Innovative Materials Science via Machine Learning

Chaochao Gao, Xin Min, Minghao Fang, Tianyi Tao, Xiaohong Zheng, Yangai Liu, Xiaowen Wu, and Zhaohui Huang**

Nowadays, the research on materials science is rapidly entering a phase of data-driven age. Machine learning, one of the most powerful data-driven methods, have been being applied to materials discovery and performances prediction with undoubtedly tremendous application foreground. Herein, the challenges and current progress of machine learning are summarized in materials science, the design strategies are classified and highlighted, and possible perspectives are proposed for the future development. It is hoped this review can provide important scientific guidance for innovating materials science and technology via machine learning in the future.

1. Introduction

At present stage of human society, artificial intelligence (AI) technology is becoming a tendency of developing all over the world. Especially, the successful application in vision recognition,[1] speech recognition,[2] natural language understanding,[3] and man-machine game[4] also accelerates the integration of different knowledge and the cross of multi-disciplinary. For example, the so-called AlphaGo Zero achieved a score of 100:0 in the game with the previous champion AlphaGo Lee in 2017;[5] A Nature Biotechnology paper published in 2019 reported six molecules based on a generative model, four of which have measurable biochemical activity;[6] One Nature cover article reported a tireless mobile robotic chemist in 2020, which was able to operate autonomously over eight days, performing 688 experiments within a ten-variable experimental space, and finally selected a novel and efficient photocatalyst.[7]

Generally, the implementation process of AI would be divided into three stages: handwritten knowledge, statistical learning, and context adaptation,[8] which is currently in the statistical learning level. The key core technology depends on statistical machine learning, such as deep learning,[9] reinforcement learning,[10] adversarial learning,[11] etc. which could double improve the research efficiency in the fields of information data mining, data classification, and new data prediction. In present, AI based on this machine learning process has become the mostly important driving force for science and industrial revolution.

Material innovation always plays a great role in the process of industrial revolution and human social development.[12] For example, the advanced alloys, semiconductor materials, polymer materials, composite materials, superconducting materials and biocompatible materials have been promoting the technology revolution of new energy, microelectronics, bioengineering technology and space technology, which opened the door to the unprecedented information society.[13,14] In the nearly future, the develop of novel material science and technology is accelerating to the interconnected and intelligent direction, which will further promote the fourth industrial revolution – green intelligent industry.[15,16] In comparison with the traditional experiment- and computation-driven research methods for materials science, the data-driven mode, which integrates high-throughput experiments, high-throughput computing and material data based on data mining and AI, would be more revolutionary in the future development.[17–20]

As one of the most essential AI methods, machine learning is becoming an excellent tool for material innovation due to its low computing cost, short development cycle, strong data analysis and prediction ability (**Figure 1**). Nowadays, the machine learning has been used and shown great potential in materials prediction for magnetocaloric effect,[21] bandgap,[22] dielectric constant,[23] quantum chemistry,[24] thermal properties,[25] new materials design, and discovery.[26] Summing up the present development of data-driven machine learning, it is still at primary stage in material innovation. As far as we know, few comprehensive review papers on machine learning and material applications have attracted close attention in recent years, focusing on machine learning in chemical discovery, energy materials, solid-state materials science and so on.[17–20] Recently, there has been a sharp increase in the number of articles on machine learning and materials science, which suggests that the research direction and content have been widely extended. Thus, it is significantly important to update the summarize the prospective in this field. Herein, this review attempts to highlight the main challenges of machine learning in materials science, focus on the research progress of existing strategies, and possible directions and perspective of machine learning in

C. Gao, X. Min, M. Fang, X. Zheng, Y. Liu, X. Wu, Z. Huang
Beijing Key Laboratory of Materials Utilization of Nonmetallic Minerals
and Solid Wastes
National Laboratory of Mineral Materials
School of Materials Science and Technology
China University of Geosciences (Beijing)
Beijing 100083, China
E-mail: minx@cugb.edu.cn; huang118@cugb.edu.cn

T. Tao, X. Zheng
Division of Environment Technology and Engineering
Institute of Process Engineering
Chinese Academy of Sciences
Beijing 100190, China

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/adfm.202108044.

DOI: 10.1002/adfm.202108044

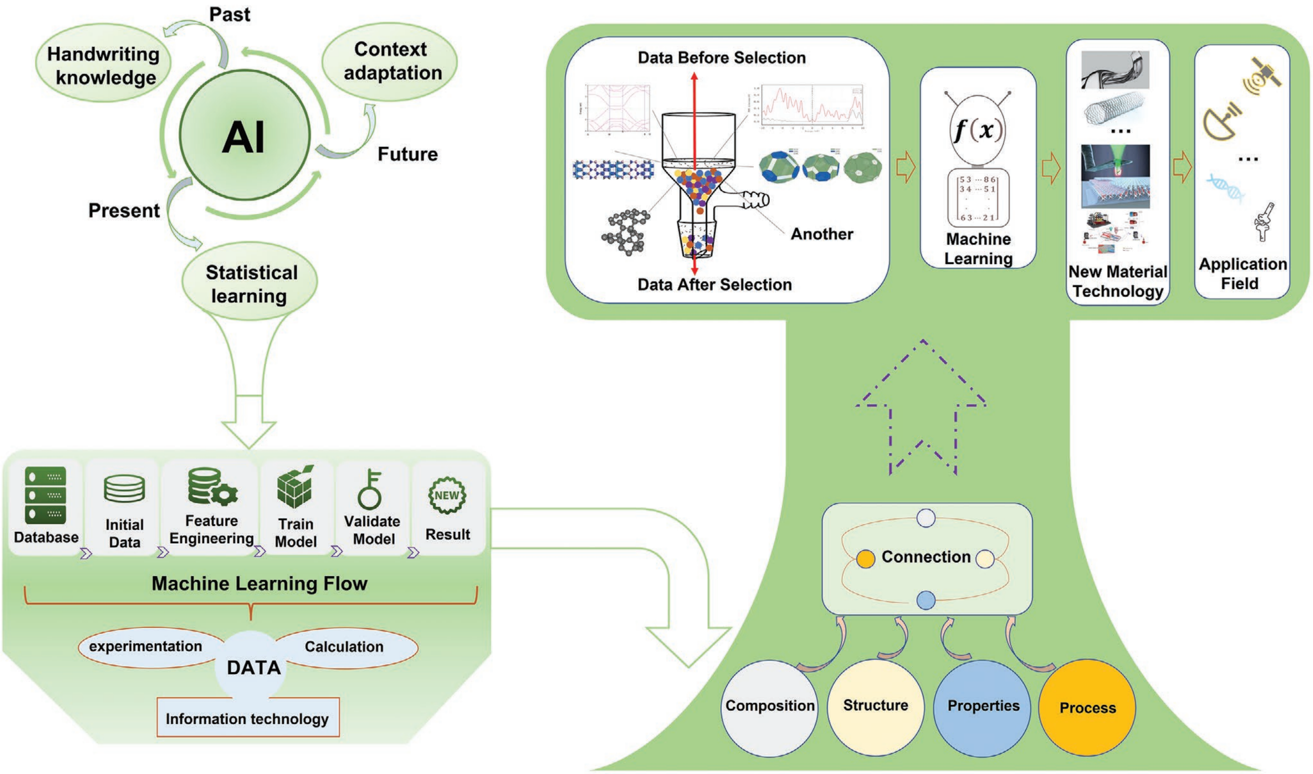

Figure 1. The introduction of AI and machine learning in materials science.

materials innovation. Our main objective is to illustrate main strategies for machine learning in materials, provide possible directions for next-generation machine learning, and suggest methodologies for the future development of materials sciences.

2. Challenges of Machine Learning in Materials Science

In the development process of materials science, the traditional methods slow down the pace of materials research because of the long research cycle, potential data information loss and low efficiency. The up-to-date machine learning expectations based on material data start from three components, including data, algorithms, and models. Firstly, the material data should be collected and obtained, then preprocess these data. For different material data, the operations, such as encoding conversion, feature crossover, calculation optimization, filtering and deleting, may be performed, and the optimized data would be divided into training set and test set. After that, the appropriate algorithm model for machine learning would be selected and trained based on the training set. The test set is used for evaluating and adjusting the algorithm model to achieve the best evaluation. Finally, the optimized algorithm model would be effective for further data prediction to obtain the desired target. Therefore, combined with the machine learning technology, the efficiency of materials research could be significantly improved. For example, machine learning can give some guiding conditions in many fields for predicting crystal structure, understanding of thermal properties of amorphous solids, identification of high temperature thermal conductors, classification of crystal structure, magnetocaloric effect, etc.[27–54] As well as the promotion of the material Genome Project,[55] machine learning has been combined with density functional theory (DFT) calculation to establish a development model for material research, such as prediction of thermodynamic stability, bandgaps, AB_2C_2 compounds,[44,56–58] and graphene-based bimetallic catalysts.[59,60] Furthermore, machine learning could even replace DFT to achieve the desired goals, including the prediction of crystal structure, adsorption energy on metal alloys and so on.[40,43,61] Besides, the failed experimental data about crystallization of templated vanadium selenite was also been used for information mining by machine learning.[62] Although the cross-research of materials science associated with machine learning has shown special advantages, there are still several challenges (**Figure 2**) which need to be properly address and deeply understood as shown below.

2.1. Insufficient Data

Data is not only the basic premise of the fourth paradigm of materials science, but also the first challenge to be solved in application of materials.[63] Unlike most disciplines, materials science and technology owns many different data categories. More severely, the data output of each category is less, and the feature dimension is lower. For example, even a simple experimental data always depends on different controllable factors, such as raw materials, contents, temperatures, times, humidity,

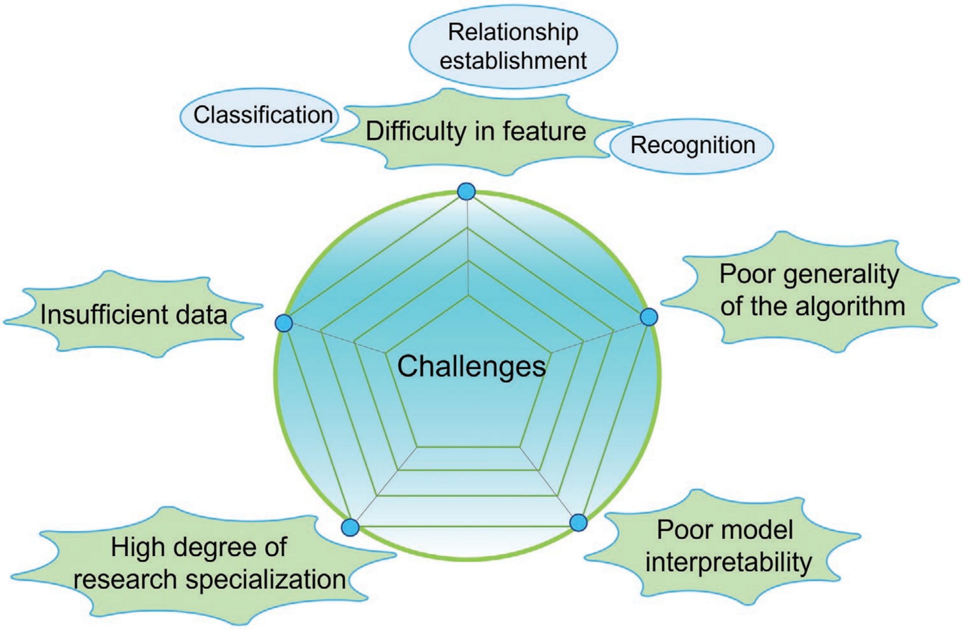
Relationship establishment
Classification
Difficulty in feature
Recognition
Poor generality of the algorithm
Insufficient data
Challenges
High degree of research specialization
Poor model interpretability

Figure 2. Summary of challenges for the machine learning in materials science.

etc. In most studies, the research data are obtained via limited experimental conditions, which takes a long time in the meantime. Thus, these measured results are just a small sampling, which also lacks comparability in different published studies. It could be concluded that sufficient data can not be obtained quickly to meet the operation of machine learning, which limit their further development in materials science.

2.2. Difficult to Identify, Classify, and Correlate the Features

In the process of material research, the key point is to explore the relationship among the four elements of materials, which could be controlled by the features of experimental parameters.[64–66] The quantity, quality, form, and relationship of these features are essential for machine learning. In general, the initial data features can not fit the ideal service conditions for machine learning, thus the feature engineering operations, such as adding, deleting, and modifying, are necessary. Herein, four challenges related to features are shown as below:

1) Processing difficulties because of the large number of features. With the promotion of high-throughput computing[67] and Materials Genome Initiative (MGI),[68] the number of features, as well as the noise and redundancy of the original data, continue to increase. Finally, the computational and storage became more and more complex, so that the traditional learning methods are no longer suitable to gain satisfactory results.
2) Challenges to explore the relationship among various features. It is mentioned before that researcher can get many data with a variety of features via experiments. When studying the performances, the control variable method is used to explore the optimized parameters. However, it is complicated to establish the relationship in multiple features.
3) Difficulty in feature classification. Since the relationship among the features remains unclear, their influences on materials science should be systematically discussed. When combining with the machine learning technology, it is still difficult to scientifically clarify these features as needed.
4) Difficulty in feature recognition. The initially collected data often cannot be directly identified for machine learning. Many modifying strategies are needed to convert them into data features to be well recognized. However, each conversion is different, which also makes the feature recognition become a huge limitation of data-driven development in materials science.

2.3. The Specific Algorithm with Poor Universality

Machine learning shows different abilities based on various algorithms, for example, different results would be obtained by the same algorithm to deal with different objects or different algorithms to deal with the same object. In materials science, the machine learning algorithms needs to be especially considered in different directions because of the rich research fields, great differences, and small relevance. Moreover, the performance optimization, structure regulation, and some other directions also should be considered during the new materials development, so that the selection and optimization of algorithms is an issue of concern. In the field of materials science, the relationship and interaction mechanism between the four elements “component-structure-performance-application” usually needs to be explored firstly, which could provide theoretical support for the materials research.[69–71] However, that raised much higher challenges to the targeted selection and optimization of machine learning algorithms. Therefore, different objects and different situations demand different algorithms without universality.

2.4. The Abstract Model with Poor Interpretability

The deep neural network, which belongs to the representative machine learning model, has developed rapidly and shown powerful ability in intelligent products. For example: self-driving cars,[72] image recognition,[73] dialogue systems.[74] However, the internal working mechanism is complicated to understand the decision-making model in human inborn thought processes. In the field of materials science, machine learning is always regard as a "black box" and the principles behind data-driven are ignored in most studies. For strict research for materials, the prediction of properties might not be well trusted in this situation. Thus, in addition to final decision, both the knowledge that learns from the data and the features that attributes most to the results should be presented in detail. Therefore, the interpretability of the machine learning model is also the focus of attention in AI for materials science.

2.5. The Highly Specialized Research in Materials Science

With the trend of AI technology, many researchers begin to combine the materials science with machine learning intelligent technology, relying on the data-driven strategy to predict and analyze material properties.[75–77] As known to all, the code programs should be firstly designed to achieve the goals for machine learning, which includes both the feature engineering and model building after data collection. Many programming languages can be used for machine learning, such as the most popular one "python" with simple but rich expansion package. However, due to the lack of deep mathematical theory and programming ability for researchers in materials science, the progress of machine learning in materials still going slowly.

3. Strategies for Machine Learning Applied in Materials Science

3.1. Increasing the Amount of Data

Since the foundation in 2011, the Materials Genome Initiative (MGI) continually accelerates the pace of material discovery, design and deployment by training to collaborate experiment, theory and computing for data generation, analysis and sharing.[78] Zhang et al. combined the MGI strategy with high-throughput measurement and CALPHAD software to accelerate the development and design of new biological titanium alloys.[79] MGI strategy is used to enhance the mining of insensitive high energy density materials by Zhang et al.[80] The authors discussed how a materials genome approach could be used to accelerate the discovery of promising insensitive high explosive (IHE) molecules. By rationalizing the relationship between structure and properties, the "genetic" features are firstly identified and extracted. Then, the computation-guided molecular designing and rapid screening are carried out, which includes library construction of molecular fragments and structural selection of candidate molecules based on filter conditions of "genetic" features. Finally, the ideal target molecule, 2,4,6-triamino-5-nitropyrimidine-1,3-dioxide, is obtained and successfully synthesized, which exhibits a high measured density, high thermal decomposition temperature, high detonation velocity, and extremely low sensitivities. While Pablo and the co-workers put forward the application of information materials, functional materials, energy and catalytic materials according to MGI in ref. [68]. Through the MIG strategy, both the data generation and the ability to identify material attributes are strengthened in cooperation with machine learning. Besides, several strategies, including high-throughput calculation,[67] DFT calculation[81] and traditional experimental data accumulation in published papers,[27] also provide a large number of valuable data for machine learning in materials science. High-throughput and DFT calculation can provide abundant data without the cost of human resources in experiment, so that more theoretical data could be directly produced in a limited time.[57,82] At the same time, there are many open high-throughput material databases nowadays, such as the open quantum material database,[83] open inorganic material database,[84] crystallographic open database,[85] thermoelectric open data resource,[86] two-dimensional material database,[87] novel material discovery database,[88] high-throughput combination database of electronic band structure for inorganic scintillator materials,[89] inorganic amorphous database,[90] and so on. All these open databases could provide much more conveniences for researchers in machine learning. Moreover, Dong et al. develop a deep neural network (DNN) to predict material defects based on small data, which also achieve very good results.[91] By above strategies, the open-source of database for materials science has gradually increased for machine learning in recent years, as well as the types of databases. This could provide more choices for machine learning research in materials and promote more researchers in different fields to participate. However, the collection efficiency for databases still needs to be further improved, and more strategies should be developed. Meanwhile, optimizing the algorithm model for large sample data to promote the improvement of machine learning capabilities and making more breakthroughs for small sample data to adapt machine learning are still the main challenges.

3.2. Dealing with the Perplexity of Research in Features by Machine Learning

As shown in **Figure 3**, the way to deal with the perplexity in features could be concluded in feature selection and dimensionality reduction with a subsequent machine learning. In addition, the feature contact information mining, classification, and reconstruction based on machine learning are also summarized herein.

3.2.1. Strengthening the Analysis and Optimization on Features

The methods by selecting and transforming multi-dimensional features into low-dimensional features have been proposed to improve the quality of features, reduce computational complexity, and improve recognition accuracy.[92–98] In present machine learning feature engineering, these two powerful technologies have shown great potential in various applications,

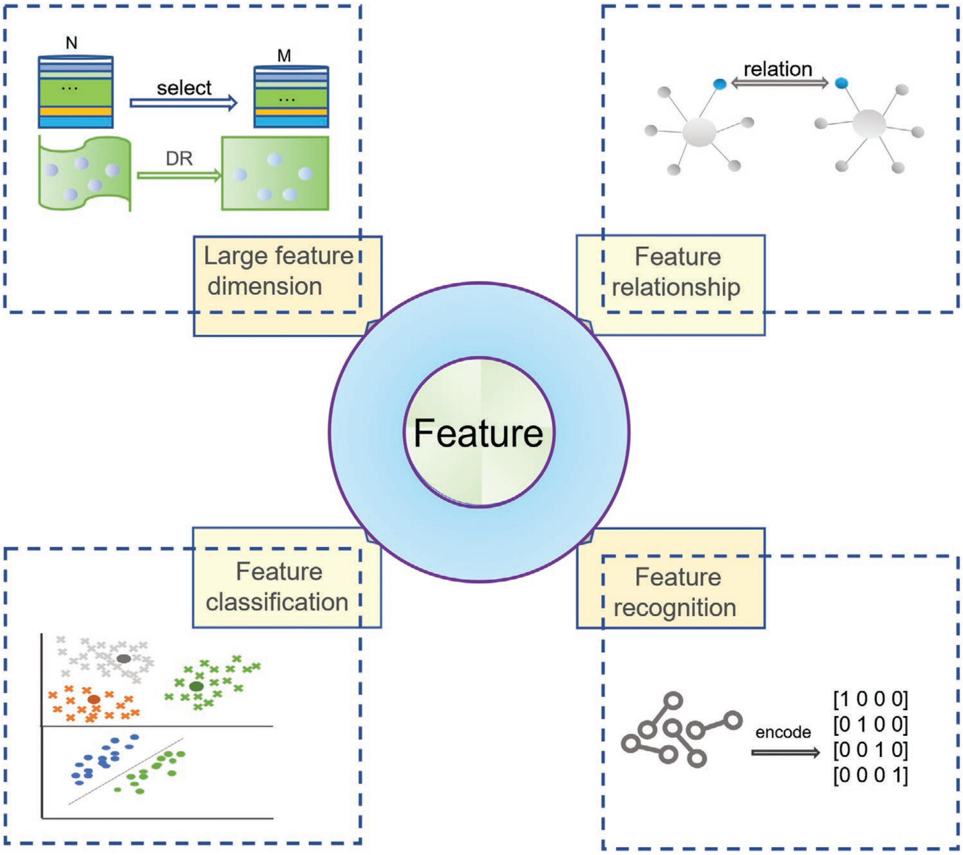

Figure 3. The strategies to deal with perplexity of research in features.

such as the text data,[99] planetary system metrics,[100] clinical medicine,[101] human hand motion classification,[102] human activity recognition,[103] etc. In materials science, the transformation in various features is also important, which might directly affect the analytical accuracy in structure and properties. The familiar machine learning algorithms, e.g. support vector machine and clustering,[104,105] have exhibited some solutions for feature selection. For example, Brgoch et al. selects and classifies the features of structure and polymorphism of equiatomic ternary ABC phase based on these two methods.[105] In predicting the performances of zeolites, Evans's group constructs a machine learning method to automatically predict the mechanical properties without any calculation of chemical properties, just based on their local geometry, structure, and porosity features.[106] According to the exceptional insight into the mechanics of zeolitic frameworks, the development of this model has highlighted the correlations between characteristic features and elastic properties, which is further trained with a DFT data set. This methodology is employed to predict the elastic response of 590 448 hypothetical zeolites, which are more accurate than those calculated by the force field method. Besides, some methods can not only find the minimum of the surface free energy, but also screen out the special structure to obtain these properties, develop novel feature classification methods, and propose complete structure descriptors,[107] which could overcome the problems that can not be realized by common descriptors. Choudhary et al have proved that the combination of paired radial, nearest neighbor, bond angle, dihedral angle and core charge distribution plays a vital role in predicting formation energy, bandgap, static refractive index, and magnetic properties via this similar method.[108] In material research, the irrelevant and redundant features are also eliminated by feature selection in most cases, to improve the accuracy of the model and reduce the running time. Meanwhile, the real relevant feature simplification model could be selected to help understanding the process of data generation. In the publication reported by Adibi, the dimensionality reduction technique has been used to reduce the dimension of the response space while keep the number of identical points in the response space for electromagnetic nanostructures design and optimization.[109] Thus the dimension of these features are reduced based on the correlation between them, which could reduce the computational complexity and avoid more mistakes.

3.2.2. Exploring the Relationship among Features

In present research of machine learning in materials science, the built-in methods are often used to explore the relationship among different features. Artrith and co-workers automatically construct atomic interaction potentials by a Behler-Parrinello approach, which is based on artificial neural networks (ANNs).[110] They try to establish a free and open-source atomic energy network package for modeling electrocatalytic CuAu and

Cu-doped ceria nanoparticles in water. Taking the well-studied TiO_2 as the example, the construction of realistic structure models is optimized in consideration of defected crystals and nano structures for catalytic activity. Thus, the lattice parameters, energies, and bulk moduli are accurately predicted by TiO_2 ANN potential. The researchers also propose a generalized crystal graph convolution neural network framework to present the periodic crystal system at the example of perovskite.[42] By predicting the properties according to the relationship between atoms in crystal, a theoretically optimal crystal material could be obtained. The neural network potential and Gaussian approximate potential have also been used to simulate the interatomic potential.[29] After unifying the different parameter factors into a free energy by the graph neural network model, the relationship between the features could be automatically linked together associating with the information mining process, which finally help to predict performances.[111] Furthermore, machine learning has also shown significant value in constructing the correlation between physical features in phase change material system.[112]

3.2.3. Classifying Features into Different Categories

Machine learning has also been used for data classification to provide more convenience and improve the learning efficiency. In ref. [31], Sohn et al. reported a deep machine learning technique based on convolution neural network, which was used to classify powder X-ray diffraction (XRD) patterns according to crystal system, extinction group, and spatial group. The XRD data are gained from the inorganic crystal structure database. Meanwhile, convolution neural network is also used to classify the crystal structure automatically by Ziletti et al.[32] The crystal structures are firstly transformed into diffraction patterns, then small amounts of data are extracted as the training set to generate a final classification model. Besides, visualization is also developed in the neural network's internal operation to ensure the classification decision. Finally, through the deployment of crystal structure classification model, the highly efficient classification is realized without any additional model optimization. These studies would pave the way to optimize the noise and incomplete three-dimensional structure data for machine learning in materials science.

3.2.4. Recognizing the Attributes of Material Features

Because of the particularity of material data and features, the data was always encoded and transformed to be easily identified by machine learning.[34,113] For example, Dimiduk et al. simplified the material topics into digital fingerprint vectors, and then developed appropriate measure of chemical (dis)similarity or chemical distance in the learning scheme to map the features' distance.[114] The principle is to encode and map the attributes to form the basis of further operation. Besides, Chen and co-workers also propose a training-saving method based on transfer learning, combining the encoder-decoder process via deep convolution network and feature matching optimization.[115] This method mainly aims at the microstructure of a single given target, which firstly encodes the initial microstructure, and then pre-train through the convolution neural network to further obtain the optimized microstructure. After unsupervised learning, the required reconstructed microstructure is finally obtained.

Looking at the above strategies at this stage, various feature engineering methods are usually used to deal with the feature challenges. Automatic selection of features based on deep learning and automatic coding could make the feature processing much easier. However, these are still in the static space stage, and the relationship among the material phase, composition, morphology, and performances still cannot be well studied in parallel at the same time. Due to the particularity, variability, polymorphism, and uncertain factors of materials science, more attention should be paid to obtain the real relationships in dynamic space in the future. For data processing methods, the internal response of data preprocessing to material properties and their intersection should also be considered with more influencing factors.

3.3. Developing Professional Algorithms for Materials Science

In order to fully solve the challenges that machine learning algorithms are not universal to materials science, many researchers draw lessons from the successful cases or empirical analysis in similar fields and choose the appropriate algorithm model to efficiently optimize the material object. As shown in **Figure 4**, the currently popular machine learning algorithms are summarized, which could be divided into four major directions: classical learning,[116] reinforcement learning,[117–119] ensemble learning,[120,121] neural network and deep learning.[122]

These four kinds of machine learning algorithms have been widely concerned by researchers in materials science. As shown in **Table 1**, ten classical algorithms are applied for materials information mining with huge application potential in machine learning.[116] Among them, the decision tree and SVM model are the main classification algorithms. The advantage of decision tree model is that no domain knowledge or parameter setting is needed in the construction process, so that the decision tree is more suitable for knowledge discovery in material research. While the SVM could avoid over-fitting to some extent, since does it only depends on support vectors. Thus, SVM is still effective in medium to small samples. Riley et al. reported a Molecule Deep Q-Networks model for molecular optimization by combining chemical knowledge and reinforcement learning,[123] which exhibit advantages including higher sampling efficiency and better molecules exploitation. Sparks et al. simulate the experimental data effectively by an ensemble learning method by merging different data sources into the modeling of sparsely represented experimental data. In the case of bandgap prediction, the root mean square error could be reduced by over 9%.[57] Yu et al. reported a decision tree-based ensemble learning model to accurately predict the material removal rate for chemical mechanical planarization.[124]

Apart from the three categories mentioned above, the neural network and deep learning are currently the focus of most attention. In 2006, Hinton et al. alleviated the local optimal solution problem by using a pre-training method, which pushed

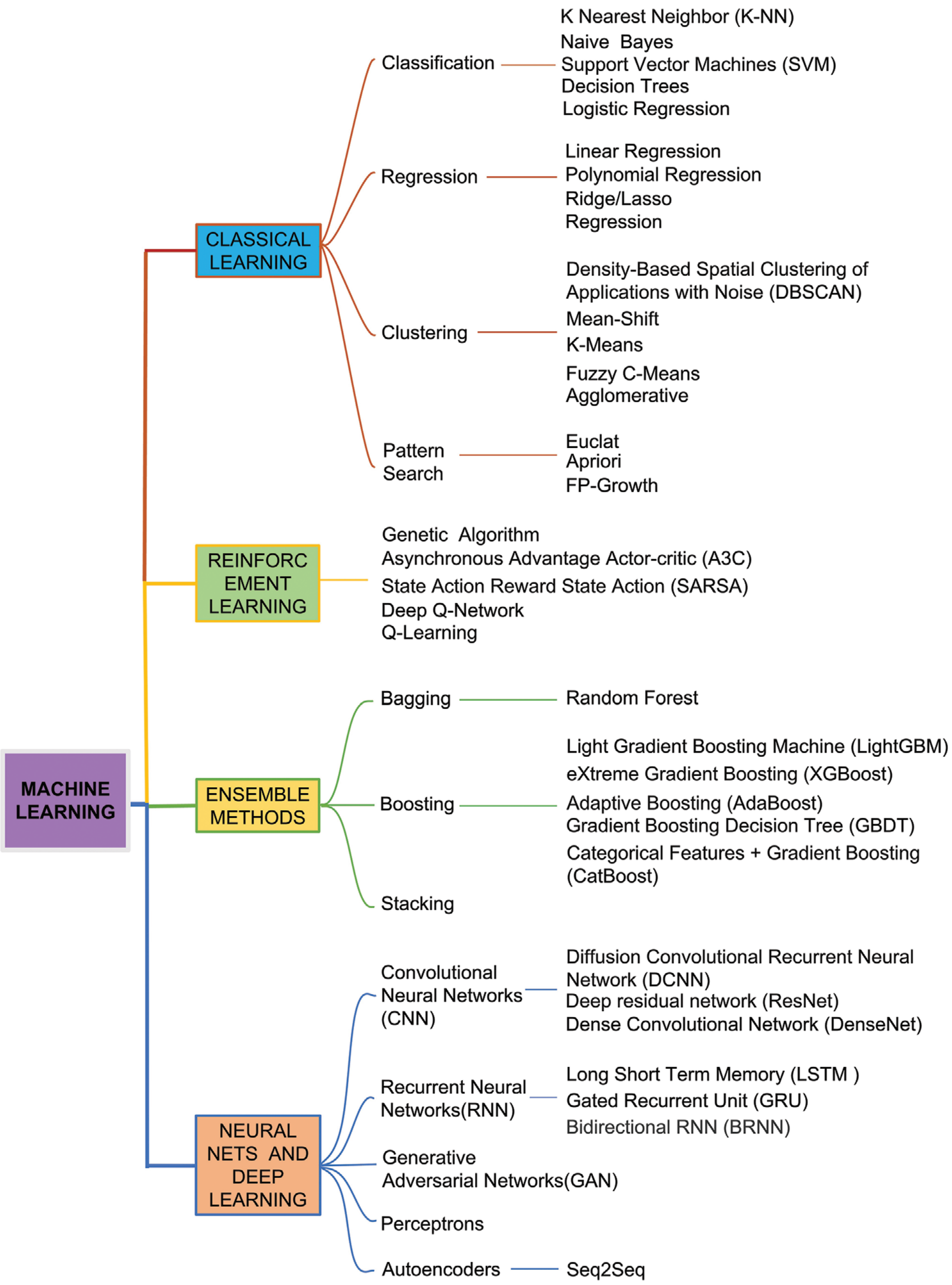
MACHINE LEARNING
CLASSICAL LEARNING
Classification
K Nearest Neighbor (K-NN)
Naive Bayes
Support Vector Machines (SVM)
Decision Trees
Logistic Regression
Regression
Linear Regression
Polynomial Regression
Ridge/Lasso Regression
Clustering
Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
Mean-Shift
K-Means
Fuzzy C-Means
Agglomerative
Pattern Search
Euclat
Apriori
FP-Growth
REINFORCEMENT LEARNING
Genetic Algorithm
Asynchronous Advantage Actor-critic (A3C)
State Action Reward State Action (SARSA)
Deep Q-Network
Q-Learning
ENSEMBLE METHODS
Bagging
Random Forest
Boosting
Light Gradient Boosting Machine (LightGBM)
eXtreme Gradient Boosting (XGBoost)
Adaptive Boosting (AdaBoost)
Gradient Boosting Decision Tree (GBDT)
Categorical Features + Gradient Boosting (CatBoost)
Stacking
NEURAL NETS AND DEEP LEARNING
Convolutional Neural Networks (CNN)
Diffusion Convolutional Recurrent Neural Network (DCNN)
Deep residual network (ResNet)
Dense Convolutional Network (DenseNet)
Recurrent Neural Networks(RNN)
Long Short Term Memory (LSTM)
Gated Recurrent Unit (GRU)
Bidirectional RNN (BRNN)
Generative Adversarial Networks(GAN)
Perceptrons
Autoencoders
Seq2Seq

Figure 4. The classification for machine learning algorithms.

the hidden layer to 7 layers.[149] After that, the neural network become the real depth learning, such as the subsequent Deep Belief Nets,[150] Convolutional Neural Network,[151–153] and Recurrent neural network.[154–156]

The rapid development of deep learning has also solved many thorny problems in materials science. For example, Rao and Liu reported a three-dimensional depth convolution neural network to predict the anisotropic effective material properties for representative volume elements (RVEs) with random inclusions.[157] A radial basis function artificial neural network model is developed by Ince et al. to predict the propagation and growth behavior of fatigue crack.[158] In ref. [48], Agrawal et al. established a highly accurate materials property prediction model by leveraging the large DFT-computational data sets, smaller DFT-computed data sets, and available experimental observations. Besides, the neural networks and deep learning methods have also demonstrated their strong ability in many other material research.

Although the diversity of machine learning algorithms could provide much selectivity, the algorithms applicable to materials science are still limited since it is obviously different from the time series and context in the text. Moreover, it also contains many indeterminate factors, such as ion valence and conditional chemical reaction, which need more hyperparameters

Table 1. Ten common algorithms of machine learning for application in materials science.

Algorithm type	Algorithm Model	Examples in Materials Science
Classification algorithm	C4.5[125]	Analysis of the causes of Coffee defects by decision Tree[126]
	Naive Bayes[127]	Classification of metal binders[128]
	SVM[129]	Material monitoring and defect diagnosis[130] Prediction of rock brittleness[131]
	KNN[132]	Prediction of process parameters of reinforced metal casting[133] Analysis of welding modeling of different materials[134]
	Adaboost[135]	Temperature compensation of Silicon Piezoresistive pressure Sensor[136]
	Cart[137]	Differential diagnosis of mucosanase[138]
Clustering algorithm	K-Means[139]	Structural texture similarity recognition of materials[140] Establishment of parametric homogenized crystal plasticity model of single crystal Ni-base superalloy[141]
	EM[142]	Estimation of dose distribution from positron emitter distribution combined with filtering[143]
Correlation distribution	Apriori[144]	Identify the frequency trajectory of material transportation[145]
Connection analysis	PageRank[146]	Measurement of hyperelastic materials[147] Remote protein homology detection[148]

to be set for algorithm model. At present, most materials researchers rely on the existing algorithms, traditional machine learning models, and simple neural networks. In the future, more suitable algorithm models should be developed in actual learning process for materials science.

3.4. Optimizing and Evaluating the Learning Models

The most straightforward interpretability for machine learning is using the interpretable models, such as logical regression, linear model, and decision tree.[159] Ribeiro et al. reported the LIME explanation technique to explain the predictions of any classifier in an interpretable and faithful manner by learning an interpretable model.[160] Besides, by transforming the task framework into the sub-module optimization, it could help to decide if one should trust a prediction, choose between different models, improve an untrustworthy classifier, and identify why a classifier should not be trusted. Murdoch et al. tried to address these concerns by defining interpretability in the context of machine learning, and the Predictive, Descriptive, Relevant framework are also introduced for interpretations.[161] Yuma Iwasaki and co-workers predicted the spin-driven thermoelectric materials with anomalous Nernst effect by an interpretable machine learning method called factorized asymptotic Bayesian inference hierarchical mixture of experts (FAB/HMEs), according to the prior knowledge of material science and physics. After practical material synthesis, a new type of spin-driven thermoelectric material is finally demonstrated.[162] Moreover, Kailkhura et al. proposed a new evaluation metric and a trust score to better quantify the confidence in the predictions, associating with a rationale generator component to provide both model-level and decision-level explanations. As a result, the properties of crystalline compounds and potentially stable solar cell materials could be primely predicted.[163] And other studies on interpretable machine learning[164–166] can also provide reference to machine learning for principal research in materials. However, the research of machine learning in materials science is still insufficient, and the interpretability is also a problem, which may lead to the omission of the internal influence factors of deep learning on materials research in the future.

3.5. Developing Open-Source Material Packages and Machine Learning Frameworks

To promote the application of machine learning in materials science, many open-source machine learning tools have been provided. For example, SchNet is a deep learning framework for molecular materials, which is specifically designed to model atomic systems by using continuous filter convolution layers.[167] DScribe is a software package for machine learning that could provide widespread feature transformation for atomic material simulations.[168] Matminer is an open-source software based on Python, which can analyze and predict material properties by the data-driven method.[169] Pymatgen is a robust open-source Python Materials Genomics library for material analysis,[170] which could the initial setup and original calculated data for high-throughput computing materials science. COMBO designs an effective Bayesian optimization scheme, which is also an as an open-source python library by combining Thompson sampling, random feature graph, first-order Cholesky update and automatic hyperparameter adjustment.[171] COMBO is available on the website (https://github.com/tsudalab/combo). AFLOW-ML provides an open RESTfulAPI with direct access to continuously updated algorithms, which can be transparently integrated into any workflow, such as electronic retrieval and thermal or mechanical performances prediction.[172] In addition to the open-source packages mentioned above, the open-source framework including TensorFlow,[173] Pytorch,[174] Keras,[175] Scikit-Learn[176] can also reduce the time of building models from the beginning and improve work efficiency.[177] For example, the python library-based Scikit-Learn is used to predict thermal stability of perovskite materials by machine learning, feature selection, and model evaluation.[44] We believe that these types can further accelerate the adoption of machine learning methods in material development, if the open-source material packages and machine learning frameworks could be effectively connected by the cloud-based interconnected applications. At present, in the research of machine learning in materials science, the materials-related open-source toolkits and programming language frameworks have been well designed by programming tools, which can provide great convenience for non-professional programming researchers, such as materials researchers. However, more development in open-source toolkits are still needed in all directions of materials science to promote the further popularization of feature engineering and machine learning.

4. Summary and Perspectives

After the above discussion, we fully believe that machine learning exhibits huge ability to mine new materials in the data driven era. Based on the challenges of machine learning in materials, the current solutions and research progress have been summarized and discussed. Although these pioneering studies have been conducted to promote the machine learning in materials science, novel algorithm model technologies, efficient data preprocessing methods, information mining, material structure and performance prediction, and new function prediction, as well as the interaction among these features, should be further prospected to improve work efficiency and scientific research progress in materials science. Thus, it is believed that notable advances still need to be developed to meet the requirements of practical applications. Herein, as shown in **Figure 5**, we outline several possible directions for machine learning and hope that these perspectives might be useful for researchers in materials science.

4.1. Machine Learning with Small Sample Size

In the process of material research, it will be a huge step if the studies could be impelled by a small amount of data. For example, the DNN regression has been used to predict defects and solidification cracking susceptibility of stainless steels based on 487 data points.[91] In future study, DNN could also be trained and tuned to convert the scattered small dataset into high-precision maps in high-dimensional chemical, processing parameter, and performances space. In addition, in order to learn from the supervised information of limited samples, zero shot learning (ZSL),[178] one shot learning (OSL),[179] and few shot learning (FSL)[180] would be mostly potential research objects for small sample learning because of the learning based on a small sample size. The typical example in character generation indicates that a computer program is required to break the characters into smaller parts, which can be transferred among each character, and then aggregate these smaller components into

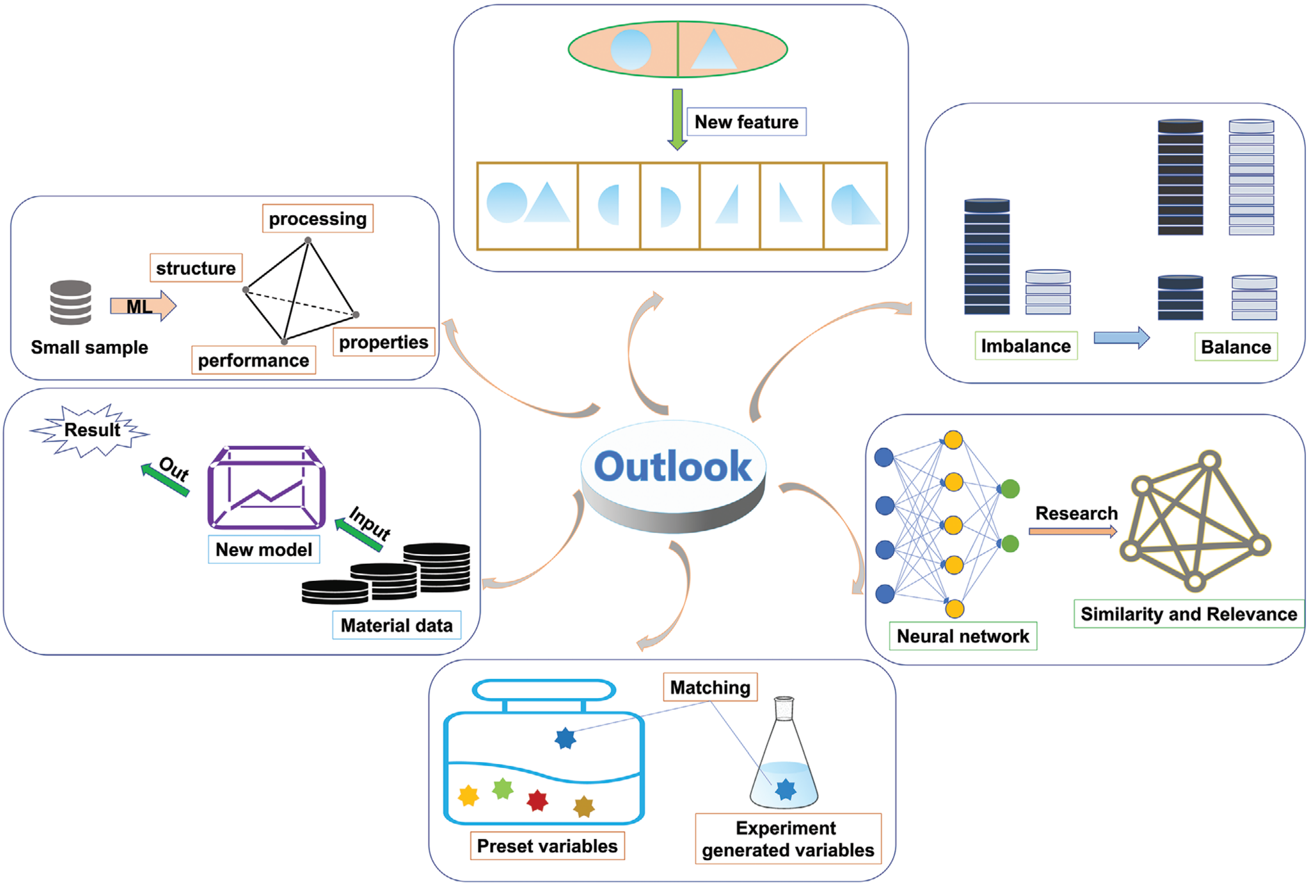

Figure 5. The perspective of machine learning in materials science.

new characters.[181] We think this effective strategy could also be similar for potential application in material detection, defect prediction, classification and so on. Accordingly, the major problems of machine learning in materials science will be solved, such as learning like human beings, learning rare situations, reducing the sample collection and computational costs.

4.2. Constructing New Features by Combining Multiple Features

Resembling the deep learning of graphics and audio synchronous recognition in the time dimension, the multi-feature combination effects among phase, structure, performance, and application based on time and space dimensions should also be developed during the whole life cycle of material prediction, material synthesis, material service, recycling, and reusing.[182] If the number of size parameters is too large, the common matrix decomposition can be used as fair idea for reference. The combination of multi-features can not only expand the amount of sample data, but also improve the fitting ability of the complex relationship of experimental features. Here are our recommendations for multi-features combination. Firstly, the initial features should be linearly combined to form new features, and the influence of combined features on the whole research process should be explored. In addition, the initial features could also be disassembled into new features, and then add them into other features to form new features.

4.3. Controlling Data to Ensure Feature Balance

Due to the inevitable data bias of the material experiment and the neglect of the collection method, the problem of data imbalance always occurs in the process of data collection which would further lead to the situation of judgment confusion. Herein, we try to provide some suggestions to solve the unbalanced data. The first one is data sampling, which could be divided into up-sampling and down-sampling. Up-sampling is the process of duplicating a small amount of data to equalize the proportion of each category, in which adding random disturbance into the newly generated data is necessary to overcome the overfitting. While the down-sampling is to select a small part from most categories to keep the data proportion of each category at a normal level, in which multiple random samples are required to ensure the integrity of information. The second one is data combination, which could generate more samples by using the similar features of existing samples. The third one is weighting, which refers to applying different weights to each samples according to the corresponding errors, so that machine learning could focus more on fewer and error-prone samples. And finally, one classification, such as One-class SVM,[183] is also a very key strategy for controlling data to ensure feature balance, when the proportion of positive and negative samples is out of balance.

4.4. Discovering the Relevance and Similarity between Features

In materials science, many factors could affect the features of labeled training samples, such as different preparation methods, preparation parameters, etc. The limited correlation between different samples and the lack of unified standard contribute to a poor comparability. Thus, the exploration on the relevance and similarity between features is still very important in materials science. We can try to establish the relationship between nodes and edges by using graph neural network to explore the relationship between the child nodes and some others.[184] The materials can be divided into different categories, and each category has many different research directions, which could have many similarities such as time, space, structure, and performance. Furthermore, the transfer learning to transfer knowledge from the source domain rich in training data to the target domain lacking in training data or clustering in graph embedding might also be effective for seeking similarity.

4.5. Presetting Experimental Variable Parameters

There are always many uncertain parameters in material research, such as temperature, humidity, time and so on, which would affect the phase composition and morphology of the final materials. Traditional machine learning also has limitations in dealing with the related features. Therefore, it is necessary to provide pre-parameters for these variable parameters, so that a conditional database should be established to meet the requirements. In addition, we believe that we can consider strengthening the generative adversarial network,[185] transformer[186] or other ideas in the process of deep learning, which can be used as references to deal with the variable parameter.

4.6. Developing Novel Learning Models

Based on the material prediction according to the theoretical calculation, several new laws would be found by machine learning. And novel models, algorithms or integrated algorithms for material computing could be developed based on the paradigm named “theory-oriented, data-driven, and parameter learning optimization”, which could help to mine unknown new theories, develop new mechanisms, and in turn promote the progress of machine learning in materials science. For example, it is suggested to develop novel learning models based on the material genome initiative via machine learning, to verify them in combination with the experimental results, and finally to optimize the genome initiative for materials science.

In conclusion, we believe that machine learning is emerging as a great strategy for materials science, which would speed up the process of material exploration. The day, when machine learning is maturely applied to materials science and technology, would be the time for opening a new chapter of human civilization. Therefore, summarizing all possible strategies and providing in-depth perspectives will be necessary to make the machine learning viable in the future.

Acknowledgements

This work was supported by the National Key R&D Program of China (No. 2018YFC1901500).

Conflict of Interest

The authors declare no conflict of interest.

Keywords

data-driven, machine learning, materials discovery, materials science, performances prediction, perspective

Received: August 12, 2021
Revised: September 15, 2021
Published online: October 4, 2021

[1] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, R. Socher, *NPJ Dig. Med.* **2021**, *4*, 5.

[2] O. Yakovenko, I. Bondarenko, *Soc. Networks Texts* **2021**, *1357*, 115.

[3] S. X. Tang, R. Kriz, S. Cho, S. J. Park, J. Harowitz, R. E. Gur, M. T. Bhati, D. H. Wolf, J. Sedoc, M. Y. Liberman, *npj Schizophr.* **2021**, *7*, 25.

[4] J. W. Crandall, M. Oudah, Tennom, F. Ishowo-Oloko, S. Abdallah, J.-F. Bonnefon, M. Cebrian, A. Shariff, M. A. Goodrich, I. Rahwan, *Nat. Commun.* **2018**, *9*, 233.

[5] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. Van Den Driessche, T. Graepel, D. Hassabis, *Nature* **2017**, *550*, 354.

[6] A. Zhavoronkov, Y. A. Ivanenkov, A. Aliper, M. S. Veselov, V. A. Aladinskiy, A. V. Aladinskaya, V. A. Terentiev, D. A. Polykovskiy, M. D. Kuznetsov, A. Asadulaev, Y. Volkov, A. Zholus, R. R. Shayakhmetov, A. Zhebrak, L. I. Minaeva, B. A. Zagribelnyy, L. H. Lee, R. Soll, D. Madge, L. Xing, T. Guo, A. Aspuru-Guzik, *Nat. Biotechnol.* **2019**, *37*, 1038.

[7] B. Burger, P. M. Maffettone, V. V. Gusev, C. M. Aitchison, Y. Bai, X. Wang, X. Li, B. M. Alston, B. Li, R. Clowes, N. Rankin, B. Harris, R. S. Sprick, A. I. Cooper, *Nature* **2020**, *583*, 237.

[8] J. Launchbury, *Retrieved November* **2017**, *11*, 2019.

[9] Y. Lecun, Y. Bengio, G. Hinton, *Nature* **2015**, *521*, 436.

[10] R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed., The MIT Press, Almere **2018**.

[11] S. Lloyd, C. Weedbrook, *Phys. Rev. Lett.* **2018**, *121*, 040502.

[12] S. L. Sass, *The Substance of Civilization: Materials and Human History from the Stone Age to the Age of Silicon*, Arcade Publishing, New York **1998**.

[13] P. S. Anton, R. Silberglitt, J. Schneider, *The Global Technology Revolution: Bio/Nano/Materials Trends and Their Synergies with Information Technology by 2015*, Rand Corporation, Santa Monica **2001**.

[14] C. Freeman, F. Louçã, *As Time Goes By: From the Industrial Revolutions to the Information Revolution*, Oxford University Press, New York **2001**.

[15] K. Schwab, *The Fourth Industrial Revolution*, Currency Books, New York **2017**.

[16] A. Agrawal, A. Choudhary, *APL Mater.* **2016**, *4*, 053208.

[17] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, *npj Comput. Mater.* **2017**, *3*, 54.

[18] K T. Butler, D W. Davies, H. Cartwright, O. Isayev, A. Walsh, *Nature* **2018**, *559*, 547.

[19] J. Schmidt, M R. G. Marques, S. Botti, M A. L. Marques, *npj Comput. Mater.* **2019**, *5*, 83.

[20] R. Batra, L. Song, R. Ramprasad, *Nat. Rev. Mater.* **2020**, *6*, 655.

[21] B. Zhang, X.-Q. Zheng, T.-Y. Zhao, F.-X. Hu, J.-R. Sun, B.-G. Shen, *Chin. Phys. B* **2018**, *27*, 067503.

[22] L. Weston, C. Stampfl, *Phys. Rev. Mater.* **2018**, *2*, 085407.

[23] C. Kim, G. Pilania, R. Ramprasad, *J. Phys. Chem. C* **2016**, *120*, 14575.

[24] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, A. Tkatchenko, *Nat. Commun.* **2017**, *8*, 13890.

[25] S. A. Tawfik, O. Isayev, M. J. S. Spencer, D. A. Winkler, *Adv. Theory Simul.* **2020**, *3*, 1900208.

[26] J. Schmidt, J. Shi, P. Borlido, L. Chen, S. Botti, M. A. L. Marques, *Chem. Mater.* **2017**, *29*, 5090.

[27] Y.-J. Wu, L. Fang, Y. Xu, *npj Comput. Mater.* **2019**, *5*, 56.

[28] P. V. Balachandran, B. Kowalski, A. Sehirlioglu, T. Lookman, *Nat. Commun.* **2018**, *9*, 1668.

[29] G. C. Sosso, V. L. Deringer, S. R. Elliott, G. Csányi, *Mol. Simul.* **2018**, *44*, 866.

[30] J.-P. Correa-Baena, K. Hippalgaonkar, J. Van Duren, S. Jaffer, V. R. Chandrasekhar, V. Stevanovic, C. Wadia, S. Guha, T. Buonassisi, *Joule* **2018**, *2*, 1410.

[31] W. B. Park, J. Chung, J. Jung, K. Sohn, S. P. Singh, M. Pyo, N. Shin, K.-S. Sohn, *IUCrJ* **2017**, *4*, 486.

[32] A. Ziletti, D. Kumar, M. Scheffler, L. M. Ghiringhelli, *Nat. Commun.* **2018**, *9*, 2775.

[33] B. Meredig, E. Antono, C. Church, M. Hutchinson, J. Ling, S. Paradiso, B. Blaiszik, I. Foster, B. Gibbons, J. Hattrick-Simpers, A. Mehta, L. Ward, *Mol. Syst. Des. Eng.* **2018**, *3*, 819.

[34] G. Pilania, X.-Y. Liu, *J. Mater. Sci.* **2018**, *53*, 6652.

[35] Y. Liu, B. Guo, X. Zou, Y. Li, S. Shi, *Energy Storage Mater.* **2020**, *31*, 434.

[36] T. Lookman, P. V. Balachandran, D. Xue, R. Yuan, *npj Comput. Mater.* **2019**, *5*, 21.

[37] P V. Balachandran, *Comput. Mater. Sci.* **2019**, *164*, 82.

[38] P. Avery, X. Wang, C. Oses, E. Gossett, D. M. Proserpio, C. Toher, S. Curtarolo, E. Zurek, *npj Comput. Mater.* **2019**, *5*, 89.

[39] X. Zhai, M. Chen, W. Lu, *Comput. Mater. Sci.* **2018**, *151*, 41.

[40] E. V. Podryabinkin, E. V. Tikhonov, A. V. Shapeev, A. R. Oganov, *Phys. Rev. B* **2019**, *99*, 064114.

[41] S. Kim, J. Noh, G. Ho Gu, A. Aspuru-Guzik, Y. Jung, *ACS Cent. Sci.* **2020**, *6*, 1412.

[42] T. Xie, J. C. Grossman, *Phys. Rev. Lett.* **2018**, *120*, 145301.

[43] W. Tong, Q. Wei, H.-Y. Yan, M.-G. Zhang, X.-M. Zhu, *Front. Phys.* **2020**, *15*, 63501.

[44] W. Li, R. Jacobs, D. Morgan, *Comput. Mater. Sci.* **2018**, *150*, 454.

[45] H. Wei, S. Zhao, Q. Rong, H. Bao, *Int. J. Heat Mass Transfer* **2018**, *127*, 908.

[46] F. Musil, S. De, J. Yang, J. E. Campbell, G. M. Day, M. Ceriotti, *Chem. Sci.* **2018**, *9*, 1289.

[47] D. Padula, J. D. Simpson, A. Troisi, *Mater. Horiz.* **2019**, *6*, 343.

[48] D. Jha, K. Choudhary, F. Tavazza, W.-K. Liao, A. Choudhary, C. Campbell, A. Agrawal, *Nat. Commun.* **2019**, *10*, 5316.

[49] A. Mangal, E. A. Holm, *Int. J. Plast.* **2019**, *114*, 1.

[50] S. K. Kauwe, J. Graser, A. Vazquez, T. D. Sparks, *Integr. Mater. Manuf. Innov.* **2018**, *7*, 43.

[51] Y. Zhuo, A. Mansouri Tehrani, J. Brgoch, *J. Phys. Chem. Lett.* **2018**, *9*, 1668.

[52] S. Lu, Q. Zhou, Y. Ouyang, Y. Guo, Q. Li, J. Wang, *Nat. Commun.* **2018**, *9*, 3405.

[53] W. Ye, C. Chen, Z. Wang, I.-H. Chu, S. P. Ong, *Nat. Commun.* **2018**, *9*, 800.

[54] P. B. Jørgensen, K. W. Jacobsen, M. N. Schmidt, *Neural message passing with edge updates for predicting properties of molecules and materials.* arXiv preprint arXiv:1806.03146, **2018**.

[55] A. Jain, K A. Persson, G. Ceder, *APL Mater.* **2016**, *4*, 053102.

[56] J. Lee, A. Seko, K. Shitara, K. Nakayama, I. Tanaka, *Phys. Rev. B* **2016**, *93*, 115104.

[57] S K. Kauwe, T. Welker, T D. Sparks, *Integr. Mater. Manuf. Innov.* **2020**, *9*, 213.

[58] J. Schmidt, L. Chen, S. Botti, M A. L. Marques, *J. Chem. Phys.* **2018**, *148*, 241728.

[59] X. Zhu, J. Yan, M. Gu, T. Liu, Y. Dai, Y. Gu, Y. Li, *J. Phys. Chem. Lett.* **2019**, *10*, 7760.

[60] M. Misawa, S. Fukushima, A. Koura, K. Shimamura, F. Shimojo, S. Tiwari, K.-.I. Nomura, R K. Kalia, A. Nakano, P. Vashishta, *J. Phys. Chem. Lett.* **2020**, *11*, 4536.

[61] T. Toyao, K. Suzuki, S. Kikuchi, S. Takakusagi, K.-I. Shimizu, I. Takigawa, *J. Phys. Chem. C* **2018**, *122*, 8315.

[62] P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S A. Friedler, J. Schrier, A. J. Norquist, *Nature* **2016**, *533*, 73.

[63] C. Draxl, *Stepping Stones towards the Fourth Paradigm of Materials Science*. in *Seminar, Research Training Group QM3, Universität Bremen*.

[64] X. Zhang, *Polymer* **2001**, *42*, 8179.

[65] P. H. C. Camargo, K. G. Satyanarayana, F. Wypych, *Mater. Res.* **2009**, *12*, 1.

[66] G. Krauss, *Steels: Processing, Structure, and Performance*, ASM International, Almere, **2015**.

[67] S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, O. Levy, *Nat. Mater.* **2013**, *12*, 191.

[68] J. J. De Pablo, N. E. Jackson, M. A. Webb, L.-.Q. Chen, J. E. Moore, D. Morgan, R. Jacobs, T. Pollock, D. G. Schlom, E. S. Toberer, J. Analytis, I. Dabo, D. M. Delongchamp, G. A. Fiete, G. M. Grason, G. Hautier, Y. Mo, K. Rajan, E. J. Reed, E. Rodriguez, V. Stevanovic, J. Suntivich, K. Thornton, J.-C. Zhao, *npj Comput. Mater.* **2019**, *5*, 41.

[69] S. Rana, R. Fangueiro, *Advanced Composite Materials for Aerospace Engineering: Processing, Properties and Applications*, Woodhead Publishing, Cambridge **2016**.

[70] J. K. Pandey, K. R. Reddy, A. K. Mohanty, M. Misra, *Handbook of Polymernanocomposites: Processing, Performance and Application*, Springer, Myrtle Beach **2013**.

[71] W. D. Callister, D. G. Dhavalikar, *Materials Science and Engineering: An Introduction*, Wiley, Hoboken, NJ **2021**.

[72] R. Kulkarni, S. Dhavalikar, S. Bangar, *Traffic Light Detection and Recognition for Self Driving Cars Using Deep Learning*. in *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, **2018**, IEEE.

[73] R. Wu, et al., *Deep image: Scaling up image recognition*. arXiv preprint arXiv:1501.02876, **2015**, *7*.

[74] I. Serban, et al. *Building end-to-end dialogue systems using generative hierarchical neural network models*, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **2016**.

[75] L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, *npj Comput. Mater.* **2016**, *2*, 16028.

[76] J. Im, S. Lee, T.-W. Ko, H. W. Kim, Y. Hyon, H. Chang, *npj Comput. Mater.* **2019**, *5*, 37.

[77] T.-T. Le, *J. Compos. Mater.* **2021**, *55*, 787.

[78] N. Science, T. Council, *Materials genome initiative for global competitiveness*, **2011**, Executive Office of the President, National Science and Technology Council.

[79] J. Ling, W. Chen, Y. Sheng, W. Li, L. Zhang, Y. Du, *Mater. Sci. Eng., C* **2020**, *106*, 110265.

[80] Y. Wang, Y. Liu, S. Song, Z. Yang, X. Qi, K. Wang, Y. Liu, Q. Zhang, Y. Tian, *Nat. Commun.* **2018**, *9*, 2444.

[81] D. Sholl, J. A. Steckel, *Density Functional Theory: A Practical Introduction*, John Wiley & Sons, Hoboken, NJ **2011**.

[82] J. Behler, *J. Chem. Phys.* **2016**, *145*, 170901.

[83] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, C. Wolverton, *npj Comput. Mater.* **2015**, *1*, 15010.

[84] A. Zakutayev, N. Wunder, M. Schwarting, J D. Perkins, R. White, K. Munch, W. Tumas, C. Phillips, *Sci. Data* **2018**, *5*, 180053.

[85] S. Gražulis, A. Daškevič, A. Merkys, D. Chateigner, L. Lutterotti, M. Quirós, N R. Serebryanaya, P. Moeck, R T. Downs, A. Le Bail, *Nucleic Acids Res.* **2012**, *40*, D420.

[86] P. Gorai, D. Gao, B. Ortiz, S. Miller, S A. Barnett, T. Mason, Q. Lv, V. Stevanović, E S. Toberer, *Comput. Mater. Sci.* **2016**, *112*, 368.

[87] S. Haastrup, M. Strange, M. Pandey, T. Deilmann, P. S. Schmidt, N. F. Hinsche, M. N. Gjerding, D. Torelli, P. M. Larsen, A. C. Riis-Jensen, J. Gath, K. W. Jacobsen, J. Jørgen Mortensen, T. Olsen, K. S. Thygesen, *2D Mater.* **2018**, *5*, 042002.

[88] C. Draxl, M. Scheffler, *MRS Bull.* **2018**, *43*, 676.

[89] W. Setyawan, R M. Gaume, S. Lam, R S. Feigelson, S. Curtarolo, *ACS Comb. Sci.* **2011**, *13*, 382.

[90] M. Hellenbrandt, *Crystallogr. Rev.* **2004**, *10*, 17.

[91] S. Feng, H. Zhou, H. Dong, *Mater. Des.* **2019**, *162*, 300.

[92] C. O. S. Sorzano, J. Vargas, A. P. Montano, *A Survey of Dimensionality Reduction Techniques.* arXiv preprint arXiv:1403.2877, **2014**.

[93] M. Taradeh, M. Mafarja, A. A. Heidari, H. Faris, I. Aljarah, S. Mirjalili, H. Fujita, *Inf. Sci.* **2019**, *497*, 219.

[94] K. Yu, X. Guo, L. Liu, J. Li, H. Wang, Z. Ling, X. Wu, *ACM Comput. Surveys (CSUR)* **2020**, *53*, 111.

[95] Y. Liang, D. Niu, W.-C. Hong, *Energy* **2019**, *166*, 653.

[96] Z. Liu, Z. Lai, W. Ou, K. Zhang, R. Zheng, *Signal Process.* **2020**, *170*, 107456.

[97] M. Nixon, A. Aguado, *Feature Extraction and Image Processing for Computer Vision*, Academic press, Waltham **2019**.

[98] X. Huang, L. Wu, Y. Ye, *Int. J. Pattern Recognition Artif. Intell.* **2019**, *33*, 1950017.

[99] A. Karami, *Int. J. Knowledge Engineering Data Mining* **2019**, *6*, 289.

[100] Y. Alibert, *Astron. Astrophys.* **2019**, *624*, A45.

[101] M. Modaresnezhad, A. Vahdati, H. Nemati, A. Ardestani, F. Sadri, *Comput. Biol. Med.* **2019**, *106*, 84.

[102] N. Rabin, M. Kahlon, S. Malayev, A. Ratnovsky, *Exp. Syst. Appl.* **2020**, *149*, 113281.

[103] I. El Moudden, M. Ouzir, B. Benyacoub, S. El Bernoussi, *Contemp. Eng. Sci.* **2016**, *9*, 1031.

[104] A. O. Oliynyk, L. A. Adutwum, J. J. Harynuk, A. Mar, *Chem. Mater.* **2016**, *28*, 6672.

[105] A. O. Oliynyk, L. A. Adutwum, B. W. Rudyk, H. Pisavadia, S. Lotfi, V. Hlukhyy, J. J. Harynuk, A. Mar, J. Brgoch, *J. Am. Chem. Soc.* **2017**, *139*, 17870.

[106] J. D. Evans, F.-X. Coudert, *Chem. Mater.* **2017**, *29*, 7833.

[107] A. Mansouri Tehrani, A. O. Oliynyk, M. Parry, Z. Rizvi, S. Couper, F. Lin, L. Miyagi, T. D. Sparks, J. Brgoch, *J. Am. Chem. Soc.* **2018**, *140*, 9844.

[108] K. Choudhary, B. Decost, F. Tavazza, *Physs Rev. Mater.* **2018**, *2*, 083801.

[109] Y. Kiarnejad, S. Abdollahramezani, A. Adibi, *npj Computat. Mater.* **2020**, *6*, 12.

[110] N. Artrith, A. Urban, *Comput. Mater. Sci.* **2016**, *114*, 135.

[111] C. Chen, W. Ye, Y. Zuo, C. Zheng, S. P. Ong, *Chem. Mater.* **2019**, *31*, 3564.

[112] M. Attarian Shandiz, R. Gauvin, *Comput. Mater. Sci.* **2016**, *117*, 270.

[113] T. Lam Pham, H. Kino, K. Terakura, T. Miyake, K. Tsuda, I. Takigawa, H. Chi Dam, *Sci. Technol. Adv. Mater.* **2017**, *18*, 756.

[114] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, R. Ramprasad, *Sci. Rep.* **2013**, *3*, 2810.

[115] X. Li, Y. Zhang, H. Zhao, C. Burkhart, L. Catherine Brinson, W. Chen, *Sci. Rep.* **2018**, *8*, 13461.

[116] X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. Mclachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D J. Hand, D. Steinberg, *Knowl. Inf. Syst.* **2008**, *14*, 1.

[117] K. Arulkumaran, M. P. Deisenroth, M. Brundage, A. A. Bharath, *IEEE Signal Process. Mag.* **2017**, *34*, 26.

[118] C. Szepesvári, *Synth. Lectures Artif. Intell. Mach. Learn.* **2010**, *4*, 1.

[119] J. Oh, et al., *Discovering reinforcement learning algorithms*. arXiv preprint arXiv:2007.08794, **2020**.
[120] T. G. Dietterich, *Ensemble Methods in Machine Learning*, in *International Workshop on Multiple Classifier Systems*. **2000**, Springer.
[121] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*, CRC Press, Boca Raton, FL **2012**.
[122] J. Schmidhuber, *Neural Networks* **2015**, *61*, 85.
[123] S. Kearnes, L. Li, Richard N. Zare, P. Riley, *Sci. Rep.* **2019**, *9*, 10752.
[124] Z. Li, D. Wu, T. Yu, *J. Manuf. Sci. Eng.* **2019**, *141*, 031003.
[125] J. R. Quinlan, *C4. 5: Programs for Machine Learning*, Elsevier, Amsterdam **2014**.
[126] I. Rizkya, et al. *Analysis of Defective Causes in Coffee Product Using Decision Tree Approach*. in *IOP Conference Series: Materials Science and Engineering*, **2020**, IOP Publishing.
[127] M. Martinez-Arroyo, L. E. Sucar, *Learning an optimal naive bayes classifier*. in *18th International Conference on Pattern Recognition (ICPR'06)*, **2006**, IEEE.
[128] V. Solov'ev, A. Tsivadze, G. Marcou, A. Varnek, *Mol. Inf.* **2019**, *38*, 1900002.
[129] C. Cortes, V. Vapnik, *Mach. Learn.* **1995**, *20*, 273.
[130] S. H. Lee, J. Mazumder, J. Park, S. Kim, *J. Manuf. Processes* **2020**, *55*, 307.
[131] D. Jahed Armaghani, P. G. Asteris, B. Askarian, M. Hasanipanah, R. Tarinejad, V. V. Huynh, *Sustainability* **2020**, *12*, 2229.
[132] T. Cover, P. Hart, *IEEE Trans. Inf. Theory* **1967**, *13*, 21.
[133] T. Adithiyaa, D. Chandramohan, T. Sathish, *Mater. Today: Proc.* **2020**, *21*, 1000.
[134] T. Adithiyaa, D. Chandramohan, T. Sathish, *Mater. Today: Proc.* **2020**, *21*, 108.
[135] R. E. Schapire, *Explaining Adaboost*, in *Empirical Inference*, **2013**, Springer, 37–52.
[136] Ji Li, C. Zhang, X. Zhang, H. He, W. Liu, C. Chen, *IEEE Access* **2020**, *8*, 12413.
[137] W. Y. Loh, *Wiley Interdiscip. Rev.: Data Mining Knowl. Discovery* **2011**, *1*, 14.
[138] S. Kadali, S. M. Naushad, A. Radha Rama Devi, V. L. Bodiga, *Mol. Cell. Biochem.* **2019**, *458*, 27.
[139] S. Na, L. Xumin, G. Yong. *Research on k-means clustering algorithm: An improved k-means clustering algorithm*. in *2010 Third International Symposium on intelligent information technology and security informatics*, **2010**, Ieee.
[140] J. Lin, T. N. Pappas, *Structural Texture Similarity for Material Recognition*. in *2019 IEEE International Conference on Image Processing (ICIP)*, **2019**, IEEE.
[141] G. Weber, M. Pinz, S. Ghosh, *JOM* **2020**, *72*, 4404.
[142] I. C. Gormley, T. B. Murphy, *Encyclop. Stat. Quality Reliab.* **2008**, *2*.
[143] T. Masuda, T. Nishio, J. Kataoka, M. Arimoto, A. Sano, K. Karasawa, *Phys. Med. Biol.* **2019**, *64*, 175011.
[144] R. Agrawal, R. Srikant, *Fast Algorithms for Mining Association Rules*, in *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, **1994**, Citeseer.
[145] S. Ren, X. Zhao, B. Huang, Z. Wang, X. Song, *J. Ambient Intell. Humanized Comput.* **2019**, *10*, 1093.
[146] A. Langville, C. Meyer, *Internet Math.* **2004**, *1*, 335.
[147] G. Bastos, et al., *Development of an inverse identification method for identifying constitutive parameters by metaheuristic optimization algorithm: Application to hyperelastic materials*, in *Residual Stress, Thermomechanics & Infrared Imaging and Inverse Problems*, **2020**, Springer, p. 141.
[148] B. Liu, S. Jiang, Q. Zou, *Brief. Bioinf.* **2020**, *21*, 298.
[149] G. E. Hinton, S. Osindero, Y.-W. Teh, *Neural Comput.* **2006**, *18*, 1527.
[150] G. E. Hinton, *Deep Belief Networks*, **2009**, *4*, 5947.
[151] J. Bouvrie, *Notes on Convolutional Neural Networks.* **2006**.
[152] R. Yamashita, M. Nishio, R. K. G. Do, K. Togashi, *Insights Imaging* **2018**, *9*, 611.
[153] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, T. Chen, *Pattern Recognition* **2018**, *77*, 354.
[154] S. Hochreiter, J. Schmidhuber, *Neural Comput.* **1997**, *9*, 1735.
[155] M. Schuster, K. K. Paliwal, *IEEE Trans. Signal Process.* **1997**, *45*, 2673.
[156] Y. Yu, X. Si, C. Hu, J. Zhang, *Neural Comput.* **2019**, *31*, 1235.
[157] C. Rao, Y. Liu, *Comput. Mater. Sci.* **2020**, *184*, 109850.
[158] S. N. S. Mortazavi, A. Ince, *Comput. Mater. Sci.* **2020**, *185*, 109962.
[159] C. Rudin, *Nat. Mach. Intell.* **2019**, *1*, 206.
[160] M. T. Ribeiro, S. Singh, C. Guestrin. *"Why should i trust you?" Explaining the predictions of any classifier*. in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, **2016**.
[161] W. J. Murdoch, et al., *Interpretable machine learning: definitions, methods, and applications.* arXiv preprint arXiv:1901.04592, **2019**.
[162] Y. Iwasaki, R. Sawada, V. Stanev, M. Ishida, A. Kirihara, Y. Omori, H. Someya, I. Takeuchi, E. Saitoh, S. Yorozu, *npj Comput. Mater.* **2019**, *5*, 103.
[163] B. Kailkhura, B. Gallagher, S. Kim, A. Hiszpanski, T. Y.-J. Han, *npj Comput. Mater.* **2019**, *5*, 108.
[164] S. Hooker, et al., *A benchmark for interpretability methods in deep neural networks.* arXiv preprint arXiv:1806.10758, **2018**.
[165] B. Mittelstadt, C. Russell, S. Wachter. *Explaining explanations in AI*. in *Proceedings of the conference on fairness, accountability, and transparency*. **2019**.
[166] L. H. Gilpin, et al. *Explaining explanations: An overview of interpretability of machine learning*. in *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, **2018**, IEEE.
[167] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, K.-R. Müller, *J. Chem. Phys.* **2018**, *148*, 241722.
[168] L. Himanen, M. O. J. Jäger, E V. Morooka, F. Federici Canova, Y S. Ranawat, D Z. Gao, P. Rinke, A S. Foster, *Comput. Phys. Commun.* **2020**, *247*, 106949.
[169] L. Ward, A. Dunn, A. Faghaninia, N. E. R. Zimmermann, S. Bajaj, Qi Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K A. Persson, G. J Snyder, I. Foster, A. Jain, *Comput. Mater. Sci.* **2018**, *152*, 60.
[170] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, G. Ceder, *Comput. Mater. Sci.* **2013**, *68*, 314.
[171] T. Ueno, T. D. Rhone, Z. Hou, T. Mizoguchi, K. Tsuda, *Mater. Discovery* **2016**, *4*, 18.
[172] E. Gossett, C. Toher, C. Oses, O. Isayev, F. Legrain, F. Rose, E. Zurek, J. Carrete, N. Mingo, A. Tropsha, S. Curtarolo, *Comput. Mater. Sci.* **2018**, *152*, 134.
[173] M. Abadi, et al. *Tensorflow: A system for large-scale machine learning*. in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*. **2016**.
[174] A. Paszke, et al., *Pytorch: An imperative style, high-performance deep learning library.* arXiv preprint arXiv:1912.01703, **2019**.
[175] A. Gulli, S. Pal, *Deep learning with Keras*, Packt Publishing Ltd, Birmingham **2017**.
[176] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, *J. Mach. Learn. Res.* **2011**, *12*, 2825.
[177] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*, **2019**, O'Reilly Media.
[178] V. K. Verma, et al. *Towards zero-shot learning with fewer seen class examples*. in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. **2021**.
[179] V. Garcia, J. Bruna, *Few-shot learning with graph neural networks.* arXiv preprint arXiv:1711.04043, **2017**.

[180] Z. Li, et al., *Meta-sgd: Learning to learn quickly for few-shot learning.* arXiv preprint arXiv:1707.09835, **2017**.
[181] B. M. Lake, R. Salakhutdinov, J. B. Tenenbaum, *Science* **2015**, *350*, 1332.
[182] E A. Olivetti, J M. Cullen, *Science* **2018**, *360*, 1396.
[183] T. Bai, D. Li, K. Sun, Y. Chen, W. Li, *Remote Sens.* **2016**, *8*, 715.
[184] F. Scarselli, M. Gori, Ah Chung Tsoi, M. Hagenbuchner, G. Monfardini, *IEEE Trans. Neural Networks* **2008**, *20*, 61.
[185] X. Yi, E. Walia, P. Babyn, *Med. Image Anal.* **2019**, *58*, 101552.
[186] A. Vaswani, *Advances in Neural Information Processing Systems*, **2017**.

Chaochao Gao received his Bachelor's degree from School of Materials Science and Technology, Taiyuan University of Science and Technology in 2018. He is currently pursuing his Master's degree under the supervision of Assoc. Prof. Xin Min and Prof. Zhaohui Huang. His research focuses on the machine learning for material design optimization.

Xin Min received his BS degree (2011) and Ph.D. degree (2016) from School of Materials Science and Technology, China University of Geosciences (Beijing). Then, he joined the faculty of School of Materials Science and Technology, China University of Geosciences (Beijing). He is a visiting scholar in Department of Materials and Metallurgy, University of Cambridge. His main research interests include designing high-performance functional materials for energy storage and conversion.

Zhaohui Huang received his Ph.D. degree from School of Materials Science and Engineering, University of Science and Technology Beijing in 2000. He joined the faculty of China University of Geosciences (Beijing) (CUGB) in 2004. He has been the associate Dean of the School of Materials Science and Engineering in CUGB. He has hosted the State Key Program of National Natural Science of China, the "12th Five-Year" Plan key projects supported by National Science and technology, and the National Key R&D Program of China. Currently, his main research interests include advanced ceramics, refractories, and material utilization of solid waste.