Machine learning: Accelerating materials development for energy storage and conversion

An Chen¹ | Xu Zhang¹ | Zhen Zhou¹,²,³,⁴,⁵

¹School of Materials Science and Engineering, Nankai University, China
²Computational Centre for Molecular Science, Nankai University, China
³Institute of New Energy Material Chemistry, Nankai University, China
⁴Key Laboratory of Advanced Energy Materials Chemistry (Ministry of Education), Nankai University, China
⁵Renewable Energy Conversion and Storage Center (ReCast), Nankai University, China

Abstract
With the development of modern society, the requirement for energy has become increasingly important on a global scale. Therefore, the exploration of novel materials for renewable energy technologies is urgently needed. Traditional methods are difficult to meet the requirements for materials science due to long experimental period and high cost. Nowadays, machine learning (ML) is rising as a new research paradigm to revolutionize materials discovery.

In this review, we briefly introduce the basic procedure of ML and common algorithms in materials science, and particularly focus on latest progress in applying ML to property prediction and materials development for energy-related fields, including catalysis, batteries, solar cells, and gas capture. Moreover, contributions of ML to experiments are involved as well. We highly expect that this review could lead the way forward in the future development of ML in materials science.

KEYWORDS
big data, energy storage and conversion, machine learning, property prediction

1 | INTRODUCTION

Nowadays, many challenges¹ in the 21st century including low carbon energy and sustainability are mainly materials-related issues. Materials with specific chemical and physical properties for efficient energy storage and conversion are urgently needed to achieve sustainable development of human society.

As shown in Figure 1, for a long time, the discovery of novel materials relies on the trial-and-error process, indicating long timeline and high cost, which cannot meet the requirements for more advanced materials. Thanks to the development of theoretical and computational chemistry, quantum mechanics (QM), and molecular mechanics become mature methods to obtain quantitative structure-property relationships before experiments. With rapid progress in high-performance computations, high-throughput computational screening has extremely accelerated the investigations on materials science,²-⁴ making it possible to compute the properties of thousands of compounds. Density functional theory (DFT)⁵,⁶ has widely been used for the computation of...
structures and properties of materials,7 and accelerates the development of materials databases with calculated properties of a huge number of systems, such as Materials Project (MP) database,8 AFLOWLIB consortium,9 Open Quantum Materials Database (OQMD),10 and MaterialGo (MG).11 With state-of-the-art supercomputers and algorithms, researchers can calculate compounds with thousands of interacting ions and electrons on basis of QM methods.12,13 However, the high computational cost of QM-based methods limits the applications to large-scale complex systems. Besides, it is unrealistic to exhaust all possible systems through QM methods.

As Materials Genome Initiative (MGI)14 progresses, the era of big materials data is coming and more efforts have been made to collect materials properties and build more materials databases. The effective management and utilization of big data is the key basis to accelerate materials design. Nowadays, quickly and effectively assessing and analyzing big data to find hidden rules is challenging in current materials science.

The emergence of artificial intelligence (AI) provides a new chance to bring breakthroughs in science and engineering. The combination of AI and big data is hailed as “the fourth paradigm of science.”15 Machine learning (ML) is the core of AI, the fundament to make computers intelligent. Currently, advances in ML have been making great influence on various fields,16 as ML is a powerful tool to find the statistical law hidden behind high-dimensional data. Recently, ML has also been used increasingly in materials science17,18 with the rapid growth of materials databases, the gradual spread of ML toolkits, such as TensorFlow,19 Pytorch,20 and scikit-learn,21 the development of workflow toolkits such as Atomate,22 and the progress of algorithms. Combining with big data,23 ML techniques have successfully made many breakthroughs in the field of energy storage and conversion materials, such as catalysts16,24 and battery materials.25,26

Several early reviews have introduced the applications of ML to materials science, including materials discovery and design,27-32 catalysts,24,33 and structure prediction.34,35 Very recently, ML investigations on energy storage and conversion materials have rapidly increased, which have not been comprehensively summarized. Therefore, in this review, we would focus on various directions where ML have been successfully applied, and hope to promote further developments in the field of energy storage and conversion.

## 2 | BASIC PROCEDURE OF ML IN MATERIALS SCIENCE

Based on ML technology, computers can automatically learn from empirical data (training data) and then determine the linear or nonlinear relationships between feature factors and materials properties. Different from traditional hard-coded approaches, in which the algorithm was preprogrammed by a human expert, ML approaches could learn from the dataset, obtain the rules dominating the dataset and build a model to make predictions. Therefore, sufficient data are very important for ML. Feature engineering, including feature extraction and selection, is the critical step to extract inputs to train a ML model from data, which is the fundamental to ensure the applicability and feasibility of ML models, since only related features could be meaningful for the construction of ML models.30,36 Then the ML algorithms would be selected to build a model and learn from the data. Finally, the model would be evaluated and optimized.

### 2.1 | Data collection

For the applications of ML to materials science, the training data can be obtained from high-throughput computations or experiments. However, these data might be incomplete, inconsistent or even spurious; therefore, data cleaning which means that identifying inaccurate data and then replacing, revising or deleting them plays a key role for the accuracy of ML models.
Fortunately, databases, which contain the materials accumulated in the past century, bring great convenience for ML in materials science. Inorganic Crystal Structure Database (ICSD),37 which contains more than 210 000 crystal structures, is one of the most widely used materials databases. Crystallography Open Database (COD)38 and Pauling File39 are also widely used as data sources, with over 400 000 and 330 000 entries, respectively. There are also several computed materials databases such as MPdatabase,8 the AFLOWLIB consortium,9 OQMD,10 and MG.11 Remarkably, the band structures in MG were calculated by Heyd-Scuseria-Ernzerhof (HSE) hybrid functional,40 which could improve the accuracy of band structures. Besides, there are many materials databases for specific applications, such as MaterialsWeb online database,41 computational materials repository,42 Materials Cloud platform2 for two-dimensional (2D) materials, Harvard Clean Energy Project (CEP) for organic photovoltaics materials,43 and Thermoelectrics Design Lab for thermoelectric materials.44 Data in these databases are generally checked for technical errors; however, there are still systematic or measurement errors which should be identified and removed before training ML models.

The fast access of materials databases is essential for data collection. Therefore, most materials databases provide application programming interfaces (API), such as Materials Project RESTful API, which could make users directly access MP data and query materials information programmatically.

In the past, it was very difficult for non-ML experts to run ML programs and train the model. Nowadays, the advances in ML frameworks, such as TensorFlow (Python),19 Pytorch (Python),20 scikit-learn (Python),21 Torch (Lua),45 Caffe (Protobuf),46 and Deeplearning4J (JAVA), ensure researchers to build high quality ML models more easily. The performances of these frameworks are different in speed and accuracy and researchers can choose one as needed.47

2.2 Feature engineering

When there are enough available data, the transformation of the raw data into some quantitative parameters, which are the most influential for the modeling of targeted properties without redundancy determines the accuracy of ML models. Therefore, choosing an appropriate feature selection method is crucial to obtain a practical ML model.48

Deep understanding of underlying scientific issues and ML algorithms is the foundation to select suitable features. Typically, the features are encoded with structure and property parameters, such as electronic properties (band gap, dielectric constant, work function, electron density, electron affinity, etc), structure properties (atomic radial distribution functions,49 configuration, property-labeled materials fragments,50 Voronoi tessellations,51 etc) and magnetic properties. Reasonable selection of features is difficult and expensive.52 In previous investigations, researchers selected different dimensions and types of features to build different ML models and adopted the model with the best performance.

As high-performance computations and deep learning are developed, automated feature engineering is more widely used. Compared with manual feature engineering, automated feature engineering is more efficient and repeatable, and allows researchers to build a better ML model faster. For deep learning, the function in each layer can transform the input data into another mathematical expression as input data for the latter layer. Deep learning provides an approach for computers to automatically obtain features learned from data and incorporate them into the process of model building, which can reduce the incompleteness of manual feature engineering. Nowadays, deep learning which could handle thousands of features without the need of feature selection is widely applied to various fields such as drug53 and nanomaterials discovery.30

However, the datasets are generally small so that it might be difficult to automatically extract features by deep learning frequently. Moreover, some property features, such as Goldschmidt tolerance factor, which has been widely used to represent the stability of perovskites, is not ideal for accurate prediction due to inherent inaccuracy. Bartel et al54 reported a novel stability tolerance factor $\tau$ for perovskites, and the performance of this new tolerance factor maintains consistent higher accuracy across a wide range of perovskites. Moreover, not limited to perovskites, $\tau$ can also estimate the stability of perovskite-like structures. Ekin et al55 demonstrated that the standard ML approach cannot establish suitable models from small data. To solve the problem of limited data, they created a transfer learning approach which combines structural models and elemental models. The final elemental-structural ML model performs acceptable results with relatively small datasets, compared with the structural model which is only trained by experimental and DFT computational data. Notably, the number of features also influences the establishment of models; the lack of feature quantity cannot describe materials comprehensively, and the redundant feature number will make the establishment of the model more complicated.56

2.3 ML methods

The selection of an appropriate ML model also plays a key role since it significantly affects the predicted performance.57 There is no single best method for all cases. ML
methods can be divided into supervised, semisupervised and unsupervised learning relying on the amount of features (input data) and corresponding labels (output data) of training data. For supervised learning, input data correspond to output ones. By using supervised model, the computer can find the relationship between input and output and predict the output value when giving a specific input value. In semisupervised learning, the amount of input data is larger than that of output ones. The ratio of unlabeled input data and labeled input data is always high. The quality of models mainly relates with the automatic training of unlabeled data. For unsupervised learning, the labels of training data are unknown. Unsupervised learning can be used to reveal the intrinsic regularity of data. The widest application of unsupervised learning is classification, such as K-means algorithm.

Currently, supervised learning is the most widely used effective tool among those methods in materials science. Therefore, we will focus on supervised learning models below.

### 2.3.1 k-Nearest neighbors

k-Nearest neighbors (kNN) algorithm is one of the most theoretically mature and simplest ML methods. The basic principle of this method is that a sample is identified by the majority of its nearest k neighbors in feature space. As shown in Figure 2, for \( k = 1 \), the yellow sample would be classified as a member of green class. For \( k = 3 \), the sample belongs to red class, since red triangles are more. For \( k = 5 \), the sample is identified as a member of green class. kNN could be used in both classification and regression. The distance between samples and training data in feature space is the basis for classification. In n-dimensional real vector space, Euclidean distance is always used in more general cases, and Minkowski distance could also be used. Once the sample positions in the feature space are available, the distance can be calculated and no explicit training phase is needed. In other words, the generalization of training data is delayed until a query is made to the system; therefore, kNN is a lazy learning method. Because of this, the prediction of kNN is time-consuming and memory footprint is large if the training dataset is large. Besides, imbalance of training data could also influence the performance of kNN.

There is no fixed rule on the selection of \( k \). A small value of \( k \) would always be selected according to the distribution of samples and then cross-validation is used to optimize the value of \( k \).

### 2.3.2 Naïve Bayes

Naïve Bayes classifier is a series of classification algorithms based on Bayes’ theorem and the hypothesis that the features are conditionally independent of each other, for example, in Figure 3, the color of feature represented by triangles is not related to feature represented by squares. The sample would be classified into the class with the highest estimated probability. This method is often used to predict biological properties. However, in reality it is difficult to satisfy the assumption that the features are conditionally independent, since the features are always interrelated. Figure 3 represents the main mathematical principle and a simple example of Naïve Bayes classifier. The people are classified into two categories \( y_1 \) and \( y_2 \). Each person has four main features, and the value of each feature is represented by different colors.

\[
f(x) = \arg \max_{y_k} P(y_k) \prod_{i=1}^{n} P(x_i | y_k)
\]

**FIGURE 2** Diagram of k-nearest neighbors algorithm

**FIGURE 3** Mathematical function and a simple example of Naïve Bayes classifier. The blue and green humanoid figures represent two types of samples, and the gray one is an unassorted sample. Four different geometric figures represent the main features of each sample, and the color of geometric figures is the optional value of features
Decision tree (DT)\cite{62} is a ML predictive model which contains nodes and directed edges. The nodes include internal nodes and leaf nodes. The internal nodes represent distinguishing conditions for features while the leaf nodes indicate different classes as shown in Figure 4. However, there are also several limitations on DT, such as sometimes nonrobust of the tree. The biggest weakness is that this method might create over-complex trees, and cause overfitting. In order to avoid overfitting and reduce complexity of trees, pruning is usually used which uses statistical methods to delete unreliable branches to improve the speed and ability of the classification of new data.

In 2001, Breiman\cite{63} proposed Random Forest (RF) technique for classification or regression based on the combination of multiple DTs into “forest.” Each tree in the forest is built on the basis of recursive partitioning. When a new instance enters, each DT would make judgment. The instance would be classified by the majority vote for classification or the average for each DT will be calculated for regression. RF can be used to effectively handle datasets with great quantity of features and reduce overfitting.

Kernel methods are a collection of pattern recognition algorithms. The most widely used kernel methods include support vector machine (SVM),\cite{64} Gaussian process (GP),\cite{65} and kernel ridge regression (KRR).\cite{66} Kernel function is the inner product of mapping relationship, which is used in these methods to transform input data into a higher-dimensional space, reduce computing complexity, and even make impossible calculations possible as shown in Figure 5.

Artificial neural network (ANN),\cite{67} also called neural network (NN), is a mathematical model for ML and pattern recognition based on the principles of neural networks in biology. With understanding and then abstracting the structure of brain and response mechanism, the model simulates the mechanisms for processing complex data in nervous systems on basis of network topology. As shown in Figure 6, a network contains an input layer, an output layer and n hidden layers (n ≥ 1). Each node includes a specific output function, named activation function. The connection between two neurons carries a weight, which is modified during the training phase and then will be evaluated with test datasets. ANN methods exhibit strong ability to capture nonlinear complex relations from large-scale datasets. However, there are still some limitations; ANNs usually require much more training data and are time-consuming. It is difficult to understand why an ANN makes a particular decision, also called “black boxes.” Besides, ANNs are susceptible to overfitting, and this method should be carefully designed.

A good ML model should be predictive. It could not only fit known data, but also generalize unknown data. It is necessary to validate the ML model. In order to evaluate the model, the total dataset is generally divided into two parts: training set and test set. The training set is used to train the model and its output data are known to the model while the test set is applied to evaluate the model and the corresponding output data are not given to the algorithm. Besides parameters of ML models, which are obtained by learning from training set, many ML models possess hyperparameters which can be selected manually, such as the k value in kNN and the number of trees in RF. Models only constructed on basis of training data might have a set of overfitting hyperparameters. Therefore, it is helpful and necessary to take part of training set as the validation set to optimize the hyperparameters for the best prediction. Notably, the test set should be close to the population distribution; therefore, it is necessary to take the test set at random from the whole population.

Cross-validation\cite{68} is a common and effective approach to evaluate ML models. K-fold cross-validation
is a widely used cross-validation method, in which the data are distributed into K separate folds with one fold as the initial test set and the others as the initial training set. Then the process is circulated until each fold serves as the test set once as shown in Figure 7. The process indicates that each sample will be predicted by the model built without the entry of its corresponding output value. Therefore, if the cross-validation error is low, the model could effectively generalize all samples in the whole dataset. A special case is that the number of samples equals to K and the method is called leave-one-out cross-validation, which is used when the data size is very small. Bootstrapping method, sampling with replacement method, is also effective for small datasets. However, bootstrapping method changes the distribution of the dataset, and then potentially introduces estimation bias. Besides, Monte Carlo cross-validation which is an asymptotically consistent method for model selection and exhibits larger probability than leave-one-out cross-validation for selecting models with good prediction ability.

For classification problems, a confusion matrix with the correct and incorrect predictions on diagonal and off-diagonal is introduced, respectively. The model performance can be evaluated by classification accuracy, which is calculated by the sum of diagonal elements divided by the sum of off-diagonal elements. Receiver operating characteristic (ROC) curve, and area under ROC curve (AUC) are also applied to evaluate the accuracy of classifiers. ROC curve can well analyze the classification performance of the classifiers of samples with uneven distribution. The value of AUC implies the ability of models. ROC curve is often used in combination with precision recall (PR) curve. Meanwhile, the mean absolute percent error (MAPE), root mean squared error (RMSE), average absolute error (AAE), correlation coefficient (R²), and cross-validated counterpart (Q²) are
widely used to evaluate the prediction accuracy for regression models.

3 | ACHIEVEMENTS OF ML IN ENERGY STORAGE AND CONVERSION MATERIALS

ML is increasingly used in materials science and has proven effective. Via ML, the properties can be accurately predicted and novel materials with specific functions could be designed. The gap between materials science and computer science has gradually narrowed. In this section, we would introduce the recent advances in applications of ML to the development of materials for energy storage and conversion.

3.1 | Promotion of theoretical chemistry

3.1.1 | Property prediction

Due to the superiority, ML methods have been applied to property prediction for energy storage and conversion materials to overcome the shortcomings of DFT computations, such as high consumption of computational resource. The utilization of ML methods provides effective and novel tools for the domain of materials science. Various ML methods have been proposed to build models for rapid property prediction. The commonly used methods, such as kernel methods (including SVM, GP, and KRR), ANN, DTs and RFs, succeed in predicting various properties of many kinds of systems.

Predicting microscopic properties

It is generally known that the bandgap is one of the most basic but important electronic character for materials. Although traditional computations could provide comparatively accurate results of bandgaps, it is still difficult to gain accurate analysis of large amounts of complex systems. However, this problem can be effectively solved by ML methods. For example, Dong and co-workers\textsuperscript{71} trained the model to simulate different configurations of hybridized graphene and h-BN by convolutional neural networks (CNNs) and SVM for comparison. Those systems have at most \(6 \times 6\) supercells, but many possible atomic configurations still exist and are hard to exhaust all cases through traditional methods. The trained CNN models have high prediction accuracy for bandgaps of any given structure while the SVM ones show relatively low correlation between the predicted results and the DFT computed ones.

Zhou et al\textsuperscript{72} gained an ideal ML model by support vector regression (SVR) for accurately predicting bandgaps of inorganic solids. Interestingly, the establishment of this model only depends on the composition. The selected compositional features and a training set that relies entirely on experimental data bring excellent accuracy, and make the model more reliable. Their prediction model eradicated errors caused by DFT computational bandgap data. Nevertheless, the experimental data are far from enough due to the difficult preparation of high quality single crystals.

Lee et al\textsuperscript{73} used a different expression of bandgaps, quasiparticle (QP) gap, as their predicting property. The difference between QP gap and experimental band gap is subtle. The predicting framework is composed of DFT computations and ML, and demonstrates better practicality toward materials science. Such strategy relies on generous data and then brings a chance for complete utilization of computational data. The final model trained by nonlinear SVR shows low RMSE of 0.24 eV, indicating the potential of this method for further materials screening. Other electronic properties, such as density of states,\textsuperscript{49} band structures,\textsuperscript{74} and optical absorption coefficients,\textsuperscript{75} have also been reported as the target predicted results of ML.

Another significant microscopic property for crystal materials is lattice parameters. Not only can lattice parameters identify materials, but also assist the design of composites. To construct a supercell model for composites, the lattice mismatch related to lattice parameters of all the constituents, has great influence on the product.\textsuperscript{27} ML methods such as SVR, ANN, and linear regression (LR) have been used to predict lattice parameters for crystal materials. Javed et al\textsuperscript{76} combined dataset generation method and SVR model to generate a lattice parameter predicting model for orthorhombic ABO\textsubscript{3} perovskites. Moreover, the resulted model shows better predicting performance on both the training data and testing data, compared with the model learned by ANN, and the SVR model shows better operational efficiency. The average percentage of absolute difference is no more than 1% in this work.

ML methods have also been applied to predict properties of molecular materials. Many models can successfully predict thermodynamic properties such as atomization energies\textsuperscript{77-79} and formation energies\textsuperscript{80} of molecular materials. Wang\textsuperscript{81} proposed a new approach to reduce the prediction error of molecular atomization energies. Such method named stacked generalization approach is composed of multiple algorithms, and those algorithms belong to different types. In this work, the author used five different ML methods, NNs, ridge regression (RR), RF, extremely randomized trees, and gradient-boosting trees to build several stacked generalization approaches. Benefiting from the unique framework, the stacked
generalization model is endowed with advantages of all the constituent algorithms. Notably, molecular crystals are also a kind of materials that cannot be neglected. Due to the various competitive noncovalent intermolecular interactions, computational prediction is necessary for the design of molecular crystals. Musil et al. reported a novel ML framework for high-accuracy property prediction of polymorphs. With Gaussian Process Regression (GPR) built on the SOAP (Smooth Over of Atomic Positions)-REMatch kernel, the model can be applied to predict relative energetics of crystal materials and the transfer integrals that are calculated to predict charge mobility of molecular crystals, and the promise of high accuracy can be satisfied which is demonstrated by cross-validated predictions. Thus, as a reference tool to design molecular crystals, the model can make the design more reliable and economic.

**Predicting macroscopic properties**

In addition to the microscopic properties of crystal materials and molecular structures, ML methods also play the important role in the macroscopic property prediction, such as mechanical properties and other physical functions. For efficient prediction of the thermodynamic stability of cubic perovskites, Schmidt and coworker applied RR, RF, extremely randomized trees and NNs to accelerate their calculations. They associated the predicting accuracy with the information of the periodic table. Another study proposed by Evans et al. discussed the use of gradient boosting regression (GBR) for forecasting the mechanical properties involving bulk (K) and shear (G) modulus. According to the structural characteristics of zeolites, only structural information is selected as the feature. Comparable RMSE was gained with the values of ~0.102 and ~0.0847 for the two property presentations log (K) and log (G). Kim et al. developed a ML model to improve the unsatisfactory accurate predictions for dielectric breakdown strength of complicated systems. Based on their previous work, the model was updated by large database of perovskite materials. Through careful preliminary screening that includes energy band selection, structural reoptimization and dynamic stability determination, 209 structures were screened to compose the dataset. KRR, random forest regression (RFR) and least absolute shrinkage and selection operator methods, were chosen to be trained by a training set of 82 octet crystalline insulator perovskites. By using the model trained with the least absolute shrinkage and selection operator method, boron-containing perovskites were selected due to their excellent performances in high electric environment, and the breakdown fields of two of these materials are ~2 GV/m, which means good feasibility for experiments. Considering the blank of the utilization of ML to energetic property prediction, Elton et al. further broadened the application scope of ML methods in the prediction of detonation pressure, explosive energy and other energetic properties of molecular structures. Even in the case of a small dataset, errors of the results calculated through KRR model are within an acceptable range. Another meaningful and successful example of ML is structure prediction and classification. Pilania and coworkers reported a new crystal structure classification method to fit the relative small dataset. Moreover, the accuracy of the model was effectively improved by a novel material property, which is called excess Born effective charge. Another ML classifier reported by Musil can help researchers to understand the packing and the self-assembly mechanism of molecular materials. Such automatic classification tool shows superiority compared with the heuristic classifications.

**Improvement in property prediction**

The accuracy of the features of the input data, a suitable training set with sufficient data, ideal learning models, and suitable features are necessary for successful ML frameworks of property prediction. There are also many reports concentrated on the steps of design and optimization for predicting ML models. In order to gain the best model to predict properties, researchers always try diverse feature collocations and ML methods. Many property features have been explored during prior time, and the most commonly used features are chemical compositions, structures, Coulomb matrix, chemical environment descriptions, and other complicated features. In order to select the appropriate features, representing compounds intuitively and maintaining the most concise expression on the premise that the basic calculation needs are essential. In the work of Seko et al., they developed a method to generate a feature set which has strong applicability. The procedure based only on compositions and structures of materials, and the strong applicability of this method is given by the great applicability of those features that can represent different element compositions and crystal structures in an identical dimensional form. Through the regression methods (KRR and GP) and the Bayesian optimization, the applications of the approach for three disparate datasets show good performance. Pham et al. reported an advanced representation, “orbital field matrix (OFM),” which shows an excellent fit for data mining. Such novel feature representation is inspired by those previous studies about features generated by the effect between the center atom and neighboring atoms, and it combines the information of the local structure and the valence orbitals of the center atom. OFM focuses on the local environment and
valence atomic configuration that represent the valence orbital structure, and it can be used in predicting the properties of both periodic crystal materials and molecules with high accuracy.

A good set of features needs to be used in conjunction with appropriate data. The size and type of datasets are both important for the construction of ML models and require careful selection. It is well known that exceptional datasets play a crucial role in the success of ML models. Although this standpoint may be one-sided, efforts to optimize datasets are necessary for training models. Generally, in order to gain a universal training result, the random and independent data are favored. However, such purposeless choice may guide the final model to the wrong way. Browning and coworkers \(^97\) gave efforts to optimize the training sets of ML models for molecular property prediction. Genetic algorithms (GA) were applied to create the optimal training dataset from multiple parent and children population iterations. Combining the use of Coulomb matrix representation, models trained by GA-optimized datasets show better performance than those based on randomly selected ones. Moreover, they pointed out that the size of training set can influence the accuracy, relative mean absolute errors (RMAE), and the speed of optimization. For different ML models, each model corresponds to distinct optimal distribution of datasets. Hence, the selection of training datasets is also a multifaceted problem.

The final factor that must be considered for improving property prediction frameworks is the ML model. Researchers always use several different types of ML algorithms for the same target predicting properties, given features and datasets, and compare their performance from a comprehensive perspective. Faber et al \(^98\) not only investigated the influence of the representation choice, but also assessed various regression models, such as Bayesian ridge regression (BR), LR, RF, KRR, and NNs, by out-of-sample errors. They illustrated the selection of ML models from the point of view of matching, which means using the minimum error regression or NNs model for different molecular properties. For example, the best prediction performance to the highest vibrational frequency is gained from RF model while the models of graph convolutions (GC), gated graph networks (GG), and KRR demonstrate high-quality prediction for all the properties mentioned in their study. However, there is no model that can satisfy all the predicted properties in an acceptable precision degree.

From the above discussion, it is obvious that numerous ML methods have been integrated with the investigation on materials properties predicted in various ways, and those technologies were renovated with some new characteristics. A motley variety of properties control abundant applications of materials and contribute to new materials design \(^99\). Hence, the utilization of ML methods plays an important role in the field of materials science, especially energy storage and conversion materials. In order to enlighten the future studies and accelerate the development of energy storage and conversion materials, we will summarize successful cases of ML applications to energy storage and conversion materials in the following sections.

### 3.1.2 Exploring energy storage and conversion materials

#### Catalysts

Since the 1990s, ML tools, especially ANNs, have been used in catalysis. These studies usually focus on the relationship between the catalytic performance and reaction conditions based on experimental data \(^100,101\). Then the synthetic conditions and corresponding compositions of the catalysts were often used as features of ML models to guide the synthesis of catalysts with better performance \(^102,103\). The experimental data need high-throughput experimentation which is time-consuming, high cost and limited, and then make ML models non-general. Compared with experiments, QM methods can be used to obtain larger databases. Recently, researchers have integrated ML and QM methods to overcome the limitations of pure QM methods, accelerating the accurate screening of catalysts \(^24,33\).

For example, QM methods are computationally expensive and therefore limit the applications to large-scale complex systems. Using ML to develop interatomic potentials which are trained by the data obtained from QM methods is an effective method to improve numerical efficiency \(^104\). ML potentials could accelerate computations by orders of magnitude with comparable accuracy to QM methods \(^105\). There are several codes for ML potentials such as AMP \(^106\), AENet \(^107,108\), PROPhet \(^74\), and TensorMol \(^109\). Besides, there are other ML schemes to predict atomic potentials which are generally based on high-dimensional neural networks (HDNNP), GPR, or KRR \(^110-114\).

The need for accurate dynamics of chemical reactions has been growing rapidly in recent years. However, ab initio molecular dynamics (AIMD) simulations are limited to hundreds of atoms and ~10 ps time scale, restricting its applicability. The low time cost of ML potentials can extend applications to more and larger systems at longer time scale compared with QM methods. Shakouri et al \(^115\) developed a high-dimensional neural network potential (HDNNP) for N\(_2\) on Ru(0001) surface from a dataset of 25 000 structures calculated with RPBE
functional. The method can accurately describe the coupling of N₂ and surface atom motion and vibrational properties of the slab model of Ru(0001). Owing to the low reaction probability, a great quantity of trajectories would be required. The HDNNP makes it possible to perform quasi-classical trajectory (QCT) calculations, simulating both molecular and surface atom motion for the accurate computation of reactions with sticking probability as small as 10⁻⁵. Importantly, the sticking probabilities computed by HDNNP are consistent with experimental results.

For complicated practical catalytic systems, many factors influence catalytic performances. For example, the solvent environment plays a major role of catalysis. Combining first-principles computations and Monte Carlo simulations with NN potentials, Artrith et al.¹¹⁶,¹¹⁷ investigated the equilibrium structure and composition of Au/Cu nanoparticles containing several thousands of atoms in water for CO₂ reduction. NN potentials were also applied to investigate the structural and dynamical properties of interfacial water at Cu surface.¹¹⁸

Besides, the structure, size and composition of catalyst nanoparticles or surface are extremely difficult to be considered with QM methods owing to the size of simulations. ML potentials can tackle this problem. Ouyang et al.¹¹⁹ used NN potentials to search for global minima of Au nanoclusters with the basin-hopping method. Sun et al.¹²⁰ employed HDNNP to explore the Pt₁₃ structure under a pressure of hydrogen and the results indicate that low energy metastable structures play a major role of catalytic performance. The method can be used to systematically investigate the impact of isomers, and take reaction conditions into consideration. Moreover, NN potentials were applied to search for global minima of Pt₁₃ and Pt₉ clusters under realistic temperatures.¹²¹ Besides clusters, NN potentials were used for multicomponent alloy surfaces and the predicted mean surface compositions for AuPd alloys showed good agreement with reported experimental results.¹²² Other alloys, such as NbMoTaW, were investigated by ML potentials in combination with Monte Carlo simulations.¹²³ For oxides, Jacobsen et al.¹²⁴ investigated the surface reconstruction of SnO₂(110)-(4 × 1) based on ML potentials. The results indicate that the ML potential energies are meaningful and can greatly speed up the search when the evolutionary search runs. It is an important foundation to reasonably model the catalyst under reaction conditions. These results reveal the value of ML potentials in catalysis.

However, one of the drawbacks of ML potentials is that the combined space of chemical species and atomic coordinates will increase rapidly with the number of chemical species and cause the complexity of ML potentials. To resolve this issue, Artrith et al.¹²⁵ proposed a descriptor with constant complexity. Moreover, the huge demand for training data, usually thousands of conformational geometries, would take a great deal of time. Chmiela et al.¹²⁶ reported a gradient-domain ML approach, which adopts exclusively atomic gradient information instead of atomic energies. The method can reproduce global potential energy surfaces for intermediate sized molecules with a high accuracy by using only 1000 geometries for training. The gradient-domain ML model is three orders of magnitude faster than DFT, enabling long-time scale path integral molecular dynamics simulations.

In addition, ML techniques can be used to predict catalytic properties and screen catalysts with good performance. A representative example is the prediction of d-band center, an extensively used catalytic descriptor for metals and alloys, with ML techniques.¹²⁷ Besides, empirical relationships such as Brønsted-Evans-Polanyi (BEP) relations which relate the activation barrier and enthalpy of reactions¹²⁸ and the scaling relationships for adsorption energies¹²⁹ could simplify computations. Therefore, the adsorption energy is also an important descriptor to evaluate the catalytic performance. Ma et al.¹³⁰ proposed a ML augmented chemisorption model which could describe the adsorbate-substrate interactions with ANNs of ~0.1 eV error and identified promising <100>-terminated Cu multimetallics for electroreduction of CO₂ to C₂ species with low overpotential and high selectivity. Using CO adsorption as a metric, Li et al.¹³¹ presented a ML chemisorption model for rapid screening of transition metal catalysts for electrochemical CO₂ reduction. CO binding energy was adopted as a descriptor to screen active facets since materials with weaker CO binding have lower barriers.¹³² This model exhibited excellent performance for screening <100>-terminated multi-metallic alloys and predicted several promising candidates with low overpotential. Then they further reported a ML framework to rapidly screen bimetallic catalysts for methanol electrooxidation from over 1000 alloy surfaces.¹³³ The ML model was trained with ~1000 model alloys and could describe the adsorbate/metal interactions with a RMSE of ~0.2 eV.

Combining NN models and DFT computations, Ulissi et al.¹³⁴ screened Ni₅Gaₙ bimetallic surfaces for the electrochemical reduction of CO₂. It is difficult to simulate Ni₅Gaₙ surfaces and adsorption configurations through pure DFT computations because several compositions will form and are stable at reducing potentials while each structure has dozens of exposed facets with hundreds of unique adsorption sites. The application of NN models can speed up the calculations by an order of magnitude. The results indicate that NiGa(210), NiGa(110) and
Ni$_2$Ga$_x$(021) would exhibit promising catalytic performance for the electrochemical reduction of CO$_2$ to C1 and C2 products.

Modeling the arrangement of surface atoms with well-defined single crystal surfaces has achieved success in predicting catalytic performance. However, for highly inhomogeneous atomic configurations including nanoparticles with atomic scale defects, this method will be limited. To resolve this problem, Jinnouchi et al$^{135}$ proposed a model based on Bayesian linear regression and BEP relations to predict the catalytic activity of NO decomposition on Rh$_{1-x}$Au$_x$ alloy nanoparticles. A kernel called SOAP$^{136}$ was used in Bayesian linear regression to evaluate the similarity between two local atomic configurations on basis of overlap integrals between three-dimensional atomic distributions. The model could predict energetics of catalytic reactions on alloy nanoparticles with DFT data of single crystals and can give detailed information about the structures of active sites and size- and composition-dependent catalytic performance. Recently, Jäger et al$^{137}$ have evaluated the performance of SOAP, Many-Body Tensor Representation (MBTR),$^{138}$ and Atom-Centered Symmetry Functions (ACSF)$^{139}$ for the prediction of hydrogen adsorption free energy on the surface of nanoclusters to estimate hydrogen evolution reaction (HER) catalytic performance. The results indicate that SOAP performed significantly better and is good choice for adsorption energy predictions. Moreover, the surface phase diagrams involving adsorbates and coverages under varying applied potentials are essential for electrocatalysis. By means of a GPR model, Ulissi et al$^{140}$ predicted the free energy of possible adsorbate coverages for IrO$_2$(110) surface. By using this model, the Pourbaix diagram is reconstructed for the IrO$_2$(110) surface with the adsorption of H, OH, and O.

These investigations indicate that ML models combined with QM methods make it possible to exhaustively screen large catalyst spaces with lower time consumption compared with pure QM calculations. Recently, Ahneman et al$^{141}$ have used high-throughput experimental data instead of QM data to train a RF algorithm. The ML model was successfully applied to predict the specific palladium catalysts with high tolerance to isoxazole during C-N cross-coupling. The results proved the value of ML in complex molecule synthesis.

The intermediates and rate-limiting transition states (TS) are also very important for the design of catalysts. The stable reactants, products and intermediates are in the local or global minima of the potential energy surface (PES) while the TS locates on the first-order saddle points of PES. There are several commonly used computational algorithms to search TS such as climbing-image nudge elastic band (CINEB),$^{142}$ dimer method,$^{143}$ single-ended growing string method$^{144}$ and force reversed method.$^{145}$

The activation barriers can be calculated once the TS is obtained. However, searching for TS is extremely time-consuming due to the complexity of PES. ML technology can be used to accelerate TS search and prediction of activation barriers. Peterson$^{146}$ reported a NN model trained by DFT to reduce the number of intermediates that need ab initio calculations to locate saddle points, therefore greatly accelerating search for saddle points. Besides NN, Koistinen et al$^{147}$ used GPR to speed up NEB calculations to find minimum energy path. This model can reduce the number of necessary energy and force evaluations to a half and has been evaluated in a benchmark involving 13 rearrangement transitions of a heptamer island on a solid surface.

Takahashi et al$^{148}$ revealed the descriptors for determining the activation energy based on 788 activation energies constructed from first-principles computations with ML models. The activation energy can be instantly predicted with high accuracy during cross validation. Choi et al$^{149}$ used thermodynamic and structural properties as input features to establish ML models for the prediction of activation energies of gas phase reactions. The results indicate that the tree boosting algorithm exhibited excellent performance with low errors. However, the reactions involving the change of the number of chemical bonds greater than 4 have large errors due to the complicated reaction mechanisms.

ML technologies could also be used to study reaction mechanisms by reducing the complexity of reaction networks. For a simple reaction network, such as the reaction of syngas over Rh(111) would possess hundreds of reactions and over 2000 potential pathways as shown in Figure 8.$^{151}$ QM methods can be used to investigate the reaction mechanisms, but it is impractical owing to the expensive computations. Ulissi et al$^{151}$ proposed a ML framework based on GPR and DFT computations to accurately and fast predict the reaction networks of syngas over Rh(111) under experimentally relevant thermal conditions (Figure 8). From a few DFT computations of intermediates, the GPR scheme was then used to predict the free energy of the intermediates. The enthalpies of TS were predicted with linear scaling relations$^{152}$ to estimate the activation energy and a simple classifier was applied to determine the rate-limiting step. Then CINEB was used to evaluate the TS energy of the most likely reaction mechanism. Finally, the most likely reaction networks to specific products can be identified.

In addition to heterogeneous catalysts, the ML technology can be used to investigate homogeneous catalysts. An example is that Burello et al$^{153}$ applied NN models to search for Heck cross-coupling catalysts in which a set of steric and electronic descriptors was defined. The models
were used to predict the catalytic performance of 60,000 combinations of virtual catalysts and reaction conditions with high accuracy.

Lithium ion batteries
Currently, Li-ion batteries (LIBs) are commercially successful energy storage devices due to high operation voltage, large energy capacity, long cycle life, and low self-discharge.\textsuperscript{150,154} The cathode, electrolyte and anode are the main components of LIBs. Li is oxidized to Li\textsuperscript{+} in anodes and moves to cathodes through electrolytes. The electrolytes should be both ionic conductors and electronic insulators. For liquid electrolytes, solvation to and desolvation from Li\textsuperscript{+} at the electrolyte/electrode interface play an important role of Li\textsuperscript{+} transport. Therefore, the coordination energy of the solvent to Li\textsuperscript{+} is a key parameter. Besides, the melting point of the electrolyte is extremely important for fast transport under low temperature. Sodeyama et al\textsuperscript{155} predicted the coordination energy and melting point of liquid electrolytes by using three information techniques including exhaustive search with linear regression (ES-LiR), multiple linear regression (MLR) and least absolute shrinkage and selection operator. The results indicate that ES-LiR has the most accurate estimation.

Over the past years, various electrolyte additives have been reported and widely applied to improve the battery performances such as enhancing the ionic conductivity of electrolytes, reducing the irreversible capacity and gas generation, improving the thermal stability of electrolytes and protecting cathode materials from dissolution and overcharging.\textsuperscript{156} Therefore, the development of electrolyte additives is also important. The redox potential is a key measure to evaluate whether the materials can be used as electrolyte additives. Okamoto et al\textsuperscript{157} built regression models from the redox potentials of 149 representative molecules calculated by ab initio molecular orbital calculations. The constituent elements and coordination numbers were used to construct features and GBR exhibited good performance in the prediction of redox potentials.

The traditional liquid organic electrolytes suffer from potential safety problems due to the flammability and volatility. The design of novel solid electrolytes is important since they are generally less flammable and therefore safer than liquid electrolytes. However, insufficient ionic conductivity limits the application of solid electrolytes and the exploration of solid electrolytes with high ionic conductivity has gained increasing attention. Jalem et al\textsuperscript{158} proposed a NN model to predict the Li diffusion

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{image.png}
\caption{(A) Reaction network for the reaction of syngas (CO + H\textsubscript{2}) to CO\textsubscript{2}. (B) The reduced network for syngas reactivity on Rh(111). The green, gray, red and white balls represent Rh, C, O and H atoms, respectively. Reprinted from Reference 150}
\end{figure}
barrier and cohesive energy of olivine-type LiMxOy (M: main group elements, X: group XIV and XV) for solid electrolytes. Then they further used NN models to investigate the structure and Li ion transport of LiM(T célavor system (M**:T** + and M**:T**- pairs, M: non-transition metals) with chemical substitutions at M and T sites for solid electrolytes.159 Several potential solid electrolyte candidates were predicted and the Li migration energy of LiMgSeO4F is only 0.11 eV. Recently, they have proposed a GP model by combining Bayesian sampling to accelerate the exploration of compounds with low ion migration energies.160

Fujimura et al25 used ML technique learning from theoretical and experimental datasets to predict the conductivity of LISICON-type materials at 373 K. Then the transport properties of garnet structured oxides were evaluated by using SVR, revealing the composition-structure-ionic conductivity relationship.161 Understanding Li diffusion mechanism is essential and Chen et al162 reported a density-based clustering method to compute the trajectories by using molecular dynamics for Li2La2Zr2O12 (LLZO). This is a kind of unsupervised learning method which can recognize lattice sites, give the site type and identify Li hopping events. The results indicate that the low vacancy concentration limits Li diffusion in LLZO and substituting cations with higher valence can increase the vacancy concentration, consistent with experimental observations.

High-throughput screening methods have been used to explore ideal solid electrolytes. For small training data, Ekin et al55 revealed a kind of transfer learning methods to screen potential solid lithium-ion conductors. In order to reduce the consumption of DFT calculations, Sokseih and the coworkers163 proposed a new predictor related with high ionic conductivity that can be used in high-throughput screening. They used this novel approach to predict new Li-ion conductors successfully. Combining ML and high-throughput screening based on materials databases is efficient to explore novel materials. Sendek et al26 proposed a large-scale computational screening framework to identify promising candidates for solid state electrolytes for LIBs from MP database. With this framework, 12,831 Li-containing candidate materials were screened. The structural and chemical stability, ionic and electronic conductivity and cost were chosen as screening criteria. Logistic regression was used to build a data-driven ionic conductivity classification model to evaluate the Li conduction based on experimental measurements reported in literature. After screening MP database, 21 best candidates were obtained with robust stability, low electronic conductivity, high ionic conductivity and low cost. For the further exploration of solid state electrolytes, they applied DFT molecular dynamics (DFT-MD) simulations to validate the most potential materials from nominee structures that were screened in previous work. Moreover, they compared the efficacy between the ML-guided search, random selection and artificial work, and the results show obvious superiority of ML-based model.164

In addition to electrolytes, the electrode materials are crucial in LIBs. The crystal system has a major effect on the physical and chemical properties of Li silicate cathodes. Shandiz et al165 predicted three types of crystal systems including monoclinic, orthorhombic and triclinic, of cathode materials with Li−Si−(Mn,Fe,Co)−O compositions. RF and extremely randomized trees exhibited high accuracy. Eremin et al166 employed ridge regression to predict the energy of LiNiO2 (LNO) and LiNi0.8Co0.15Al0.05O2 (NCA) cathode materials. The ML results indicate that the topology of Li layers and relative disposition of Li and dopants in NCA are the most important descriptors in energy balance estimations. Moreover, ML methods have been also applied to predict the potential of electrode materials for other metal-ion batteries.167

Organic electrode materials, which only contain light and earth abundant elements, can be used as an alternative to traditional inorganic materials owing to their low cost and high energy density.168 Allam et al169 combined DFT and ML to establish a high-throughput screening method to predict redox potentials of carbon-based molecular electrode materials. Both electronic properties and structural information were chosen as input variables for ML models. The predicted redox potentials are in good agreement with DFT results, indicating the high accuracy.

Accurately predicting the lifetime of LIBs is important for accelerating the development of LIBs. However, it has remained a challenge due to the complex aging mechanisms and working conditions. By using ML approaches, Severson et al170 developed a data-driven model, which can accurately predict the lifetime of commercial lithium iron phosphate/graphite cells based on early cycle data without prior knowledge of degradation mechanisms. A dataset of 124 commercial cells cycled under fast-charging conditions with cycle life from 150 to 2300 cycles was built. The best models achieve a low test error of 9.1% for predicting cycle life with the first 100 cycles and 4.9% with the first 5 cycles for classifying cycle life into low- and high-lifetime groups. This work provides a promising approach to understand and develop LIBs.

Also, ML methods contribute to the development of LIBs through the advanced data extraction and collection technologies. The first book written by machine was published by Springer in 2019, which helps researchers to quickly and easily understand the current frontiers in the
Solar cells
Photovoltaic solar energy conversion is considered as one of the most promising approaches to resolve the global energy crisis and environment pollution. Perovskites have been attracting wide attention for solar cells owing to their high solar absorption, the ease of fabrication and low non-radiative carrier recombination rates. However, two obvious challenges limit the large-scale commercial applications. One is the toxicity due to the element of Pb and the other is poor environmental stability. Therefore, it is important to search for stable and environmentally friendly perovskites with high power conversion efficiency (PCE). The conversion efficiency relies on multiple factors but the bandgap is widely used as a screening criterion. It is extremely time-consuming to accurately compute bandgaps with QM methods, which is impractical for high-throughput investigations. ML approaches can be a promising alternative. Pilania et al. demonstrated a ML framework to efficiently and accurately predict bandgaps of double perovskites. More than 1.2 million features were evaluated and the lowest occupied Kohn-Sham levels and elemental electronegativity of the constituent atomic species were identified as the most important predictors. Then they proposed a multifidelity co-kriging statistical ML model to predict the bandgaps of double perovskites. The bandgaps were calculated with both the low fidelity (Perdew-Burke-Ernzerhoff, PBE) and high fidelity (HSE). The HSE bandgaps were predicted by considering different numbers of PBE bandgaps in a combined dataset. The results indicate that the prediction accuracy increases as the number of HSE and PBE bandgaps increases in the training set.

High-throughput screening is effective to accelerate the development of perovskites. Allam et al. reported a high-throughput screening method to search for ABX₃ inorganic 2D perovskites. NN model was used to evaluate the importance of parameters with respect to the bandgaps. The results suggest that the oxidation state of anion, the number of layer of perovskite and ionic radii of cations are the most important factors that influence the bandgap. Takahashi et al. built a RF model to predict the bandgaps of perovskites based on 18 physical descriptors. The model was used to predict the undiscovered perovskites with ideal bandgap range for solar light absorption and 9328 perovskites were obtained. Then DFT computations were performed to evaluate the stability. Finally, 10 thermodynamically stable undiscovered perovskites were revealed with ideal bandgaps. Lu et al. developed a high-throughput framework to screen stable Pb-free hybrid organic-inorganic perovskites (HOIPs) with high PCE and sustainable air stability as shown in Figure 9. To achieve this, they built a GBR-based ML model with 212 reported bandgap values as training data to predict the bandgaps of 5158 unexplored. Then the stability was also evaluated with AIMD. Finally, six orthorhombic Pb-free HOIPs were obtained with ideal bandgaps for solar cells and excellent room temperature thermal stability.

Besides bandgaps, the stability of perovskites can be computed by using ML technology. Generally, the stability of materials can be evaluated by the energy above the convex hull ($E_{\text{hull}}$). Li et al. established a ML model to predict $E_{\text{hull}}$ of perovskite oxides learning from over 1900 DFT computed perovskite oxides. Then the model was used to predict 15 novel perovskite compounds. Schmidt et al. applied several ML algorithms to predict the $E_{\text{hull}}$ of cubic perovskites ABX₃, including ridge regression, LASSO, and other methods.

**FIGURE 9** High-throughput screening framework for HOIPs
Reprinted from Reference 179
RF, extremely randomized trees and NN. The results indicate that extremely randomized trees exhibited the lowest mean absolute error (MAE) for 230,000 perovskite structures and 500 thermodynamically stable unreported structures were obtained. Besides, deep neural networks based on only two descriptors, ionic radii and Pauling electronegativity can be used to predict formation energies of ABO₃ perovskites and garnets with low MAEs. Recently, ML technology has also been used to identify and classify perovskites for solar cells, further speeding up the development of perovskite-based solar cells.

In addition to perovskites, organic solar cells have been extensively investigated owing to their easy fabrication, low cost, light weight, and large area. The donor/acceptor bulk-heterojunction structure has been considered as one of the most effective strategies to achieve high solar cell performance due to the good exciton dissociation and charge carrier transport in active layers. Therefore, the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO) of donors and acceptors are extremely important. Pereira et al trained ML models with over 88,000 molecules to estimate the HOMO and LUMO energies with accuracies of 0.15 and 0.16 eV, respectively. Then Jørgensen et al reported a screening procedure for promising donor-acceptor polymers based on a grammar variational autoencoder model. The model was trained with a dataset of 3989 monomers obtained from DFT computations and could predict LUMO and the lowest optical transition energy for unknown molecules without knowledge of the atomic positions with MAEs of 43 and 74 meV, respectively. Besides, the deep tensor neural network model was also applied and exhibited higher predictive accuracy but the atomic positions calculated by DFT were needed as inputs.

The PCE is an extremely important parameter for solar cells. Troisi and coworkers used various ML approaches, including RF, GBR and NN to predict PCE by using 13 microscopic properties of organic materials as features. Among these models, GBR exhibited an impressive performance with a Person’s coefficient of 0.79. Recently, they used kNN and KRR models to predict photovoltaic efficiency of organic solar cells. When considering both electronic and structural parameters as the input, the KRR model can give good predictive capability with a Person’s coefficient of ~0.7. By using 1000 experimental parameters including molecular weight, PCE and electronic properties as training data, Nagasawa et al built RF and NN models to predict the PCE for organic solar cells. The results indicate that the RF model exhibited better performance and therefore was further used for the design, synthesis and characterization of a conjugated polymer.

The current acceptor materials are generally fullerene derivatives, suffering from high cost and difficult functionalization. To overcome these shortcomings, Lopez et al reported an automated workflow to explore the conformational space of each candidate with molecular mechanics. The QM technology was used to calculate the electronic structure of the molecule and GP was adopted to calibrate the computed HOMO and LUMO to experimentally determined values. The workflow was used to screen over 51,000 molecules to identify potential non-fullerene acceptor materials.

Metal oxides are another choice for solar cells if their PCEs can be improved. Yosipof et al developed a data mining and ML workflow to analyze two solar cell libraries based on titanium and copper oxides. This workflow can effectively highlight the descriptors that most affect the photovoltaic properties and exhibit good predictive ability.

**CO₂ capture**

Anthropogenic CO₂ emission is the key factor for global climate change. The development of materials for CO₂ capture and sequestration from the atmosphere is one of the grand challenges of the 21st century. Metal-organic frameworks (MOFs) with large pore volumes, ultra-high surface areas, and tunable porosities can provide abundant adsorption sites for CO₂ capture. The combination of an extraordinary variety of metal ions or clusters and organic ligands is desirable and can lead to countless combinations. Thus, it is not feasible to evaluate each MOF through QM computations or experiments.

Woo and coworkers made efforts in the field of CO₂ capture based on ML. They reported large-scale quantitative structure-property relationship (QSPR) analysis of gas adsorption of MOFs. The gas adsorption capacity, including methane, N₂ and CO₂, has been investigated. Atomic property weighted radial distribution function (AP-RDF) was introduced to capture the chemical and geometric features. Nonlinear regression models (such as SVM) can give good prediction of the absolute CO₂ capture values under different pressures. QSPR was trained by the database of 32,450 MOFs and then used to screen 292,050 MOFs. This model can capture 945 of the 1000 top MOFs, indicating its accuracy. Very recently, they have further used QSPR models to predict CO₂ working capacity and CO₂/H₂ selectivity in MOFs for precombustion carbon capture. Similarly, Borboudakis et al used Just Add Data to produce ML models to predict the CO₂ and H₂ adsorption properties of MOFs with experimental data of 100 MOFs as the training set. The results indicate that the chemical properties of MOFs are predictable through ML methods with increasing prediction accuracy with the sample size. In
another work, Anderson et al.\textsuperscript{197} trained six ML algorithms, including SVM, DT, RF, NN, GBR and MLR, to investigate the role of pore chemical and topological features in the improvement of CO\textsubscript{2} capture of MOFs. The predictive results suggest that functionalization with hydroxyl, thiol, amino, cyano, or nitro would improve CO\textsubscript{2} capture.

### 3.2 Success in experimental explorations

One of the core problems that has plagued experimenters for many years is unpredictable chemical reaction routes. The quantity of reaction pathways is so large that researchers need to set special conditions to reduce the difficulty of analysis. However, though the compounds can be prepared successfully, the principles are still not understood clearly. For controllable synthesis, computational methods were utilized in assisting experimental processes as early as 1969.\textsuperscript{198} Different from the backward computer technology in the past, the modern computing ability is strong enough to undertake the goal of “big data” learning. Thus, novel programs assisting experiments are more credible.

The success of ML methods for organic materials evokes the research enthusiasm of other materials. Although the early attempts were initiated by organic chemists, computational simulations demonstrate great applied potential for inorganic, or hybrid structures. In particular, there are a large number of nonorganic compounds in the family of energy storage and conversion materials. The urgency of energy shortage and environmental problems makes the fruits of laboratory need to be put into practical application as soon as possible. Computer presimulation accelerates the realization of this expectation. However, generally, larger systems are the most similar systems to practical objects, so it is difficult to maintain the balance between the cost and effectiveness when using the general computational methods such as ab initio methods. Fortunately, the combination of ML methods and materials science has been helping to realize the goal.

For materials design in terms of considering experimental realization, the research intention of the correlative algorithms is different from the ML methods for property prediction. The starting point of the model design should incline to assist experiments. Perovskites, a kind of most potential energy conversion materials, have permeated solar cells, catalysts, batteries and other energy fields.\textsuperscript{174} Balachandran et al.\textsuperscript{199} highlighted the key factor, stability of materials, is critical for constructing meaningful ML and active learning methods for experimental assist. Notably, they also used both the failed and successful experimental data to train their ML models. Moreover, through the approach of classification algorithm combined with regression methods, the model showed superiority because of constrain of classified demarcation. Thus, the candidate structures screened by the process are guaranteed in the possibility of experimental synthesis and application. In addition, the successful results of such complicated kinds of materials, perovskite solid solutions, show the power of ML methods. To another kind of novel and important energy materials, two-dimensional (2D) materials, the successful experimental preparation is also a barrier toward wide applications. Frey and coworkers\textsuperscript{200} used positive and unlabeled (PU) ML method to solve this problem. Considering the large chemical search space and many successfully synthesized examples, the materials with high representativeness at present, 2D transition metal carbides, carbonitrides and nitrides, MXenes, were explored by the trained models. The inputs of simple elemental information are classified by principal component analysis (PCA), isolating the most correlative representations. Based on the data of experiments, the models were verified by k-fold cross-validation, and this kind of validation method ensures the reliability. Other validated methods were also applied to remedy the shortcoming of small positive sample size. 111 MAX structures and 18 MXenes are screened with the possibility of experimental preparation.

Materials characterization is an indispensable part in experiments. Computational methods have been used to simulate the figures of various characterization technologies, such as scanning electron microscope (SEM), solid-state nuclear magnetic resonance (NMR), Infrared spectra (IR),\textsuperscript{201} X-ray absorption near-edge spectroscopy (XANES), and X-ray absorption fine structure spectra (EXAFS).\textsuperscript{202} The utilization of those novel technologies makes materials characterization enter a new era. Moreover, compared with traditional computational methods, the ML methods can provide cheap and highly accurate simulation processes. Paruzzo et al.\textsuperscript{203} reported a ML workflow for simulating the results of chemical shifts in solids which are always gained by NMR. ML models are based on computational data from Cambridge Structural Database (CSD), and the models have acceptable accuracy even though no data of experimental shifts are used for the training step. After greatly shortening calculation time, large molecular crystals with more than 1000 atoms in a unit cell were calculated. Hu et al.\textsuperscript{204} used RF method to assist the signal prediction of surface-enhanced Raman spectroscopy (SERS), which also remedied the drawbacks of ab initio methods in spectroscopy simulations.
The rapid development of science and technology leads to the explosive growth of data, taking MGI as an example, which provides opportunities for further breakthroughs in ML. Especially, combining with computations or experiments, ML technologies have made significant achievements in the development of materials for energy storage and conversion. A major application of ML in this field is to reveal the relationships between structure, property and performance, further guiding the discovery and design of novel materials.

Moreover, ML potentials allow the simulations of larger systems at longer time scale with higher accuracy compared with QM methods. The successful fruits of experiment-assisted ML methods not only encourage the developments of this domain, but such incomplete achievements also challenge researchers. Especially, although ML is data-based science and the quality of models unusually depends on the size of the database, the establishment of ML models trained by relative small size databases is urgent for materials science due to the lack of experimental databases. However, there are still many challenges to face.

Generally, ML requires very large amount of data for learning to guarantee the accuracy. However, in materials science, the data size is usually limited to hundreds or even dozens. With the progress of MGI, several databases have been built but more published data were not embodied by databases and even more “failed data” that can be used to train ML models were not reported. In future, researchers could report the data in a computer readable form to further share data. Another solution to the problem is allowing computers to process and understand human languages. Natural language processing, a branch of AI, is good choice. The text mining technology has been widely used in chemical and materials science.

Besides, machines thirst for the ability of one shot learning, learning a class from a handful or even a single labeled example, which originates in human beings and can solve the problem of limited datasets. The key is learning to learn, also called meta learning, which has been used in image recognition. Several approaches were developed for meta learning and it is worth introducing to materials science for limited datasets. Moreover, the quality of training data is extremely important. If the materials data are collected from different publications, more or less noise or bias would be brought.

The success of ML models heavily relies on the selection of features. Most present feature selection processes are often determined by the experience and intuition of researchers. It is a commonly used method to iterate initial feature sets until the performances become acceptable. However, some useful features might be omitted by artificial selection. Automated feature engineering could help nonexpert users to train models and significantly reduce artificial errors.

Currently, ML models are usually a “black box” which links the inputs and outputs, hindering the physical significance. Therefore, it is difficult to extract the knowledge from ML models and then summarize it into scientific laws for general cases. Up to now, the interpretability of ML models is also a key challenge for several reasons. It is extremely difficult to translate the connection weights in ML models to formulas. Furthermore, the scientific laws beyond the models might be too complex to be understood. Many efforts have been devoted to improving the interpretability of ML models. Developing more interpretable algorithms is an effective method. Taking an example, Yang et al reported an approach that can identify variables for ANNs and the black box mechanics can be significantly illuminated. Besides, it is also possible to extract the physical significance from the results. For example, Suntivich et al reported that the electrocatalytic performance of oxygen evolution reactions is related to the occupancy and then Zhou et al revealed that the physical significance of occupancy controls the electronic conductivity.

Various ML algorithms have been widely used in materials science. There is no single algorithm fitting for all issues. The selection of ML algorithms relies on the internal correlation, distribution and size of the dataset, the linearity or nonlinearity of the issue, and some other important factors. As an example, kNN is a simple and effective method for the situation in which the local data are relevant, while RF and SVM might be effective for nonlocal problems. For linear issues, linear regression might be fast and credible. Moreover, the time consumption should be considered. For example, NN algorithms need very long training time while kNN methods have fast training speed but low test speed. As mentioned above, since the data size in materials science is quite small, the time consumption of ML in materials science is of no great importance presently. However, with the development of MGI, the data size would rapidly increase and the time consumption will be more important. Therefore, the reasonable selection of ML algorithms lies at the center of ML applications. Until now, the ML investigations in materials science primarily rely on supervised learning. Besides supervised learning, semisupervised, unsupervised learning and other novel ML methods also have widespread applications in materials science. For instance, Tran et al applied active learning method to predict the electrocatalytic
performance for CO$_2$ reduction and H$_2$ evolution. Zhang et al$^{18}$ used unsupervised learning to develop a helpful method for investigations with small datasets and successfully propose potential solid electrolytes for Li batteries. Sun et al$^{217}$ applied unsupervised method to classify different ternary nitrides. More effective semisupervised and unsupervised learning algorithms would be developed and widely applied to materials science.

With the development of multidisciplinary fields, such as probability, statistics, computer and materials science, ML technology has the potential to create changes in materials science and a strong AI for the materials development will come true.

ACKNOWLEDGMENT
This work was supported by NSFC (21933006) and China Postdoctoral Science Foundation (2019M660055) in China.

CONFLICT OF INTEREST
The authors declare no conflicts of interest.

ORCID
Zhen Zhou https://orcid.org/0000-0003-3232-9903

REFERENCES
1. Royal Geographical Society, 21st Century Challenges; 2015. https://21stcenturychallenges.org/challenges/. Accessed xxxx.
2. Mounet N, Gibertini M, Schwaller P, et al. Two-dimensional materials from high-throughput computational exfoliation of experimentally known compounds. Nat Nanotechnol. 2018;13: 246-252.
3. Aykol M, Kim S, Hegde VI, et al. High-throughput computational design of cathode coatings for Li-ion batteries. Nat Commun. 2016;7:13779.
4. Zhang X, Chen A, Zhou Z. High-throughput computational screening of layered and two-dimensional materials. WIREs Comput Mol Sci. 2019;9:e1385.
5. Hohenberg P, Kohn W. Inhomogeneous electron gas. Phys Rev. 1964;136:B864-B871.
6. Kohn W, Sham LJ. Self-consistent equations including exchange and correlation effects. Phys Rev. 1965;140:A1133-A1138.
7. Lejaeghere K, Bihlmayer G, Bjorkman T, et al. Reproducibility in density functional theory calculations of solids. Science. 2016;351:aad3000.
8. Jain A, Ong SP, Hautier G, et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. APL Mater. 2013;1:011002.
9. Calderon CE, Plata JJ, Toher C, et al. The AFLOW standard for high-throughput materials science calculations. Comput Mater Sci. 2015;108:233-238.
10. Kirklin S, Saal JE, Meredith B, et al. The open quantum materials database (OQMD): assessing the accuracy of DFT formation energies. npj Comput Mater. 2015;1:15010.
11. Jie J, Weng M, Li S, et al. A new MaterialGo database and its comparison with other high-throughput electronic structure databases for their predicted energy band gaps. Sci China Technol Sci. 2019;62:1423-1430. https://doi.org/10.1007/s11431-11019-19514-11435.
12. Havu V, Blum V, Havu P, Scheffler M. Efficient O(N) integration for all-electron electronic structure calculation using numeric basis functions. J Comput Phys. 2009;228:8367-8379.
13. Arita M, Bowler DR, Miyazaki T. Stable and efficient linear scaling first-principles molecular dynamics for 10000+ atoms. J Chem Theory Comput. 2014;10:5419-5425.
14. https://www.mgi.gov/. Accessed xxxx.
15. Agrawal A, Choudhary A. Perspective: materials informatics and big data: realization of the "fourth paradigm" of science in materials science. APL Mater. 2016;4:053208.
16. Beck DAC, Carothers JM, Subramanian VR, Pfendtner J. Data science: accelerating innovation and discovery in chemical engineering. AIChE J. 2016;62:1402-1416.
17. Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nature. 2018;559:547-555.
18. Ramprasad R, Batra R, Pliania G, Mannodi-Kanakkithodi A, Kim C. Machine learning in materials informatics: recent applications and prospects. npj Comput Mater. 2017;3:54.
19. Abadi M, Barham P, Chen J, et al. Tensorflow: a system for large-scale machine learning. 12th Symposium on Operating Systems Design and Implementation. Chatham; 2016.
20. Paszke A, Gross S, Chintala S, et al. Automatic differentiation in pytorch. Proceedings of the 31st Conference on Neural Information Processing Systems. Long Beach, CA; 2017.
21. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825-2830.
22. Mathew K, Montoya JH, Faghaninia A, et al. Atomate: a high-level interface to generate, execute, and analyze computational materials science workflows. Comput Mater Sci. 2017;139:140-152.
23. Ghiringelli LM, Vybiral J, Levchenko SV, Draxl C, Scheffler M. Big data of materials science: critical role of the descriptor. Phys Rev Lett. 2015;114:105503.
24. Kitchin JR. Machine learning in catalysis. Nat Catal. 2018;1:230-232.
25. Fujimura K, Seko A, Koyama Y, et al. Accelerated materials design of lithium superionic conductors based on first-principles calculations and machine learning algorithms. Adv Energy Mater. 2013;3:980-985.
26. Sendek AD, Yang Q, Cubuk ED, et al. Holistic computational structure screening of more than 12000 candidates for solid lithium-ion conductor materials. Energy Environ Sci. 2017;10:306-320.
27. Liu Y, Zhao T, Ju W, Shi S. Materials discovery and design using machine learning. J Materiomics. 2017;3:159-177.
28. Sanchez-Lengeling B, Aspuru-Guzik A. Inverse molecular design using machine learning: generative models for matter engineering. Science. 2018;361:360-365.
29. Correa-Baena J-P, Hippalgaonkar K, van Duren J, et al. Accelerating materials development via automation, machine learning, and high-performance computing. Joule. 2018;2:1410-1420.
30. Wang M, Wang T, Cai P, Chen X. Nanomaterials discovery and design through machine learning. Small Methods. 2019;3:1900025.
31. Oliynyk AO, Mar A. Discovery of intermetallic compounds from traditional to machine-learning approaches. Acc Chem Res. 2018;51:59-68.
32. Zhu X, Yan J, Gu M, et al. Activity origin and design principles for oxygen reduction on dual-metal-site catalysts: a combined density functional theory and machine learning study. J Phys Chem Lett. 2019;10:7760-7766.
33. Goldsmith BR, Esterhuizen J, Liu J-X, Bartel CJ, Sutton C. Machine learning for heterogeneous catalyst design and discovery. AIChE J. 2018;64:2311-2323.
34. Graser J, Kauwe SK, Sparks TD. Machine learning and energy minimization approaches for crystal structure predicitions: a review and new horizons. Chem Mater. 2018;30:3601-3612.
35. Li F, Peng X, Wang Z, et al. Machine learning (ML)-assisted design and fabrication for solar cells. Energy Environ Mater. 2019;2:1-12.
36. Sun B, Fernandez M, Barnard AS. Machine learning for silver nanoparticle electron transfer property prediction. J Chem Inf Model. 2017;57:2413-2423.
37. Inorganic Crystal Structure Database (ICSD). http://www2.fiz-karlsruhe.de/icsd_web.html. Accessed February 11, 2020.
38. Grazulis S, Daskevic A, Merkys A, et al. Crystallography open database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. Nucleic Acids Res. 2012;40:D420-D427.
39. Villars P, Onodera N, Iwata S. The Linus Pauling file (LPF) and its application to materials design. J Alloy Compd. 1998;279:1-7.
40. Heyd J, Scuseria GE, Ernzerhof M. Hybrid functionals based on a screened Coulomb potential. J Chem Phys. 2003;118:8207-8215.
41. Ashton M, Paul J, Sinnott SB, Hennig RG. Topology-scaling identification of layered solids and stable exfoliated 2D materials. Phys Rev Lett. 2017;118:106101.
42. Rasmussen FA, Thygesen KS. Computational 2D materials database: electronic structure of transition-metal dichalcogenides and oxides. J Phys Chem C. 2015;119:13169-13183.
43. Hachmann J, Olivares-Amaya R, Atahan-Evrenk S, et al. The Harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid. J Phys Chem Lett. 2011;2:2241-2251.
44. Gorai P, Gao D, Ortiz B, et al. TE Design Lab: a virtual laboratory for thermoelectric material design. Comput Mater Sci. 2016;112:368-376.
45. Collobert R, Bengio S, Mariethoz J. Torch: A Modular Machine Learning Software Library. Technical report, Idiap, 2002.
46. Jia Y, Shelhamer E, Donahue J, et al. Caffe: convolutional architecture for fast feature embedding. Proceedings of the 22nd ACM International Conference on Multimedia. Orlando, FL: ACM; 2014.
47. Kovalev V, Kalinovsky A, Kovalev S. Deep learning with theano, torch, caffe, tensorflow, and deeplearning4j: Which one is the best in speed and accuracy? The 13th International Conference on Pattern Recognition and Information Processing, Minsk, Belarus; 2016.
48. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3:1157-1182.
49. Schütt KT, Glawe H, Brockherde F, et al. How to represent crystal structures for machine learning: towards fast prediction of electronic properties. Phys Rev B. 2014;89:205118.
50. Isayev O, Oses C, Toher C, et al. Universal fragment descriptors for predicting properties of inorganic crystals. Nat Commun. 2017;8:15679.
51. Ward L, Liu R, Krishna A, et al. Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. Phys Rev B. 2017;96:024104.
52. Ramakrishna S, Zhang T-Y, Lu W-C, et al. Materials informatics. J Intell Manuf. 2018;30:2307-2326.
53. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. Drug Discov Today. 2018;23:1241-1250.
54. Bartel CJ, Sutton C, Goldsmith BR, et al. New tolerance factor to predict the stability of perovskite oxides and halides. Sci Adv. 2019;5:eaaav069.
55. Cubuk ED, Sendek AD, Reed EJ. Screening billions of candidates for solid lithium-ion conductors: a transfer learning approach for small data. J Chem Phys. 2019;150:214701.
56. Lu S, Zhou Q, Ma L, Guo Y, Wang J. Rapid discovery of ferroelectric photovoltaic perovskites and material descriptors via machine learning. Small Methods. 2019;3:1900360.
57. Mitchell JB. Machine learning methods in chemoinformatics. Wires Comput Mol Sci. 2014;4:468-481.
58. Cover TM, Hart P. Nearest neighbor pattern classification. IEEE Trans Inform Theory. 1967;13:21-27.
59. Hand DJ, Yu K. Idiot’s Bayes—not so stupid after all? Int Stat Rev. 2001;69:385-398.
60. Koutsoukas A, Lowe R, Kalantarrovatamied Y, et al. In silico target predictions: defining a benchmarking data set and comparison of performance of the multiclass Naive Bayes and Parzen-Rosenblatt window. J Chem Inf Model. 2013;53:1957-1966.
61. Nigsch F, Bender A, Jenkins J, Mitchell JB. Ligand-target prediction using winnow and Naive Bayesian algorithms and the implications of overall performance statistics. J Chem Inf Model. 2008;48:2313-2325.
62. Quinlan JR. Induction of decision trees. Mach Learn. 1986;1:81-106.
63. Breiman L. Random forests. Mach Learn. 2001;45:5-32.
64. Noble WS. What is a support vector machine? Nat Biotechnol. 2006;24:1565-1567.
65. Rasmussen CE. Gaussian processes in machine learning. Summer School on Machine Learning. Berlin, Heidelberg: Springer; 2003.
66. Shawe-Taylor J, Cristianini N. Kernel Methods for Pattern Analysis. Cambridge, UK: Cambridge University Press; 2004.
67. Jain AK, Mao J, Mohiuddin K. Artificial neural networks: a tutorial. Computer. 1996;29:31-44.
68. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of 14th International Joint Conference on Artificial Intelligence, Montreal; 1995.
69. Efron B. Bootstrap methods: another look at the jackknife. Breakthroughs in Statistics. New York, NY: Springer; 1992:569-593.
70. Xu Q-S, Liang Y-Z, Du Y-P. Monte Carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration. J Chemom. 2004;18:112-120.
71. Dong Y, Wu C, Zhang C, et al. Bandgap prediction by deep learning in configurationally hybridized graphene and boron nitride. npj Comput Mater. 2019;5:26.
improved dimer method and partitioned rational function optimization method. J Chem Phys. 2005;123:224101.

144. Zimmerman PM. Single-ended transition state finding with the growing string method. J Comput Chem. 2015;36: 601-611.

145. Sun K, Zhao Y, Su H-Y, Li W-X. Force reversed method for locating transition states. Theor Chem Acc. 2012;131:1118.

146. Peterson AA. Acceleration of saddle-point searches with machine learning. J Chem Phys. 2016;145:074106.

147. Koistinen OP, Dagbjartsdottir FB, Asgeirsson V, Vehtari A, Jonsson H. Nudged elastic band calculations accelerated with Gaussian process regression. J Chem Phys. 2017;147:152720.

148. Takahashi K, Miyazato I. Rapid estimation of activation energy in heterogeneous catalytic reactions via machine learning. J Comput Chem. 2018;39:2405-2408.

149. Choi S, Kim Y, Kim JW, Kim Z, Kim WY. Feasibility of activation energy prediction of gas-phase reactions by machine learning. Chem A Eur J. 2018;24:12354-12358.

150. Zhang X, Chen A, Zhong M, et al. Metal-organic frameworks (MOFs) and MOF-derived materials for energy storage and conversion. Electrochem Energy Rev. 2018;2:29-104.

151. Ulissi ZW, Medford AJ, Bligaard T, Norskov JK. To address surface reaction network complexity using scaling relations machine learning and DFT calculations. Nat Commun. 2017;8:14621.

152. Norskov JK, Bligaard T, Rossmeisl J, Christensen CH. Towards the computational design of solid catalysts. Nat Chem. 2009;1:37.

153. Burello E, Farrusseng D, Rothenberg G. Combinatorial explosion in homogeneous catalysis: screening 60,000 cross-coupling reactions. Adv Synth Catal. 2004;346:1844-1853.

154. Hu M, Pang X, Zhou Z. Recent progress in high-voltage lithium ion batteries. J Power Sources. 2013;237:229-242.

155. Sodeyama K, Igarashi Y, Nakayama T, Tateyama Y, Okada M. Liquid electrolyte informatics using an exhaustive search with linear regression. Phys Chem Chem Phys. 2018;20:22585-22591.

156. Haregewoin AM, Wotango AS, Hwang B-J. Electrolyte additives for lithium ion battery electrodes: progress and perspectives. Energy Environ Sci. 2016;9:1955-1988.

157. Okamoto Y, Kubo Y. Ab initio calculations of the redox potentials of additives for lithium-ion batteries and their prediction through machine learning. ACS Omega. 2018;3:7868-7874.

158. Jalem R, Nakayama M, Kasuga T. An efficient rule-based screening approach for discovering fast lithium ion conductors using density functional theory and artificial neural networks. J Mater Chem A. 2014;2:720-734.

159. Jalem R, Kimura M, Nakayama M, Kasuga T. Informatics-aided density functional theory study on the Li ion transport of tavorite-type LiMTO4F (M^{3+}-T^{5+}, M^{2+}-T^{6+}). J Chem Inf Model. 2015;55:1185-1186.

160. Jalem R, Kanamori K, Takeuchi I, et al. Bayesian-driven first-principles calculations for accelerating exploration of fast ion conductors for rechargeable battery application. Sci Rep. 2018;8:5845.

161. Kirueva N, Pervov VS. Materials space of solid-state electrolytes: unraveling chemical composition-structure-ionic conductivity relationships in garnet-type metal oxides using cheminformatics virtual screening approaches. Phys Chem Chem Phys. 2017;19:20904-20918.

162. Chen C, Lu Z, Ciucci F. Data mining of molecular dynamics data reveals Li diffusion characteristics in garnet Li1-La2Zr2O12. Sci Rep. 2017;7:40769.

163. Muy S, Voss J, Schlem R, et al. High-throughput screening of solid-state Li-ion conductors using lattice-dynamics descriptors. iScience. 2019;16:270-282.

164. Sendek AD, Cubuk ED, Antoniuik ER, et al. Machine learning-assisted discovery of solid Li-ion conducting materials. Chem Mater. 2018;31:342-352.

165. Attarian Shandiz M, Gauvin R. Application of machine learning methods for the prediction of crystal system of cathode materials in lithium-ion batteries. Comput Mater Sci. 2016;117:270-278.

166. Eremin RA, Zolotarev PN, Ivanshina OY, Bobrikov IA. Li(Ni1-x,Co,Al)2O2 cathode delithiation: a combination of topological analysis, density functional theory, neutron diffraction, and machine learning techniques. J Phys Chem C. 2017;121:28293-28305.

167. Joshi RP, Eickholt J, Li L, et al. Machine learning the voltage of electrode materials in metal-ion batteries. ACS Appl Mater Interface. 2019;11:18494-18503.

168. Wu D, Xie Z, Zhou Z, Shen P, Chen Z. Designing high-voltage carbonyl-containing polycyclic aromatic hydrocarbon cathode materials for Li-ion batteries guided by Clar's theory. J Mater Chem A. 2015;3:19137-19143.

169. Allam O, Cho BW, Kim KC, Jang SS. Application of DFT-based machine learning for developing molecular electrode materials in Li-ion batteries. RSC Adv. 2018;8:39414-39420.

170. Severson KA, Attia PM, Jin N, et al. Data-driven prediction of battery cycle life before capacity degradation. Nat Energy. 2019;4:383-391.

171. Writer B. Lithium-ion Batteries: A Machine-Generated Summary of Current Research. Switzerland: Springer; 2019.

172. Correa-Baena J-P, Saliba M, Buonassisi T, et al. Promises and challenges of perovskite solar cells. Science. 2017;358:739-744.

173. Wang Z, Lin Q, Wenger B, et al. High irradiance performance of metal halide perovskites for concentrator photovoltaics. Nat Energy. 2018;3:855-861.

174. Chen A, Zhang X, Zhang Z, Yao S, Zhou Z. Band engineering of two-dimensional Ruddlesden-Popper perovskites for solar utilization: the relationship between chemical components and electronic properties. J Mater Chem A. 2019;7:11530-11536.

175. Pilania G, Gubernatis JE, Lookman T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. Comput Mater Sci. 2017;129:156-163.

176. Allam O, Holmes C, Greenberg Z, Kim KC, Jang SS. Density functional theory—machine learning approach to analyze the bandgap of elemental halide perovskites and Ruddlesden-Popper phases. ChemPhysChem. 2018;19:2559-2565.

177. Takahashi K, Takahashi I, Miyazato I, Tanaka Y. Searching for hidden perovskite materials for photovoltaic systems by combining data science and first principle calculations. ACS Photon. 2018;5:771-775.

178. Lu S, Zhou Q, Ouyang Y, et al. Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. Nat Commun. 2018;9:3405.

179. Li W, Jacobs R, Morgan D. Predicting the thermodynamic stability of perovskite oxides using machine learning models. Comput Mater Sci. 2018;150:454-463.
180. Ye W, Chen C, Wang Z, Chu IH, Ong SP. Deep neural networks for accurate predictions of crystal stability. Nat Commun. 2018;9:3800.

181. Im J, Lee S, Ko T-W, et al. Identifying Pb-free perovskites for solar cells by machine learning. npj Comput Mater. 2019;5:37.

182. Balachandran PV, Emery AA, Gubernatis JE, et al. Predictions of new ABO$_3$ perovskite compounds by combining machine learning and density functional theory. Phys Rev Mater. 2018;2:043802.

183. An Q, Zhang F, Zhang J, et al. Versatile ternary organic solar cells: a critical review. Energy Environ Sci. 2016;9:281-322.

184. Xue R, Zhang J, Li Y, Li Y. Organic solar cell materials toward commercialization. Small. 2018;14:1801793.

185. Pereira F, Xiao K, Latino DA, et al. Machine learning methods by random forest. Phys Chem Chem Phys. 2016;18:4501.

186. Padula D, Simpson JD, Troisi A. Combining electronic and structural features in machine learning models to predict organic solar cells properties. Mater Horiz. 2019;6:343-349.

187. Nagasawa S, Al-Naamani E, Saeki A. Organic solar cell materials toward commercialization. J Mater Chem C. 2017;5:11-21.

188. Jorgensen PB, Mesta M, Shil S, et al. Machine learning-based screening of complex molecules for polymer solar cells. J Chem Phys. 2018;148:241735.

189. Sahu H, Rao W, Troisi A, Ma H. Toward predicting efficiency of organic solar cells via machine learning and improved descriptors. Adv Energy Mater. 2018;8:1801032.

190. Padula D, Simpson JD, Troisi A. Combining electronic and structural features in machine learning models to predict organic solar cells properties. Mater Horiz. 2019;6:343-349.

191. Lopez SA, Sanchez-Lengeling B, de Goes SJ, Aspuru-Guzik A. Design principles and top non-fullerene acceptor candidates for organic photovoltaics. Joule. 2017;1:857-870.

192. Yosipof A, Nahum OE, Anderson AY, et al. Data mining and machine learning tools for combinatorial material science of all-oxide photovoltaic cells. Mol Inform. 2015;34:367-379.

193. Yu J, Xie LH, Li JR, et al. CO$_2$ capture and separations using MOFs: computational and experimental studies. Chem Rev. 2017;117:9674-9754.

194. Fernandez M, Woo TK, Wilmer CE, Snurr RQ. Large-scale quantitative structure-property relationship (QSPR) analysis of methane storage in metal-organic frameworks. J Phys Chem C. 2013;117:7681-7689.

195. Fernandez M, Trefiak NR, Woo TK. Atomic property weighted radial distribution functions descriptors of metal-organic frameworks for the prediction of gas uptake capacity. J Phys Chem C. 2013;117:14095-14105.

196. Dureckova H, Krykunov M, Aghaji MZ, Woo TK. Robust machine learning models for predicting high CO$_2$ working capacity and CO$_2$/H$_2$ selectivity of gas adsorption in metal organic frameworks for precombustion carbon capture. J Phys Chem C. 2019;123:4133-4139.

197. Borboudakis G, Stergiannakos T, Frysali M, et al. Chemically intuited, large-scale screening of MOFs by machine learning techniques. npj Comput Mater. 2017;3:40.

198. Anderson R, Rodgers J, Argueta E, Biong A, Gomez-Gualdrén DA. Role of pore chemistry and topology in the CO$_2$ capture capabilities of MOFs: from molecular simulation to machine learning. Chem Mater. 2018;30:6325-6337.

199. Corey EJ, Wipke WT. Computer-assisted design of complex organic syntheses. Science. 1969;166:178-192.

200. Balachandran PV, Kowalski B, Sehrliglou A, Lookman T. Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning. Nat Commun. 2018;9:1668.

201. Frey NC, Wang J, Vega Bellido GI, et al. Prediction of synthesis of 2D metal carbides and nitrides (MXenes) and their precursors with positive and unlabeled machine learning. ACS Nano. 2019;13:3031-3041.

202. Ünver Y, Sancak K, Tanak H, et al. 5-Benzyl-4-[3-(1H-imidazol-1-yl)propyl]-2H-1,2,4-triazol-3(4H)-ones: synthesis, spectroscopic characterization, crystal structure and a comparison of theoretical and experimental IR results by DFT calculations. J Mol Struct. 2009;936:46-55.

203. Bare SR, Kelly SD, Vila FD, et al. Experimental (XAS, STEM, TPR, and XPS) and theoretical (DFT) characterization of supported rhenium catalysts. J Phys Chem C. 2011;115:5740-5755.

204. Paruzzo FM, Hofstetter A, Musil F, et al. Chemical shifts in molecular solids by machine learning. Nat Commun. 2018;9:4501.

205. Hu W, Ye S, Zhang Y, et al. Machine learning protocol for surface-enhanced Raman spectroscopy. J Phys Chem Lett. 2019;10:6026-6031.

206. Raccuglia P, Elbert KC, Adler PD, et al. Machine-learning-assisted materials discovery using failed experiments. Nature. 2016;533:73-76.

207. Kralinger M, Rabal O, Lourenco A, Oyarzabal J, Valencia A. Information retrieval and text mining technologies for chemistry. Chem Rev. 2017;117:7673-7761.

208. Kim E, Huang K, Saunders A, et al. Materials synthesis insights from scientific literature via text extraction and machine learning. Chem. Mater. 2017;29:9436-9444.

209. Lake BM, Salakhutdinov R, Tenenbaum JB. Human-level concept learning through probabilistic program induction. Science. 2015;350:1332-1338.

210. Vinyals O, Blundell C, Lillicrap T, Wierstra D. Matching networks for one shot learning. Advances in Neural Information Processing Systems 29, Annual Conference on Neural Information Processing Systems 2016; December 5-10, 2016, Barcelona, Spain.

211. Sung F, Zhang L, Xiang T, Hospedales T, Yang Y. Learning to learn: meta-critic networks for sample efficient learning. arXiv preprint arXiv:170609529; 2017.

212. Duan Y, Andrychowicz M, Stadie B, et al. One-shot imitation learning. Advances in Neural Information Processing Systems 30, Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA.

213. Yang L, Wang P, Jiang Y, Chen J. Studying the explanation of metasizing perovskite cobaltite through surface catalysis from molecular orbital principles. Science. 2011;334:1383-1385.

214. Zhou S, Miao X, Zhao X, et al. Engineering electrocatalytic activity in nanosized perovskite cobaltite through surface spin-state transition. Nat Commun. 2016;7:11510.
215. Tran K, Ulissi ZW. Active learning across intermetallics to guide discovery of electrocatalysts for CO$_2$ reduction and H$_2$ evolution. Nat Catal. 2018;1:696-703.

216. Zhang Y, He X, Chen Z, et al. Unsupervised discovery of solid-state lithium ion conductors. Nat Commun. 2019;10:5260.

217. Sun W, Bartel CJ, Arca E, et al. A map of the inorganic ternary metal nitrides. Nat Mater. 2019;18:732-739.

AUTHOR BIOGRAPHIES

**An Chen** gained her bachelor's degree in materials chemistry at Northeast Forestry University, China in 2018. She is studying for her master degree in Professor Zhen Zhou's group at Nankai University, China. Her research focuses on computational investigation of energy storage materials and devices.

**Xu Zhang** was born in Shandong, China. He obtained his BSc in chemistry at Nankai University, China in 2014. Then he obtained his PhD at Nankai University with Professor Zhen Zhou in 2019, majoring in materials physics and chemistry. He is now working as a postdoctoral fellow in Professor Zhen Zhou's group. His research interest mainly focuses on materials design for energy storage and conversion.

**Zhen Zhou** received his BSc (applied chemistry, in 1994) and PhD (inorganic chemistry, in 1999) from Nankai University, China. He joined the faculty at Nankai University as a lecturer in 1999. Two years later, he began to work as a postdoctoral fellow in Nagoya University, Japan. In 2005, he returned to Nankai University as an associate professor and was promoted as a full professor in 2011. In 2014, he was appointed as the Director of Institute of New Energy Material Chemistry, Nankai University. His main research interest is the design, preparation, and application of nanomaterials for energy storage and conversion.

**How to cite this article:** Chen A, Zhang X, Zhou Z. Machine learning: Accelerating materials development for energy storage and conversion. InfoMat. 2020;1–24. https://doi.org/10.1002/inf2.12094