Engineering 27 (2023) 70–83

Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Smart Process Manufacturing toward Carbon Neutrality—Review

Machine Learning for Chemistry: Basics and Applications

Yun-Fei Shi [a,#], Zheng-Xin Yang [a,#], Sicong Ma [b], Pei-Lin Kang [a], Cheng Shang [a], P. Hu [c,*], Zhi-Pan Liu [a,b,*]

[a] Collaborative Innovation Center of Chemistry for Energy Material, Shanghai Key Laboratory of Molecular Catalysis and Innovative Materials, Key Laboratory of Computational Physical Sciences of the Ministry of Education, Department of Chemistry, Fudan University, Shanghai 200433, China
[b] Key Laboratory of Synthetic and Self-Assembly Chemistry for Organic Functional Molecules, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai 200032, China
[c] School of Chemistry and Chemical Engineering, Queen's University Belfast, Belfast BT9 5AG, UK

ARTICLE INFO

Article history:
Received 31 August 2022
Revised 19 January 2023
Accepted 6 April 2023
Available online 31 July 2023

Keywords:
Machine learning
Atomic simulation
Catalysis
Retrosynthesis
Neural network potential

ABSTRACT

The past decade has seen a sharp increase in machine learning (ML) applications in scientific research. This review introduces the basic constituents of ML, including databases, features, and algorithms, and highlights a few important achievements in chemistry that have been aided by ML techniques. The described databases include some of the most popular chemical databases for molecules and materials obtained from either experiments or computational calculations. Important two-dimensional (2D) and three-dimensional (3D) features representing the chemical environment of molecules and solids are briefly introduced. Decision tree and deep learning neural network algorithms are overviewed to emphasize their frameworks and typical application scenarios. Three important fields of ML in chemistry are discussed: ① retrosynthesis, in which ML predicts the likely routes of organic synthesis; ② atomic simulations, which utilize the ML potential to accelerate potential energy surface sampling; and ③ heterogeneous catalysis, in which ML assists in various aspects of catalytic design, ranging from synthetic condition optimization to reaction mechanism exploration. Finally, a prospect on future ML applications is provided.

1. Introduction

It has long been a dream in history for humans to invent machines with human-like intelligence that can automatically complete complex tasks. This dream has never come so true as it has in the past decade, which has witnessed the rapid applications of machine learning (ML) techniques and artificial intelligence (AI) machines in various areas of human activity. The development of new ML models—particularly deep learning methods [1]—and sharply increased data storage capability are key to the recent surge in ML cases. Apart from successful ML achievements in everyday life, such as image recognition [2] and speech recognition [3], ML has drawn a great deal of attention in modern scientific research; for example, the AlphaFold algorithm for predicting protein structure has demonstrated its power as a game-changer in structural biology [4,5]. This review will focus on recent advances of ML applications in chemistry research, which inherently contains a huge amount of data, relating to the material complexity and the huge variety of organic molecules.

Chemists are educated to perform experiments and collect data but are generally much less familiar with modern ML algorithms [6]. Unlike the computer-aided chemical research in the 1990s that was largely based on theoretical/empirical rules [7], current ML applications rely on big datasets carrying all the essential information [8,9]. Poor quality of datasets may well create unnecessary difficulties for ML applications that should in principle be feasible and straightforward [10]. A common problem with chemistry datasets is the heavy bias toward successful experiments. In fact, not only good data (e.g., producing the desired products) but also bad data (e.g., failed experiments) are required in order to provide a balanced view of the chemical space. In addition, due to the complexity of chemical experiments, the synthetic conditions documented in the literature are often incomplete, with important variables being overlooked. For these reasons, it is no wonder that—compared with experimental fields—ML applications are much more popular in computational chemistry, where datasets

* Corresponding authors.
E-mail addresses: p.hu@qub.ac.uk (P. Hu), zpliu@fudan.edu.cn (Z.-P. Liu).
# These authors contributed equally to this work.

https://doi.org/10.1016/j.eng.2023.04.013

can be reliably and consistently constructed from quantum mechanics (QM) calculations. These computed datasets can be utilized to directly benchmark the physicochemical properties of molecules and materials and to develop advanced computational methods. Therefore, it is imperative for chemists to equip a basic knowledge of ML, which would benefit them profoundly, from data recording to practicing ML-guided experiments.

For this purpose, this review will first introduce popular chemistry databases, which provide a basis for practicing ML models. Second, some widely-used two-dimensional (2D) and three-dimensional (3D) features are presented, which transform molecular structures into acceptable inputs for ML models. Third, popular ML algorithms are briefly overviewed, with a focus on their basic theoretical framework and suitable application scenarios. Finally, three chemistry fields with important progress in ML are described in more detail, including retrosynthesis in organic chemistry, ML-potential-based atomic simulation, and ML for heterogeneous catalysis. These applications either greatly expedite the original research by reducing the experimental/simulation cost or provide a new route for solving complex problems in a rational way. An outlook of future challenges is provided at the end.

2. Data

There is no artificial intelligence (AI) without data. Thus, the availability of data is the prerequisite for modern ML applications, where both the size and the quality of the dataset matter. In the field of chemistry, there has been a long tradition of collecting and compiling data, ranging from element atomic spectra to material macroscopic properties. The data science in chemistry has created the subject of chemical informatics, which further greatly benefits the applications of ML in chemistry. In fact, although it may appear to be daunting to build a large dataset from scratch, many chemical databases were available well before the ML era. Table 1 lists selected popular databases in chemistry, many of which have a long history of data collection and compilation. The sources of these data include open patents and research articles, high-throughput experiments toward specific properties, and QM calculations, typically based on density functional theory (DFT).

2.1. Chemical reaction databases

Chemical reaction databases hold high value for experimentalists in the design of synthetic routes and are particularly useful

Table 1
A list of popular chemical databases commonly used in ML.

Classification	Name	Content	URL
Chemical reaction databases	SciFinder	Information on chemical compounds, bibliographic data, and chemical reactions (commercial database)	https://scifinder.cas.org/
	Reaxys	Chemical reaction and bibliographic information (commercial database)	https://www.reaxys.com/
	USPTO	Chemical structure and reaction	https://www.repository.cam.ac.uk/handle/1810/244727
	ORD	Organic chemical reaction data	https://github.com/open-reaction-database
	NextMove	Chemical reaction data	https://www.nextmovesoftware.com/about.html
Chemical property databases	PubChem	Chemical and physical properties, biological activities, and toxicity of substances	https://pubchem.ncbi.nlm.nih.gov/
	NIST	Standard physicochemical properties of compounds	https://webbook.nist.gov/chemistry/
	ChemSpider	Structure and property of compounds	https://www.chemspider.com
	ChemBL	Drug-like properties of bioactive molecules	https://www.ebi.ac.uk/chembl/
	DrugBank	Properties of drug molecules	https://go.drugbank.com/releases/latest
	Tox21	Toxic effects of substances	https://ntp.niehs.nih.gov/whatwestudy/tox21/index.html
	ESOL	Water solubility of compounds	https://doi.org/10.1021/ci034243x
	FreeSolv	Water solubility of small neutral molecules	https://github.com/MobleyLab/FreeSolv
	Lipophilicity	Lipid solubility of organic compounds	https://doi.org/10.1002/cem.2718
Material databases	CSD	Organic and metal–organic crystal structures	https://www.ccdc.cam.ac.uk/
	ICSD	Inorganic and metal–organic crystal structures	https://icsd.products.fiz-karlsruhe.de/
	PDF	Diffraction data of inorganic and organic compounds	https://www.icdd.com/pdfsearch/
	MatWeb	The thermoplastic and thermoset of polymers, metals, and other engineering materials	https://matweb.com/
	Li-ion Battery Aging Datasets	Charge and discharge curves of lithium batteries	https://data.nasa.gov/dataset/Li-ion-Battery-Aging-Datasets/uj5r-zjdb
	HTEM	Experimental information of inorganic thin-film materials	https://htem.nrel.gov/
Computational chemistry databases	GDB-17	Structures of organic molecules up to 17 atoms	https://www.gdb.unibe.ch/downloads/
	QM9	Quantum chemical properties of organic molecules	https://quantum-machine.org/datasets/
	ANI-1	Energy and force of non-equilibrium molecules	https://github.com/isayev/ANI1_dataset
	Materials Project	DFT relaxed material structures and their thermal, electronic, and elastic properties	https://next-gen.materialsproject.org/
	OQMD	DFT relaxed material structures and their thermal, electronic, and elastic properties	https://oqmd.org/
	Aflowlib	DFT relaxed material structures and their thermal, electronic, and elastic properties	https://aflowlib.org/
	MD17/ISO-17	Energy and force of non-equilibrium molecules	https://quantum-machine.org/datasets/
	LASP	Global PES dataset of molecules/materials	https://www.lasphub.com
	OC20	Adsorption energy of molecules in catalysts	https://opencatalystproject.org/
	Atom3D	3D structure of molecules, RNA, and proteins	https://www.atom3d.ai/

URL: uniform resource locator; USPTO: United States Patent and Trademark Office; ORD: Open Reaction Database; NIST: National Institute of Standards and Technology; CSD: Cambridge Structural Database; ICSD: Inorganic Crystal Structure Database; PDF: Powder Diffraction File; HTEM: High-Throughput Experimental Materials; OQMD: Open Quantum Materials Database; OC20: Open Catalyst 2020; DFT: density functional theory; PES: potential energy surface.

in organic chemistry. Before the Internet was available, reactions in the literature had already been indexed by the Chemical Abstracts Service (CAS). These data can now can be accessed from SciFinder, which includes chemical and bibliographic information from journals, patents, books, and other sources. However, SciFinder, along with a similar commercial database, Reaxys, are unable to export large amounts of chemical compound and chemical reaction data in batches, which limits the size of the training datasets required for deep ML. For this reason, researchers use text processing techniques to extract reaction information from United States Patent and Trademark Office (USPTO) patents [11], which are open source and downloadable from the Internet. More recently, the Open Reaction Database (ORD) [12] established a data format template for chemical reaction storage that supports the data sharing of public chemical reaction datasets. It should be mentioned that an increasing number of researchers in the field of computer-aided synthesis now make their databases publicly available—such as by using NextMove software [13], which provides open-source text mining tools for identifying chemicals—and share their datasets for downloading and online querying.

2.2. Chemical property databases

There are many databases in the category of chemical property databases, due to the wide variety of chemical properties. PubChem [14] is an open chemical database that focuses on chemical and physical properties, biological activities, and the toxicity of substances. Since 1996, the National Institute of Standards and Technology (NIST) has released the Chemistry WebBook [15], which collects the spectroscopic and thermodynamic data initially published in handbooks and tables; it also includes other basic data on physics and chemistry, such as ionization energetics, solubility, and spectroscopic, chromatographic, and computational data. These datasets are available for batch download on the website. Similarly, ChemSpider [16] compiles publicly available web databases that provide the structure and properties of molecules. Apart from general databases, there are also a number of datasets focusing on specific properties, such as the biological activity of drugs in ChemBL [17] and DrugBank [18], the toxic effects of compounds in the Tox21 dataset [19] (covering 12 707 representative chemical compounds and 12 different toxic effects) obtained via high-throughput toxicity assays, the experimental solubility of small molecules in ESOL [20] (covering the water solubility data for organic small molecules), data on the solubility and calculated hydration free energy of small molecules in water in FreeSolv [21], and experimental data on the octanol–water partition coefficient for organic small molecules in Lipophilicity [22].

2.3. Material databases

For solid materials, the Cambridge Structural Database (CSD) [23] is the most recognized; it collects organic crystal structure information from the literature, including X-ray or neutron diffraction data, crystallization conditions, and experiment records on the conformation determination. The Inorganic Crystal Structure Database (ICSD) [24] contains more than 272 000 crystal structures, along with the molecular formula, atomic coordinates, cell parameters, space groups, and other information, mostly determined by experiments. The Powder Diffraction File (PDF) [25] database provides the diffraction and crystallographic data of 1 143 236 materials (Release 2023). The PDF was originally a collection of single-phase X-ray powder diffraction patterns; however, in recent years, it has also partly included atomic coordinates entries from the CSD, ICSD, NIST, and so forth. The MatWeb database covers a wide range of engineering materials, such as thermoplastic and thermoset polymers, metallic materials, and ceramic materials, recording the physical properties (e.g., water absorption, specific gravity), mechanical properties (e.g., modulus of elasticity), thermodynamic properties (e.g., melting point), and electrical properties (e.g., dipole moment, electrical resistance). Other more specific databases include the Li-ion Battery Aging Datasets [26] for lithium (Li)-ion battery materials from the National Aeronautics and Space Administration (NASA) Ames Prognostics Center and the High-Throughput Experimental Materials (HTEM) dataset [27] for inorganic thin-film materials. The former collects operating profiles, such as the charging, discharging, and electrochemical impedance spectroscopy of the battery material, while the latter includes information on the synthetic conditions, chemical composition, crystal structure, and characteristics of thin-film materials.

2.4. Computational chemistry databases

For the ease of first-principles calculations, computational chemistry databases are becoming a major source of chemistry data nowadays. The obvious advantages of computational data include their high accuracy, self-consistency, and good reproducibility (even for compounds that are difficult to synthesize in experiments). The GDB-17 database [28] has often been utilized in the literature for ML applications, as it contains 166.4 billion organic molecules with up to 17 atoms of carbon (C), nitrogen (N), oxygen (O), sulfur (S), and halogens. These molecules are enumerated and filtered by the strain topology and stability criteria, which are indexed using the simplified molecular-input line-entry system (SMILES) [29] name to differentiate by molecular composition and connection. The QM9 dataset [30] is a benchmark dataset for quantum chemical properties; it is made up of equilibrium organic compounds from the GDB-17 database with up to nine "heavy" atoms from the range of C, N, O, and fluorine (F) [30]. It also offers comparable harmonic frequencies, dipole moments, polarizabilities, energies, enthalpies, and free energies, in addition to energy minima, which are calculated at DFT B3LYP/6-31G (2df, p) level. In parallel with small-molecule databases, there are many material datasets as well, including the Materials Project [31], the Open Quantum Materials Database (OQMD) [32], and the Aflowlib database [33,34], which provide web-based open access to the DFT-optimized (mostly Perdew–Burke–Ernzerhof (PBE) functional) structures and computed properties of millions of known or predicted materials. These projects are often accompanied by Python packages, such as pymatgen [35] for the Materials Project, qmpy [32] for OQMD, and AFLOW [33] for Aflowlib, which offer a high-throughput DFT calculation framework to expand the dataset, as well as post-processing tools to analyze the data.

To expand the chemical space, significant efforts have been made to create off-equilibrium datasets, such as by using molecular dynamics (MD) simulations. The ANI-1 dataset [36], which is one such example, contains 20 million non-equilibrium molecules. This dataset was created from 57 000 different molecular configurations comprising the chemical components C, hydrogen (H), N, and O. The MD17 [37] and ISO-17 datasets [38] are other examples of the benchmark for quantum chemical properties; they contain off-equilibrium molecules, which are obtained from finite-temperature MD simulations of molecules with different conformations. Moreover, LASP software [39] provides a global potential energy surface (PES) dataset for molecules and materials obtained from stochastic surface walking (SSW) global PES exploration, and contains reaction configurations and high-energy structures. These datasets have been utilized to construct ML potentials (see below). In addition to general datasets of molecules, datasets for specific applications are available, such as the Open Catalyst 2020 (OC20) dataset [40], with 872 000 adsorption states of saturated or unsaturated molecular fragments on a wide variety of surfaces, and the

Atom3D database [41], which has 3D structures of biomolecules, including molecules, RNA, and protein.

3. Features

Data and features determine the upper limit of ML models. Features—also commonly known as representations or descriptors—that are preprocessed from the source data are the input for the ML model. The selection of important features (called feature engineering) used to be the most time-consuming and labor-intensive work in the training of ML models. Although deep learning techniques can allow an ML model to learn how to extract features itself, they generally require a relatively large training dataset and model parameter space; thus, they have a higher computational cost and finally create an ML model with poor interpretability. In chemistry, the input features for different ML models may be different [42–44], but the molecular/crystal structure representation is a general task of feature engineering. As excellent review articles have already been published on this topic [45,46], we only briefly introduce a few related to the applications mentioned in Sections 4 and 5.

There are basically two categories of molecule descriptors—namely, 2D and 3D features. 2D features focus on the bonding pattern in molecules and neglect the spatial conformation. The features are derived from molecule graphs (with atoms as nodes and bonds as edges) or adjacency matrices (i.e., bond matrices). For example, SMILES describes a saturated molecule using a human-readable string (e.g., "CCO" for ethanol), and the International Union of Pure and Applied Chemistry (IUPAC) International Chemical Identifier (InChI) [47] represents a compound using a strictly unique but less human-readable string. Apart from strings, the topology of a molecule can also be abstracted as a vector of float numbers. The extended-connectivity fingerprint (ECFP) [48], which was developed using the Morgan algorithm, iteratively searches substructures in the molecule and encodes them to a hash value.

3D features are encoded from atomic coordinates, which can hardly be a direct input for an ML model due to the lack of permutation, translation, and rotation invariance [49]. Elegant methods have been designed to preserve the permutation, translation, and rotation invariance and sensitively distinguish among different structures in 3D. These methods are generally based on the numerical functions derived from interatomic distances and the angles among atoms, such as the minimum percent buried volume [50], atom-centered symmetry functions (ACSFs) [51], Steinhardt-type order parameters [52], and power-type structure descriptors (PTSDs) [53,54]. Other methods are based on atomic density alike functions, including but not limited to average steric occupancy (ASO) [55], smooth overlap of atomic positions (SOAP) [56], and Gaussian-type orbital based density vectors [57].

4. ML models

After features encode data into machine-readable input, the ML model transforms the input into output—that is, the predicted properties. Instead of deriving physical laws from theory, ML models build a numerical connection between easily accessible variables relating to how a dataset is generated and the concerned properties, which are often too complex to solve by theory. Broadly speaking, ML algorithms—depending on how the dataset is learned—can be divided into three main categories: supervised learning to fit labeled data, unsupervised learning to classify unlabeled data, and reinforcement learning, which utilizes a reward mechanism to guide the data learning. Among these, supervised learning is the most widely utilized in scientific research, due to its better numerical predictability for specific targets. Although there are many recipes and categories in ML, it is not difficult to implement ML in practice, thanks to many openly available software packages such as scikit-learn [58], PyTorch [59], and TensorFlow [60]. In the following, we will introduce the frequently used algorithms in supervised learning, especially those involving (deep) neural networks (NNs) developed in the past decade. Readers should refer to advanced ML books for mathematical details.

4.1. Decision trees

A decision tree can be visualized as a map of the possible consequences of a series of related choices, as shown in Fig. 1(a), with the consequences shown as terminal nodes (classes A, B, and C in Fig. 1(a)) and the choices as the nodes in branches (the attribute; e.g., $\boldsymbol{x}[2]$ in Fig. 1(a)). To train a decision tree, the dataset is recursively split by a selected attribute to maximally classify subgroups to have the same consequence [61]. This algorithm is popularly utilized for classification and prediction due to its advantages, which include being explainable, having few hyperparameters, having a low computation cost, and being suitable for relatively small datasets (e.g., 200 samples). However, the prediction may vary significantly with a tiny change in data.

To enhance the model robustness, the random forest (RF) [62] has been developed, which trains multiple trees independently and collects all results to make a final prediction by voting or averaging. Each tree is trained on a different sub-dataset randomly sampled from the source data, known as bootstrap aggregating or bagging. Through an ensemble of decision trees, an RF model achieves enhanced robustness and thus better predictability. Such models are more suitable for predicting discrete target values; thus, the typical application is to optimize experimental variables [63] by correlating synthetic conditions with the selectivity of the desired products [64,65].

4.2. Feedforward neural networks

A feedforward neural network (FFNN), also known as multilayer perception (MLP) [66], consists of multiple fully connected layers of neurons (i.e., nodes) that perform both linear and nonlinear operations. As plotted in Fig. 1(b), from the input $\boldsymbol{x}$ to the output $\boldsymbol{y}$, each fully connected layer performs a linear operation, as written in Eq. (1), where the weight $\boldsymbol{W}_{m\times n}$ and bias $\boldsymbol{b}_{m\times 1}$ are trainable parameters, and m and n are the dimensions of the output and input, respectively.

$$\boldsymbol{y}_{m\times 1} = \boldsymbol{W}_{m\times n}\boldsymbol{x}_{n\times 1} + \boldsymbol{b}_{m\times 1} \tag{1}$$

A nonlinear transformation, the activation, can be performed on the received data at each node. There are many possible activation functions, such as hyperbolic tangent, sigmoid, and rectified linear unit (ReLU). The training of an FFNN is achieved by minimizing the error between the predicted value and the true value, known as the cost function, as shown in Eq. (2).

$$\boldsymbol{W}^*, \boldsymbol{b}^* = \underset{\boldsymbol{W},\boldsymbol{b}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{y}_i - \mathrm{FFNN}(\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{x}_i)\|^2 \tag{2}$$

where $\boldsymbol{y}_i$ and $\boldsymbol{x}_i$ are the labels and features of the i-th sample in the training set. A variety of gradient-based optimization methods, such as stochastic gradient descent [67], Adam optimization [68], and limited-memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) [69], can be utilized to find the optimum parameters in an FFNN. With an increase in the number of intermediate layers (hidden layers), there are more fitting parameters, and the model could thus in principle have a higher fitting ability [1]. In an FFNN, the number of hidden layers is typically up to three, due to the gradient vanishing problem that manifests as a slow rate of improvement in training. However, with the help of residual

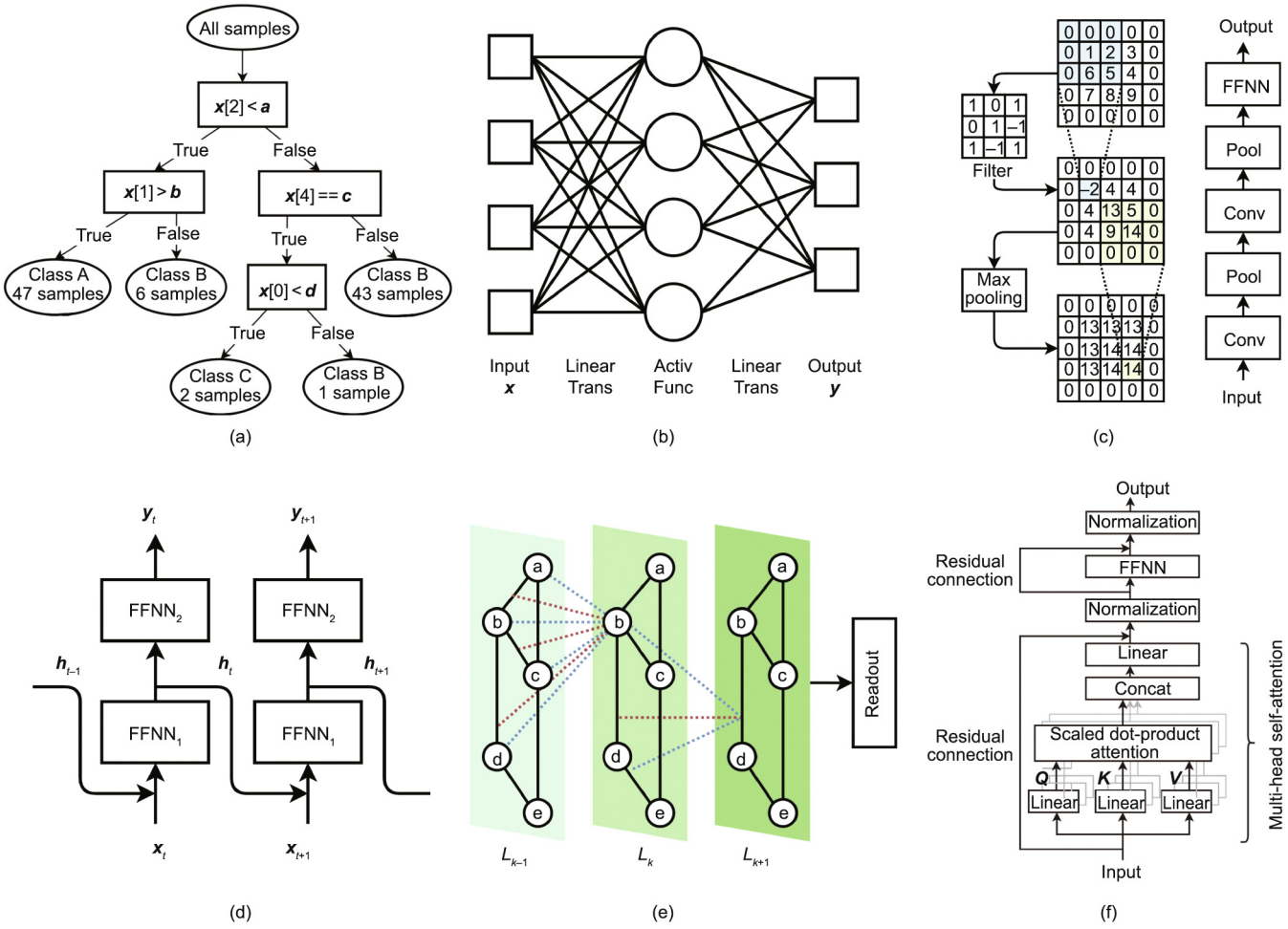

Fig. 1. Six popular machine learning models. (a) Decision tree; (b) feedforward neural network (Trans: transformation; Activ Func: activation functions); (c) convolution neural network (Conv: convolution; Pool: pooling); (d) recurrent neural network; (e) graph neural network; (f) transformer neural network.

connection [70] (i.e., skip connection), this problem can be relieved, although the fitting of a large network is computationally demanding.

4.3. Convolution neural networks

Developed upon an FFNN, a convolutional neural network (CNN) is a deep learning method that adds multiple convolution layers and pooling layers to an FFNN, as plotted in Fig. 1(c). The CNN was first introduced for image recognition with great success, and thus is particularly powerful for learning grid-like data [2]. Taking a single-channel (grayscale) image as an example (Fig. 1(c)), a convolution layer focuses on small windows of a predefined size (e.g., 3 × 3 pixels) inside the image. By performing a convolution (actually a cross-correlation) between a weight matrix, called a filter, with the small-window input data (3 × 3 matrix), and by sliding the small window over the whole image, the features of the image from the local windows are extracted to a 2D map. In practice, multiple filters are applied in a CNN to capture different features and generate multiple 2D maps. Following the convolution layer, a pooling layer further scans over the 2D map with a predefined pattern, such as a 3 × 3 window, and computes the average or maximum value in the region, with the aim of aggregating and coarsening features. In a CNN, the fitting parameters include not only those used in an FFNN but also the weights of filters in the convolution layers.

A CNN can be utilized for chemistry problems with 2D data, such as gas leak detection with infrared cameras [71]; it is also the basic unit in AlphaFold1 [4]. In practice, one-dimensional (1D) data, such as signals from chemical sensors, can also be taken as input, allowing the application of 1D CNNs for fault detection and diagnosis in chemical engineering [72–75].

4.4. Recurrent neural networks

The recurrent neural network (RNN) is another class of artificial NN that allows output from some nodes to re-feed to the same nodes as additional inputs, as shown in Fig. 1(d). This makes the RNN applicable to tasks with sequential events [76], such as speech recognition [3]. For sequential data at time t, $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ are the input and output, respectively. From $\boldsymbol{x}_t$ to $\boldsymbol{y}_t$, a simple RNN model can be expressed as follows:

$$\boldsymbol{h}_t = \phi(\boldsymbol{W}_{h\times h}\boldsymbol{h}_{t-1} + \boldsymbol{W}_{h\times n_x}\boldsymbol{x}_t + \boldsymbol{b}_{h\times 1}) \tag{3}$$

$$\boldsymbol{y}_t = \phi(\boldsymbol{W}_{n_y\times h}\boldsymbol{h}_t + \boldsymbol{b}_{h\times 1}) \tag{4}$$

where $\boldsymbol{h}_t$ is the hidden variable at time t; $\boldsymbol{W}_{h\times h}$, $\boldsymbol{W}_{h\times n_x}$, and $\boldsymbol{W}_{n_y\times h}$ are trainable weight matrices; and h, n_x, and n_y are the dimensions of the hidden variables, input, and output, respectively. Obviously, $\boldsymbol{W}_{h\times h}\boldsymbol{h}_{t-1}$ is the additional term from the previous time $t-1$, which will affect the output at time t. Without the additional term, an RNN degenerates to a standard FFNN. RNNs are particularly suited

for learning sequential-like data, such as a string of chemical names. By using the SMILES name of the reactant as input, RNNs have been utilized to predict the products of organic reactions [77] (Section 5.1).

4.5. Graph neural networks

A graph neural network (GNN) is a class of deep learning methods that can process graph data via pairwise message passing between nodes in a graph; it is also commonly known as a message passing neural network (MPNN) [78,79]. A GNN typically stacks several message passing layers, as shown in Fig. 1(e); thus, one node in the graph can communicate with other nodes that are several neighbors away. In each MPNN layer, L_k, the node N_k^{b} (i.e., node b in the k-th layer) representation is updated based on the information from the previous layer L_{k-1}, including the node itself (N_{k-1}^{b}), its first neighbor nodes (N_{k-1}^{a}, N_{k-1}^{c}, and N_{k-1}^{d}), and the edges it connects to (E_{k-1}^{ab}, E_{k-1}^{bc}, and E_{k-1}^{bd}). The edge representation can be updated with similar method. The updating strategy in MPNN can be designed quite freely, such as by using a sum of neighbor representations followed by a nonlinear activation. After the message passing layers, a readout function (e.g., an FFNN) is utilized to obtain the output based on the last message passing layer.

GNNs are of particular interest to chemists, since molecules can naturally be represented by graphs. As a class of cutting-edge but slightly underdeveloped methods, GNNs have been successfully applied to predict the properties of molecules [78] and crystals [80]. Attempts have also been made to fit the PES of materials with GNNs [38,81] (as detailed in Section 5.2).

4.6. Transformer neural networks

A transformer is a novel deep learning model that was initially designed to process sequential data (e.g., natural language processing) [82] and demonstrated great potential to replace RNN models. The key feature of transformers is a multi-head self-attention mechanism, which allows the processing of the whole input sequence all at once. As plotted in Fig. 1(f), a transformer layer can be expressed as Eq. (5).

$$\mathrm{Atten}\left(\boldsymbol{Q}_{d_{\mathrm{k}}\times d_{\mathrm{m}}}, \boldsymbol{K}_{d_{\mathrm{k}}\times d_{\mathrm{m}}}, \boldsymbol{V}_{d_{\mathrm{k}}\times d_{\mathrm{m}}}\right) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}}\right)V \tag{5}$$

$$\mathrm{softmax}(\boldsymbol{z})_i = \frac{\exp(z_i)}{\sum_{i=1}^{K}\exp(z_i)} \text{ for } i = 1, \cdots, K \tag{6}$$

This equation calculates the inner product of the query vectors $\boldsymbol{Q}$ and key vectors $\boldsymbol{K}$, which is sent to the softmax function defined in Eq. (6) to obtain a group of weights for the value $\boldsymbol{V}$ vector. Here, d_{k} and d_{m} are the dimensions of the key vector and model, respectively. The three matrices $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ are generated from the same input by a linear transformation, where the linear transformation weights $\boldsymbol{W_Q}$, $\boldsymbol{W_K}$, and $\boldsymbol{W_V}$ are parameters to learn (thus, the method is called self-attention). By using parallel multiple attention units with different sets of ($\boldsymbol{W_Q}$, $\boldsymbol{W_K}$, $\boldsymbol{W_V}$), the so-called multi-head attention, the model can jointly attend to the feature information at different positions. The output of the multi-head self-attention layer is further processed by an FFNN. Because a transformer model can be deep, with many layers, the residual connection [70] is utilized to avoid gradient vanishing; this adds the input of a certain layer (e.g., FFNN) directly with its output, and takes the sum as the input for the next layer. With its powerful feature-extraction ability due to multi-head self-attention, the transformer model has been shown to be successful for both sequential text data [83,84] and grid image data [85], thereby unifying two important application fields of ML.

Benefiting from its powerful ML framework, the transformer has had a few notable applications in recent years. For example, AlphaFold2 utilizes a variant of the transformer, the so-called Evoformer [5], to replace the residual-connected CNN in AlphaFold1 [4]. Graphormer [86], an improved transformer for graphs, showed high accuracy in predicting the relaxed energy from the unrelaxed structure in Open Catalyst Challenge 2021, outperforming classic MPNNs. Schwaller et al. [87] used a transformer to learn the atom-mapping relationship between the products and reactants of organic reactions without supervision or human labeling, thus identifying the reaction rules.

5. Applications

In the following section, we provide a few important applications of ML to illustrate how these ML techniques are used to solve chemistry problems, including retrosynthesis in organic chemistry, ML potential in computational chemistry, and heterogeneous catalysis in physical chemistry. Some related literature is summarized in Table 2 [38,56,57,63,88–106], which lists information on ML tasks, input data, features, ML models, and the prediction target.

5.1. Retrosynthesis

Synthesis planning, also known as retrosynthesis, is at the core of chemistry, answering the question of how to synthesize desired chemical compounds from available materials. Over its long history, this task has relied heavily on the knowledge of experienced chemists; thus, computer-assisted synthesis planning (CASP)—proposed by Corey et al. [107,108] as early as in the 1960s—always ranks at the top of hot topics in chemistry. Since then, many successful CASP programs have been developed, such as LHASA [109], simulation and evaluation of chemical synthesis (SECS) [110], Chematica [111], IBM RXN [112], 3N Monte-Carlo tree search (MCTS) [88], and AiZynthFinder [113] (Table 2). Since organic reactions are abundant and such databases are relatively easy to access, retrosynthesis has been actively developed through the years, particularly with the help of ML techniques in the past decade [88,111–117].

Reaction prediction and retrosynthesis are two key modules in CASP. Reaction prediction is the basis of retrosynthesis, with a focus on one-step reactions, aiming to establish a one-to-one correspondence between reactants and products under certain reaction conditions. Prediction must select the correct reaction rules (i.e., the template), which depend on both the molecular structures and the reaction conditions. Therefore, reaction prediction can be divided into two categories: the template-based method and the template-free method [89–92,118]. The former requires an *a priori* template library that can either be codified by experts using chemical informatics [108,109] or be extracted from reaction databases by the recently popular atom-mapped algorithm [93]. The template-free method generally focuses on the prediction of the reaction center in a molecule and thus identifies the bonds most suitable for (dis)connection.

In the template-based method, there are often too many likely products from one reactant, yielding overloaded candidate reactions. In 2016, Wei et al. [94] made attempts to use ML to predict template applicability. Based on a fingerprint-based NN algorithm, they predicted the most promising reaction type out of 16 basic reactions of alkyl halides and alkenes, given only the reactants and reagents as inputs. The final reactions were generated by

Table 2
A summary of the application of ML in retrosynthesis, ML potentials, and heterogeneous catalysis.

Application	Task	Input data	Feature	Model	Prediction target	Refs.
Retrosynthesis	Template-based reaction prediction	Reactant molecule	ECFP	FFNN	The most probable reaction type	[93,94]
	Template-free reaction prediction	Product molecule, reaction type	SMILES	RNN	SMILES of reactant	[89]
	Template-free reaction prediction	Reactant molecule	SMILES	RNN	SMILES of product	[90]
	Template-free reaction prediction	Reactant molecule	SMILES	Transformer	SMILES of product	[91]
	Template-free reaction prediction	Reactant molecule	Molecule graph	GNN	Reaction center and product	[92]
	Retrosynthesis	Product molecule	ECFP	FFNN	SCScore	[95]
	Retrosynthesis	Product molecule	ECFP	MCTS	Retrosynthetic route	[88]
ML potentials	ML potential	Atomic coordinates	SOAP	Gaussian process regression	DFT energy	[56]
	ML potential	Atomic coordinates	ACSF/PTSD	FFNN	DFT energy	[99]
	ML potential	Atomic coordinates	Interatomic distance	CNN	DFT energy	[96,97]
	ML potential	Atomic coordinates	Interatomic distance	GNN	DFT energy	[38,98]
	ML potential	Atomic coordinates	Gaussian-type-orbital based atomic density vector	FFNN	DFT energy	[57]
	ML potential	Atomic coordinates	ACSF	FFNN	DFT energy by atomic charge	[100]
Heterogeneous catalysis	Optimizing catalysts	Experimental data	Experiment condition	FFNN, RF	Product yield, selectivity	[101,102]
	Optimizing catalysts	Literature experimental data	Experiment condition, the characteristic results	RF	Product yield, selectivity	[63]
	Optimizing catalysts	Robot-produced experimental data	Experiment condition	Bayesian	Catalyst activity	[103]
	Predicting reactivity	Atomic coordination environments	Coordination number, element type	RF	Adsorption energy	[104]
	Predicting reactivity	Element information	Elementally, the atomic radius, number of valence electrons	RF	d–p band center	[105]
	Research reaction mechanism	Atomic coordinates	PTSD	FFNN	DFT energy	[106]

MCTS: Monte-Carlo tree search; SCScore: synthetic complexity score.

applying the SMARTS transformations to the reactants. Their model achieved an accuracy of 85% in their test reactions and 80% in selected textbook questions. Later, Segler and Waller [93] applied the approach to a more complex experimental dataset from Reaxys. As shown in Fig. 2(a) [93], each reactant fingerprint yielded a probability distribution over a library of 8720 algorithmically extracted templates, and the accuracy reached 78%. It should be mentioned that the template-based method is relatively mature in CASP, with concerns mainly including the relevance of the prediction and the scope of the template library. Rare templates generally have to be excluded in the training of the ML model.

The template-free method that has emerged in recent years holds the potential to break the limitations of the template-based method due to quality and completeness. The seq2seq model based on an RNN is the most representative template-free ML model [89–91,118]. In the seq2seq model, reaction prediction is solved as a machine translation problem between SMILES strings [29] of reactants and products and the output SMILES code of the precursors/products followed by a graphic transformation module to regenerate real chemical structures, as shown in Fig. 2(b) [89]. It is worth mentioning that the seq2seq model only outputs the SMILES sequence, so the SMILES sequence outputs sometimes cannot be converted into a reasonable structural formula, due to a misunderstanding of the grammar of the SMILES representation. In 2017, Liu et al. [89] trained a seq2seq model on 50 000 experimental reaction examples from the USPTO and were able to achieve 37.4% top-1 accuracy and 70.7% top-50 accuracy on the test dataset. More recently, Schwaller et al. [91] replaced the RNN in the seq2seq model with a transformer and achieved a top-1 accuracy of 90.4% (93.7% top-2 accuracy) on a common benchmark dataset. Similarly, a GNN can be used for template-free prediction [92,119]. A study by Jin et al. [92] using the Weisfeiler–Lehman network (WLN), a kind of MPNN, achieved 76% top-1 accuracy on the USPTO-15K dataset and 79% top-1 accuracy on the USPTO dataset.

Retrosynthesis is more complex, as its aim is to provide a global optimum synthetic pathway, which is not as simple as connecting the best one-step reactions or picking the shortest route. Traditionally, CASP programs (e.g., LHASA and SECS) suggest a few candidates, and the final choice is made by experienced chemists [107,109]. One step further, Coley et al. [95] proposed the synthetic complexity score (SCScore) as a metric for ranking molecules in retrosynthesis. As shown in Fig. 2(c) [95], an FFNN model was constructed to compute the SCScore from an ECFP [48] and was trained on over 12 million reactions from the Reaxys database. Based on the premise that, on average, the products of published chemical reactions should be more synthetically complex than their corresponding reactants, a hinge loss function was utilized in the training to encourage a separation of the SCScore between the reactant and the product. Under this scheme, a high-valued synthetic route should exhibit a monotonic increase in SCScore.

Instead of using the SCScore to evaluate the synthetic route, Segler et al. [88] developed an MCTS-based method (Fig. 2(d) [120]) to grow asymmetrically promising sub-synthetic trees, where an in-scope filter network is utilized to predict whether or not a reaction is actually feasible. The filter network takes the product and the reaction fingerprints as inputs and works as a classifier to filter out nonsensical reactions in the expansion stage of the MCTS. By combining with two other NN models (i.e., policy models) for predicting reaction patterns, the researchers showed that, in a double-blinded A/B test of nine routes to different molecules, the computer-generated reaction routes were as good as the reported literature routes on average (57% preference of MCTS and 43% of the literature, as judged by 45 organic chemists).

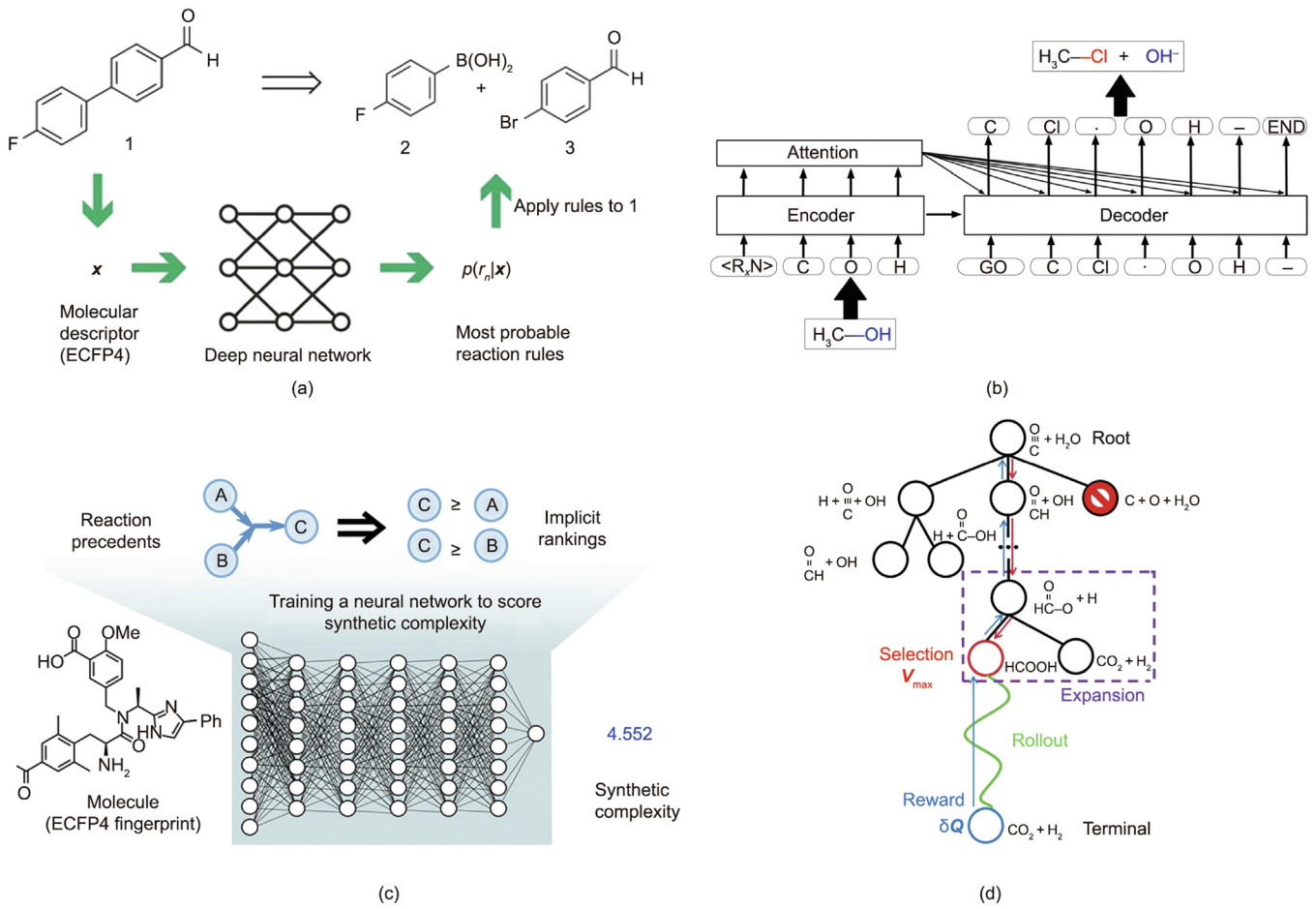

Fig. 2. (a) Overview of the neural-symbolic approach for template-based reaction prediction, which predicts possible reaction rules through reactant's ECFP4 descriptors. (b) Seq2seq model architecture for template-free reaction prediction, which translates the SMILES name of the reactant into the product. (c) Scheme of an SCScore model to guide retrosynthesis. (d) Illustration of the MCTS algorithm, which is composed of four steps: selection, expansion, rollout, and reward. (a) Reproduced from Ref. [93] with permission; (b) reproduced from Ref. [89] with permission; (c) reproduced from Ref. [95] with permission; (d) reproduced from Ref. [120] with permission.

Despite these successes, the synthesis of natural products remains a challenge. Aside from the sparsity of the training data on complex molecules, the quantitative yield of enantiomers is generally missing in most models but is important for properly evaluating a synthetic route.

5.2. ML potentials

Another important application of ML in chemistry is related to the atomic simulation of complex systems, where ML potentials [121] replace computationally demanding QM calculations for evaluating PES. Because ML potentials are trained on a dataset from QM calculations, ML potential calculations can achieve an accuracy that is comparable to that of QM, but with a speed that is several orders of magnitude faster. The ML potential method thus significantly expands the territory of atomic simulation to multi-element systems with thousands of atoms, which may only be possible to simulate traditionally by means of an empirical force field, although the availability of a force field is highly limited to systems with a relatively simple PES. Since the advent of the first ML potential in 1995 [122], many different types of ML models have been proposed, and two classes of ML architecture (Table 2)—namely, NN potentials [81,123,124] and kernel-based potentials [125–127]—are the most popular. Although kernel-based potentials, such as the Gaussian approximation potential (GAP) [128,129] and updated versions with the smooth overlap of atomic positions kernel (SOAP-GAP) [56], have much fewer hyperparameters than NN potentials, their calculation speed is restricted by the size of the training samples. Hence it is intrinsically difficult to use kernel-based potentials to go beyond big training sets, and they are more suitable for single-element systems, such as carbon and silicon [128–133]. In the following, we focus on the NN potential, which is becoming the mainstream in ML potential calculations.

Despite numerous early applications in molecular systems, the NN potential for complex systems started from the high-dimensional NN (HDNN) framework proposed by Behler and Parrinello [123] in 2007. By assuming the total energy of the structure as a sum of individual atomic energies, HDNNs establish an FFNN to correlate the local chemical environment of an atom with the atomic energy. Behler and Parrinello further invented a set of ACSFs that are invariant to the translation, rotation, and permutation of structure, as the structural descriptors for the input layer of the NN. A major virtue of the HDNN framework is its satisfaction of the extensity of the total energy, allowing different structural configurations in the dataset with variable atom numbers and chemical compositions to be treated on an equal footing.

The HDNN architecture has since been actively researched and improved, particularly regarding the structure descriptor. For example, the local atomic environment can be extracted using a CNN architecture, as implemented in Deep Potential [96,97], where the atom-centered pairwise distances are utilized as the grid data. Similarly, the MPNN [78] of a GNN can also be utilized to extract

descriptors from pairwise atomic distances, which have been implemented in deep tensor NN (DTNN) [38] for molecules and in SchNet [98] for periodic solids. The embedded atom NN potential proposed by Zhang et al. [57] utilizes a Gaussian-type orbital-based density vector as the input for the NN, which has been demonstrated to yield as good accuracy as other ML models.

The global NN (G-NN) potential method (plotted in Fig. 3) proposed by the Liu's group [39,134] realizes an automatic data generation procedure for predicting reaction systems and improves the structure descriptor and network architecture. The G-NN potential is iteratively trained upon the global PES dataset collected from SSW global PES exploration [135,136]. To better fit the global PES data, a new set of structure descriptors—namely, PTSDs [53,54]—have been developed that better describe the local chemical environment of the atom. A multi-net architecture is also implemented for the fast generation of multi-element G-NN potentials by reusing the dataset and the pre-trained NN potential in subsystems. The SSW-NN method (Fig. 3(a)) [134] is now implemented in the LASP software [39,99] and has been applied to solve many complex PES problems, such as catalyst structure determination and reaction network predictions [137–141].

To provide an example of a G-NN potential, we refer to the first Ti–O–H G-NN potential, which is constructed to describe the PES of amorphous TiO_2 structures treated under H_2 [142]. The G-NN potential adopts a large set of PTSDs that contains 201 descriptors for every element, including 77 two-body, 108 three-body, and 16 four-body descriptors, and the network involves two hidden layers (201–50–50–1 net), equivalent to approximately 38 000 network parameters in total. The final energy and force criteria of the root mean square errors (RMSEs) are around 9.8 meV per atom and 0.22 eV·Å^{-1}, respectively, for a large TiO_xH_y global dataset of 140 000 structures. Using this Ti–O–H G-NN potential, Ma et al. [142] resolved the formation mechanism of amorphous TiO_2 during hydrogenation and found a TiH hydride-mediated pathway for hydrogen production.

The local chemical environment descriptors utilized in the above ML models are generally deficient in capturing long-range interactions, such as the charge transfer in molecules. A possible solution was proposed by Ghasemi et al. [100], who used the charge equilibration neural network technique (CENT) method to learn explicit atomic charges using the same HDNN architecture. These were then utilized to compute the long-range electrostatic interactions. Ko et al. [143] recently proposed the fourth generation HDNN potential (4G-HDNNP) method for studying conjugated long-chain organic molecules and non-neutral metal and ionic clusters [143]. This method can include non-local electrostatic interactions via a special charge equilibration scheme.

5.3. ML for heterogeneous catalysis

Due to the complexity of catalyst structures and the great significance of catalysts in industry, heterogeneous catalysis has always been a major testing ground for new techniques. Early ML applications dating back to the 1990s [144,145] were generally at the phenomenological level, learning experimental data using simple ML models to optimize the catalyst synthetic and reaction conditions [101,102]. These ML applications seem to have been restricted by the availability of experimental datasets and, due to a lack of fundamental understanding, may well have overlooked key variables hidden in the experiment, leading to the failure of ML models. With the advent of deep learning and ML methods, many more exciting application scenarios have emerged, such as ML-assisted literature analysis [65,146–148] and AI robots [103] (Table 2).

ML-assisted literature analysis exploits the data mining ability of natural language processing models to abstract experimental data from the literature. Further data analysis will help to reveal the key recipes among different experiments. For example, Suvarna et al. [63] collected 1425 experimental datapoints from the literature related to CO_2 hydrogenation to methanol on Cu-, Pd-, In_2O_3-, and ZnO/ZrO_2-based catalysts. As shown in Fig. 4 [63], an RF model ($R^2 > 0.85$) was then established to correlate the methanol space–time yield with 12 descriptors relating to the experimental operation conditions, from which the top-ranking factors (e.g., the space velocity, pressure, and metal content) were identified. Experimental validation was then performed and showed a small RMSE of 0.11 $g_{MeOH}·h^{-1}·g_{cat}^{-1}$ and a high R^2 value of 0.81, demonstrating the validity of the ML model.

Chemist robots are believed to be the future of chemistry, as they will automatically perform experiments with high efficiency, while maintaining maximal data consistency between experiments [103,149,150]. For example, Burger et al. [103] developed

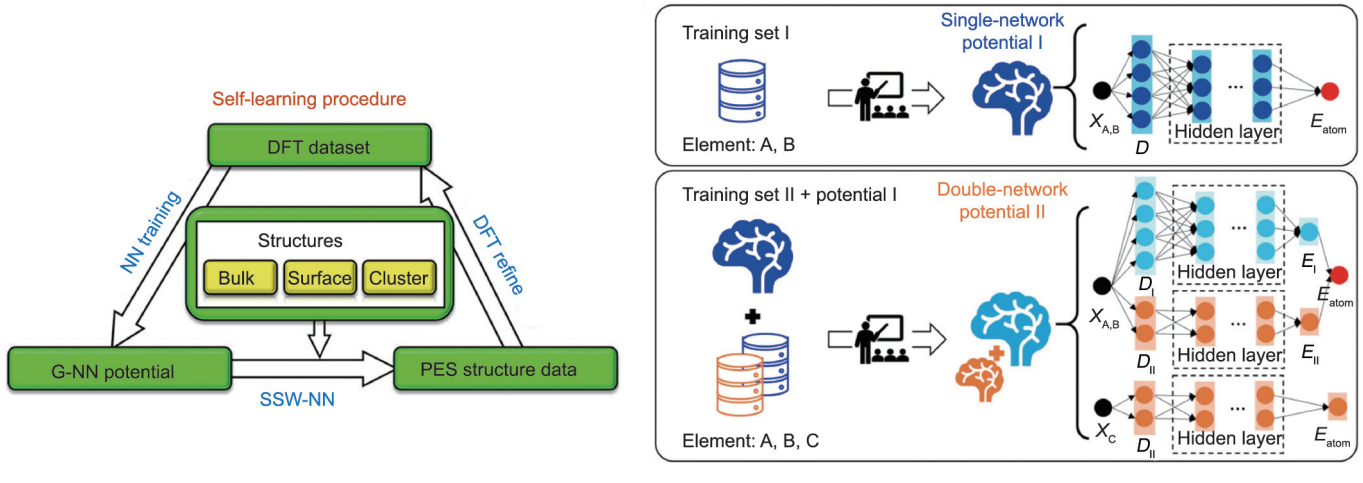

Fig. 3. (a) Scheme of the SSW-NN self-learning procedure of a G-NN potential. The G-NN is iteratively improved through cycles of SSW sampling, DFT refining, and NN training. (b) Scheme of the double-network framework implemented in LASP. With potential I trained for elements A and B, it is reused as a starting point of a subnet in potential II, whose dataset contains the elements A, B, and C. *X*: Cartesian coordinates of each atom; E_{atom}: atomic energy of each atom; D: PTSD used in NN. (a) Reproduced from Ref. [134] with permission; (b) reproduced from Ref. [39] with permission.

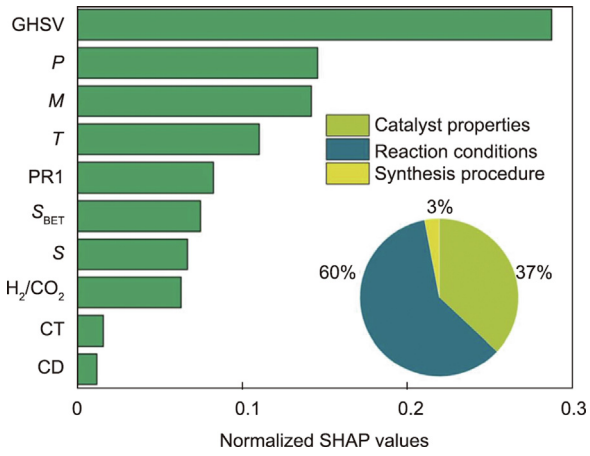

Fig. 4. Feature importance analysis for CO_2 hydrogenation to methanol. SHAP: Shapley additive explanations; GHSV: gas-hourly space velocity; *P*: pressure; *M*: metal content; *T*: temperature; PR1: promoter 1 content; S_{BET}: catalyst surface area; *S*: support content; CT: calcination temperature; CD: calcination time. Reproduced from Ref. [63] with permission.

a mobile robot to search for improved photocatalysts for hydrogen production by splitting water. In eight days, the robot performed 688 experiments within a ten-variable experimental space, guided by a batched Bayesian search algorithm (preferentially selecting beneficial components according to previous experiments). It successfully identified a catalyst synthesized from a new recipe containing P10 (5 mg), NaOH (6 mg), *L*-cysteine (200 mg), and $Na_2Si_2O_5$ (7.5 mg) in water (5 mL) that is six times more active than those using the initial formula.

From a theoretical point of view, an ML model can also be utilized to learn low-cost computable quantities, such as the adsorption energy of molecules and the electronic band structures, which are known to be important for catalysis [151,152]. Tran and Ulissi [104] used an RF-based pipeline to correlate structural fingerprints with CO and H adsorption energies on alloys based on a database containing alloys with 31 different elements. Finally, 131 candidate surfaces from 54 bulk alloys for CO_2 reduction and 258 surfaces from 102 bulk alloys for H_2 evolution were identified. From that, a CuAl alloy with near-optimal CO binding was further experimentally demonstrated to be a good catalyst for selective CO_2 reduction [153]. Sun et al. [105] recently found that the oxygen evolution reaction (OER) activity on spinel oxides is intrinsically determined by the covalency competition between tetrahedral and octahedral sites, which can be quantified using the distance between the centers of the metal d and oxygen p bands, denoted as D_M. They thus developed an RF model to predict the D_M, and a predicted $[Mn]_T[Al_{0.5}Mn_{1.5}]_OO_4$ mixed oxide was experimentally confirmed to have high OER activity, with a 240 mV (vs reversible hydrogen electrode (RHE)) overpotential at 25 $\mu A \cdot cm_{ox}^{-2}$.

On the other hand, ML atomic simulations can provide atomic-level knowledge about the catalyst structure and reaction mechanism, which benefits the rational design of catalysts. For example, Shi et al. [106] proposed a microkinetics-guided ML pathway search method (MMLPS), which can automatically explore the reaction network and determine the low-energy pathways with the help of a G-NN potential. Each branch of MMLPS independently samples different parts of the reaction PES, starting from different molecules and surface coverages. A reaction pair dataset is thus established by merging reactions from all branches, from which the lowest-barrier reaction pathway can be identified. As illustrated in Fig. 5(a) [106], a complete 2D reaction map of CO and CO_2 hydrogenation on a Cu and Zn-alloyed Cu surface is plotted using MMLPS to sample 14 958 reaction pairs. On all surfaces, CO_2 hydrogenates via the formate pathway (CO_2–HCOO*–HCOOH*–H_2COOH*–HCHO*–CH_3O*–CH_3OH*–CH_3OH) and CO hydrogenates via the formyl pathway (CO–CO*–CHO*–HCHO*–CH_3O*–CH_3OH*–CH_3OH), as shown by the free energy profile in Figs. 5(b) and (c) [106]. The overall barrier of CO_2 hydrogenation is only

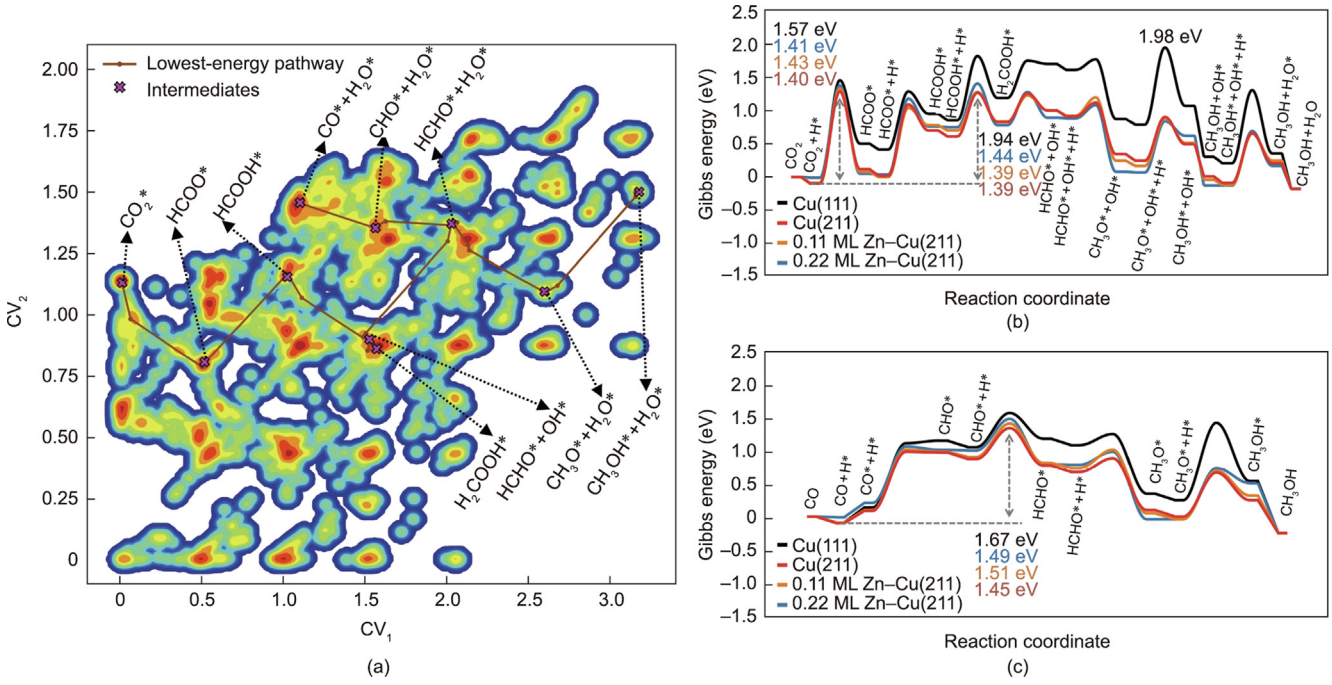

Fig. 5. (a) Contour plot for 14 958 reaction pairs obtained via MMLPS on Cu(211). The color indicates the occurrence frequency of the state in the reaction pair collection. All structures are projected onto the plot with two collective variables (CV_1 and CV_2). Key intermediates along the lowest-energy pathway are highlighted by brown lines. (b) CO_2 and (c) CO hydrogenation Gibbs energy profiles on a Cu(211) and Zn-alloyed Cu(211) surface. The "ML" here stands for monolayer. Reproduced from Ref. [106] with permission.

1.40 eV on Cu(211), while the barrier is 1.45 eV for CO, indicating that CO_2 is the main carbon source in the methanol product. A subsequent microkinetics simulation shows that Zn alloying has no significant effect on the reaction rate or even deactivates the reaction.

6. Perspective

This review summarized the key ingredients in recent ML applications for chemistry, from popular databases to common features, modern ML models, and standard application scenarios. Along with the success of recent ML applications, it must be recognized that the use of ML in chemistry presents many challenges. For example, a major obstacle is the lack of high-quality data, especially data involving experiments. Even with high-throughput experimental technology and experiment robots, there are still many fields in chemistry in which humans must produce the experimental data. In addition, chemists are often unfamiliar with state-of-the-art ML methods and related computer science techniques, which leads to difficulty in designing appropriate features for target applications. How to automatically extract features for different chemical problems remains challenging. Finally, most ML research based on FFNNs is poorly interpretable and is thus difficult to transfer to new chemistry problems.

With the fast updating of computing facilities and the development of new ML algorithms, it can be expected that more exciting ML applications are coming, and the future of chemical research will surely be reshaped in the ML era. While the future is difficult to predict, especially in such a fast-evolving field, there is no doubt that the development of ML models will lead to better accessibility, more generality, better accuracy, more intelligence, and thus higher productivity. The integration of ML models with the Internet is a good way to share ML predictions across the world. An interesting contribution was made by Yoshikawa et al. [154], who established a retrosynthetic analysis bot on Twitter that can automatically reply to retrosynthesis results if a SMILES of the target molecule is given as input. The bot utilizes the AiZynthFinder [113] package for retrosynthesis analysis.

Because of the many element types and great material complexity, the transferability of ML models in chemistry is a common problem. A prediction usually has to be restricted to the applied database, which is simply a local dataset in the vast chemistry space. The accuracy of prediction drops rapidly beyond the dataset. This issue may be solved with the advent of new techniques that can perform more efficient data collection, as shown by the G-NN potential that can learn SSW global PES data, or with better ML models that can learn more complex systems with a significant number of fitting parameters. In fact, a variety of ML competitions are held by data scientists, such as Kaggle [98], leading to the birth of many outstanding algorithms. In this regard, open ML contests on chemistry problems are still limited [40], and more efforts are needed to promote the growth of young talents in the field.

Toward more intelligent ML applications, end-to-end learning is a promising direction, as it generates the final output from raw input rather than from manually designed descriptors. AlphaFold2 [5] is a typical end-to-end learning framework that processes the 1D structure of the protein as input and finally outputs the 3D structure of the protein. This framework has provided great convenience for experimental biologists in using ML models. Similarly, in heterogeneous catalysis, an end-to-end AI model for resolving reaction pathways was recently shown by Kang et al. [120], demonstrating a bright future of solving complex problems in a single attempt by combining multiple ML models. These advanced ML models should also help in the construction of more intelligent experiment robots for performing high-throughput experiments [103,149,150].

Acknowledgments

This work received financial support from the National Key Research and Development Program of China (2018YFA0208600), the National Natural Science Foundation of China (12188101, 22033003, 91945301, 91745201, 92145302, 22122301, and 92061112), the Tencent Foundation for XPLORER PRIZE, and Fundamental Research Funds for the Central Universities (20720220011).

Compliance with ethics guidelines

Yun-Fei Shi, Zheng-Xin Yang, Sicong Ma, Pei-Lin Kang, Cheng Shang, P. Hu, and Zhi-Pan Liu declare that they have no conflict of interest or financial conflicts to disclose.

References

[1] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521(7553):436–44.
[2] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017;60(6):84–90.
[3] Li X, Wu X. Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition. In: Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing; 2015 Apr 19–24; South Brisbane, QLD, Australia. Piscataway: IEEE; 2015. p. 4520–4.
[4] Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Improved protein structure prediction using potentials from deep learning. Nature 2020;577(7792):706–10.
[5] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596(7873):583–9.
[6] Dobbelaere MR, Plehiers PP, Van de Vijver R, Stevens CV, Van Geem KM. Machine learning in chemical engineering: strengths, weaknesses, opportunities, and threats. Engineering 2021;7(9):1201–11.
[7] Venkatasubramanian V. The promise of artificial intelligence in chemical engineering: is it here, finally? AIChE J 2019;65(2):466–78.
[8] Zhou T, Song Z, Sundmacher K. Big data creates new opportunities for materials research: a review on methods and applications of machine learning for materials design. Engineering 2019;5(6):1017–26.
[9] Chen W, Iyer A, Bostanabad R. Data centric design: a new approach to design of microstructural material systems. Engineering 2022;10:89–98.
[10] Thebelt A, Wiebe J, Kronqvist J, Tsay C, Misener R. Maximizing information from chemical engineering data sets: applications to machine learning. Chem Eng Sci 2022;252:117469.
[11] Lowe DM. Extraction of chemical structures and reactions from the literature [dissertation]. Cambridge: University of Cambridge; 2012.
[12] Kearnes SM, Maser MR, Wleklinski M, Kast A, Doyle AG, Dreher SD, et al. The open reaction database. J Am Chem Soc 2021;143(45):18820–6.
[13] Akhondi SA, Klenner AG, Tyrchan C, Manchala AK, Boppana K, Lowe D, et al. Annotated chemical patent corpus: a gold standard for text mining. PLoS One 2014;9(9):e107477.
[14] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res 2019;47(D1):D1102–9.
[15] Olver FW, Lozier DW, Boisvert RF, Clark CW, editors. NIST handbook of mathematical functions hardback and CD-ROM. Cambridge: Cambridge University Press; 2010.
[16] Ayers M. ChemSpider: the free chemical database. Ref Rev 2012;26(7):45–6.
[17] Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 2012;40(D1):D1100–7.
[18] Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 2008;36(D1):D901–6.
[19] Huang R, Xia M, Nguyen DT, Zhao T, Sakamuru S, Zhao J, et al. Tox21Challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs. Front Environ Sci 2016;3:85.
[20] Delaney JS. ESOL: estimating aqueous solubility directly from molecular structure. J Chem Inf Comput Sci 2004;44(3):1000–5.
[21] Mobley DL, Guthrie JP. FreeSolv: a database of experimental and calculated hydration free energies, with input files. J Comput Aided Mol Des 2014;28(7):711–20.
[22] Wang JB, Cao DS, Zhu MF, Yun YH, Xiao N, Liang YZ. *In silico* evaluation of $logD_{7.4}$ and comparison with other prediction methods. J Chemometr 2015;29(7):389–98.
[23] Groom CR, Bruno IJ, Lightfoot MP, Ward SC. The Cambridge Structural Database. Acta Cryst B 2016;72(Pt 2):171–9.

[24] Zagorac D, Müller H, Ruehl S, Zagorac J, Rehme S. Recent developments in the Inorganic Crystal Structure Database: theoretical crystal structure data and related features. J Appl Cryst 2019;52(Pt 5):918–25.
[25] Gates-Rector S, Blanton T. The Powder Diffraction File: a quality materials characterization database. Powder Diffr 2019;34(4):352–60.
[26] Lucu M, Martinez-Laserna E, Gandiaga I, Camblong H. A critical review on self-adaptive Li-ion Battery Ageing Models. J Power Sources 2018;401:85–101.
[27] Zakutayev A, Wunder N, Schwarting M, Perkins JD, White R, Munch K, et al. An open experimental database for exploring inorganic materials. Sci Data 2018;5(1):180053.
[28] Ruddigkeit L, van Deursen R, Blum LC, Reymond JL. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. J Chem Inf Model 2012;52(11):2864–75.
[29] Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 1988;28(1):31–6.
[30] Ramakrishnan R, Dral PO, Rupp M, von Lilienfeld OA. Quantum chemistry structures and properties of 134 kilo molecules. Sci Data 2014;1(1):140022.
[31] Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, et al. Commentary: the Materials Project: a materials genome approach to accelerating materials innovation. APL Mater 2013;1(1):011002.
[32] Kirklin S, Saal JE, Meredig B, Thompson A, Doak JW, Aykol M, et al. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. npj Comput Mater 2015;1(1):15010.
[33] Curtarolo S, Setyawan W, Hart GLW, Jahnatek M, Chepulskii RV, Taylor RH, et al. AFLOW: an automatic framework for high-throughput materials discovery. Comput Mater Sci 2012;58:218–26.
[34] Calderon CE, Plata JJ, Toher C, Oses C, Levy O, Fornari M, et al. The AFLOW standard for high-throughput materials science calculations. Comput Mater Sci 2015;108:233–8.
[35] Ong SP, Richards WD, Jain A, Hautier G, Kocher M, Cholia S, et al. Python Materials Genomics (pymatgen): a robust, open-source Python library for materials analysis. Comput Mater Sci 2013;68:314–9.
[36] Smith JS, Isayev O, Roitberg AE. ANI-1, a data set of 20 million calculated off-equilibrium conformations for organic molecules. Sci Data 2017;4(1):170193.
[37] Bowman JM, Qu C, Conte R, Nandi A, Houston PL, Yu Q. The MD17 datasets from the perspective of datasets for gas-phase "small" molecule potentials. J Chem Phys 2022;156(24):240901.
[38] Schütt KT, Arbabzadah F, Chmiela S, Müller KR, Tkatchenko A. Quantum-chemical insights from deep tensor neural networks. Nat Commun 2017;8(1):13890.
[39] Kang P, Shang C, Liu Z. Recent implementations in LASP 3.0: global neural network potential with multiple elements and better long-range description. Chin. J Chem Phys 2021;34(5):583–90.
[40] Kolluru A, Shuaibi M, Palizhati A, Shoghi N, Das A, Wood B, et al. Open challenges in developing generalizable large-scale machine-learning models for catalyst discovery. ACS Catal 2022;12(14):8572–81.
[41] Townshend RJL, Vögele M, Suriana P, Derry A, Powers A, Laloudakis Y, et al. ATOM3D: tasks on molecules in three dimensions. 2022. arXiv:2012.04035.
[42] Tolman CA. Steric effects of phosphorus ligands in organometallic chemistry and homogeneous catalysis. Chem Rev 1977;77(3):313–48.
[43] Al Hasan NM, Hou H, Sarkar S, Thienhaus S, Mehta A, Ludwig A, et al. Combinatorial synthesis and high-throughput characterization of microstructure and phase transformation in Ni–Ti–Cu–V quaternary thin-film library. Engineering 2020;6(6):637–43.
[44] Plehiers PP, Symoens SH, Amghizar I, Marin GB, Stevens CV, Van Geem KM. Artificial intelligence in steam cracking modeling: a deep learning algorithm for detailed effluent prediction. Engineering 2019;5(6):1027–40.
[45] Musil F, Grisafi A, Bartók AP, Ortner C, Csányi G, Ceriotti M. Physics-inspired structural representations for molecules and materials. Chem Rev 2021;121(16):9759–815.
[46] Durand DJ, Fey N. Computational ligand descriptors for catalyst design. Chem Rev 2019;119(11):6561–94.
[47] Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D. InChI, the IUPAC International Chemical Identifier. J Cheminform 2015;7(1):23.
[48] Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model 2010;50(5):742–54.
[49] Braams BJ, Bowman JM. permutationally invariant potential energy surfaces in high dimensionality. Int Rev Phys Chem 2009;28(4):577–606.
[50] Newman-Stonebraker SH, Smith SR, Borowski JE, Peters E, Gensch T, Johnson HC, et al. Univariate classification of phosphine ligation state and reactivity in cross-coupling catalysis. Science 2021;374(6565):301–8.
[51] Behler J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. J Chem Phys 2011;134(7):074106.
[52] Steinhardt PJ, Nelson DR, Ronchetti M. Bond-orientational order in liquids and glasses. Phys Rev B 1983;28(2):784–805.
[53] Huang SD, Shang C, Kang PL, Liu ZP. Atomic structure of boron resolved using machine learning and global sampling. Chem Sci 2018;9(46):8644–55.
[54] Huang SD, Shang C, Zhang XJ, Liu ZP. Material discovery by combining stochastic surface walking global optimization with a neural network. Chem Sci 2017;8(9):6327–37.
[55] Zahrt AF, Henle JJ, Rose BT, Wang Y, Darrow WT, Denmark SE. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. Science 2019;363(6424):eaau5631.
[56] Bartók AP, Kondor R, Csányi G. On representing chemical environments. Phys Rev B 2013;87(18):184115.
[57] Zhang Y, Hu C, Jiang B. Embedded atom neural network potentials: efficient and accurate machine learning with a physically inspired representation. J Phys Chem Lett 2019;10(17):4962–7.
[58] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-Learn: machine learning in Python. J Mach Learn Res 2011;12(85):2825–30.
[59] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: an imperative style, high-performance deep learning library. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems; 2019 Dec 8–14; Vancouver, BC, Canada. Red Hook: Curran Associates Inc.; 2019. p. 8026–37.
[60] TensorFlow Developers. TensorFlow. Version 2.8.2 [software]. 2022 May 23 [cited 2022 Jun 8]. Available from: https://zenodo.org/record/6574269.
[61] Quinlan JR. Induction of decision trees. Mach Learn 1986;1(1):81–106.
[62] Ho TK. Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition; 1995 Aug 14–16; Montreal, QC, Canada. Piscataway: IEEE; 1995. p. 278–82.
[63] Suvarna M, Araújo TP, Pérez-Ramírez J. A generalized machine learning framework to predict the space-time yield of methanol from thermocatalytic CO_2 hydrogenation. Appl Catal B 2022;315:121530.
[64] Muraoka K, Sada Y, Miyazaki D, Chaikittisilp W, Okubo T. Linking synthesis and structure descriptors from a large collection of synthetic records of zeolite materials. Nat Commun 2019;10(1):4459.
[65] Baysal M, Günay ME, Yıldırım R. Decision tree analysis of past publications on catalytic steam reforming to develop heuristics for high performance: a statistical review. Int J Hydrogen Energy 2017;42(1):243–54.
[66] Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev 1958;65(6):386–408.
[67] Bottou L. Large-scale machine learning with stochastic gradient descent. In: Lechevallier Y, Saporta G, editors. Proceedings of COMPSTAT'2010; 2010 Aug 22–27; Paris, France. Heidelberg: Physica-Verlag HD; 2010. p. 177–86.
[68] Kingma DP, Ba J. Adam: a method for stochastic optimization. 2017. arXiv:1412.6980.
[69] Liu DC, Nocedal J. On the limited memory BFGS method for large scale optimization. Math Program 1989;45(1):503–28.
[70] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA. Piscataway: IEEE; 2016. p. 770–8.
[71] Wang J, Tchapmi LP, Ravikumar AP, McGuire M, Bell CS, Zimmerle D, et al. Machine vision for natural gas methane emissions detection using an infrared camera. Appl Energy 2020;257:113998.
[72] Wang N, Li H, Wu F, Zhang R, Gao F. Fault diagnosis of complex chemical processes using feature fusion of a convolutional network. Ind Eng Chem Res 2021;60(5):2232–48.
[73] Wen L, Li X, Gao L, Zhang Y. A new convolutional neural network-based data-driven fault diagnosis method. IEEE Trans Ind Electron 2018;65(7):5990–8.
[74] Xing J, Xu J. An improved convolutional neural network for recognition of incipient faults. IEEE Sens J 2022;22(16):16314–22.
[75] Ge X, Wang B, Yang X, Pan Y, Liu B, Liu B. Fault detection and diagnosis for reactive distillation based on convolutional neural network. Comput Chem Eng 2021;145:107172.
[76] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput 1997;9(8):1735–80.
[77] Bort W, Baskin II, Gimadiev T, Mukanov A, Nugmanov R, Sidorov P, et al. Discovery of novel chemical reactions by deep generative recurrent neural network. Sci Rep 2021;11(1):3178.
[78] Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural message passing for quantum chemistry. In: Precup D, Teh YW, editors. Proceedings of the 34th International Conference on Machine Learning; 2017 Aug 6–11; Sydney, NSW, Australia; 2017. p. 1263–72.
[79] Sanchez-Lengeling B, Reif E, Pearce A, Wiltschko AB. A gentle introduction to graph neural networks. Distill 2021;6(9):e33.
[80] Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Phys Rev Lett 2018;120(14):145301.
[81] Schütt KT, Sauceda HE, Kindermans PJ, Tkatchenko A, Müller KR. SchNet—a deep learning architecture for molecules and materials. J Chem Phys 2018;148(24):241722.
[82] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: von Luxburg U, Guyon I, Bengio S, Wallach H, Fergus R, editors. Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA, USA. Red Hook: Curran Associates, Inc.; 2017. p. 6000–10.
[83] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, editors. Advances in neural information processing systems 33. Red Hook: Curran Associates, Inc.; 2020. p. 1877–901.
[84] Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2019. arXiv:1810.04805.
[85] Parmar N, Vaswani A, Uszkoreit J, Kaiser L, Shazeer N, Ku A, et al. Image transformer. In: Dy J, Krause A, editors. Proceedings of the 35th International

Conference on Machine Learning; 2018 Jul 10–15; Stockholm, Sweden. Red Hook: Curran Associates, Inc.; 2018. p. 4055–64.
[86] Ying C, Cai T, Luo S, Zheng S, Ke G, He D, et al. Do transformers really perform badly for graph representation? In: Ranzato M, Beygelzimer A, Dauphin Y, Liang PS, Wortman Vaughan J, editors. Advances in neural information processing systems 34. Red Hook: Curran Associates, Inc.; 2021. p. 28877–88.
[87] Schwaller P, Hoover B, Reymond JL, Strobelt H, Laino T. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. Sci Adv 2021;7(15):eabe4166.
[88] Segler MHS, Preuss M, Waller MP. Planning chemical syntheses with deep neural networks and symbolic AI. Nature 2018;555(7698):604–10.
[89] Liu B, Ramsundar B, Kawthekar P, Shi J, Gomes J, Luu Nguyen Q, et al. Retrosynthetic reaction prediction using neural sequence-to-sequence models. ACS Cent Sci 2017;3(10):1103–13.
[90] Schwaller P, Gaudin T, Lányi D, Bekas C, Laino T. "Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. Chem Sci 2018;9(28):6091–8.
[91] Schwaller P, Laino T, Gaudin T, Bolgar P, Hunter CA, Bekas C, et al. Molecular Transformer: a model for uncertainty-calibrated chemical reaction prediction. ACS Cent Sci 2019;5(9):1572–83.
[92] Jin W, Coley C, Barzilay R, Jaakkola T. Predicting organic reaction outcomes with Weisfeiler–Lehman network. In: Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, editors. Advances in neural information processing systems 30. Red Hook: Curran Associates, Inc.; 2017. p. 2604–13.
[93] Segler MHS, Waller MP. Neural-symbolic machine learning for retrosynthesis and reaction prediction. Chemistry 2017;23(25):5966–71.
[94] Wei JN, Duvenaud D, Aspuru-Guzik A. Neural networks for the prediction of organic chemistry reactions. ACS Cent Sci 2016;2(10):725–32.
[95] Coley CW, Rogers L, Green WH, Jensen KF. SCScore: synthetic complexity learned from a reaction corpus. J Chem Inf Model 2018;58(2):252–61.
[96] Zhang L, Han J, Wang H, Car R, E W. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. Phys Rev Lett 2018;120(14):143001.
[97] Han J, Zhang L, Car R, E W. Deep Potential: a general representation of a many-body potential energy surface. Commun Comput Phys 2018;23(3):629–39.
[98] Schütt K, Kindermans PJ, Sauceda Felix HE, Chmiela S, Tkatchenko A, Müller KR. SchNet: a continuous-filter convolutional neural network for modeling quantum interactions. In: Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, editors. Advances in neural information processing systems 30. Red Hook: Curran Associates, Inc.; 2017. p. 992–1002.
[99] Huang SD, Shang C, Kang PL, Zhang XJ, Liu ZP. LASP: fast global potential energy surface exploration. WIREs Comput Mol Sci 2019;9(6):e1415.
[100] Ghasemi SA, Hofstetter A, Saha S, Goedecker S. Interatomic potentials for ionic systems with density functional accuracy based on charge densities obtained by a neural network. Phys Rev B 2015;92(4):045131.
[101] Kito S, Hattori T, Murakami Y. Estimation of catalytic performance by neural network—product distribution in oxidative dehydrogenation of ethylbenzene. Appl Catal A 1994;114(2):L173–8.
[102] Abdul Rahman MB, Chaibakhsh N, Basri M, Salleh AB, Abdul Rahman RNZR. Application of artificial neural network for yield prediction of lipase-catalyzed synthesis of dioctyl adipate. Appl Biochem Biotechnol 2009;158(3):722–35.
[103] Burger B, Maffettone PM, Gusev VV, Aitchison CM, Bai Y, Wang X, et al. A mobile robotic chemist. Nature 2020;583(7815):237–41.
[104] Tran K, Ulissi ZW. Active learning across intermetallics to guide discovery of electrocatalysts for CO_2 reduction and H_2 evolution. Nat Catal 2018;1(9):696–703.
[105] Sun Y, Liao H, Wang J, Chen B, Sun S, Ong SJH, et al. Covalency competition dominates the water oxidation structure–activity relationship on spinel oxides. Nat Catal 2020;3(7):554–63.
[106] Shi YF, Kang PL, Shang C, Liu ZP. Methanol synthesis from CO_2/CO mixture on Cu–Zn catalysts from microkinetics-guided machine learning pathway search. J Am Chem Soc 2022;144(29):13401–14.
[107] Corey EJ, Wipke WT. Computer-assisted design of complex organic syntheses: pathways for molecular synthesis can be devised with a computer and equipment for graphical communication. Science 1969;166(3902):178–92.
[108] Corey EJ, Cramer III RD, Howe WJ. Computer-assisted synthetic analysis for complex molecules. Methods and procedures for machine generation of synthetic intermediates. J Am Chem Soc 1972;94(2):440–59.
[109] Corey EJ, Long AK, Rubenstein SD. Computer-assisted analysis in organic synthesis. Science 1985;228(4698):408–18.
[110] Wipke WT, Ouchi GI, Krishnan S. Simulation and evaluation of chemical synthesis—SECS: an application of artificial intelligence techniques. Artif Intell 1978;11(1–2):173–93.
[111] Mikulak-Klucznik B, Gołębiowska P, Bayly AA, Popik O, Klucznik T, Szymkuć S, et al. Computational planning of the synthesis of complex natural products. Nature 2020;588(7836):83–8.
[112] Schwaller P, Petraglia R, Zullo V, Nair VH, Haeuselmann RA, Pisoni R, et al. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. Chem Sci 2020;11(12):3316–25.
[113] Genheden S, Thakkar A, Chadimová V, Reymond JL, Engkvist O, Bjerrum E. AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. J Cheminform 2020;12(1):70.
[114] Coley CW, Green WH, Jensen KF. Machine learning in computer-aided synthesis planning. Acc Chem Res 2018;51(5):1281–9.
[115] Wang Z, Zhang W, Liu B. Computational analysis of synthetic planning: past and future. Chin J Chem 2021;39(11):3127–43.
[116] Badowski T, Gajewska EP, Molga K, Grzybowski BA. Synergy between expert and machine-learning approaches allows for improved retrosynthetic planning. Angew Chem Int Ed Engl 2020;59(2):725–30.
[117] Jiang Y, Yu Y, Kong M, Mei Y, Yuan L, Huang Z, et al. Artificial intelligence for retrosynthesis prediction. Engineering 2023;25:32–50.
[118] Lin K, Xu Y, Pei J, Lai L. Automatic retrosynthetic route planning using template-free models. Chem Sci 2020;11(12):3355–64.
[119] Coley C, Jin W, Rogers L, Jamison TF, Jaakkola TS, Green WH, et al. A graph-convolutional neural network model for the prediction of chemical reactivity. Chem Sci 2019;10(2):370–7.
[120] Kang PL, Shi YF, Shang C, Liu ZP. Artificial intelligence pathway search to resolve catalytic glycerol hydrogenolysis selectivity. Chem Sci 2022;13(27):8148–60.
[121] Kocer E, Ko TW, Behler J. Neural network potentials: a concise overview of methods. Annu Rev Phys Chem 2022;73(1):163–86.
[122] Blank TB, Brown SD, Calhoun AW, Doren DJ. Neural network models of potential energy surfaces. J Chem Phys 1995;103(10):4129–37.
[123] Behler J, Parrinello M. Generalized neural-network representation of high-dimensional potential-energy surfaces. Phys Rev Lett 2007;98(14):146401.
[124] Lorenz S, Groß A, Scheffler M. Representing high-dimensional potential-energy surfaces for reactions at surfaces by neural networks. Chem Phys Lett 2004;395(4–6):210–5.
[125] Bartók AP, Csányi G. Gaussian approximation potentials: a brief tutorial introduction. Int J Quantum Chem 2015;115(16):1051–7.
[126] Bartók AP, Payne MC, Kondor R, Csányi G. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. Phys Rev Lett 2010;104(13):136403.
[127] Chmiela S, Sauceda HE, Poltavsky I, Müller KR, Tkatchenko A. sGDML: constructing accurate and data efficient molecular force fields using machine learning. Comput Phys Commun 2019;240:38–45.
[128] Szlachta WJ, Bartók AP, Csányi G. Accuracy and transferability of Gaussian approximation potential models for tungsten. Phys Rev B 2014;90(10):104108.
[129] Deringer VL, Csányi G. Machine learning based interatomic potential for amorphous carbon. Phys Rev B 2017;95(9):094203.
[130] Unruh D, Meidanshahi RV, Goodnick SM, Csányi G, Zimányi GT. Gaussian approximation potential for amorphous Si : H. Phys Rev Mater 2022;6(6):065603.
[131] Deringer VL, Caro MA, Csányi G. A general-purpose machine-learning force field for bulk and nanostructured phosphorus. Nat Commun 2020;11(1):5461.
[132] Bartók AP, Kermode J, Bernstein N, Csányi G. Machine learning a general-purpose interatomic potential for silicon. Phys Rev X 2018;8(4):041048.
[133] Bernstein N, Bhattarai B, Csányi G, Drabold DA, Elliott SR, Deringer VL. Quantifying chemical structure and machine-learned atomic energies in amorphous and liquid silicon. Angew Chem Int Ed Engl 2019;131(21):7131–5.
[134] Ma S, Shang C, Liu ZP. Heterogeneous catalysis from structure to activity via SSW-NN method. J Chem Phys 2019;151(5):050901.
[135] Shang C, Zhang XJ, Liu ZP. Stochastic surface walking method for crystal structure and phase transition pathway prediction. Phys Chem Chem Phys 2014;16(33):17845–56.
[136] Shang C, Liu ZP. Stochastic surface walking method for structure prediction and pathway searching. J Chem Theory Comput 2013;9(3):1838–45.
[137] Liu QY, Shang C, Liu ZP. *In situ* active site for Fe–catalyzed Fischer–Tropsch synthesis: recent progress and future challenges. J Phys Chem Lett 2022;13(15):3342–52.
[138] Liu QY, Shang C, Liu ZP. *In situ* active site for CO activation in Fe-catalyzed Fischer–Tropsch synthesis from machine learning. J Am Chem Soc 2021;143(29):11109–20.
[139] Li XT, Chen L, Shang C, Liu ZP. *In situ* surface structures of PdAg catalyst and their influence on acetylene semihydrogenation revealed by machine learning and experiment. J Am Chem Soc 2021;143(16):6281–92.
[140] Kang PL, Shang C, Liu ZP. Large-scale atomic simulation via machine learning potentials constructed by global potential energy surface exploration. Acc Chem Res 2020;53(10):2119–29.
[141] Kang PL, Shang C, Liu ZP. Glucose to 5-hydroxymethylfurfural: origin of site-selectivity resolved by machine learning based reaction sampling. J Am Chem Soc 2019;141(51):20525–36.
[142] Ma S, Huang SD, Fang YH, Liu ZP. TiH hydride formed on amorphous black titania: unprecedented active species for photocatalytic hydrogen evolution. ACS Catal 2018;8(10):9711–21.
[143] Ko TW, Finkler JA, Goedecker S, Behler J. A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer. Nat Commun 2021;12(1):398.
[144] Sasaki M, Hamada H, Kintaichi Y, Ito T. Application of a neural network to the analysis of catalytic reactions analysis of NO decomposition over Cu/ZSM-5 zeolite. Appl Catal A 1995;132(2):261–70.
[145] Mohammed ML, Patel D, Mbeleck R, Niyogi D, Sherrington DC, Saha B. Optimisation of alkene epoxidation catalysed by polymer supported Mo(VI) complexes and application of artificial neural network

for the prediction of catalytic performances. Appl Catal A 2013;466:142–52.
[146] Günay ME, Yildirim R. Knowledge extraction from catalysis of the past: a case of selective CO oxidation over noble metal catalysts between 2000 and 2012. ChemCatChem 2013;5(6):1395–406.
[147] Günay ME, Yildirim R. Neural network analysis of selective CO oxidation over copper-based catalysts for knowledge extraction from published data in the literature. Ind Eng Chem Res 2011;50(22):12488–500.
[148] Omata K. Screening of new additives of active-carbon-supported heteropoly acid catalyst for Friedel–Crafts reaction by Gaussian process regression. Ind Eng Chem Res 2011;50(19):10948–54.
[149] Rohrbach S, Šiaučiulis M, Chisholm G, Pirvan PA, Saleeb M, Mehr SHM, et al. Digitization and validation of a chemical synthesis literature database in the ChemPU. Science 2022;377(6602):172–80.
[150] Perera D, Tucker JW, Brahmbhatt S, Helal CJ, Chong A, Farrell W, et al. A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. Science 2018;359(6374):429–34.
[151] Ulissi ZW, Tang MT, Xiao J, Liu X, Torelli DA, Karamad M, et al. Machine-learning methods enable exhaustive searches for active bimetallic facets and reveal active site motifs for CO_2 reduction. ACS Catal 2017;7(10):6600–8.
[152] Liu X, Xiao J, Peng H, Hong X, Chan K, Nørskov JK. Understanding trends in electrochemical carbon dioxide reduction rates. Nat Commun 2017;8(1):15438.
[153] Zhong M, Tran K, Min Y, Wang C, Wang Z, Dinh CT, et al. Accelerated discovery of CO_2 electrocatalysts using active machine learning. Nature 2020;581(7807):178–83.
[154] Yoshikawa N, Kubo R, Yamamoto KZ. Twitter integration of chemistry software tools. J Cheminform 2021;13(1):46.