J Materiomics 3 (2017) 159–177

Contents lists available at ScienceDirect

J Materiomics

journal homepage: www.journals.elsevier.com/journal-of-materiomics/

# Materials discovery and design using machine learning

Yue Liu [a], Tianlu Zhao [a], Wangwei Ju [a], Siqi Shi [b, c, *]

[a] *School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China*
[b] *School of Materials Science and Engineering, Shanghai University, Shanghai 200444, China*
[c] *Materials Genome Institute, Shanghai University, Shanghai 200444, China*

ARTICLE INFO

*Article history:*
Received 20 June 2017
Received in revised form
6 August 2017
Accepted 6 August 2017
Available online 8 August 2017

*Keywords:*
New materials discovery
Materials design
Materials properties prediction
Machine learning

ABSTRACT

The screening of novel materials with good performance and the modelling of quantitative structure-activity relationships (QSARs), among other issues, are hot topics in the field of materials science. Traditional experiments and computational modelling often consume tremendous time and resources and are limited by their experimental conditions and theoretical foundations. Thus, it is imperative to develop a new method of accelerating the discovery and design process for novel materials. Recently, materials discovery and design using machine learning have been receiving increasing attention and have achieved great improvements in both time efficiency and prediction accuracy. In this review, we first outline the typical mode of and basic procedures for applying machine learning in materials science, and we classify and compare the main algorithms. Then, the current research status is reviewed with regard to applications of machine learning in material property prediction, in new materials discovery and for other purposes. Finally, we discuss problems related to machine learning in materials science, propose possible solutions, and forecast potential directions of future research. By directly combining computational studies with experiments, we hope to provide insight into the parameters that affect the properties of materials, thereby enabling more efficient and target-oriented research on materials discovery and design.

## Contents

[*] Corresponding author. School of Materials Science and Engineering, Shanghai University, Shanghai 200444, China.
*E-mail address:* sqshi@shu.edu.cn (S. Shi).
Peer review under responsibility of The Chinese Ceramic Society.

http://dx.doi.org/10.1016/j.jmat.2017.08.002

## 1. Introduction

The screening of high-performance materials, the modelling of quantitative structure-activity relationships (QSARs) and other issues related to the chemical structures of materials and their various biological effects or activities of interest are not only scientifically important but also critical to the development of many technologically relevant fields [1,2]. Unfortunately, the repetitive experimental and theoretical characterization studies are often time-consuming and inefficient because significant progress tends to require a combination of chemical intuition and serendipity. For example, the time frame for discovering new materials is remarkably long, typically approximately 10–20 years from initial research to first use. As shown in Fig. 1, new materials research comprises seven discrete stages, namely, discovery, development, property optimization, system design and integration, certification, manufacturing and deployment, and the different stages may be conducted by different engineering or scientific teams at different institutions. Although experienced teams are involved in each stage of the process, there are few opportunities for feedback between earlier and later stages, which could accelerate the process as a whole [3].

It is well known that computational simulation and experimental measurement are two conventional methods that are widely adopted in the field of materials science. However, it is difficult to use these two methods to accelerate materials discovery and design because of the inherent limitations of both experimental conditions and theoretical foundations. Generally speaking, experimental measurement, which usually includes microstructure and property analysis, property measurement, synthetic experiments, and so on, is an easy and intuitive method of materials research, although it is usually conducted in an inefficient manner over a long time period. In addition, this kind of approach poses high requirements in terms of the equipment, the experimental environment, and the expertise of the researcher. Alternatively, computational simulation, ranging from electronic structure calculations based on density functional theory [4,5], molecular dynamics [6,7], Monte Carlo techniques [8], and the phase-field method [9–11] to continuum macroscopic approaches, is another approach in which existing theory is exploited for analysis using computer programs. Materials design guided by computation is expected to lead to the discovery of new materials and reductions in materials development time and cost [12]. Compared with experimental measurement, computational simulation requires less time and is advantageous for supplying real experiments in that one has full control over the relevant variables. Nevertheless, there are also many challenges related to computational simulation; e.g., it strongly depends on the microstructures of the materials involved; it requires high-performance computing equipment, usually in large computing clusters, on which the computational simulation programs can run; and no explicit use can be made of previous calculation results when a new system is studied. Modern materials research often requires close integration between computation and experiment to yield a fundamental understanding of the structures and properties of the materials of interest and how they are related to the synthesis and processing procedures. In particular, some experiments can be performed virtually using powerful and accurate computational tools, and thus, the corresponding time frame can be decreased from 10 or 20 years, as is required based on traditional methods, to 18 months [2,30].

Given the sophisticated requirements involved in understanding the basic physicochemical properties of materials and accelerating their technological applications, both experimental measurement and computational simulation are often incapable of addressing newly emerging issues. For instance, it is very complicated and inefficient to investigate the transition temperature of glass [13] through experimental measurements because the transition occurs over a wide temperature range. However, the glass transition temperature also cannot be exactly simulated using computer programs because it depends on a variety of internal and external conditions, such as pressure, structure, and constitutive and conformational features [14]. Therefore, numerous attempts have been made in the field of materials science to develop ways to overcome the shortcomings of these two common methods.

With the launch of the Materials Genome Initiative (MGI) [15] in 2011 and the coming of the “big data” era, a large effort has been made in the materials science community to collect extensive datasets of materials properties to provide materials engineers with ready access to the properties of known materials, such as the Inorganic Crystal Structure Database (ICSD) [16], the superconducting critical temperatures (SuperCon) [17], the Open Quantum Materials Database (OQMD) [18], the Cambridge Structural

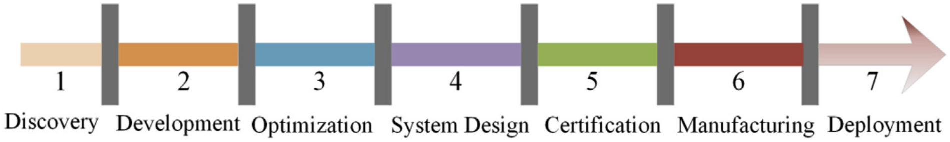


**Fig. 1.** The process of finding new materials using traditional methods [3].

Databases [19], the Harvard Clean Energy Project (HCEP) [20], the Materials Project (MP) [21], the Materials Commons [22], and the Materials Data Facility [23]. A generic data management and sharing platform could provide a powerful impetus to accelerate materials discovery and design. Advanced materials characterization techniques, with their ever-growing rates of data acquisition and storage capabilities, represent a challenge in modern materials science, and new procedures for quickly assessing and analyzing the collected data are needed [24]. Machine learning (see Section 2.1 for the detailed definition of this term) is a powerful tool for finding patterns in high-dimensional data; it employs algorithms by which a computer can learn from empirical data by modelling the linear or nonlinear relationships between the properties of materials and related factors [25]. In recent years, machine learning techniques [26] and big data methods [27] have successfully resolved the difficulties of modelling the relationships between materials properties and complex physical factors. Notably, the successful applications of machine learning, such as assisting in materials discovery based on failed experiments [28] and screening for efficient molecular organic light-emitting diodes [34], are regarded as an innovative mode of materials development, in which inverting the machine learning models reveals new hypotheses regarding the conditions for successful product formation.

Over the past 20 years, the computational activities related to materials science have been steadily shifting from technique development and purely computational studies of materials toward the discovery and design of new materials guided by computational results, machine learning and data mining or by close collaboration between computational predictions and experimental validation [29,30]. The advantages of modern materials research strategies lie in their ability to find a good balance between reasonable experimental requirements and a low error rate, to make full use of the extensive data available and to speed up the materials research process. Great efforts are being devoted to developing more suitable methods that combine traditional experimental methods with intelligent data analysis techniques to improve experimental efficiency and reduce the error rate. For example, Sumpter et al. [31] proposed a novel integrated method for guiding the synthesis of new inorganic materials, which is achieved through the incorporation of big data approaches in imaging and scattering coupled with scalable first principles. The techniques related to big data and deep data serve to link first-principles theoretical predictions with high-veracity imaging and scattering data from materials with microscopic degrees of freedom, thereby greatly accelerating rational materials design and synthesis. Haughtier et al. [32] and Meredig et al. [33] reported that an iterative combination of machine learning methods and first-principles calculations can greatly accelerate the discovery process for novel ternary compounds. The above examples indicate that machine learning can be effectively combined with theoretical computational methods to solve various problems related to materials science and that the corresponding experimental results have proven to be reliable. It is worth emphasizing that the successes of the above examples are all founded on a basis of extensive available data. In other words, the ability to make full use of extensive data is the key to the application of machine learning to materials science research. Moreover, seen from the perspective of big data, many failed experiments nevertheless provide valuable information that can be used to determine the boundaries between success and failure and can also be useful for new materials discovery, as shown in Ref. [28]. In addition, the application of intelligent data analysis methods in materials science research can greatly assist in optimizing the virtual screening space for new materials and speed up the process considerably [34].

Currently, several excellent review articles on materials informatics are available in the literatures [36–40]. Ref. [36] describes data-driven techniques for deciphering processing-structure-property-performance relationships. Ref. [37] presents a review of the current state of affairs with respect to data and data analytics in the materials community, with a particular emphasis on thorny challenges and promising initiatives that exist in the field. Ref. [38] provides a description of recent developments in materials informatics, concentrating specifically on relating a material's crystal structure and its composition to its properties. Ref. [39] presents a vision for data and informatics in the future materials innovation ecosystem. Ref. [40] discusses the intersection between materials informatics and atomistic calculations, with a particular focus on solid inorganic materials. The main purposes of the current paper are to review the applications of machine learning in Materials Discovery and Design and to analyze the successful experiences and the common existing problems. It is anticipated that this review can establish a new horizon toward which to conduct materials discovery and design. The remainder of this paper is organized as follows: Section 2 introduces the formal description of machine learning in materials science and classifies the commonly used machine learning methods into three broad categories. Sections 3–5 detail the research status with respect to the applications of machine learning in material property prediction, in new materials discovery and for various other purposes, respectively. Section 6 discusses the limitations and drawbacks from which machine learning suffers in the field of materials science and then proposes corresponding methods of improvement. Finally, the most significant conclusions from this review are summarized in Section 7.

## 2. Description of machine learning methods in materials science

As a scientific endeavor, machine learning grew out of the quest for artificial intelligence [41]. In the 1950s, attempts were made to approach the problem of acquiring knowledge by machine using various symbolic methods [42], and later, methods based on the connection principle, such as neural networks and perceptrons, were broadly studied [43]. Subsequently, several statistical learning theory (SLT)-based methods, such as support vector machines (SVMs) [44] and decision trees (DTs) [45], were proposed. Currently, several new machine methods, such as deep learning for big data analysis, have attracted attention in both academia and industry. Machine learning is a method of automating analytical model building. Using algorithms that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed where to look [46].

Machine learning shows good applicability in classification, regression and other tasks related to high-dimensional data. Aimed at extracting knowledge and gaining insight from massive databases, machine learning learns from previous computations to produce reliable, repeatable decisions and results and thus has played an important role in many fields, especially speech recognition, image recognition [47], bioinformatics [48], information security [49], and natural language processing (NLP) [50]. Many of our day-to-day activities are powered by machine learning algorithms, including fraud detection, web searches, text-based sentiment analysis, credit scoring and next-best offers. The pioneering applications of machine learning in materials science can be traced back to the 1990s, when machine learning methods such as symbol methods and artificial neural networks (ANNs) were employed to predict the corrosion behavior and the tensile and compressive strengths of the fiber/matrix interfaces in ceramic-matrix composites [51–53]. Subsequently, machine learning has been used to address various topics in materials science, such as new materials discovery and material property prediction.

### 2.1. Paradigms of machine learning in materials science

A classical definition of machine learning is as follows: <*P*,*T*,*E*>, where *P*, *T* and *E* denote performance, task and experience, respectively. The main interpretation is that a computer program is said to learn from experience *E* with respect to some class of tasks *T* and a performance measure *P* if its performance on tasks in *T*, as measured by *P*, improves with experience *E* [54]. In general, a machine learning system should be constructed when using machine learning to address a given problem in materials science. The general paradigm of such machine learning systems is given as follows:

$$Goal + Sample + Algorithm = Model \tag{1}$$

here, the ultimate *Goal* represents the given problem, which is usually expressed in the form of an objective function. The *Sample* is a subset of the population that is selected for study in some prescribed manner [55]. In general, data preprocessing, including data cleaning and feature engineering, is employed to convert the original data into the *Sample*. Data cleaning refers to identifying incomplete, incorrect, inaccurate and irrelevant parts of the data and then replacing, modifying, or deleting these dirty or coarse data [56]. Feature engineering, including feature extraction, feature selection, feature construction and feature learning, is the process of using domain knowledge regarding the data to create features that allow machine learning algorithms to function. Feature engineering is fundamental to the application of machine learning and is both difficult and expensive. The *Algorithm*, which mainly includes the machine learning algorithm and the model optimization algorithm, is a self-contained step-by-step set of operations to be performed [57]. The most commonly used machine learning algorithms are SVM, DT and ANN algorithms. Model optimization algorithms mainly include genetic algorithms (GAs), simulated annealing algorithms (SAAs) and particle swarm optimization (PSO) algorithms. The *Model* is a description of a system in terms of mathematical concepts and refers to the algorithm that has been learned from the *Sample*.

### 2.2. Basic steps of machine learning in materials science

As shown in Fig. 2, the construction of a machine learning system is divided into three steps: sample construction, model building and model evaluation.

#### 2.2.1. The first step is sample construction

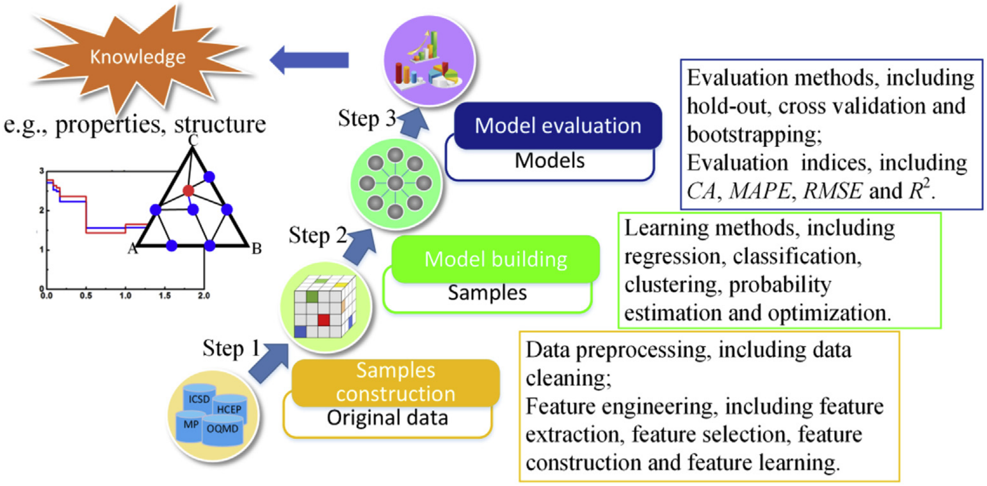
In materials science, the original data are collected from computational simulations and experimental measurements. These data are typically incomplete, noisy and inconsistent, and thus, data cleaning should be performed when constructing a sample from the original data. Furthermore, there are several conditional factors affecting any obtained sample, and some of them are not relevant to the decision attributes. For example, in research on the prediction of Li ionic conductivity [58], although several internal and external factors are expected to affect the ionic conductivity, only the four most relevant factors, namely, ion diffusivity, average volume, transition temperature and experimental temperature, are of interest in comparative experiments. Therefore, it is important to use a proper feature selection [59] method to determine the subset of attributes to be used in the final simulation.

#### 2.2.2. The second step is model building

It is essentially a black box linking input data to output data using a particular set of nonlinear or linear functions. For typical research in materials science, complex relationships usually exist between the conditional factors and the target attributes, which traditional methods have difficulty handling. Machine learning provides a means of using examples of a target function to find the coefficients with which a certain mapping function approximates the target function as closely as possible. For example, in an experiment on predicting the glass transition temperature $T_g$ [60], it is difficult to find a formula that can accurately describe the relationship between $T_g$ and the four relevant factors of rigidness, chain mobility, average molecular polarizability and net charge; however, a machine learning method can be used to model the relationships between conditional factors and decision attributes based on a given sample. This is where machine learning plays a role and where the "core" algorithms lie. The knowledge obtained through machine learning is stored in a format that is readily usable and, in turn, can be used for materials discovery and design.

#### 2.2.3. The last step is model evaluation

A data-driven model should achieve good performance not only on existing data but also on unseen data. Generally, we can evaluate the generalization errors of models by means of calculation-based tests and use the results to select the best one. To this end,



Evaluation methods, including hold-out, cross validation and bootstrapping;
Evaluation indices, including *CA*, *MAPE*, *RMSE* and $R^2$.

Learning methods, including regression, classification, clustering, probability estimation and optimization.

Data preprocessing, including data cleaning;
Feature engineering, including feature extraction, feature selection, feature construction and feature learning.

**Fig. 2.** The general process of machine learning in materials science.

testing data are needed to test the discriminative capabilities of models on a new dataset; then, the test error achieved on the testing data can be taken as an approximation of the generalization error. When there is only one dataset $D=\{(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)\}$ containing $m$ samples, we can partition $D$ into a training dataset $S$ and a testing dataset $T$ for training and testing, respectively, using several evaluation methods, as shown in Table 1.

In the hold-out method, the dataset $D$ is partitioned into a training dataset $S$ and a testing dataset $T$, such that $D=S\cup T$ and $S\cap T=\varnothing$. A category-preserving sampling method, known as "stratified sampling", is often used to maintain the consistency of the data distribution to the greatest possible extent, avoiding the introduction of additional deviations. Because there is no perfect solution for deciding the relative proportions of $S$ and $T$, 2/3–4/5 of the samples in $D$ are often designated as $S$, while the remainder are assigned to $T$.

In the cross-validation method, the original dataset $D$ is partitioned into $k$ mutually exclusive subsets of the same size, $D=D_1\cup D_2\cup\ldots\cup D_k, D_i\cap D_j=\varnothing(i\neq j)$, where each $D_i$ is generated through "stratified sampling". Then, the union set of $k$-1 of the subsets is taken as the training dataset $S$, and the remaining subset is used as the testing dataset $T$. When $k$ is equal to the number of samples $m$ in $D$, this method is known as leave-one-out cross-validation (LOOCV), which is not affected by random sample partitioning. Note that this process is repeated for each subset of the data, and thus, a total of $k$ experiments are conducted. Therefore, the cross-validation method may take a long time and is unsuitable for large datasets.

Based on bootstrapping sampling [61], the bootstrapping method involves copying one sample from $D$ into a new dataset $D'$ randomly and repeatedly until $D'$ contains $m$ samples. Finally, $D'$ is designated as the training dataset, and $D\backslash D'$ is used as the testing dataset. Because it maintains a number of training samples equal to the size of the original dataset, the bootstrapping method is effective when the data volume is small and it is difficult to properly partition the training/testing data. However, the bootstrapping method changes the distribution of the original dataset, potentially introducing estimation bias.

When evaluating an algorithm, one tests it to see how well it performs. The accuracy of the model's predictions is examined by comparing experimental values with the corresponding predicted ones. The evaluation standards depend on the type of problem at hand.

The classification accuracy (CA) is used to evaluate models used in classification problems:

$$CA=S/N \tag{2}$$

where $S$ and $N$ denote the number of samples that are correctly classified and the total number of samples, respectively.

As shown in expressions (3)–(5), the mean absolute percent error (*MAPE*), the root mean square error (*RMSE*) and the correlation coefficient ($R^2$) are all used to evaluate models applied to solve regression problems.

$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\frac{|y'_i-y_i|}{y_i} \tag{3}$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y'_i-y_i)^2} \tag{4}$$

$$R^2=\frac{\left[\sum_{i=1}^{n}(y_i-\bar{y})(y'_i-\bar{y}')\right]^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2\bullet\sum_{i=1}^{n}(y'_i-\bar{y}')^2} \tag{5}$$

where $y_i$ and $y_i'$ represent an original value and the corresponding predicted value, respectively, and $\bar{y}$ and $\bar{y}'$ are the averages of the original and predicted values, respectively.

In addition, other indices for evaluating classification model accuracy include the precision, the recall, the receiver operating characteristic (ROC) curve, the logistic regression loss, the hinge loss, the confusion matrix, Cohen's kappa, the Hamming distance, the Jaccard similarity coefficient, the coverage error, the label ranking average precision and the ranking loss. For regression problems, the explained variance and the coefficient of determination are two additional commonly used indices. The Rand index, mutual information (MI), and the silhouette coefficient are indices used for model evaluation in clustering problems.

### 2.3. Commonly used machine learning algorithms in materials science

The selection of an appropriate machine learning algorithm is a key step in the construction of a machine learning system, as it greatly affects the prediction accuracy and generalization ability [62]. Each algorithm has its own scope of application, and thus, there is no algorithm that is suitable for all problems. As shown in Fig. 3, the commonly used machine learning algorithms in materials science can be divided into four categories: probability estimation, regression, clustering, and classification. Specifically, probability estimation algorithms are mainly used for new materials discovery, whereas regression, clustering and classification algorithms are used for material property prediction on the macro- and micro-levels. In addition, machine learning methods are commonly combined with various intelligent optimization algorithms [63] [64], such as GAs, SAAs or PSO algorithms, which are mainly used to optimize the model parameters. Furthermore, these optimization algorithms can also be employed to perform other difficult optimization tasks [65], such as the optimization of spatial

**Table 1**
The comparison of three evaluation methods.

| Method | Advantages | Disadvantages | Applicable situation |
|---|---|---|---|
| Hold-out | Low computational complexity. | The proper relative proportions of training/testing data are difficult to determine;<br>The volume of the training data is smaller than that of the original dataset. | The data volume is sufficient. |
| Cross-validation | Not greatly influenced by changes in the volume of training data. | The computational complexity is high, especially on a large dataset;<br>The volume of the training data is smaller than that of the original dataset. | The data volume is sufficient. |
| LOOCV | | | The data volume is small, and the training and testing data can be partitioned effectively. |
| Bootstrapping | Effective partitioning of training/testing data. | The distribution of the training data differs from that of the original dataset. | The data volume is small, and the training and testing data are difficult to properly partition. |

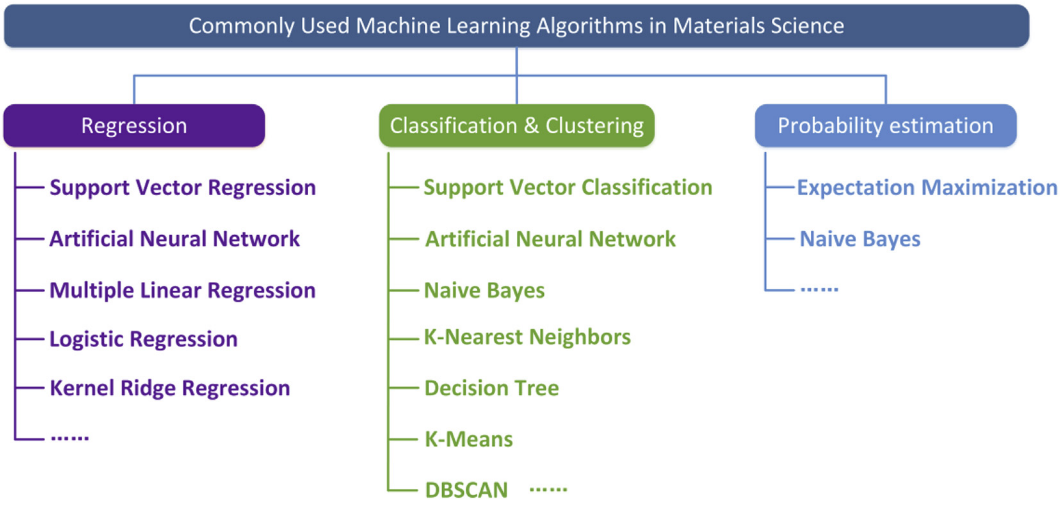

Commonly Used Machine Learning Algorithms in Materials Science
Regression
Support Vector Regression
Artificial Neural Network
Multiple Linear Regression
Logistic Regression
Kernel Ridge Regression
……
Classification & Clustering
Support Vector Classification
Artificial Neural Network
Naive Bayes
K-Nearest Neighbors
Decision Tree
K-Means
DBSCAN ……
Probability estimation
Expectation Maximization
Naive Bayes
……

**Fig. 3.** Commonly used machine learning algorithms in materials science.

configurations and materials properties [66] [67].

### 2.4. Overview of the application of machine learning in materials science

Machine learning is widely used in materials science and demonstrates superiority in both time efficiency and prediction accuracy. As shown in Fig. 4, the applications of machine learning in materials discovery and design can be divided into three main classes: material property prediction, new materials discovery and various other purposes. In research on material property prediction, regression analysis methods are typically used, and both macroscopic and microscopic properties can be predicted. The main idea underlying the application of machine learning in new materials discovery is to use a probabilistic model to screen various combinations of structures and components and finally to select a material with good performance from the candidate set by means of density functional theory (DFT)-based validation. In addition, machine learning is also used for other purposes in materials science, such as process optimization and density function approximation.

## 3. The application of machine learning in material property prediction

The properties of materials, such as hardness, melting point, ionic conductivity, glass transition temperature, molecular atomization energy, and lattice constant, can be described at either the macroscopic or microscopic level. There are two common methods of studying materials properties: computational simulation and experimental measurement. These two methods involve complicated operations and experimental setup. Therefore, it is quite difficult to build computational simulations that fully capture the complicated logical relationships between the properties of a material and their related factors, and some of these relationships may even be unknown. Moreover, the experiments that are performed to measure the properties of compounds generally occur in the later stages of materials selection. Consequently, if the results are not satisfactory, the enormous amounts of time and experimental resources invested up to that point prove to have been wasted [68]. In addition, in many cases, it is difficult or nearly impossible to study the properties of materials even through massive computational or experimental efforts. Therefore, there is in urgent need to develop intelligent and high-performance prediction models that can correctly predict the properties of materials at a low temporal and computational cost. Machine learning concerns the construction and study of algorithms that can learn patterns from data. The basic idea of using machine learning methods for material property prediction is to analyze and map the relationships (nonlinear in most cases) between the properties of a material and their related factors by extracting knowledge from existing empirical data.

Fig. 5 shows the fundamental framework for the application of machine learning in material property prediction. First of all, hand tuning or feature engineering (including feature extraction and selection) is conducted to identify the conditional attributes that are related to property prediction. Second, the mapping relationship between these conditional factors and the decision attributes is found through model training. Finally, the trained model can be used for property prediction. For instance, Isayev et al. [69] proposed a calculation tool called Property-Labelled Materials Fragments (PLMF) for building machine learning models to predict the properties of inorganic crystals. In PLMF, low-variance and highly correlated features are first filtered out to obtain a feature vector. Using the gradient boosting decision tree (GBDT) technique [70], a novel candidate material will first be classified as either a metal or an insulator, and the band gap energy will be predicted if the material is an insulator. Regardless of the material's metal/insulator classification, six thermomechanical properties (bulk modulus, shear modulus, Debye temperature, heat capacity at constant pressure, heat capacity at constant volume, and thermal expansion coefficient) are then predicted. Before model training, fivefold cross-validation is used to partition the dataset. The ROC curve, $RMSE$, $MAE$ and $R^2$ are all used to evaluate the prediction accuracy of the trained models. Depending on the scale of the analysis, the applications of machine learning in material property prediction can be divided into two broad categories: macroscopic performance prediction and microscopic property prediction.

### 3.1. Macroscopic performance prediction

Research on the macroscopic performance of materials mainly focuses on the structure-activity relationship between the macroscopic (e.g., mechanical and physical) properties of a material and its microstructure [71,72]. Because of their good performance in solving problems related to regression and classification, machine learning approaches involving ANN and SVM algorithms in combination with optimization algorithms have been widely applied in the study of macroscopic performance prediction.

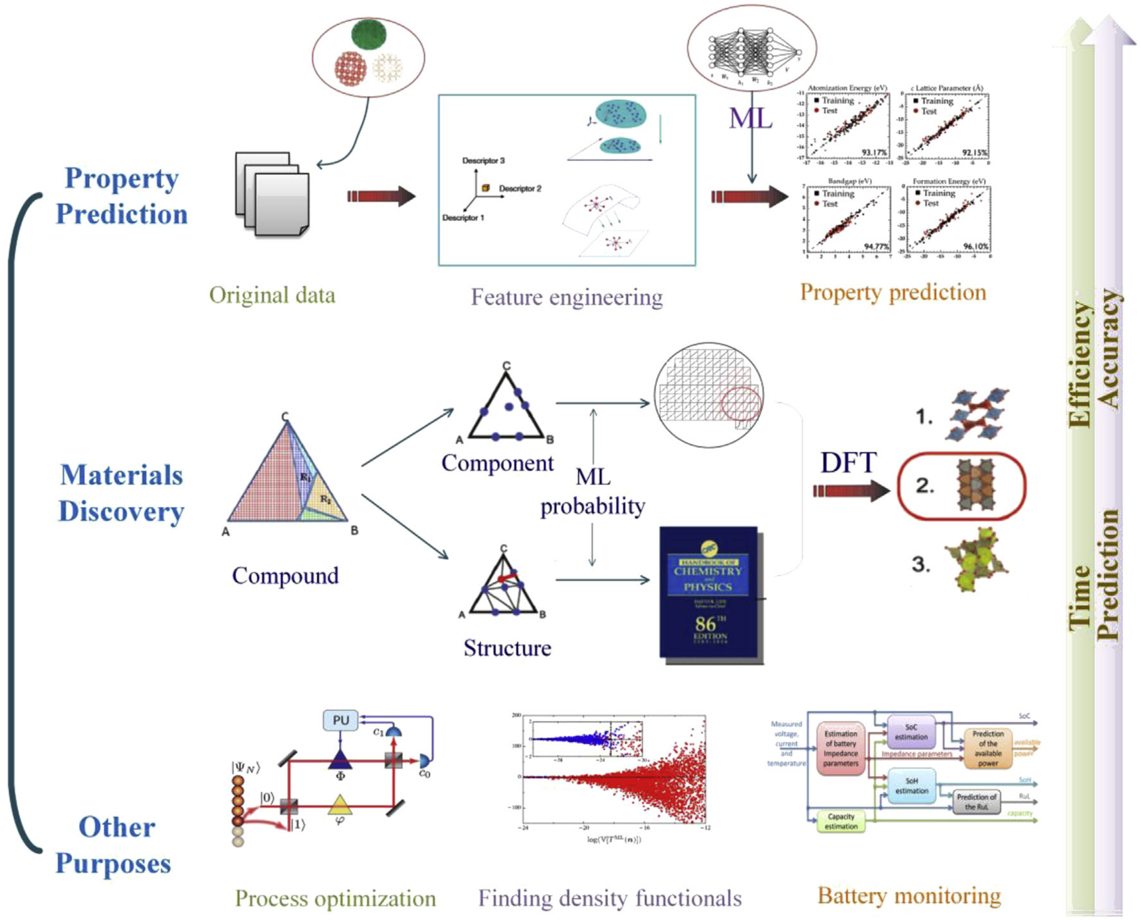


**Fig. 4.** An overview of the application of machine learning in materials science.

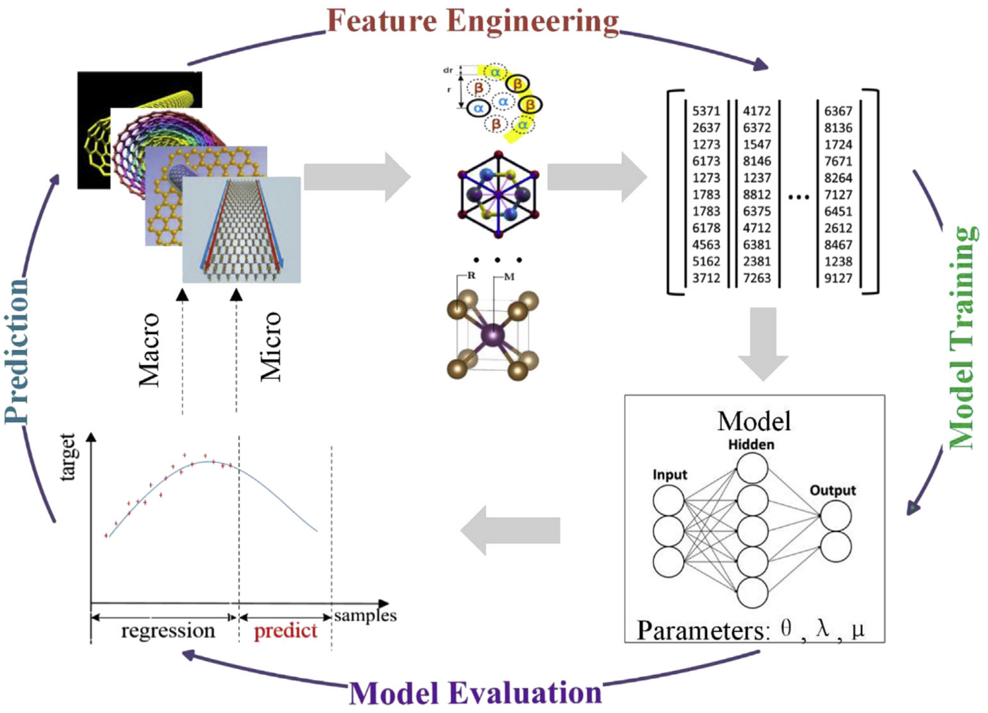


**Fig. 5.** The fundamental framework for the application of machine learning in material property prediction.

In machine learning, ANNs are a family of models inspired by biological neural networks (the central nervous systems of animals, particularly the brain) that are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown. The ANN approach is essentially a nonlinear statistical analysis technique with strong self-learning and adaptive capabilities [73]. Back propagation ANNs (BP-ANNs) [74–76] have been used to predict material behaviors such as temperature responses and tensile, elongation, wastage, corrosion and compressive properties. Chen et al. [77] employed a BP-ANN and a linear regression model to predict the glass transition temperatures of polymers and found that the average prediction error of the former approach (17 K) was much smaller than that of the latter (30 K). Since no comprehensive physical background knowledge is required, BP-ANNs can provide dramatic benefits to the industry by allowing such problems to be solved with acceptable prediction errors and a good generalization ability. However, they have a slow convergence rate and sometimes may fall into local minima. Alternatively, RBF-ANNs are another type of ANN that, by combining the ANN concept with the radial basis function, can fundamentally overcome the problem of local minima and also have the advantage of a high convergence rate. When using an RBF-ANN to investigate crack propagation in a bituminous layered pavement structure, Gajewski and Sadowski [78] reported that cracking considerably increases as the thickness of bituminous layer B2 increases. Moreover, ANN modelling has found a place in other applications, such as the prediction of melting points [79], the density and viscosity of biofuel compounds [80], excited-state energies [81], diffusion barriers [82] and other functional properties [83–86]. The greatest advantage of ANNs is their ability to be used as an arbitrary function approximation mechanism that ‘learns’ from observed data. The utility of an ANN model lies in the fact that little prior knowledge about the target material is required and the rules governing property variations can be exactly learned from empirical data. However, a common criticism of ANNs is that they require a highly diverse training dataset with sufficient representative examples for property prediction in order to capture the underlying structure to a sufficient extent that their results can be generalized to new cases. Furthermore, an inherent deficiency of neural networks is that the learned knowledge is concealed in a large number of connections, which leads to poor comprehensibility, i.e., poor transparency of knowledge and poor explainability.

The SVM technique is an SLT-based supervised learning method for analyzing data and recognizing patterns that can be used for classification and regression [87]. Compared with ANNs, SVM models are more suitable for application to small samples and can successfully overcome the problems of “the curse of dimensionality” and “overlearning”. They exhibit many unique advantages for solving nonlinear and high-dimensional problems. Fang et al. [64] proposed a novel hybrid methodology combining GAs and with support vector regression (SVR), which is capable of forecasting the atmospheric corrosion behaviors of metallic materials such as zinc and steel. Results show that this hybrid model provides a better predictive capability than other methods, and thus, it is regarded as a promising alternative method of forecasting the atmospheric corrosion behaviors of zinc and steel. To overcome the lower performance of a single SVR model for the prediction of the temperature of the As-Se glass transition, which involves radical structure changes, Ref. [88] proposed a Feature Selection based Two-Stage SVR (FSTS-SVR) forecasting method for investigating the turning point based on structural analysis and constructed a prediction model for each stage. The experimental results indicated that the prediction accuracy of the FSTS-SVR method is higher than that of other methods based on SVR in cases in which a turning point exists. In Ref. [89], an SVM approach was used to estimate the exposure temperatures of fire-damaged concrete structures. Based on the output results from the SVM analysis, the most effective parameter for improving the estimation accuracy was identified to be the ultrasonic pulse velocity of the concrete. In addition, SVM models can also be employed to predict ionic conductivities [58,90], glass transition temperatures [91–93] and various behaviors of functional materials [94,95].

### 3.2. Microscopic property prediction

The macroscopic performance of a material is determined by its microscopic properties, including its atomic and structural characteristics, such as the lattice constant. To the best of our knowledge, the applications of machine learning in microscopic property prediction tend to concentrate on several aspects, including the lattice constant, band energy, electron affinity and molecular atomization energy.

In the design of new substrate materials or buffer materials for semiconductor epitaxy, the lattice mismatch between layers consisting of different materials is one of the main concerns. Therefore, lattice constant prediction can help to accelerate the evaluation of the lattice constants for all possible combinations of elements. In this regard, a fast and reliable solution for predicting the lattice constants of a large number of unknown compounds is becoming a key requirement for high technology development. Machine learning can be applied in solving this prediction problem because of its good practicability in regression analysis.

As shown in Table 2, there have been several studies related to the use of machine learning for lattice constant prediction for a variety of perovskite crystal materials, in which machine learning regression methods such as the SVR, ANN and logistic regression (LR) techniques have been used. By using the LR and ANN techniques and a sample set of 157 known $GdFeO_3$-type $ABO_3$ perovskites, lattice constants in $GdFeO_3$-type $ABO_3$ perovskites were correlated with their constituent elemental properties [96]. The LR models were first obtained using only two elemental ionic radii, and the ANN models were generated based on five elemental properties; the ionic radii, the electro-negativities of cations A and B, and the valence of ion A were further considered to improve the predictive capabilities of the models. The results indicated that the LR (ANN) method achieved percentages of absolute difference (PADs) of 0.93, 0.82 and 0.77 (0.35, 0.34 and 0.44) for constants *a*, *b* and *c*, respectively. The ANN obviously had a better accuracy than the LR method, achieving an error within 2%. Furthermore, by comparing the prediction accuracies of the ANN method on the training and testing data, it was found that the prediction accuracy on the training data was high, whereas that on the testing data was not satisfactory. To overcome this problem of weak generalization ability that is observed for ANNs, Javed et al. [97] employed SVR to predict the lattice constants of orthorhombic $ABO_3$ perovskite compounds. It was found that the SVR model performed better than the ANN model on both the training and testing datasets. The average PAD values were less than 1% for all lattice constants. In addition, the SVR model required less training and testing time compared with the ANN model by virtue of the fast learning capability of the former. In Refs. [98], SVR, generalized regression neural network (GRNN) and ANN methods were used to predict the lattice constants of perovskite compounds. They achieved mean PAD values of 0.43%, 0.54% and 0.96%, respectively, indicating that the SVR method showed the best prediction performance. Majid et al. [99] used several machine learning methods, including SVR, random forests (RF), a GRNN and multiple linear regression (MLR), with the ionic radius as the only conditional attribute for predicting the lattice constants of complex cubic perovskites. The results indicated that these four methods yielded prediction accuracies of

**Table 2**
Applications of machine learning for lattice constant prediction.

| Reference | Conditional attributes | Compounds | ML methods | PAD | | |
|---|---|---|---|---|---|---|
| | | | | a | b | c |
| Li et al. [96] | ionic radii (r), valence (z) , electro-negativities (x) | $GdFeO_3$-type $ABO_3$ perovskites | LR/ANN | 0.93/0.35 | 0.82/0.44 | 0.77/0.34 |
| Javed et al. [97] | ionic radii (r), valence (z), electro-negativities (x) | $ABO_3$ perovskites | SVR | 0.52 | 0.54 | 0.58 |
| Majid et al. [98] | ionic radii (r), oxidation state (z), electro-negativities (x) | perovskites | SVR/GRNN/ANN | average of a, b, and c<br>0.43/0.54/0.96 | | |
| Majid et al. [99] | ionic radii (r), oxidation state (z), electro-negativities (x) | cubic perovskites | SVR/RF/GRNN/MLR | average of a, b, and c<br>0.26/0.41/0.70/0.51 | | |

0.26, 0.41, 0.70 and 0.51, respectively, all of which are higher than that of the well-known SPuDS software [100], which is extensively used in crystallography. The performances of the prediction models were also compared with each other in terms of linear correlation. Fig. 6 shows the performance curves relating the experimental and predicted values for the considered dataset. It can be seen that the experimental $R^2$ values for all four methods were larger than 0.95, indicating that these four machine learning methods all performed well in lattice constant prediction. Furthermore, the SVR method proved to have a higher generalization ability than the ANN, RF and GRNN methods when the training sample was small.

Hansen et al. [101] outlined five machine learning methods, including linear ridge regression (LRR), kernel ridge regression (KRR), SVR, the K-nearest-neighbor (KNN) method and the ANN method, and investigated the influence of the molecular representation on the methods' performance. It was found that all five methods allow quantum-chemical computations to be performed in a matter of milliseconds rather than hours or days when using *ab initio* calculations and that the KRR method achieves the lowest prediction error of 3 kcal/mol for the atomization energies of a wide variety of molecules. Using the Pearson coefficient, RF and SVR, Liu et al. [102] constructed a two-phase hybrid system including feature extraction and regression. Experimental results on a total of 2500 microscale volume elements (MVEs) showed that this system can build elastic localization linkages with a training (testing) mean absolute strain error (*MASE*) of 7.17% (13.02%). More recently, these authors constructed a multi-agent data-driven predictive model by introducing K-means clustering and principal component analysis (PCA) to improve upon the performance of a single-agent system [103]. The results proved that the multi-agent model could achieve better *MASE*

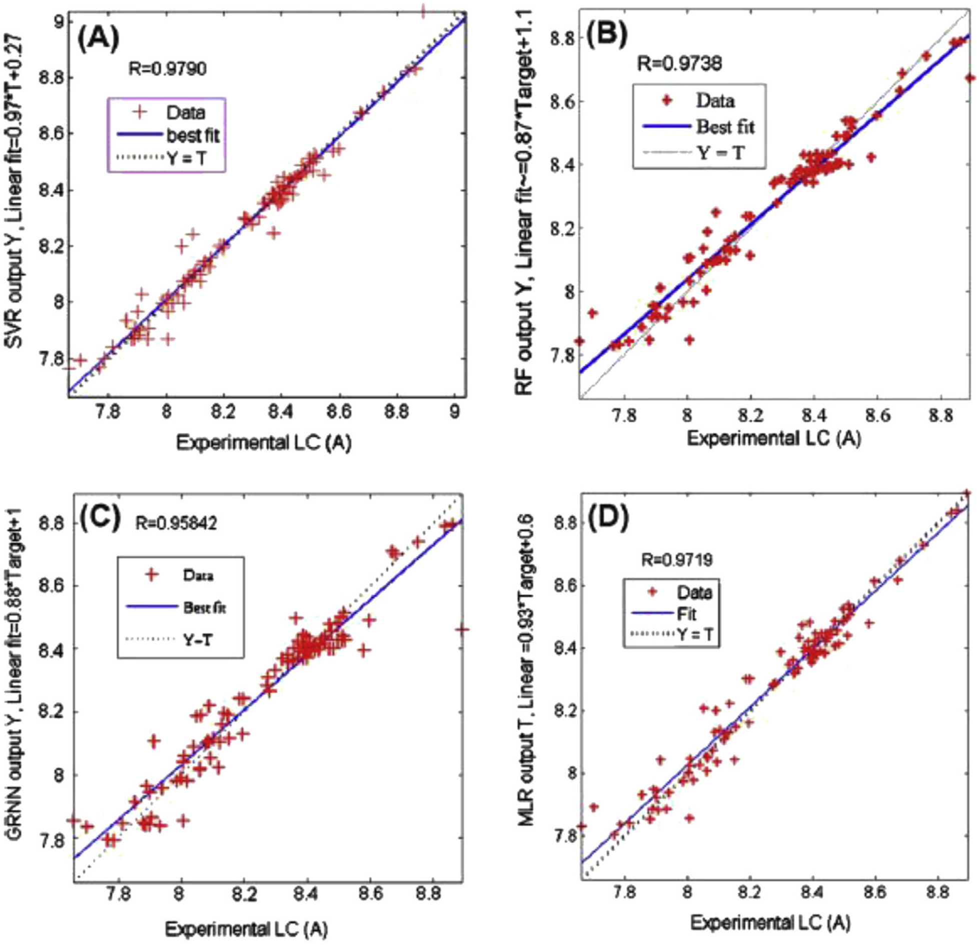


**Fig. 6.** (A-D) Show the linear correlation between the experimental and predicted values for each machine learning model, where the red points in each plot represent the original data, the blue line is the curve fitted to the results, and the dotted line is provided for reference.

values for both training and testing. To improve the time efficiency and prediction accuracy of machine learning methods for predicting the band gap energies and glass-forming ability of inorganic materials, Ward et al. [104] applied three key strategies to design a general-purpose machine learning framework with improved efficiency and accuracy. First, they summarized a general-purpose attribute set containing 145 elemental properties that could effectively capture the decision properties. Second, an ensemble learning method was used to overcome the disadvantages of each individual method. Third, they utilized a partitioning strategy in which the elements of the dataset were grouped into chemically similar subsets and trained a separate model on each subset. By employing machine (or statistical) learning methods trained on quantum mechanical computations in combination with the notion of chemical similarity, Pilania et al. [105] reported that it is possible to efficiently and accurately predict a diverse set of microscopic properties of material systems, as shown in Fig. 7. This is the first example in which a distance-measurement-based approach has been used to predict the microscopic properties of polymer materials. The proposed machine learning methodology for microscopic property prediction consists of four steps. First, material motifs within a class are reduced to numerical fingerprint vectors. Next, a suitable measure of chemical similarity or chemical distance is used to evaluate each material. Finally, a learning scheme—in this case, KRR is employed to map the relationship between the distances and properties. Fingerprints based on either chemo-structural (compositional and configurational) information or the electronic charge density distribution can be used to make ultra-fast, yet accurate, property predictions. Recently, machine learning methods have also been applied in material property prediction for lithium-ion batteries. Chen et al. [106] combined the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) method with information theory to elucidate the Li diffusion mechanism in the $Li_7La_3Zr_2O_{12}$ crystal lattice. By clustering the trajectories computed using molecular dynamics simulations, it was found that DBSCAN can be used to recognize lattice sites, determine the site type, and identify Li hopping events. Wang et al. [107] constructed QSAR formulations for cathode volume changes in lithium-ion batteries using the partial least squares (PLS) method based on data obtained from *ab initio* calculations. A variable importance in projection (VIP) analysis showed that the radius of the $X^{4+}$ ion and the *X* octahedron descriptors strongly affect the volume changes of a cathode during delithiation. In addition, ANNs have been successfully used for the prediction of vacancy migration and formation energies [105], electron affinities [105], vacancy migration energies [108] and potential energies [109], thereby fostering a deeper physical understanding of the microscopic properties of complex chemical systems.

As seen from the above analysis of their various application cases, machine learning methods exhibit higher accuracy and robustness than traditional methods of material property prediction (macro and micro), although the performance strongly depends on the machine learning method selected and the sample construction. It is worth mentioning that no single machine learning method can achieve good results for all applications. Therefore, selecting the best solutions always requires a comparison of various machine learning methods. Data cleaning and feature engineering are key steps of sample construction [110] that help to improve the prediction accuracy of a trained model through optimization of the features on which it is based. To verify the importance of feature selection, we conducted experiments on five commonly used material property datasets [58,79,80,96] using Multi-Layer Filtering Feature Selection (MLFFS). MLFFS is a novel method proposed by us in Ref. [111], which employs variance filtering, the Pearson coefficient, RF and PCA to sequentially and automatically delete sparse redundant and irrelevant features. The results are listed in Table 3. It can be seen that the numbers of features in all five datasets decrease after feature selection and that the model prediction accuracies are increased in terms of the *RMSE*, *MAPE* and $R^2$ values. In particular, the numbers of features in Dataset 4 and Dataset 5 are reduced by almost half, with lower *RMSE* and *MAPE* values and higher $R^2$ values. Thus, it is obvious that feature selection can improve the accuracy of material property prediction, and MLFFS is expected to become an efficient tool for materials science researchers.

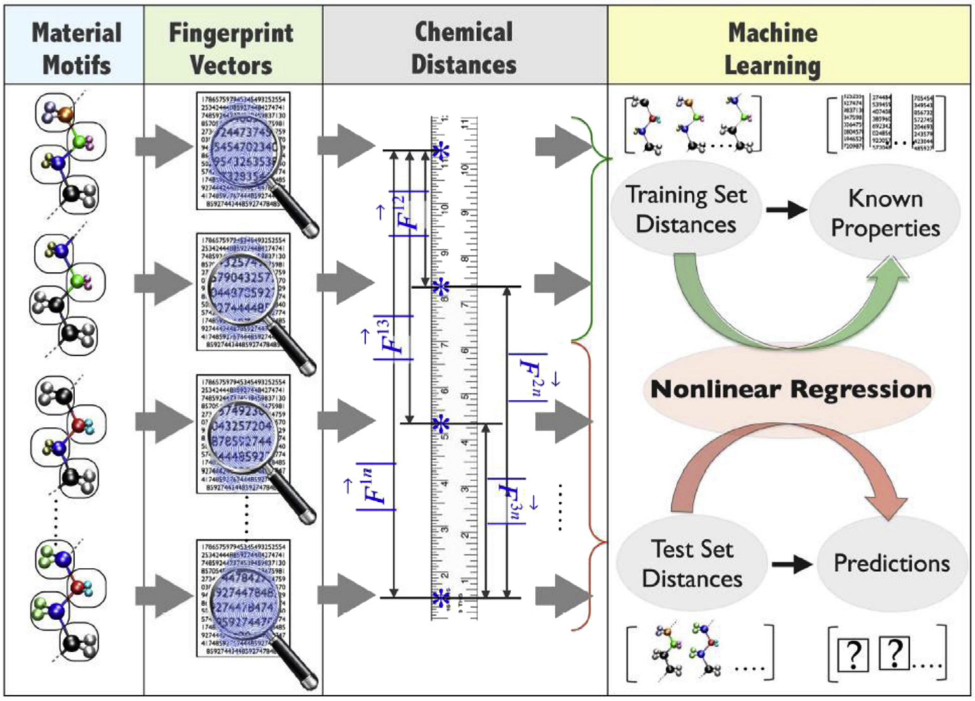


**Fig. 7.** A machine learning methodology for using distance measurements to predict the microscopic properties of materials.

## 4. The application of machine learning in new materials discovery

Finding new materials with good performance is the eternal theme in materials science. Currently, experimental and computational screenings for new materials discovery involve element replacement and structure transformation. However, the compositional search space, structural search space, or both tend to be sharply constrained [33]. Both screening methods may also require massive amounts of computation or experimentation and usually result in effort being directed in incorrect directions in an “exhaustive search”, which consumes considerable time and resources. In consideration of this fact and the advantages of machine learning, a completely adaptive method combining machine learning with computational simulation is proposed for the evaluation and screening of new materials “in silico” to provide suggestions for new and better materials.

The general process of machine learning in the discovery of new materials is shown in Fig. 8. The machine learning system for discovering new materials includes two parts, i.e., a learning system and a prediction system. The learning system performs the operations of data cleaning, feature selection, and model training and testing. The prediction system applies the model that is obtained from the learning system for component and structure prediction. New materials are often “predicted” through a suggestion-and-test approach: candidate structures are selected by the prediction system through composition recommendation and structure recommendation, and DFT calculations are used to compare their relative stability.

At present, various machine learning methods are used for finding new materials with good performance (see Table 4). They can be mainly divided into methods focused on crystal structure prediction and methods focused on composition prediction, which will be discussed in more detail in the following subsections. However, Ref. [112] represents an exception, in which the integration of ANN and GA approaches failed to accelerate new materials discovery since the relevant descriptors are not well known; nevertheless, this work can be regarded as an implementation of a novel concept for new materials discovery.

### 4.1. Crystal structure prediction

The prediction and characterization of the crystal structure of materials constitute a key problem that forms the basis for any rational materials design. Through crystal structure prediction, some unnecessary structure-related experiments can be avoided, which will greatly reduce the consumption of DFT calculation and computing resources while also helping to discover new materials. Predicting crystal structures following a chemical reaction is even more challenging because it requires an accurate potential energy surface for the entire reaction. First-principles crystal structure prediction is fundamentally difficult even for simple crystallization because of the need to consider a combinatorically enormous set of component arrangements using high-level quantum chemistry methods [113]. By contrast, machine learning is typically focused on empirical rules that have been extracted from a large amount of experimental information by algorithms, and this approach is attracting increasing attention.

Research on crystal structure prediction received essentially no attention before the 1980s. Maddox described the lack of crystal structure prediction in materials science as “one of the stigmas in physics” [114]. Over the past 10 years, machine learning has been used for crystal structure prediction. In 2003, Curtarolo et al. [115] transferred the concept of heuristic rule extraction to a large library of *ab initio* calculated information and creatively combined machine learning with quantum mechanical calculations to predict the crystal structures of binary alloys. More specifically, the machine learning model was used to predict several candidate crystal structures, and DFT calculations were used to determine their thermodynamic stability. Nevertheless, the disadvantage of this machine learning method was that it only predicted crystal structures existing in the database instead of novel structures. By using electronegativity, atomic size and atomic location points to describe the crystal structure, Ceder et al. [116] investigated the structure prediction problem by employing principal component regression and Bayesian probability to relate electronegativity and atom size to crystal structure, thereby gaining considerable insight into the physical mechanism that dominates structure prediction. From the perspective of knowledge extraction from computational or experimental data, Fischer et al. [117] constructed an informatics-based structure suggestion model for structure prediction, Data Mining Structure Predictor (DMSP), which rigorously mines correlations embodied within experimental data and uses them to efficiently direct quantum mechanical techniques toward stable crystal structures. Phillips and Voth [118] introduced two point-clustering algorithms, DBSCAN and OPTICS, with which new types of crystalline structures can be automatically identified from large datasets of coordinates. To address the challenges of the high dimensionality of the microstructure space, multi-objective design requirements and the non-uniqueness of solutions in structure prediction, Liu et al. [119] proposed a systematic machine learning framework consisting of random data generation, feature selection and classification algorithms to predict the microstructures of magnetoelastic Fe-Ga alloys. The results showed that the framework outperforms traditional computational methods, with an average running time that is reduced by as much as 80% and an optimality that cannot be achieved otherwise. In 2016, by combining finite-temperature phonon calculations with PCA and regression analysis, Roekeghem et al. [120] calculated the mechanical stability of approximately 400 semiconducting oxides and fluorides with cubic perovskite structures at 0 K, 300 K and 1000 K. They ultimately found 92 perovskite compounds that are mechanically stable at high temperatures, including 36 new compounds. Targeting novel emissive layers for organic light-emitting diodes (OLEDs), Rafael et al. [34] employed a machine learning method to screen efficient OLED molecules, in which multi-task neural networks were used as the training algorithm and each molecule was converted into a fixed-dimensional vector using extended connectivity fingerprints (ECFPs). From 400,000 candidate molecules, they identified 2500

**Table 3**
Prediction results for five datasets obtained when using MLFFS for feature selection.

| Dataset | Original Features | Original Prediction Result | | | Selected Features | Final Prediction Result | | |
|---|---|---|---|---|---|---|---|---|
| | | *RMSE* | *MAPE* | $R^2$ | | *RMSE* | *MAPE* | $R^2$ |
| 1 [58] | 6 | 0.59 | 0.26 | 0.96 | 4 | 0.58 | 0.26 | 0.97 |
| 2 [96] | 6 | 0.08 | 0.01 | 0.93 | 5 | 0.07 | 0.01 | 0.90 |
| 3 [79] | 5 | 8.69 | 0.09 | 0.96 | 4 | 7.24 | 0.07 | 0.96 |
| 4 [80] | 47 | 28.36 | 0.03 | 0.97 | 25 | 25.16 | 0.03 | 0.98 |
| 5 [80] | 18 | 0.64 | 0.27 | 0.91 | 9 | 0.36 | 0.15 | 0.97 |

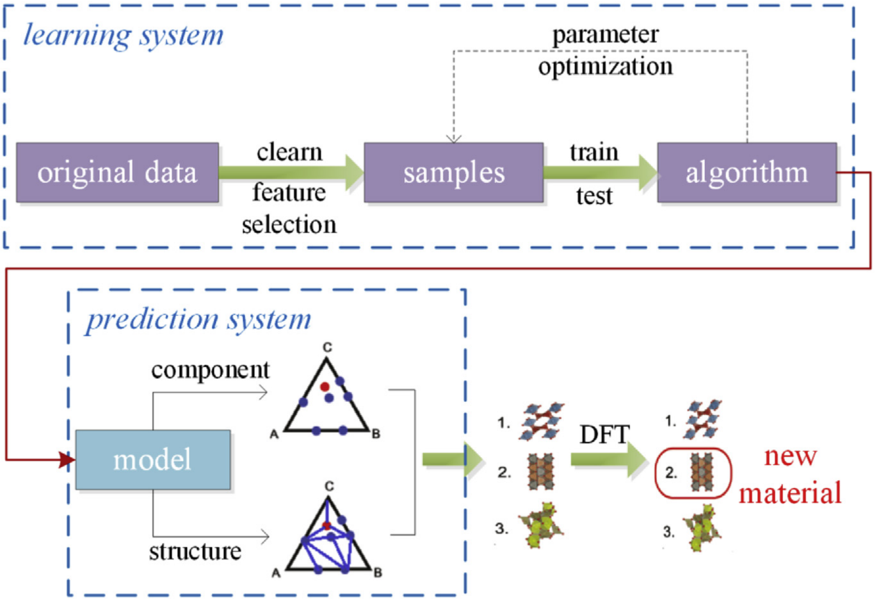


**Fig. 8.** The general process of machine learning in the discovery of new materials.

**Table 4**
Applications of machine learning in the discovery of new materials.

| Application description | Reference | ML method | Achievement |
|---|---|---|---|
| The design of new guanidinium ionic liquids | [71] | ANN | 6 new guanidinium salts |
| Finding nature's missing ternary oxide compounds | [32] | Bayesian | 209 new compounds |
| Discovery of new compounds with ionic substitutions | [123] | Bayesian | substitution rates of 20 common ions |
| Discovering crystals | [118] | DBSCAN & OPTICS | acceleration of finding new materials |
| Screening new materials in an unconstrained composition space | [33] | Bayesian | 4500 new stable materials |
| Machine-learning-assisted materials discovery using failed experiments | [28] | SVM | success rate of 89% |
| Virtual screening of materials | [112] | ANN | failed |

promising novel OLED molecules through machine learning prescreening and collaborative decision-making. The excellent predictive power achieved resulted in the report of devices with over 22% efficiency. Sendek et al. [121] used an LR model for the screening of solid lithium-ion conductor materials. By screening the MP database for materials that satisfied specific requirements, they first cut the number of candidate materials down from 12,831 to 317, a reduction of 92.2%. Then, they applied LR to develop an ionic conductivity classification model for further screening, and the 21 most promising materials were finally obtained, corresponding to an overall reduction of 99.8%.

From previous studies, the vast majority of unreported ‘dark’ (failed) chemical reactions are archived in laboratory notebooks that are generally inaccessible. In fact, however, these reactions contain valuable information for determining the boundaries between success and failure, and they may also be useful for new materials discovery. Raccuglia et al. [28] made full use of failed reaction data and demonstrated an alternative approach in which an SVM-derived DT algorithm trained on reaction data was used to predict reaction outcomes for the crystallization of templated vanadium selenites. The proposed method outperformed traditional human strategies, successfully predicting conditions for new organically templated inorganic product formation with a success rate of 89%.

### *4.2. Component prediction*

Component prediction is another way to discover new materials. In brief, one must decide which chemical compositions are likely to form compounds. Machine learning is more widely applied in component prediction than in crystal structure prediction. The bottlenecks for empirical or semi-empirical methods are that the search space for components is very limited and such searches require many verification calculations and experiments, which can severely affect the new materials discovery progress. Currently, the research on machine-learning-based component prediction can be divided into two main classes: 1) the recommendation of element combinations from a pool of elements for a given structure and 2) the presentation of ionic substitutions for the discovery of new compounds. Whereas crystal structure prediction can be performed by means of regression analysis without any prior knowledge, component prediction proceeds by means of solving for a posteriori probabilities through a Bayesian statistical model.

The difference between classical statistical models and Bayesian statistical models lies in whether prior information is used [122]. A Bayesian statistical model attaches great importance not only to the use of total sample information and single sample information but also to the collection, mining and processing of prior information. Bayesian statistical models are used to predict material components because they exhibit good performance for posterior probability estimation. Hautier et al. [32] used a Bayesian statistical method to extract knowledge from 183 common oxides in the ICSD database and successfully predicted 209 new ternary oxides. The calculation cost was reduced by nearly a factor of 30 compared with the traditional (exhaustive) method. Fig. 9(a) shows the distribution of the new compounds for every *A*-*B*-O system across chemical classes, where *A* and *B* are plotted on the *x* and *y* axes, respectively. The elements are ordered according to their Mendeleev numbers. This ordering of the elements allows us to spot the different chemical classes in which new compounds have been directly found. Meredig et al. [33] used the same method for the component prediction of ternary compounds, instead of only ternary oxides; they successfully predicted 4500 kinds of ternary compounds with thermodynamic stability, and the computation time was reduced by six orders of magnitude compared with a single first-principles calculation.

In a study of ionic-substitution-based component recommendation, Hautier et al. [123] proposed a model assessing the likelihood for ionic species to substitute for each other while preserving the crystal structure and showed how such a model can be used to predict new compounds and their crystal structures. The predictive power of this model was demonstrated via cross-validation on 2967 quaternary ionic compounds provided by the ICSD. As shown in Fig. 9(b), positive values indicate a tendency to substitute, whereas negative values indicate a tendency not to substitute, and the substitution rules embedded in the model can be used to predict compounds in the much less populated quaternary space.

As reviewed above, powerful machine learning tools can be developed for finding novel materials via both crystal structure prediction and component prediction. However, some difficulties still exist in the data collection stage when machine learning methods are used for crystal structure and component prediction. One of the main reasons for these difficulties is the lack of availability of large and consistent datasets due to the high cost of library synthesis.

## 5. The application of machine learning for various other purposes

Machine learning has been applied for material property prediction and new materials discovery, which has yielded many remarkable results. In addition, it is also employed to solve other problems related to materials science that involve massive computations and experiments. Note that some of these problems cannot be solved at all via traditional methods.

### 5.1. Process optimization

Process optimization is the design of the process parameters in material synthesis. In past production practice, reasonable material processing procedures were formulated through theoretical analysis and the accumulation of experience. Fuzzy neural networks (FNNs) represent a machine learning method that integrates the excellent learning capability of neural networks with fuzzy inference for deriving the initial rules of a fuzzy system. Han et al. [124] used MLR and FNN techniques to establish models of the relationship between the technological parameters and mechanical properties for the titanium alloy Ti-10V-2Fe-3Al; using these models, the optimal processing parameters for achieving the desired mechanical properties could be quickly selected. Compared with the MLR method, the FNN method achieved much better agreement with the experimental results, with a relative error below 7%, and showed better prediction precision in terms of the *RMSE*, $R^2$ and *MAPE* indices. In Refs. [125,126], neuro-fuzzy and physically based models were jointly used to predict the flow stress and microstructural evolution during the thermomechanical processing of aluminum alloys, achieving reasonable agreement with the experimental data. In addition, in Ref. [127], a model based on least squares support vector machines (LSSVMs) was considered as a powerful modelling tool for optimizing the aging process of aluminum alloys, and a desirable solution was obtained.

### 5.2. Finding density functionals

More than 10,000 papers each year report solutions to electronic structure problems obtained using Kohn-Sham (KS) DFT. All approximate the exchange-correlation (XC) energy as a functional of the electronic spin densities. The quality of the results crucially depends on these density functional approximations. For example, the present approximations often fail for strongly correlated systems, rendering the methodology useless for some of the most interesting problems. Recently, by defining the key technical concepts that are needed to apply machine learning to DFT problems, Snyder et al. [128] adopted machine learning to address a prototype

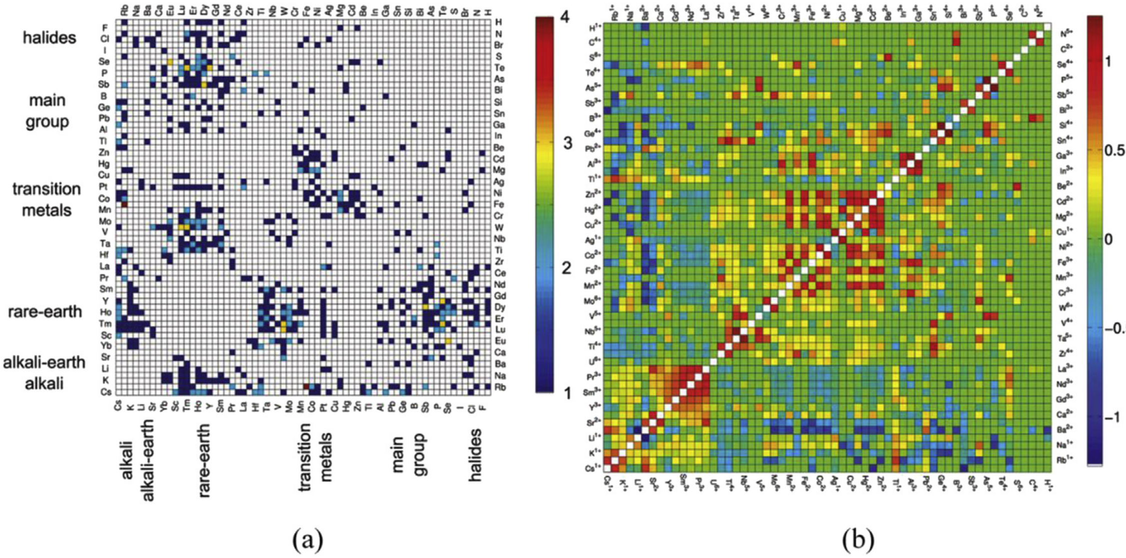


**Fig. 9.** (a) Distribution of the new compounds for every *A*-*B*-O system across chemical classes, where *A* is plotted on the *x* axis and *B* is on the *y* axis [32]. (b) Logarithm (base 10) of the pair correlation $g_{ab}$ for each ion couple (a, b) [123].

density functional problem: non-interacting spinless fermions confined to a 1D box, subject to a smooth potential. The accuracy achieved in approximating the kinetic energy (KE) of this system, with mean absolute errors below 1 kcal/mol on test densities similar to the training set when trained on fewer than 100 densities, is far beyond the capabilities of any present approximations. Moreover, it is even sufficient to produce highly accurate self-consistent densities. This machine learning approximation (MLA) approach uses many more inputs to achieve chemical accuracy but requires far less insight into the underlying physics.

### 5.3. *Battery monitoring*

Battery monitoring refers to the continuous determination of the state of a battery during operation. It is a challenging task in battery management systems (BMSs) because the state of the battery is affected by various internal and external conditions, and the relationship between these conditions and the battery state is nonlinear and changes over the lifetime of the battery. Impedance spectroscopy, voltage pulse response, and Coulomb counting are three of the main traditional methods used for battery monitoring, all of which have the same drawbacks: each is suitable only for a certain type of battery and works only for estimating the state of charge (SoC) of a battery in the steady state but fails for a battery that is being charged or discharged. Machine learning provides a superior means of predicting battery parameters because of its advantage of capturing the relationship between the battery state and related factors by constructing a trained model [129]. Great efforts have been devoted to employing machine learning methods to monitor various battery state parameters, such as the SoC, capacity, impedance parameters, available power, state of health (SoH) and remaining useful life (RUL), in real time.

Li et al. [130] established an FNN model for SoC estimation in which the internal resistance, voltage and current of the battery are used as the primary input variables. Lee et al. [131] proposed a machine learning system implemented based on learning controllers, FNNs and cerebellar-model-articulation-controller networks for SoC estimation. By using voltage and discharge efficiency as the input variables, this machine learning system not only can generate an estimate of how much residual battery power is available but also can provide users with additional useful information, such as an estimated travel distance at a given speed. As part of the prediction of available power, a BP-ANN optimized using the Gauss-Newton (GN) method with gradient descent (GD) has been applied to predict the current aging state of a battery [132]. The trade-off between the computational cost of batch learning and the accuracy achieved during on-line adaptation was optimized, resulting in a real-time system with a time forward voltage prognosis (TFVP) absolute error of less than 1%. The SoH is a more powerful performance indicator that is closely related to the ability of a cell/battery to perform a particular discharge or charge function at a given instant in the charge-discharge-stand-cycle regime. SVM methods have been found to be a powerful means of SoH estimation. An SoH estimation method based on SVM models and virtual standard performance tests has been validated and evaluated [133]. In addition, machine learning has also been coupled with the extended Kalman filter (EKF) for monitoring battery states. Charkhgard and Farrokhi [134] used a neural network and the EKF to model the SoCs of LIBs, achieving good estimation of the SoC and fast convergence of the EKF state variables.

### 5.4. *Other classification problems*

In Ref. [135], in which the reconstruction of three-dimensional (3D) microstructures was posed and solved as a pattern recognition problem, a combination of classification methodology with PCA for the effective reduced-order representation of 3D microstructures was demonstrated. This pattern recognition technique uses two-dimensional microstructure signatures to generate 3D realizations in nearly real time, thus accelerating the prediction of materials properties and contributing to the development of materials by design. Addin et al. [136] introduced Bayesian networks in general and a Naïve Bayes classifier in particular as one of the most successful classification systems for simulating damage detection in engineering materials. In Ref. [137], a new three-step method that involves the most common descriptors, multivariate analysis techniques and DT models was proposed and successfully tested for assessing the characteristics of nanoparticle aggregates and classifying them into the four most widely accepted shapes: spheroidal, ellipsoidal, linear and branched. Furthermore, using the SVM approach, Fernandez et al. [138] developed the Quantitative Structure-Property Relationship (QSPR) model, which can rapidly and accurately recognize high-performing metal-organic framework (MOF) materials for $CO_2$ capture. The SVM classifier could recover 945 of the top 1000 MOFs in the test set while flagging only 10% of the entire library for compute-intensive screening.

## 6. Analysis of and countermeasures for common problems

### 6.1. *Sample construction*

A sample is a subset of the original data that is selected for study in some prescribed manner. In the context of machine learning, the term sample refers to the basic data, which usually include training data and test data. At present, problems related to sample construction can be mainly divided into three types: the source of the sample data, the construction of feature vectors and the determination of the sample size.

The sample data in materials science usually originate from computational simulations and experimental measurements, are collected by different research institutions or schools and lack a system of centralized management. The development of materials data infrastructures (as mentioned in section 1) has alleviated this problem, although each database is separate and not unified in data format, which still limits the applicability of machine learning.

Feature vectors, which largely determine the accuracy of model prediction, are critical. Ideally, the feature vectors should provide a simple physical basis for the extraction of major structural and chemical trends and thus enable rapid predictions of new material chemistries. The most commonly used feature vectors in materials research mainly include the composition, structure, electron density and Coulomb matrix. Since each eigenvector is intended for a specific application, no unified eigenvector exists that is effective for all applications in materials research. Hansen et al. [101] employed different kinds of Coulomb matrices as feature vectors in the prediction of molecular atomization energies and demonstrated that different representations greatly influence the forecasting results. Schütt et al. [110] also emphasized the importance of materials representation throughout the entire problem-solving process based on several experiments. Ghiringhelli et al. [139] analyzed the critical role of feature vectors, employed the Least Absolute Shrinkage and Selection Operator (LASSO) technique for feature selection and revealed how a meaningful descriptor can be systematically found. Balachandran et al. [140] used PCA combined with DT and SVM methods to uncover the functional forms that mimic the most frequently used features and provided a mathematical basis for feature set construction without a priori assumptions. Furthermore, these authors applied the proposed method to study two broad classes of materials, namely, (1) wide-band-gap *AB* compounds and (2) rare earth-main group RM intermetallics, and demonstrated that the proposed method can be used to rapidly design new ductile materials.

The determination of sample size is also a key factor during sample construction that is related to dimension reduction in machine learning. The sample size determines whether the sample data include the implied information about the inherent laws governing the sample, which strongly depends on the research project and the machine learning method chosen. Given that some methods with few parameters and low complexity, such as the SVM approach [88], can perform well when the sample size is small, complicated models such as ANNs [51,101] can also achieve high prediction accuracy on high-quality sample data regardless of the sample size. For instance, Liu et al. [88] applied SVR to predict the $T_g$ of $Ge_xSe_{1-x}$ glass based on 9 samples, and Rao and Mukherjee [51] used only 14 samples for the prediction of the shear strength and percentage reinforcement of a ceramic-matrix composite. By contrast, Hansen et al. [101] used 7165 samples for the prediction of molecular atomization. Note that a good prediction effect was achieved on both types of samples, although the question of how to determine a research-topic-matched sample size is always one of the difficulties that must be faced.

### 6.2. Generalization ability

The ability to correctly predict new examples that differ from those used for training is known as generalization ability [52], which is one of the important evaluation criteria for machine learning. The generalization ability of a model $\widehat{f}$ in SLT is usually expressed in terms of the generalization error, which is given in Eq. (6).

$$R_{\exp}\left(\widehat{f}\right) = E_p\left[L\left(Y,\widehat{f}(x)\right)\right] = \int xyL\left(y,\widehat{f}(x)\right)P(x,y)dxdy \tag{6}$$

where $x$ is the conditional attribute, $y$ is the decision attribute, $L$ is the loss function $\widehat{f}(x)$ represents the prediction value generated by the model, and $P(x,y)$ is the joint probability distribution of $x$ and $y$.

Improving the generalization ability of a model, which means minimizing the structure risk of the model (Eq. (7)), is the universal goal of machine learning.

$$R_{sm}\left(\widehat{f}\right) = \frac{1}{N}\sum\nolimits_{i=1}^{N} L\left(y_i,\widehat{f}(x_i)\right) + \lambda J\left(\widehat{f}\right) \tag{7}$$

where $N$ is the total sample size, $J(\widehat{f})$ represents the complexity of the model and is a functional defined in the hypothesis space F, and $\lambda$ is a coefficient used to weight the empirical risk and the model complexity. The more complex the model $\widehat{f}$ is, the greater the complexity $J(\widehat{f})$ is.

As shown in Fig. 10, the tendencies of the training and prediction errors with increasing model complexity are different from each other. The generalization ability of a trained model is often related to the issues of under-fitting and/or over-fitting. Under-fitting means that the sample data are insufficient or the learning algorithm is not suitable, whereas over-fitting generally occurs when a model is excessively complex, such as having too many parameters relative to the number of observations. The generation abilities achieved in different applications of materials discovery and design vary widely. The *RMSE* value can vary over a large range, for instance, from 0.004467 [52] (for a model with a good generalization ability) to 20.195 [14] (for one with a poor generalization ability). Therefore, the question of how to improve the generalization ability of a model is an urgent problem that must be addressed, for which the sample quality, sample size and training algorithm should all be taken into consideration.

### 6.3. Understandability

At present, most machine learning models are treated as a “black box”, which means that the knowledge extracted by such a model is difficult to understand. For example, when an SVM model is used to solve a classification or regression problem, the optimal classification plane and/or the parameters of the fitted curve obtained through training are invisible and hidden in the model. The intelligibility of knowledge representation is one of the important metrics for evaluating a learning algorithm. In most areas, a machine learning model is expected to be intelligible because it will tend to be treated as a model with intelligible patterns and rules. Applications of machine learning in materials research similarly require models with good understandability. In the early days of applying machine learning to predict the behavior of materials, symbolic machine learning methods with good intelligibility were used. However, with the development of statistical learning methods, a problem of poor intelligibility arises. Therefore, the question of how to turn a “black box” into a “white box” and improve the intelligibility of the model is currently a problem that needs to be solved immediately. The most commonly used methods for solving problems of this kind are as follows: 1) Attempt to develop a more intelligible algorithm and avoid using algorithms with poor intelligibility. In Ref. [141], Yang et al. proposed a method for studying the explanatory capacity of ANNs, and thus, the “black box” problem was successfully overcome. 2) Extract knowledge from the results of a poorly intelligible algorithm. For instance, the knowledge hidden in a neural network can be clearly expressed in an intelligible manner. Zhou [142] presented a critique of the prevailing rule quality evaluation framework in the context of rule extraction from neural networks, known as the FACC framework, arguing that two different goals of rule extraction have been confused. Liu et al. [143] proposed a new method called the impulse-force-based ART (IFART) neural network for earthquake prediction and extracted IF-THEN rules from the trained IFART neural network based on its architecture. As a result, both the prediction accuracy and reliability of earthquake prediction were improved.

### 6.4. Usability

Usability is the degree of complexity of using machine learning methods to solve practical problems. The complexity of applying machine learning in materials science manifests in two aspects. 1) The machine learning process is complex and cannot be performed

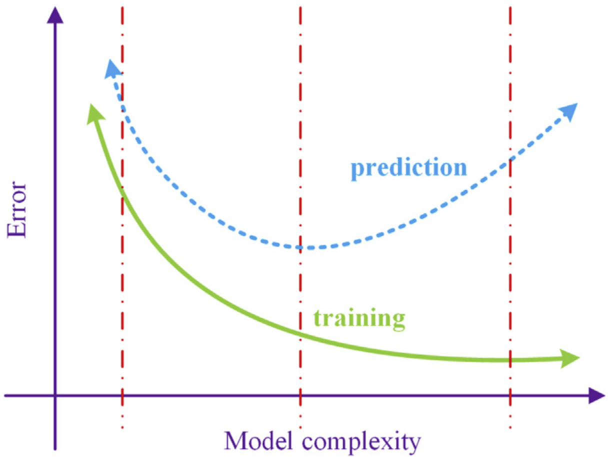


**Fig. 10.** Generalization ability of a machine learning model.

without professional knowledge and instructions. For example, dimension reduction and correlation analysis should be applied to increase the prediction accuracy of a model when using machine learning for material property prediction. Ref. [115] reported a study on crystal structure prediction in which PCA was used to reduce the high dimensionality of the problem due to the high dimensionality of the sample, which helped to improve the prediction accuracy. Ref. [105] used a conditional attribute correlation analysis to explain the prediction results for the properties of organic polymer materials. 2) The determination of parameters is also a complex task. Because machine learning methods are very sensitive to these parameters and kernel functions, parameter determination is a key step in the machine learning process. The parameters of the machine learning methods used in materials science are largely determined through manual adjustment or based on experience. In addition, some optimization algorithms are adopted to optimize these parameters. For instance, Pei et al. [91] utilized a PSO algorithm to optimize SVM and ANN parameters. Therefore, determining how to improve the usability of machine learning methods is a problem that urgently needs to be solved.

### 6.5. *Learning efficiency*

The speed of machine learning is directly related to its practical application. Although high speeds are always pursued in model training and testing, it is impossible to achieve both simultaneously. For instance, the KNN method features a high training speed but a low testing speed, whereas neural network models have a low training speed but a high testing speed. Currently, the problem of learning efficiency is not of great concern in machine learning applications in materials science because the sample sizes for these machine learning applications are quite small, ranging from dozens to thousands. However, with the advancement of materials genome projects in countries around the world, materials science will enter the era of “big data”, and the data volume will become enormous, which will pose tremendous challenges in learning efficiency. Therefore, the question of how to improve the learning efficiency of machine learning will also become a problem that urgently needs to be solved. To this end, we will need to investigate the possibility of adopting high-performance computing methods, such as parallel computing and cloud computing, in this field.

## 7. Concluding remarks

As a branch of artificial intelligence and one of the hottest families of analysis techniques, machine learning is an important means through which computers can acquire knowledge. The various applications of machine learning span several hot topics in materials science, including new materials discovery, material property prediction and other purposes ranging from the macroscopic to the microscopic scale. The objects of the application of machine learning in materials science are very extensive, including inorganic oxide materials, electrolyte materials, metallic materials, and functional materials. The broad variety of related studies demonstrates that machine learning can be used to develop efficient and accurate tools for materials science. With the continuous development of theories and methods, the topics to which machine learning can be applied in materials science will become broader. Meanwhile, machine learning methods themselves are also worthy of further study.

- Machine learning is applied in materials design and discovery mainly to solve problems of regression, classification, clustering and probability estimation. In addition, machine learning also exhibits good performance in solving problems involving correlation, sorting, and so forth. Therefore, the application of machine learning methods can still be expanded to solve more problems in materials science and will likely enable even greater developments.
- Applications of machine learning in materials science usually involve the use of one specific method, such as an SVM, ANN and DT approach, for one kind of material or the comparison of the results of multiple machine learning methods to select the most suitable method for a given problem. This makes the application range of any particular model very limited. Therefore, if one could develop a unified framework and apply it to multiple problem-solving strategies, the applicability of machine learning methods in materials science would be greatly enhanced, and the efficiency and generalization ability of the trained models could also be improved.
- Big data is a hot topic at present and is attracting extensive attention in various fields. The questions of how to store, manage and analyze high-volume data are challenging problems that need to be solved, in materials science research as well as other fields. Therefore, investigating the applications of machine learning in materials science against the background of big data is expected to be a crucial research direction for the future. In particular, deep learning has performed very well in the processing of large volumes of data and has enabled considerable breakthroughs in the fields of image processing, speech recognition and so forth. Consequently, it is worth considering the adoption of deep learning methods for intelligent big data analysis in materials science research.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. U1630134, 51622207 and 51372228), the National Key Research and Development Program of China (Grant Nos. 2017YFB0701600 and 2017YFB0701500), the Shanghai Institute of Materials Genome from the Shanghai Municipal Science and Technology Commission (Grant No. 14DZ2261200), the Shanghai Municipal Education Commission (Grant No. 14ZZ099) and the Natural Science Foundation of Shanghai (Grant No. 16ZR1411200).

## References

[1] Nantasenamat C, Isarankura-Na-Ayudhya C, Naenna T, Prachayasittikul V. A practical overview of quantitative structure-activity relationship. Excli J 2009;8:74–88.
[2] Rajan K. Materials informatics. Mater TodCay 2005;8:38–45.
[3] About J. Materials genome initiative for global competitiveness. USA: National Science and Technology Council; 2011.
[4] Hohenberg P, Kohn W. Inhomogeneous electron gas. Phys Rev 1964;136:864–71.
[5] Kohn W, Sham LJ. Self-consistent equations including exchange and correlation effects. Phys Rev 1965;140:1133–8.
[6] Alder BJ, Wainwright TE. Studies in molecular dynamics. I. General method. J Chem Phys 1959;31(2):459–66.
[7] Rahman A. Correlations in the motion of atoms in liquid argon. Phys Rev 1964;136:405–11.
[8] Binder K, Baumgärtner A. The Monte Carlo method in condensed matter physics, vol. 71. Springer-Verlag; 1995 (2).
[9] Chen LQ. Phase-field models for microstructure evolution. Annu Rev Mater Res 2002;32:113–40.
[10] Steinbach I. Phase-field models in materials science. Modell Simul Mater Sci Eng 2009;17:073001.
[11] Boettinger WJ, Warren JA, Beckermann C, Karma A. Phase-field simulation of solidification. Annu Rev Mater Res 2002;32:163–94.
[12] Olson GB. Designing a new material world. Science 2000;288:993–8.
[13] Camacho-Zuñiga C, Ruiz-Treviño FA. A new group contribution scheme to estimate the glass transition temperature for polymers and diluents. Ind Eng Chem Res 2003;42:1530–4.
[14] Yu XL, Yi B, Wang XY. Prediction of the glass transition temperatures for

polymers with artificial neural network. J Theor Comput Chem 2008;7: 953–63.

[15] https://www.mgi.gov.

[16] Belsky A, Hellenbrandt M, Karen VL, Luksch P. New developments in the inorganic crystal structure database (ICSD): accessibility in support of materials research and design. Acta Crystallogr, Sect B Struct Sci 2002;58: 364–9.

[17] http://supercon.nims.go.jp.

[18] Kirklin S, Saal JE, Meredig B, Thompson A, Doak JW, Aykol M, et al. The open quantum materials database (OQMD): assessing the accuracy of DFT formation energies. npj Comput Mater 2015;1:15010.

[19] Allen FH. The cambridge structural database: a quarter of a million crystal structures and rising. Acta Crystallogr, Sect B Struct Sci 2002;58:380–8.

[20] Hachmann J, Olivaresamaya R, Atahanevrenk S, Amadorbedolla C, Sanchezcarrera RS, Goldparker A, et al. The Harvard clean energy project. large-scale computational screening and design of molecular motifs for organic photovoltaics on the world community grid. J Phys Chem Lett 2011;2:2241–51.

[21] Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. Apl Mater 2013;1(1):011002.

[22] Puchala B, Tarcea G, Marquis EA, Hedstrom M, Jagadish HV, Allison JE. The materials commons: a collaboration platform and information repository for the global materials community. JOM 2016;68(8):2035–44.

[23] Blaiszik B, Chard K, Pruyne J, Ananthakrishnan R, Tuecke S, Foster I. The materials data facility: data sservices to advance materials science research. JOM 2016;68(8):2045–52.

[24] Kusne AG, Gao T, Mehta A, Ke L, Nguyen MC, Ho K-M, et al. On-the-fly machine-learning for high-throughput experiments: search for rare-earth-free permanent magnets. Sci Rep 2014;4:6367.

[25] Murphy KP. Machine learning: a probabilistic perspective, vol. 58. MIT Press; 2012. p. 27–71.

[26] Mueller T, Kusne AG, Ramprasad R. Machine learning in materials science: recent progress and emerging applications. Rev Comput Chem 2016;29:186.

[27] Ghiringhelli LM, Vybiral J, Levchenko SV, Draxl C, Scheffler M. Big data of materials science: critical role of the descriptor. Phys Rev Lett 2015;114(10): 105503.

[28] Raccuglia P, Elbert KC, Adler PDF, Falk C, Wenny MB, Mollo A, et al. Machine-learning-assisted materials discovery using failed experiments. Nature 2016;533:73–6.

[29] Chen L-Q, Chen L-D, Kalinin SV, Klimeck G, Kumar SK, Neugebauer J, et al. Design and discovery of materials guided by theory and computation. npj Comput Mater 2015;1:15007.

[30] Hautier G, Jain A, Ong SP. From the computer to the laboratory: materials discovery and design using first-principles calculations. J Mater. Sci 2012;47: 7317–40.

[31] Sumpter BG, Vasudevan RK, Potok T, Kalinin SV. A bridge for accelerating materials by design. npj Comput Mater 2015;1:15008.

[32] Hautier G, Fischer CC, Jain A, Mueller T, Ceder G. Finding nature's missing ternary oxide compounds using machine learning and density functional theory. Chem Mater 2010;22:3762–7.

[33] Meredig B, Agrawal A, Kirklin S, Saal JE, Doak JW, Thompson A, et al. Combinatorial screening for new materials in unconstrained composition space with machine learning. Phys Rev B 2014;89:094104.

[34] Gómez-bombarelli R, Aguilera-Iparraguirre J, Hirzel TD, Duvenaud D, Maclaurin D, Blood-Forsythe MA, et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. Nat Mater 2016;15:1120–8.

[36] Agrawal A, Choudhary A. Perspective: materials informatics and big data: realization of the “fourth paradigm” of science in materials science. Apl Mater 2016;4(5):1–17.

[37] Hill J, Mulholland G, Persson K, Seshadri R, Wolverton C, Meredig B. Materials science with large-scale data and informatics: unlocking new opportunities. MRS Bull 2016;41(5):399–409.

[38] Jain A, Hautier G, Ong SP, Persson K. New opportunities for materials informatics: resources and data mining techniques for uncovering hidden relationships. J Mater. Res 2016;31(8):977–94.

[39] Kalidindi SR, Medford AJ, Mcdowell DL. Vision for data and informatics in the future materials innovation ecosystem. JOM 2016;68(8):2126–37.

[40] Ward L, Wolverton C. Atomistic calculations and materials informatics: a review. Curr Opin Solid State Mater. Sci 2016;21(3):167–76.

[41] Russell SJ, Norvig P. Artificial intelligence: a modern approach, second edition. Pearson Educ 2003;263(5):2829–33.

[42] Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev 1958;65:386–408.

[43] Vahed A, Omlin CW. Rule extraction from recurrent neural networks using a symbolic machine learning algorithm, ICONIP. IEEE; 1999. p. 712–7.

[44] Vapnik V. The nature of statistical learning theory. Springer science & business media; 2013.

[45] De'ath G, Fabricius KE. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecol 2000;81(11):3178–92.

[46] http://www.sas.com/en_us/home.html.

[47] Joze HRV, Drew MS. Improved machine learning for image category recognition by local color constancy. IEEE; 2010. p. 3881–4.

[48] Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, et al. Machine learning in bioformatics. Briefings Bioinforma 2005;7:86–112.

[49] Eminagaoglu M, Eren S. Implementation and comparison of machine learning classifiers for information security risk analysis of a human resources department. In: International conference on computer information systems and industrial management applications, vol. 3; 2011. p. 391–8.

[50] Olsson F. A literature survey of active machine learning in the context of natural language processing. Swedish Institute of Computer Science; 2009.

[51] Rao HS, Mukherjee A. Artificial neural networks for predicting the macro mechanical behaviour of ceramic-matrix composites. Comput Mater Sci 1996;5:307–22.

[52] Reich Y, Travitzky N. Machine learning of material behaviour knowledge from empirical data. Mater Des 1996;16:251–9.

[53] Li CH, Guo J, Qin P, Chen RL, Chen NY. Some regularities of melting points of *AB*-type intermetallic compounds. J Phys Chem Solids 1996;57:1797–802.

[54] Mitchell TM. Machine learning and data mining. Commun Acm 1999;42: 31–6.

[55] Peck RL, Olsen C, Devore JL. Introduction to statistics and data analysis. Cengage Learning; 2015.

[56] Wu SM. A review on coarse warranty data and analysis. Reliab Eng Syst Saf 2013;114:1–11.

[57] Kohavi R, Provost F. Glossary of terms. Mach Learn 1998;30:271–4.

[58] Fujimur K, Seko A, Koyama Y, Kuwabara A, Kishida I, Shitara K, et al. Accelerated materials design of lithium superionic conductors based on first principles calculations and machine learning algorithms. Adv Energy Mater 2013;3:980–5.

[59] Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157–82.

[60] Hutchinson JM. Determination of the glass transition temperature. J Therm Anal Calorim 2009;98:579–89.

[61] Efron B, Tibshirani RJ. An introduction to the bootstrap. CRC press; 1994.

[62] Bishop CM. Pattern recognition and machine learning. New York: Springer; 2007.

[63] Pei JF, Cai CZ, Zhu YM, Yan B. Modeling and predicting the glass transition temperature of polymethacrylates based on quantum chemical descriptors by using hybrid PSO-SVR. Macromol Theor Simul 2013;22:52–60.

[64] Fang SF, Wang MP, Qi WH, Zheng F. Hybrid genetic algorithms and support vector regression in forecasting atmospheric corrosion of metallic materials. Comput Mater Sci 2008;44:647–55.

[65] Paszkowicz W, Harris KDM, Johnston RL. Genetic algorithms: a universal tool for solving computational tasks in materials science. Comput Mater Sci 2009;45:ix–x.

[66] Zhang XJ, Chen KZ, Feng XA. Material selection using an improved genetic algorithm for material design of components made of a multiphase material. Mater Des 2008;29:972–81.

[67] Mohn CE, Kob W. A genetic algorithm for the atomistic design and global optimisation of substitutionally disordered materials. Comput Mater Sci 2009;45:111–7.

[68] Ning X, Walters M, Karypisxy G. Improved machine learning models for predicting selective compounds. J Chem Inf Model 2012;52:38–50.

[69] Isayev O, Oses C, Toher C, Gossett E, Curtarolo S, Tropsha A. Universal fragment descriptors for predicting properties of inorganic crystals. Nat Commun 2017;8:15679.

[70] Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat 2001;29:1189–232.

[71] Carrera GVSM, Branco LC, Aires-de-Sousa J, Afonso CAM. Exploration of quantitative structure-property relationships (QSPR) for the design of new guanidinium ionic liquids. Tetrahedron 2008;64:2216–24.

[72] Bertinetto C, Duce C, Micheli A, Solaro R, Starita A, Tiné MR. Evaluation of hierarchical structured representations for QSPR studies of small molecules and polymers by recursive neural networks. J Mol Graph Model 2009;27: 797–802.

[73] Wang SC. Artificial neural network, interdisciplinary computing in Java programming. Kluwer Academic Publishers; 2003. p. 81–100.

[74] Guo Z, Malinov S, Sha W. Modelling beta transus temperature of titanium alloys using artificial neural network. Comput Mater. Sci 2005;32:1–12.

[75] Altun F, Kişi Özgür, Aydin K. Predicting the compressive strength of steel fiber added lightweight concrete using neural network. Comput Mater. Sci 2008;42:259–65.

[76] Topçu B, Sarıdemir M. Prediction of properties of waste AAC aggregate concrete using artificial neural network. Comput Mater Sci 2007;41:117–25.

[77] Chen X, Sztera L, Cartwright HM. A neural network approach to prediction of glass transition temperature of polymers. Int J Intell Syst 2008;23:22–32.

[78] Gajewski J, Sadowski T. Sensitivity analysis of crack propagation in pavement bituminous layered structures using a hybrid system integrating artificial neural networks and finite element method. Comput Mater Sci 2014;82: 114–7.

[79] Salahinejad M, Le TC, Winkler DA. Capturing the crystal: prediction of enthalpy of sublimation, crystal lattice energy, and melting points of organic compounds. J Chem Inf Model 2013;53:223–9.

[80] Saldana DA, Starck L, Mougin P, Rousseau B, Ferrando N, Creton B. Prediction of density and viscosity of biofuel compounds using machine learning methods. Energy Fuels 2012;26:2416–26.

[81] Häse F, Valleau S, Pyzer-Knapp E, Aspuru-Guzik A. Machine learning exciton dynamics. Chem Sci 2016;7(8):5139–47.

[82] Wu H, Lorenson A, Anderson B, Witteman L, Wu HT, Meredig B, et al. Robust

FCC solute diffusion predictions from ab-initio machine learning methods. Comput Mater Sci 2017;134:160–5.

[83] Scott DJ, Coveney PV, Kilner JA, Rossiny JCH, Alford N Mc N. Prediction of the functional properties of ceramic materials from composition using artificial neural networks. J Eur Ceram Soc 2007;27:4425–35.

[84] Raj RE, Daniel BSS. Prediction of compressive properties of closed-cell aluminum foam using artificial neural network. Comput Mater Sci 2008;43:767–73.

[85] Sivasankaran S, Narayanasamy R, Ramesh T, Prabhakar M. Analysis of workability behavior of Al-SiC P/M composites using backpropagation neural network model and statistical technique. Comput Mater Sci 2009;47:46–59.

[86] Cavaliere P. Flow curve prediction of an Al-MMC under hot working conditions using neural networks. Comput Mater. Sci 2007;38:722–6.

[87] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20:273–97.

[88] Liu Y, Zhao TL, Yang G, Ju WW, Shi SQ. A feature selection based two-stage support vector regression method for forecasting the transition temperature ($T_g$) of $Ge_xSe_{1-x}$ glass. Comput Mater. Sci Submitt 2017.

[89] Chen BT, Chang TP, Shih JY, Wang JJ. Estimation of exposed temperature for fire-damaged concrete using support vector machine. Comput Mater Sci 2009;44:913–20.

[90] Liu X, Lu WC, Peng CR, Su Q, Guo J. Two semi-empirical approaches for the prediction of oxide ionic conductivities in $ABO_3$ perovskites. Comput Mater Sci 2009;46:860–8.

[91] Pei JF, Cai CZ, Zhu YM. Modeling and predicting the glass transition temperature of vinyl polymers by using hybrid PSO-SVR method. J Theor Comput Chem 2013;12:1350002.

[92] Gharagheizi F, Ilani-Kashkouli P, Mohammadi AH. A group contribution method for estimation of glass transition temperature ionic liquids. Chem Eng Sci 2012;81:91–105.

[93] Ahmad A, Amjad A, Denny M, Bergström CAS. Experimental and computational prediction of glass transition temperature of drugs. J Chem Inf Model 2014;54:3396–403.

[94] Yu K, Cheng YY. Machine learning techniques for the prediction of the peptide mobility in capillary zone electrophoresis. Talanta 2007;71:676–82.

[95] Helma C, Cramer T, Kramer S, DeRaedt L. Data mining and machine learning techniques for the identification of mutagenicity inducing substructures and structure activity relationships of noncongeneric compounds. J Chem Inf Comput Sci 2004;44:1402–11.

[96] Li CH, Thing YH, Zeng YZ, Wang CM, Wu P. Prediction of lattice constant in perovskites of $GdFeO_3$ structure. J Phys Chem Solids 2003;64:2147–56.

[97] Javed SG, Khan A, Majid A, Mirza AM, Bashir J. Lattice constant prediction of orthorhombic $ABO_3$ perovskites using support vector machines. Comput Mater. Sci 2007;39:627–34.

[98] Majid A, Khan A, Javed G, Mirza AM. Lattice constant prediction of cubic and monoclinic perovskites using neural networks and support vector regression. Comput Mater Sci 2010;50:363–72.

[99] Majid A, Khan A, Choi T. Predicting lattice constant of complex cubic perovskites using computational intelligence. Comput Mater Sci 2011;50:1879–88.

[100] Lufaso MW, Woodward PM. Prediction of the crystal structures of perovskites using the software program SpuDS. Acta Crystallogra, Sect B Struct Sci 2001;57:725–38.

[101] Hansen K, Montavon G, Biegler F, Fazli S, Rupp M, Scheffler M, et al. Assessment and validation of machine learning methods for predicting molecular atomization energies. J Chem Theory Comput 2013;9:3404–19.

[102] Liu RQ, Yabansu CY, Agrawal A, Kalidindi SR, Choudhary AN. Machine learning approaches for elastic localization linkages in high-contrast composite materials. Integrating Mater Manuf Innov 2015;4(1):13.

[103] Liu RQ, Yabansu CY, Yang ZJ, Choudhary AN, Kalidindi SR, Agrawal A. Context aware machine learning approaches for modeling elastic localization in three-dimensional composite microstructures. Integrating Mater Manuf Innov 2017:1–12.

[104] Ward L, Agrawal A, Choudhary A, Wolverton C. A general-purpose machine learning framework for predicting properties of inorganic materials. npj Comput Mater 2016;2:16028.

[105] Pilania G, Wang C, Jiang X, Rajasekaran S, Ramprasad R. Accelerating materials property predictions using machine learning. Sci Rep 2013;3:2810.

[106] Chen C, Lu Z, Ciucci F. Data mining of molecular dynamics data reveals Li diffusion characteristics in garnet $Li_7La_3Zr_2O_{12}$. Sci Rep 2017;7:40769.

[107] Wang X, Xiao R, Li H, Chen LQ. Quantitative structure-property relationship study of cathode volume changes in lithium ion batteries using ab-initio and partial least squares analysis. J Materiomics 2 March 2017. http://dx.doi.org/10.1016/j.jmat.2017.02.002.

[108] Castin N, Fernández JR, Pasianot RC. Predicting vacancy migration energies in lattice-free environments using artificial neural networks. Comput Mater. Sci 2014;84:217–25.

[109] Anatole von Lilienfeld O, Ramakrishnan R, Rupp M, Knoll A. Fourier series of atomic radial distribution functions: a molecular fingerprint for machine learning models of quantum chemical properties. Int J Quantum Chem 2015;115:1084–93.

[110] Schütt KT, Glawe H, Brockherde F, Sanna A, Müller KR, Gross EKU. How to represent crystal structures for machine learning: towards fast prediction of electronic properties. Phys Rev B 2014;89:205118.

[111] Ju WW. Research on materials properties prediction based on machine learning method. Master Degree Thesis. Shanghai University; 2016.

[112] Farrusseng D, Clerc F, Mirodatos C, Rakotomalala R. Virtual screening of materials using neuro-genetic approach: concepts and implementation. Comput Mater. Sci 2009;45:52–9.

[113] Beran GJO. A new era for ab initio molecular crystal lattice energy prediction. Angew Chem Int Ed 2015;54:396–8.

[114] Maddox J. Crystals from first principles. Nature 1988;335:201.

[115] Curtarolo S, Morgan D, Persson K, Rodgers J, Ceder G. Predicting crystal structures with data mining of quantum calculations. Phys Rev Lett 2003;91:135503.

[116] Ceder G, Morgan D, Fischer C, Tibbetts K, Curtarolo S. Data-mining-driven quantum mechanics for the prediction of structure. MRS Bull 2006;31:981–5.

[117] Fischer CC, Tibbetts KJ, Dane M, Ceder G. Predicting crystal structure by merging data mining with quantum mechanics. Nat Mater 2006;5:641–6.

[118] Phillips CL, Voth GA. Discovering crystals using shape matching and machine learning. Soft Matter 2013;9:8552–68.

[119] Liu R, Kumar A, Chen Z, Agrawal A, Sundararaghavan V, Choudhary A. A predictive machine learning approach for microstructure optimization and materials design. Sci Rep 2015;5:11551.

[120] Roekeghem AV, Carrete J, Oses C, Curtarolo S, Mingo N. High throughput thermal conductivity of high temperature solid phases: the case of oxide and fluoride perovskites. 2016.

[121] Sendek AD, Yang Q, Cubuk ED, Duerloo KD, Cui Y, Reed EJ. Holistic computational structure screening of more than 12000 candidates for solid lithium-ion conductor materials. Energy Environ Sci 2017;10(1):306–20.

[122] Bolstad WM. Introduction to Bayesian statistics. John Wiley & Sons; 2004.

[123] Hautier G, Fischer C, Ehrlacher V, Jain A, Ceder G. Data mined ionic substitutions for the discovery of new compounds. Inorg Chem 2011;50:656–63.

[124] Han YF, Zeng WD, Shu Y, Zhou YG, Yu HQ. Prediction of the mechanical properties of forged Ti-10V-2Fe-3Al titanium alloy using FNN. Comput Mater Sci 2011;50:1009–15.

[125] Abbod MF, Linkens DA, Zhu Q, Mahfouf M. Physically based and neuro-fuzzy hybrid modelling of thermomechanical processing of aluminium alloys. Mater Sci Eng A 2002;333:397–408.

[126] Zhu Q, Abbod MF, Talamantes-Silva J, Sellars CM, Linkens DA, Beynon JH. Hybrid modelling of aluminium-magnesium alloys during thermomechanical processing in terms of physically-based, neuro-fuzzy and finite element models. Acta Mater 2003;51:5051–62.

[127] Fang SF, Wang MP, Song M. An approach for the aging process optimization of Al-Zn-Mg-Cu series alloys. Mater Des 2009;30:2460–7.

[128] Snyder JC, Rupp M, Hansen K, Muller KR, Burke K. Finding density functionals with machine learning. Phys Rev Lett 2012;108:253002.

[129] Shi SQ, Gao J, Liu Y, Zhao Y, Wu Q, Ju WW, et al. Multi-scale computation methods: their applications in lithium-ion battery research and development. Chin Phys B 2016;25:018212.

[130] Li GC, Wang HY, Yu ZL. New method for estimation modeling of SOC of battery. World Congr Softw Eng 2009;2:387–90.

[131] Lee DT, Shiah SJ, Lee CM, Wang YC. State-of-charge estimation for electric scooters by using learning mechanisms. IEEE Trans Veh Technol 2007;56:544–56.

[132] Fleischer C, Waag W, Bai Z, Sauer DU. Adaptive on-line state-of-available-power prediction of lithium-ion batteries. J Power Electron 2013;13:516–27.

[133] Klass V, Behm M, Lindbergh G. A support vector machine-based state-of-health estimation method for lithium-ion batteries under electric vehicle operation. J Power Sourc 2014;270:262–72.

[134] Charkhgard M, Farrokhi M. State-of-charge estimation for lithium-ion batteries using neural networks and EKF. IEEE Trans Ind Electron 2010;57:4178–87.

[135] Sundararaghavan V, Zabaras N. Classification and reconstruction of three-dimensional microstructures using support vector machines. Comput Mater Sci 2005;32:223–39.

[136] Addin O, Sapuan SM, Mahdi E, Othman M. A Naïve-Bayes classifier for damage detection in engineering materials. Mater Des 2007;28:2379–86.

[137] Martinez RF, Okariz A, Ibarretxe J, Iturrondobeitia M, Guraya T. Use of decision tree models based on evolutionary algorithms for the morphological classification of reinforcing nano-particle aggregates. Comput Mater Sci 2014;92:102–13.

[138] Fernandez M, Boyd PG, Daff TD, Aghaji MZ, Woo TK. Rapid and accurate machine learning recognition of high performing metal organic frameworks for $CO_2$ capture. J Phys Chem Lett 2014;17:3056–60.

[139] Ghiringhelli 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.

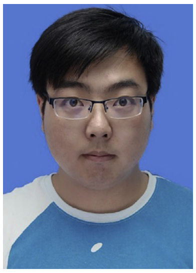
[140] Balachandran PV, Theiler J, Rondinelli JM, Lookman T. Materials prediction via classification learning. Sci Rep 2015;5.

[141] Yang L, Wang P, Jiang Y, Jian C. Studying the explanatory capacity of artificial neural networks for understanding environmental chemical quantitative structure-activity relationship models. J Chem Inf Model 2005;45:1804–11.

[142] Zhou ZH. Rule extraction: using neural networks or for neural networks? J Comput Sci Technol 2004;19:249–53.

[143] Liu Y, Liu H, Zhang BF, Wu GF. Extraction of if-then rules from trained neural network and its application to earthquake prediction. In: IEEE International Conference on Cognitive Informatics, vol. 56; 2004. p. 37–44.

Y. Liu obtained her B.S. and M.S. in computer science from Jiangxi Normal University in 1997 and 2000. She finished her Ph.D. in control theory and control engineering from Shanghai University (SHU) in 2005. She has been working with the School of Computer Engineering and Science of SHU since July 2000. During that time, she has been a curriculum R&D manager at the Sybase-SHU IT Institute of Sybase Inc. from July 2003 to July 2004 and a visiting scholar at the University of Melbourne from Sep. 2012 to Sep. 2013. Her current main research interests are focused on machine learning and its applications in materials science and demand forecasting.



T. Zhao is a graduate candidate in the School of Computer Engineering and Science, Shanghai University, China. He received his Bachelor of Engineering degree in computer science from the School of Computer and Software, Nanjing University of Information Science & Technology, China, in 2015. His main research interests are focused on machine learning for predicting the properties of lithium-ion batteries.

W. J. received his B.S. in computer science from Anhui Normal University in 2013. He finished his M.S. in computer science from Shanghai University in 2016. His main research interests are focused on machine learning for predicting the properties of lithium-ion batteries.

S. S. obtained his B.S. from Jiangxi Normal University in 1998. He finished his Ph.D. from the Institute of Physics, Chinese Academy of Sciences, in 2004. After that, he joined the National Institute of Advanced Industrial Science and Technology of Japan and Brown University in the USA as a senior research associate, where he remained until joining Shanghai University as a professor in early 2013. His research interests are focused on the fundamentals and microscopic design of energy storage and conversion materials related to lithium-ion batteries and CeO2-based solid-state oxide fuel cells.