Nanoscale Advances

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: J. Cai, X. Chu, K. Xu, H. Li and J. Wei, *Nanoscale Adv.*, 2020, DOI: 10.1039/D0NA00388C.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/nanoscale-advances

View Article Online
DOI: 10.1039/D0NA00388C

REVIEW

Machine Learning Driven New Material Discovery

Jiazhen Cai[a], Xuan Chu[a], Kun Xu[a], Hongbo Li[b], Jing Wei[a,b]*

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

Abstract

New materials can bring tremendous technology and application progress. But now the commonly used trier-and-error method can't meet the need today. Now, a newly proposed idea using machine learning to explore new materials is becoming popular. In this paper, we review this research paradigm of applying machine learning in material discovery, including data preprocessing, feature engineering, machine learning algorithms and cross-validation procedures. Furthermore, we propose to assist traditional DFT calculation with this idea for material discovery. Many experiments and literatures have shown the great effect and prospect of this idea. It's now showing its potential and advantages in property prediction, material discovery, inverse design, corrosion detection and many other aspects in life.

Key words: machine learning, materials discovery.

1. Introduction

Machine learning(ML)[1, 2] is a new subfield of artificial intelligence that focuses on optimizing computer programs to improve algorithms by data and researching experience. It has also become an efficient and important tool of analysing existing materials in new material discovery field.[3, 4] The traditional trial-and-error method relies on the personal experience. Therefore, it often takes decades from experiment to marketing.[5, 6] Considering the consumption of experimental development, the traditional method can hardly adapt to the large-scale demand of high-performance new-type materials.

Under this situation, the United States of America launched the "Material Genome Initiative" project (MGI) in 2011.[7] The MGI proposes that the researchers should focus on the "material design" instead of "trial-and-error", which requires researchers to deepen the comprehension of materials; collect enormous material data to build the databases and computing platforms; and above all, use high-throughput screening upon materials to eventually achieve the propose of reducing research costs and speeding up development. In 2016, China launched the "Material Genetic Engineering and Support Platform" program as a national strategy. Being different from MGI, Chinese government concerns about building a high-throughput computing platform[8] that can serve the majority of researchers. In this condition, the machine learning driven new material discovery method is thriving.

When solving material problems by ML, we need datasets to help us detecting target features, properties or unknown materials. These datasets and all the messages inside are called as "input", and the targets are called as "output". Now with these two definitions, this ML-aided method is defined as "using inputs and appropriate ML algorithms to build a numerical predicting model, and detecting unknown outputs by the predicting ability of this model" (Figure 1). Because outputs are fitted by inputs, it's reasonable that outputs have similar chemical structures as inputs, and can be evaluated in the way evaluating inputs.[9] With this model, we can enhance the comprehension of material properties and predict unknown demanding materials. At present, this method is still confronted with many challenges: the messy datasets must be preprocessed; the accuracy of model is limited by algorithms; and the pressure of high-intensity computation on the computing resources etc.[10]

Machine learning has been widely used in many aspects of material science (Figure 2). And in this review, we focus on the model construction, computational algorithms, model verification procedures, the role ML plays in material science field and the prospect of machine learning. The Section 2 describes the data preprocessing and feature engineering steps, which can systemically reconstruct the datasets and help understanding material properties and physicochemical relationships. In the Section 3, some high-performance algorithms in material-discovery field are introduced in this part, and some practical applying examples in this field are followed. The Section 4 describes several cross-validation procedures for material-discovery ML models. the Section 5 explains how ML assisting traditional density functional theory way in the field of materials science. In Section 6, some other artificial intelligence methods that ML involves in are discussed. In Section 7, we will summarize the current development condition of ML methods, and briefly explain the prospects of ML in new material discovery.

[a] State Key Laboratory of Information Photonics and Optical Communications, Beijing University of Posts and Telecommunications, Beijing, China.
[b] Experimental Center of Advanced Materials, School of Materials Science & Engineering, Beijing Institute of Technology, Beijing 100081, China.
*weijing@bit.edu.cn.

View Article Online
DOI: 10.1039/D0NA00388C

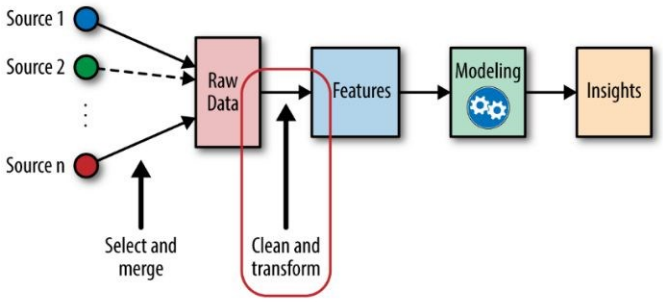


Figure 1. The machine learning workflow, and the place of feature engineering is in the red circle.[11]

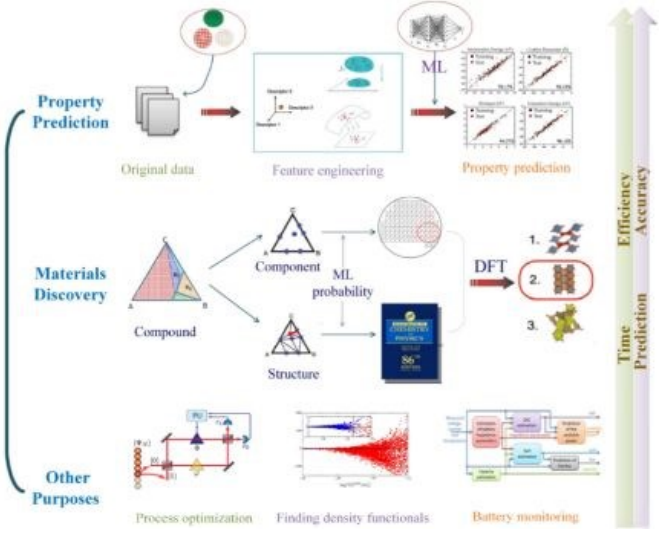


Figure 2. An overview of the application of machine learning in materials science.[12]

2. Data preprocessing and feature engineering

Data is very important in ML procedures. Generally, the final ML results can be directly affected by the volume and reliability of data, and that's where data preprocessing and feature engineering way in. These two steps can reconstruct datasets so it's more convenient for computers to understanding material physicochemical relationships, detecting material properties and establishing material predicting models.[13-15] For example, Paul Raccuglia once proposed that the predicting model can be trained by failed data from failed experiment. They integrate experimental data by using failed or less successful hydrothermal synthesis reaction information to train a machine learning model for predicting the crystallization of template vanadium selenite crystals. It turns out that this model is obviously superior compared with traditional manual analysis method. The prediction accuracy for the formation condition of new organic template inorganic products can reach 89%.[5]

In the following parts, we will introduce how these two steps work in material discovery, and we will also give some successful examples in material field.

1) Data preprocessing

Data preprocessing is mainly consisted of two steps: data collecting and data cleaning.[16]

i. Data collecting

Researchers always hope to collect representative data. Therefore, it is necessary for researchers to select the appropriate data for specific problems. Nowadays, a lot of open-source databases like Harvard Clean Energy Project, the Open Quantum materials database and the Materials Project[17] have been established. These databases are reliable and accessible, which can be used as a bedrock in researching work. Some of the most authoritative and reliable databases are listed below for reference (Table 1).

Table 1. An overview of some of the most authoritative databases in material science.

Title	Website Address	Brief Introduction
AFLOWLIB	http://www.aflowlib.org	A globally database of 3,249,264 material compounds and over 588,116,784 calculated properties.
ASM Alloy Database	https://www.asminternational.org/materials-resources/online-databases	An authoritative database focusing on alloys, mechanical, alloy phase and failure experiment data.
Cambridge Crystallographic Data Centre	http://www.ccdc.cam.ac.uk	It focuses on structural chemistry and contains over 1,000,000 structures.
ChemSpider	http://www.chemspider.com	A free chemical structural database providing fast searching access to over 67,000,000 structures.
Harvard Clean Energy Project	http://cepdb.molecularspace.org/	A massive database of organic solar cell materials.
High Performance Alloys Database	https://cindasdata.com/products/hpad	The high performance alloy database addresses the needs of chemical processing, power generation and transportation industries.
Materials Project	https://materialsproject.org/	It offers more than 530,000 nanoporous materials, 124,000 inorganic compounds and power analysis tools for researchers.
NanoHUB	https://nanohub.org/resources/databases	An open source database focusing on the nanomaterials.
Open Quantum Materials Database	http;//oqmd.org/	It contains huge amount of data of thermodynamic and structural properties of 637,644 materials

View Article Online
DOI: 10.1039/D0NA00388C

Springer Materials	http://materials.springer.com	A comprehensive database covering multiple material classes, properties and applications.

For example, Edward O. Pyzer-Knapp's team uses the data from the Harvard Clean Energy Database for ML model training to solve the over-provisioning problem, which computers mistakenly consider noise as useful features. According to the results, the trained ML model can return predictions on the verification set within 1.5 seconds, and over-provisioning problem is successfully solved.[18] Kamal Choudhary's team establishes a criterion to identify 2D materials based on comparison about lattice constants obtained from experiments and Materials Project database. And to test this criterion, they calculated the exfoliation of many layered materials. The result shows that in 88.9% of the cases the criterion is satisfied.[19] T. Björkman's team screens International Crystallographic Structural Database (ICSD) and discover 92 possible 2D compounds based on symmetry, packing ratio, structural gaps and covalent radii.[20] Other cases like Michael Ashton and colleagues use the topology-scaling algorithm to detect the layered materials from ICSD and successfully find 680 stable monolayers;[21] Nicolas Mounet and coworkers use data mining upon ICSD and Crystallographic Open Database to search for the layered compounds;[22] and Sten Haastrup's team establishes one of the largest 2D materials database.[23] All the researches mentioned before strongly support the availability and superiority of high-quality data in the practical application. From this point of view, it's a necessary step for ML method in material discovery.

ii. Data cleaning

After the data collecting, there are still many problems in the collected data, such as data redundancy or abnormal values. In order to get an efficient ML predicting model, and also reduce the amount of calculation, data cleaning is necessary. In this paper, we define data cleaning as a data-operation consisted by four steps: data sampling, abnormal value processing, data discretization, and data normalization.[24, 25]

First of all, data sampling ensures that researchers can get high performance prediction models by less data without compromising predicting accuracy.[26] And in order to ensure the ability and accuracy of the predicting model, researchers have to eliminate the abnormal values to keep the accuracy of predicting models.[27] What's more, data discretization can significantly reduce the number of possible values of continuous feature. And data normalization can adjust the magnitudes of data to a suitable and the same level, which is crucial for many machine learning algorithms. As an example, in the research of photovoltaic organic-inorganic hybrid perovskites by Shuai Hua Lu's team, all the input data is obtained from reliable databases made up of high throughput first-principles calculations. And for the data consistency and accuracy of ML predictions, they carefully construct their own training sets and validation sets with appropriately processed data. Only orthorhombic-like crystal structures with bandgap calculated using the Perdew-Burke-Ernzerh (PBE) functional are selected[28].

And after four steps above, the dataset would be divided into training sets and testing sets. As we can see that after data preprocessing, the noise, data redundancy and abnormal values are all largely reduced. However, the dataset is still a mess. We need to reconstruct it for computers to better understand the data inside. And that's where feature engineering comes.

2) Feature engineering

Feature engineering is the process of extracting the most appropriate features in data and tasks. And that feature, is the so-called descriptor. Figure 1 has shown the position of feature engineering in the machine learning workflow. Figure 4 presents a typical workflow from data to fingerprinting descriptors. Its purpose is to obtain the features from training data, so that the ML algorithms can approach its best performance. In the latest work of Ming Wang et al, researchers introduce the automated feature engineering as a new trend in nanomaterial discovery. Automated feature engineering uses deep learning algorithms to automatically develop a set of features that are relevant to predict desired output. It would significantly minimize the amount of domain knowledge used in training model, and accelerate its application among non-expert users.

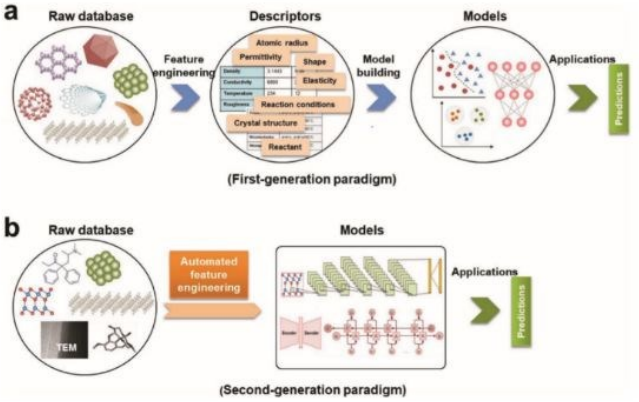


Figure 3. Evolution of the workflow of machine learning in nanomaterials discovery and design. (a) First-generation approach. In this paradigm, there are two main steps: feature engineering from raw database to descriptors; model building from descriptors to target model. (b) Second-generation approach. The key characteristic that distinguishes itself from first-generation approach is eliminating human-expert feature engineering, which can directly learn from raw nanomaterials.[29]

We can see that this new technic has very good application prospect[29], and it's also a very significant applying paradigm of deep learning. As an example, Figure 3 shows the evolution of the workflow of machine learning in nanomaterials discovery and design.

Turning back to descriptors. In particular, a descriptor refers to a set of meaningful parameters that describe a mechanism or a feature. For example, in the chemical periodic table, all the elements are sorted by rows (period) and column (family). The rows and columns can be considered as a set of descriptors. The appropriate descriptors can integrate essential information, and the quality of predicting model is also directly related to the quality of descriptors. High-performance descriptors can effectively organize independent variables, express various mechanisms and find hidden relationships with less volume of data. Now there are basically two mainstream ideas of designing descriptors. The first one is manually creating a set of descriptors depending on relevant physical chemical properties of experiment candidates. The second one is using related

View Article Online
DOI: 10.1039/D0NA00388C

mathematical and physical theories to project candidates' features into numerical vectors as descriptors.[34] Luca M. Ghiringhelli presumed four basic standards of descriptors after research.[30]

1. The dimensions of the descriptor should be as low as possible.
2. The descriptor uniquely characterizes the material as well as property-relevant elementary process.
3. Materials that are very different (similar) should be characterized selected by very different (similar) descriptors values.
4. The determination of the descriptor must not involve calculations as intensive as those needed for the evaluation of the property to be predicted.

Despite the four basic standards above, it still bothers when selecting the right descriptors. Daniel C. Elton believes that when facing a small dataset, the characterization of data is more important than the construction of model, and a set of highly efficient descriptors can ensure the accuracy of results. When dealing with large databases, a large dataset is enough for the ML algorithm to extract complex or potential features from the usual traits. But in this case, researchers suggest to select descriptors with superior computational efficiency and experimental performance. And to ensure the results' accuracy, transparent and abstract descriptors are also preferred.[31]

And when it comes to practical application, we need to choose suitable descriptors depending on different situations. Anton O. Oliynyk and coworkers use machine learning methods to detect potential Heusler compounds and properties.[32] They focus on some specific dimensions where material patterns are most likely to be found. In these dimensions, the utility of descriptors can be maximized.

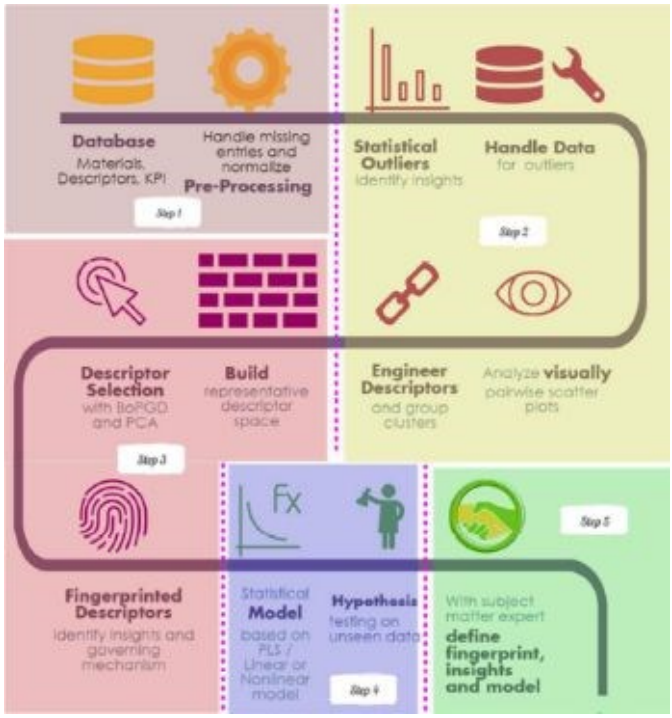


Figure 4. A recipe for going from data to fingerprinting descriptors to insights to models and discovery.[35]

They finalize 22 descriptors through experiments to help computers discover the hidden relationships. And after the verification, the predicting model they got by this set of descriptors can conduct prediction and calculation on over 400,000 groups of substances within 45 minutes. And the results after ten times of cross-validations, which proves the correctness of this prediction. Fleur Legrain's team tries predict vibrational energies and entropies of compounds by ML method. In this case, they choose chemical composition-based descriptors to guarantee the quality and accuracy of results in small datasets. In the experiment, they get a conclusion that the descriptors based on chemical composition and elemental properties of atoms of materials have excellent perform in small datasets. The predictive power of this idea is validated by comparing experimental results with measured values from National Institute of Standards and Technology.[33] What's more, Seiji Kajita and colleagues develop a universal 3D voxel descriptor that can represent three dimensionality of field quantities in solid-state ML. And in the experimental validation, they associate the convolutional neural network with solid-state ML by this novel descriptor, and test the experimental performance of the descriptor using 680 oxides data. The result shows that this descriptor outperforms other existing descriptors in prediction of Hartree energies and are relevant to the long-wavelength distribution of valence electrons.[34]

Besides the basic descriptors, there are some cases when the descriptor itself needs to be explored deeply. Ying Zhang's team finds that the increase in model predicting accuracy is at the expense of Degree of Freedom. To solve this problem, they introduce a so-called "crude estimate of property" descriptor in the feature space to improve accuracy without increasing the degree of freedom. Compared with the conventional method, the ML with new descriptors has better performance and smaller standard deviation in the tests.[36] Another classical case is that Ankit Jain and Thomas Bligaard develop a universal atomic-position independent descriptor for periodic inorganic solids. Generally speaking, the high-throughput for periodic inorganic solids requires essential atomic positions to encode structural and compositional details into appropriate material descriptors. However, when exploring the novel materials, the atomic-position information is usually not available. So they develop this descriptor system. And in the application for formation energies of bulk crystalline solids, the descriptors achieve prediction mean absolute error of only 0.07 eV/at on a test dataset of more than 85000 materials.[37]

As we can see, feature engineering and descriptors can largely reduce the workload in experiments. But it's still a serious question about how to choose the suitable descriptors. From this aspect, further study is still needed.

3. Basic machine learning algorithms

After building a database, it is necessary to select the appropriate machine learning algorithms. Some mathematical theories such as Markov chain, least squares method and Gaussian process[38] constructed the foundation of ML algorithms. Nowadays, ML algorithms can be divided into four types: supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. The reinforcement learning usually focuses on the interactions between algorithms and environment, but not finding specific patterns from datasets, which is not appropriate for material discovery. So, we would not discuss it in this section. As the "no free lunch theorems"[39] goes, there is no absolutely superior ML

View Article Online
DOI: 10.1039/D0NA00388C

algorithm. Each algorithm has its own advantages and disadvantages. Here we list several commonly used machine learning algorithms for reference.

1) Regression Analysis

The regression analysis can be divided into two categories: the supervised regression analysis and the unsupervised regression analysis. The regression analysis algorithms can quantify the magnitude of the dependent variable affected by the independent variable through analyzing a large volume of data. Then, it would find matching linear or nonlinear regression equations, and predict the dependent variable through regression equations. Depending on this feature, researchers can use the regression equations to analysis properties or discovery new materials.

For example, the team of Atsuto Seko individually uses the ordinary least squares regression (OLSR)[40], partial least-square regression (PLSR)[41-43], support vector regression (SVR, which is also a kind of Support Vector Machine algorithms)[44, 45] and Gaussian process regression (GPR)[46, 47] to predict the melting temperature of monobasic compounds and binary compounds. They select four simple physical properties and ten element properties as features, then construct two datasets to analyze the performance of four algorithms. The result shows that the support vector regression has the lowest root mean square error and the best performance.[48] Stefano Curtarolo and coworkers try to use simple ML algorithms such as PLSR to predict formation energies and optimize high-throughput calculation.[49] And in another experiment, to monitor the spatial distribution of CaO in cement materials, Bulent Tutmez and Ahmet Dag use the regression kriging model and geographically weighted regression to evaluate spatially varying relationships in a cement quarry. It turns out that regression kriging model outperforms the geographically weighted regression.[50]

2) Naïve Bayes Classifiers

Bayesian classification is a general term for a class of algorithms which are all established based on the Bayes theory, and the naïve Bayes classifier[51, 52] is the simplest one. Naïve Bayes Classifiers assume that all features are independent of each other. This assumption greatly simplifies the sample space and the amount of solving calculation, which makes it basically the simplest one in all Bayes classifications. It can choose the assumption which has the highest probability of correctly representing the data, and it's widely used because of its efficient algorithm, fast classification, and the ability to be applied to the field of big data.[53] There is a case that O. Addin and colleagues try to detect the damage of laminated composite materials by using naïve Bayes classifier. The naïve Bayes classifier shows great classification accuracy, which can reach 94.65% in the experiment.[54]And in an experiment that using specially designed robotic finger to recognizing the object surface material, Hongbin Liu uses naive Bayes classification, k-nearest neighbor classification and radial basis function network to identify the surfaces. The results indicate that the naïve Bayes classification outperforms the other two classification methods with the average successful rate of 84.2%.[55]

3) Support Vector Machine (SVM)

Support vector machine[56, 57] is a kind of supervised learning method. For a group of points in the N dimension, the support vector machine would find a hyperplane of the N-1 dimension and divide this group into two categories. SVM is built on the statistical learning theory, and the essence of SVM is to minimize the actual errors of ML.

After long-term of development, the support vector machine has already been able to greatly simplify the problem, reduce the dimensions of data, and eliminate the noise, which shows great generalization ability in the unknown samples. Xintao Qiu and the colleagues combine SVM and recursive feature elimination together to modeling the atmospheric corrosion upon materials. This brand new method can extract the most influential factors and build a reliable analyzing model. And when selecting corrosion factors in small sample sizes, it greatly outperforms other algorithms in regression and prediction performance.[58] To detecting the molecular functions and biological activities of phage virion proteins, Balachandran Manavalan trains a SVM predicting model by 136 features called PVP-SVM. The performance of PVP-SVM is consistent in both training and testing datasets, and the cross validation shows that the accuracy of PVP-SVM is 0.870, which is higher than any other control SVM predictors[59] (Figure5).

Nowadays, SVM has also made good progresses in the medical field. Manfred K. Warmuth's team uses SVM to classify and screen the compounds related to the target drug, and successfully find the most satisfying drug in the screening. In this procedure, the classification characteristics of SVM can effectively judge the correlation between the compound and the target, and shows very good accuracy in the final test[60]

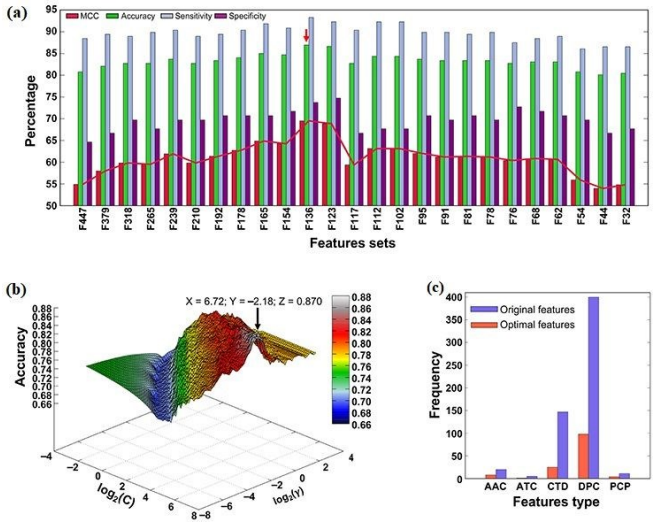


Figure 5. (a) Performance of SVM-based classifiers in distinguishing between PVPs and non-PVPs. The red arrow denotes the final selected model. (b) Classification accuracy of the final selected model. (c) Distribution of each feature type in the optimal feature set (136 features) and original feature set (583 features).[59]

4) Decision Tree and Random Forest

The decision tree[61-63] is a kind of supervised learning. When generating a decision tree, the source dataset (which constitutes the root node) is split into several subsets (which constitute the successor children) according to a series of splitting rules that based on classification features. By repeating this procedure, a decision tree grows (Figure 6a). The decision tree is also used in classification problems, but it still has some shortcomings such as overfitting and generalizing weakness. Therefore, researchers created random

View Article Online
DOI: 10.1039/D0NA00388C

forest algorithm, lifting tree algorithm, and many other algorithms based on the decision tree. The random forest algorithm[64, 65] is a classifier composed of multiple decision trees. The random forest solves the problems of a single decision tree by the voting mechanism of multiple decision trees, which greatly improves its generalizing ability.

In the experiment by Anton O. Oliynyk and coworkers, they attempt to synthesize AB_2C Heusler compounds by using the random forest algorithm.[66, 67] By averaging the predictions of these decision trees, the random forest aggregates the different trends of each tree, and results in a complex and stable model. They select two gallium-containing compounds (MRu_2Ga and RuM_2Ga (M=Ti-Ni)) to test its predictive ability. After the prediction, they are considered to have more than 60% probability of forming a Heusler compound, and Co_2RuGa has a higher probability around 92%.[32] Derek T. Ahneman and colleagues try to predict the Buchwald-Hartwig amination reaction against Pd catalysis, the root mean square error (RMSE) of the test set of the random forest mode was 7.8%, which is much better than Kernel Ridge Regression(KRR), support vector machine, Bayes generalized linear model and single layer neural network.[68] In another case, Junfei Zhang and his colleagues use a beetle antennae search algorithm based random forest (BAS-RF) method to detect the uniaxial compressive strength (UCS) of lightweight self-compacting concrete. The result shows that this algorithm has a high predictive accuracy indicated by the high correlation coefficient of 0.97. In Figure 6b, it's clear to see that BAS-RF has lower root-mean-square error value and the higher correlation coefficien than multiple linear regression (MLR) and logistic regression (LR), indicating the better performance. [69]

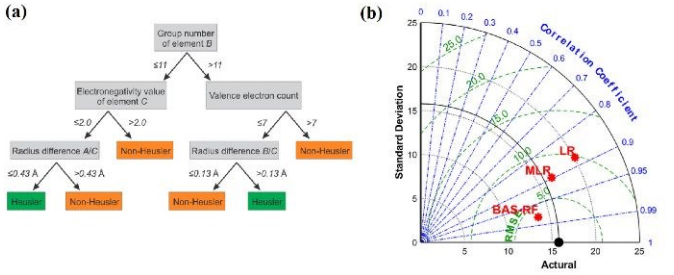


Figure 6. (a) Schematic representation of a single decision tree.[32] (b) The taylor diagram of different models for UCS prediction.[69]

5) Artificial Neural Network (ANN)

The artificial neural network[70, 71], which constructed by neurons, is a kind of ML algorithms that simulating biological cranial nerves. The neural network can date back to 1949, when a Canadian psychologist Donald Hebb developed a theory of learning known as Hebbian Learning.[72] But it was not until nearly two decades that ANN has been largely developed. In the early period of ANN, the single-layer neural network was first proposed. But due to some of its specific limitations, researchers generally turned to multiple-layer neural network in latter studies, which is consisted of input layer, hidden layer and output layer. Figure 7a is a typical example of multiple layer neural network. Neural network algorithms can analyze problems efficiently by nonlinear complex interactions between neurons. It is confirmed that ANN is very useful in the following three situations[35]:

1.The number of samples or experiments is very large.

2.There is not much a priori information about the dataset.

3.In the above case, the target only predicts within the model domain.

Edward O. Pyzer-Knapp used a special form of neural network called "multilayer perceptron" to discover new compounds. Edward and his colleagues use the multilayer perceptron to continuously filter datasets, eliminate data with small correlations after each iteration, and then put the remaining data into the next round of screening, thereby lock the target compound while minimizing computational effort.[18] Tian Zhang and colleagues propose a novel approach by using ANN to realize spectrum prediction, performance optimization and inverse design for plasmonic waveguide-coupled with cavities structure.[73] Tarak Patra and colleagues build a neural network-biased genetic algorithm, which can discover materials with extremal properties in the absence of pre-existing data. And it's shown to outperform both a nonbiased genetic algorithm and a neural-network-evaluated genetic algorithm based on a pre-existing dataset in a number of test applications. Figure 7b is the schematic of ANN evaluated genetic algorithm applied in this experiment[74] (Figure 7b). Tian Xie and Jeffrey C. Grossman develop a crystal graph convolutional neural network to learn material properties from the connection of crystal atoms and accelerate the design of crystalline materials. This neural network can realize crystal materials design with a very high accuracy. In the practical application, it successfully discovered 33 perovskites in dataset, and it significantly reduced the search space of high throughput screening.[75]

Neural network algorithms are also used for drug development. In the past few decades, the connection of computing technology and drug development has become more and more close.[76] At present, many ML algorithms like the neural network have already been applied in the field of drug design, which can help predicting structure, predicting the biological activity, and specifically studying the pharmacokinetics and toxicology of target compounds.[77-79]

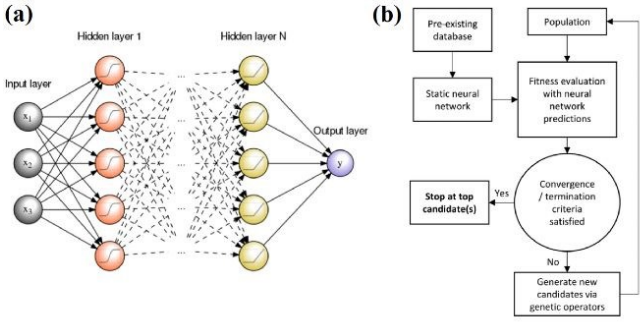


Figure 7. (a) An example of a feed-forward ANN with N hidden layers and a single neuron in the output layer.[80] (b) Schematic representation of an ANN evaluated genetic algorithm.[74]

6) The Deep Learning

The idea of deep learning[81, 82] originated from the research of multiple-layer neural network of ANN. In some way, deep learning is a branch subject of ANN. But now, deep learning has already developed a series of researching ideas and methods of its own. It is a method of characterizing learning based on data. As a new field in machine learning, the deep learning aims to build a neural network that mimics the human brain to analyze data. In some way, deep learning is similar to a neural network with multiple layers. The deep learning algorithm can extract the underlying features in the underlying network, and combine the underlying features in the

upper network layers to obtain more abstract high-level features. As the number of neural network layer increases, the features obtained by the algorithm will be more abstract. At the top level, the final advanced features would be combined enabling the neural network to correctly recognize the object. For example, for a square, deep learning algorithm will first extract features such as "four line segments" and "four right angles". As the neural network layer increase, the algorithm will get abstract features like "four line segments connected" and "four line segments of equal length". And finally, these features are combined and the algorithm correctly recognizes the square. One thing to note here is that, unlike the more primitive machine learning methods, deep learning does not require the researcher to manually implement the feature engineering processing of the input data; instead, the algorithm will self-adjust and choose suitable features independently in continuous learning. This can be considered as a major advancement in machine learning.

The idea of deep learning was proposed in 2006. After more than a decade of research, many important algorithms have been developed, such as Convolutional Neural Networks (CNN), Automatic Encoder, Sparse Coding, Restricted Boltzmann Machine. (RBM) limits Boltzmann machines, Deep Belief Networks (DBN), Recurrent neural network (RNN), and so on. Today, deep learning has been widely used in many fields, such as computer vision, image recognition and natural language recognition. For example, using convolutional neural networks to detect the corrosion of many facilities[83]; Maxim Signaevsky and his coworkers proposes to use the deep learning algorithms to judge the accumulation of abnormal protein TAU to help diagnose neurodegenerative diseases[84] In Figure 8, we can see how they extract image patches for network training and test network's robustness and reliablity with naïve images. Izhar Wallach uses deep learning to predict the biological activity of small molecules in drug discovery[85]. All in all, as a new kind of machine learning method, the deep learning has an excellent prospect of development.

Table 2 summarised some basic algorithms used in material science. Except from the algorithms mentioned above, there are also many methods that have been experimentally tested. Generally, the practical processes are often based on the supervised learning, in which the researchers usually combine personal experience and ML algorithms together. Narrowing down to the specific project, the research idea is not limited to a certain method, and algorithms are also selected and designed individually according to the practical situation.

Table 2. An overview of some basic machine learning algorithms

Algorithms	Brief introduction	Advantages	Disadvantages	Representative applications
Regression Analysis	It could find regression equations and predict depende	Deeply developed and widely used in many occasion.	Needs lots of data, and may be overfitting in practical applications.	Machine Learning with Systematic Density-Functional Theory Calculations: Application to Melting Temperatures of Single- and Binary-component Solids
Naïve Bayes Classifier	It can classify data into several categories following the highest possibility.	Only small amount of data is needed to get essential parameters.	The feature independence hypothesis is not always accurate.	A Naïve-Bayes Classifier for Damage Detection in Engineering Materials
Support Vector machine	SVM can find a hyperplane to divide a group of points into two categories.	It has great generalization ability and can properly handle high-dimension datasets.	SVM is not that appropriate for multiple classification problems.	PVP-SVM: Sequence-Based Prediction of Phage Virion Proteins Using a Support Vector Machine
Decision Tree and Random Forest	By splitting source datasets into several subsets, all data will be judged and classified.	The calculating processes are easy to be comprehended. And it can handle great amount of data.	It's uneasy to gain a high performance decision tree or a random forest. And the overfitting problem may occur.	High-throughput Machine-Learning-Driven Synthesis of Full-Heusler Compounds
Artificial Neural Network	By imitating neuron activities, ANN can automatically find underlying patterns in inputs.	ANN has great self-improving ability, great robustness and high fault-tolerance ability.	Its inner calculation progresses are very difficult to be understood.	Learning from the Harvard Clean Energy Project: the Use of Neural Networks to Accelerate Materials Discovery

Deep Learning	Originated from ANN. It aims to build a neural network to analyze data by imitating human brain.	It has the best self-adjusting and self-improving ability compared with other ML methods.	As a new trend in ML, it has not been well studied yet. Many defects are still unclear.	Artificial Intelligence in Neuropathology: Deep Learning-Based Assessment of Tauopathy

4. Cross-validation

The main goal of machine learning is material predicting, so it is necessary to test the quality of the predicting model. If the model is not flexible enough or the volume of input data is not sufficient enough to find the appropriate physical chemical rules, the predicting results will be not reliable. If the model is too complex, the results may be over-fitted. In order to avoid such possible risks, the researchers need to verify the correctness and reliability of predicting

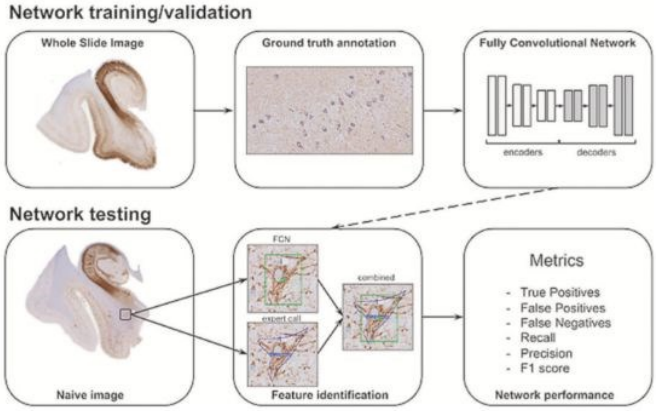


Figure 8. Schematic overview of data annotation and deep learning pipeline in neurodegenerative disease diagnose.[84]

model, and the key to verification is using unknown data to test the model and determine its accuracy. Here, we will briefly introduce several methods of cross-validation. And apart from that, it is also important to know that cross-validation is reliable only when the training sets and validation sets can represent of the entire datasets.[86]

1) The average cross-validation on multiple Hold-out estimates

The average cross-validation[87] is developed on the basis of Hold-out estimation method. The accuracy of original Hold-out validation method is usually lower than the expectation. So after the improvement by Geisser, it was transformed into the average cross-validation method. The average cross-validation method can avoid the random effects caused by one-time division that may occur in the original method. But if the volume of data continues to increase, it will lead to very large computational cost and unaffordable computational complexity.[88, 89]

2) Leave-p-out cross-validation and Leave-one-out cross-validation

In order to reduce the computational complexity, researchers proposed a leave-p-out cross validation (LPO)[90]. In Hold-out estimation, the number of subsets for validating calculation is $\sum_{p=1}^{n} C_n^p$, but in LPO, it is decreased to C_n^p. In this way, the computational complexity is successfully reduced, but the high computational costs caused by huge amount of data is still unacceptable.

Leave-one-out cross validation (LOO)[92] is a special form of leave-p-out cross validation. In LOO, the number p=1, which decreases the number of subsets from C_n^p to n. After years of development, it's now widely used due to its decrease in the volume of computation. However, LOO still have some defects. It may underestimate the predicting error[93] and may also lead to overfitting.[94]

3) Repeated learning test cross-validation

The repeated learning test (RTL) cross-validation[95] was introduced by Breiman and got further studied by Burman and Zhang. It only divides a part of dataset instead of the whole dataset.[95-98] Compared with the previous verification methods, the computational complexity of RLT is significantly reduced.

4) Monte Carlo cross-validation

The Monte Carlo cross validation (MCCV)[99, 100] is similar to RLT, but is easier to operate. The MCCV leaves out a part of the samples every time for validation, then repeats this procedure for many times. Khaled Haddad tried to verify the regional hydrological regression model by LOO and MCCV. Compared with LOO, MCCV can select the more simplified models, and it can also better estimate the predicting ability of models.[101]

5) The K-fold cross-validation

The K-fold cross-validation[102] is proposed as an alternative solution for LOO. It is now the simplest and most widely used generalization error estimation method. The most obvious advantage is that only K times calculation is required, and its calculation cost is far less than the cost of LOO and LPO. But it should be noticed that when the K number is not that big, it may have larger biases.[95]

6) Bootstrap cross-validation

Since K-fold cross-validation tends to have great variations in the cases of small volume of sample data, researchers proposed the bootstrap cross-validation (BCV)[95, 103]. Compared with traditional validation methods, BCV has less variability and less biases under small amount of samples. However, it must be noticed that the calculation amount of BCV will increase sharply under large samples, so it's not recommended using BCV in this situation.[103]

All the analysis shows many different cross validation methods and their unique characteristics. As we can see, more researches are needed to further improve the cross validation methods.[98, 104, 105]

5. Assisting DFT calculation with machine learning

In this section, we will introduce a novel idea that assisting traditional DFT calculation with ML. The first part will thoroughly state the theoretical basis of this idea, and how it works in experiments. And in the second part, we will give serval applying cases of this new idea, and show the great effect and prospect of this idea.

1) The theoretical basis of assisting DFT with ML

View Article Online
DOI: 10.1039/D0NA00378C

Density Functional Theory (DFT) is a quantum mechanical method for studying the electronic structure of multi-electron systems. It has been widely used in material science and computational chemistry because of its high calculating accuracy. However, some defects of DFT are quite obvious in practical applying: the calculation method is too complicated, the calculation processes occupy a lot of computing resources, and the ability of DFT itself is limited by the exchange function that describes the non-classical interaction between electrons. These problems are even more serious when handling complex materials. At present, the new material discovery mode based on DFT may be considered too expensive in computing costs and experimental costs. So, researchers try to assist or partially replace DFT with machine learning in material discovery procedures. As a kind of auxiliary means, machine learning can help avoid the defects in traditional DFT calculation and improve the experimental results in practical applying. In fact, it has been proved that when the amount of data is sufficient, machine learning can reproduce the properties of DFT calculation, and the deviation from the DFT value is less than the deviation between the DFT calculation and the experiment. [80, 106, 107] And when facing with small amount of data, researchers should focus on the dataset itself and try to construct a more delicate and highly efficient dataset to eliminate this deviation. Although Daniel C. Elton has proved that it's possible to reach high accuracy using small amount of data, but the amount of data is still a limitation for ML methods. There are still many details unclear in this field, and more essential researches are need.[31]

From the theoretical point, the study by Rampi Ramprasad reveals an important phenomenon that machine learning based on data prediction are actually consistent with the nature of scientific processes: starting with basic observations (data analysis), then intuition (predictive), and finally build quantitative theory (feedback corrections and learning) that explains observational phenomena. Because of this theoretical support, it is reasonable to assist DFT calculation with machine learning.[9]

Specific to the actual research methods, first, researchers need to represent the various materials in the dataset digitally. Each material (input) should be represented as a string of numbers, which is called the fingerprint vector. The highly distinctive fingerprint vectors are organized by the descriptors and represent the features of this material. Secondly, researchers need to map between input and target features, and this mapping relation is also totally digitized. Many ML algorithms mentioned before can be applied here. When this mapping is established, researchers have the objective conditions to predict new materials with similar materials.

In addition, completely digital mapping means that researchers don't need to consider complex physicochemical relationships when discovering new materials. Since the original ones (input) have all the above physicochemical properties, correspondingly, the target materials naturally conform to these physicochemical properties. This is an essential theory that focuses on data, and can greatly reduce the computational pressure of existing methods.

2) The practical application cases

There are already many researchers who have been involved in this work. For example, I E Castelli and K W Jacobsen conduct a study about the perovskite crystals of the ABO_2N cubic structure. They use machine learning instead of traditional DFT calculation to calculate the tendency of various elements forming perovskites, and they successfully perform simple data manipulation for bandgap evaluation.[108] Ghanshyam Pilania and the colleagues try to use chemical structures and electron charge density as "fingerprint vectors" to find new material properties. And the results show same-level accuracy and lower experimental consumption compared with DFT.[109] However, this method has a certain limitation, that the maps corresponding to each type of materials are totally different, which can only be applied on specific type of materials. Under this circumstance, only finding a special map for each type of material can meet the research needs. And what is worth mentioning is that Logan Ward attempts to replace the dataset with a set of universal material feature, then they use the features set as "fingerprints" to find a broad mapping property for the vast majority. This method can analyze most materials with a general model and avoid calculating a mapping for each material, thus greatly saving research costs. In this research, they select the random forest algorithm to build the model and then test it for twice. In the first experiment (using a DFT-validated property database to predict new solar cell materials), the model showed an excellent performance with the lowest average absolute error in ten times of cross-validations. In the second experiment (exploring new metal-glass alloys using experimentally measured data on the ability of crystalline compounds to form glass), it showed a fairly high 90.1% accuracy in ten times of cross-validations.[110]

In addition, in the field of assisting DFT with machine learning algorithms, one of the most typical case is that Shuaihua Lu's team used this method to find the suitable mixed organic-inorganic calcium-titanium (HOIP) as a new type of photovoltaic material. The researchers used the machine learning method to train the model of bandgap, and selected four ideal solar cell properties as the measurement indicators, then they successfully selected six suitable materials from 5158 kinds of undetected target compounds. The results are listed in Figure 9. As the busting iteration numbers increasing, the deviation between training and testing sets are decreasing (Figure 9a). The distribution of post predicted bandgap is also very close to original input sets (Figure 9b). And other hidden trends and periodicities are also unraveled by data visualization. The results are showed in Figure 9c-f, which is divided according to the position of X-site element as F, Cl, Br and I. Being different from the traditional methods, DFT is only used here to help calculating the most probable target materials rather than all target substances, thus greatly reducing the computational costs.[28]

In practical application, the experimental accuracy of ML can also be maintained. For example, Prasanna Balachandran and colleagues construct a new ML material discovery method by constructing orbital radii based on data, and applied this method from relatively simple AB compounds, to more electronically complex RM intermetallic. The results show great agreement of classification rules extracted from both ML and DFT calculation.[111] In another case, Daniele Dragoni constructs a Gaussian Approximation Potential (GAP) model, and trained this model by DFT data from 150k atoms. After the finishing the construction, researchers verified the accuracy of model by another group of DFT data. The results show that this model can reproduce the DFT calculation and keep a very high accuracy. For example, the value of bulk modulus by GAP is 198.2, while it by DFT is 199.8 ± 0.1[7]. The deviation between these

two methods is quite small.[112] What's more, there are also many other cases proving the high accuracy of ML method, such as Felix Faber uses 13 kinds of quantum properties from over 117k atoms to test the accuracy of ML and DFT calculation, and it shows that ML can definitely reach the accuracy of DFT.[113] Albert P. Bartók and coworkers build a ML model which can distinguish active and inactive protein ligands with more than 99% reliability. This experience dependent method can reach same level of accuracy of DFT with quite low cost.[114] Ryosuke Jinnouchi and Ryoji Asahi try use ML to detect the catalytic activity of nanoparticles with DFT data on single crystals. And the accuracy can meet the expect of practical application.[115]

Except from all the practical examples above, researchers have also attempted to optimize the DFT calculation process by ML. Solving the Kohn-Sham equation is a necessary step in the DFT calculations process, and this very step is also the most time-consuming part of the DFT calculation process. In this field, now we can see that John C. Snyder and Felix Brockherde have already made some significant results.[116, 117] The results have shown that taking the advantage of the flexibility and efficiency of ML can highly reduce the computational cost of DFT, which can decrease the calculation time and increase the calculation speed.

From these examples we can see that the methods and ideas of ML has brought great convenience to the material research. Since it is a pure data operation, the computer can quickly determine all physical and chemical rules by sufficient data and the correct algorithm, no matter they are hidden or discovered, which may back-feed the theoretical chemistry someday.[118] The method of ML combined with data operation has less computational complexity and computational cost compared with traditional DFT calculation. However, the accuracy of this new idea is now behind the expectation. Ying Zhang and his team try to introduce additional new parameters into the calculation to improve the accuracy of the models, but the best ones still don't have the high accuracy as the traditional DFT calculations.[36] From this point of view, although the idea of ML combined with data operation can bring great changes to today's material science research, but there are also defects to be overcame.

6. Artificial intelligence assisting new material development

Artificial intelligence (AI) is subject of computer science simulating human intelligence. Since its birth in 1950, it went through many ups and downs. Today, thanks to the development of big data and computer technology, the theoretical system of AI is hugely enriched. Now AI has many subfields like data mining, computer vision, machine and so on. Moreover, it has shown great potential and ability in material science, and it is widely used in material design, corrosion detection, material screening and so many other fields of material science. In this section, we will introduce some subfields of AI and their applications in material science.[119]

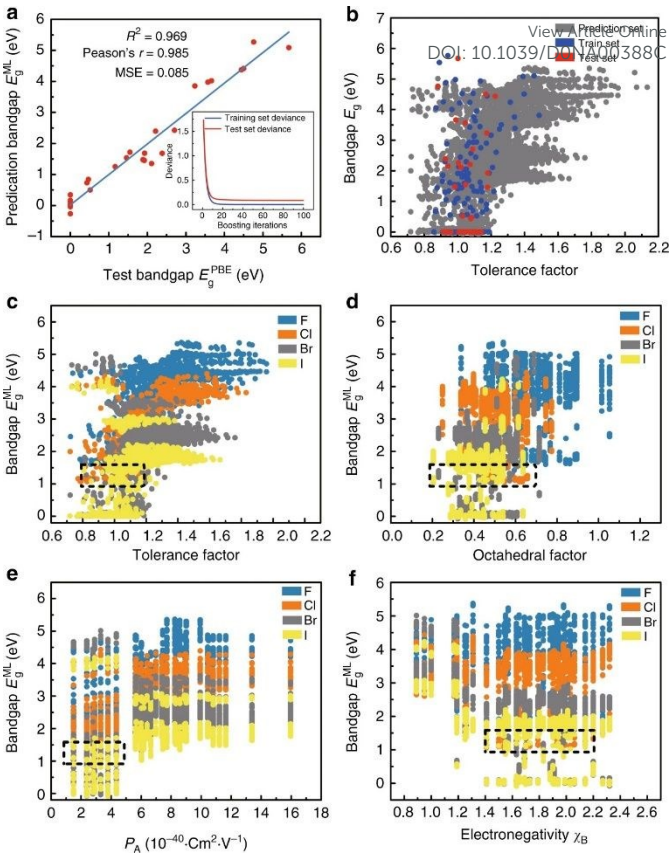


Figure 9. Results and insights from ML model. (a) The fitting results of test bandgaps EPBE g and predicted bandgaps EML g. (b) Scatter plots of tolerance factors against the bandgaps for the prediction dataset from trained ML model (blue, red and dark gray plots represent train, test and prediction set, respectively). Data visualization of predicted bandgaps for all possible HOIPs (one color represents a class of halogen perovskites) with (c) tolerance factor, (d) octahedral factor, (e) ionic polarizability for the A-site ions, and (f) electronegativity of B-site ions. The dotted box represents the most appropriate range for each feature.[28]

1) Inverse design for desired compounds

Inverse design aims to find materials from desired particular or material functionalities. Inverse design is significantly different from forward development. The traditional forward design is to obtain the target materials through experiments, and then further judge the functionalities of the materials. The inverse design shows a more obvious goal-oriented characteristic. It starts from desired properties and ends in chemical space. We can see their difference in Figure 10a. But inverse design faces a very obvious problem. As inverse design is to find suitable materials based on functionalities, we can consider this process as starting from the specific conditions to find the possible solutions in a large range of candidate materials and material combinations. According to the current researches, the optimal solution may or may not exist, and there may be one or more

solutions. However, in the process of searching for the most optimal solution, the number of candidates waiting for analyzing will also increase tremendously. For example,

for the pharmacologically relevant small molecules, the number of structures is considered to be nearly 10.[60, 120] Data of this volume can't be verified manually. Therefore, we introduce ML in this process to help researchers for analyzing.

A basic idea of inverse design originates from high-throughput virtual screening. It's a data-driven experimental idea. There are many successful application examples. For example, Benjamin Sanchez-Lengeling1 and Alán Aspuru-Guzik once thoroughly explored the inverse design method. From the current point of view, the research idea and implementation method of inverse design are quite diversiform, but still not mature. Some specific procedures such as the digital representation of molecules, the selection of ML methods, and the design of inverse design tools, still need further study.[121]

2) Computer vision for material-picture analysis

Computer vision is defined as an artificial intelligence system that can extract information from images or multidimensional data. Computer vision technology originated very early, but it's not until in the 1970s, when the ability of computers was improved high enough to handle large-scale data such as images, computer vision got attention and development. For example, in the late 1980s, Y. LeCun used the letters in the US postal system as an example, and applied computer vision technology to analyze and determine the handwritten zip codes on the letters.[122] Today, computer vision has become an interdisciplinary technology widely used in various related fields, and it has also been applied in many subfield of material science. In material science, computer vision can analyze unclear or unknown material properties from huge amount of figures, which can greatly help scientists understanding the physical chemical properties and their inner relationships of similar materials. In this way, scientists can better construct and design the target novel materials. For example, detecting corrosion of concrete ties of railway by analyzing rail pictures;[123] or exploring particle morphology and surface textures of powder by analyzing microstructure pictures of powder for additive manufacturing processes.[124] In the first case, we can see that computer divides every tie in four pictures and separately evaluate them to detect the condition of crumbling and chipping (Figure 10b). In the last case, researchers used a "bag of visual word" method, and the final classification accuracy was 80%-85%. Another typical case is that M.X. Bastidas-Rodriguez uses machine learning algorithms (ANN and SVM) to classify the fracture of metal materials for failure analysis. The results show that the performance of the ANN application is slightly stronger than the SVM, and the highest accuracy percentage is 84.95%.[125] In addition, some researchers have further explored the theory of computer vision. For example, Brian L. DeCost and Elizabeth A. Holm have tried to develop a general method to find useful characteristics and relationships of massive microstructures in large and diverse microstructural image databases. In this research, a "visual dictionary" method was created to detect the characteristics of images by extracting meaningful details and features in pictures to construct "visual words" and expressing pictures as the probability distribution of visual words.[126] Moreover, computer vision also plays an important role in material classification. For example, Hua Chen and the colleagues propose a polarization phase-based method in computer vision for material classification according to intrinsic electrical conductivity. This method is computationally efficient and can be achieved with existing image technology.[127]

3) High-throughput screening & Big data in material discovery

The high-throughput screening in the field of novel material discovery means using tremendous volume of data to conduct computational tasks for detecting materials properties and helping design target material. And the big data can be defined as a research method extracting information and detecting the relationships from the extraordinary huge datasets. These two ideas are now often used in the novel material discovery.[128-130] Researchers collect huge volume of data about target materials, and use high-throughput screening to analyze material properties or possibility of synthetizing target materials. Considering the need of data when applying ML, these two methods has literally become the foundations of novel material discovery field. And there are lots of cases proving it, for example, Courtney R. Thomas's team builds a high-throughput screening to estimate the safety, nanotoxicology, ecotoxicology, environment assessment and other properties of engineered nanomaterials.[131] Jeff Greeley and his colleagues use a DFT-based high-throughput screening to estimate activity of over 700 binary surface alloys to find an appropriate electro-catalyst for the hydrogen evolution reaction, and they successfully find the BiPt as the desired target material[132] (Figure 10). Kirill Sliozberg's team creates an automated optical scanning droplet cell for high-throughput screening and high-throughput preparation. And they apply this tool on the evaluation on thin-film Ti-W-O and thin-film Fe-W-O to seek the efficient semiconductors.[133, 134] And in another case, Ankit Agrawala and Alok Choudhary systemically state the important role big data plays in materials informatics nowadays.[135]

Except from ML, other AI algorithms are also widely used in the field of novel material discovery, and there are tons of cases showing the importance, effect, and development prospect of AI. In summary, AI has strong effect and bright future in materials science, but it still needs further development.

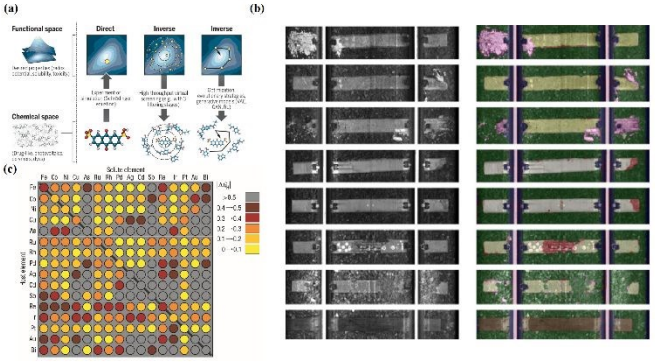


Figure 10. (a) Schematic of different approaches toward molecular design. Inverse design starts from desired properties and ends in chemical space, unlike the direct approach that leads from chemical space to the properties.[121] (b) Computer vision analyzes pictures to detect rail defects.[123] (c) The computational high-throughput screening for $|\Delta G_H|$ on 256 pure metals and surface alloys. The rows indicate the identity of the pure metal substrates, the columns indicate the identity of the solute embedded in the surface layer of the substrate, and the diagonal of the plot corresponds to the hydrogen-adsorption free-energy on the pure metals.[131]

7. Prospect & Conclusion

Leading by MGI, the era of data driven material discovery has come. Over the past decades, by the development of computing technology, the progress of AI science, and the abundant experimental results, this new material discovery method has generally become a research paradigm. As the research deepening, it has shown many advancing abilities, like low experimental consumption, low time consumption, high generalizing ability and high density analysis. Now it's used in basic chemistry, pharmaceutical science and materials science, such as terahertz spectral analysis and recognition[136], prediction of the melting temperature of binary compounds[48], band gap energy of certain crystals[137, 138], and analysis of complex reaction networks[139] etc.

However, this method also has certain defects. The result accuracy highly depends on the quality of data and algorithms. Some defects like computing consumption, reliability of algorithms and dependence of data need to be overcame. We need to take measures to complete this method, including establish reliable databases, enhance the combination between ML and other material theories, and explore other new material researching methods like inverse design. Now we are making progress in this field, and gradually improve this method. With the development of researches, it will reveal more important effect.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

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).

Notes and references

1. M. I. Jordan and T. M. Mitchell, *Science*, 2015, **349**, 255-260.
2. A. L. Blum and P. Langley, *Artificial Intelligence*, 1997, **97**, 245-271.
3. E. Lopez, D. Gonzalez, J. V. Aguado, E. Abisset-Chavanne, E. Cueto, C. Binetruy and F. Chinesta, *Archives of Computational Methods in Engineering*, 2016, **25**, 59-68.
4. W. Lu, R. Xiao, J. Yang, H. Li and W. Zhang, *Journal of Materiomics*, 2017, **3**, 191-201.
5. P. Raccuglia, K. C. Elbert, P. D. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier and A. J. Norquist, *Nature*, 2016, **533**, 73-76.
6. Yang Xiaoyu, Wang Juan, Ren Jie, Song Jianlong, Wang Zongguo, Zeng Zhi, Zhang Xiaoli, Huang Sunchao, Zhang Ping and Lin Haiqing, *Chinese Journal of Computational Physics*, 2017, **34**, 697-704.
7. Lin Hai, Zheng Jiaxin, Lin Yuan and P. Feng, *Energy Storage Science and Technology*, 2017, **6**, 990-999.
8. Yang Xiaoyu, Ren Jie, Wang Juan, Zhao Xushan, Wang Zongguo and S. Jianlong, *Science & Technology Review*, 2016, **34**, 62-67.
9. R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi and C. Kim, *npj Computational Materials*, 2017, **3**, 1-13.
10. P. De Luna, J. Wei, Y. Bengio, A. Aspuru-Guzik and E. Sargent, *Nature*, 2017, **552**, 23-27.
11. A. Zheng and A. Casari, *Feature engineering for machine learning: principles and techniques for data scientists*, O'Reilly Media, Inc., Sebastopol, State of California, USA, 1 edn., 2018.
12. Y. Liu, T. Zhao, W. Ju and S. Shi, *Journal of Materiomics*, 2017, **3**, 159-177.
13. L. Yang and G. Ceder, *Physical Review B*, 2013, **88**, 224107.
14. K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller and E. K. U. Gross, *Physical Review B*, 2014, **89**, 205118.
15. H. K. D. H. Bhadeshia, R. C. Dimitriu, S. Forsik, J. H. Pak and J. H. Ryu, *Materials Science and Technology*, 2013, **25**, 504-510.
16. Yin Haiqing, Jiang Xue, Zhang Ruijie, Liu Guoquan, Zheng Qingjun and Q. Xuanhui, *Materials China*, 2017, **36**, 401-405+454.
17. A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner and G. Ceder, *Apl Materials*, 2013, **1**, 011002.
18. E. O. Pyzer-Knapp, K. Li and A. Aspuru-Guzik, *Advanced Functional Materials*, 2015, **25**, 6495-6502.
19. K. Choudhary, I. Kalish, R. Beams and F. Tavazza, *Scientific Reports*, 2017, **7**, 1-16.
20. S. Lebègue, T. Björkman, M. Klintenberg, R. M. Nieminen and O. Eriksson, *Physical Review X*, 2013, **3**, 031002.
21. M. Ashton, J. Paul, S. B. Sinnott and R. G. Hennig, *Physical Review Letters*, 2017, **118**, 106101.
22. N. Mounet, M. Gibertini, P. Schwaller, D. Campi, A. Merkys, A. Marrazzo, T. Sohier, I. E. Castelli, A. Cepellotti, G. Pizzi and N. Marzari, *Nature Nanotechnology*, 2018, **13**, 246-252.
23. 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 and K. S. Thygesen, *2D Materials*, 2018, **5**, 042002.
24. S. Kotsiantis, D. Kanellopoulos and P. Pintelas, *International Journal of Computer Science*, 2006, **1**, 111-117.
25. A. Holzinger, 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA), Kosice, 2018.
26. J. H. Friedman, *Computing Science and Statistics*, 1998, **29**, 3-9.
27. K. Lakshminarayan, S. A. Harp and T. Samad, *Applied Intelligence*, 1999, **11**, 259-275.
28. S. Lu, Q. Zhou, Y. Ouyang, Y. Guo, Q. Li and J. Wang, *Nature Communications*, 2018, **9**, 1-8.

29. M. Wang, T. Wang, P. Cai and X. Chen, *Small Methods*, 2019, **3**, 1900025.
30. L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl and M. Scheffler, *Physical Review Letters*, 2015, **114**, 105503.
31. D. C. Elton, Z. Boukouvalas, M. S. Butrico, M. D. Fuge and P. W. Chung, *Scientific Reports*, 2018, **8**, 9059.
32. A. O. Oliynyk, E. Antono, T. D. Sparks, L. Ghadbeigi, M. W. Gaultois, B. Meredig and A. Mar, *Chemistry of Materials*, 2016, **28**, 7324-7331.
33. F. Legrain, J. Carrete, A. van Roekeghem, S. Curtarolo and N. Mingo, *Chemistry of Materials*, 2017, **29**, 6220-6227.
34. S. Kajita, N. Ohba, R. Jinnouchi and R. Asahi, *Scientific Reports*, 2017, **7**, 1-9.
35. P. Pankajakshan, S. Sanyal, O. E. de Noord, I. Bhattacharya, A. Bhattacharyya and U. Waghmare, *Chemistry of Materials*, 2017, **29**, 4190-4201.
36. Y. Zhang and C. Ling, *npj Computational Materials*, 2018, **4**, 1-8.
37. A. Jain and T. Bligaard, *Physical Review B*, 2018, **98**, 214112.
38. M. Seeger, *International Journal of Neural Systems*, 2004, **14**, 69-106.
39. D. H. Wolpert and W. G. Macready, *IEEE Transactions on Evolutionary Computation*, 1997, **1**, 67-82.
40. G. W. Haggstrom, *Journal of Business Economic Statistics*, 1983, **1**, 229-238.
41. S. Wold, M. Sjöström and L. Eriksson, *Chemometrics and Intelligent Laboratory Systems*, 2001, **58**, 109-130.
42. R. Wehrens and B.-H. Mevik, *Journal of Statistical Software*, 2007, **18**, 1-23.
43. V. Esposito Vinzi and G. Russolillo, *Wiley Interdisciplinary Reviews: Computational Statistics*, 2013, **5**, 1-19.
44. K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda and B. Schölkopf, *IEEE Transactions on Neural Networks*, 2001, **12**, 181-201.
45. C.-C. Chang and C.-J. Lin, *ACM Transactions on Intelligent Systems and Technology*, 2011, **2**, 1-27.
46. C. K. Williams, in *Learning in graphical models*, ed. M. I. Jordan, Springer Science & Business Media, Dordrecht, The Netherlands, 1 edn., 1998, vol. 89, ch. 23, pp. 599-621.
47. C. E. Rasmussen, Summer School on Machine Learning, Tübingen,Germany, 2003.
48. A. Seko, T. Maekawa, K. Tsuda and I. Tanaka, *Physical Review B*, 2014, **89**, 054303.
49. S. Curtarolo, D. Morgan, K. Persson, J. Rodgers and G. Ceder, *Physical Review Letters*, 2003, **91**, 135503.
50. B. Tutmez and A. Dag, *Computers & Concrete*, 2012, **10**, 457-467.
51. I. Rish, IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, Seattle,State of Washington,USA, 2001.
52. D. D. Lewis, European Conference on Machine Learning, Chemnitz, Germany, 1998.
53. Q. Wang, G. M. Garrity, J. M. Tiedje and J. R. Cole, *Applied and Environmental Microbiology*, 2007, **73**, 5261-5267.
54. O. Addin, S. M. Sapuan, E. Mahdi and M. Othman, *Materials & Design*, 2007, **28**, 2379-2386.
55. H. Liu, X. Song, J. Bimbo, L. Seneviratne and K. Althoefer, 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, Algarve, Portugal, 2012.
56. C. J. Burges, *Data Mining and Knowledge Discovery*, 1998, **2**, 121-167.
57. M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt and B. Scholkopf, *IEEE Intelligent Systems & Their Applications*, 1998, **13**, 18-28.
58. X. Qiu, D. Fu, Z. Fu, K. Riha and R. Burget, 2011 34th International Conference on Telecommunications and Signal Processing (TSP), Budapest, Hungary, 2011.
59. B. Manavalan, T. H. Shin and G. Lee, *Front Microbiol*, 2018, **9**, 476.
60. M. K. Warmuth, J. Liao, G. Ratsch, M. Mathieson, S. Putta and C. Lemmen, *Journal of Chemical Information and Modeling*, 2003, **43**, 667-673.
61. J. R. Quinlan, *Machine learning*, 1986, **1**, 81-106.
62. A. Ehrenfeucht, D. J. I. Haussler and Computation, *Information and Computation*, 1989, **82**, 231-246.
63. S. R. Safavian and D. Landgrebe, *IEEE Transactions on Systems, Man, and Cybernetics*, 1991, **21**, 660-674.
64. L. Breiman, *Machine learning*, 2001, **45**, 5-32.
65. A. Liaw and M. Wiener, *R news*, 2002, **2**, 18-22.
66. B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary and C. Wolverton, *Physical Review B*, 2014, **89**, 094104.
67. J. Carrete, W. Li, N. Mingo, S. Wang and S. Curtarolo, *Physical Review X*, 2014, **4**, 011019.
68. D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher and A. G. Doyle, *Science*, 2018, **360**, 186-190.
69. J. Zhang, G. Ma, Y. Huang, J. sun, F. Aslani and B. Nener, *Construction and Building Materials*, 2019, **210**, 713-719.
70. G. P. Zhang, *IEEE Transactions on Systems, Man, and Cybernetics, Part C-Applications and Reviews*, 2000, **30**, 451-462.
71. G. B. Goh, N. O. Hodas and A. Vishnu, *Journal of Computational Chemistry*, 2017, **38**, 1291-1307.
72. H. D. Olding, *The organization of behavior: A neuropsychological theory*, Psychology Press, Mahwah, State of New Jersey, USA, 1 edn., 2005.
73. T. Zhang, J. Wang, Q. Liu, J. Zhou, J. Dai, X. Han, Y. Zhou and K. Xu, *Photonics Research*, 2019, **7**, 368-380.
74. T. K. Patra, V. Meenakshisundaram, J.-H. Hung and D. S. Simmons, *ACS Combinatorial Science*, 2017, **19**, 96-107.
75. T. Xie and J. C. Grossman, *Physical Review Letters*, 2018, **120**, 145301.
76. C. Nantasenamat, C. Isarankura-Na-Ayudhya and V. Prachayasittikul, *Expert Opinion on Drug Discovery*, 2010, **5**, 633-654.
77. V. G. Maltarollo, J. C. Gertrudes, P. R. Oliveira and K. M. Honorio, *Expert Opinion on Drug Metabolism & Toxicology*, 2015, **11**, 259-271.
78. T. Fox and J. M. Kriegl, *Current Topics in Medicinal Chemistry*, 2006, **6**, 1579-1591.
79. A. N. Lima, E. A. Philot, G. H. Trossini, L. P. Scott, V. G. Maltarollo and K. M. Honorio, *Expert Opinion on Drug Discovery*, 2016, **11**, 225-239.

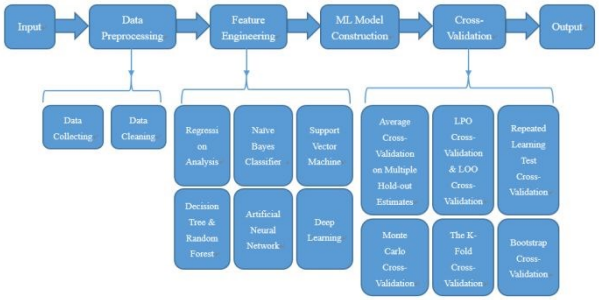
80. G. R. Schleder, A. C. M. Padilha, C. M. Acosta, M. Costa and A. Fazzio, *Journal of Physics: Materials*, 2019, **2**, 032001.
81. Deng Li and D. Yu, *Signal Processing*, 2014, **7**, 197-387.
82. Y. LeCun, Y. Bengio and G. Hinton, *Nature*, 2015, **521**, 436-444.
83. W. Nash, T. Drummond and N. Birbilis, *npj Materials Degradation*, 2018, **2**, 1-12.
84. M. Signaevsky, M. Prastawa, K. Farrell, N. Tabish, E. Baldwin, N. Han, M. A. Iida, J. Koll, C. Bryce, D. Purohit, V. Haroutunian, A. C. McKee, T. D. Stein, C. L. White, 3rd, J. Walker, T. E. Richardson, R. Hanson, M. J. Donovan, C. Cordon-Cardo, J. Zeineh, G. Fernandez and J. F. Crary, *Laboratory Investigation*, 2019, **99**, 1019.
85. I. Wallach, M. Dzamba and A. Heifets, *Abstracts of Papers of the American Chemical Society*, 2016, **251**, 1.
86. K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, *Nature*, 2018, **559**, 547-555.
87. S. Geisser, *Journal of the American Statistical Association*, 1975, **70**, 320-328.
88. A. Luntz, *Technicheskaya Kibernetica*, 1969, **3**.
89. L. Bo, L. Wang and L. Jiao, *Neural Computation*, 2006, **18**, 961-978.
90. C. R. Rao and Y. Wu, *Journal of Statistical Planning and Interfence*, 2005, **128**, 231-240.
91. A. Celisse and S. Robin, *Computational Statistics & Data Analysis*, 2008, **52**, 2350-2368.
92. M. Kearns and D. Ron, *Neural Computation*, 1999, **11**, 1427-1453.
93. B. Efron, *Journal of the American Statistical Association*, 1986, **81**, 461-470.
94. A. K. Smilde, *Journal of Quality Technology*, 2018, **34**, 464-465.
95. P. Burman, *Biometrika*, 1989, **76**, 503-514.
96. C. Nadeau and Y. Bengio, Advances in Neural Information Processing Systems, Denver, Colorado, USA, 2000.
97. P. Zhang, *Annals of Statistics*, 1993, **21**, 299-313.
98. S. Arlot and A. Celisse, *Statistics Surveys*, 2010, **4**, 40-79.
99. R. R. Picard and R. D. Cook, *Journal of the American Statistical Association*, 1984, **79**, 575-583.
100. Q.-S. Xu and Y.-Z. Liang, *Chemometrics and Intelligent Laboratory Systems*, 2001, **56**, 1-11.
101. K. Haddad, A. Rahman, M. A Zaman and S. Shrestha, *Journal of Hydrology*, 2013, **482**, 119-128.
102. J. D. Rodriguez, A. Perez and J. A. Lozano, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, **32**, 569-575.
103. W. J. Fu, R. J. Carroll and S. Wang, *Bioinformatics*, 2005, **21**, 1979-1986.
104. W. Y. Yang Liu, *Application Research of Computers*, 2015, **32**, 1287-1290+1297.
105. S. Borra and A. Di Ciaccio, *Computational Statistics & Data Analysis*, 2010, **54**, 2976-2989.
106. L. Ward, R. Liu, A. Krishna, V. I. Hegde, A. Agrawal, A. Choudhary and C. Wolverton, *Physical Review B*, 2017, **96**, 024104.
107. F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley and O. A. von Lilienfeld, *Journal of Chemical Theory and Computation*, 2017, **13**, 5255-5264.
108. I. E. Castelli and K. W. Jacobsen, *Modelling and Simulation in Materials Science and Engineering*, 2014, **22**, 055007.
109. G. Pilania, C. Wang, X. Jiang, S. Rajasekaran and R. Ramprasad, *Scientific Reports*, 2013, **3**, 2810.
110. L. Ward, A. Agrawal, A. Choudhary and C. Wolverton, *npj Computational Materials*, 2016, **2**, 16028.
111. P. V. Balachandran, J. Theiler, J. M. Rondinelli and T. Lookman, *Scientific reports*, 2015, **5**, 13285.
112. D. Dragoni, T. D. Daff, G. Csányi and N. Marzari, *Physical Review Materials*, 2018, **2**, 013808.
113. F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley and O. A. von Lilienfeld, *arXiv preprint*, 2017, arXiv:05532.
114. A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi and M. Ceriotti, *Science Advances*, 2017, **3**, e1701816.
115. R. Jinnouchi and R. Asahi, *The Journal of Physical Chemistry Letters*, 2017, **8**, 4279-4283.
116. J. C. Snyder, M. Rupp, K. Hansen, K. R. Muller and K. Burke, *Physical Review Letters*, 2012, **108**, 253002.
117. F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke and K. R. Muller, *Nature Communications*, 2017, **8**, 872.
118. S. V. Kalinin, B. G. Sumpter and R. K. Archibald, *Nature Materials*, 2015, **14**, 973-980.
119. M. Haenlein and A. Kaplan, *California Management Review*, 2019, **61**, 5-14.
120. A. M. Virshup, J. Contreras-García, P. Wipf, W. Yang and D. N. Beratan, *Journal of the American Chemical Society*, 2013, **135**, 7296-7303.
121. B. Sanchez-Lengeling and A. Aspuru-Guzik, *Science*, 2018, **361**, 360-365.
122. Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel, *Neural Computation*, 1989, **1**, 541-551.
123. X. Gibert, V. M. Patel and R. Chellappa, *IEEE Transactions on Intelligent Transportation Systems*, 2017, **18**, 153-164.
124. B. L. DeCost and E. A. Holm, *Computational Materials Science*, 2017, **126**, 438-445.
125. M. X. Bastidas-Rodriguez, F. A. Prieto-Ortiz and E. Espejo, *Engineering Failure Analysis*, 2016, **59**, 237-252.
126. B. L. DeCost and E. A. Holm, *Computational materials science*, 2015, **110**, 126-133.
127. H. Chen and L. B. Wolff, *International Journal of Computer Vision*, 1998, **28**, 73-83.
128. C. L. Philip Chen and C.-Y. Zhang, *Information Sciences*, 2014, **275**, 314-347.
129. C. Xue-Wen and L. Xiaotong, *IEEE Access*, 2014, **2**, 514-525.
130. L. Zhou, S. Pan, J. Wang and A. V. Vasilakos, *Neurocomputing*, 2017, **237**, 350-361.
131. C. R. Thomas, S. George, A. M. Horst, Z. Ji, R. J. Miller, J. R. Peralta-Videa, T. Xia, S. Pokhrel, L. Mädler and J. L. Gardea-Torresdey, *ACS Nano*, 2011, **5**, 13-20.

View Article Online
DOI: 10.1039/D0NA00388C

132. J. Greeley, T. F. Jaramillo, J. Bonde, I. Chorkendorff and J. K. Nørskov, *Nature Materials*, 2006, **5**, 909-913.
133. K. Sliozberg, D. Schäfer, T. Erichsen, R. Meyer, C. Khare, A. Ludwig and W. Schuhmann, *ChemSusChem*, 2015, **8**, 1270-1278.
134. R. Meyer, K. Sliozberg, C. Khare, W. Schuhmann and A. Ludwig, *ChemSusChem*, 2015, **8**, 1279-1285.
135. A. Agrawal and A. Choudhary, *APL Materials*, 2016, **4**, 053208.
136. Z. Yue, S. Ji, Y. Sigang, C. Hongwei and X. Kun, *Radio Engineering*, 2019, **49**, 1031-1036.
137. P. Dey, J. Bible, S. Datta, S. Broderick, J. Jasinski, M. Sunkara, M. Menon and K. Rajan, *Computational Materials Science*, 2014, **83**, 185-195.
138. G. Pilania, A. Mannodi-Kanakkithodi, B. P. Uberuaga, R. Ramprasad, J. E. Gubernatis and T. Lookman, *Scientific Reports*, 2016, **6**, 19375.
139. Z. W. Ulissi, A. J. Medford, T. Bligaard and J. K. Norskov, *Nature Communications*, 2017, **8**, 1-7.

Table of Contents:

View Article Online
DOI: 10.1039/D0NA00388C



This paper summarizes the idea, operations and workflows of how machine learning driven new material discovery.