Trends in Chemistry

CellPress

Feature Review

When machine learning meets molecular synthesis

João C.A. Oliveira [1,2], Johanna Frey [1,2], Shuo-Qing Zhang,[3] Li-Cheng Xu,[3] Xin Li,[3] Shu-Wen Li,[3] Xin Hong [3,4,5,*] and Lutz Ackermann [1,2,*]

The recent synergy of machine learning (ML) with molecular synthesis has emerged as an increasingly powerful platform in organic synthesis and catalysis. This merger has set the stage for key advances in *inter alia* reaction optimization and discovery as well as in synthesis planning. The creation of predictive ML models relies on chemical databases, molecular descriptors, and the choice of the ML algorithms. Chemical databases provide a crucial support of chemical knowledge contributing to the development of an accurate and generalizable ML model. Molecular descriptors translate the chemical structure into digital language, so that substrates or catalysts in molecular synthesis and catalysis are represented in a numerical fashion. ML algorithms achieve an effective mapping between the molecular descriptors and the target properties, enabling an efficient prediction based on readily available or calculated descriptors. Herein, we highlight the key concepts and approaches in ML and their major potential towards molecular synthesis with emphasis in catalysis, pointing out additionally the most successful cases in the field.

Highlights

Key advances were made possible in catalysis thanks to the synergy between machine learning (ML) and organic synthesis.

The chemical system is converted into a computer-readable language through the use of molecular representations and descriptors.

Databases are crucial for the development of ML predictive models in molecular synthesis, but they still largely lack open-access, scale, and structured chemical databases, particularly with respect to unsuccessful experimental data.

The synergy between experimentation and ML allows a significant acceleration of reaction optimization.

The rapid exploration of the chemical space through deep learning makes it an efficient strategy for the assembly of viable molecules in molecular synthesis.

Deep neural networks can assist in the choice of retrosynthesis planning, minimizing unreasonable synthetic operations.

Progress has been made on the human-free autonomous development of highly efficient chemical reactions in molecular synthesis.

Relevant aspects of ML in molecular synthesis and catalysis

The development of new synthesis strategies by catalysis is of crucial relevance to the development of novel compounds, with noteworthy applications in medicinal chemistry, materials science, and the pharmaceutical and agrochemical industries, among others [1–11].

One of the main drawbacks in the development of novel reactivities in catalysis is the need to experimentally test a vast number of reaction conditions. These include the laborious and time-consuming exploration of *inter alia* catalysts, ligands, and solvents. This constitutes a bottleneck in the development of novel catalytic synthesis approaches and a major investment in time and resources.

A recent strategy to overcome these drawbacks has gained considerable attention; namely, data-driven approaches for the performance prediction and efficient evaluation of chemical reactions. One widely used approach to establish the quantitative structure–activity relationship in organic synthesis is Sigman's multivariant linear regression method [12–17]. Through parametrization of transformation-related molecular descriptors, predictive regression models can be developed to evaluate reactivity or selectivity in molecular synthesis and catalysis, enabling reaction design in organic synthesis. ML has been used in various chemistry areas [18–22] such as drug discovery [23–25], the prediction of chemical reactions [26–31], and computer-aided synthesis [32,33]. ML approaches rely on a large set of datapoints (dataset). These datasets may comprise thousands of data points, which can be obtained, for instance, through high-throughput experimentation (HTE) campaigns or quantum chemical calculations [34–39]. Dreher, Doyle, and coworkers have

[1]Institut für Organische und Biomolekulare Chemie, Georg-August-Universität Göttingen, Tammannstraße 2, 37077 Göttingen, Germany
[2]Wöhler Research Institute for Sustainable Chemistry (WISCh), Georg-August-Universität Göttingen, Tammannstraße 2, 37077 Göttingen, Germany

illustrated the relevance of the merger of HTE and *in silico* strategies in the generation of ML predictive models for the prediction of chemical reactivity [28]. Moreover, Denmark and colleagues exploited datasets comprising 1075 transformations for the accurate prediction of reaction outcomes; namely, for enantioselectivities from out-of-sample data through the optimization of chiral phosphoric acid catalysts [40]. Corminboeuf and colleagues demonstrated the importance of ML in combination with volcano plots in determining the most suitable catalysts in a universe of 18 000 for Suzuki–Miyaura coupling [41]. Reaction optimization studies of the same reaction with the assistance of deep learning has been successfully accomplished by Zheng and coworkers [42]. Deep learning was employed in the optimization of several reaction parameters simultaneously. Such an approach presents an acceleration over other optimization procedures, such as the most commonly used 'one parameter at a time' or design of experiment (DOE) approaches. This highlights the potential of ML in the context of molecular synthesis and catalysis. Doyle and coworkers have likewise successfully applied Bayesian optimization algorithms for reaction optimizations [26]. ML has hence surfaced as an attractive tool in the optimization of chemical transformations in molecular synthesis and catalysis. Besides prediction of chemical reactivity and the exploration of conditions for reaction optimization, ML has been successfully applied to synthesis planning. The major advance in the history of computer-aided molecular synthesis can be explored, for example, in the Synthia (or Chematica) software [43], where multiple synthetic pathways can be computationally accessed taking into consideration a huge set of mechanistic rules [43]. Besides the relevant *in silico* progress of ML in molecular synthesis and catalysis (*vide supra*), the development of ML-assisted autonomous setups for the exploration of chemical reactions, as catalytic reactions, can be extremely advantageous. These would strongly contribute to significant advances in the field of molecular synthesis and in the development of future novel catalytic reactions. For example, Jensen and coworkers have significantly contributed to progress in the development of an automated robotic setup for molecular synthesis [44].

Besides the number of data points and the choice of a ML algorithm to generate a reliable ML predictive model, the selection of molecular descriptors for a given chemical space is of key importance. In the context of molecular synthesis, the descriptors describe each component of the chemical reaction. These can comprise physical chemical properties, Morgan fingerprints, the simplified molecular input line entry specification (SMILES), or other variants (*vide infra*). The choice of the most relevant descriptors is often not easy, but chemical intuition can serve as a guiding light. Recognition of the most pertinent descriptors is paramount to decrease the complexity of the system, as well as to generate a reliable predictive model.

Herein, we highlight the most successful cases of the use of ML in the field of molecular synthesis with emphasis on catalysis. We demonstrate the potential impacts and benefits of ML in assisting in the rapid development of catalysis.

Guidelines

Molecular representations and descriptors

To develop predictive ML models, the components of the catalytic system of interest and their inherent properties need to be translated into a computer-readable mode. These may include a molecular representation of the involved components of the catalytic reactions and a portrayal of their properties as descriptors (Figure 1).

String representations

For string representations, the molecular structure is simplified by a linear sequenced-based series of characters (e.g., number and letters) taking into consideration the adjacent atoms.

[3]Center of Chemistry for Frontier Technologies, Department of Chemistry, State Key Laboratory of Clean Energy Utilization, Zhejiang University, Hangzhou 310027, PR China
[4]Beijing National Laboratory for Molecular Sciences, Zhongguancun North First Street No. 2, Beijing 100190, PR China
[5]Key Laboratory of Precise Synthesis of Functional Molecules of Zhejiang Province, School of Science, Westlake University, 18 Shilongshan Road, Hangzhou 310024, Zhejiang Province, PR China

*Correspondence:
hxchem@zju.edu.cn (X. Hong) and Lutz.Ackermann@chemie.uni-goettingen.de (L. Ackermann).

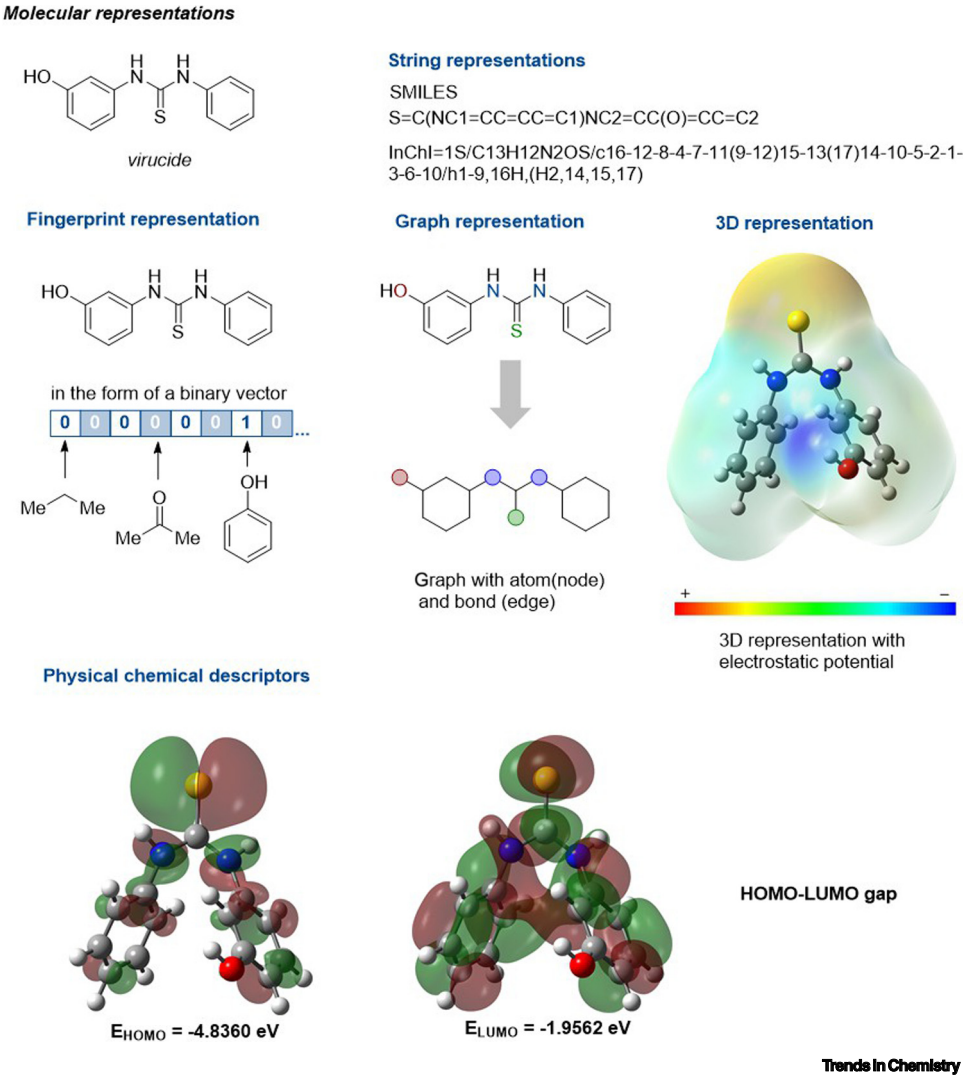


Figure 1. Examples of molecular representations used in the context of machine learning (ML).

The most commonly used string representations are the SMILES and the IUPAC International Chemical Identifier (InChI) [45,46]. However, SMILES has limitations, notably when dealing with certain stereochemistry issues (e.g., axial and planar chiralities) as well as when representing molecules with complicated topological structures (e.g., C60, carborane). Standard SMILES performs rather poorly in describing organometallic compounds given the diversity and complexity of the bonds present in such complexes compared with organic molecules. The syntactics used in the generation of SMILES strings can lead to the generation of invalid strings that do not correspond to a valid molecule. DeepSMILES provides a viable replacement for SMILES. Here, the number of balanced closed parentheses designates the branch length. Additionally, only one symbol is used at the ring closing position, which is indicative of the ring size (https://github.com/baoilleach/deepsmiles). As a robust alternative to SMILES, Aspuru-Guzik developed SELF-referencing embedded strings (SELFIES), where any SELFIES matches to a valid molecule [47].

Fingerprint representations

Fingerprint representations are one of the most common representations for chemical structures. These descriptors encode the molecular structure into a binary vector or a count vector. The more frequently used is a bit vector solely comprising 0 and 1, where 0 and 1 correspond to the absence or presence of a certain predefined structural fragment, such as a functional group or arenes. These types of fingerprints are identified as structural keys, and one can take as examples the Molecular ACCess System (MACCS) keys and PubChem fingerprints [48,49]. The latter comprise an 881-bit-long fingerprint. Alternatives to structural keys are topological and circular fingerprints, which do not require a predefined fragment library. For topological fragments, generation follows a linear path up to a certain number of bonds in the molecule. One of the most linear path-based fingerprints is the Daylight fingerprint (https://www.daylight.com/). Circular fragments are generated by exploring the circular environment around a certain atom to a designated radius. Examples are extended-connectivity fingerprints (ECFPs) and functional-class fingerprints (FCFPs) [50,51]. In ECFPs, Morgan's algorithm is used, and the fingerprint is generated taking into consideration the neighboring atoms inside an atom-centered circular limit with a defined maximum radius *n* (e.g., ECFP4). FCFPs, contrary to ECFPs, encode atom roles and not atoms (e.g., aromatic, hydrogen-bond acceptor). Fingerprint representations are user-friendly since they have very low generation costs and require limited storage.

Graph representations

Given a graph to represent molecules where atoms and bonds occupy the nodes and edges of the graph, this type of graph model can be processed using a graph convolutional neural network (GCN). In a GCN, each atom is initially represented in the same dimension and treated equally, and the bond is handled similarly. Subsequently, the atom representations are passed into the convolutional layer; each atom will obtain its own attribute updated from the adjacent atoms through the convolution function, whereas the weight of each adjacent atom in the convolution function is calculated from the attributes of the corresponding bond [52,53]. A GCN is a parameterized encoder, which can be trained for the prediction of molecular properties as well as for the generation of molecular fingerprints.

3D representations

In 3D representations, the molecular system is represented through a consideration of the distribution of the atoms in space. Such representations can be obtained experimentally through crystallographic data or by means of computation. 3D representations are therefore numerical representations of molecular properties, such as steric and electronic properties, which are derived from a 3D molecular structure. Denmark and coworkers invented and created the average steric occupancy (ASO) descriptor for the prediction of highly selective catalysts [40]. Other 3D descriptors that are widely used in catalysis and molecular synthesis are, for example, charges calculated through natural bond orbital (NBO) calculations [54], Sterimol parameters, buried volumes, and, in cases where phosphines are present in the catalysis, the Tolman angle [55]. Atom-centered symmetry functions have also been proven to be advantageous [56].

Physical chemical descriptors

Physical chemical descriptors are directly derived from the physical chemical properties of the molecular structure. They represent several levels of complexity, ranging from simply the number of atoms or molecular weight to more complex information, including highest occupied molecular orbital (HOMO) or lowest unoccupied molecular orbital (LUMO) energies, partial charges, NMR resonances, and buried volumes [57]. Experimental determination is applicable for a limited set of physical chemical descriptors (e.g., NMR resonances), while quantum mechanical calculation is often a more general approach to access these descriptors.

The choice of the most significant descriptors is of high importance as these are crucial for the generation of a reliable ML predictive model. In the context of molecular synthesis and catalysis, SMILES is the most widely used string representation, whereas between the 2D descriptors are Morgan fingerprints. 3D descriptors are especially relevant when dealing with conformers and bulky molecules like phosphines, for the latter of which Sterimol parameters, Tolman cone angles, and physical chemical descriptors like buried volumes are widely used. Atomic charges, bond dissociation enthalpies, and HOMO and LUMO energies are also the most commonly used descriptors. These tend to be calculated through DFT, however, using different levels of theory depending on the molecular system.

ML algorithm

ML algorithms can be separated into three main categories: reinforced, unsupervised, and supervised learning [58]. The reinforced learning model is trained based on reward feedback, rewarding positive stimuli while punishing undesired ones. In this way, the model is generated by the response of the learning agent to the environment. This learning method has been employed in the determination of catalytic reactions as well as in the optimization of molecular structures [59,60]. For unsupervised ML, the model performs the analysis and clustering of an unlabeled dataset. This has been employed in the generation of vector representations of molecular structures [61]. Supervised ML makes use of a labeled dataset to train ML algorithms to predict an output (e.g., yield) in an input–output duality. This strategy was applied to the prediction of chemical reaction performance [28]. In this review, the focus is directed to supervised ML approaches.

Supervised ML

Supervised ML can be subdivided into classification and regression algorithms. In classification algorithms, a provided dataset is separated into classes by categorization, whereas in regression algorithms a quantitative prediction of the relation between features and a continuous target variable is modeled.

In supervised ML algorithms, the full labeled dataset is separated into a training set, a test set, and a validation set (Figure 2A). The training and test sets are solely involved in the generation of the predictive ML model through a supervised ML algorithm. The validation set is employed in the performance evaluation of the generated predictive model. For regression models, typical performance assessments include the Pearson correlation coefficient, R^2, the mean absolute error (MAE), and the root-mean-square error (RMSE). Accuracy, precision, recall, and f1 score are typical evaluation indexes for classification models.

The model is validated given the quality of the data available using nested cross-validation. One practical cross-validation schemes is *k*-fold cross-validation. Here, the data points of the full dataset minus the validation set are divided into *k* subsets with equal dimensions. One of the subsets will be used as a test set, whereas the remaining *k* - 1 subsets will integrate the training set. This procedure will be iterated until all of the subsets were once an internal test set (i.e., nested). An extreme case is leave-one-out cross-validation, where one data point stipulated as an internal test set.

A possible workflow for the application of supervised ML to the prediction of a reaction outcome is depicted in Figure 2B. In the example, the goal is the prediction of the enantioselectivity – namely, the enantiomeric excess (*ee*) – of a selected new reaction. In this case, the *ee* is identified as the target. The first step involves the collection of experimental data, such as the identification of each reaction component, and the *ee* for very similar reactions. The second step includes the encoding of each reaction component into a computer-friendly input through the use of

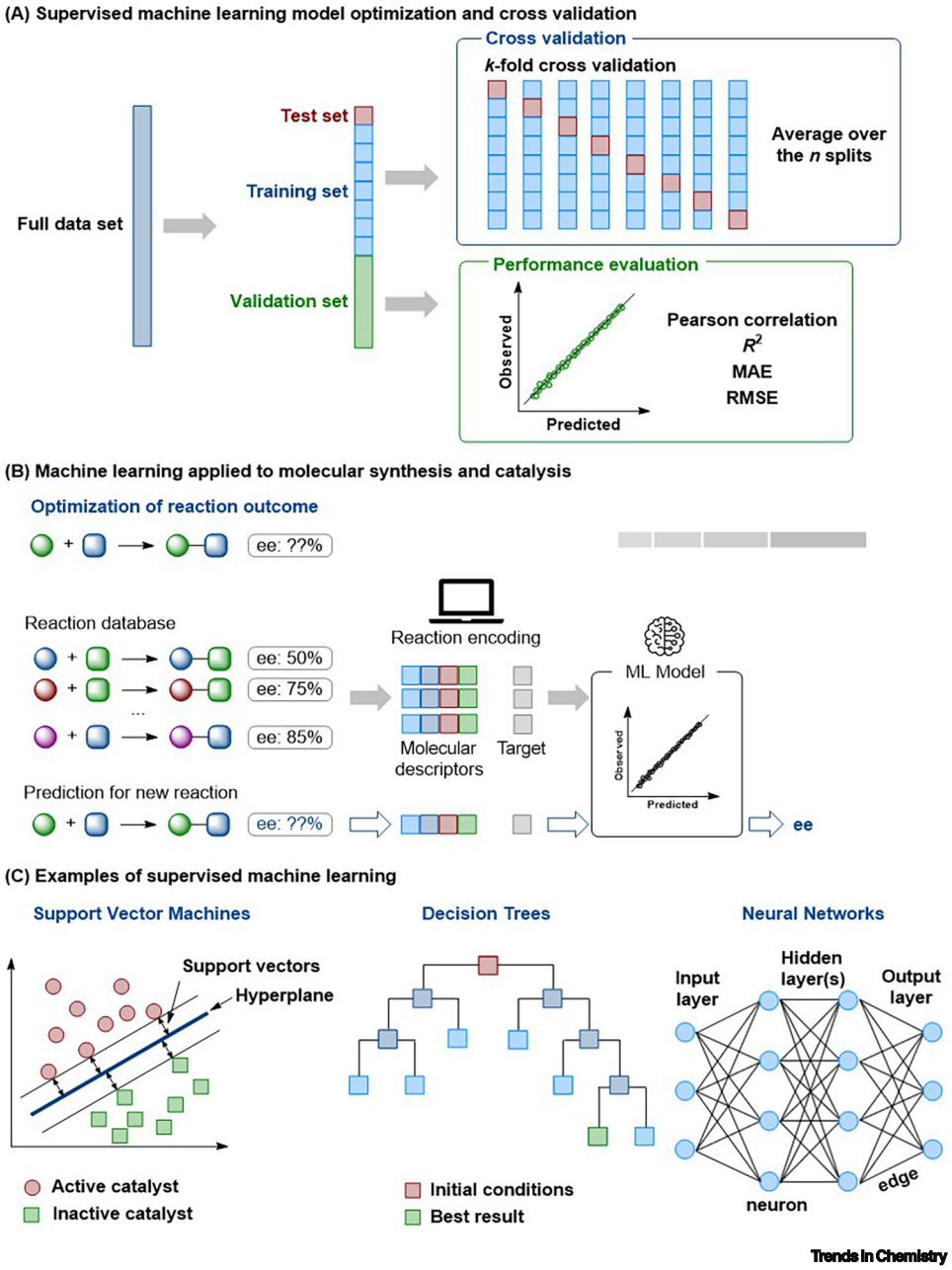


Figure 2. Supervised machine learning (ML) in molecular synthesis and catalysis. (A) Optimization and validation of supervised ML models. (B) General workflow for supervised ML applied to molecular synthesis. (C) Examples of supervised ML algorithms.

descriptors (*vide supra*). Afterwards this information is fed into a chosen ML algorithm leading to the generation of a ML model for the prediction of *ee*. It is crucial to try several ML algorithms to identify the most suitable for the desired case. Once the ML predictive model is generated, one proceeds to the encoding of each component of the new reaction, which is fed into the model, and the respective *ee* is predicted. Such a general approach can be applied to several aspects in molecular synthesis (*vide infra*).

(Multivariate) Linear regression

One of the simplest supervised ML algorithms constitutes linear regression or multivariate linear regression (MLR). This model relies on establishing a linear relationship between the target and the weighted contribution of the descriptors. The latter are optimized during the generation of the predictive model [17]. Such algorithms have set the stage for the prediction of enantioselectivity in asymmetric catalytic reactions [62,63].

k-nearest neighbors and support vector machines (SVMs)

In k-nearest neighbors, the target prediction of a specific data point is determined based on the k-nearest data points (neighbors). SVM algorithms attempt to separate the training set through a linear separation known as a hyperplane (Figure 2C). This is achieved by maximizing the distance between the data point and the hyperplane (i.e., the support vector). SVMs have been shown to perform well in catalyst selectivity prediction for asymmetric catalysis [40].

Decision trees, random forests, gradient tree boosting

In decision-tree algorithms, the target value is predicted based on decision rules (Figure 2C). Each node where the query occurs is known as a decision node. The tree grows until the best result by full separation is reached (i.e., leaves). A variation of the decision tree is the random forest. This combines a multitude of several decision trees on various subsets, where the accuracy of the target is given by the average. Random-forest algorithms have been largely employed in the context of molecular synthesis.

Boosting algorithms are used to improve the predictive power by employing a sequence of less-positive/weak learners, where each model is stronger than the previous one. Examples of boosting algorithms are the gradient boosting algorithm and the adaptive boosting (AdaBoost) algorithm. In boosting algorithms, each generated model is an improvement over the previous one by making use of a gradient descent method, which minimizes the loss function. AdaBoost uses an ensemble method, where weak and strong learners are identified through the attribution of weights. Both algorithms can be used in classification as well as in regression problems.

Boosting algorithms are commonly used in combination with tree algorithms. As an example, gradient tree boosting constructs a new tree based on the less-positive results from an initial decision tree, repeating such a process until the optimization of the predicted value is sufficient.

Artificial neural networks (ANNs)

ANNs are now widely used in the ML field for a multitude of tasks. Their nomenclature is derived from the ANN intrinsic algorithm properties, which resembles the neuron pattern from the human brain. Therefore, ANNs comprise a network of artificial neurons (nodes) organized into layers that are weightily interconnected (Figure 2C). In an ANN, an input layer is connected to a hidden layer of neurons, which are then connected to an output layer. The connection, transmission of the output from one neuron to another neuron in the next layer, occurs only if the output is above a certain threshold; otherwise, no information is transferred. The simplest architecture comprises solely one hidden layer; however, this can differ by case. When more than one hidden layer is in play, one refers to it as deep learning [22]. Compared with more traditional ML models, they have the advantages of nonlinear learning models; however, they are sensitive to feature scaling. Deep learning has been applied to numerous tasks in molecular synthesis, such as in the prediction of the outcome of chemical reactions and the planning of molecular synthesis [32,64].

Chemical data for ML

Chemical data are the source of innovation for data-driven research in chemistry. The establishment of curated databases paves the way for ML applications in chemistry, enabling the desired accurate and efficient predictions of molecular properties or reaction performance [18,65–67]. Organic chemistry, or more generally even the entire chemistry field, despite the long research history, still lacks open-access, large-scale, and standardized chemical databases, particularly of non-optimal results. Most of the literature data are stored in a 'cold' form, which requires extensive efforts for additional data collection and cleaning to meet ML purposes [68]. Remarkable efforts have been dedicated to address this data issue by Reymond [37,38], Lilienfeld [36], Mayr,[i] Cheng (http://ibond.nankai.edu.cn/), Aspuru-Guzik [69], Coley [60], Hong [70], and others [71,72], which have led to a number of widely used experimental and computational databases that are available to the community. A few representative ones in organic chemistry are highlighted below.

Experimental database

Mayr's systematic studies on nucleophilicity and electrophilicity provided a remarkable data source for the kinetic properties of molecules in nucleophilic and electrophilic reactions (Figure 3) [73–75]. These efforts led to the famous Mayr's equation, which allows quantified calculation of second-order rate constants in a simple and elegant fashion [76–78]. Mayr's reactivity parameters include three empirical components: the solvent-independent term electrophilicity parameter *E*; the solvent-dependent term nucleophilicity parameter *N*; and the solvent-dependent term susceptibility s_N. The establishment of Mayr's reactivity parameters were based on experimental measurements of rate constants; for example, by conventional UV/Vis, stopped-flow, or laser flash photolysis technologies [79]. To overcome the issue of measuring the rate constants for nucleophiles with a wide reactivity range, Mayr designed a series of diarylcarbenium ions as reference carbon electrophiles [80]. *E* of bis(4-methoxyphenyl)carbenium ion is defined as 0.00 and s_N of allyltrimethylsilane is defined as 1.00 (Figure 3) [81].

To date, Mayr's database of reactivity parameters has covered a large chemical space in organic chemistry (Figure 3)[i]. Hence, the parameters of 345 electrophiles have been measured and recorded, which are mainly carbocations (132), electron-deficient olefins (114), or other carbon electrophiles (78), as well as a limited number of sulfur (nine), fluorine (five), nitrogen (four), and chlorine (three) electrophiles; 1250 nucleophiles have been recorded, including 530 carbon, 311 nitrogen, 191 hydrogen, 138 oxygen, 33 sulfur and selenium, 29 halide anion, and 18 phosphorus nucleophiles. The distributions of the measured reactivity parameters are close to a normal distribution (Figure 3). In addition, the database has classified the confidence of the recorded data depending on the measurement details. Based on this database, one can easily predict the second-order rate constants for over 430 0000 reaction combinations.

In addition to the reactivity data for organic molecules, the information on organic transformation from reactant to product under certain conditions is also of key relevance [68,82]. This contains knowledge on chemical transformations, which can support the artificial intelligence (AI) applications for reaction outcome prediction and automatic total synthesis planning [29,32]. Although the literature contains a large number of reported synthetic transformations, these data are not available to the community as an open-source database. Interestingly, the major data source for such information is currently from US patents[ii]. Through text-mining techniques, the reaction schemes were extracted as the USPTO database by NextMove [83]. Further labeling of the USPTO database can improve its usefulness; using a fingerprint-based approach to assign reaction roles, Schneider and colleagues [84] selected a subset of USPTO data and applied their labeling technique. This dataset is now known as USPTO-50k, whose data distribution is illustrated in Figure 4. Carbon functionalization is the predominant type due to the nature of patent

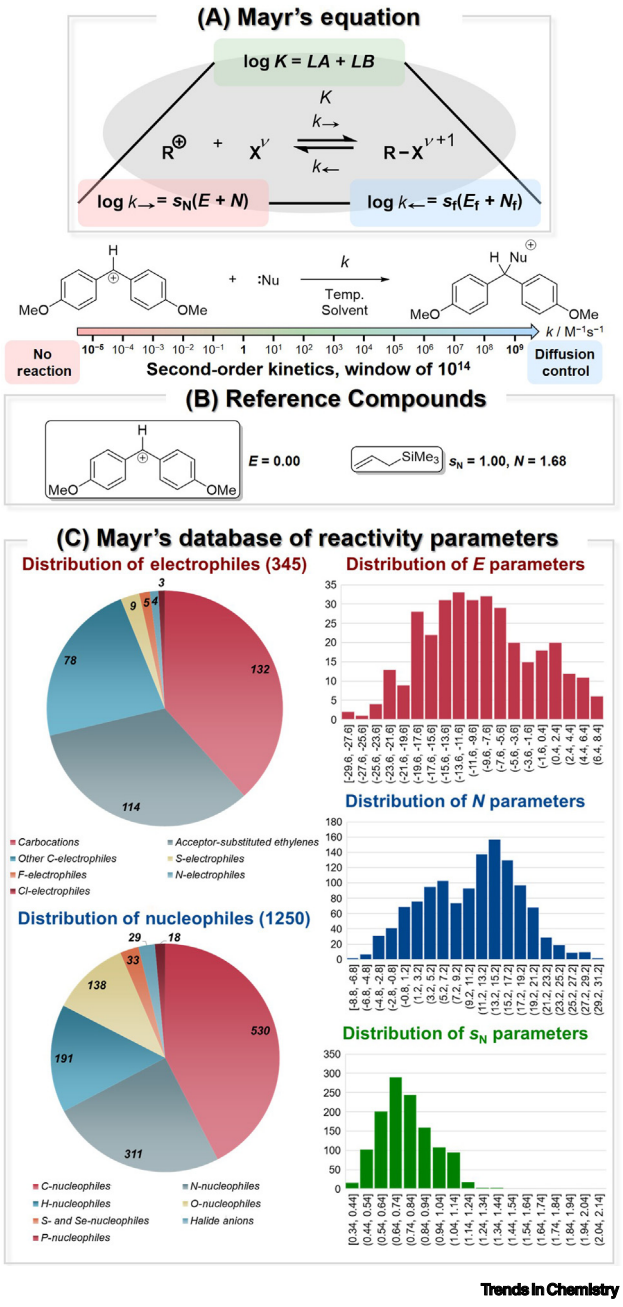


Figure 3. Herbert Mayr's database of reactivity parameters[I]. (A) Mayr's equation. (B) Reference compounds in Mayr's equation. (C) Key statistics of Mayr's database of reactivity parameters.

data. USPTO and its derivatives, like USPTO-50k, have now been widely applied in model training to predict organic reaction outcomes and for synthetic planning [82].

Computationally established database

In addition to the experimental measurement of chemical data, the rapid advancement of theoretical and computational tools allows chemists to explore organic chemistry through computational simulations [85]. Under an appropriate level of theory, modern computational approaches

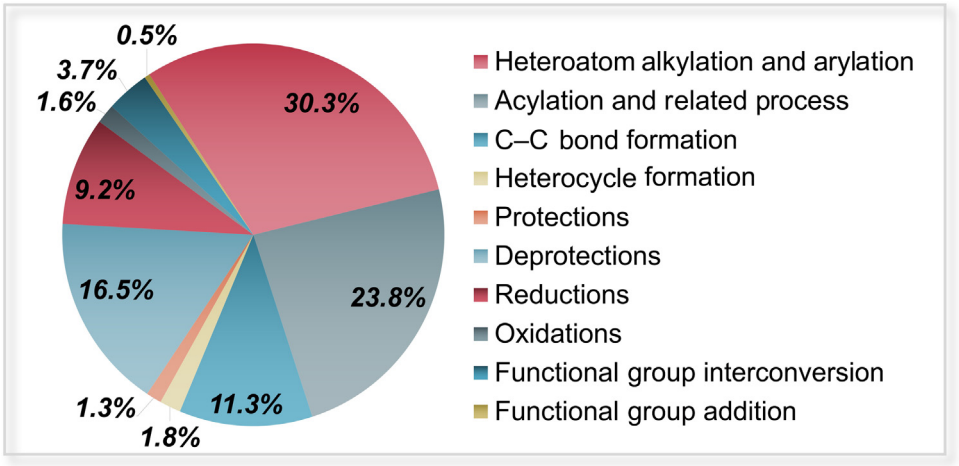
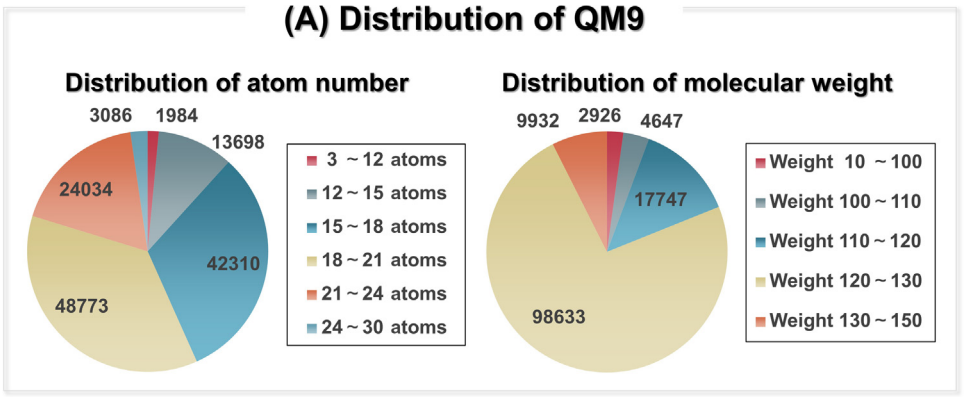


Figure 4. Coverage and data distribution of USPTO-50k database.

enabled the rationalization and even prediction of the reactivities and selectivities of organic transformations [86]. Therefore, a computational approach can serve as an increasingly viable alternative strategy to access structure–activity relationship data in organic chemistry [87]. Given the intrinsic features of computation, a computed chemical database generally has a larger scale and is typically more complete than experimental databases. Thus, a number of computational databases on molecular properties or reaction performance in organic chemistry have emerged over the recent few years [36–38,88], which have provided the needed data support for scenarios in which experimental measurements are challenging.

Under the B3LYP/6-31G(2df,p) level of theory, Lilienfeld and coworkers have systematically computed the 133 885 stable small organic molecules containing up to nine heavy atoms (C, O, N, and F), and this well-known dataset is called 'QM9' [36]. This remarkable group of molecules is a subset of the GDB-17 chemical universe [37]. The distributions of atom number and molecular weight are shown in Figure 5A. Through DFT calculations, these molecules are optimized to the stationary geometry, and the harmonic frequencies are computed. These calculations allow the recording of a series of important molecular properties that chemists use daily for the evaluation of reactivity and selectivity, including FMO energies and dipole moment (Figure 5B). The QM9 database has now been extensively used in ML training and prediction of molecular properties [88–94]. These surrogate property models have provided accurate and efficient access to synthetically relevant molecular properties, which have also been applied in synthetic performance predictions [41,95].

Modern quantum-chemical approaches can provide reasonable barrier calculations, which offer an alternative data source for molecular synthesis if based on solid mechanistic understandings [85]. Targeting the selectivity prediction of radical C–H functionalization of heteroarenes, Hong and coworkers showed that DFT calculations can serve as a useful data augmentation strategy, which enabled accurate ML modeling for selectivity prediction (Figure 6) [96]. The authors designed a representative chemical space that contains 201 distinctive arene substitution patterns and 13 radicals, involving 6114 barriers and 9438 barrier differences between competing positions. Such a large number of reactivity and regioselectivity data are challenging to obtain by experimental approaches, especially for this transformation. Based on the mechanistic understandings of the radical direct functionalization of heteroarenes [97], the authors computed the transition states of the rate-determining radical addition step in a semiautomatic fashion and



(B) Properties of QM9

Property	Description
A	Rotational constant
B	Rotational constant
C	Rotational constant
μ	Dipole moment
α	Isotropic polarizability
ϵ_{HOMO}	Energy of HOMO
ϵ_{LUMO}	Energy of LUMO
ϵ_{gap}	Gap ($\epsilon_{LUMO} - \epsilon_{HOMO}$)
$\langle R^2 \rangle$	Electronic spatial extent
zpve	Zero point vibrational energy
U_0	Internal energy at 0 K
U	Internal energy at 298.15 K
H	Enthalpy at 298.15 K
G	Free energy at 298.15 K
C_v	Heat capacity at 298.15 K

Trends in Chemistry

Figure 5. QM9 database. (A) Key statistical distribution of QM9. (B) Molecular properties recorded in QM9 [36].

created a virtual database (Figure 6). This database covers a wide range of reactivity and selectivity and the knowledge-based chemical-space design ensured that the steric and electronic properties of substituents are properly represented. Benefiting from this computational database, the authors trained a random-forest model using physical organic descriptors, which showed satisfying performance compared with experimental regioselectivities [98]. This work indicated that the combination of mechanistic knowledge and computational data augmentation can serve as a useful approach to support ML and HTE in organic synthesis.

Open reaction database

To address the data availability issue in molecular synthesis, Kearnes, Coley, and coworkers proposed an initiative called the Open Reaction Database (ORD) [68]. This database aims to

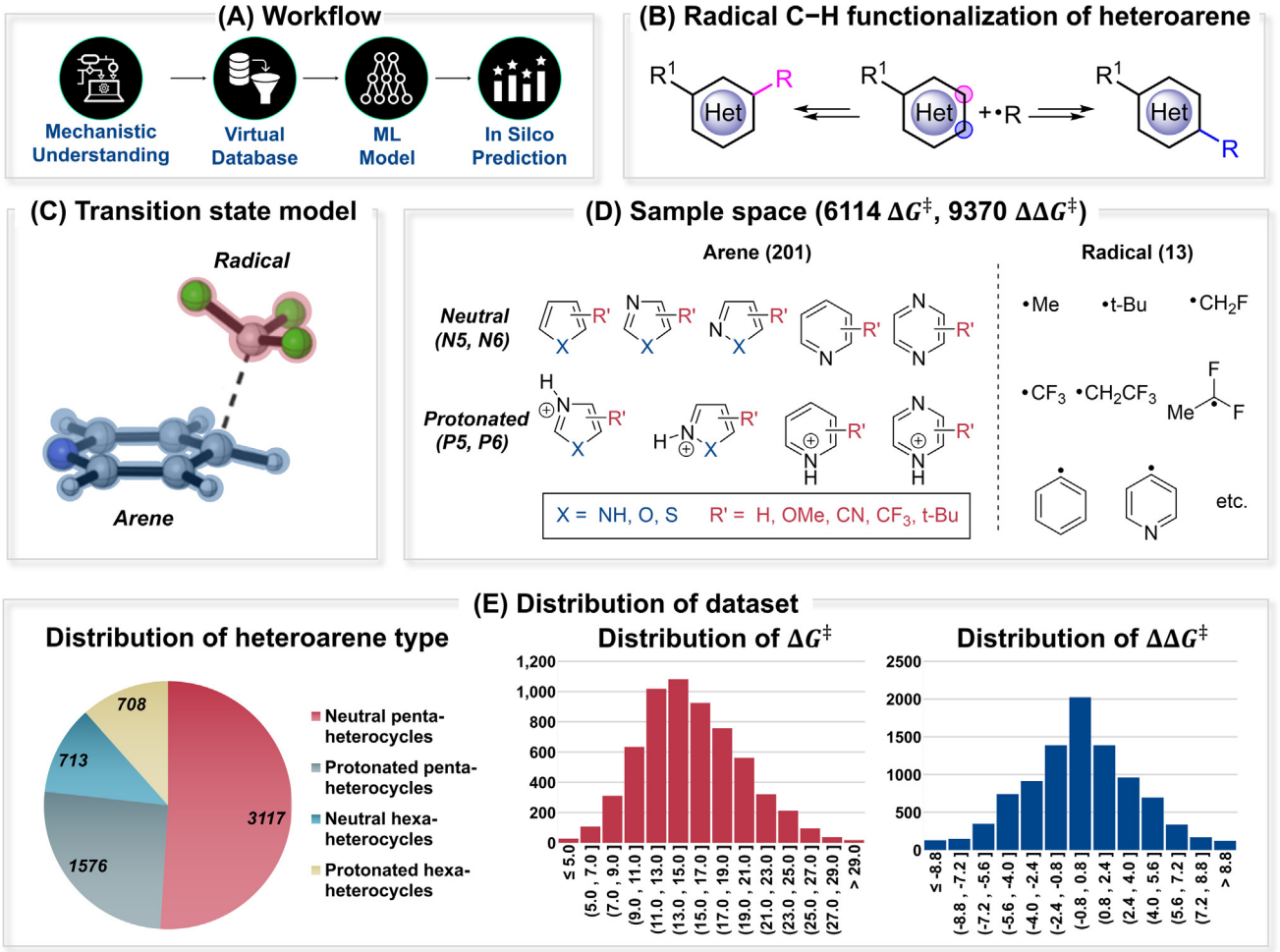


Figure 6. Computational approach and database for radical C–H functionalization of heteroarenes [96]. (A) Workflow for data generation. (B) Regioselectivity of radical C–H functionalization of heteroarene. (C) Transition-state model of selectivity-determining step. (D) Sample space of generated computational regioselectivity dataset, including 6114 Gibbs free energy barriers ($\Delta G^{\ddagger}$) and 9370 barrier differences of regioisomeric competitions ($\Delta\Delta G^{\ddagger}$). (E) Key statistical distribution of generated dataset based on heteroarene type, barrier height, and selectivity.

provide an open-access infrastructure for structuring and sharing organic reaction data. ORD allows a structured schema, submission mechanism, and search/retrieval tools, which significantly removes the barriers of data sharing and downstream ML applications. ORD has an amazing compatibility to support conventional and emerging synthetic approaches, including benchtop reactions to automated HTE and flow chemistry. ORD is fully open access, with data and code available on its GitHub project (https://github.com/open-reaction-database). ORD has already accumulated a remarkable set of data from various sources, as highlighted in Table 1. Its establishment will provide strong momentum to create a data community for molecular synthesis.

ML for molecular synthesis and catalytic reactions

Quantitative prediction of reaction outcomes

When optimizing the reaction conditions for a given reaction, it is often necessary to test a wide array of catalyst systems. However, only a small number of possible combinations can be

Table 1. Example datasets currently available in the ORD

Category	Description	Size	Refs
Literature extracted	Reactions extracted by text mining US published patents; imported from CML documents	1 771 017	[ii]
High throughput	Suzuki coupling reaction performance as a function of aryl halide, boronic acid, ligand, base, and solvent performed under pseudoflow conditions	5760	[34]
High throughput	C–N cross-coupling reaction yields varying aryl halide, additive, Pd catalyst, and base identities	4312	[28]
High throughput	Combinatorial nanochemistry screen of a complex aryl halide library using dual-metal photoredox C–N coupling	1728	[99]
High throughput	C–N cross-coupling reaction performance of 3-bromopyridine with various nucleophiles, varying precatalysts and bases	1536	[35]
Photochemistry	Ir-catalyzed debromination conversions as a function of photocatalyst ligands	1152	[100]
High throughput	Subset of 'chemistry informer' screen of copper-catalyzed Buchwald–Hartwig aminations	90	[101]
Single-step batch	Deoxyfluorination reaction screening as a function of substrate, base, and fluoride source	80	[27]
Single-step batch	Microwave synthesis of a small library using the Biginelli multicomponent condensation reaction	48	[102]
Flow chemistry	Sulfonamide library synthesis in flow	39	[103]
Electrochemistry	Electroreductive coupling of alkenyl and benzyl halides via nickel catalysis	27	[104]
Photochemistry	Substrate scope tables regarding coupling of α-carboxyl sp^3 carbons with aryl halides	24	[105]
Kinetic profiling	Online monitoring of a Suzuki coupling reaction by HPLC	7	[106]
Enzymatic	Multistep biocatalytic cascade for the manufacture of islatravir	3	[107]
Multistep	Copper-catalyzed enantioselective hydroamination of alkenes	3	[108]

experimentally tested compared with the entire chemical space. In this regard, ML can play an important role in developing the best possible catalytic system by limiting the number of experiments to be performed. Various methods have been used.

The group of Corminboeuf used ML in combination with Nørskov's famous concept of the volcano plot, which enabled the identification of attractive candidates from a thermodynamic point of view (Figure 7A) [41,109]. The number of catalytic systems that can be tested with this method is considerably greater than those found in more traditional experimental approaches. They were able to determine the best catalysts for a Suzuki–Miyaura coupling among 18 062 candidates, comprising six different metals, first from a thermodynamic point of view and then by adding a price consideration. Two years later, they estimate the activity of over 143 000 prospective nickel catalysts for the reductive cleavage of the 2-methoxynaphthalene C–O bond with trimethylsilane [109]. The choice of substituents on the phosphine was extremely important, since it allowed the stabilization of intermediates of the catalytic cycle through noncovalent intramolecular interactions and/or destabilization of the metal–ligand interactions.

In a distinct approach, Denmark and colleagues used a retrospective analysis to guide the optimization of BINOL-phosphoric acid to catalyze the enantioselective formation of *N*,*S*-acetals and enantioselective alkylation reactions with cinchona alkaloid-derived phase-transfer catalysts (Figure 7B) [40,110]. They developed a unique workflow that is applicable to any reaction and

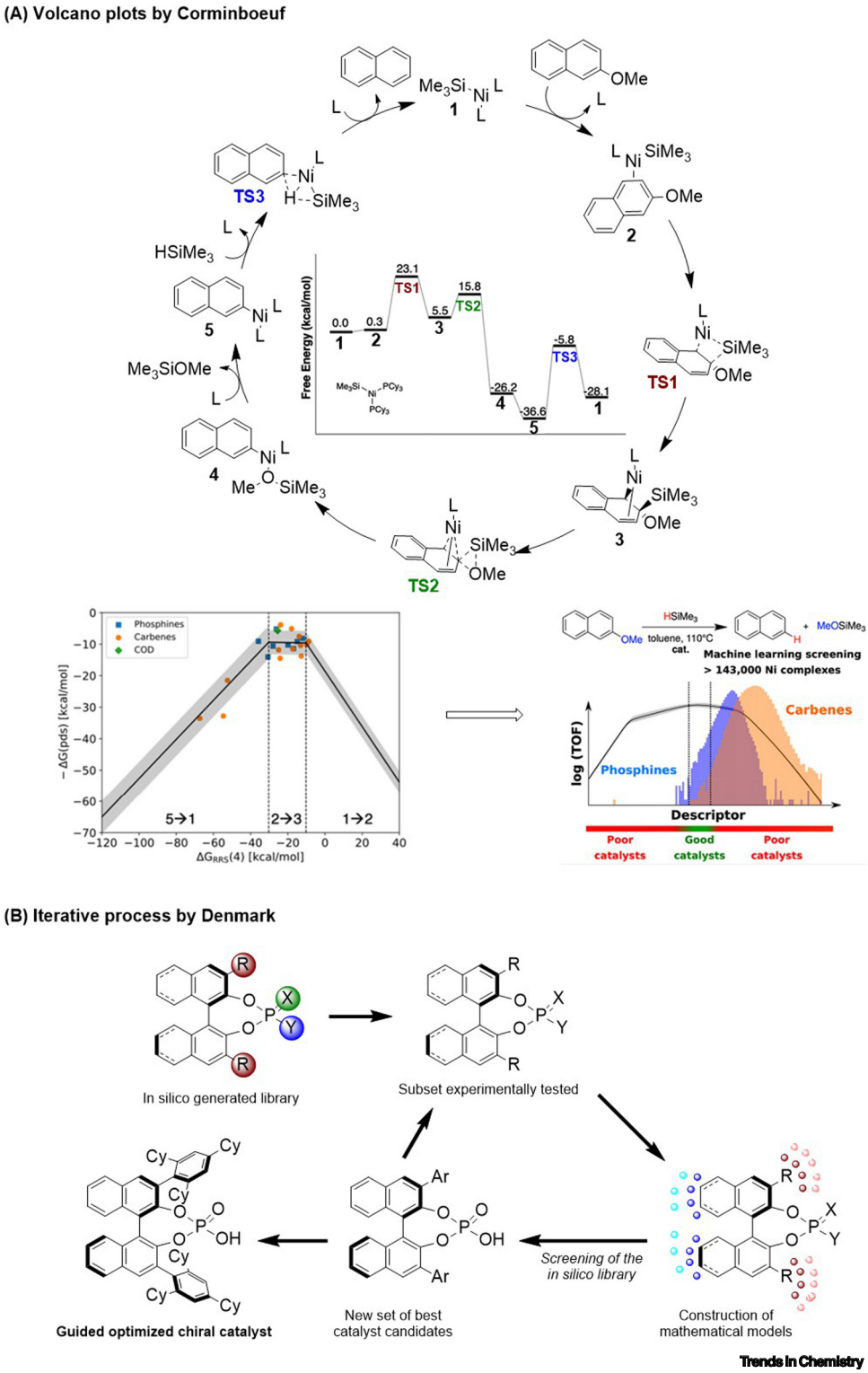


Figure 7. Machine learning (ML) applied to the prediction of reaction outcomes. (A) Identification of attractive catalyst from a thermodynamic point of view though volcano plots. (B) Retrospective analysis for the formation of chiral phosphoric acids [109,110].

mechanism called the universal training set (UTS). The UTS strategy selects the most representative molecules from the candidate chemical space based on the Kennard–Stone algorithm. They showed that their conformer-dependent representation of the catalyst outperforms the single-conformer analog and were able to perform reliable predictions for catalyst and substrate structures unknown to the model, including outside the selectivity range of the training set points. First, a large library of catalysts is constructed *in silico* and then descriptors are calculated for them. A representative subset is algorithmically selected, which can be synthesized and evaluated in any reaction of interest. Using these results, a mathematical model is constructed to match the calculated descriptors and the experimental outcome. With this model, the *in silico* library is virtually screened, and the best catalyst candidates are identified. This process is finally performed interactively, with each subsequent iteration being added to the training data until an ideal catalyst is identified, ML thus assisting in achieving it much more quickly. In addition, unsupervised learning was used to extend the database and improve the predictions.

The same strategy of iterative supervised analysis for ligand design was also employed by Reiner and Tonks to achieve a regioselective synthesis of pyrrole catalyzed by titanium [111].

Doyle and coworkers have successfully employed ML to assess the performance of C–N cross-coupling reactions based on data obtained by HTE [28]. The authors explored the palladium-catalyzed Buchwald–Hartwig cross-coupling of aryl halides with 4-methylaniline. For the latter, DFT-calculated molecular, atomic, and vibrational descriptors were considered for the generation of the ML model, with the chemical yields as the targets. Several ML algorithms were assessed by linear regression, k-nearest neighbors, SVMs, Bayes GLM, neural networks, and random forest. The authors demonstrated that random-forest algorithms exhibited a greater predictive performance while linear regression models fell short. Furthermore, the first was effectively used in the prediction of out-of-sample data. The same authors have also demonstrated the predictive power of ML for the identification of reactivity cliffs in 11 nickel- and palladium-catalyzed cross-coupling datasets with monodentate phosphine ligands (Figure 8) [55]. Descriptors commonly used to describe the electronic and steric properties of monodentate phosphine ligands fall short in justifying different reactivities for similar phosphine ligands. Thus, this gave rise to ligand reactivity discontinuities. The use of the Tollman cone angle was found not as appropriate to give a distinct reactivity cut-off. However, the use of the smallest percentage of buried volume [$\%V_{bur}$(min)] as a descriptor was shown to be propitious to divide the data set into active and inactive regions at a comparable threshold value. An algorithm for the threshold analysis was developed by making use of a decision-tree algorithm. The use of $\%V_{bur}$(min) was shown to correctly predict the distinction between monoligated metal (L_1M) and bisligated metal (L_2M), which will greatly impact the reactivity.

More recently, Schoenebeck and coworkers reported an unsupervised ML workflow for the speciation of palladium(I) complexes, for which mechanistic foundations remain largely absent [112]. Their approach made use of only five experimental data points in combination with a 'ligand knowledge base' database for phosphine ligands (348 ligands) developed by Fey, Harvey, Orpen, and coworkers [113–115]. Additionally the database was complemented with problem-specific descriptors calculated *in silico* for the palladium(I) dimer and problem-specific clustering through the k-means algorithm. From 21 promising predicted ligands that could afford the palladium(I) dimer *in lieu* of palladium(0) and palladium(II) species, eight were successfully verified experimentally.

In the above studies, the expert-guided molecular properties are often used as encoding descriptors, whose generation may require electronic structure calculation. Thanks to the fast development

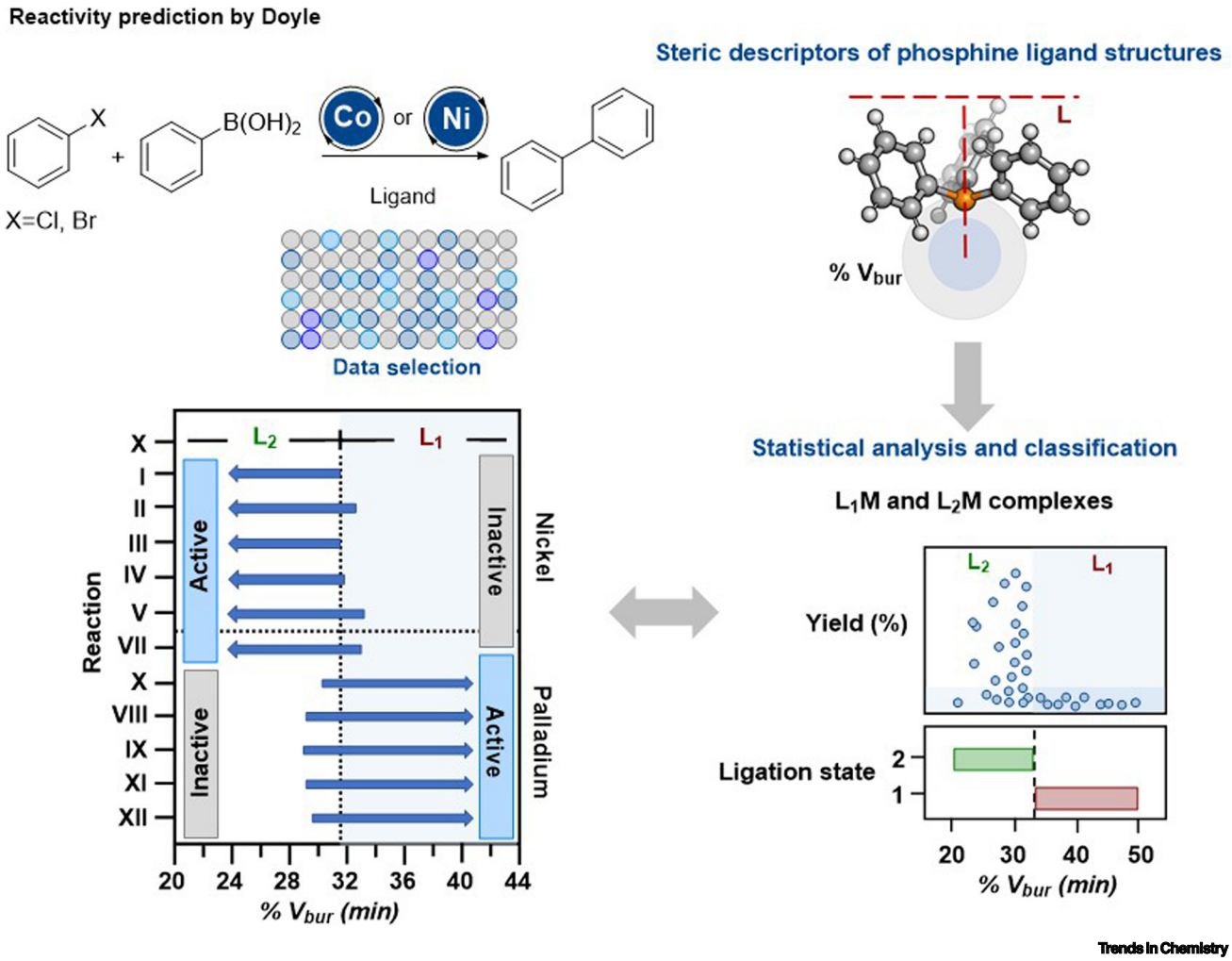


Figure 8. Machine learning (ML) for the prediction of chemical reactivity [55].

of ML models for molecular property prediction, accurate ML predictions were realized for both global molecular properties (i.e., HOMO and LUMO energies) [88,90–94,116] and site-specific properties including pK_a, [117,118], BDE [119], and atomic charge [120,121]. This allows on-the-fly generation of ML-predicted molecular properties and subsequent prediction of synthetic outcomes, which is demonstrated in a recent study from the Jensen group [95]. The authors trained a multitask directed message-passing neural network (D-MPNN) that can accurately generate molecular property descriptors on the fly and eventually achieved an end-to-end regioselectivity prediction in a series of organic transformations. The fusion model took advantage of the predictive value of molecular properties while greatly reducing the required computational cost.

Recently, Corminboeuf and coworkers developed NaviCatGA, a genetic algorithm for the optimization of homogeneous catalysts [122]. Moreover, this genetic algorithm is compatible with several structural representations, such as SMILES and 3D structures. It allows the identification of the most promising optimized tailored catalysts for a given catalytic system. It amazingly has the ability to trace down the origin of the catalyst components such as ligands and side chains. NaviCatGA also exhibits a broad applicability to several catalytic systems.

Reaction optimization

In the previous section, only the catalytic system was evaluated to correlate it with the yield or enantioselectivity of a transformation. However, deep learning models were also used to optimize several parameters of a reaction at once (Figure 9A). This accelerates the optimization process, which is traditionally conducted by changing one parameter at a time or by DOE. Moreover, this tedious experimental method does not necessarily allow access to the optimal conditions. Recently, a neural network model was trained to find the best catalyst but also to optimize catalyst loading, the residence time, and the temperature of a Suzuki–Miyaura cross-coupling reaction [42]. Furthermore, this deep learning model shows excellent generalizability to unexplored reaction-condition space and is reliable and robust even with unseen reactants.

The employment of Bayesian optimization algorithms has been proposed for chemical problems [123]. Thus, Doyle and coworkers recently demonstrated the power of Bayesian optimization algorithms applied to reaction optimizations [26]. The developed Bayesian reaction optimization framework was reported to be compatible with automated systems. The advantage of such implementations is the variable number of experiments per batch (parallelization), which allows fast screening and exploration of reaction conditions. The development of the optimizer was conducted with the exploration of palladium-catalyzed direct arylation reactions and applied to the optimization of Mitsunobu and deoxyfluorination reactions. In the development of the optimizer, chemical-based descriptors calculated through DFT, fingertip-based descriptors provided by Mordred, and one-hot-encoded (OHE) representations were considered. DFT-calculated descriptors were found to give more stable results. Interestingly, the Bayesian reaction optimizer surpassed human decision making in terms of both consistency and efficiency. This guidance can strongly facilitate and improve the performance of experimental chemists.

Synthesis planning

Computer scientists have also long tried to tackle the task of retrosynthesis. Interestingly, deep neural networks and symbolic AI were successfully used to avoid the typical errors in retrosynthetic systems of applying over-general rules, leading to unreasonable steps [32].

Monte Carlo tree search (MCTS) was used in combination with three different neural networks to propose chemical synthesis planning. The first one, the expansion policy, aims to propose a limited number of automatically extracted transformations to focus the research on promising directions. Then, the second is to determine the feasibility of the reactions being considered. The last one estimates the position value. These neural networks were trained on all of the reactions from the Reaxys database.

Through deep neural networks, MCTS searches by iterating over four phases (Figure 9B): (i) the most promising node is selected considering the current position values; (ii) the expansion procedure is used (including the first two neural networks); (iii) the rollout phase allows the evaluation of new positions (the third neural network); and (iv) the new position values are incorporated in the search tree. This system has shown promising results; however, there are still limitations. In particular, it does not allow the quantitative prediction of enantiomeric or diastereomeric ratios.

The assembly of natural products is extremely challenging for synthetic chemists. Recently, Grzybowski and coworkers achieved retrosynthesis planning of complex natural products implemented in the Synthia (former Chematica) software [43]. The program allows the exploration of multiple possible synthetic steps based on causal relationships between organic chemistry knowledge and AI routines. Synthia now encodes more than 100 000 mechanistic reaction rules.

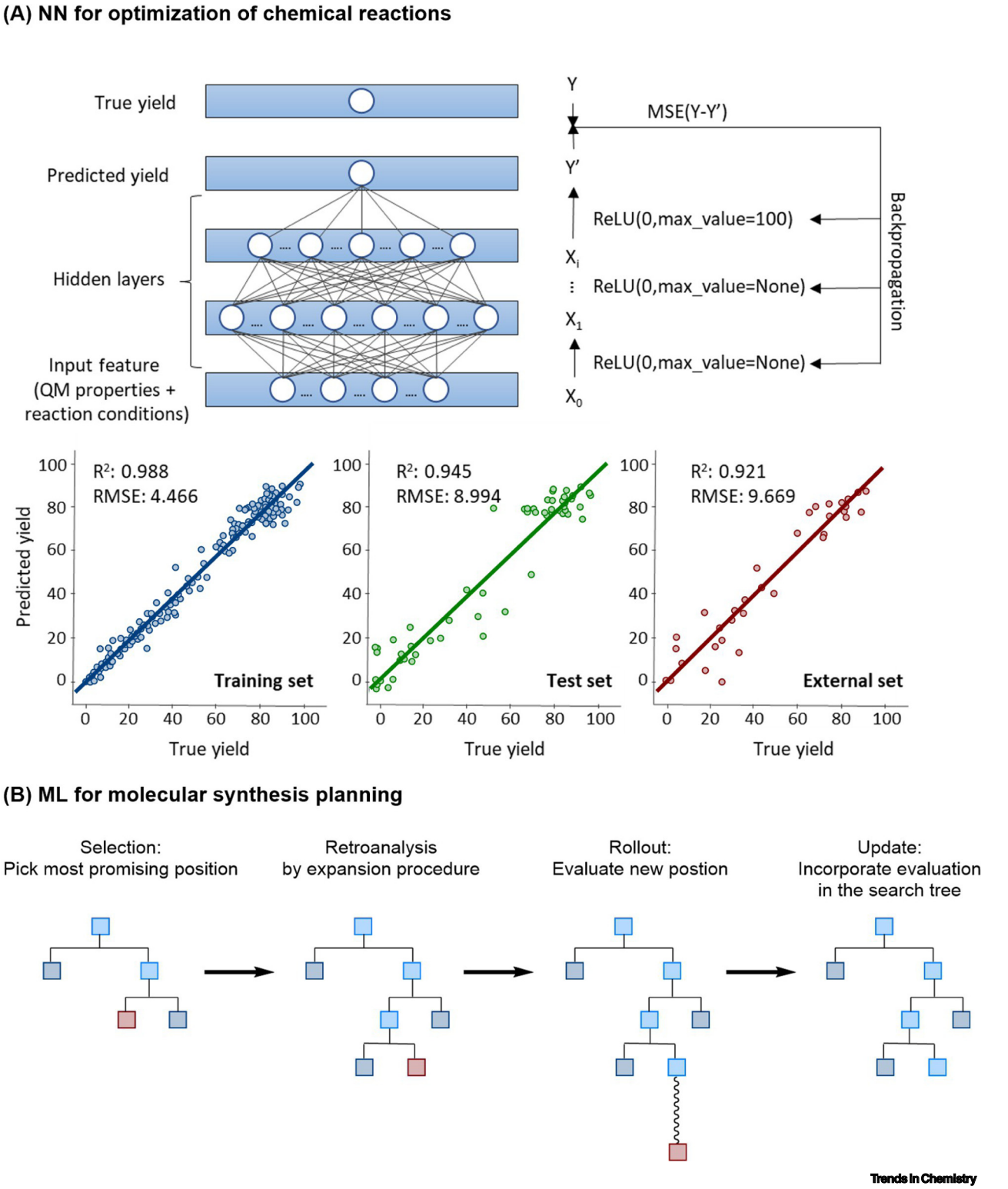


Figure 9. Machine learning (ML) as an attractive approach to chemical reaction optimization and molecular synthesis planning. (A) Neural networks applied to the optimization of chemical reactions [42]. (B) Application of Monte Carlo tree search (MCTS) to chemical synthesis planning [32].

Autonomous reaction exploration

One of the milestones in the symbiosis of AI and molecular synthesis is the human-free autonomous development of highly efficient novel chemical reactions, such as in catalysis. A key step into this direction has been achieved by Jensen and coworkers (Figure 10) [44]. The authors developed an automated robotic setup for the synthesis of organic compounds in flow. In such a process, ANNs and Monte Carlo tree algorithms were involved in the synthetic route exploration. The *in silico* synthetic routes are supported by millions of published chemical reactions.

Progress in the field of autonomous reaction optimization, also in flow, has been accomplished by Felpin and coworkers for direct C–H arylation reactions of indole-3-acetic acid derivatives [124]. Progress has also been archived in the development of user-friendly and open-source software modules for the automation of chemical reactions, such as Rxn Rover [125]. Also, Laino and coworkers recently developed RoboRXN for the design and autonomous production of compounds with reduced human intervention. (https://www.ibm.com/blogs/research/2020/08/roborxn-automating-chemical-synthesis/). One key aspect in the development of autonomous processes is the retrieval of published chemical reaction conditions. Several automated procedures have been reported [126–128].

The development of autonomous processes is highly adventurous, since they allow the exploration of a greater number of reactions, minimization of the number of routine tasks performed by bench chemists, and a significant reduction of chemical waste and the corresponding experimental processual costs.

Summary and outlook: pitfalls in ML

The exciting development of ML applications in organic chemistry is just the silhouette of the potential of data-driven approaches in the physical sciences. Through the accumulation and

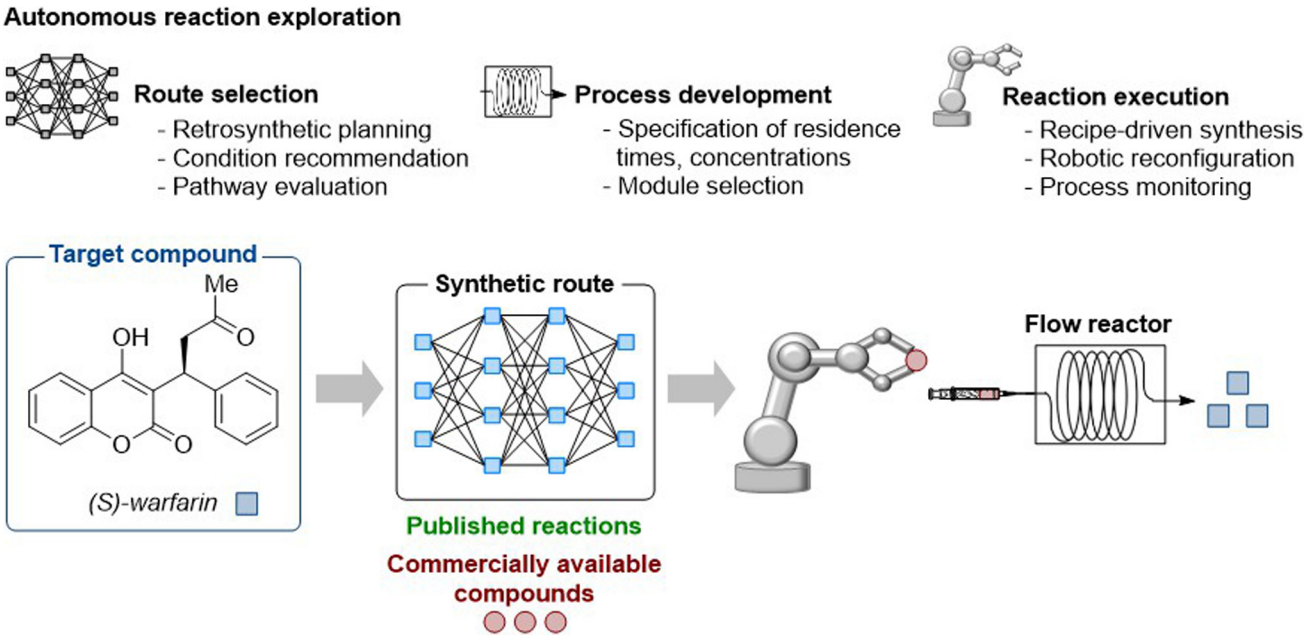


Figure 10. Artificial intelligence (AI)-assisted reconfigurable robot for multistep chemical syntheses in flow [44].

compilation of chemical data and the advance of ML frameworks for molecular science, we have the strong belief that AI will play a key role in driving innovation in organic chemistry. However, it should also be noted that the data-driven approach is not the 'panacea' for organic chemistry. There are certain pitfalls of ML, which require attention from interested users.

First, ML models typically lack a generalization ability to handle the infinite chemical space of organic chemistry. In contrast to the strong capability of interpolative prediction, the extrapolative prediction of ML models is rather weak. How to train a ML model to facilitate the design of innovative functional molecules or organic transformations remains elusive. To tackle this challenging problem, there are recent exciting explorations [28,40,96,129–132]. Grzybowski and coworkers showed that physical organic chemistry and transition-state knowledge-based descriptor design can improve the model's prediction performance on compounds that are not present in the training set [130,131]. The ability to achieve satisfying prediction on such 'unseen' compounds is highly desirable, especially to predict and design new catalysts or reagents for organic synthesis.

In addition, the available data in organic chemistry is a just tiny fraction of the potential chemical space. This differentiates ML modeling in chemical science from traditional AI areas such as computer vision or natural language processing, where big databases are available (e.g., ImageNet [133], WordNet [134]). In ML in organic chemistry, this problem of few-shot learning will be a longstanding issue for the foreseeable future. To provide ML predictions in the early stage of catalysis screening and design, Hong and colleagues developed a new ML approach called hierarchical learning [70]. Based on the structural similarity between the target compound and the compounds in the available dataset, the chemical database is divided into a number of hierarchies to enable the training of an ensemble model. The hierarchical learning model allows a chemical heuristics-based connection between the big literature data and the small sample of an ongoing study, which can serve as a useful approach for ML-assisted catalyst screening in organic synthesis.

Concluding remarks

The application of ML to molecular synthesis has gained substantial momentum in the past several years. It has been shown to be a powerful tool for the prediction of molecular properties as well as the outcome of catalytic reactions, most notably yields and selectivities. To achieve desired synthesis predictions, the success of ML models is strongly related to the quantity and particularly the quality of the available databases, as well as the choice of molecular descriptors. Currently, the number of datapoints needed for the generation of a reliable predictive model, however, is largely still elusive. Key questions remain, among others (see Outstanding questions). What is the minimum number of datapoints necessary to generate a reliable ML model? How can the rich knowledge of chemistry be implemented in a ML model to maximize the impact of synthetic data? In addition, digital representations of organic molecules and chemical transformations remain underdeveloped. This may lead to out-of-distribution issues in ML applications in molecular synthesis, limiting the extrapolative power of the trained model to make useful predictions for new reactants or catalysts. Only through the synergistic improvement of good data, representations, and algorithms can a reliable and robust predictive ML model for chemical innovation be generated. Given the advances witnessed in ML and HTE, along with the topical interest in molecular synthesis and catalysis, major progress is expected in this rapidly evolving arena.

Outstanding questions

How can a molecular system be turned into a computer-friendly input for the generation of a ML predictive model for catalysis and molecular synthesis?

What is the bottleneck limiting the availability of ML tools for the daily practice of synthetic discovery?

Which are the most effective ML algorithms for catalysis and molecular synthesis?

Can AI design new types of organic reactions?

How can we digitalize synthetic knowledge and create knowledge-based prediction models?

How can ML assist in the development of novel catalysts?

How accurately can ML predict catalytic reaction outcomes?

What is the minimal amount of synthetic data necessary to create a reliable ML predictive model?

What are the drawbacks of solely relying on positive synthetic literature results in the generation of ML models?

What are the pitfalls of ML in molecular synthesis and catalysis?

Acknowledgments

Generous support by the DFG (SPP1807, SPP2363, and Gottfried-Wilhelm-Leibniz award to L.A.), the ERC (ERC Advanced Grant to L.A.), the ITN CHAIR, the National Natural Science Foundation of China (21873081 and 22122109, X.H.; 22103070, S-Q.Z.), the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study (SN-ZJU-

SIAS-006, X.H.), Beijing National Laboratory for Molecular Sciences (BNLMS202102, X.H.), the Center of Chemistry for Frontier Technologies and Key Laboratory of Precise Synthesis of Functional Molecules of Zhejiang Province (PSFM 2021-01, X.H.), the State Key Laboratory of Clean Energy Utilization (ZJUCEU2020007, X.H.), and CAS Youth Interdisciplinary Team (JCTD-2021-11, X.H.) is gratefully acknowledged.

Declaration of interests

No interests are declared.

Resources

[i]https://doi.org/10.1055/s-0041-1737327

[ii]https://doi.org/10.6084/m9.figshare.5104873.v1

References

1. Ackermann, L. (2009) *Modern arylation methods*, Wiley
2. Beller, M. and Bolm, C. (2004) *Transition metals for organic synthesis*, Wiley
3. Nicolaou, K.C. and Chen, J.S. (2009) The art of total synthesis through cascade reactions. *Chem. Soc. Rev.* 38, 2993–3009
4. Rej, S. *et al.* (2020) Bidentate directing groups: an efficient tool in C–H bond functionalization chemistry for the expedient construction of C–C bonds. *Chem. Rev.* 120, 1788–1887
5. Gandeepan, P. *et al.* (2019) 3d transition metals for C–H activation. *Chem. Rev.* 119, 2192–2452
6. Park, Y. *et al.* (2017) Transition metal-catalyzed C–H amination: scope, mechanism, and applications. *Chem. Rev.* 117, 9247–9301
7. Moir, M. *et al.* (2019) An overview of late-stage functionalization in today's drug discovery. *Expert Opin. Drug Discovery* 14, 1137–1149
8. Cernak, T. *et al.* (2016) The medicinal chemist's toolbox for late stage functionalization of drug-like molecules. *Chem. Soc. Rev.* 45, 546–576
9. Koy, M. *et al.* (2021) *N*-Heterocyclic carbenes as tunable ligands for catalytic metal surfaces. *Nat. Catal.* 4, 352–363
10. Chen, H. *et al.* (2020) The progress and outlook of bioelectrocatalysis for the production of chemicals, fuels and materials. *Nat. Catal.* 3, 225–244
11. Kar, S. *et al.* (2022) Green chemistry in the synthesis of pharmaceuticals. *Chem. Rev.* 122, 3637–3710
12. Tang, T. *et al.* (2021) Analyzing mechanisms in Co(I) redox catalysis using a pattern recognition platform. *Chem. Sci.* 12, 4771–4778
13. Santiago, C.B. *et al.* (2018) Predictive and mechanistic multivariate linear regression models for reaction development. *Chem. Sci.* 9, 2398–2412
14. Reid, J.P. and Sigman, M.S. (2018) Comparing quantitative prediction methods for the discovery of small-molecule chiral catalysts. *Nat. Rev. Chem.* 2, 290–305
15. Niemeyer, Z.L. *et al.* (2016) Parameterization of phosphine ligands reveals mechanistic pathways and predicts reaction outcomes. *Nat. Chem.* 8, 610–617
16. Milo, A. *et al.* (2015) A data-intensive approach to mechanistic elucidation applied to chiral anion catalysis. *Science* 347, 737–743
17. Zhao, S. *et al.* (2018) Enantiodivergent Pd-catalyzed C–C bond formation enabled through ligand parameterization. *Science* 362, 670–674
18. Zuranski, A.M. *et al.* (2021) Predicting reaction yields via supervised learning. *Acc. Chem. Res.* 54, 1856–1865
19. Jorner, K. *et al.* (2021) Organic reactivity from mechanism to machine learning. *Nat. Rev. Chem.* 5, 240–255
20. Strieth-Kalthoff, F. *et al.* (2020) Machine learning the ropes: principles, applications and directions in synthetic chemistry. *Chem. Soc. Rev.* 49, 6154–6168
21. Cova, T.G.G.C. and Pais, A.A.C.C. (2019) Deep learning for deep chemistry: optimizing the prediction of chemical patterns. *Front. Chem.* 7, 809
22. Mater, A.C. and Coote, M.L. (2019) Deep learning in chemistry. *J. Chem. Inf. Model.* 59, 2545–2559
23. Chen, H. *et al.* (2018) The rise of deep learning in drug discovery. *Drug Discov. Today* 23, 1241–1250
24. Ma, J. *et al.* (2015) Deep neural nets as a method for quantitative structure–activity relationships. *J. Chem. Inf. Model.* 55, 263–274
25. Lavecchia, A. (2015) Machine-learning approaches in drug discovery: methods and applications. *Drug Discov. Today* 20, 318–331
26. Shields, B.J. *et al.* (2021) Bayesian reaction optimization as a tool for chemical synthesis. *Nature* 590, 89–96
27. Nielsen, M.K. *et al.* (2018) Deoxyfluorination with sulfonyl fluorides: navigating reaction space with machine learning. *J. Am. Chem. Soc.* 140, 5004–5008
28. Ahneman, D.T. *et al.* (2018) Predicting reaction performance in C–N cross-coupling using machine learning. *Science* 360, 186–190
29. Coley, C.W. *et al.* (2017) Prediction of organic reaction outcomes using machine learning. *ACS Cent. Sci.* 3, 434–443
30. Wei, J.N. *et al.* (2016) Neural networks for the prediction of organic chemistry reactions. *ACS Cent. Sci.* 2, 725–732
31. Kayala, M.A. *et al.* (2011) Learning to predict chemical reactions. *J. Chem. Inf. Model.* 51, 2209–2222
32. Segler, M.H.S. *et al.* (2018) Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* 555, 604–610
33. Coley, C.W. *et al.* (2018) Machine learning in computer-aided synthesis planning. *Acc. Chem. Res.* 51, 1281–1289
34. Perera, D. *et al.* (2018) A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science* 359, 429–434
35. Buitrago Santanilla, A. *et al.* (2015) Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science* 347, 49–53
36. Ramakrishnan, R. *et al.* (2014) Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* 1, 140022
37. Ruddigkeit, L. *et al.* (2012) Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* 52, 2864–2875
38. Blum, L.C. and Reymond, J.L. (2009) 970 Million druglike small molecules for virtual screening in the chemical universe database GDB-13. *J. Am. Chem. Soc.* 131, 8732–8733
39. Friis, S.D. *et al.* (2020) Cobalt-catalysed C–H methylation for late-stage drug diversification. *Nat. Chem.* 12, 511–519
40. Zahrt, A.F. *et al.* (2019) Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science* 363, eaau5631
41. Meyer, B. *et al.* (2018) Machine learning meets volcano plots: computational discovery of cross-coupling catalysts. *Chem. Sci.* 9, 7069–7077
42. Fu, Z.Y. *et al.* (2020) Optimizing chemical reaction conditions using deep learning: a case study for the Suzuki–Miyaura cross-coupling reaction. *Org. Chem. Front.* 7, 2269–2277
43. Mikulak-Klucznik, B. *et al.* (2020) Computational planning of the synthesis of complex natural products. *Nature* 588, 83–88
44. Coley, C.W. *et al.* (2019) A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* 365, eaax1566
45. Heller, S.R. *et al.* (2015) InChI, the IUPAC International Chemical Identifier. *J. Cheminform.* 7, 23

46. Weininger, D. (1988) Smiles, a chemical language and information-system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* 28, 31–36
47. Krenn, M. *et al.* (2020) Self-referencing embedded strings (SELFIES): a 100% robust molecular string representation. *Mach. Learn. Sci. Technol.* 1, 045024
48. Bolton, E.E. *et al.* (2008) PubChem: integrated platform of small molecules and biological activities. In *Annual reports in computational chemistry* (Wheeler, R.A. and Spellmeyer, D.C., eds), pp. 217–241, Elsevier
49. Durant, J.L. *et al.* (2002) Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* 42, 1273–1280
50. Rogers, D. and Hahn, M. (2010) Extended-connectivity fingerprints. *J. Chem. Inf. Model.* 50, 742–754
51. Morgan, H.L. (2002) The generation of a unique machine description for chemical structures – a technique developed at Chemical Abstracts Service. *J. Chem. Doc.* 5, 107–113
52. Feinberg, E.N. *et al.* (2018) PotentialNet for molecular property prediction. *ACS Cent. Sci.* 4, 1520–1530
53. Coley, C.W. *et al.* (2019) A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* 10, 370–377
54. Lau, S.H. *et al.* (2021) Ni/Photoredox-catalyzed enantioselective cross-electrophile coupling of styrene oxides with aryl iodides. *J. Am. Chem. Soc.* 143, 15873–15881
55. Newman-Stonebraker, S.H. *et al.* (2021) Univariate classification of phosphine ligation state and reactivity in cross-coupling catalysis. *Science* 374, 301–308
56. Eckhoff, M. and Behler, J. (2021) High-dimensional neural network potentials for magnetic systems using spin-dependent atom-centered symmetry functions. *npj Comput. Mater.* 7, 170
57. Falivene, L. *et al.* (2019) Towards the online computer-aided design of catalytic pockets. *Nat. Chem.* 11, 872–879
58. Bonaccorso, G. (2018) *Mastering machine learning algorithms*, Packt
59. Lan, T. and An, Q. (2021) Discovering catalytic reaction networks using deep reinforcement learning from first-principles. *J. Am. Chem. Soc.* 143, 16804–16812
60. Zhou, Z. *et al.* (2019) Optimization of molecules via deep reinforcement learning. *Sci. Rep.* 9, 10752
61. Jaeger, S. *et al.* (2018) Mol2vec: unsupervised machine learning approach with chemical intuition. *J. Chem. Inf. Model.* 58, 27–35
62. Werth, J. and Sigman, M.S. (2020) Connecting and analyzing enantioselective bifunctional hydrogen bond donor catalysis using data science tools. *J. Am. Chem. Soc.* 142, 16382–16391
63. Reid, J.P. and Sigman, M.S. (2019) Holistic prediction of enantioselectivity in asymmetric catalysis. *Nature* 571, 343–348
64. Fooshee, D. *et al.* (2018) Deep learning for chemical reaction prediction. *Mol. Syst. Des. Eng.* 3, 442–452
65. Rinehart, N.I. *et al.* (2021) Dreams, false starts, dead ends, and redemption: a chronicle of the evolution of a chemoinformatic workflow for the optimization of enantioselective catalysts. *Acc. Chem. Res.* 54, 2041–2054
66. Walters, W.P. and Barzilay, R. (2021) Applications of deep learning in molecule generation and molecular property prediction. *Acc. Chem. Res.* 54, 263–270
67. Gallegos, L.C. *et al.* (2021) Importance of engineered and learned molecular representations in predicting organic reactivity, selectivity, and chemical properties. *Acc. Chem. Res.* 54, 827–836
68. Kearnes, S.M. *et al.* (2021) The Open Reaction Database. *J. Am. Chem. Soc.* 143, 18820–18826
69. Lopez, S.A. *et al.* (2016) The Harvard Organic Photovoltaic Dataset. *Sci. Data* 3, 160086
70. Xu, L.C. *et al.* (2021) Towards data-driven design of asymmetric hydrogenation of olefins: database and hierarchical learning. *Angew. Chem. Int. Ed.* 60, 22804–22811
71. Richard, A.M. *et al.* (2021) The Tox21 10K Compound Library: collaborative chemistry advancing toxicology. *Chem. Res. Toxicol.* 34, 189–216
72. Burley, K.H. *et al.* (2019) Enhancing side chain rotamer sampling using nonequilibrium candidate Monte Carlo. *J. Chem. Theory Comput.* 15, 1848–1862
73. Mayr, H. *et al.* (2003) π-Nucleophilicity in carbon–carbon bond-forming reactions. *Acc. Chem. Res.* 36, 66–77
74. Streidl, N. *et al.* (2010) A practical guide for estimating rates of heterolysis reactions. *Acc. Chem. Res.* 43, 1537–1549
75. Mayr, H. and Ofial, A.R. (2016) Philicities, fugalities, and equilibrium constants. *Acc. Chem. Res.* 49, 952–965
76. Mayr, H. and Patz, M. (1994) Scales of nucleophilicity and electrophilicity – a system for ordering polar organic and organometallic reactions. *Angew. Chem. Int. Ed.* 33, 938–957
77. Mayr, H. and Ofial, A.R. (2008) Do general nucleophilicity scales exist? *J. Phys. Org. Chem.* 21, 584–595
78. Mayr, H. and Ofial, A.R. (2015) A quantitative approach to polar organic reactivity. *SAR QSAR Environ. Res.* 26, 619–646
79. An, F. *et al.* (2020) Basicities and nucleophilicities of pyrrolidines and imidazolidinones used as organocatalysts. *J. Am. Chem. Soc.* 142, 1526–1547
80. Mayr, H. (2015) Reactivity scales for quantifying polar organic reactivity: the benzhydrylium methodology. *Tetrahedron* 71, 5095–5111
81. Ammer, J. *et al.* (2012) Free energy relationships for reactions of substituted benzhydrylium ions: from enthalpy over entropy to diffusion control. *J. Am. Chem. Soc.* 134, 13902–13911
82. Thakkar, A. *et al.* (2020) Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. *Chem. Sci.* 11, 154–168
83. Schneider, N. *et al.* (2016) Big data from pharmaceutical patents: a computational analysis of medicinal chemists' bread and butter. *J. Med. Chem.* 59, 4385–4402
84. Schneider, N. *et al.* (2016) What's what: the (nearly) definitive guide to reaction role assignment. *J. Chem. Inf. Model.* 56, 2336–2346
85. Cheong, P.H.-Y. *et al.* (2011) Quantum mechanical investigations of organocatalysis: mechanisms, reactivities, and selectivities. *Chem. Rev.* 111, 5042–5137
86. Lam, Y.H. *et al.* (2016) Theory and modeling of asymmetric catalytic reactions. *Acc. Chem. Res.* 49, 750–762
87. Zahrt, A.F. *et al.* (2020) Quantitative structure–selectivity relationships in enantioselective catalysis: past, present, and future. *Chem. Rev.* 120, 1620–1689
88. Unke, O.T. and Meuwly, M. (2019) PhysNet: a neural network for predicting energies, forces, dipole moments, and partial charges. *J. Chem. Theory Comput.* 15, 3678–3693
89. Liu, Z. *et al.* (2021) Transferable multilevel attention neural network for accurate prediction of quantum chemistry properties via multitask learning. *J. Chem. Inf. Model.* 61, 1066–1082
90. Schütt, K.T. *et al.* (2021) Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 9377–9388, PMLR
91. Liu, Y. *et al.* (2021) Spherical message passing for 3D graph networks. *arXiv* Published online February 9, 2021. https://doi.org/10.48550/arXiv.2102.05013
92. Klicpera, J. *et al.* (2020) Directional message passing for molecular graphs. In *International Conference on Learning Representations*, ICLR
93. Klicpera, J. *et al.* (2020) Fast and uncertainty-aware directional message passing for non-equilibrium molecules. In *Advances in neural information processing systems* (Vol. 33), Curran Associates
94. Anderson, B.M. *et al.* (2019) Cormorant: covariant molecular neural networks. In *Advances in neural information processing systems* (Vol. 32), pp. 14537–14546, Curran Associates
95. Guan, Y. *et al.* (2021) Regio-selectivity prediction with a machine-learned reaction representation and on-the-fly quantum mechanical descriptors. *Chem. Sci.* 12, 2198–2208
96. Li, X. *et al.* (2020) Predicting regioselectivity in radical C–H functionalization of heterocycles through machine learning. *Angew. Chem. Int. Ed.* 59, 13253–13259
97. Baxter, R.D. *et al.* (2015) Mechanistic insights into two-phase radical C–H arylations. *ACS Cent. Sci.* 1, 456–462
98. Smith, J.M. *et al.* (2019) Alkyl sulfinates: radical precursors enabling drug discovery. *J. Med. Chem.* 62, 2256–2264
99. Dreher, S.D. and Krska, S.W. (2021) Chemistry informer libraries: conception, early experience, and role in the future of cheminformatics. *Acc. Chem. Res.* 54, 1586–1596

100. Mdluli, V. *et al.* (2020) High-throughput synthesis and screening of iridium(III) photocatalysts for the fast and chemoselective dehalogenation of aryl bromides. *ACS Catal.* 10, 6977–6987
101. Kutchukian, P.S. *et al.* (2016) Chemistry informer libraries: a chemoinformatics enabled approach to evaluate and advance synthetic methods. *Chem. Sci.* 7, 2604–2613
102. Stadler, A. and Kappe, C.O. (2001) Automated library generation using sequential microwave-assisted chemistry. application toward the Biginelli multicomponent condensation. *J. Comb. Chem.* 3, 624–630
103. Gioiello, A. *et al.* (2013) Building a sulfonamide library by eco-friendly flow synthesis. *ACS Comb. Sci.* 15, 235–239
104. DeLano, T.J. and Reisman, S.E. (2019) Enantioselective electroreductive coupling of alkenyl and benzyl halides via nickel catalysis. *ACS Catal.* 9, 6751–6754
105. Zuo, Z. *et al.* (2014) Merging photoredox with nickel catalysis: coupling of α-carboxyl sp3-carbons with aryl halides. *Science* 345, 437–440
106. Christensen, M. *et al.* (2019) Development of an automated kinetic profiling system with online HPLC for reaction optimization. *React. Chem. Eng.* 4, 1555–1558
107. Huffman, M.A. *et al.* (2019) Design of an *in vitro* biocatalytic cascade for the manufacture of islatravir. *Science* 366, 1255–1259
108. Liu, R.Y. and Buchwald, S.L. (2018) Copper-catalyzed enantioselective hydroamination of alkenes. *Org. Synth.* 95, 80–96
109. Cordova, M. *et al.* (2020) Data-driven advancement of homogeneous nickel catalyst activity for aryl ether cleavage. *ACS Catal.* 10, 7021–7031
110. Henle, J.J. *et al.* (2020) Development of a computer-guided workflow for catalyst optimization. descriptor validation, subset selection, and training set analysis. *J. Am. Chem. Soc.* 142, 11578–11592
111. See, X.Y. *et al.* (2020) Iterative supervised principal component analysis driven ligand design for regioselective Ti-catalyzed pyrrole synthesis. *ACS Catal.* 10, 13504–13517
112. Hueffel, J.A. *et al.* (2021) Accelerated dinuclear palladium catalyst identification through unsupervised machine learning. *Science* 374, 1134–1140
113. Durand, D.J. and Fey, N. (2019) Computational ligand descriptors for catalyst design. *Chem. Rev.* 119, 6561–6594
114. Fey, N. *et al.* (2006) Development of a ligand knowledge base, part 1: computational descriptors for phosphorus donor ligands. *Chem. Eur. J.* 12, 291–302
115. Jover, J. *et al.* (2010) Expansion of the ligand knowledge base for monodentate P-donor ligands (LKB-P). *Organometallics* 29, 6245–6258
116. Schütt, K.T. *et al.* (2018) SchNet – a deep learning architecture for molecules and materials. *J. Chem. Phys.* 148, 241722
117. Roszak, R. *et al.* (2019) Rapid and accurate prediction of pK_a values of C–H acids using graph convolutional neural networks. *J. Am. Chem. Soc.* 141, 17142–17149
118. Yang, Q. *et al.* (2020) Holistic prediction of the pK_a in diverse solvents based on a machine-learning approach. *Angew. Chem. Int. Ed.* 59, 19282–19291
119. St John, P.C. *et al.* (2020) Prediction of organic homolytic bond dissociation enthalpies at near chemical accuracy with sub-second computational cost. *Nat. Commun.* 11, 2328
120. Bleiziffer, P. *et al.* (2018) Machine learning of partial charges derived from high-quality quantum-mechanical calculations. *J. Chem. Inf. Model.* 58, 579–590
121. Nebgen, B. *et al.* (2018) Transferable dynamic molecular charge assignment using deep neural networks. *J. Chem. Theory Comput.* 14, 4687–4698
122. Laplaza, R. *et al.* (2022) Genetic optimization of homogeneous catalysts. *Chem. Methods* 2, e202100107
123. Hase, F. *et al.* (2018) Phoenics: a Bayesian optimizer for chemistry. *ACS Cent. Sci.* 4, 1134–1145
124. Vasudevan, N. *et al.* (2020) Direct C–H arylation of indole-3-acetic acid derivatives enabled by an autonomous self-optimizing flow reactor. *Adv. Synth. Catal.* 363, 791–799
125. Crandall, Z. *et al.* (2022) Rxn Rover: automation of chemical reactions with user-friendly, modular software. *React. Chem. Eng.* 7, 416–428
126. Guo, J. *et al.* (2021) Correction to automated chemical reaction extraction from scientific literature. *J. Chem. Inf. Model.* 61, 4124
127. Vaucher, A.C. *et al.* (2020) Automated extraction of chemical synthesis actions from experimental procedures. *Nat. Commun.* 11, 3601
128. Burger, B. *et al.* (2020) A mobile robotic chemist. *Nature* 583, 237–241
129. Tomberg, A. *et al.* (2019) A predictive tool for electrophilic aromatic substitutions using machine learning. *J. Org. Chem.* 84, 4695–4703
130. Moskal, M. *et al.* (2021) Scaffold-directed face selectivity machine-learned from vectors of non-covalent interactions. *Angew. Chem. Int. Ed.* 60, 15230–15235
131. Beker, W. *et al.* (2019) Prediction of major regio-, site-, and diastereoisomers in Diels–Alder reactions by using machine-learning: the importance of physically meaningful descriptors. *Angew. Chem. Int. Ed.* 58, 4515–4519
132. Yang, L.C. *et al.* (2021) Machine learning prediction of hydrogen atom transfer reactivity in photoredox-mediated C–H functionalization. *Org. Chem. Front.* 8, 6187–6195
133. Russakovsky, O. *et al.* (2015) ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* 115, 211–252
134. Fellbaum, C. (1998) *WordNet: an electronic lexical database*, MIT Press