Received: 14 June 2019 | Revised: 10 July 2019 | Accepted: 14 July 2019

DOI: 10.1002/inf2.12028

REVIEW ARTICLE

WILEY

Machine learning in materials science

Jing Wei[1] | **Xuan Chu**[1] | **Xiang-Yu Sun**[1] | **Kun Xu**[1] | **Hui-Xiong Deng**[2] | **Jigen Chen**[3] | **Zhongming Wei**[2,4] | **Ming Lei**[1]

[1]State Key Laboratory of Information Photonics and Optical Communications, Beijing University of Posts and Telecommunications, Beijing, China

[2]State Key Laboratory of Superlattices and Microstructures, Institute of Semiconductors, Chinese Academy of Sciences, Center of Materials Science and Optoelectronics Engineering, University of Chinese Academy of Sciences, Beijing, China

[3]Zhejiang Provincial Key Laboratory for Cutting Tools, Taizhou University, Taizhou, China

[4]Beijing Academy of Quantum Information Sciences, Beijing, China

Correspondence
Zhongming Wei, State Key Laboratory of Superlattices and Microstructures, Institute of Semiconductors, Chinese Academy of Sciences, Center of Materials Science and Optoelectronics Engineering, University of Chinese Academy of Sciences, Beijing 100083, China.
Email: zmwei@semi.ac.cn

Ming Lei, State Key Laboratory of Information Photonics and Optical Communications, Beijing University of Posts and Telecommunications, Beijing 100876, China.
Email: mlei@bupt.edu.cn

Funding information
China Postdoctoral Science Foundation, Grant/Award Number: 2017M620694; National Natural Science Foundation of China, Grant/Award Number: 61622406 61571415; Beijing Academy of Quantum Information Sciences, Grant/Award Number: Y18G04; Strategic Priority Research Program of Chinese Academy of Sciences, Grant/Award Numbers: XDB30000000, 2016YFB0700700, 2017YFA0207500; National Key Research and Development Program of China; National Postdoctoral Program for Innovative Talents, Grant/Award Number: BX201700040

Abstract

Traditional methods of discovering new materials, such as the empirical trial and error method and the density functional theory (DFT)-based method, are unable to keep pace with the development of materials science today due to their long development cycles, low efficiency, and high costs. Accordingly, due to its low computational cost and short development cycle, machine learning is coupled with powerful data processing and high prediction performance and is being widely used in material detection, material analysis, and material design. In this article, we discuss the basic operational procedures in analyzing material properties via machine learning, summarize recent applications of machine learning algorithms to several mature fields in materials science, and discuss the improvements that are required for wide-ranging application.

KEYWORDS

data processing, deep learning, machine learning, modeling, validation

1 | INTRODUCTION

The guiding ideology of materials science can be summarized in four paradigms[1]: the first paradigm is the empirical trial and error method, the second paradigm is physical and chemical laws, the third paradigm is computer simulation, and the fourth paradigm is big data-driven science. Among them, with the continuous development of data mining technology and

artificial intelligence, the fourth paradigm can perfectly unify the other three paradigms in the aspects of theory, experiment, and computer simulation. New methods that are based on big data, such as machine learning, have emerged in and from the study of materials science.

Rapid developments in information science, energy, national defense, and other fields have imposed crucial and diverse requirements for materials. However, traditional methods for discovering new materials, such as the empirical trial and error method and the density functional theory (DFT)-based method, typically require a long research and development cycle, are of high cost with low efficiency, and have difficulty keeping pace with the development of materials science today. Machine learning can substantially reduce the computational costs and shorten the development cycle; hence, it is one of the most efficient ways of replacing DFT calculations or even repetitive laboratory experiments.

Machine learning was proposed by Samuel[2] in 1959 and has been widely applied in computer vision, general game playing, economics, data mining, and bioinformatics, among other areas.[3-11] With artificial intelligence and machine learning coming of age, important advances are being made not only by researchers in the mainstream artificial intelligence field but also by experts in other fields who are employing these methods to realize their own objectives. Early in the last century, machine learning was used to detect the solubility of C_{60} in materials science,[12] and it has now been used to discover new materials, to predict material and molecular properties, to study quantum chemistry, and to design drugs.[13-17] As the resources and tools for machine learning are abundant and easy to access, the barrier to entry for applying machine learning in materials science is lower than ever.

In this article, we not only stated the basic operational procedures in analyzing the materials' properties of machine learning but also summarized its algorithms application on several mature fields in materials science recently and discussed the improvement required for wild-ranging application. This work committed to popularize the basic knowledge of machine learning and promote its use in materials science.

As a branch of artificial intelligence, machine learning uses large amounts of data to continuously optimize models and to make reasonable predictions under the guidance of algorithms.[17,18] A complete process of machine learning, including data processing, modeling, and validation, will be discussed in detail below. Figure 1 shows a simple workflow of machine learning.

2 | DATA PROCESSING

Information in materials science is substantially enriched by the development of big data. Agrawal and Choudhary[1] summarized the changes that have been caused by big data into seven Vs: volume, velocity, variety, variability, veracity, value, and visualization. Together, these hindered the application of data processing in materials science, which, as a crucial step of machine learning, will directly affect the performance of the resulting machine learning model. Typically, data processing consists of two parts: data selection and feature engineering.

2.1 | Data selection

In data selection, data are selected comprehensively by considering their type, quality, and format.

The use of high-quality data can prevent the consideration of erroneous, missing or redundant information; hence, researchers must collect data from authoritative databases. In 2011, the United States proposed the Materials Genome Initiative for highlighting the importance of massive data in the development of materials science, which strongly encouraged the establishment of a high-quality material database.[19] Various material databases, such as the Open Quantum Material Database, Material Project, Computational Materials Repository, Harvard Clean Energy Project, Inorganic Crystal Structure Database and AFLOWLIB, have already been used for computational materials science.[20-27] In addition, text mining technology has been used to retrieve

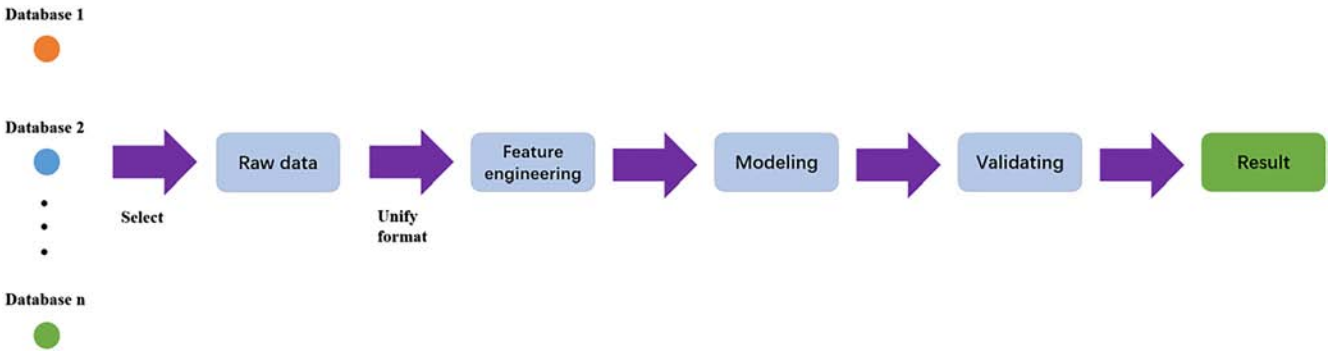

FIGURE 1 Workflow of machine learning

relevant literature on materials to enrich existing databases. Raccuglia et al[28] proposed training a machine learning model on failure data from failed experiments as an application of data processing in materials science. They integrated experimental data that were gathered from failed or less successful hydrothermal synthesis reactions to train a machine learning model to predict the crystallization of template vanadium selenite crystals. This model outperforms traditional manual analysis and the accuracy can reach 89% in predicting the formation conditions of new organic template inorganic products.

At present, the data for materials science can be roughly classified into four types: material properties from experiments and simulations (physical, chemical, structural, thermodynamics, dynamics, etc.), chemical reaction data (reaction rate, reaction temperature, etc.), image data (scanning electron microscope images of materials, photos of material surfaces, etc.), and data from the literature.[1] These data are discrete (eg, texts), continuous (eg, vectors and tensors), or in the form of weighted graphs.[29] Because the data are stored in various databases in various formats, it is difficult to consider data from multiple databases. In addition, the required data format depends on the machine learning algorithm that is applied. Therefore, it is necessary to unify the data in terms of format and to select a suitable data representation for machine learning algorithms in data processing. Fingerprint, SMILES, weighted graph, and the Coulomb matrix are common data representations.[27,29-39] Figure 2 illustrates various representations of a molecule.[29]

2.2 | Feature engineering

After data selection, one should extract suitable characteristics for the predicted target, which is called feature engineering. Feature engineering is the process of extracting features from raw data to enable the application of algorithms. It is crucial to the whole machine learning model and sometime determines the upper limit of its performance. The position of

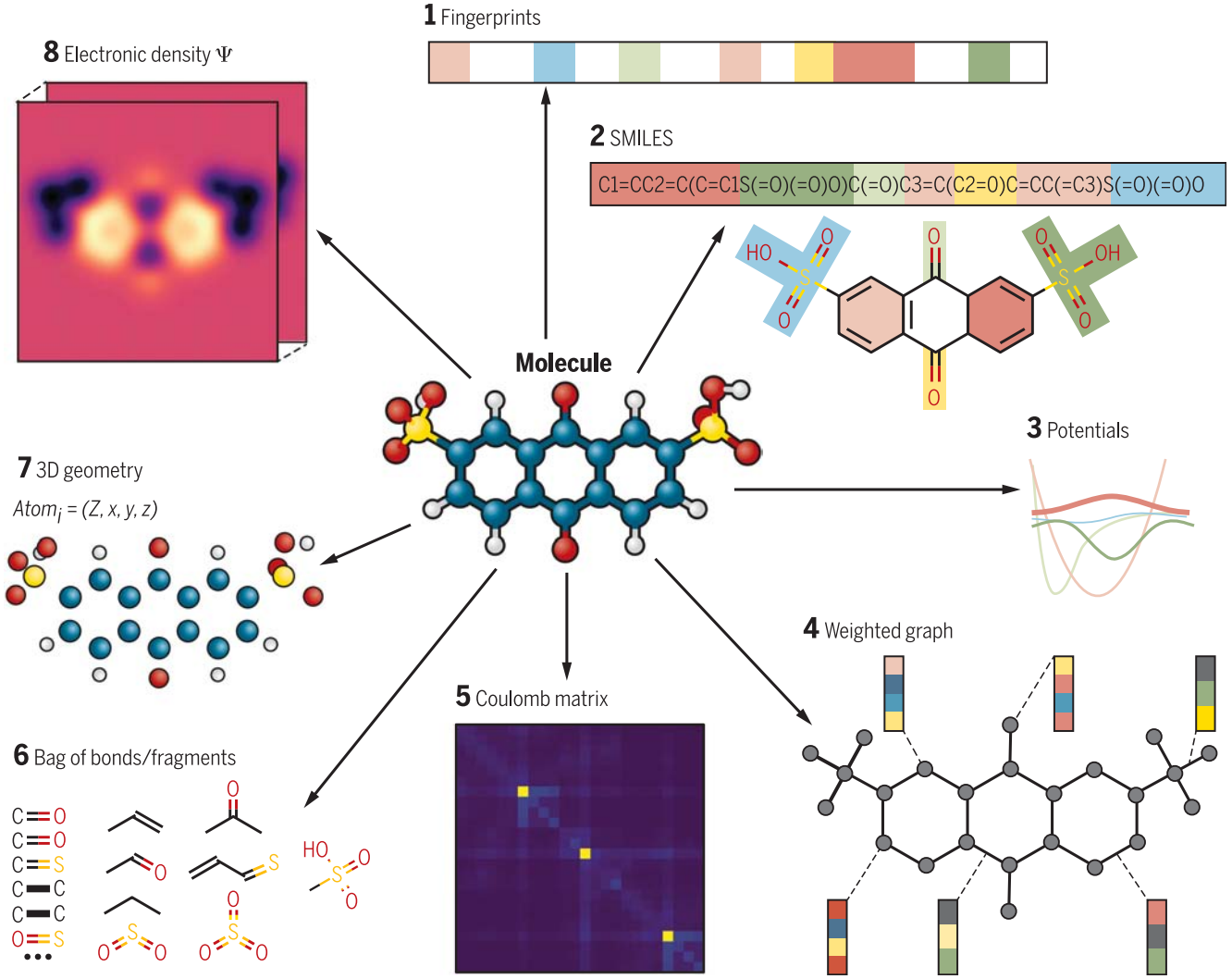

FIGURE 2 Representations of a molecule.[29] Copyright 2018, The American Association for the Advancement of Science

feature engineering in the machine learning workflow is specified in Figure 1.

Traditional machine learning methods (shallow learning) require features to be selected manually. For example, Oliynyk et al[40] used machine learning methods to study potential Heusler compounds and properties. In this study, they finalized 22 features (eg, the group number of element *B*, the total number of p valence electrons, and the radius difference *A*/*B*) via experiments to accelerate the discovery of hidden relationships by computers. Lu et al[41] selected 14 (eg, the tolerance factor, the total ionic charge, the p orbital electrons, and the sum of the s and p orbital radii) out of 30 initial features for training a machine learning model to predict undiscovered hybrid organic-inorganic perovskites (HOIPs) for photovoltaics. Figure 3 illustrates the workflow of their study and the most representative features, which were selected for machine learning.[41]

However, manual feature engineering is not an ideal solution. The limitations of human experience render

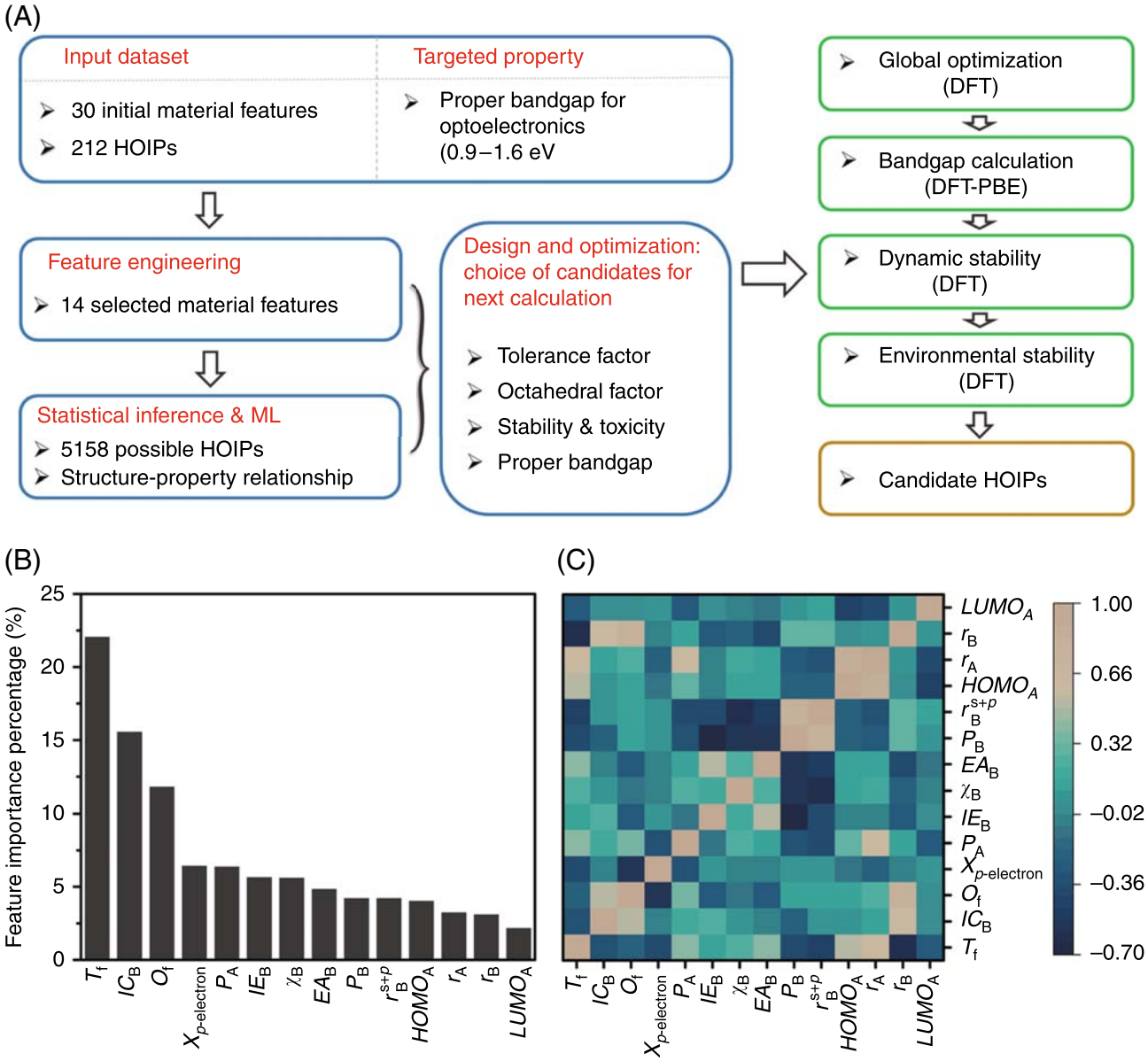

FIGURE 3 Machine learning for discovering stable lead-free hybrid organic-inorganic perovskites. [41] A. A lead-free HOIP design framework. The data-driven method combines machine learning with DFT calculations to identify stable lead-free HOIPs with suitable bandgaps. Copyright 2018, Springer Nature. B. The 14 most representative features, as ranked by the gradient boosting regression (GBR) algorithm. Copyright 2018, Springer Nature. C. A heat map of the Pearson correlation coefficient matrix for the 14 selected features for HOIPs.[41] Copyright 2018, Springer Nature. DFT, density functional theory; HOIP, hybrid organic-inorganic perovskite

difficult the identification of the most representative features for the prediction target. In addition, manual feature engineering requires higher labor and computational costs. More recently, the development of deep learning has eliminated the need for manually featured engineering,[42,43] which may become a trend in machine learning for materials science.

3 | MODELING

With sufficient data in a suitable format, one can build a model for analyzing materials. The modeling steps include selecting appropriate algorithms, training from training data, and making accurate predictions.

Machine learning can be divided into supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning.[13,17] Supervised learning is also known as "learning with a teacher", namely, the corresponding outputs of the training data have been labeled. In contrast, the corresponding outputs of the training data in unsupervised learning are unlabeled. For semisupervised learning, some of the training data are labeled and the remaining data are unlabeled; the amount of unlabeled data often well exceeds the number of labeled data. In reinforcement learning, instead of specifying to the model how to produce correct actions, reinforcement signals that are provided by the environment are used to evaluate the quality of the generated actions and to improve the strategies for adapting to the environment.

Algorithms are available for implementing the four types of machine learning methods that are described above, which can be divided into two types: shallow learning and deep learning.

3.1 | Shallow learning

Referring to traditional machine learning models, shallow learning methods, such as support vector machine (SVM), decision tree (DT), and artificial neural network (ANN), are mainly used in linear classification.

3.1.1 | Support vector machine

The SVM[44,45] is a generalized linear classifier for the binary classification of data. The SVM identifies an $N-1$-dimensional hyperplane for a group of data points in the N-dimensional data dimension. Consider the classification of two-dimensional (2D) data sets as an example: the hyperplane can correctly divide the training data into two categories. If unknown data are encountered, the algorithm will apply this classification model to the unknown data. Figure 4A illustrates the strategy of a linear SVM.[46] Proposed in 1964, the SVM has been rapidly developed since 1990s and has spawned a series of improved algorithms. The SVM has been applied in face recognition, text categorization, biomedicine, and other pattern recognition problems.[46-51]

Due to its excellent classification performance, SVM has been used to classify compounds that are related to a target drug and to successfully identify the drug that is most similar to the target drug in primary screening.[52] In addition, the SVM is well suited for identifying structure–property relationships. Corma et al[53] used the synthesis variables (eg, the initial gel concentration, the reaction process, the temperature, and the time) in the process of zeolite synthesis as the inputs and accurately predicted the structural characteristics and thermodynamic properties of the synthesized products.

3.1.2 | Naive Bayes classifier

The naive Bayes classifier,[13,54] which has been extensively studied since 1950s, is a series of simple probability classifiers that are based on Bayes theorem under a strong independence assumption on the features. It is not a single algorithm for training such classifiers, but a series of algorithms that are based on the following principle: each feature of a sample is unrelated to the others. An object to be classified shall be deemed to belong to the category that corresponds to the maximum probability if the probability of each category has been obtained. In many practical applications, the naive Bayes model parameter uses the maximum likelihood estimation

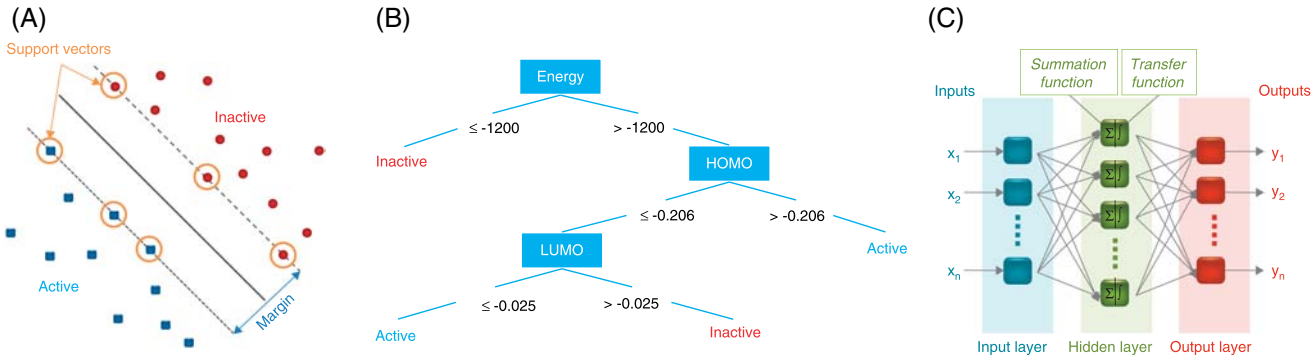

FIGURE 4 Schematic diagrams of (A) support vector machine, (B) decision tree, and (C) artificial neural network. Copyright 2010, Informa Healthcare[14]

method; hence, the naive Bayes model can be applied without using the Bayes probability or any Bayes model in this scenario. One advantage of a naive Bayes classifier is that it only requires the estimation of necessary parameters (the mean and the variance of each variable) based on a small number of training data to make predictions; therefore, it is often used to predict whether a strategy (such as a synthetic recipe for a new molecule) will realize the desired result.

3.1.3 | Decision tree

The DT[55] is a method for approximating discrete function values. It is a typical classification method that induces a set of classification rules from the training set with the objective of correctly classifying examples. Figure 4B illustrates the architecture of a DT. DT was first developed in 1960s by J. Ross Quinlan, who proposed the iterative dichotomiser 3 (ID3) algorithm. Then, the C4.5 algorithm was improved based on the ID3 algorithm in terms of its pruning technology and derivation rules, among other aspects. The DT method typically consists of three steps: feature selection, DT generation, and DT pruning. Among them, the objective of feature selection is to retain features that exhibit sufficient classification performance; the objective of pruning is to make the tree simpler and, thus, more generalizable. There may be multiple DTs that can correctly classify the training data; hence, it is crucial to choose a DT that is less inconsistent with the training data and is sufficiently generalizable.

Carrete used DTs to synthesize new AB_2C Heusler compounds.[40] The training data were collected from Pearson's Crystal Data and the ASM Alloy Phase Diagram Database with the following conditions: (a) the phases do not contain hydrogen, noble gases, or radioactive or actinide elements and (b) the phases exhibit exact 1:2:1 stoichiometry, contain three components and are thermodynamically stable. Twenty-two features (eg, the group number of element *B*, the total number of p valence electrons, the radius difference *A*/*B*, and the electronegativity value of each element) were selected for representing the Heusler compounds. The random forest algorithm (which consists of multiple DTs) was applied to train multiple predictors and to combine their output to yield a single final prediction. Each subpredictor is a DT that has been trained on a small subset of the training data.

3.1.4 | Artificial neural network

In 1943, W. S. McCulloch and W. Pitts established a mathematical model: the MP model. Through the MP model, they proposed a formal mathematical description of neurons and the network structure and proved that a single neuron can perform the logical function, thereby initiating the era of ANNs.

An ANN refers to a nonlinear and adaptive information processing system that is formed by many interconnected processing units (neurons). It is a type of nonprogramed and adaptive information processing architecture, which obtains parallel and distributed information via network transformation and dynamic behavior and imitates the information processing function of the human nervous system. The ANN is an interdisciplinary system that involves neuroscience, artificial intelligence, computer science, and other fields; the architecture is illustrated in Figure 4C.

In an ANN, neurons connect with each other to form layers and represent objects, such as features, words, concepts, or meaningful abstract patterns. There are three types of layers in an ANN: the input layer, the output layer, and the hidden layer. The input layer receives signals or data from the outside world. The output layer outputs the system processing results. The hidden layer lies between the input and output layers and cannot be viewed from outside the network. It performs calculations based on the core function. The input data of an ANN are joined into a new vector and transformed into a matrix prior to flowing into the network. As the data stream passes through the network, the *i*th neuron in the input layer multiplies the input data by the weight W_{ij} and outputs it to the *j*th neuron in the next layer. The weight between neurons reflects the connection strength of the network and the ANN can improve the performance of whole model by adjusting the weight. In the hidden layer, the weighted inputs from the neurons are summed and added to the bias prior to being delivered to the next layer via the activation function. Finally, the output is estimated via a suitable transformation function in the output layer. The main advantages of ANN are as follows: it has a self-learning function, it has associative storage function, and it can search for optimal solutions at a high speed.

Nearly 40 types of neural network models have been proposed, which include backpropagation networks, perceptrons, self-organizing maps, the Hopfield network, and Boltzmann machines. The ANN has been used in many fields of materials science, such as nanomaterial synthesis, quantum computing, and material property analysis.[56-60]

3.2 | Deep learning

Although shallow learning yields satisfactory results in various areas of materials science, several problems are encountered: First, shallow learning algorithms cannot realize the same accuracy as DFT in various tasks, although they have reduced the computational cost. Second, shallow learning algorithms require manual feature engineering; hence, their application requires researchers with domain knowledge to develop suitable representations for the input data. This directly leads to the decline of the model accuracy.[42,43,61-63]

In recent years, the development of deep learning has made new progress in the application of data-driven methods in the field of materials science. As discussed above, shallow learning, which is based on manual feature extraction and linear classification, is highly suitable for linear classification tasks. However, the performance is insufficient for nonlinear classification tasks. By using a nonlinear cascade processing unit for automatic feature extraction and deriving low-level features to obtain more abstract high-level representations of attribute categories, deep models typically outperform shallow models on nonlinear tasks.

Currently, deep learning exhibits powerful performance in image recognition, speech recognition, natural language understanding, biomedicine, and other fields.[64-73] In materials science, various architectures (e.g., convolutional neural network [CNN], recurrent neural network [RNN], deep belief network [DBN], and deep coding network] have demonstrated excellent performance in material detection, material analysis, material design, and quantum chemistry.[42,43,74-78] CNN and RNN will be introduced in the following section. The following methods will not be discussed in detail: DBN, which can be used not only to classify and recognize data but also to generate data; deep stacking network, which consists of multiple blocking models and is easy to train via a supervised approach; and deep Q network, which is a new algorithm that combines deep learning with reinforcement learning to realize end-to-end learning from perception to action.

Due to insufficient material database availability, deep learning still cannot solve many problems in materials science. Moreover, due to the long training time and low interpretability of deep neural networks (DNNs), they may not outperform shallow learning in solving some problems. Therefore, the algorithm that is used for modeling must be selected according to the task.

3.2.1 | Convolutional neural network

The CNN is a feedforward ANN. Inspired by the classical concepts of simple cells and complex cells in visual neuroscience, a CNN combines the ANN with discrete convolution for image processing, which can accept images directly as the input of the network to avoid the complex processes of feature extraction and data reconstruction that are carried out in traditional image recognition algorithms.

In 1980, Kunihiko Fukushima proposed a neural network model, namely, neocognitron, which was the predecessor of the CNN, for visual pattern recognition. After that, although scientists tried utilizing many methods to train multilayer networks, the performance of the CNN was limited due to lack of computing resources when the network depth increased. After 2006, the high-efficiency graphics processing unit became a universal computing device, which facilitated further development of the CNN.[79,80]

A typical CNN network model is illustrated in Figure 5A. Typically, neurons in two adjacent layers are fully connected, whereas neurons in the same layer are not. Each layer of a CNN accepts the output of the layer above as input. Three types of layers are used to construct the CNN architecture between input and output: the convolutional layer, the pooling layer, and the fully connected layer. The convolutional layer is used to extract the characteristics of the input data and to reduce noise. The pooling layer subsamples the input data and divides the input into small regions to apply functions on each region, such as an average function or a maximum function. Figure 5B,C illustrates operations of the convolutional layer and the max pooling layer. Sometimes, other types of layers, such as the drop-out layer, are used to control the size of the CNN.[15,43,61,80,82]

3.2.2 | Recurrent neural network

As there are no connections between neurons in the same layer of the CNN model, data flow from the input layer to the hidden layer and, finally, are output by the output layer. Therefore, it is difficult for a CNN to process data that are related.[82] Hence, a RNN should be used to process sequential data.

In an RNN, the output of the previous step is stored and used to calculate the current output, together with the data from the input layer. If S_i is defined as the state of the RNN, x_i and y_i are the input and output, respectively, of the network, and i is the number of steps, then the current state of network S_i can be calculated as follows:

$$S_i = \mathrm{f}(U*x_i + W*S_{i-1}),$$

where f is a nonlinear function, and U and W are parameters of the RNN.

According to the above equation, the parameters of each layer in the RNN can be shared. Network state S_i is often regarded as a memory unit of the hidden layer that stores the output from the previous step and updates the current input of a nonlinear function each time. Therefore, an RNN requires less parameter learning than CNN.[82]

The RNN has been applied widely in machine translation, speech recognition, and other fields of natural language processing.[83-85] In the field of materials science, the use of an RNN to imitate a similar reward mechanism to design new materials with specified properties has been proposed.[13]

However, errors that may occur during the training process, such model bias and model variance. Model bias is typically caused by errors in the assumptions of the algorithm, whereas model variance is caused by sensitivity to small fluctuations in the training set. In addition to these two types of errors, errors can be caused by calculation limitations or missing data.[13] Moreover, in some cases, overfitting occurs:

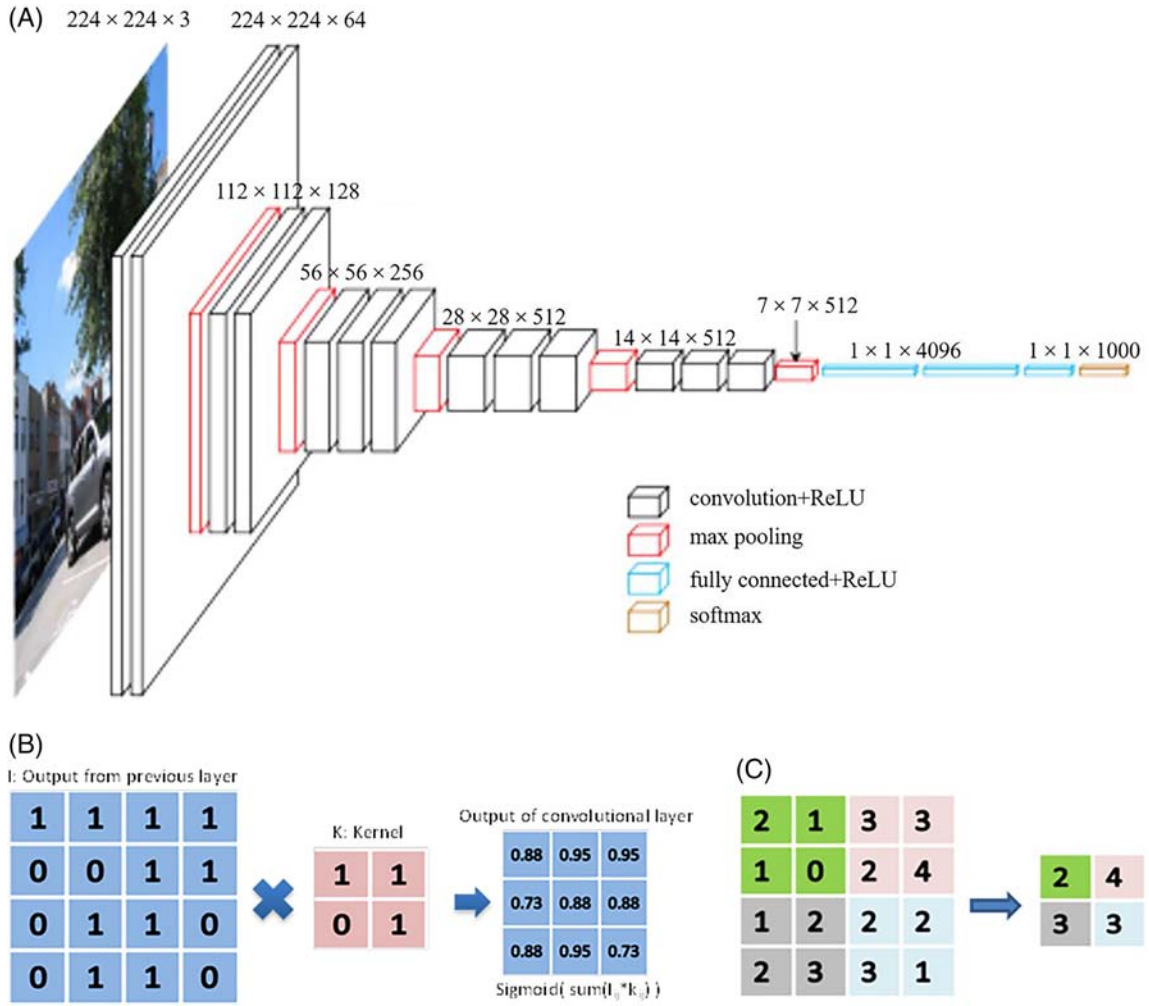

FIGURE 5 A. The architecture of the VGG-16 model (a typical convolutional neural network).[81] This image is from the Internet. B. A schematic diagram of a convolutional layer. A 2 × 2 kernel moves from the top-left to the bottom-right to calculate the weighted sum of the covered area and transforms the result into a fraction that is between 0 and 1 via a sigmoid function. Copyright 2016, New Jersey: World Scientific Pub. Co., 2007.[82] C. A schematic diagram of max a pooling layer. [82] The pooling layer divides the input into four regions and outputs the maximum value of each region. Copyright 2016, New Jersey: World Scientific Pub. Co., 2007

to maintain a consistent hypothesis, the hypothesis becomes excessively strict and, consequently, the performance of the model cannot be guaranteed if the data set to be predicted differs from the training set. Common causes of overfitting are as follows: the modeling sample selection is unsuitable (eg, too few samples, an unsuitable sample selection method, incorrect sample labels), thereby resulting in the selected sample being insufficient for representing the predetermined classification rules; the interference of the model is so large that the computer considers part of the noise as a feature; the model hypothesis cannot reasonably hold or the conditions under which the hypothesis is true are not satisfied; and too many parameters and too high of a model complexity. In addition, for the DT model, if the growth is not restricted rationally, free growth may only include simple event data or no event data, such that although it can perfectly fit the training data, it cannot adapt to other data sets. For the ANN, first, the decision surface may not be unique to the sample, which causes the back propagation algorithm to converge the weights to a more complex decision surface; second, overtraining may cause the model to fit noise or features that are not representative. Therefore, model validation is necessary for reducing errors and avoiding overfitting.[86]

4 | MODEL VALIDATION

When the training of a model has been completed, model validation is conducted to evaluate the accuracy of the model by using the unseen data, which differ from the data in the training data set. Many machine learning methods divide the original data into a training set and a test set and use the training set for model training and the test set for model validation.

K-fold cross-validation is a common validation method. K-fold cross-validation refers to randomly splitting the original data into K parts and using K – 1 parts for model training in each round while retaining one part for model validation. All these parts were used for model validation and the average value of the verification results was calculated as the final estimate.[87-89] One disadvantage of K-fold cross-validation is that it requires the construction of K models, which can be highly time-consuming for a large data set.[86]

Another common validation method is leave-one-out cross-validation (LOOCV). Similar to K-fold cross-validation, if there are N samples in the original data set, each sample is used individually as the verification set and the remaining $N - 1$ samples are used as the training set; hence, LOOCV considers N models and the average classification accuracy of the final validation set of these N models is used as the performance index of the classifiers. Compared with K-fold cross-validation, LOOCV has two main advantages: first, almost all the samples in each round are used to train the model; hence, the distribution is closer to the distribution of the original samples and the obtained results are more reliable. Second, during the experiment, no random factors affect the experimental data, thereby ensuring that the experimental process can be repeated. Similar to K-fold cross-validation, LOOCV has the disadvantage of high computing cost. If the number of original data samples is large, LOOCV has difficulties in implementation, unless the calculations are parallelized to reduce the calculation time.

Methods such as repeating learning test (RLT) cross-validation and bootstrap cross-validation are also used to validate models. In contrast to LOOCV, RLT cross-validation divides part of the data set for validation. Therefore, the computational complexity is substantially reduced. However, the optimal amount of data for model validation is difficult to determine and the test set usually must be selected according to the scenario in practice. Bootstrap cross-validation is a generalization error method that is based on multiple sampling. This method is effective in reducing the variance of K-fold cross-validation; however, the computational cost increases.

5 | APPLICATIONS

Compared with computational simulation, machine learning can identify patterns in large high-dimensional data sets effectively, extract useful information quickly and discover hidden laws. Therefore, it is well suited for material discovery and can accelerate the process of predicting the properties of materials, which typically requires computationally expensive theoretical calculations. In this section, we will introduce three main applications of machine learning in the field of materials science.

5.1 | Material property analysis

5.1.1 | Degradation detection

Machine learning is more accurate and convenient than human judgment in material analysis for the detection of metal corrosion and asphalt pavement cracking and the determination of concrete strength.[59,77,86,90-112] Agrawal et al[86] explored various applications in which machine learning methods (such as feature selection and predictive modeling) are used to predict the fatigue strength of steel by studying the relationship among various properties of the alloy and its composition and manufacturing process parameters. They focused on 25 features that are associated with fatigue strength and found that the tempering temperature was the most important feature that affects the fatigue strength. The process can be divided into four steps: first, the raw data were preprocessed using domain knowledge. Second, ranking-based feature selection methods were applied to select features with high relevance. Then, machine learning algorithms were used to predict the fatigue strength. Finally, LOOCV was used to evaluate the accuracies of the models. The results show that many machine learning methods, such as ANNs, SVM, and linear regression, realized high prediction accuracy, with R^2 values >.98 and error rates <4%.

Zhang proposed an efficient architecture with pixel-level accuracy that is based on CNN. From three-dimensional (3D) images of an asphalt surface, it could detect pavement cracks automatically with a high accuracy of 90.13%. A model that was built by Gibert et al[94] proposed an efficient model for railway track inspection. The model was based on a fully convolutional network. Four convolutional layers were used for material classification and five convolutional layers were used for fastener detection. Researchers used an artificially illuminated car to collect 203 287 track images along 85 miles of track and annotated the data using a customized software tool. The data set was divided into five parts, with 80% of the images used for training and 20% for testing. For each data segmentation, 50 000 patches of each class were randomly sampled. Therefore, each model was trained on 2 million patches. The architecture of the model and the semantic segmentation results are presented in Figure 6.

5.1.2 | Nanomaterials analysis

With the development of artificial intelligence, machine learning plays an increasingly important role in the field of nanomaterials.[58,113-115] As early as 1993, the use of machine learning to study the solubility of C_{60} was proposed.[12] Machine learning has been widely used to predict the toxicity of nanomaterials, to discover new nontoxic nanoparticles, to develop multistructure/single-property relationships of nanoparticles, to study quantum-mechanical observables of

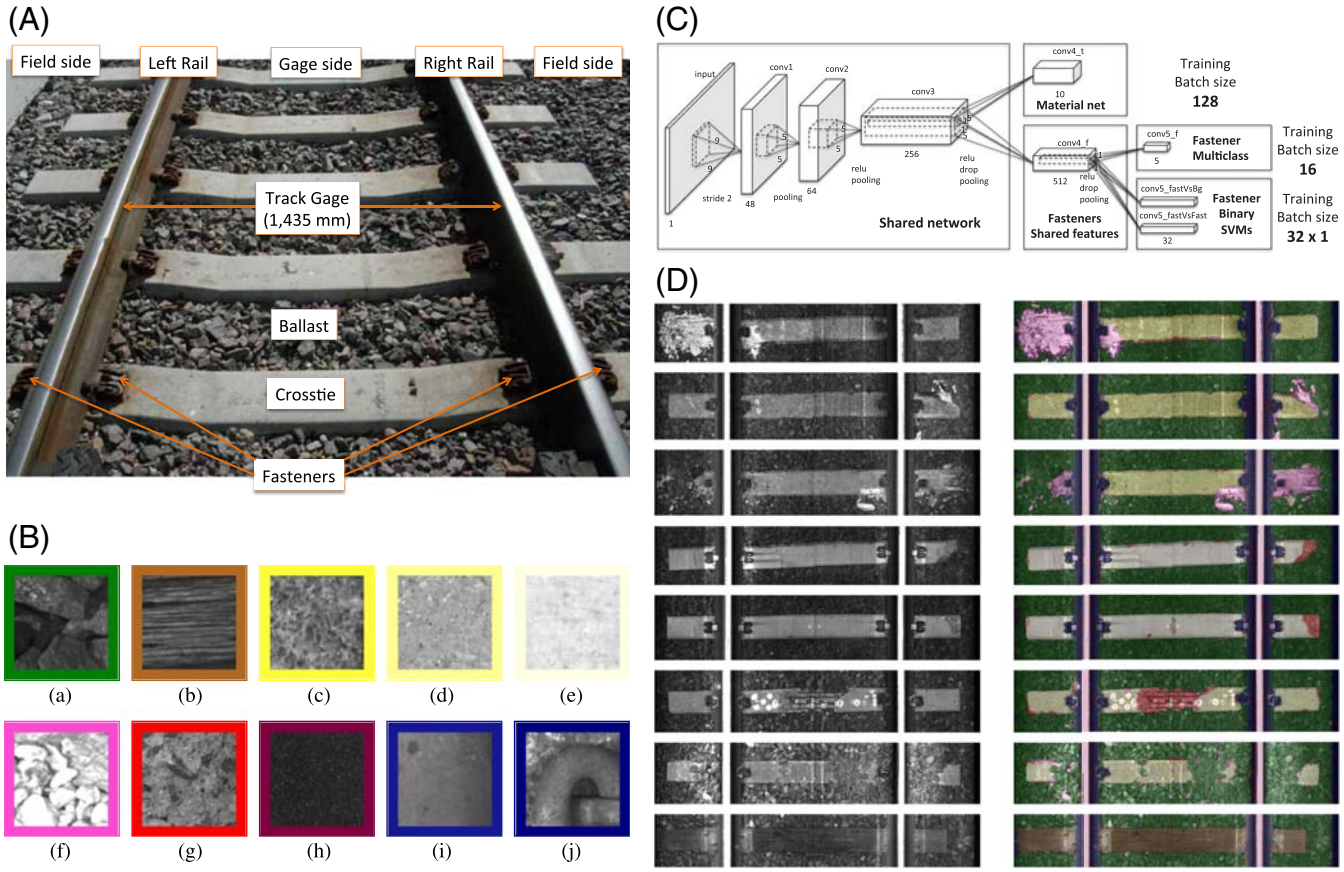

FIGURE 6 Deep learning for railway track inspection.[94] A. Definitions of basic elements of the railway. Copyright 2017, IEEE. B. Material categories that were identified via CNN: (a) ballast, (b) wood, (c) rough concrete, (d) medium concrete, (e) smooth concrete, (f) crumbling concrete, (g) chipped concrete, (h) lubricator, (i) rail, and (j) fastener. Copyright 2017, IEEE. C. The architecture of a fully convolutional neural network. Copyright 2017, IEEE. D. The semantic segmentation results of a railway. The first three rows correspond to an example of crumbling ties. Copyright 2017, IEEE

molecular systems, to analyze chemical reactions of nanomaterials and to solve kinetic systems.[16,29,58,60,76,113-118] Oh et al successfully applied meta-analysis to study the chemical toxicity of quantum dots (QDs). They used text mining to extract relevant QD toxicity data from 307 studies in the literature and applied the random forest regression model to analyze the data. According to their results, the toxicity was closely related to the surface properties (such as the shell, ligand, and surface modification), the diameter, the assay type, and the exposure time of QDs. Figure 7 presents the characteristics of 14 species that substantially influence QD toxicity.[113]

5.1.3 | Molecular property prediction

High-throughput density functional calculations for molecular property prediction are highly time-consuming. As an alternative, machine learning is a feasible approach for the fast prediction of structures or properties of molecules, compounds and materials; in addition, it can realize high accuracy.[13,14,42,43,62,74,75,119-128] ElemNet is a model that is based on a DNN that takes elements as input for predicting material properties.[42] It extracts the physical and chemical interactions and similarities between elements automatically and makes fast and precise predictions. The framework of ElemNet is illustrated in Figure 8.[42] Similar to ElemNet, Chemception is a model that is based on a CNN that transforms raw data of compounds into 2D images to predict toxicity, activity, and solvation properties.[43] In addition, Ward et al[122] proposed a general machine learning framework for predicting properties of inorganic materials. This model can be applied to predict diverse properties of crystalline and amorphous materials, such as band gap energy and glass-forming ability.

5.2 | Discovering new materials

5.2.1 | Structure-oriented design

Many classification and regression algorithms can be applied to predict the chemical composition of a material from its

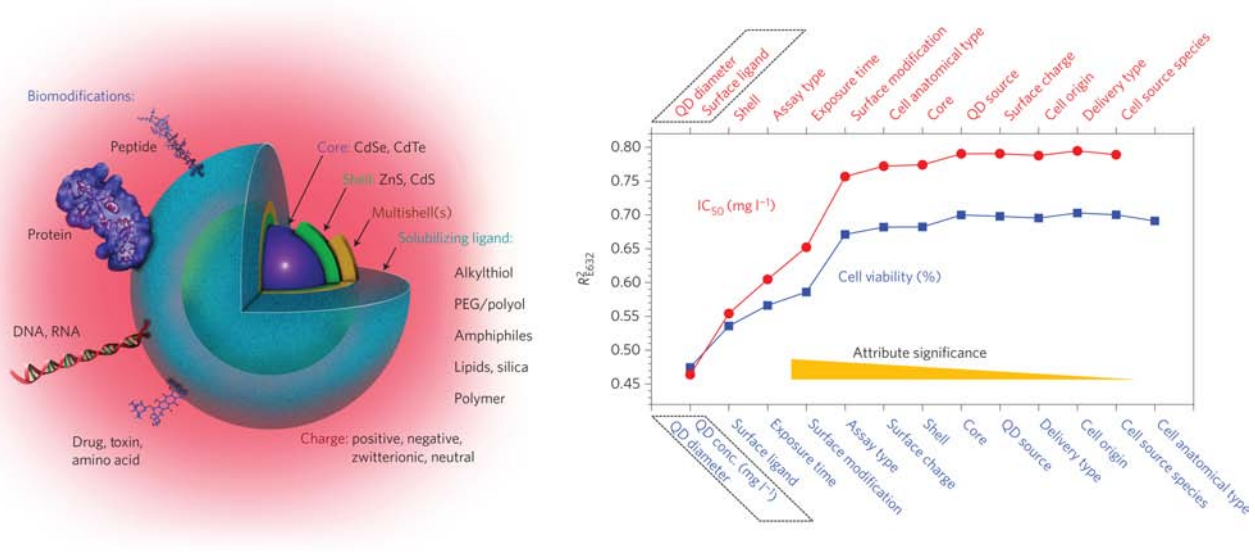

FIGURE 7 Meta-analysis of cellular toxicity for cadmium-containing quantum dots (QDs). Researchers used data mining to collect toxicity data of QDs and random forest was used to identify relevant QD data attributes and to develop robust data-driven models of QD toxicity.[113] Copyright 2016, Springer Nature

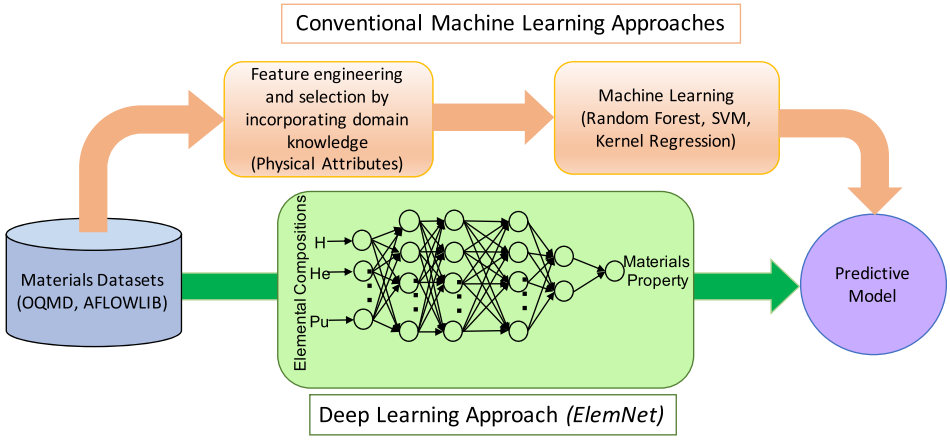

FIGURE 8 Deep neural networks for property prediction of materials.[42] Copyright 2018, Springer Nature

structure.[28,40,41,129,130] Perovskite is an important crystal structure in many fields.[131,132] Shuaihua et al used various regression algorithms (Gradient boosting regression, kernel ridge regression [KRR], support vector regression, Gaussian process regression, DT regression, and multilayer perceptron regression) to predict stable lead-free HOIPs from 5158 unexplored HOIPs and successfully identified six stable compounds ($C_2H_5OInBr_3$, $C_2H_6NInBr_3$, $NH_3NH_2InBr_3$, $C_2H_5OSnBr_3$, NH_4InBr_3, and $C_2H_6NSnBr_3$).[41] Pilania et al[121] developed a systematic feature engineering approach and an efficient machine learning model for predicting electronic bandgaps of double perovskites and for selecting stable perovskite candidates. The structure of double perovskite crystal and the chemical space of cations in the A-site and B-site are illustrated in Figure 9A,B. The KRR-based learning model was trained and tested on a data set of more than 1300 double perovskites. From this learning framework, it is concluded that the lowest occupied energy levels of the A-site elements and the electronegativities of the B-site elements are key factors that primarily control the bandgaps of double perovskites.

Oliynyk et al[40] proposed a Heusler discovery engine that is based on the random forest algorithm for identifying new full-Heusler compounds. This model realized a high true-positive rate of .94 and successfully predicted 12 novel gallides, namely, MRu_2Ga and RuM_2Ga (M = Ti − Co), as Heusler compounds. Legrain et al[130] trained a model using experimentally reported compounds to predict the stability of half-Heusler compounds. The model, which was based on the random forest algorithm, retrieved 71 178 compositions

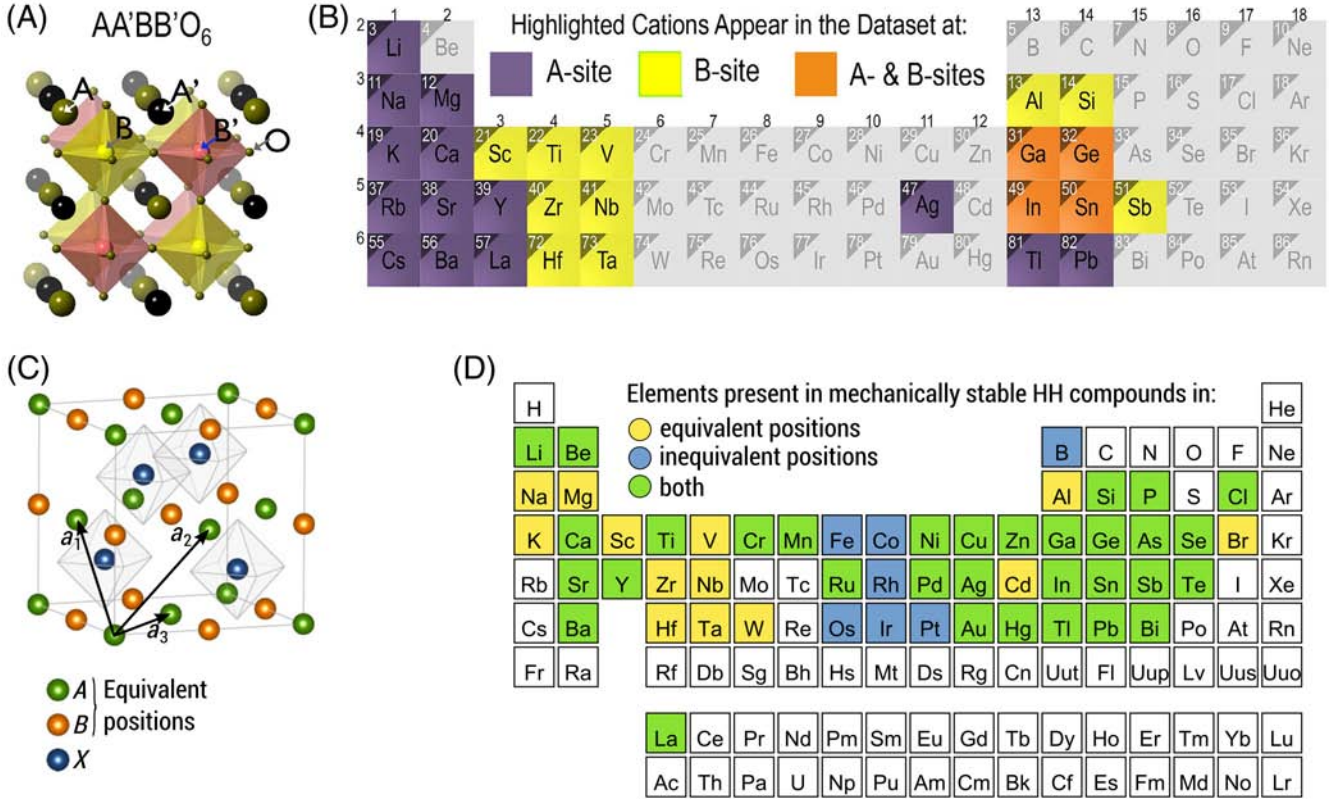

FIGURE 9 A. The structure of a double perovskite crystal.[121] Copyright 2016, Springer Nature. B. The chemical space of cations in the A-site and B-site for double perovskite prediction.[121] Copyright 2016, Springer Nature. C. The structure of a half-Heusler compound. [133] Copyright 2014, American Physical Society. D. The chemical space for low-thermal-conductivity half-Heusler compound prediction.[133] Copyright 2014, American Physical Society

and yielded 30 results, which mostly matched half-Heusler compounds, for further exploration. Another similar study was reported on identifying low-thermal-conductivity half-Heusler semiconductors.[133] The random forest algorithm was used to scan more than 79 000 half-Heusler entries in the AFLOWLIB database. Possible half-Heusler compounds from all nonradioactive combinations of elements in the periodic table were considered. Figure 9C,D illustrates the structure of a half-Heusler compound and the chemical space of this study.

5.2.2 | Element-oriented design

Machine learning can be used to predict new compounds and their structures from an input composition.[134-136] A probabilistic model that was based on an experimental crystal structure database was used to identify 209 new ternary oxides.[134] In addition, Meredig et al[136] constructed a machine learning model that was based on thousands of DFT calculation results for predicting the thermodynamic stability of arbitrary compositions. The researchers built a large database of DFT calculation results and two predictive formation energy models (one heuristic and one machine learning based). Then, they used these two models to scan 1.6 million candidate ternary compositions, ranked the most probable results by combining the two models and finally identified 4500 new stable materials.

In a similar study, machine learning was used to study binary compounds.[135] First, researchers used a unsupervised learning algorithm to separate 67 octet compounds into distinct classes according to their crystal structures; second, the supervised learning algorithm was applied to identify the correct crystal structures of 55 compounds; finally, a regression algorithm was used to predict the melting points of 44 AB suboctet compounds by mining a combination of 16 properties of the constituent atoms of each binary compound. In the studies that are discussed above, machine learning has demonstrated high accuracy and its potential in discovering new compounds.

Recently, a model that was based on objective-reinforced generative adversarial networks was proposed for generating new organic molecules with specified chemical features and physical responses via a reward mechanism.[13,137] The model consists of a generator and a discriminator: the generator captures the distribution of the data while the discriminator compares the molecular structure that is obtained by the

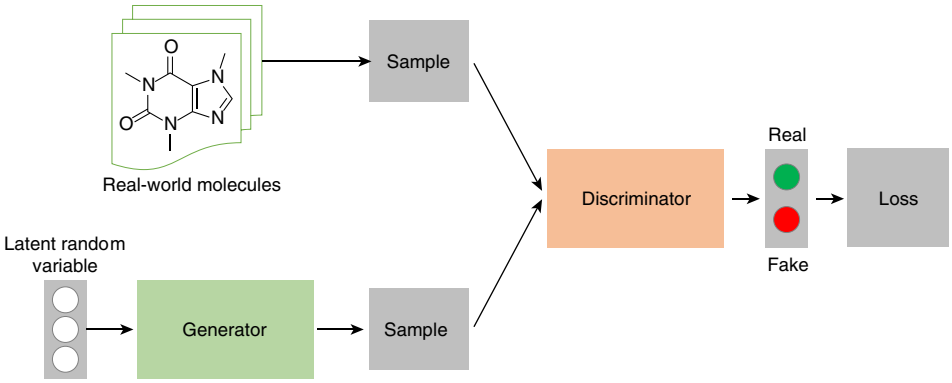

FIGURE 10 Framework of an objective-reinforced generative adversarial network.[13] Copyright 2018, Springer Nature

generator with real molecular structures to determine whether the molecular structure that was obtained by the generator can exist. The generator is trained to maximize the error probability of the discriminator. Repeating this process will increase the discrimination performance of discriminator between real and fake data. Therefore, such "reward mechanism" reinforcement learning network can be used to design chemical structures with special physical or biological characteristics. Figure 10 illustrates the framework of this objective-reinforced generative adversarial network.

5.2.3 | Inverse design

Deep learning has high potential in the inverse design of materials.[13,29,78,137] Inverse design begins from the required functionality and searches for the ideal molecular structure that exhibits this functionality. The method takes a functionality as input and outputs a molecular structure.[29] Figure 11A compares three material design schemes.

The successful realization of inverse-designed organic molecules via a model that was based on DNN and RNN was reported.[78] In contrast to reconstructing molecule structures directly from molecule descriptors, the inverse design model translates molecule descriptors into molecule identifiers. Therefore, a DNN and a RNN are adopted as the encoder and the decoder, respectively. The DNN identifies the relationship between molecule structures and material properties and encodes the molecule descriptors x into encoded molecule information z; the RNN decodes the encoded molecule information z into molecule structure identifier y to reconstruct the encoded molecular descriptors into a molecular structure. If the molecular structure of the RNN output is stable and the molecular properties are consistent with the target, the inverse design of molecules is successful. Figure 11B,C illustrates the workflow of the molecular inverse design method and the gradient-based optimization of molecular properties in continuous latent space.

5.2.4 | Drug design

Drug design is one of the mature fields in which machine learning is utilized.[52,63,79,139] Machine learning approaches that are modeled on small molecules can handle the structural complexity of proteins and can predict structure-activity relationships accurately, which facilitates the discovery of target drugs.[79] A typical process of machine learning for drug discovery begins with compounds that have already been tested. Then, batches of compounds are repeatedly designed and selected for parallel testing. The activity model is refined at each step to select the most promising compound for the next batch and the process is repeated until an active drug has been identified.[52] This is an effective method for discovering active drugs; however, it has a long development cycle and is computationally expensive. Figure 12B illustrates the workflow of this drug discovery approach, which is based on machine learning.

Modern drug discovery strategies include ligand-based drug design (LBDD) and structure-based drug design (SBDD) techniques,[139] some steps of which use machine learning to simplify calculations or to build statistical validation models. LBDD includes similarity search (the use of 2D or 3D information from one or more compounds to calculate a similarity index for sorting compounds in the unknown database) and the construction of a classification or regression model for predicting biological activity. Figure 12A illustrates the framework of LBDD. Compared with SBDD, LBDD has lower computational cost and is easier to use. SBDD can be used when information about biological target structures is available. Here, machine learning can be applied to predict the tertiary structures of receptors through predictions of secondary structures, solvent accessibility, and disordered regions, among other factors.[139] Many common machine learning algorithms, such as ANN, SVM, DTs, random forest, and k-nearest neighbor, can be used in two drug discovery strategies.

Recently, AtomNet, which is the first architecture that is based on deep CNNs for molecular binding affinity

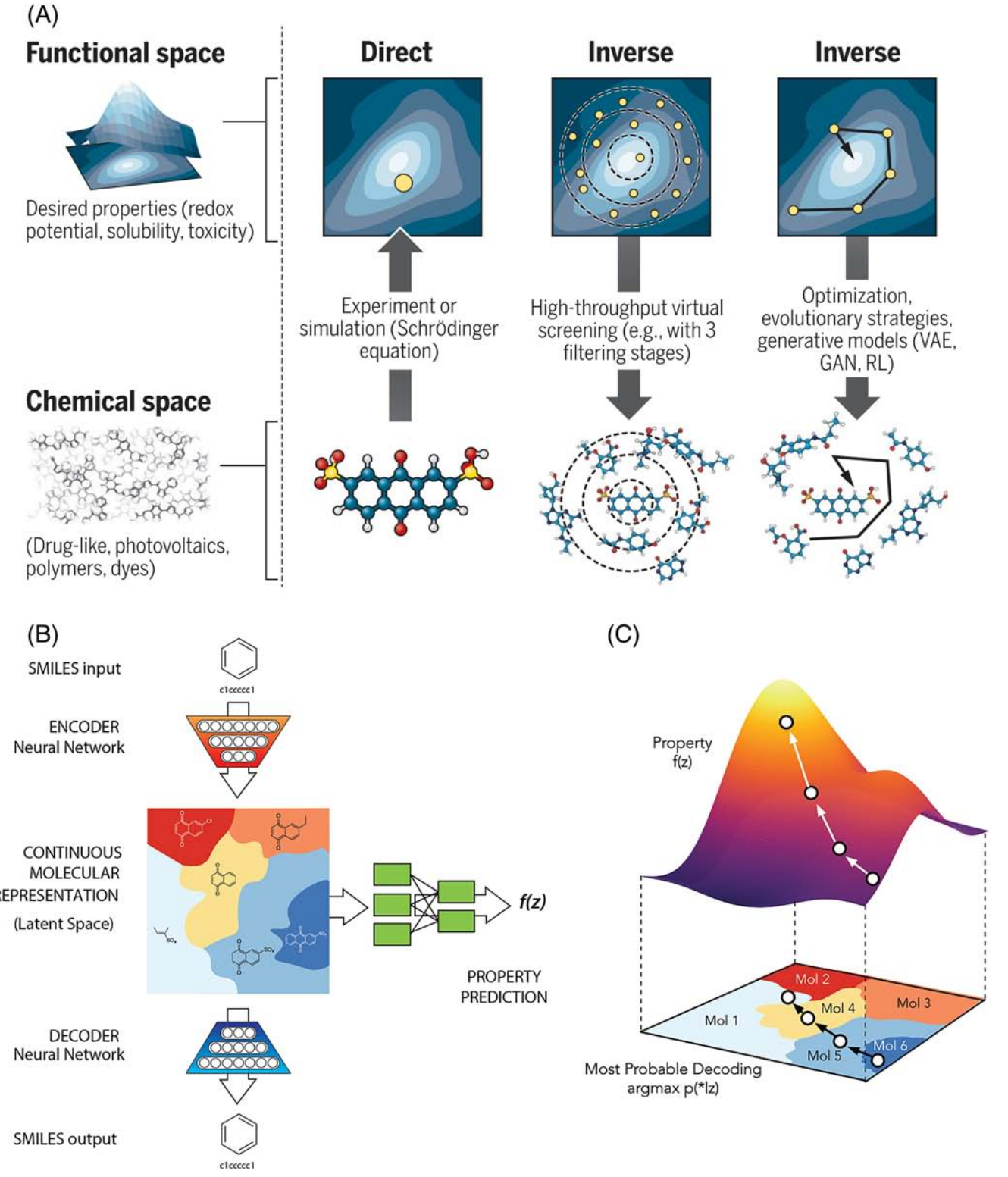

FIGURE 11 Inverse design of molecules via machine learning. A. Three molecule design methods. Comparing with inverse design that is based on high-throughput screening, the machine-learning-based inverse design method is more effective in chemical space exploration.[29] Copyright 2018, The American Association for the Advancement of Science. B. The workflow of the inverse design method.[138] Copyright 2018, American Chemical Society. C. Gradient-based optimization in continuous latent space.[138] Copyright 2018, American Chemical Society

prediction, was applied for bioactivity prediction in structure-based drug discovery.[79] It predicts the structure of a protein directly and yields outstanding results in SBDD. Because the filters of AtomNet cannot be easily visualized, the researchers proposed an indirect way of verifying that the model had learned relevant features. They applied filters

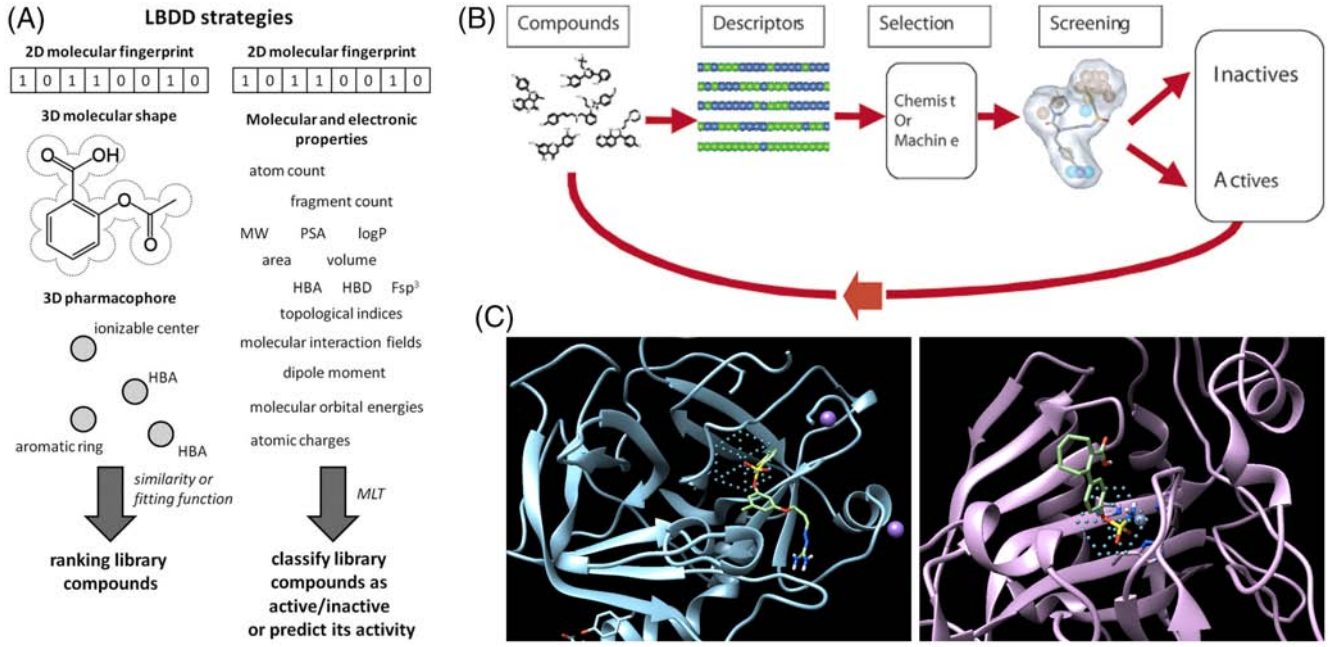

FIGURE 12 A. LBDD strategies for drug design.[139] Copyright 2016, Taylor & Francis. B. The drug discovery approach.[52] Copyright 2003, American Chemical Society. C. Sulfonyl/sulfonamide detection with autonomously trained convolutional filters.[79] Copyright 2015, by the author. LBDD, ligand-based drug design

to input data and examined the location where they maximally fire. Figure 12C illustrates the first convolutional layer, which specializes as a sulfonyl/sulfonamide detector. According to the results, the filter can infer a meaningful spatial arrangement of input atomic types without any prior chemical knowledge.

5.3 | Quantum chemistry

Machine learning, in place of or combined with computer simulation (DFT), is often used to simplify the computations of complex problems in the field of quantum chemistry.[41,134,140-142] By using DFT data to train a machine learning model, Seko et al substantially reduced the calculation cost without sacrificing the accuracy of the model. Models that have been trained on DFT data have been used to predict the melting temperatures of single and binary compounds.[128]

As a data-driven method, machine learning can bypass the solution of complex equations (eg, the Kohn-Sham equation or Schrödinger equation) to determine the properties that are related to the energy, geometry, and curvature of the potential energy surfaces of molecules.[29,76,115,120,121,126,127,143-145] A group developed a model that is based on a deep tensor neural network for predicting atomic energies and local chemical potentials in molecules, reliable isomer energies, and molecules with peculiar electronic structures, thereby resulting in insights into quantum-mechanical observables of molecular systems.[76] In this work, each atomic type corresponds to a coefficient vector $\mathbf{c}_i^0$, which is progressively refined by interaction v_{ij}. The interaction v_{ij} is determined by the current coefficient vector $\mathbf{c}_j^t$ and the distance D_{ij} between atoms i and j. After T interactions, the energy of the final coefficient vector $\mathbf{c}_i^T$ is predicted to be E_i and the molecular energy E is the sum of all the atomic energies. Figure 13 illustrates their work in detail.

6 | PROSPECTS

Machine learning has been widely used in the prediction of properties, the discovery of new materials and the exploration of quantum chemistry due to its powerful prediction performance and relatively low computational cost. However, the application of machine learning in materials science still faces many problems. For example, the available high-quality data that are related to materials are insufficient, the properties of materials are difficult to represent perfectly, and the prediction accuracy is lower than that of DFT calculations. Herein, we propose further directions that may contribute to the application of machine learning in materials science.

First, accelerating the construction of a material database is highly important for the future development of machine learning. As a data-driven method, the quantity and quality of data directly affect the accuracy of machine learning. The scientific literature and experimental records contain a large amount of material data to which machine learning can be applied, such as molecular properties, reaction conditions, and synthetic formulations. Using text mining, these useful

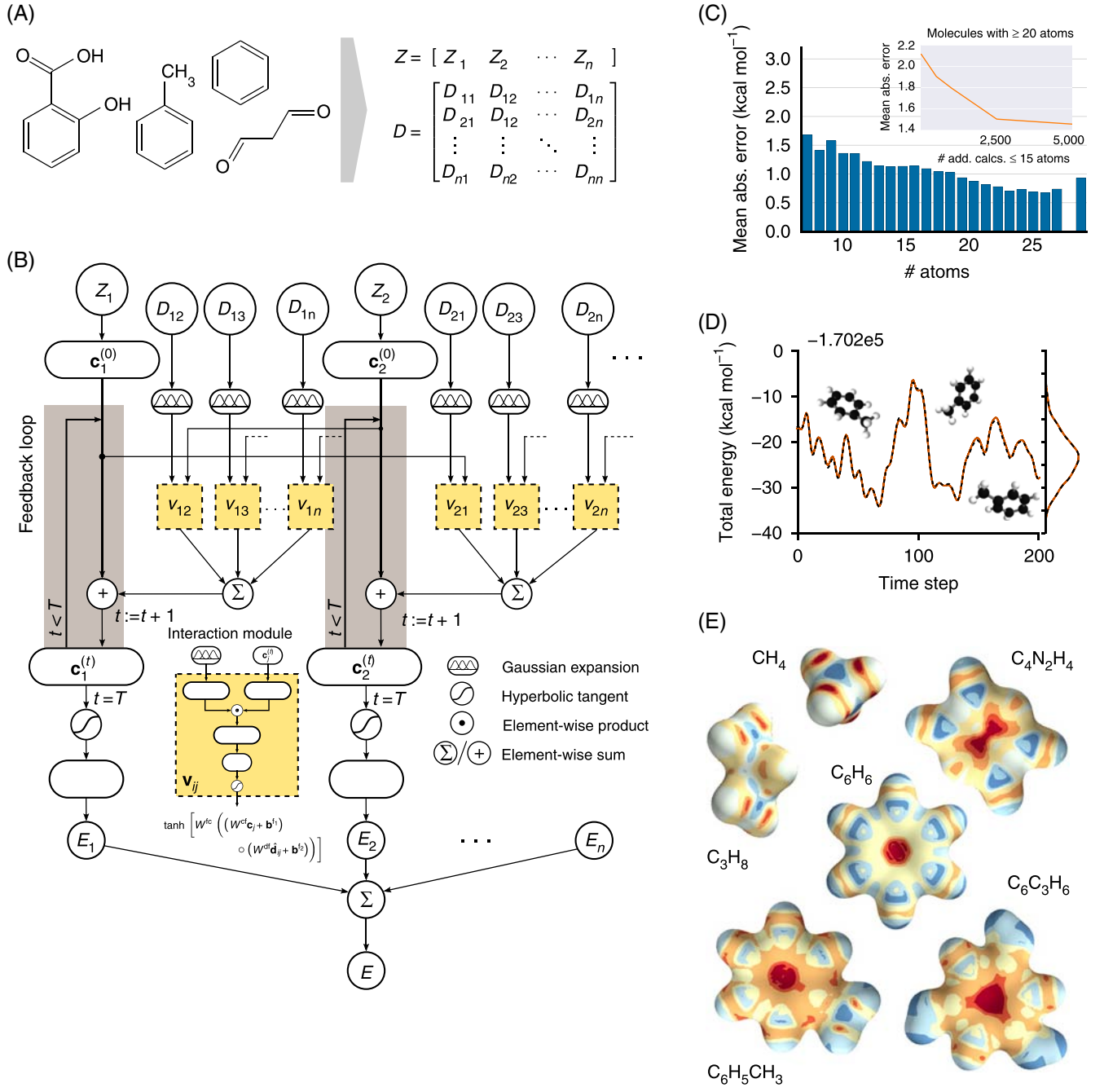

FIGURE 13 Deep tensor neural network for predicting atomic energies and local chemical potentials in molecules. A. Encoding of a molecule into a vector of nuclear charges and an interatomic distance matrix as input of a neural network.[76] Copyright 2017, Springer Nature. B. The architecture of a deep tensor neural network.[76] Copyright 2017, Springer Nature. C. The total energies that are extracted from the calculated (black) and predicted (orange) molecular dynamics trajectories of toluene.[76] Copyright 2017, Springer Nature. D. The energy contribution of a hydrogen test charge on the isosurfaces of various molecules.[76] Copyright 2017, Springer Nature

data, which are scattered among articles, journals and magazines, can be quickly collected, which will substantially enrich the existing material databases and enable the creation of specialized databases.[13,16,146]

Second, establishing new principles for machine learning is essential. With the development of deep learning and the replacement of manual feature engineering, raw data will be represented more effectively in the future. However, experts still do not understand the basis on which DNNs select features and the meanings of the selected features. This renders the results of deep learning insufficiently convincing and fails to yield a widely suitable theory. Trying to understand what is going on inside the "black box" not only enhances the generalizability of machine learning in materials science but also facilitates the identification of laws of nature that are unknown to humans.

Third, quantum chemistry could be another key application of machine learning. Machine learning's powerful data-processing capability enables it to solve many problems in quantum chemistry. Combining DFT with machine learning can substantially increase the prediction accuracy of the model. This could be a powerful tool for predicting complicated properties and structures of molecules, for investigating quantum multibody systems and for discovering new materials.

Machine learning still cannot realize the expected accuracy when applied to some tasks due to insufficient material data. Therefore, a more accurate model that was trained on a small but accurate data set is absolutely necessary in some scenarios. The performance of a deep learning model that has been trained on a small data set size of 4000 samples has been demonstrated to be sufficient.[42] In addition, the method that was discussed above of training a model with failure data that were collected from failed experiments may be helpful in such scenarios.[28]

ACKNOWLEDGMENTS

Dr J. Wei and X. Chu contributed equally to this work. Dr J. Wei acknowledges that this project was funded by China Postdoctoral Science Foundation (no. 2017 M620694) and National Postdoctoral Program for Innovative Talents (BX201700040). This work was also financially supported by the National Natural Science Foundation of China (grant nos. 61622406 and 61571415), the National Key Research and Development Program of China (grant nos. 2017YFA0207500 and 2016YFB0700700), the Strategic Priority Research Program of Chinese Academy of Sciences (grant no. XDB30000000), and Beijing Academy of Quantum Information Sciences (grant no. Y18G04). All authors agree with the content of manuscript.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Jing Wei https://orcid.org/0000-0003-2991-0123
Zhongming Wei https://orcid.org/0000-0002-6237-0993

REFERENCES

1. Agrawal A, Choudhary A. Perspective: materials informatics and big data: realization of the "fourth paradigm" of science in materials science. *APL Mater*. 2016;4:053208.
2. Kohavi R, Provost F. Glossary of terms. *Machine Learn*. 1998;30:271-275.
3. Silver D, Huang A, Maddison CJ, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*. 2016;529:484-504.
4. Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge. *Nature*. 2017;550:354-359.
5. Shin HC, Roth HR, Gao M, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans Med Imag*. 2016;35:1285-1299.
6. Cambria E, White B. Jumping NLP curves: a review of natural language processing research [review article]. *IEEE Comput Intell Mag*. 2014;9:48-57.
7. Liu H, Xu C, Liang J. Dependency distance: a new perspective on syntactic patterns in natural languages. *Phys Life Rev*. 2017;21:171-193.
8. Tsai C-W, Lai C-F, Chiang M-C, Yang LT. Data mining for internet of things: a survey. *IEEE Commun Surv Tutor*. 2014;16:77-97.
9. Cully A, Clune J, Tarapore D, Mouret JB. Robots that can adapt like animals. *Nature*. 2015;521:503-516.
10. Kononenko I. Machine learning for medical diagnosis—history, state of the art and perspective. *Artif Intell Med*. 2001;23:89-109.
11. Feng N, Wang HJ, Li M. A security risk analysis model for information systems: causal relationships of risk factors and vulnerability propagation analysis. *Inform Sciences*. 2014;256:57-73.
12. Ruoff RS, Tse DS, Malbotra R, Lorents DC. Solubility of fullerene (C_{60}) in a variety of solvents. *Phys Chem*. 1993;97:3379-3384.
13. Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. *Nature*. 2018;559:547-555.
14. Nantasenamat C, Isarankura-Na-Ayudhya C, Prachayasittikul V. Advances in computational methods to predict the biological activity of compounds. *Expert Opin Drug Discov*. 2010;5:633-654.
15. Nash W, Drummond T, Birbilis N. A review of deep learning in the study of materials degradation. *npj Mater Degrad*. 2018;2:37-49.
16. Wang M, Wang T, Cai P, Chen X. Nanomaterials discovery and design through machine learning. *Small Method*. 2019;1900025:1-7.
17. Wu W, Sun Q. Applying machine learning to accelerate new materials development. *Sci Sin Phys Mech Astron*. 2018;48:107001.
18. de Mantaras RL, Armengol E. Machine learning from examples: inductive and lazy method. *Data Knowl Eng*. 1998;99:99-123.
19. Andrea Widener CEW. Materials genome initiative. *Govern Policy*. 2011;1-3.
20. Hachmann J, Olivares-Amaya R, Atahan-Evrenk S, et al. The Harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid. *J Phys Chem Lett*. 2011;2:2241-2251.
21. Hachmann J, Olivares-Amaya R, Jinich A, et al. Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry—the Harvard Clean Energy Project. *Energ Environ Sci*. 2014;7:698-704.
22. Jain A, Ong SP, Hautier G, et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL Mater*. 2013;1:011002.

23. Saal JE, Kirklin S, Aykol M, et al. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). *TMS*. 2013;65:1501-1509.
24. Kirklin S, Saal JE, Meredig B, et al. The open quantum materials database (OQMD): assessing the accuracy of DFT formation energies. *npj Comput Mater*. 2015;1:1-15.
25. Curtarolo S, Setyawan W, Wang S, et al. AFLOWLIB.ORG: a distributed materials properties repository from high-throughput ab initio calculations. *Comp Mater Sci*. 2012;58:227-235.
26. Allen FH. The Cambridge Structural Database a quarter of a million crystal structures and rising. *Struc Sci*. 2002;58:380-388.
27. Kalidindi SR, De Graef M. Materials data science: current status and future outlook. *Annu Rev Mat Res*. 2015;45:171-193.
28. Raccuglia P, Elbert KC, Adler PD, et al. Machine-learning-assisted materials discovery using failed experiments. *Nature*. 2016;533:73-78.
29. Sanchez-Lengeling B, Aspuru-Guzik A. Inverse molecular design using machine learning generative models for matter engineering. *Science*. 2018;361:360-366.
30. Bartók AP, Kondor R, Csányi G. On representing chemical environments. *Phys Rev B*. 2013;87:184115.
31. Duvenaudy, D., Maclauriny, D., Aguilera-Iparraguirre, J., et al Convolutional networks on graphs for learning molecular fingerprints. arXiv:1509.09292 [cs.LG].
32. Gilmer, J., Schoenholz, S. S., Riley, P. F., et al Neural message passing for quantum chemistry. arXiv:1704.01212 [cs.LG].
33. Hansen K, Biegler F, Ramakrishnan R, et al. Machine learning predictions of molecular properties: accurate many-body potentials and nonlocality in chemical space. *J Phys Chem Lett*. 2015;6:2326-2331.
34. Hirn M, Mallat S, Poilvert N. Wavelet scattering regression of quantum chemical energies. *Multiscale Model Sim*. 2017;15:827-863.
35. Huang, B. Lilienfeld, O. A. V. The DNA of chemistry-scalable quantum machine learning with amons. arXiv:1707.04146 [physics.chem-ph].
36. Kearnes S, McCloskey K, Berndl M, Pande V, Riley P. Molecular graph convolutions: moving beyond fingerprints. *J Comput Aided Mol des*. 2016;30:595-608.
37. Rogers D, Hahn M. Extended-connectivity fingerprints. *J Am Chem Soc*. 2010;50:742-755.
38. Rupp M, Tkatchenko A, Muller KR, Von Lilienfeld OA. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys Rev Lett*. 2012;108:058301.
39. Weininger D, Smiles A. Chemical language and information system. 1. Introduction to methodology and encoding rules. *J Am Chem Soc*. 1988;28:31-37.
40. Oliynyk AO, Antono E, Sparks TD, et al. High-throughput machine-learning-driven synthesis of full-Heusler compounds. *Chem Mater*. 2016;28:7324-7331.
41. Lu S, Zhou Q, Ouyang Y, Guo Y, Li Q, Wang J. Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. *Nat Commun*. 2018;9:3405-3413.
42. Jha D, Ward L, Paul A, et al. ElemNet: deep learning the chemistry of materials from only elemental composition. *Sci Rep*. 2018;8:17593.
43. G.B. Goh, C. Siegel, A. Vishnu, et al. Chemception:a deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models. arXiv:1706.06689 [stat.ML].
44. Cortes C, Vapnik V. Support-vector networks. *Mach Leam*. 1995;20:273-298.
45. Burges CJC. A tutorial on support vector machines for pattern recognition. *Data Min Knowl Disc*. 1998;2:121-167.
46. Lal TN, Schroder M, Hinterberger T, et al. Support vector channel selection in BCI. *IEEE Trans Biomed Eng*. 2004;51:1003-1010.
47. Brown MPS, Grundy WN, Lin D, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. *PNAS*. 2000;97:262-267.
48. Cifarelli C, Patrizi G. Solving large protein secondary structure classification problems by a nonlinear complementarity algorithm with {0, 1} variables. *Optim Method Softw*. 2007;22:25-49.
49. Molina GNG, Ebrahimi T, Vesin J-M. Joint time-frequency-space classification of EGG in a brain-computer Interface application. *Eurasip J Adv Sig Pg*. 2003;7:713-728.
50. Joachims T. Text categorization with support vector machines—learning with many relevant features. *ECML*. 1998;1398:137-142.
51. Qin J, He Z-S. A SVM face recognition method based on Gabor-featured key point. *IEEE*. 2005;8:5144-5149.
52. Warmuth MK, Liao J, Ratsch G, et al. Active learning with support vector machines in the drug discovery process. *J Am Chem Soc*. 2003;43:667-674.
53. Serra JM, Baumes LA, Moliner M, et al. Zeolite synthesis modelling with support vector machines-a combinatorial approach. *Comb Chem High Throughput Screen*. 2007;10:13-24.
54. Hand DJ, Yu K. Idiot's Bayes—not so stupid after all? *Int Stat*. 2001;69:385-399.
55. Quinlan JR. Introduction of decision tree. *Mach Leam*. 1986;1:81-107.
56. Kwang-Hwi Choa KTN, Scheraga HA. A polarizable force field for water using an artificial neural network. *J Mol Struct*. 2002;641:77-91.
57. Akbarpour H, Mohajeri M, Moradi M. Investigation on the synthesis conditions at the interpore distance of nanoporous anodic aluminum oxide: a comparison of experimental study, artificial neural network, and multiple linear regression. *Comp Mater Sci*. 2013;79:75-81.
58. Amato F, Gonzalez-Hernandez JL, Havel J. Artificial neural networks combined with experimental design: a “soft” approach for chemical kinetics. *Talanta*. 2012;93:72-78.
59. Butcher JB, Day CR, Austin JC, Haycock PW, Verstraeten D, Schrauwen B. Defect detection in reinforced concrete using random neural architectures. *Comput Aid Civil Infrastruct*. 2014;29:191-207.
60. Maghsoudi M, Ghaedi M, Zinali A, Ghaedi AM, Habibi MH. Artificial neural network (ANN) method for modeling of sunset yellow dye adsorption using zinc oxide nanorods loaded on activated carbon: kinetic and isotherm study. *Spectrochim Acta A*. 2015;134:1-9.
61. Deng L. Deep learning: methods and applications. *Found Trends Signal Process*. 2014;7:197-387.
62. Feinberg EN, Sur D, Wu Z, et al. PotentialNet for molecular property prediction. *ACS Cent Sci*. 2018;4:1520-1530.
63. Wainberg M, Merico D, Delong A, Frey BJ. Deep learning in biomedicine. *Nat Biotechnol*. 2018;36:829-838.

64. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM*. 2017; 60:84-90.
65. Farabet CE, Couprie C, Najman L, LeCun Y. Learning hierarchical features for scene labeling. *IEEE*. 2013;35:1915-1929.
66. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. *IEEE*. 2015;1:1-9.
67. Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Proc Lett*. 2012;29:82-97.
68. Ma J, Sheridan RP, Liaw A, Dahl GE, Svetnik V. Deep neural nets as a method for quantitative structure-activity relationships. *J Chem Inf Model*. 2015;55:263-274.
69. Ciodaro T, Deva D, de Seixas JM, Damazio D. Online particle detection with neural networks based on topological calorimetry information. *JPCS*. 2012;368:012030.
70. Helmstaedter M, Briggman KL, Turaga SC, Jain V, Seung HS, Denk W. Connectomic reconstruction of the inner plexiform layer in the mouse retina. *Nature*. 2013;500:168-174.
71. Leung MK, Xiong HY, Lee LJ, Frey BJ. Deep learning of the tissue-regulated splicing code. *Bioinformatics*. 2014;30:i121-i129.
72. Xiong HY, Alipanahi B, Lee LJ, et al. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science*. 2015;347:1254806.
73. Sutskever, I., Vinyals, O. Le, Q. V. Sequence to sequence learning with neural. arXiv:1409.3215 [cs.CL].
74. Cecen A, Dai H, Yabansu YC, Kalidindi SR, Song L. Material structure-property linkages using three-dimensional convolutional neural networks. *Acta Mater*. 2018;146:76-84.
75. Schutt KT, Sauceda HE, Kindermans PJ, et al. Sch Net—a deep learning architecture for molecules and materials. *J Chem Phys*. 2018;148:241722.
76. Schutt KT, Arbabzadah F, Chmiela S, et al. Quantum-chemical insights from deep tensor neural networks. *Nat Commun*. 2017;8: 13890.
77. Jang K, Kim N, An Y-K. Deep learning–based autonomous concrete crack evaluation through hybrid image scanning. *Struct Health Monit*. 2019;00:1-16.
78. Kim K, Kang S, Yoo J, et al. Deep-learning-based inverse design model for intelligent discovery of organic molecules. *npj Comput Mater*. 2018;4:1-7.
79. Wallach, I., Dzamba, M. Heifets, A. AtomNet a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. arXiv:1510.02855 [cs.LG].
80. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015; 521:436-444.
81. *A Brief Report of the Heuritech Deep Learning Meetup #5*. https://blog.heuritech.com/2016/02/29/a-brief-report-of-the-heuritech-deeplearning-meetup-5/.
82. Hao X, Zhang G, Ma S. Deep learning. *IJSC*. 2016;10:417-439.
83. Zhao G, Huang W, Liang S, Tang Y. Modeling MongoDB with relational model. *Fourth International Conference on Emerging Intelligent Data and Web Technologies*. Xi'an; 2013:115-121.
84. Liu B, Lane I. Joint online spoken language understanding and language modeling with recurrent neural networks. *Proc Sigdial*. 2016;123:22-30.
85. S. Venugopalan, Xu, H. Donahue, J. Translating videos to natural language using deep recurrent neural networks. arXiv:1412.4729 [cs.CV].
86. Agrawal A, Deshpande PD, Cecen A, et al. Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters. *IMMI*. 2014;3:8.
87. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. *Stat Surv*. 2010;4:40-79.
88. Zhang P. Model selection via multifold cross validation. *Ann Stat*. 1993;21:299-314.
89. Burman P. A comparative study of ordinary cross-validation, v-foldcross-validation and the repeated learning testing methods. *Biometrika*. 1989;73:501-514.
90. Atha DJ, Jahanshahi MR. Evaluation of deep learning approaches based on convolutional neural networks for corrosion detection. *Struct Health Monit*. 2017;17:1110-1128.
91. Barton TF, Tuck DL, Wells DB. The identification of pitting and crevice corrosion using a neural network. *IEEE*. 1993;1:325-326.
92. Cha Y-J, Choi W, Suh G, Mahmoudkhani S, Büyüköztürk O. Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types. *Comput AidCivil Infrastruct*. 2018;33:731-747.
93. Gibert X, Patel VM, Chellappa R. Material classification and semantic segmentation of railway track images with deep convolutional neural networks. *IEEE*. 2015;987:621-625.
94. Gibert X, Patel VM, Chellappa R. Deep multitask learning for railway track inspection. *IEEE Trans Intell Transport*. 2017;18: 153-164.
95. Hou W, Wei Y, Guo J, et al. Automatic detection of welding defects using deep neural network. *JPCS*. 2018;933:012006.
96. Janssens O, Slavkovikj V, Vervisch B, et al. Convolutional neural network based fault detection for rotating machinery. *J Sound Vib*. 2016;377:331-345.
97. Jia F, Lei Y, Guo L, Lin J, Xing S. A neural network constructed by deep learning technique and its application to intelligent fault diagnosis of machines. *Neurocomputing*. 2018;272:619-628.
98. Jing L, Zhao M, Li P, Xu X. A convolutional neural network based feature learning and fault diagnosis method for the condition monitoring of gearbox. *Measurement*. 2017;111:1-10.
99. Lin Y-Z, Nie Z-H, Ma H-W. Structural damage detection with automatic feature-extraction through deep learning. *Comput Aid Civil Infrastruct*. 2017;32:1025-1046.
100. Liu L, Tan E, Zhen Y, et al. AI-facilitated coating corrosion assessment system for productivity enhancement. *ICIEA*. 2018; 987:607-612.
101. Meng M, Chua YJ, Wouterson E, Ong CPK. Ultrasonic signal classification and imaging system for composite materials via deep convolutional neural networks. *Neurocomputing*. 2017;257:128-135.
102. Nguyen, T., Zaslan, T. O., Miller, I. D., et al U-Net for MAV-based penstock inspection an investigation of focal loss in multi-class segmentation for corrosion identification. arXiv:1809.06576 [cs.CV].
103. Petricca L, Moss T, Figueroa G, Broen S. Corrosion detection using AI a comparison of standard computed vision techniques and deep learning model. *JCSIT*. 2016;6:91-99.
104. Postolache O, Ramos HG, Ribeiro AL. Detection and characterization of defects using GMR probes and artificial neural networks. *Comput Stand Inter*. 2011;33:191-200.
105. Prabhu DR, Winfree WP. Neural network based processing of thermal NDE data for corrosion detection. *RPQM*. 1993;12: 775-782.

106. Sadowski L. Non-destructive investigation of corrosion current density in steel reinforced concrete by artificial neural networks. *Arch Civil Mech Eng*. 2013;13:104-111.
107. Uruchurtu-Chavarin J, Malo-Tamayo M, Hernandez-Perez JA. J. Artificial intelligence for the assessment on the corrosion conditions diagnosis of transmission line tower foundations. *Recent Patent Corros Sci*. 2012;2:98-111.
108. Wang X, Hu Z. Grid-based pavement crack analysis using deep learning. *ICTIS*. 2017;978:917-924.
109. Zhang A, Wang KCP, Li B, et al. Automated pixel-level pavement crack detection on 3D asphalt surfaces using a deep-learning network. *Comput Aid Civil Infrastruct*. 2017;32:805-819.
110. Zhao R, Wangy J, Yanz R, Mao K. Machine health monitoring with LSTM networks. *ICST*. 2016;978:1-6.
111. Zhao R, Yan R, Chen Z, Mao K, Wang P, Gao RX. Deep learning and its applications to machine health monitoring. *Mech Syst Signal Process*. 2019;115:213-237.
112. Zhao R, Yan R, Wang J, Mao K. Learning to monitor machine health with convolutional bi-directional LSTM networks. *Sensors (Basel)*. 2017;17:1-18.
113. Oh E, Liu R, Nel A, et al. Meta-analysis of cellular toxicity for cadmium-containing quantum dots. *Nat Nanotechnol*. 2016;11:479-486.
114. Sun B, Fernandez M, Barnard AS. Machine learning for silver nanoparticle electron transfer property prediction. *J Chem Inf Model*. 2017;57:2413-2423.
115. Zhu Q, Samanta A, Li B, Rudd RE, Frolov T. Predicting phase behavior of grain boundaries with evolutionary search and machine learning. *Nat Commun*. 2018;9:467-476.
116. Pyrgiotakis G, Kundakcioglu OE, Pardalos PM, Moudgil BM. Raman spectroscopy and support vector machines for quick toxicological evaluation of titania nanoparticles. *J Ranman Spectrosc*. 2011;42:1222-1231.
117. Chandana Epa V, Burden FR, Tassa C, et al. Modeling biological activities of nanoparticles. *Nano Lett*. 2012;12:5808-5812.
118. Zhou Z, Li X, Zare RN. Optimizing chemical reactions with deep reinforcement learning. *ACS Cent Sci*. 2017;3:1337-1344.
119. Pilania G, Wang C, Jiang X, Rajasekaran S, Ramprasad R. Accelerating materials property predictions using machine learning. *Sci Rep*. 2013;3:2810-2816.
120. Schütt KT, Glawe H, Brockherde F, et al. How to represent crystal structures for machine learning: towards fast prediction of electronic properties. *Phys Rev B*. 2014;89:205118.
121. Pilania G, Mannodi-Kanakkithodi A, Uberuaga BP, Ramprasad R, Gubernatis JE, Lookman T. Machine learning bandgaps of double perovskites. *Sci Rep*. 2016;6:19375.
122. Ward L, Agrawal A, Choudhary A, Wolverton C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput Mater*. 2016;2:1-7.
123. Elton DC, Boukouvalas Z, Butrico MS, Fuge MD, Chung PW. Applying machine learning techniques to predict the properties of energetic materials. *Sci Rep*. 2018;8:9059.
124. Dong J, Yao ZJ, Zhu MF, et al. Chem SAR: an online pipelining platform for molecular SAR modeling. *J Chem*. 2017;9:0215.
125. Yang L, Ceder G. Data-mined similarity function between material compositions. *Phys Rev B*. 2013;88:224107.
126. Dey P, Bible J, Datta S, et al. Informatics-aided bandgap engineering for solar materials. *Comp Mater Sci*. 2014;83:185-195.
127. Pankajakshan P, Sanyal S, de Noord OE, Bhattacharya I, Bhattacharyya A, Waghmare U. Machine learning and statistical analysis for materials science: stability and transferability of fingerprint descriptors and chemical insights. *Chem Mater*. 2017;29:4190-4201.
128. Seko A, Maekawa T, Tsuda K, Tanaka I. Machine learning with systematic density-functional theory calculations: application to melting temperatures of single- and binary-component solids. *Phys Rev B*. 2014;89:054303.
129. Castelli IE, Jacobsen KW. Designing rules and probabilistic weighting for fast materials discovery in the Perovskite structure. *Model Simulat Mater Sci Eng*. 2014;22:055007.
130. Legrain F, Carrete J, van Roekeghem A, Madsen GKH, Mingo N. Materials screening for the discovery of new half-Heuslers: machine learning versus ab initio methods. *J Phys Chem B*. 2018;122:625-632.
131. Wei J, Guo F, Wang X, et al. SnO_2-in-polymer matrix for high-efficiency Perovskite solar cells with improved reproducibility and stability. *Adv Mater*. 2018;52:1805153.
132. Waser R, Dittmann R, Staikov G, Szot K. Redox-based resistive switching memories—nanoionic mechanisms, prospects, and challenges. *Adv Mater*. 2009;21:2632-2663.
133. Carrete J, Li W, Mingo N, et al. Finding unprecedentedly low-thermal-conductivity half-Heusler semiconductors via high-throughput materials modeling. *Phys Rev X*. 2014;4:011019.
134. 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;22:3762-3767.
135. Saad Y, Gao D, Ngo T, Bobbitt S, Chelikowsky JR, Andreoni W. Data mining for materials: computational experiments with AB compounds. *Phys Rev B*. 2012;85:104104.
136. Meredig B, Agrawal A, Kirklin S, et al. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys Rev B*. 2014;89:094104.
137. Guimaraes, G., Sanchez-Lengeling, B., Outeiral, C., et al Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models. arXiv:1705.10843 [stat.ML].
138. Gomez-Bombarelli R, Wei JN, Duvenaud D, et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent Sci*. 2018;4:268-276.
139. Lima AN, Philot EA, Trossini GH, et al. Use of machine learning approaches for novel drug discovery. *Expert Opin Drug Discov*. 2016;11:225-239.
140. Faber FA, Hutchison L, Huang B, et al. Prediction errors of molecular machine learning models lower than hybrid DFT error. *J Chem Theory Comput*. 2017;13:5255-5264.
141. Smith JS, Isayev O, Roitberg AE. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem Sci*. 2017;8:3192-3203.
142. Ulissi ZW, Medford AJ, Bligaard T, Norskov JK. To address surface reaction network complexity using scaling relations machine learning and DFT calculations. *Nat Commun*. 2017;8:14621.
143. Hansen K, Montavon G, Biegler F, et al. Assessment and validation of machine learning methods for predicting molecular atomization energies. *J Chem Theory Comput*. 2013;9:3404-3419.
144. Ward L, Liu R, Krishna A, et al. Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Phys Rev B*. 2017;96:024104.

145. Wu Z, Ramsundar B, Feinberg EN, et al. Molecule Net: a benchmark for molecular machine learning. *Chem Sci*. 2018;9:513-530.
146. Kim E, Huang K, Saunders A, McCallum A, Ceder G, Olivetti E. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem Mater*. 2017;29:9436-9444.

AUTHOR BIOGRAPHIES

Jing Wei received her PhD from the Peking University in 2017. She is now an assistant researcher in the Beijing University of Posts and Telecommunications. Her research interests focused on semiconductor materials and their optoelectronic devices, information functional materials and devices, and computational materials science.

Xuan Chu is a master's degree candidate of the Institute of Information Photonics and Optical Communications at Beijing University of Posts and Telecommunications under the direction of professor Kun Xu. He received his BS degree (2016) from the Beijing University of Posts and Telecommunications. He joined professor Xu's group in the summer of 2018 and he is especially interested in machine learning and computational materials science.

Ming Lei received his PhD from the Institute of Physics, Chinese Academy of Science in 2007. He worked as a postdoctoral fellow at the Hong Kong University of Science and Technology and Chinese University of Hong Kong from 2007 to 2008 and from 2009 to 2010, respectively. He is now a professor in the Beijing University of Posts and Telecommunications. His research group focuses on synthesis of low-dimensional semiconductor and related photoelectric properties.

How to cite this article: Wei J, Chu X, Sun X-Y, et al. Machine learning in materials science. *InfoMat*. 2019;1:338–358. https://doi.org/10.1002/inf2.12028