1376-1387.

172. Dobbelaere, M. R.; Lengyel, I.; Stevens, C. V.; Van Geem, K. M. *Journal of Cheminformatics* **2024**, *16* (1), 1-14.

173. Hu, Q.-N.; Deng, Z.; Hu, H.; Cao, D. S.; Liang, Y. Z. *Bioinformatics* **2011**, *27* (17), 2465-2467.

174. Zhang, M.-L.; Zhou, Z.-H. *IEEE transactions on knowledge and data engineering* **2013**, *26* (8), 1819-1837.

175. Wigh, D. S.; Arrowsmith, J.; Pomberger, A.; Felton, K. C.; Lapkin, A. A. *Journal of Chemical Information and Modeling* **2023**.

176. Beker, W.; Roszak, R.; Wołos, A.; Angello, N. H.; Rathore, V.; Burke, M. D.; Grzybowski, B. A. *Journal of the American Chemical Society* **2022**, *144* (11), 4819-4827.

177. Wang, W.; Liu, Y.; Wang, Z.; Hao, G.; Song, B. *Chemical Science* **2022**, *13* (43), 12604-12615.

178. Griffin, D. J.; Coley, C. W.; Frank, S. A.; Hawkins, J. M.; Jensen, K. F. *Organic Process Research & Development* **2023**, *27* (11), 1868-1879.

179. ASKCOS. https://askcos.mit.edu/ (accessed May 27, 2024).
180. Seifrid, M.; Pollice, R.; Aguilar-Granda, A.; Morgan Chan, Z.; Hotta, K.; Ser, C. T.; Vestfrid, J.; Wu, T. C.; Aspuru-Guzik, A. *Accounts of Chemical Research* **2022**, *55* (17), 2454-2466.
181. Guo, J.; Yu, C.; Li, K.; Zhang, Y.; Wang, G.; Li, S.; Dong, H. *Journal of Chemical Theory and Computation* **2024**.
182. Janet, J. P.; Mervin, L.; Engkvist, O. *Current Opinion in Structural Biology* **2023**, *80*, 102575.
183. Maser, M. R.; Cui, A. Y.; Ryou, S.; DeLano, T. J.; Yue, Y.; Reisman, S. E. *Journal of Chemical Information and Modeling* **2021**, *61* (1), 156-166.
184. Genheden, S.; Mårdh, A.; Lahti, G.; Engkvist, O.; Olsson, S.; Kogej, T. *Molecular Informatics* **2022**, *41* (8), 2100294.
185. Andronov, M.; Voinarovska, V.; Andronova, N.; Wand, M.; Clevert, D.-A.; Schmidhuber, J. *Chemical Science* **2023**, *14* (12), 3235-3246.
186. Qian, Y.; Li, Z.; Tu, Z.; Coley, C. W.; Barzilay, R. *arXiv preprint arXiv:2312.04881* **2023**.
187. Sim, M.; Vakili, M. G.; Strieth-Kalthoff, F.; Hao, H.; Hickman, R. J.; Miret, S.; Pablo-García, S.; Aspuru-Guzik, A. *Matter* **2023**.
188. Hammer, A. J.; Leonov, A. I.; Bell, N. L.; Cronin, L. *JACS Au* **2021**, *1* (10), 1572-1587.
189. 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. *Journal of the American Chemical Society* **2022**, *144* (43), 19999-20007.
190. Ruan, Y.; Lin, S.; Mo, Y. *Journal of Chemical Information and Modeling* **2023**, *63* (3), 770-781.
191. Frazier, P. I. *arXiv preprint arXiv:1807.02811* **2018**.
192. Häse, F.; Aldeghi, M.; Hickman, R. J.; Roch, L. M.; Christensen, M.; Liles, E.; Hein, J. E.; Aspuru-Guzik, A. *Machine Learning: Science and Technology* **2021**, *2* (3), 035021.
193. Hickman, R.; Parakh, P.; Cheng, A.; Ai, Q.; Schrier, J.; Aldeghi, M.; Aspuru-Guzik, A. **2023**.
194. Kang, Y.; Yin, H.; Berger, C. *IEEE Transactions on Intelligent Vehicles* **2019**, *4* (2), 171-185.
195. Balandat, M.; Karrer, B.; Jiang, D.; Daulton, S.; Letham, B.; Wilson, A. G.; Bakshy, E. *Advances in neural information processing systems* **2020**, *33*, 21524-21538.
196. Griffiths, R.-R.; Klarner, L.; Moss, H.; Ravuri, A.; Truong, S.; Du, Y.; Stanton, S.; Tom, G.; Rankovic, B.; Jamasb, A. *Advances in Neural Information Processing Systems* **2024**, *36*.
197. Häse, F.; Aldeghi, M.; Hickman, R. J.; Roch, L. M.; Aspuru-Guzik, A. *Applied Physics Reviews* **2021**, *8* (3).
198. Kandasamy, K.; Vysyaraju, K. R.; Neiswanger, W.; Paria, B.; Collins, C. R.; Schneider, J.; Poczos, B.; Xing, E. P. *Journal of Machine Learning Research* **2020**, *21* (81), 1-27.
199. *Uncertainty in Artificial Intelligence*, PMLR: 2020.
200. Nambiar, A. M.; Breen, C. P.; Hart, T.; Kulesza, T.; Jamison, T. F.; Jensen, K. F. *ACS Central Science* **2022**, *8* (6), 825-836.
201. Wang, G.; Ang, H. T.; Dubbaka, S. R.; O'Neill, P.; Wu, J. *Trends in Chemistry* **2023**.
202. Clayton, A. D. *Chemistry-Methods* **2023**, *3* (12), e202300021.

203. Dietz, T.; Klamroth, K.; Kraus, K.; Ruzika, S.; Schäfer, L. E.; Schulze, B.; Stiglmayr, M.; Wiecek, M. M. *European Journal of Operational Research* **2020**, *280* (2), 581-596.
204. Papoulias, S. A.; Grossmann, I. E. *Computers & chemical engineering* **1983**, *7* (6), 723-734.
205. 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. *Angewandte Chemie* **2023**, *135* (3), e202214511.
206. Kearney, A. M.; Collins, S. G.; Maguire, A. R. *Reaction Chemistry & Engineering* **2024**, *9* (5), 990-1013.
207. Nolan, L. J.; King, S. J.; Wharry, S.; Moody, T. S.; Smyth, M. *Current Opinion in Green and Sustainable Chemistry* **2024**, 100886.
208. Climent, M. J.; Corma, A.; Iborra, S. *Chemical Reviews* **2011**, *111* (2), 1072-1133.
209. Volk, A. A.; Epps, R. W.; Yonemoto, D. T.; Masters, B. S.; Castellano, F. N.; Reyes, K. G.; Abolhasani, M. *Nature Communications* **2023**, *14* (1), 1403.
210. Betinol, I. O.; Lai, J.; Thakur, S.; Reid, J. P. *Journal of the American Chemical Society* **2023**, *145* (23), 12870-12883.
211. Kim, H.; Gerosa, G.; Aronow, J.; Kasaplar, P.; Ouyang, J.; Lingnau, J. B.; Guerry, P.; Farès, C.; List, B. *Nature Communications* **2019**, *10* (1), 770.
212. 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* (6646), 706-712.
213. 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* (6661), 965-972.
214. Wagen, C. C.; McMinn, S. E.; Kwan, E. E.; Jacobsen, E. N. *Nature* **2022**, *610* (7933), 680-686.
215. Strassfeld, D. A.; Algera, R. F.; Wickens, Z. K.; Jacobsen, E. N. *Journal of the American Chemical Society* **2021**, *143* (25), 9585-9594.
216. Strambeanu, I. I.; Diccianni, J. B. High-Throughput Experimentation in Discovery Chemistry: A Perspective on HTE Uses and Laboratory Setup. In *The Power of High-Throughput Experimentation: General Topics and Enabling Technologies for Synthesis and Catalysis (Volume 1)*; ACS Publications: 2022; pp 11-22.
217. Buglioni, L.; Raymenants, F.; Slattery, A.; Zondag, S. D.; Noël, T. *Chemical Reviews* **2021**, *122* (2), 2752-2906.
218. Taylor, C. J.; Felton, K. C.; Wigh, D.; Jeraal, M. I.; Grainger, R.; Chessari, G.; Johnson, C. N.; Lapkin, A. A. *ACS Central Science* **2023**, *9* (5), 957-968.