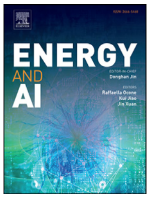
Contents lists available at ScienceDirect

Energy and AI

journal homepage: www.elsevier.com/locate/egyai



Review

Machine learning for advanced energy materials

Yun Liu, Oladapo Christopher Esan, Zhefei Pan, Liang An*

Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong SAR, China

HIGHLIGHTS

- The application roadmap to carbon neutrality is presented.
- A comprehensive review of fundamental ML tutorials is provided.
- The latest progress in data-driven materials science and engineering is discussed.
- The keys to successful ML applications and remaining challenges are highlighted.

ARTICLE INFO

Article history:
Received 31 December 2020
Received in revised form 16 January 2021
Accepted 16 January 2021
Available online 24 January 2021

Keywords:
Energy materials
Artificial intelligence
Machine learning
Data-driven materials science and engineering
Prediction of materials properties
Design and discovery of energy materials

ABSTRACT

The screening of advanced materials coupled with the modeling of their quantitative structural-activity relationships has recently become one of the hot and trending topics in energy materials due to the diverse challenges, including low success probabilities, high time consumption, and high computational cost associated with the traditional methods of developing energy materials. Following this, new research concepts and technologies to promote the research and development of energy materials become necessary. The latest advancements in artificial intelligence and machine learning have therefore increased the expectation that data-driven materials science would revolutionize scientific discoveries towards providing new paradigms for the development of energy materials. Furthermore, the current advances in data-driven materials engineering also demonstrate that the application of machine learning technology would not only significantly facilitate the design and development of advanced energy materials but also enhance their discovery and deployment. In this article, the importance and necessity of developing new energy materials towards contributing to the global carbon neutrality are presented. A comprehensive introduction to the fundamentals of machine learning is also provided, including open-source databases, feature engineering, machine learning algorithms, and analysis of machine learning model. Afterwards, the latest progress in data-driven materials science and engineering, including alkaline ion battery materials, photovoltaic materials, catalytic materials, and carbon dioxide capture materials, is discussed. Finally, relevant clues to the successful applications of machine learning and the remaining challenges towards the development of advanced energy materials are highlighted.

1. Introduction

With the increasing global environmental issues, it has become a global consensus to earnestly develop clean and renewable energy technologies to achieve carbon-neutral society in the next few decades [1,2]. One of the crucial means to attain large-scale application of green energy is the development of advanced energy materials towards enabling efficient energy conversion and stable power output [3,4]. The traditional ways to discover and design energy materials include laboratory exploration and simulation activities [5]. It is therefore a time-consuming process, while the number of explored new material samples is also limited [6]. Additionally, the success probability of these traditional methods is low [7]. During the last few decades, the density functional theory (DFT) calculation method was frequently applied to screen new materials [8]. This is majorly because the DFT is able to sustain large space searching and provides higher computational accuracy [9]. However, there are still some disadvantages of employing DFT calculation, such as high computational cost.

The recent progress of artificial intelligence (AI) technology in various research fields has demonstrated the great potentials of the application of AI in seeking new and energy-efficient materials [10,11]. While AI is a technology which enables a machine to simulate human behavior; machine learning (ML), a subset of AI, leverages algorithms and models to learn from past data or existing knowledge [12,13]. ML can therefore be used to accelerate materials development due to its inherent strong ability in processing massive data and high-dimensional analysis [14,15]. For instance, in order to obtain new polymer membrane materials, Barnett et al. [16] developed an ML model based on Gaussian regression process. Using gas permeability data of approximately 700 polymers, the ML model predicted the gas separation behavior of more than 11,000 untested homopolymers. Elsewhere, Gayon-Lombardo et al. [17] applied the Deep Convolution Generative Adver-

* Corresponding author.
E-mail address: liang.an@polyu.edu.hk (L. An).

https://doi.org/10.1016/j.egyai.2021.100049

sarial Network (DC-GAN) to generate real n-phase microstructure data of multiphase porous electrodes. The results reveal that the proposed method can greatly reduce the computational cost of electrochemical simulations. In addition to the above successful application cases, the advancements of data-driven material science in the past decade have also shown that the ML technology can significantly contribute towards the development of new materials [18,19]. One of the most common applications of ML technology in materials community is to screen high-performance materials, which highly relies on the extensive search capabilities and precise classification of ML algorithms [20–22]. Additionally, the use of ML models to realize accurate prediction for materials properties has gradually received increasing attention [23,24]. The reason is that the information predicted by ML can not only reveal the characteristics of the tested material, but also guide the next round of experiments [25,26]. Therefore, a rational materials design can be achieved with the help of ML technology [27]. In terms of the previously mentioned challenges for developing energy materials, ML was thus considered as an effective tool to address current issues, which can facilitate the design, discovery, and deployment of advanced energy materials [28–30]. Moreover, the energy materials development process can be further accelerated by integrating the ML with intelligent robots [31]. These aforementioned prospects and progresses not only verify the feasibility of materials genomics, but also present the potential of accelerating the development of zero-emission society [32].

The summary of state-of-the-art attempts on data-driven materials science can therefore promote the development of the Materials Genomics Initiative (MGI) and provide insights for future perspectives. There are a few existing review papers associated with ML for the development of advanced materials [33–35]. For example, Liu et al. [36] gave a detailed review of how ML accelerates the discovery and design of materials. However, they did not include the latest developments. Gu et al. [37] focused on the applications of ML for renewable energy materials. However, they did not provide detailed information about the tutorial for ML technologies. Li et al. [38] showed how AI strategies are applied at different stages of the development of materials, while case studies mentioned in this paper are less concentrated on energy materials. Chen et al. [39] provided an overview of ML techniques and their applications in materials research. However, the future prospective of data-driven materials science should be further expanded. Correa-Baena et al. [40] summarized state-of-the-art attempts via automation and ML for the discovery of materials from the perspective of theory, policy and investment. However, other important advances were not included. Moreover, the provision of design rules for the development of energy materials and the synthesis of materials predicted by ML have not received enough attention.

In this paper, we provide a comprehensive review of the recent progress and development in data-driven materials science and engineering, indicating the current research status and future perspectives on the fundamentals and applications of ML for the development of advanced energy materials. First, the roadmap to carbon neutrality is presented to reveal the importance and necessity of developing new energy materials. Second, a comprehensive introduction of fundamental ML tutorials is provided, including open-source databases, feature engineering, detailed introduction of typical ML algorithms, and effectiveness analysis of ML model. Third, the latest progress in data-driven materials science is introduced and discussed using real case studies on alkaline ion battery materials, photovoltaic materials, catalytic materials, and carbon dioxide capture materials. Moreover, relevant means towards successful ML applications and its remaining challenges are highlighted for each of these energy materials. Furthermore, general perspectives for future data-driven materials science are discussed, such as data infrastructures (data scarcity and standardization), ML techniques (automatic closed-loop optimization framework and visualization of black box models), experimental exploration (self-driving laboratory by robots), interdisciplinary communication and supporting policies.

2. Discovery of energy materials

2.1. Roadmap to carbon neutrality

For the purpose of achieving carbon neutrality, reducing CO_2 emissions has become a consensus worldwide. As shown in **Fig. 1**, the CO_2 emissions of global fossil from 1970 to 2019 [41], indicates that the CO_2 emissions of the power industry and transportation sectors exceeded more than half of the total CO_2 emissions. Therefore, effective measures are needed to be taken to reduce CO_2 emissions in the power industry as well as the transportation sectors. With this regard, governments and institutions have put forward many supporting policies to expedite the development of renewable energy and achieve zero-emission transportation [42–44], including the European BATTERY 2030+ [45], China's 13th Five-Year Plan for Renewable Energy [46], and the Paris Climate Agreement [47]. To clearly show the energy application scenarios of a fossil-free society in the future, **Fig. 2** illustrates the roadmap to achieve carbon neutrality which includes power generation, energy storage and conversion, and energy utilizations.

In terms of power generation, the most ideal source is renewable energy which is gleaned from natural sources such as water, sunlight, wind, and biomass [48,49]. To realize carbon neutrality, the most feasible way is to develop and apply renewable energy to replace fossil fuels on a large scale [50]. Current clean energy resources with large-scale applications potential include solar energy, wind energy, hydro energy, and nuclear energy [51]. In China, for example, to achieve carbon neutrality by 2060, it is necessary to deploy negative emissions technologies and utilize clean energy at very large scales [52]. Therefore, the development of renewable energy technology has significant impacts on the realization of carbon neutral society. Based on the above analysis, renewable energy will be the foundation of future energy development. However, renewable energy sources are susceptible to the influence of natural environment, for example, at night or under cloudy weather conditions, there will be less available solar energy due to the reduction or absence of sunlight, thereby supplying intermittent energy output. In this case, it is difficult to guarantee the maximum utilization of these energy sources when directly connected to the grid for clean electricity. Thus, exploring new technology of energy storage and conversion becomes necessary [53,54]. The typical energy storage technologies include compressed air, pumped hydro power, and flywheel, etc. During the last decade, advanced energy conversion and storage technologies, such as super capacitors, rechargeable batteries, flow batteries, and fuel cells, etc., have emerged and received rapid development [55–58]. Recently, electrochemical energy storage technologies, represented by hydrogen energy, have attracted widespread attention [59,60]. This technology converts clean electric energy into gaseous or liquid fuel, which is convenient for storage and transportation. In addition, by combining with CO_2 or N_2 in the air, hydrogen energy can be converted into energy-dense carbon-neutral liquid fuels (such as methanol and ammonia) [61–63]. In this way, the energy harvested from renewable energy sources can be converted and stored to provide unlimited green power for energy-consuming terminals such as buildings, transportation, and industries. However, the current energy conversion and storage technologies cannot meet the future energy demand. One of the best promising perspectives to address the above challenges is to develop advanced energy materials, which can greatly improve the efficiency of energy conversion as well as promote the large-scale applications of novel energy technologies. In summary, developing new energy materials with high-performance is necessary.

2.2. Development technologies

Traditional energy materials development methods include experimental analysis, theoretical calculation and simulation [40]. As shown in **Fig. 3**, the process of materials development can be accelerated by combining experiments and calculations, such as DFT calculation. How-

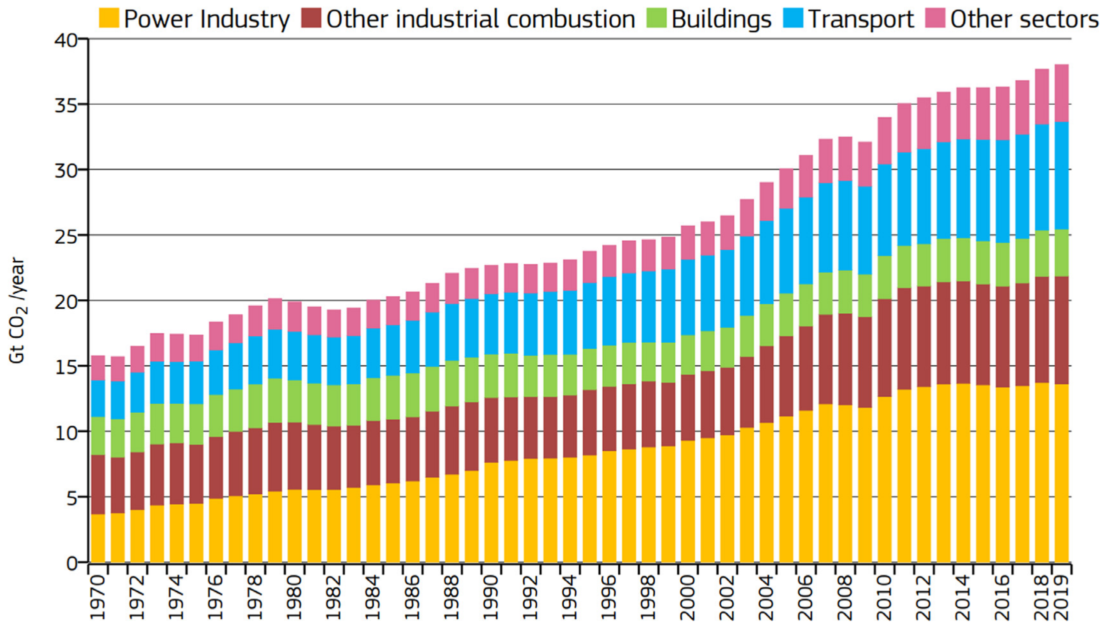


Fig. 1. Global fossil CO_2 emissions from 1970 until 2019. Reproduced from fossil CO_2 and GHG emissions of all world countries 2020 Report [41].

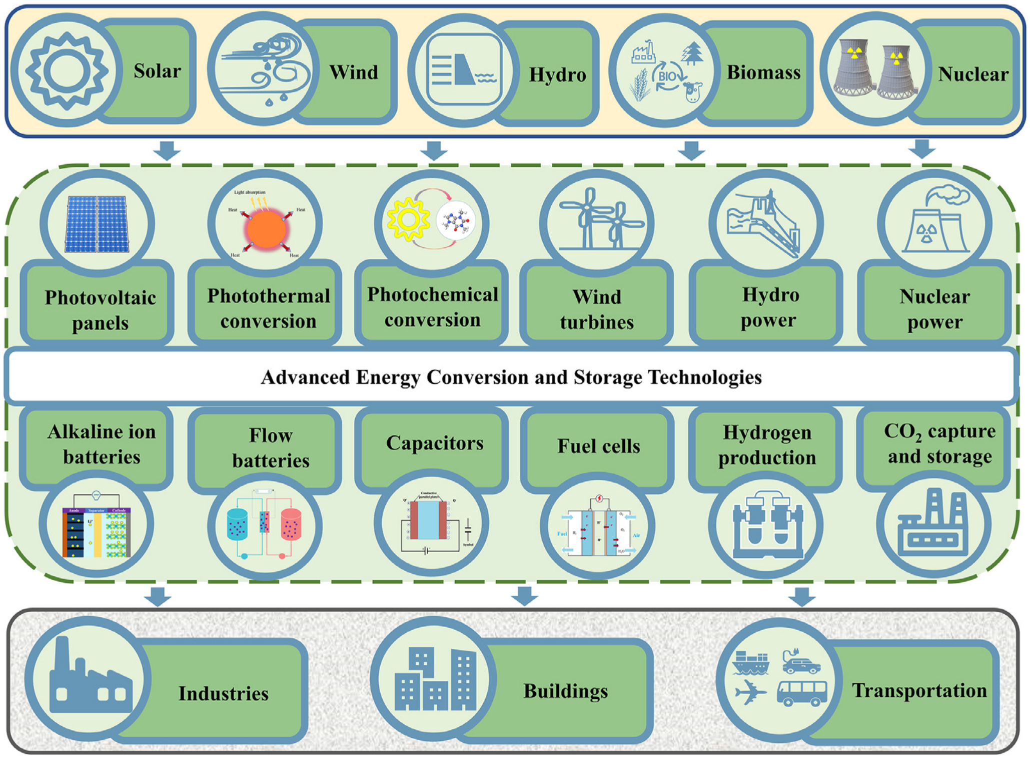


Fig. 2. The roadmap of future energy.

ever, DFT calculation has its drawbacks such as high time consumption and computational cost [18]. With the introduction and continuous advancement of AI and big data techniques, the application of ML in screening high-performance energy materials has been extensively studied [64,65]. By using data obtained from experiments or DFT calculations, a database can be developed. Based on the database and selected features, ML algorithms can implement large-scale data modeling, classification, and optimization. As a result, the promising candidates will be screened out. In addition, ML algorithms can be used to predict the macro and micro properties of energy materials. Moreover, feature engineering can be carried out for the purpose of determining the importance of different descriptors, thus providing effective guidance for the next round of modeling or classification [66]. In summary, the application of ML technology can greatly facilitate the development of advanced energy materials.

3. Machine learning tutorials

ML is a subset of algorithms in AI, which attempts to discover and infer the hidden laws in accordance with the historical data and then pre-

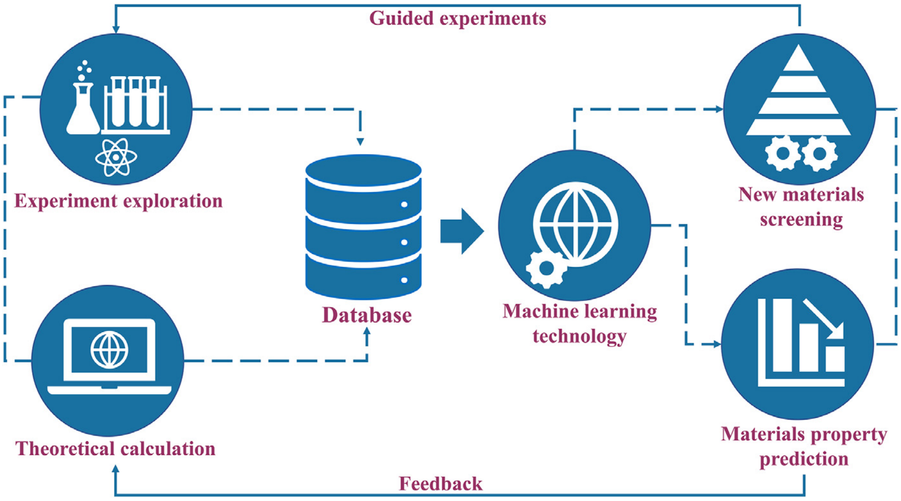


Fig. 3. Traditional and high-throughput development methods of energy materials.

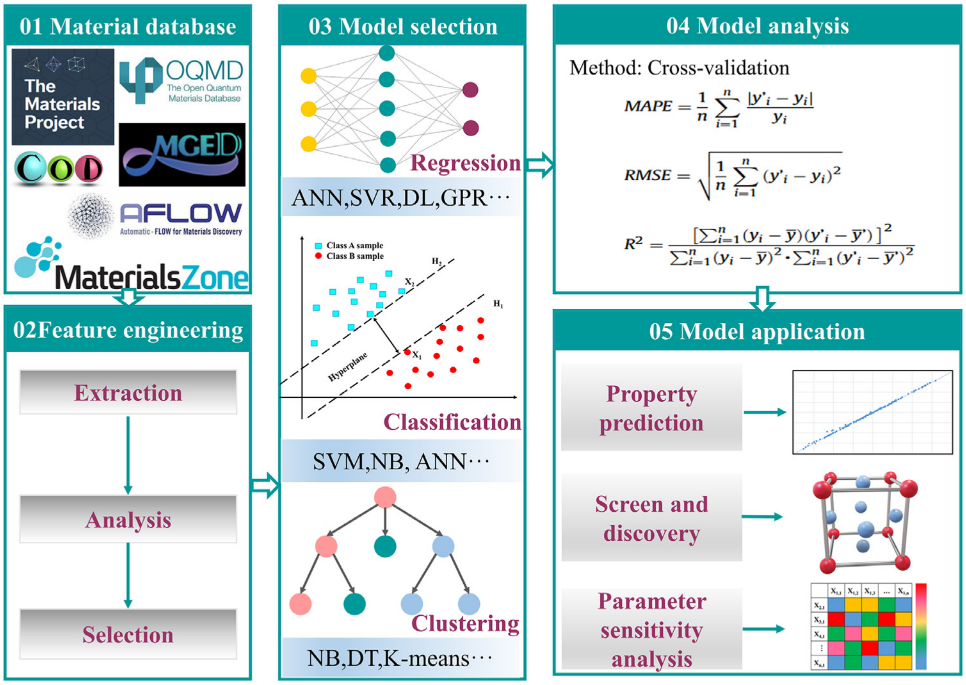

01 Material database
The Materials Project
OQMD The Open Quantum Materials Database
COD
MGED
AFLOW Automatic - FLOW for Materials Discovery
MaterialsZone
02Feature engineering
Extraction
Analysis
Selection
03 Model selection
Regression
ANN,SVR,DL,GPR···
Class A sample
Class B sample
H₂
H₁
X₂
X₁
Hyperplane
Classification
SVM,NB, ANN···
Clustering
NB,DT,K-means···
04 Model analysis
Method: Cross-validation
MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y'_i - y_i|}{y_i}
RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y'_i - y_i)^2}
R^2 = \frac{[\sum_{i=1}^{n}(y_i - \bar{y})(y'_i - \bar{y}')]^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2 \cdot \sum_{i=1}^{n}(y'_i - \bar{y}')^2}
05 Model application
Property prediction
Screen and discovery
Parameter sensitivity analysis

Fig. 4. General application procedure of ML technology in materials development.

dict or classify unlabeled data sets [67]. With the development of data technology, ML has been successfully applied in many fields. In the last two decades, the application of ML technology in screening advanced energy materials has gradually become a research focus, accelerating the discovery of new energy materials [68,69]. Fig. 4 shows the typical ML application process for energy material design and discovery, including ML database construction, feature engineering, ML algorithm selection, and ML model application. The details of ML application process will be illustrated in the following sections.

3.1. Database construction

Database plays significant roles during the application of ML for energy material development since the quality of modeling data determines the accuracy of ML model. According to the development history of material databases, most of the existing databases were developed in the past two decades. In 2006, materials scientist Ceder initiated a research project called "Materials Genome Project" at the Massachusetts Institute of Technology, which began to apply AI algorithms to the prediction and data collection of lithium-ion battery materials [70]. Four years later, approximately 20,000 forecast materials were included in the project. In 2011, since the American government launched the same name project, the Materials Genome Project therefore changed to the famous Material Project [71]. During the same period, Curtarolo, a former member of the Ceder team, established a new materials genomics centre at Duke University [72] and created another well-known material database, namely, AFLOW. Afterwards, many material databases were developed all over the world. For example, Chris Wolverton created the Open Quantum Materials Database (OQMD) in 2013, which focuses on inorganic crystal structures based on DFT calculations and includes around 400,000 hypothetical materials [73]. EPFL director Marzari developed a database named Materials Cloud, which focuses on the seamless sharing and dissemination of resources in data-driven materials science and engineering [74]. In China, in accordance with the national research and development (R&D) plan of 13th “Five-year Plan”, 40 projects related to the MGI were funded to promote the development

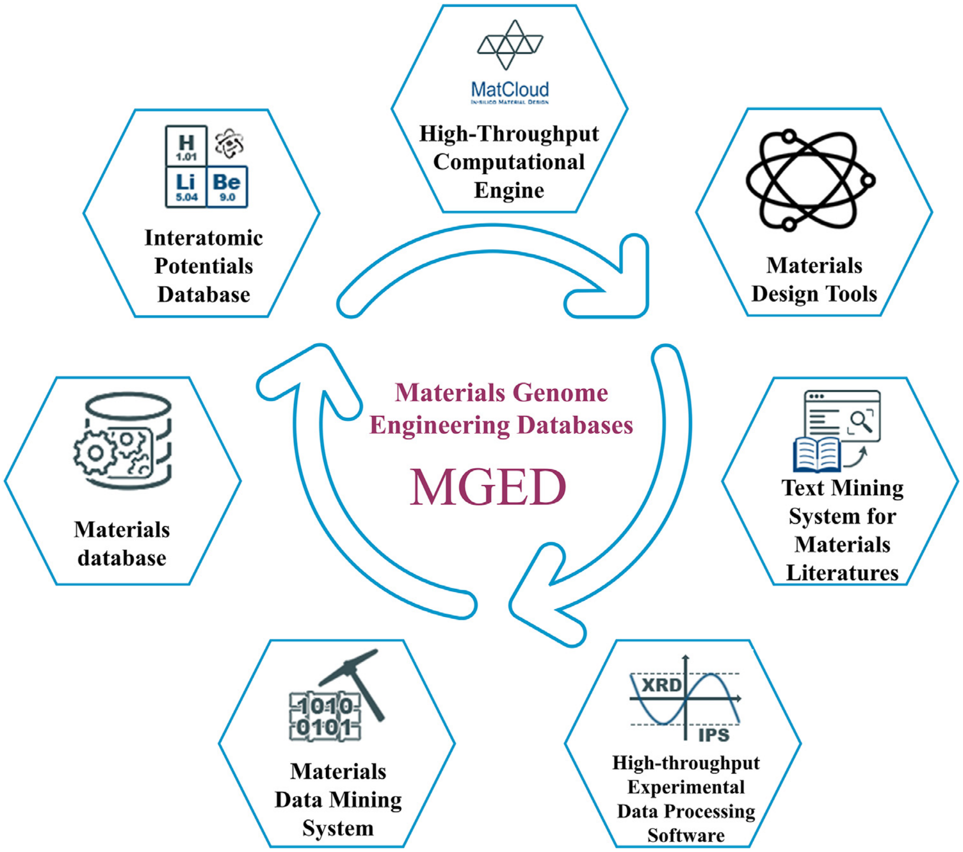

MatCloud
High-Throughput Computational Engine
Interatomic Potentials Database
Materials Design Tools
Materials Genome Engineering Databases
MGED
Materials database
Text Mining System for Materials Literatures
Materials Data Mining System
High-throughput Experimental Data Processing Software
XRD
IPS

Fig. 5. Function modules of the Materials Genome Engineering Databases (MGED). Adopted from the website of Materials Genome Engineering Databases [77].

of high-throughput material genomics [75]. Then, China established the Materials Genome Engineering Databases (MGED) in 2018, which integrates seven different functional modules (see **Fig. 5**).

Table 1 lists the newly developed material database. Most material databases are developed based on the data from experiments, scientific publications, and computer calculations. However, the data reported in the literature usually include only the results of successful experiments, and the grey data or failed data in the experiments are generally deliberately hidden. In order to effectively use the failed data, a database named Dark Reactions Project was constructed by Harvard University, which collected information on unpublished failed reactions [76]. With the continuous development of MGI, it is foreseeable that more online open-source materials databases will be established to accelerate the development of advanced energy materials.

3.2. Feature engineering

Feature selection plays a key role in data-driven materials science [78]. The reason is that, for a specific energy material, the modeling feature not only considers the structural parameters of the material, but also includes performance characteristics. In order to achieve accurate modeling of the characteristics of energy materials, it is necessary to select appropriate features [79]. Feature selection normally includes four stages (see **Fig. 6**), which are feature extraction, feature analysis, correlation and importance analysis, and feature selection [80]. As shown in **Fig. 6a**, the objective of feature extraction is to transform the materials space into descriptors space, i.e. input variables $X_{i,j}$. Based on the specific application scenarios, the number of $X_{i,j}$ is different. However, the complexity of feature selection and computational load will increase as the number of independent variables increases. In the process of applying ML to the development of energy materials, most of the existing feature extraction technologies rely on human decision-making [81]. Feature analysis is the key step after feature extraction. The main goal is to determine the importance and correlation of the extracted features. **Fig. 6b** shows the frequently used feature analysis techniques, including four typical methods: filter method, embedded method, wrapper method, and deep learning method [82]. The filter method can determine the impacts of input variables on output [79]. As a result, the importance of each $X_{i,j}$ can be calculated. Commonly used filter methods include Pearson correlation (PC) and correlation based feature selection (CFS) [83]. The embedded method can simultaneously perform feature selection and construct feature classifiers, thereby achieving higher efficiency. Typical embedded methods include least absolute shrinkage and selection operator (LASSO), random forest permutation accuracy importance (RFPAI), and least absolute shrinkage (LAS), etc. [84]. Another frequently used feature analysis technique is wrapper method. Wrapper method generally uses classification error or accuracy rate as the feature evaluation standard [85]. Wrapper method can analyse the correlation between different characteristics. Evolutionary algorithms, such as the genetic algorithm (GA) and particle swarm optimization (PSO), are often used to optimize wrapper model for subsets selection. Recently, with the rapid development of AI, deep learning neural networks have achieved great success in many fields because they are widely used in nonlinear problems and complex system modeling [86]. After the feature analysis process, the correlation and importance of selected features can be obtained through visible mapping, as shown in **Fig. 6c**. Then, according to the specific requirements of the application scenario, various feature subsets can be obtained for further research (see **Fig. 6d**). In short, feature engineering is a complex problem and directly affects the accuracy of ML models. Therefore, it is necessary to carry out feature engineering in data-driven materials science.

4. Machine learning algorithms

ML is a branch of AI that leverages algorithms and models to learn and infer from past data as well as existing knowledge [87–89]. ML algo-

Table 1.
ML database for energy materials.

Database	Descriptions	URL
Material Genome Engineering Databases (MGED)	Integrated seven functional modules, including materials database, materials design tools, data processing software, and text mining system, etc.	https://www.mgedata.cn/
Materials Scientific Data Sharing Network (MSDSN)	An online website of database (experimental and calculation data), data mining, material design, application cases, and metadata, etc.	http://www.materdata.cn/
The NIMS Materials Database (MatNavi)	MatNavi includes polymer materials, inorganic materials, metal materials and computing electronic structure information.	https://mits.nims.go.jp/en/
AFLOW	The AFLOW contains about 3 million material compounds, and the calculated properties exceed 560 million.	http://afowlib.org
American Mineralogist Crystal Structure Database	A crystal structure database containing data from different mineral journals.	http://rruff.geo.arizona.edu/AMS/amcsd.php
ChemSpider	The structural database of the Royal Society of Chemistry, containing experimental data and calculation data.	http://www.chemspider.com/
Citrination	AI-driven material data platform, containing about 4 million data sets.	https://citrination.com/
Computational Materials Repository	An integrated platform shows examples of how to use Python and the atomic simulation environment to process data.	https://cmr.fysik.dtu.dk/
Crystallography Open Database	Open-source database of inorganic, organic, mineral crystal structure, and metal organic compound.	http://www.crystallography.net/cod/
NCCR MARVEL	A materials informatics platform focusing on energy materials and organic crystals.	https://nccr-marvel.ch/
Materials Cloud	A systematic platform for computational materials, including databases, tools, and software, etc.	https://www.materialscloud.org/home
Materials Platform for Data Science (MPDS)	The MPDS presents the materials data, extracted by the project PAULING FILE team from the scientific publications.	https://mpds.io/#start
Materials Project	The Materials Project provides different material genome databases and high-throughput data analysis tools and software.	https://materialsproject.org/
National Renewable Energy Laboratory (NREL) Materials Database	The NREL database focuses on renewable energy materials, such as solar cell materials and thermoelectric materials.	https://materials.nrel.gov/
NIST Materials Data Repository	Establish data exchange protocols and mechanisms to promote data sharing and reuse.	https://materialsdata.nist.gov/
NOMAD CoE	Focuses on the systematic research and prediction of new materials to solve the energy and environmental challenges facing the future society.	https://www.nomad-coe.eu/
Open Quantum Materials Database (OQMD)	The OQMD is a database of thermodynamic and structural properties of 815654 materials calculated by DFT.	http://oqmd.org/
SUNCAT	Center for interface science and catalysis	http://suncat.stanford.edu/

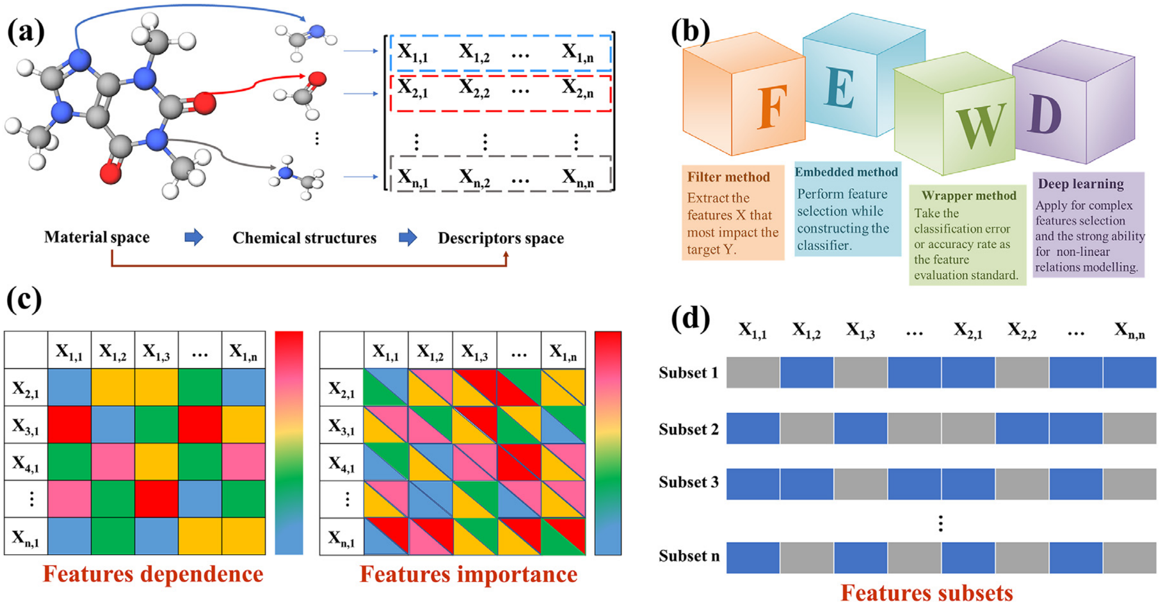

(a)
$X_{1,1}$ $X_{1,2}$ … $X_{1,n}$
$X_{2,1}$ $X_{2,2}$ … $X_{2,n}$
$X_{n,1}$ $X_{n,2}$ … $X_{n,n}$
Material space
Chemical structures
Descriptors space
(b)
F
E
W
D
Filter method
Extract the features X that most impact the target Y.
Embedded method
Perform feature selection while constructing the classifier.
Wrapper method
Take the classification error or accuracy rate as the feature evaluation standard.
Deep learning
Apply for complex features selection and the strong ability for non-linear relations modelling.
(c)
$X_{1,1}$ $X_{1,2}$ $X_{1,3}$ … $X_{1,n}$
$X_{2,1}$
$X_{3,1}$
$X_{4,1}$
$X_{n,1}$
Features dependence
Features importance
(d)
$X_{1,1}$ $X_{1,2}$ $X_{1,3}$ … $X_{2,1}$ $X_{2,2}$ … $X_{n,n}$
Subset 1
Subset 2
Subset 3
Subset n
Features subsets

Fig. 6. Feature engineering for ML applications: (a) Feature extraction process. (b) Typical ML feature analysis methods. (c) Correlation and importance analysis of selected features. (d) Various feature subsets obtained from feature engineering analysis.

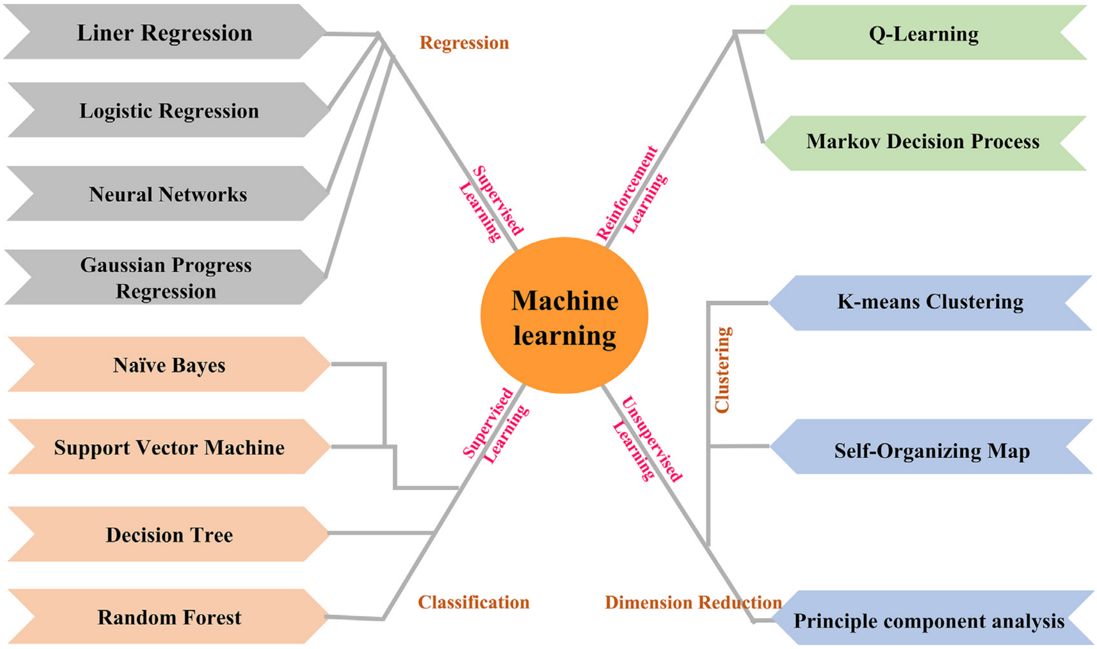


Fig. 7. Typical ML algorithms.

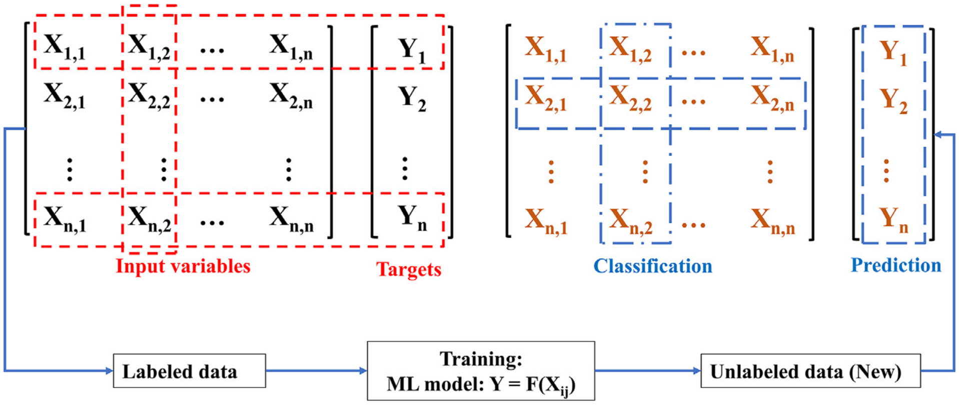

$X_{1,1}$ $X_{1,2}$ … $X_{1,n}$
$X_{2,1}$ $X_{2,2}$ … $X_{2,n}$
$X_{n,1}$ $X_{n,2}$ … $X_{n,n}$
Y_1 Y_2 Y_n
Input variables
Targets
$X_{1,1}$ $X_{1,2}$ … $X_{1,n}$
$X_{2,1}$ $X_{2,2}$ … $X_{2,n}$
$X_{n,1}$ $X_{n,2}$ … $X_{n,n}$
Classification
Y_1 Y_2 Y_n
Prediction
Labeled data
Training:
ML model: $Y = F(X_{ij})$
Unlabeled data (New)

Fig. 8. The application process of supervised learning algorithms.

rithms generally include supervised learning algorithms, unsupervised learning algorithms and reinforcement learning algorithms (see Fig. 7). In supervised learning algorithms, there are two types of ML models: regression model and classification model [90–92], such as logistic regression and neural networks. For unsupervised learning algorithms, it is mainly used for clustering and dimensionality reduction, such as K-nearest neighbors and principle components analysis [93]. In addition, reinforcement learning is also a significant part of ML, which can learn in an interactive environment through trial and error based on feedback. The commonly used reinforcement learning algorithms include Q-learning and Markov decision process [94]. The next section will introduce a detailed tutorial for each algorithm.

4.1. Supervised learning algorithms

The supervised learning algorithm is a typical ML method, which constructs a ML model by training labeled historical data (see Fig. 8) [95]. For each input variable X, there is always a corresponding target output Y (Y can be a specific data value or classification label). In other words, the expected output corresponding to the input variable is known. Generally, ML models constructed through supervised learning include regression models and classification models. After the trained ML model is obtained, it can be used for classification or prediction of unlabeled data (new data). Commonly used supervised learning algorithms are explained in subsequent sections.

4.2. Regression

4.2.1. Linear regression

Linear regression is one of the most famous and easily understood algorithms in statistics and ML. It is a linear approach which assumes a linear relationship between the input variable and the output variable. The basic formula for linear regression model is as follow [96]:

$$y = \varepsilon + \omega x \tag{1}$$

If the number of independent variables is greater than one, then Eq. 1 would change to the following format, which is known as multiple linear regression:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = X_i^T \beta + \varepsilon_i, i = 1, \ldots, n, \tag{2}$$

The typical linear regression method is ordinary least squares [97]. Advantages of linear regression include simple implementation, suitability for linearly separable data sets, and overfitting avoidance through regulation. While the disadvantages of linear regression include its vulnerability to under fitting as well as sensitivity to outliers. In data-driven materials science, linear regression is commonly used to predict and screen candidate materials to ascertain ideal properties [98]. In addition, for small data sets derived from material experiments, linear regression algorithms can be applied to provide fast and accurate prediction results.

4.2.2. Logistic regression

Logistic regression, unlike linear regression which outputs continuous values, is a typical classification algorithm that uses a logistic sigmoid function to transform its output into two discrete classes labeled 0 or 1. Thus, the efficient classification can be realized. However, if only linear regression is applied, and the estimated value of some data points based on the linear regression model may be greater than 1 or less than 0, the classification will be challenged. Therefore, logistic regression can be regarded as the promotion of linear regression model on classification problems. The logistic function is defined as follows [99]:

$$\text{logistic}(y) = \frac{1}{1 + \exp(-y)} \tag{3}$$

During the step from linear regression to logistic regression, the y can be considered as the linear regression model (eq. 2). Then the logistic function transfer to:

$$P(y(i) = 1) = \frac{1}{1 + \exp\left(-\left(\beta_0 + \beta_1 x_1^{(i)} + \cdots + \beta_p x_p^{(i)}\right)\right)} \tag{4}$$

Commonly used logistic regression methods include ordinal logistic regression (OLR), binary logistic regression (BLR), and multi logistic regression (MLR) [100]. The advantage of logistic regression model is that it is not only useful for classification model, but also applicable to probability model. However, it is difficult to capture complex relationships. The typical applications of logistic regression in high-throughput computational screening are to search for energy materials with high-performance [101]. An obvious advantage of using logistic regression is that it can quickly identify potential and suitable candidates from unknown materials, thereby reducing computational cost and time.

4.2.3. Gaussian process regression

Gaussian process regression (GPR) is a typical non-parametric model (i.e. not limited by a specific functional form) that applies Gaussian process prior to perform data regression analysis [102]. GPR can infer and discover the complex relationship between independent variables and dependent variables by theoretically using unlimited parameters and utilizing data to determine complexity level. In addition, GPR is a probabilistic model with versatility and resolvability [103]. The general application process of GPR is as follows: For the data used for GPR (take linear regression Eq. 1 as an example), according to the observed data, the prior distribution $p(\omega)$ can be first calculated and afterwards relocate probabilities based on Bayes' rule [104]:

$$p(\omega|y, X) = \frac{p(y|X, \omega)p(\omega)}{p(y|X)} \tag{5}$$

Thereafter, the updated distribution $p(\omega|y, X)$, i.e., the posterior distribution can be obtained. For the purpose of obtaining predictions at unknown points of interest, x*, the predictive distribution of data can be obtained through weighting all possible predictions values in accordance with their calculated posterior distribution [105]:

$$p(f * |\mathrm{x} *, y, X) = \int_{\omega} p(f * |x *, \omega)p(\omega|y, X)d\omega \tag{6}$$

Then, the joint multivariate Gaussian distributed for training points and test points can be obtained:

$$\begin{bmatrix} y \\ f^* \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ \mu^* \end{bmatrix}, \begin{bmatrix} K & K^* \\ K^{*T} & K^{**} \end{bmatrix}\right) \tag{7}$$

Here μ represents mean value, K represents covariance matrix. $K^{**} = K\ (X^*,\ X^*)$; $K^* = K\ (X,\ X^*)$. Then $f^* \sim N\ (\mu',\ K')$. Therefore, $\mu' = K^T K^{-1} f$, $K' = K^{**} - K^* K^{-1} K^{*T}$. After the f* has been confirmed, the predicted value for test input data can also be confirmed. Generally, GPR can be used to predict the performance of various energy materials, especially for materials with complex structures, interfaces, and compositions such as lithium-ion batteries and solar cells [106]. In addition, the predicted candidates can be screened out in accordance with the calculated probabilities.

4.2.4. Neural networks

Neural network is a mathematical or computational model that imitates the structure and function of biological neural network [107,108]. Through learning from examples, it can complete complex nonlinear modeling tasks and predictions [109]. One of the most commonly used neural network algorithms is artificial neural network (ANN) [110]. The general algorithm structure of ANN is shown in **Fig. 9a**, which includes three layers, namely the input layer, the hidden layer, and the output layer, respectively. The input layer is the independent variable x, which can be set to different variables and numbers according to the feature engineering and the expertise of a specific application. The hidden layer performs a nonlinear transformation on the input to the network, where the function applies weights to the input and directs it to the output through the activation function. The internal structure of the hidden layer changes according to the function of the neural network. The output layer consists of the dependent variable y, which is also the supervision target of the ANN. The ANN can learn, summarize and induce to produce an automatic recognition system. The advantages of ANN include fault tolerance, parallel processing capability, and strong nonlinear fitting ability. Compared with the single hidden layer of ANN, a deep learning neural network with multiple hidden layers (**Fig. 9b**) can accurately model complex nonlinear systems. Commonly used deep learning algorithm includes convolutional neural network (**Fig. 9c**), recurrent neural network (**Fig. 9d**), long short-term memory network, etc. [111]. Neural networks are normally applied to address complex modeling problems due to the strong capability of capturing complex nonlinear relationships. The typical applications of neural networks algorithm in data-driven materials science include prediction of materials properties and screening of promising candidates [112]. In particular, deep learning neural networks can not only be used to model the complex relationships between materials structure, composition, and performance, but also reveal the underlying mechanisms of various chemical reactions. Therefore, neural networks are often carried out to study materials with complex structures and multiphase reaction interfaces, such as battery materials and catalytic materials. In addition, neural networks can handle a large number of data samples, which will be a powerful tool when combining ML technology with DFT calculations to develop new energy materials.

4.3. Classification

4.3.1. Naïve Bayes

Naive Bayes (NB) classification method is one of the supervised learning algorithms based on Bayes' theorem. Given the value of class variables, it is assumed that the conditional independence between each pair of features is naïve. The Naïve Bayesian classifiers have a high degree of scalability, such that it requires the parameters to have a linear relationship with the features in the problem to be solved. With the application of NB, the maximum likelihood training can be achieved through evaluating closed-form expressions, which is more effective than iterative approximation clustering method. Although their assumptions are obviously oversimplified, the NB classifier can work well in many practical situations. For specific supervised classification problems, when the data is discrete, the mostly used formula based on Bayes' theorem is as follows [113]:

$$p(y|x_1, \ldots, x_n) = \frac{p(y)p(x_1, \ldots, x_n|y)}{p(x_1, \ldots, x_n)} \tag{8}$$

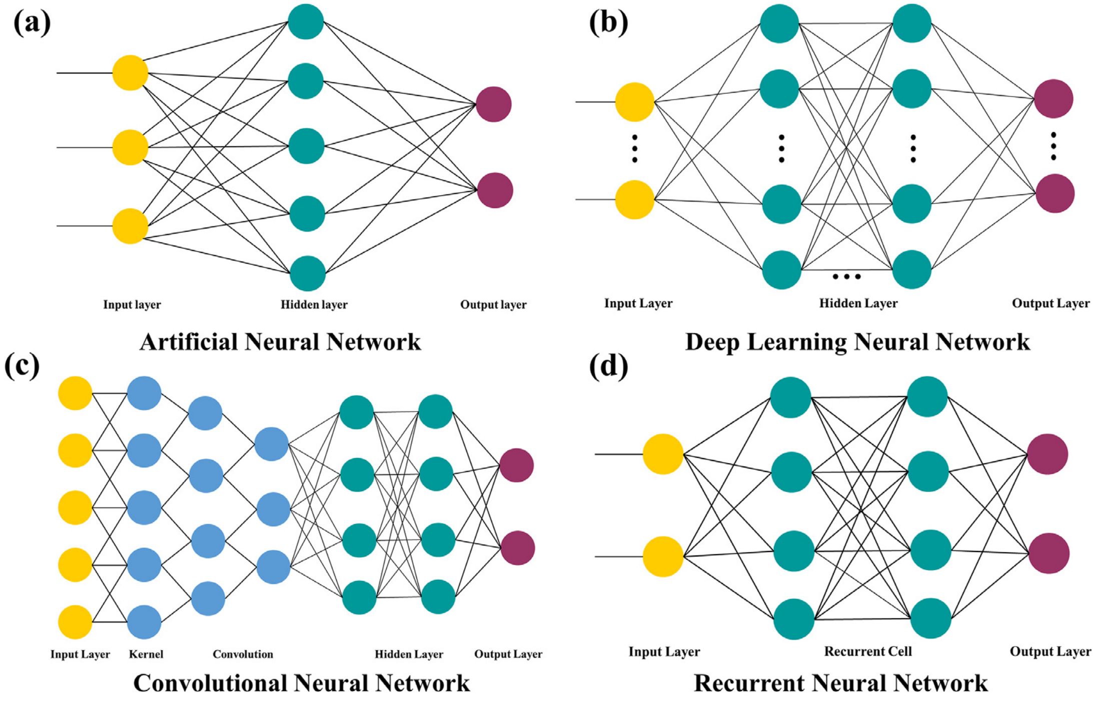


Fig. 9. Neural networks structures: (a) Artificial neural network. (b) Deep learning neural network. (c) Convolutional neural network. (d) Recurrent neural network.

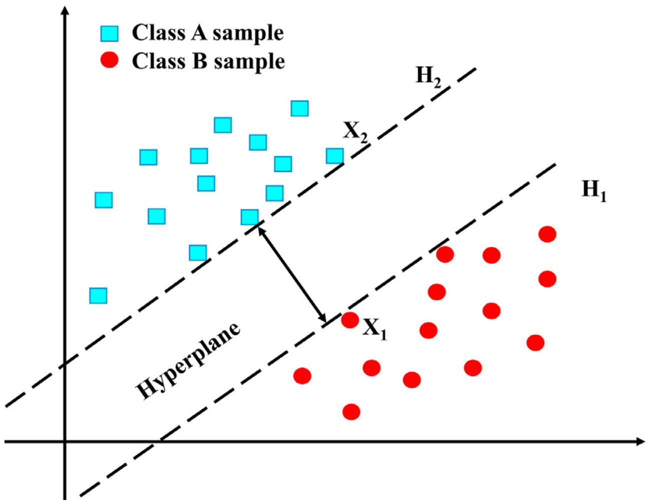


Fig. 10. Conceptual diagram of support vector machine classification.

Based on the naive assumptions between different features, it can be transformed into its final form as shown below:

$$\hat{y} = \arg\max_{y} P(y) \prod_{i=1}^{n} P\left(x_i | y\right) \tag{9}$$

As the data set becomes continuous, assumptions can be changed to adjust for clustering. The NB algorithm is recommended to be a classification tool for the development of energy materials. More importantly, the latest advancements in data-driven materials science suggest that the Bayes algorithms can be applied in the close-loop optimization to realize material design automation [114]. More relevant case studies can be found in Section 5.

4.3.2. *Support vector machine*

Support vector machine (SVM) is a supervised ML algorithm that can divide unlabeled data sets into two categories (See **Fig. 10**). The application process of SVM generally includes two stages: first, the SVM model can be constructed by training the labeled data set (i.e., the classification result of each data point is known). The trained SVM model thereafter becomes a non-probabilistic binary linear classifier [115]. Second, the trained SVM model can be used to classify the unlabeled data sets, which maps the new instance to the same space and predict its category according to the side of the interval the new instance falls on. Compared with other algorithms, the advantage of SVM includes fast classification process and higher classification accuracy in a limited number of samples. However, SVM cannot directly provide probability estimates. In addition to linear data set classification, SVM can also be used to classify non-linear disturbed data set through dimension changes based on kernel trick. With regard to the applications in materials research community, SVM can be applied to identify potential candidates through classification [116]. Another commonly used algorithm based on support vector is support vector regression (SVR), which can be used to predict materials properties as well. In addition, SVR can be used to model complex dynamic reactions, such as the migration characteristics of lithium ions in lithium batteries.

4.3.3. *Decision tree and random forest*

Decision tree (DT) is a typical and easy-to-understand ML algorithm, which represents the mapping relationship between different variables and can be used for prediction or classification [117]. The DT algorithm uses a tree structure and inference layer to achieve the final decision of modeling results. The application process of DT generally contains feature selection, generation and pruning of DT structure. The DT structure (see **Fig. 11a**) usually consists of three elements, namely the root node (all samples to be classified), internal nodes (feature attributes), and leaf nodes (decision-based classification). When applying DT for prediction, first use a certain attribute value to determine the internal routine of the tree (based on the if-then-else rule), and then determine the branch to enter and interrupt until the leaf is reached in accordance with the judgment result. Finally, the result of DT classification can be obtained. The main advantages of DT include simple algorithm structure, highly interpretable, easy implementation, etc. However, DT also has its disadvantages, such as easy overfitting, and unstable or complicated generated decision tree structure. Furthermore, a single DT model is easily affected by noisy data and is prone to over fitting. To solve this prob-

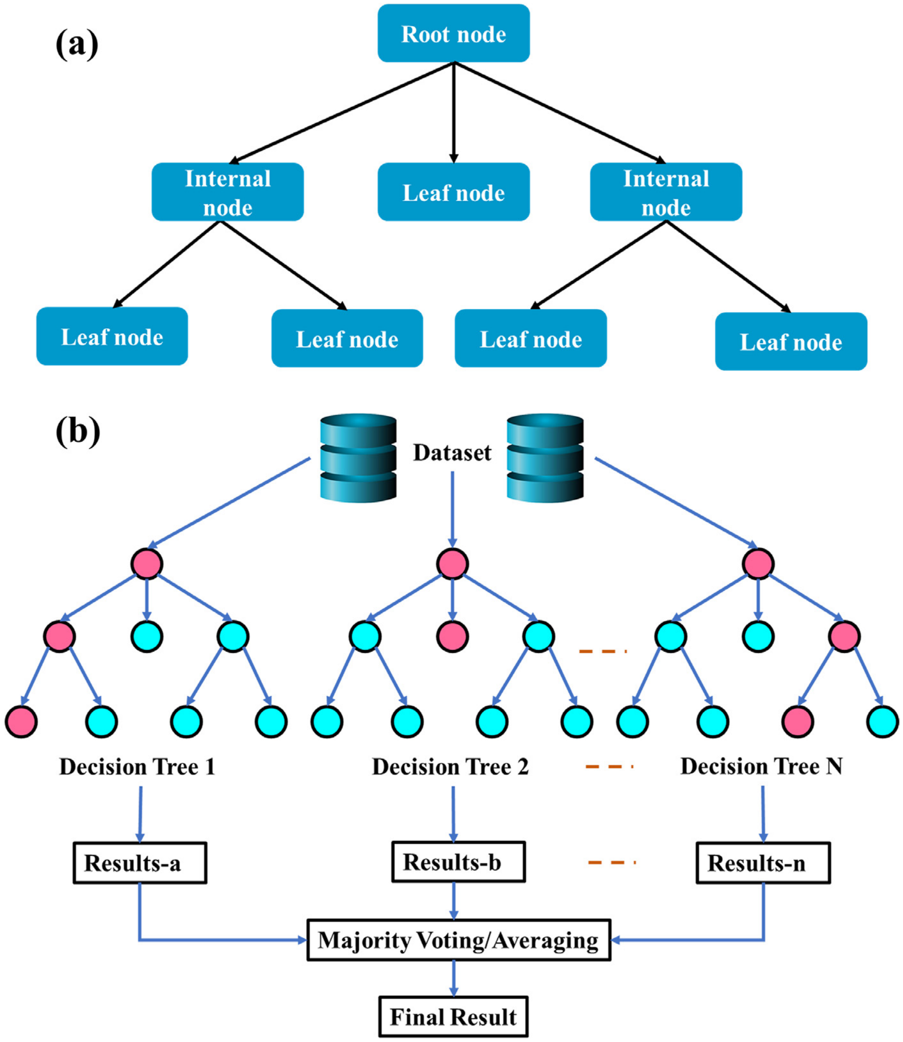


Fig. 11. The general structure of tree-based models: (a) Decision tree (b) Random forest.

lem, random forest (RF) is proposed. RF randomly selects samples and features based on DT, which is a typical integrated algorithm [118]. As shown in the **Fig. 11b**, unlike DT, RF samples several data subsets in the original training data set and performs model training on each subset. After the training is completed, the average output of all models will be taken, and then the best model will be selected based on the main vote. The typical applications of tree-based model in materials informatics include classification and regression, which can provide detailed and intuitive results with the specified tree structures. For example, based on the research of Shi et al. [119], RF was considered as an ideal ML algorithm for the development of CO_2 capture material. In addition, the commonly used tree-based algorithms in data-driven materials science include gradient-boosted regression trees (GBRT), gradient boosting decision tree (GBDT), etc.

4.4. Unsupervised learning and reinforcement learning

Unsupervised learning is a method that can automatically classify or group input data without marking training samples in advance. The main applications of unsupervised learning include cluster analysis and dimensionality reduction. It is an alternative to supervised learning and reinforcement learning strategies. K-means clustering is a commonly used unsupervised algorithm for clustering unlabeled data into different groups. The principle of k-means clustering algorithm is to classify the unlabeled data into k clusters, thereafter connect each data to its nearest cluster centre. In addition to clustering, unsupervised learning algorithms such as principal component analysis (PCA) and self-organizing mapping (SOM) can also be applied to achieve dimensionality reduction. For instance, PCA can be used to realize data set dimension reduction, improve the interpretability and minimize information loss [120]. SOM uses unsupervised learning to generate low-dimensional discrete representations of input variables. One of the prominent characteristics of SOM is that it uses a competitive learning mechanism instead of an error correction learning mechanism. Furthermore, there is no hidden layer in SOM [121]. It is worth to mention that reinforcement learning is a small branch of ML, which emphasizes how to take actions according to the environment to maximize the expected benefits. The commonly used reinforcement learning algorithms is Q-learning [122]. Q-learning is based on the record of the learning process, and then the information is expressed to the agent so that the maximum return will be obtained at a specific circumstance. The common application of unsupervised learning algorithms in MGI is to classify candidate materials into different subsets to achieve reasonable classification. Additionally, unsupervised learning algorithms can be combined with supervised learning algorithms to assist the development of energy materials. In summary, the advantages, disadvantages, and typical applications of each algorithm are summarized in Table 2.

4.5. Machine learning model analysis

After obtaining ML model, it is also necessary to evaluate the accuracy of ML model. In terms of the general procedures for the effective

Table 2.
Introduction of various ML algorithms.

Algorithms	Advantages	Disadvantages	Typical applications	Ref.
Linear regression	Simple implementation and easy to understand	Vulnerability to under fitting and sensitivity to outliers	Properties prediction and materials screening	[98]
Logistic regression	Useful for classification model and probability model	Difficult to capture complex relationships	Properties prediction and materials screening	[101]
Gaussian process regression	Simplified organization, fewer parameters, and clear probability formula	Large amount of calculations	Properties prediction	[106]
Neural networks	Powerful nonlinear modeling capabilities	The model is invisible and difficult to explain	Properties and structure prediction, as well as materials screening	[112]
Naïve Bayes	Stable classification efficiency, handles small-scale data, and the results are easy to interpret	Need to calculate the prior probability and sensitive to the expression of the input data	Optimization	[114]
Support vector machine	Fast classification process and higher classification accuracy	Cannot directly provide probability estimates	Materials screening and properties prediction	[116]
Decision tree	simple algorithm structure, highly interpretable, easy-to-implement	Easy overfitting, and unstable tree structure	Materials screening and properties prediction	[123]
Random forest	It can process high-dimensional data and directly gives the result of feature analysis	Sensitive to noise data, high computational cost	Materials screening and properties prediction	[124]
Principal component analysis	It makes the data set easier to use, reduces the calculation cost and has no parameter limitation	There is a small amount of information loss; information overlap cannot be effectively eliminated	Dimension reduction	[125]

analysis of the ML model, most of the data samples will be employed to train the ML model while a small part of data samples are reserved for testing and validation. The estimated prediction error of the trained ML model is then found and recorded. The common evaluation method of ML model is cross-validation. Cross-validation is mainly used in the process of modeling applications. The working principle of cross-validation is similar to the method mentioned above, which first divides the unlabeled data set into different subsets and then selects a certain number of subsets as the training data set while the remaining subsets will be used for validation. Commonly used validation methods include leave-one-out validation, K-fold cross-validation, and holdout validation. Typical error calculation formulas include root mean square error, variance, and average absolute error [126]:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{\left|y'_i - y_i\right|}{y_i} \tag{10}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y'_i - y_i\right)^2} \tag{11}$$

$$R^2 = \frac{\left[\sum_{i=1}^{n}\left(y^i - \bar{y}\right)\left(y'_i - \bar{y}\right)\right]^2}{\sum_{i=1}^{n}\left(y^i - \bar{y}\right)^2 \bullet \sum_{i=1}^{n}\left(y'_i - \bar{y}\right)^2} \tag{12}$$

According to the model analysis results, material prediction and discovery can be also realized through model selection. Regarding the accuracy of ML models, one point is worth noting. In the actual application of ML to develop energy materials, the established ML model should not only focus on the accuracy of model prediction, but also the effectiveness of the model in solving practical problems. In other words, the model accuracy may not necessarily be high. The reason is that other factors, such as stability, need to be considered comprehensively. For instance, Sutton et al. [127] provided a case study to investigate the applicability of ML models for developing materials. Although the accuracy of the established ML model is not satisfactory, the developed model is still applicable, i.e. it can be used to screen materials in a fixed compositional space. In addition, as mentioned before, combining ML algorithms with other techniques can further promote the development of data-driven materials science. For example, the data sources for ML modeling can be extracted from experimental data, DFT calculation, and resources collected from literature. Moreover, the development route of high-performance materials can be automated through integrated ML technology, optimized algorithms, and intelligent robots. The applications of ML in the development of energy materials will be introduced and discussed in the next section.

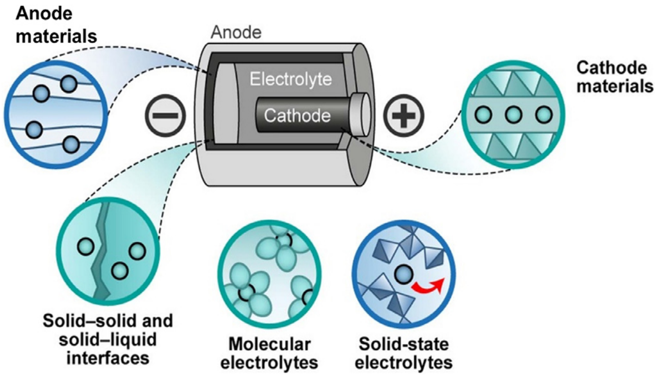


Fig. 12. Typical structure of lithium-ion batteries. Reproduced with permission from [135]. Copyright 2020, IOP Publishing Ltd.

5. Machine learning applications

Recently, the application of ML algorithms in the design and discovery of advanced energy materials has become a popular trend [128–130]. In this section, the recent advancements in data-driven materials science and engineering will be introduced and discussed, including alkaline ion battery materials, photovoltaic materials, catalytic materials, and carbon dioxide capture materials.

5.1. Alkaline ion batteries

Alkaline ion batteries have developed rapidly in the past few decades due to its high energy density and environmentally friendly features [131]. However, there are some challenges pertinent to this battery technology, such as safety issues and limited raw materials [132]. Developing advanced battery materials is therefore considered as one of the promising ways to address these challenges [133,134]. A typical structure of lithium-ion battery is shown in **Fig. 12**, consists of cathode, anode, and electrolyte [135]. The anode material generally can be made of carbon, graphite, or silicon while the cathode material is normally composed of lithium-containing metal oxide [136]. In terms of the electrolyte materials, the liquid and solid electrolytes in lithium-ion batteries are generally consisted of lithium salt, and lithium metal oxide, respectively. Recently, the applications of ML to screen high-performing lithium-ion battery materials have been extensively studied [137,138]. This section thus provides a brief summary on some of these latest investigations.

5.1.1. Electrolytes

Liquid electrolytes: in order to greatly improve the reliability and safety of lithium-ion batteries, it is urgent to develop new electrolyte systems [139]. Recently, the application of ML methods to find new electrolyte materials has received attention from researchers [140]. For instance, to measure the disordered characteristics of new electrolyte materials, Sodeyama et al. [98] proposed a multi-ML application framework on the basis of three different linear regression algorithms. The results show that exhaustive search of linear regression can provide the most accurate estimation of electrolyte liquid properties. In addition, the weight map of descriptors can also be analysed to identify the complex correlation between computational cost and prediction accuracy, thereby improving the search efficiency of a large number of new materials. During the practical application, the transfer process of various ions between the electrolyte and the electrodes is complicated. The coordination energy of ions and solvents can properly indicate the transfer of ions at the electrolyte and electrode interface. Therefore, research on coordination energy can supply effective guidance for the development of advanced electrolyte materials. Ishikawa et al. [141] later applied quantum chemical calculations to study the coordination energy of five alkali metal ions (Li, Na, K, Rb, and Cs) to electrolyte solvents. Validation results showed that the linear regression algorithms provide the highest prediction accuracy of coordination energy of 0.127 eV. In addition to the material of the electrolyte, the electrolyte additives also have significant impacts on the performance of lithium-ion batteries. For example, Yasuharu et al. [106] combined ab initio calculations and ML methods (Gaussian kernel ridge regression and gradient boosting regression) to model and analyse the redox potential of 149 electrolyte additives for lithium-ion batteries. Results show that the descriptors accurately predicted the redox potentials. Furthermore, the essential characteristics of the redox potential can be described by a small number of features derived from the analysis of feature engineering. To further speed up the scientific innovations in aqueous electrolytes for lithium-ion batteries, Dave et al. [142] developed an integrated platform that combines ML technology with intelligent robots, which can independently perform hundreds of sequential experiments to optimize battery electrolyte (see **Fig. 13)**. A database consisting 251 aqueous electrolytes was provided and a promising candidate of mixed-anion sodium electrolyte was identified.

Solid electrolytes: lithium-ion battery that uses solid electrolyte is considered as the future perspective due to its inherent safety and high energy density [143]. However, there are some critical challenges such as low conductivity and poor stability of the battery interfaces that remain unresolved, thereby hindering the development of solid-state lithium-ion batteries [144]. The recent advancements in battery technology demonstrate that the tsavorite structure can maintain a fast lithium-ion insertion rate for battery cathode applications. Following this motivation, Jalem et al. [112] therefore explored the $LiMTO_4F$ tsavorite system for solid electrolyte. The research objective is to identify potential components with very low lithium migration energy, and to explore the impact of structure parameters on migration energy. A crystal structure-based migration energy prediction model was therefore constructed through integrating ML technology (neural network) and DFT calculation. This study identified the key factors affecting migration energy, such as the covalent effect of polyanions and the competition between local lattices. By using logistic regression, Sendek et al. [101] proposed a new screening method to identify high-performance candidate materials for solid state electrolytes. This research determined 21 promising structures from 12,831 potential candidates. For the purpose of unraveling the composition–structure–ionic conductivity relationships, Kireeva et al. [116] applied the SVR method to analyse the lithium-ion migration characteristics of the garnet structure oxide. A reasonable level of predictive ability of the models was achieved. In order to screen advanced materials for solid-state Li-ion conductors, Zhang et al. [145] performed an unsupervised ML method to prioritize the candidate list from various Li-containing materials and discovered 16 new fast Li-conductors (refer to **Fig. 14a and b**). Similarly, by using a recommender system coupled with the random forest classification algorithm, Suzuki et al. [124] discovered two lithium-ion conductors for solid-state electrolyte battery which has never been reported before. Moreover, the synthesis time of newly found $Li_6Ge_2P_4O_{17}$ was 10 times less than the conventional conductor. To investigate the application potential of using non-flammable Li-conducting ceramics as solid electrolytes, Nakayama et al. [114] proposed two data-driven method for screening materials. The Bayesian optimization was applied to process data, thereby the searching efficiency was greatly improved. With the same ML algorithm, Wang et al. [146] developed an automated simulation optimization framework to design new solid polymer electrolytes. As shown in **Fig. 14c**, the materials design process of solid polymer electrolytes started from discrete conventional design space, which thereafter transferred to continue coarse graining design space by simulation and iterative exploration. Then the Bayesian optimization was applied to optimize the materials design output. In this way, the complex interactions between the conductivity of lithium and the intrinsic material properties of the molecule are determined. In summary, the application of ML technology can help to discover high-performing materials for lithium-ion batteries.

5.1.2. Electrodes

In the past few decades, the continuous development of lithium-ion battery electrode materials has laid a solid foundation for the successful commercialization of lithium-ion batteries. To speed up the development of battery electrode materials, the application of ML to explore new electrode materials has become a new research focus [147]. For example, Shandiz et al. [148] applied 8 different clustering algorithms to investigate the effects of crystal structure on the performance of battery electrode. Three typical crystal systems were studied and results showed that the highest prediction accuracy can be obtained with the application of RF model. Furthermore, the parameter sensitivity analysis results of classification model confirmed that the number of sites and the volume of crystal have a significant impact on determining the type of crystal system. For the purpose of investigating the most important parameters that affect the cathode volume of the battery, Wang et al. [149] reported a method combining ab initio calculation and partial least squares (PLS) analysis. The results of feature analysis confirmed that the X octahedron and the radius of X4þ ion are the determining factors. To accelerate the process of developing new materials of molecular electrode, a DFT-ML framework for developing a high-throughput screening was proposed by Allam et al. [150]. Both the electronic properties and structural information were selected as independent variables for the prediction of redox potentials (see **Fig. 15a**). Through the application of a linear correlation analysis, a large number of input variables were downsized to six core input variables (see **Fig. 15b**). Moreover, the results indicate that the most critical factor affecting the redox potential is electron affinity. Aiming to explore the mechanism of micro-structure design of lithium-ion battery electrode, Takagishi et al. [151] proposed a comprehensive framework using three-dimensional virtual structures and ML (refer to **Fig. 15c**). The results show that the electrode specific resistance predicted by the ANN model is in good agreement with the simulated value. In order to promote the discovery of materials for battery, Joshi et al. [152] developed a web accessible tool by integrating ML technology to predict the voltage of electrode materials in metal-ion batteries. The results show that the developed online tool can estimate the voltage of any bulk electrode material for multiple metal ions

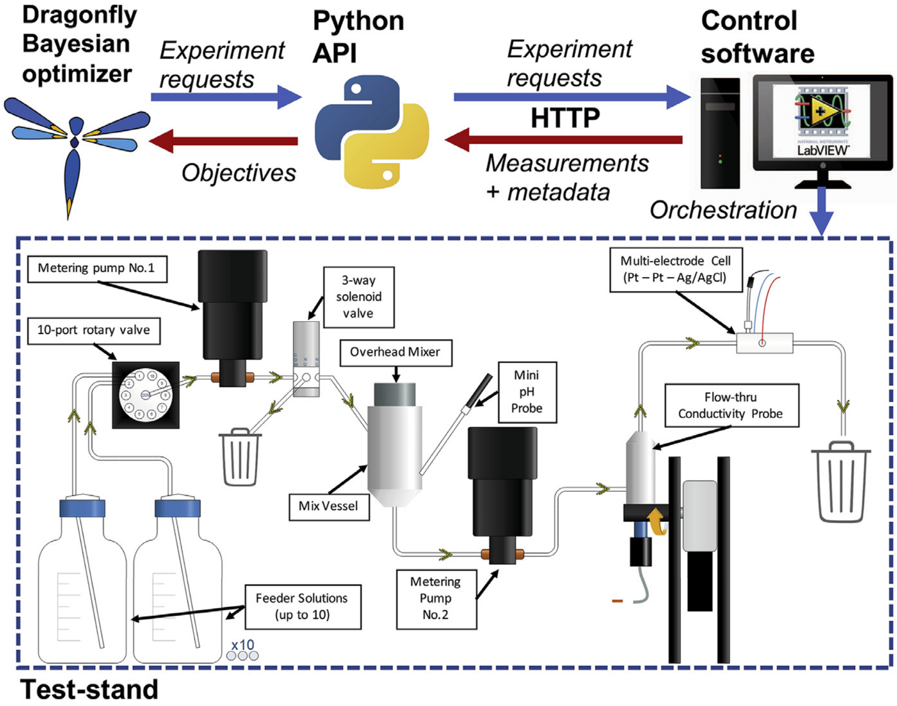


Fig. 13. Automated ML and robot integration platform for exploring liquid electrolytes. Reproduced with permission from [142]. Copyright 2020, Elsevier.

within one minute. Moreover, the online platform can be freely accessed at http://se.cmich.edu/batteries. For the purpose of studying the effect of the microstructure of the composite electrode on the charge and discharge performance of a single battery, Jiang et al. [153] proposed a comprehensive framework combining experimental exploration, convolutional neural networks, and mathematical modeling (**Fig. 15d**). The results showed that the conductivity is positively correlated with the degree of particle detachment. Furthermore, this study confirmed that balancing lithium-ion kinetics and electron diffusion have a significant impact on improving battery performance.

In terms of promoting the application of advanced battery technology, in addition to developing novel battery materials, other aspects such as the manufacturing and application are worth studying. With this consideration, Turetskyy et al. [154] applied a data-driven technology to establish a digital and intelligent battery manufacturing system and provided a successful case study. To investigate the impact of materials and battery design on the performance of lithium-sulfur (Li-S) batteries, Kilic et al. [155] developed a new ML method which coupled the association rule mining method and Apriori algorithm. Based on the data resources extracted from literatures, this study found that the type and quantity of encapsulation material play a vital role in increasing battery capacity and extending cycle life. Moreover, the latest technical progresses showed that ML can be used to predict the battery health status as well as sustainable life cycles [156]. In summary, the above case studies and analysis confirm that one of the most promising prospects in battery technology is the application of data-driven techniques such as ML and big data to accelerate the development of next-generation battery technologies.

5.2. Photovoltaic materials

Exploring materials with high conversion efficiency for solar battery is a prerequisite for the large-scale application of solar energy [157,158]. The exploration of applying ML algorithms to discover new solar materials with high performance has gradually become a future trend [159,160]. The typical application of ML technology for solar battery includes prediction of property and conversion efficiency, as well as the screening of new photovoltaic materials with high-performance [161–163].

5.2.1. Property prediction and screening

The screening of high-performance photovoltaic materials and the accurate prediction of the relationship between structure and properties are important pursuits for future solar cell research and application [164,165]. To predict the stability of perovskite structure, Sun et al. [166] applied a data screen method combined with a one-dimensional tolerance factor. Validation results illustrate that the proposed ML framework can accurately identify 92 % of the compounds in the data set of 576 ABX_3 materials. Based on the known crystal structure information of ABX_3 perovskite, Pilania et al. [167] established a classification model for prediction of new perovskite halides by using SVM algorithm. Results showed that several new ABX_3 compositions with perovskite crystal structure were discovered. In order to find lead-free perovskite materials for solar cells, Im et al. [168] applied GBRT method to predict the formation heat and band gap of candidate halide double perovskite. For the purpose of discovering advanced two-dimensional solar cell materials (see **Fig. 16a**), a data-driven screening framework was proposed by Jin et al. [169]. The searching diagram was shown in **Fig. 16b**, which integrated ML model and DFT validation to identify potential candidates from a large number of experimentally confirmed crystal structures. 26 two-dimensional photovoltaic materials were finally identified. In order to search stable and metastable perovskite materials, Liu et al. [123] developed a classification model in accordance with GBDT. 331 candidates (refer to **Fig. 16c**), which are predicted to have a perovskite structure, were screened out from 891 ABO_3.

5.2.2. Solar conversion efficiency

To accelerate the process of discovering the hybrid organic-inorganic perovskites for photovoltaics, Lu et al. [170] applied six ML algorithms combined with DFT calculations to screen solar battery materials (see **Fig. 16d**). Validation results showed that the gradient boosting regression algorithm provides the highest accuracy. Additionally, six orthorhombic lead-free hybrid organic-inorganic perovskites were discovered for the first time. Similarly, Schmidt et al. [171] constructed a data set containing DFT calculations of approximately 250,000 cubic per-

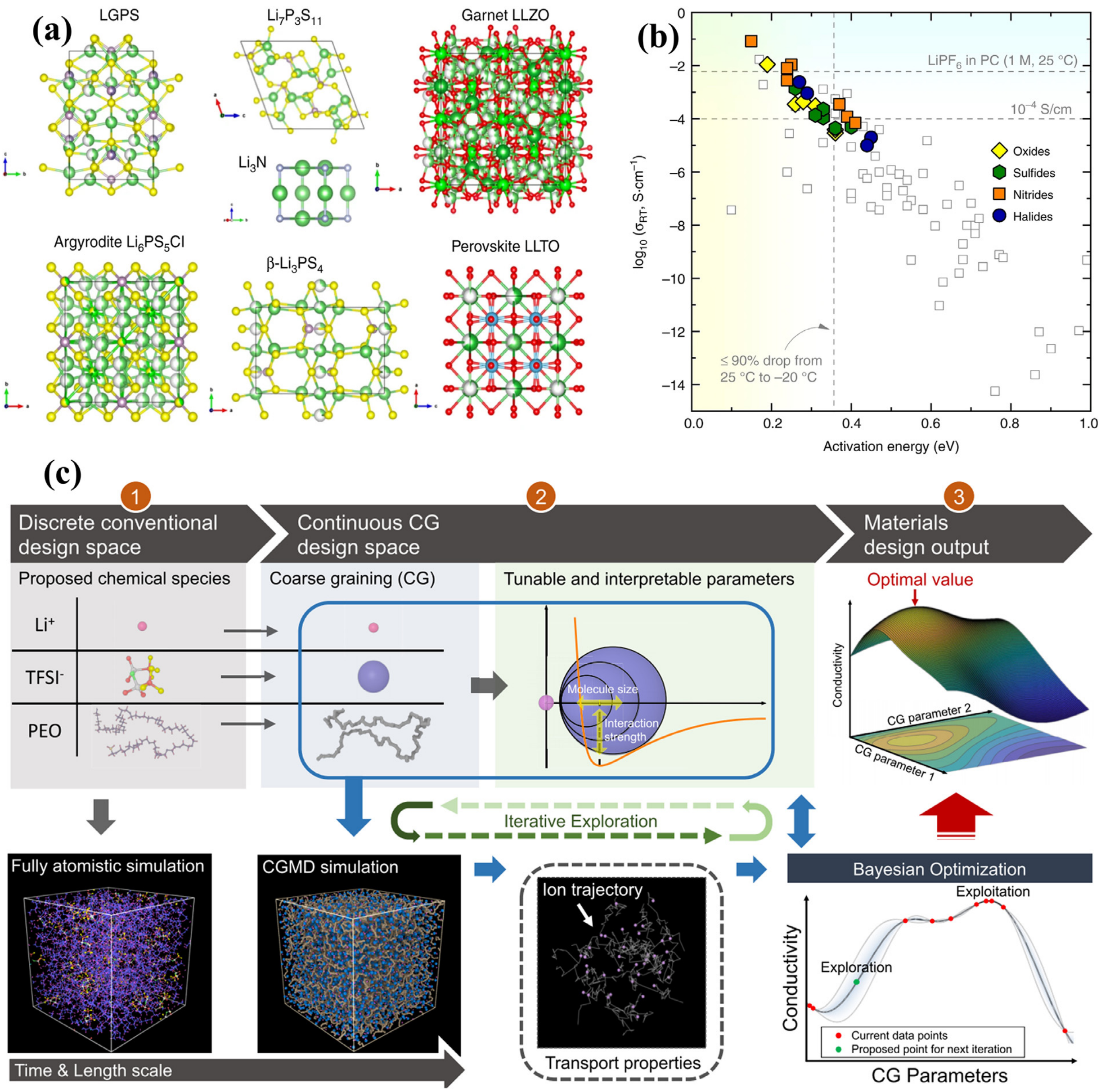


Fig. 14. Data-driven materials science for solid-state battery: (a) Typical crystal structures of existed solid-state lithium-ion conductors. (b) Comparison of ion conduction characteristics between predicted and known solid-state lithium ion conductors. Reproduced with permission from [145], Copyright 2019, Springer Nature Limited. (c) The coarse-grained molecular dynamics–Bayesian optimization framework. Reproduced with permission from [146], Copyright 2020 American Chemical Society.

ovskite systems. Four ML algorithms were thereafter applied to predict the thermodynamic stability of solids. Their results illustrated that the extremely randomized trees give the highest accuracy. It also indicated that ML can be used to significantly speed up high-throughput DFT calculations at least 5 times. Takahashi et al. [172] performed a random forest algorithm to predict the band gap of 9328 perovskite materials. 11 new perovskite materials with proper band gap and formation energy range were discovered. To find out the most critical features for prediction of the electronic band gap in double perovskites, Pilania et al. [173] adopted a ML model based on Kernel ridge regression. The results showed that the lowest Kohn-Sham level and the electronegativity of the elements constituting the atomic species are the most important predictors. Min et al. [174] developed an inorganic ABO_3 perovskite materials screening platform based on ML and active learning. Their results also indicate that the application of ML algorithms can greatly promote the development of new materials.

5.2.3. *Organic photovoltaics*

Organic solar cells are one of the promising technologies for solving the clean energy crisis in the coming decades [175]. However, searching for suitable candidates with desirable performance by laboratory exploration is a time-consuming process [176]. The latest advancements in AI show that the application of ML technology has the potential to expedite the development of organic photovoltaic materials [177,178]. Paul et al. [179] applied a deep neural network to screen organic solar cell by predicting the highest occupied molecular orbital (HOMO) value. This study verifies that the search for high-performance organic solar cells can be performed faster by using transfer learning from a larger calculated data set to a carefully planned data set. To address the challenge of high-throughput molecular design of organic photovoltaic materials, Nagasawa et al. [180] reported a supervised learning screening model of conjugated molecules for polymer-fullerene organic photovoltaic applications. Results showed that the RF model expresses the highest predic-

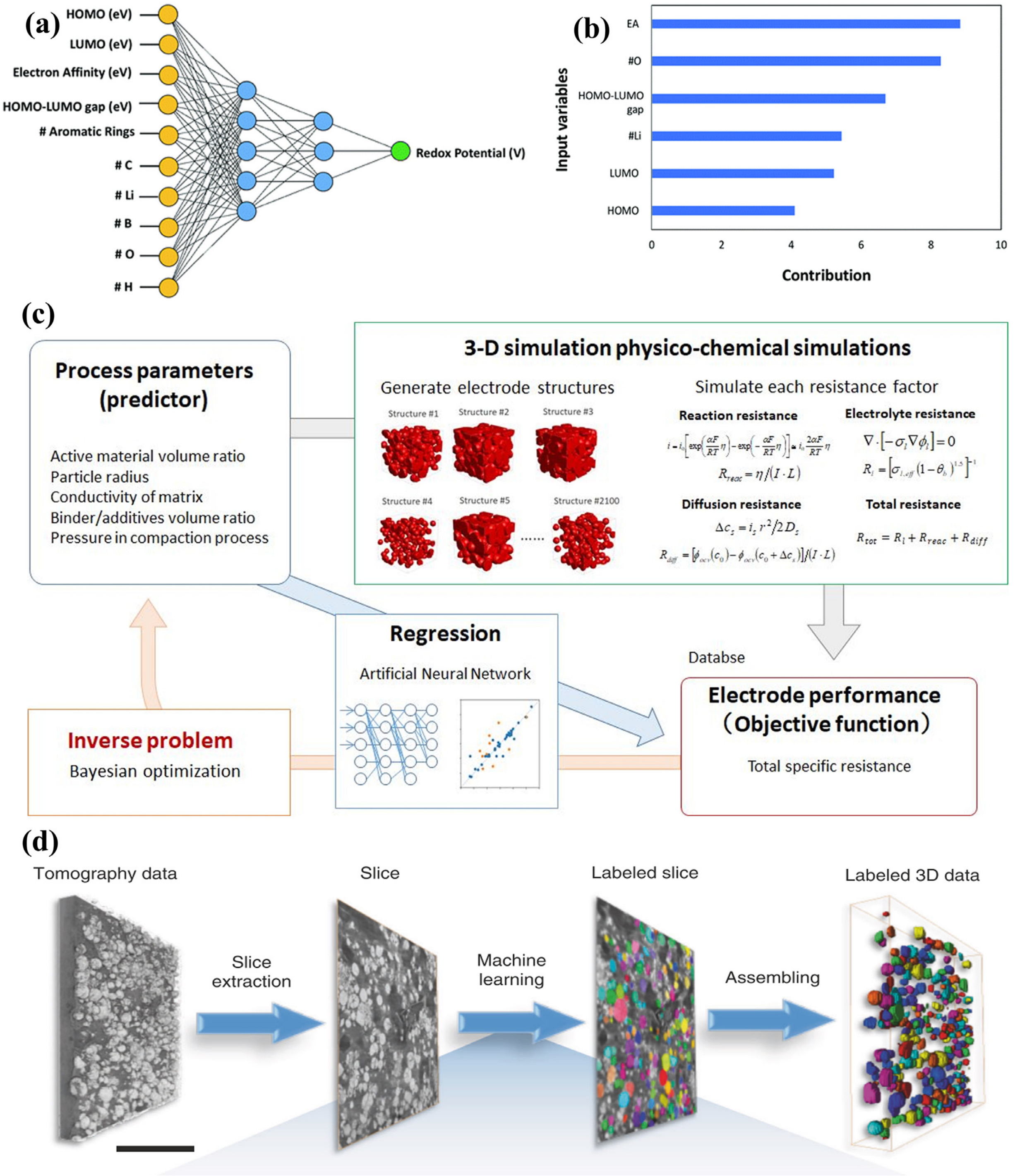

(a)
HOMO (eV)
LUMO (eV)
Electron Affinity (eV)
HOMO-LUMO gap (eV)
# Aromatic Rings
# C
# Li
# B
# O
# H
Redox Potential (V)
(b)
EA
#O
HOMO-LUMO gap
#Li
LUMO
HOMO
Input variables
0
2
4
6
8
10
Contribution
(c)
Process parameters (predictor)
Active material volume ratio
Particle radius
Conductivity of matrix
Binder/additives volume ratio
Pressure in compaction process
3-D simulation physico-chemical simulations
Generate electrode structures
Structure #1
Structure #2
Structure #3
Structure #4
Structure #5
......
Structure #2100
Simulate each resistance factor
Reaction resistance
$i = i_0\left[\exp\left(\frac{\alpha F}{RT}\eta\right) - \exp\left(-\frac{\alpha F}{RT}\eta\right)\right] \approx i_0\frac{2\alpha F}{RT}\eta$
$R_{reac} = \eta/(I \cdot L)$
Electrolyte resistance
$\nabla \cdot [-\sigma_l \nabla \phi_l] = 0$
$R_l = [\sigma_{l,eff}(1-\theta_b)^{1.5}]^{-1}$
Diffusion resistance
$\Delta c_s = i_s r^2/2D_s$
$R_{diff} = [\phi_{ocv}(c_0) - \phi_{ocv}(c_0 + \Delta c_s)]/(I \cdot L)$
Total resistance
$R_{tot} = R_l + R_{reac} + R_{diff}$
Regression
Artificial Neural Network
Databse
Electrode performance （Objective function）
Total specific resistance
Inverse problem
Bayesian optimization
(d)
Tomography data
Slice
Labeled slice
Labeled 3D data
Slice extraction
Machine learning
Assembling

Fig. 15. ML technology for the development of battery electrode materials: (a) ANN model framework with 10 input variables and two hidden layers. (b) Six core input variables determined by feature engineering. Reproduced with permission from [150]. Copyright 2018, The Royal Society of Chemistry 2018. (c) The prediction and optimization framework for porous electrode in lithium-ion battery. Reproduced with permission from [151]. Copyright 2020, MDPI. (d) Workflow of segmentation based on ML technology. Reproduced with permission from [153]. Copyright 2020, Springer Nature Limited.

tion accuracy, which can contribute to the decision making of molecular design. Padula et al. [181] studied the effects of the electronic and structural characteristics of solar cells on their performance by applying linear and nonlinear ML models. The results show that combining DFT calculations and solar cell electronic and structural features, the ML model can achieve higher prediction accuracy. Sahu et al. [182] conducted a data-driven virtual screening of 10,170 candidate molecules for organic photovoltaic cells. With the application of GBRT and ANN model, 126 promising candidates were screened out. This research demonstrates that ML-assisted virtual screening studies have the potential to reveal hidden guidelines that can be used to discover and design promising molecules. In addition to discovering new materials, the key step in practical applications is to synthesize ML screened and predicted materials to confirm the effectiveness of ML models [183]. Sun et al. [184] used a supervised learning method to study the influence of chemical structure on the performance of photovoltaic materials. The results show that the

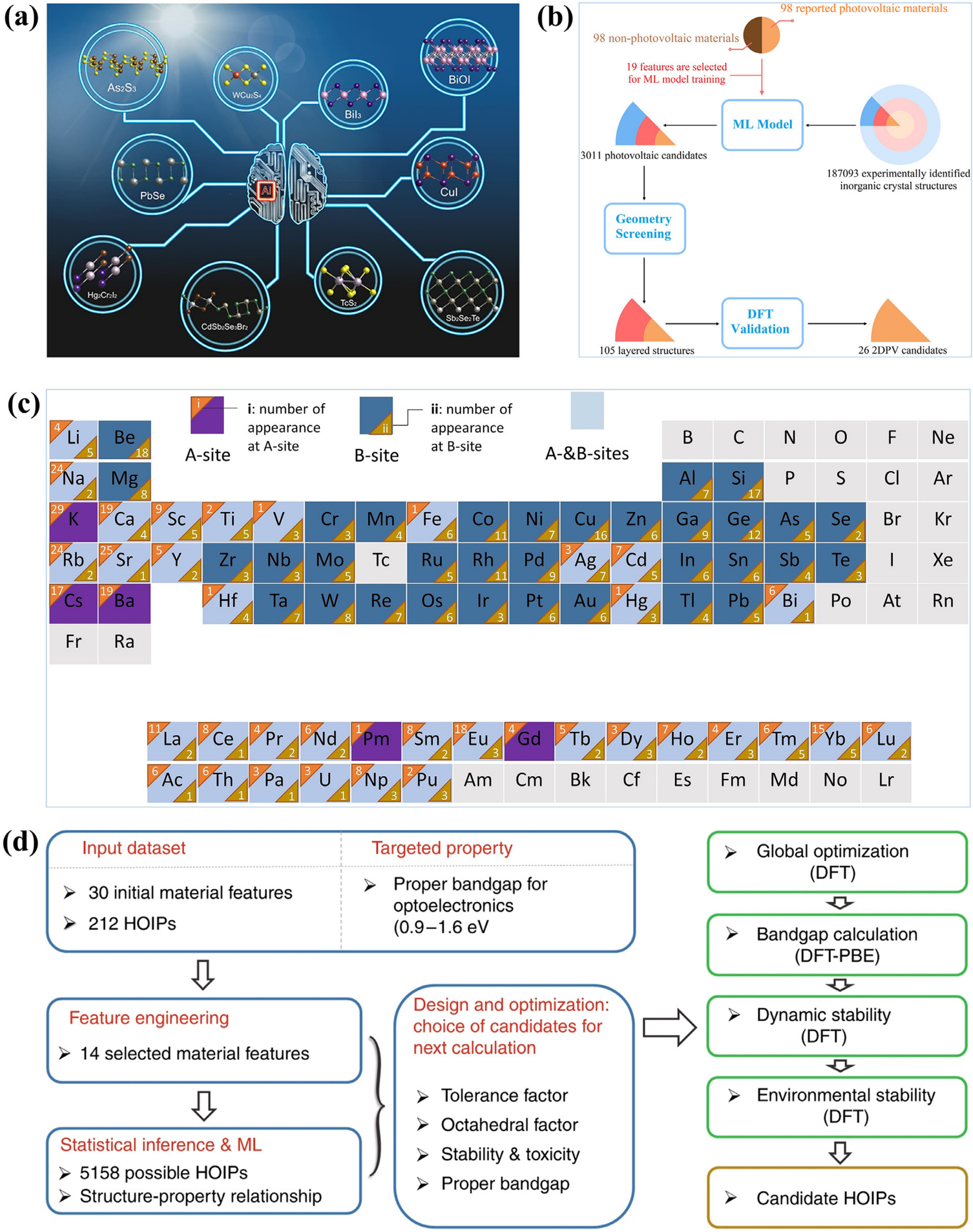


Fig. 16. The application of ML technology for property prediction and screening of photovoltaic materials: (a) two-dimensional (2D) photovoltaic materials. (b) The screening procedure for 2D photovoltaic materials based on ML model. Reproduced with permission from [169]. Copyright 2020, American Chemical Society. (c) The prediction result of 331 ABO_3 perovskites. Reproduced with permission from [123]. Copyright 2020, Elsevier. (d) ML algorithms combined with DFT calculations to screen solar battery materials. Reproduced with permission [170]. Copyright 2020, Springer Nature Limited.

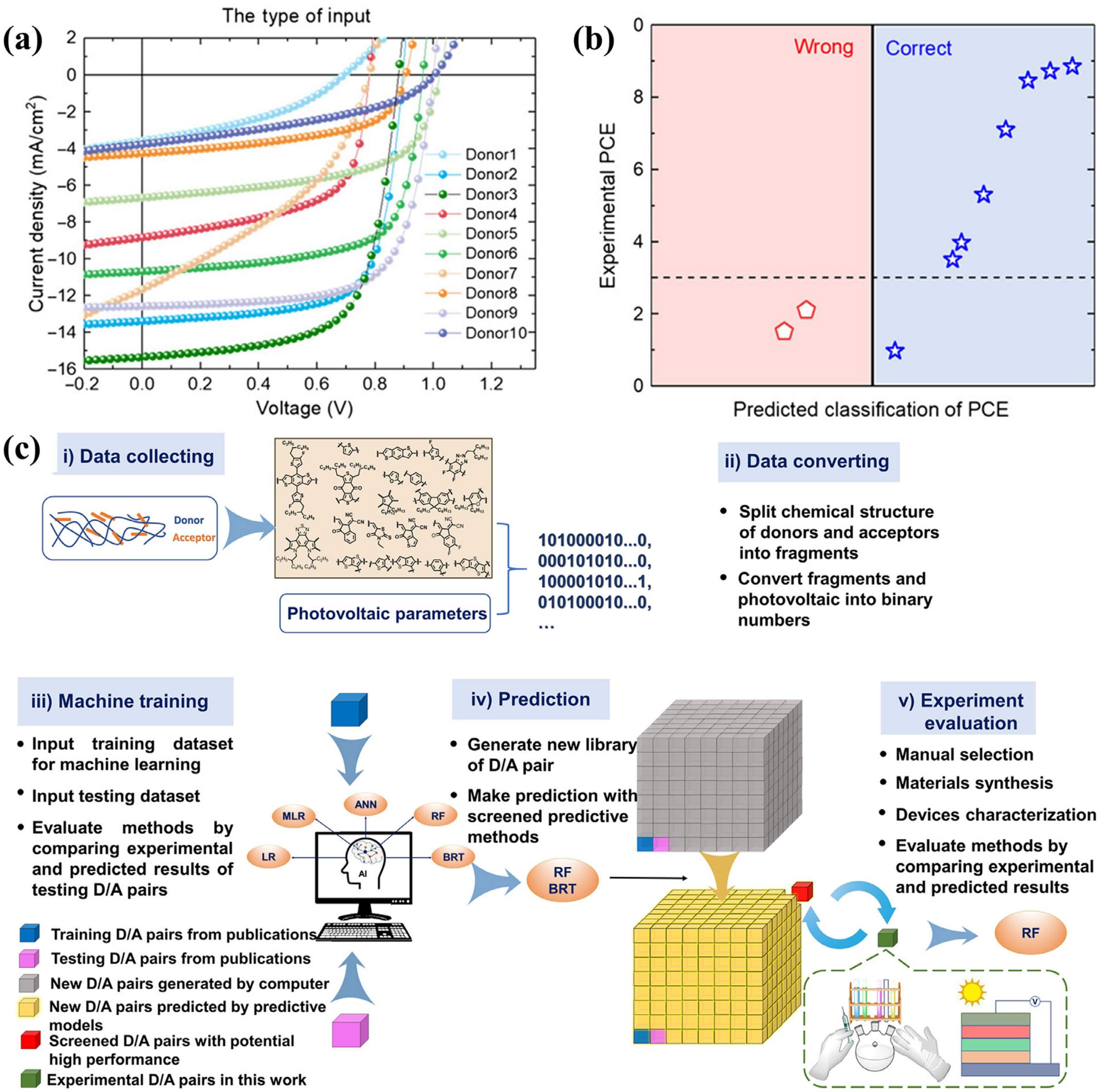


Fig. 17. ML assisted materials design for high-performing organic photovoltaic materials: (a) 10 newly developed molecular donor materials based on ML screening. (b) Prediction results versus experimental data for the discovered organic photovoltaic materials. Reproduced with permission from [184]. Copyright 2019, The Authors, some rights reserved; American Association for the Advancement of Science. (c) Workflow of ML technology in the development of advanced organic solar cells. Reproduced with permission from [185]. Copyright 2020, Springer Nature Limited.

developed ML model can correctly describe the structure-property relationship. In addition, validation experiments were conducted based on 10 newly synthesized donor materials (see Fig. 17**a**) to confirm the reliability of the ML model. The results are shown in Fig. 17**b**, indicating that the experimental results are consistent with the model prediction results. Following the same motivation, Wu et al. [185] proposed an integrated workflow which combined ML technology and experimental validation. As shown in Fig. 17**c**, five ML algorithms were used to process data. Based on the prediction results, manually experimental evaluation will be carried out to synthesize the predicted materials and validate the reliability of developed ML model. In this case study, six photovoltaic donor/acceptor pairs were selected and synthesized. Validation results confirm that the experimental power conversion efficiency is at the same level as the predicted value. In addition to above mentioned research focus, developing new data infrastructures for solar battery is also meaningful. For example, Marchenko et al. [186] developed an open-source database of perovskite materials, including ML predicted information for crystal structures, band gaps, and atomic partial charges. In summary, the application of ML can accelerate the design and discovery of photovoltaic materials, which can further extend the large application potential of renewable energy.

5.3. Catalytic reactions

Catalytic materials play a key role in the application of advanced energy technologies [187,188]. From the past to the future, the screening of new catalytic materials has been a goal pursued by industry and academia [189,190]. Traditional catalytic material development methods through trial and error cannot meet the current needs of the rapid development of industry. The latest revolution in the field of data science has raised the expectation that the application of ML technology can accelerate the development of highly efficient catalyst materials

[191–194]. Choi et al. [195] explored the feasibility of applying ML to predict the activation energy of gas phase reactions. Using molecular structure and thermodynamic properties and their differences as input features, six different ML models based on ANNs, SVR and tree boosting methods were tested. The verification results show that the tree boosting method shows the best predictive performance. Toyao et al. [196] developed a ML model to predict the adsorption energy of CH_4 related substances on Cu-based alloys. Based on the database established by the DFT calculation results, 12 features were selected to construct the ML model with the help of four supervised learning algorithms. The results show that the prediction accuracy of the extra tree regression algorithm is the highest. Based on the model, the adsorption energy can be predicted by the model without time-consuming DFT calculation. Ma et al. [197] proposed an enhanced chemical adsorption model based on ANN, which can quickly and accurately predict the surface reactivity of metal alloys in a wide chemical space. The ANN model was trained by a group of data obtained from theoretical calculation of ideal bimetallic surfaces. The trained ML model was applied to capture the nonlinear interactions of adsorbates on multi-metallic surfaces. This method is expected to facilitate the screening of high-throughput catalysts. In order to determine the active sites on the catalyst surface, Chen et al. [198] developed a comprehensive ML model. The model couples three modules: ANN, multi-scale simulation, and quantum mechanics. Taking the reduction of CO_2 as an example, the properties of all 5,000-10,000 surface parts on the surface of Au nanoparticle surface (AuNPs) and dealloyed Au were explored (refer to **Fig. 18a**). In addition, the activity of the entire surface of the catalyst is visualized by the mapping method (see **Fig. 18b)**. The results showed that ML methods can help guide the design of high-performance CO_2 recovery catalysts. Meyer et al. [199] developed a ML model to predict the oxidative addition reaction energy between transition metal complexes and substrates. It can estimate the activity of homogeneous catalyst through combing the model with a molecular volcano map. A total of 18,062 compounds were predicted, and 557 candidate catalysts that fell into the ideal thermodynamic window were selected. McCullough et al. [200] summarized the latest progress in combining AI algorithms with high-throughput experiments in catalyst discovery. The results show that the AI model can predict and discover new catalysts that do not exist in the existing experimental database. Additionally, the prediction accuracy of the ML model can be improved by considering more complex parameters such as absorption energy and band gap. Compared with the traditional catalytic reactions, the emerging electrocatalytic reactions have gradually received attention [201,202], especially in terms of green energy conversion, such as electrolysis of water to generate hydrogen [203,204], and electrocatalytic reduction of carbon dioxide to carbon-neutral fuels [205,206], electrocatalytic nitrogen reduction to ammonia [207,208], etc. Aiming to discover efficient electrocatalysts for carbon dioxide reduction, Chen et al. [209] developed an ML model to analyse a large number of calculated data sources by using the extreme gradient enhancement regression (XGBR) algorithm. This research provides clues for quickly searching for high-performance catalysts using the predicted value of Gibbs free energy (see **Fig. 18c**). To screen ideal catalysts for hydrogen evolution reaction (HER), Sun et al. [210] applied four ML algorithms coupled with DFT calculation to predict the values of Gibbs free energy of hydrogen adsorption (ΔG_{H^*}). Results show that a higher prediction accuracy was obtained through using SVR model with simple features (see **Fig. 18d**). Additionally, 28 candidates were screened out by ML and five among them were identified as the promising catalysts for HER (see **Fig. 18e**).

In addition to the work mentioned above, materials researchers have explored other aspects in the application of ML to develop catalyst materials [211]. For example, Fischer et al. [212] discussed the application of random forest regression (RFR) in the development of two-dimensional catalytic materials. The results show that the RFR model has high prediction accuracy for the binding energy of small molecules. Although most of the case studies in this section use computational data to build ML models, Smith et al. [213] created an ML model based on experimental data extracted from the literature. The proposed framework can effectively guide experiments and descriptor selection. In order to further lower the threshold for applying ML technology to develop new catalysts materials, Palkovits [125] provided a fundamental tutorial with code for the application of ML in catalysis. The programming code for various ML algorithms is directly provided, enabling a convenient usage for other materials scientists. Moreover, Toyao et al. [214] summarized the latest progresses in data-driven science for catalytic materials, including materials design, synthesis, characterization, and applications, etc. A close-loop roadmap of future catalysis research coupled with ML technology was proposed (see **Fig. 19**), which contains different modules, such as automated synthesis and analysis platform for catalytic materials, data resources and human intuition, theoretical calculation, as well as ML prediction. In summary, the application of ML technology can speed up the development of high-performance catalytic materials.

5.4. CO_2 capture technologies

Metal-organic framework (MOF) has attracted widespread attention because of its tenable structure, which can realize selective CO_2 physical adsorption. However, considering the many functions that can be changed simultaneously in thousands of MOFs so far, it has become very challenging to determine the most critical functions to improve CO_2 capture capacity. The high-throughput screening method based on ML brings hope to improve the performance of MOF [119,215]. Anderson et al. [216] used multi-scale DFT, grand canonical Monte Carlo (GCMC), and ML methods to study the role of different pore chemistry and topological characteristics in enhancing the CO_2 capture indicators of MOFs. The results show that the simple descriptors proposed by "human intuition" for training ML algorithms can be an effective simulation tool for predicting CO_2 capture indicators. Searching for electronically conductive MOFs among thousands of reported MOF structures is a difficult task. He et al. [217] used a new strategy that combined ML technology, statistical multiple selection and ab initio calculations to screen 2,932 MOFs. Six MOF crystal structures with promising performance were determined. In order to accurately determine the candidate MOFs with enhanced CO_2 adsorption capacity, Fernandez et al. [218] developed a quantitative structure-property relationship classifier. The research results show that the ML classifier can reduce the calculation time by an order of magnitude. For the purpose of determining the key factors of carbon dioxide capture capacity, Zhu et al. [219] established a quantitative structure-attribute relationship model based on the RF algorithm. The results show that there is a strong correlation between CO_2 adsorption capacity and pressure (see **Fig. 20a**). In addition, the relative importance of the three key parameters is shown in **Fig. 20b**, which provides direct clues for the selection of key parameters. To summarize the recent state-of-the-arts of applying ML technology to develop novel materials for CO_2 adsorption. Chong et al. [220] provided a systematic review of applications of ML technology in MOFs, which emphasizes the future prospects of data-driven materials science and engineering in the development of carbon dioxide adsorption materials.

6. Future perspectives

6.1. Perspectives for specific energy materials

6.1.1. Alkaline ion batteries materials

Concerning the development of advanced battery materials, there are some challenges hindering the continuous innovation and research on data-driven battery materials. First, when applying the ML algorithm to study the comprehensive influence of electrode and electrolyte materials on battery performance, the ML model established for battery materials is usually more complicated. The reason includes the complex relationship between the structure and composition of the battery, and

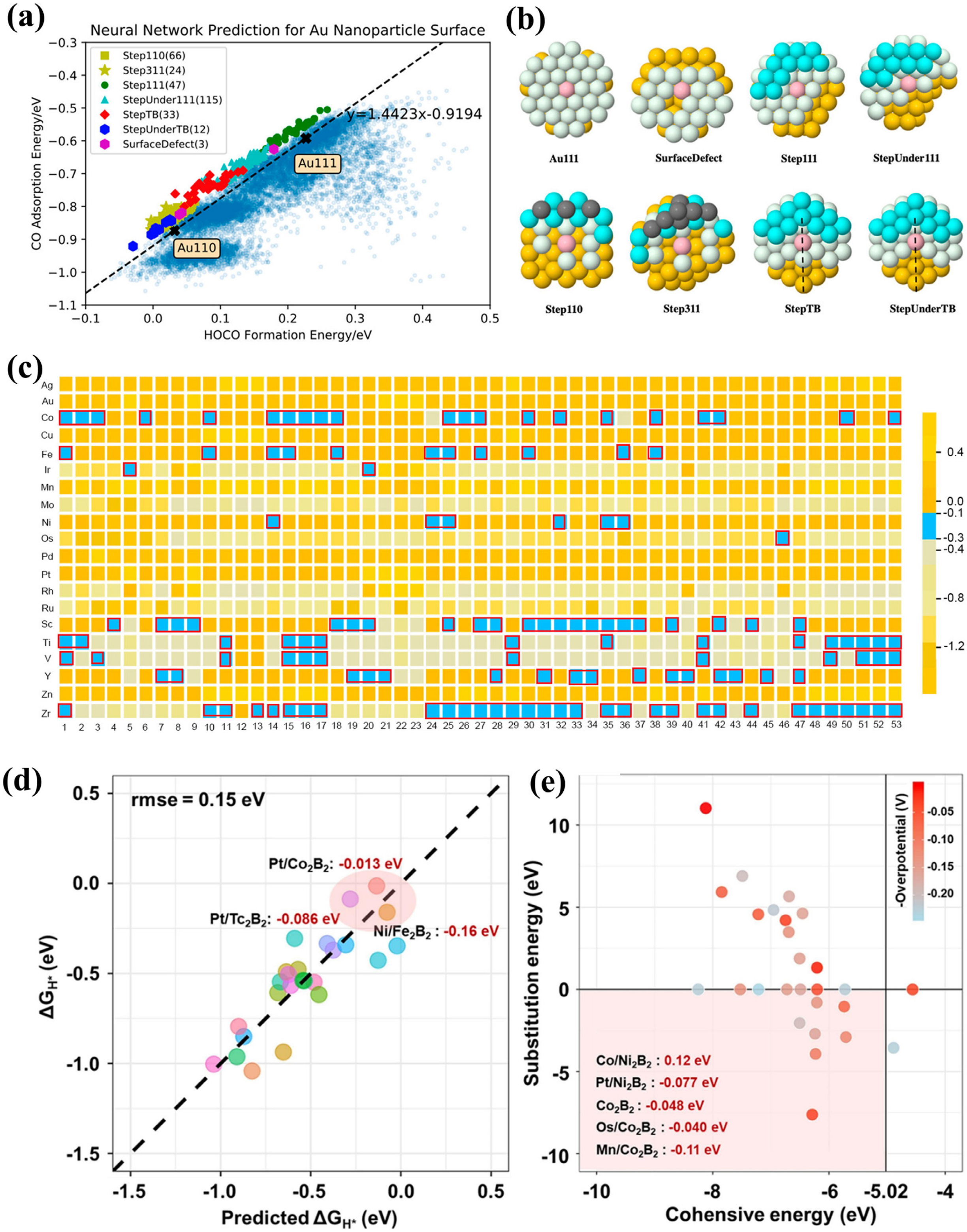


Fig. 18. The application of ML technology for screening and prediction of catalytic materials: (a) Neural network predictions for AuNPs. (b) Identified active sites for AuNPs surfaces. Reproduced with permission from [198]. Copyright 2019, American Chemical Society. (c) The predicted heat map of the Gibbs free energy change of CO adsorption (ΔG_{CO}) for 1060 designed single-atom catalysts. Reproduced with permission from [209]. Copyright 2020, American Chemical Society. (d) Predicted vs. DFT-calculated value of Gibbs free energy of hydrogen adsorption (ΔG_{H^*}). (e) 28 candidates for hydrogen evolution reaction catalysts screened by ML. Reproduced with permission from [210]. Copyright 2020, Elsevier.

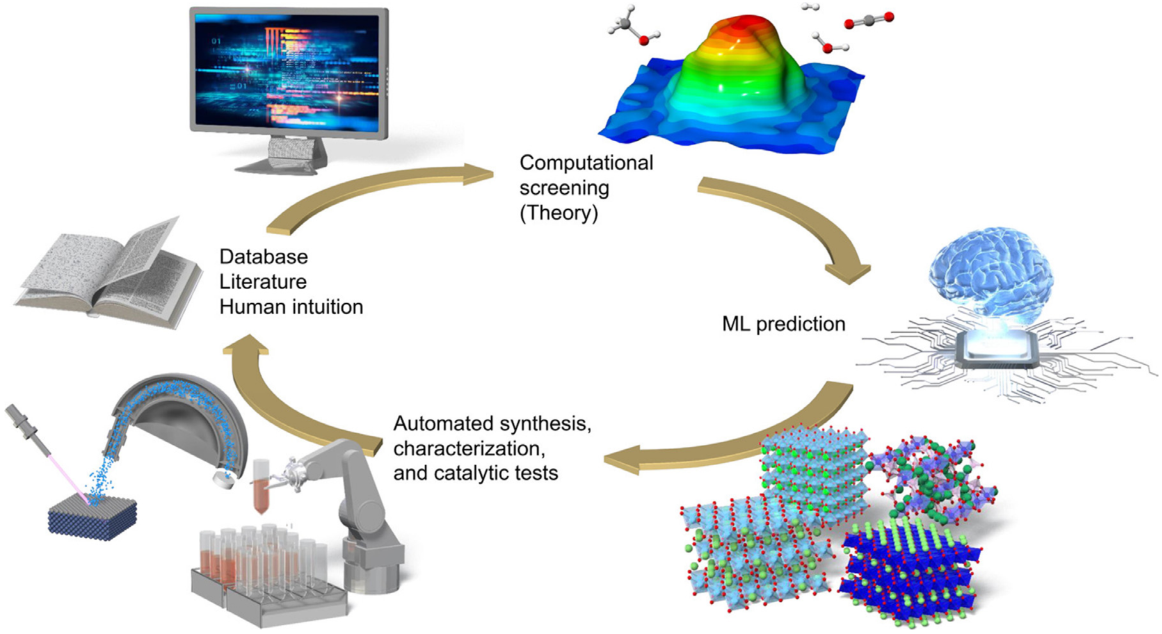


Fig. 19. Schematic diagram of ML-aided future catalysis research. Reproduced with permission from [214]. Copyright 2019, American Chemical Society.

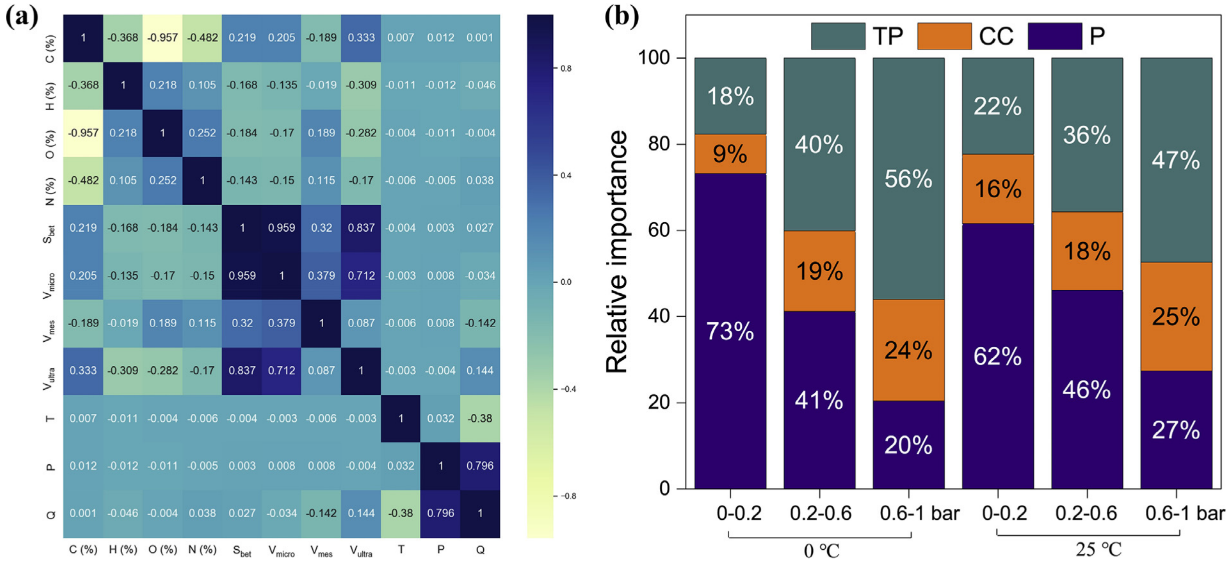


Fig. 20. Using ML technology to explore the key factors of CO2 adsorption capacity: (a) Analysis results of feature correlation. (b) The relative importance of comprehensive influencing factors under various pressure ranges. Reproduced with permission from [219]. Copyright 2020, Elsevier.

various chemical reactions that occur at the electrode-electrolyte interface. Second, battery data resources used for modeling and analysis lack systematic collection and standardization. Furthermore, data resources generated from practical applications in the battery industry are also worth collecting. To address these critical challenges mentioned above, future research can be performed from the following perspectives. On one hand, parameters of ML model can be optimized in accordance with results of feature engineering, which indicate the importance of each parameter. Meanwhile, parameters selection can be further optimized to simplify the model by integrating with optimization algorithms, such as evolutionary algorithm and Bayesian algorithm. On the other hand, it is also suggested to combine domain expert knowledge of battery technology with ML modeling process, thereby maintaining the model reliability above a certain level. Moreover, the development of battery materials can be accelerated by combining various resources, including DFT calculations, ML technologies, and experimental exploration.

6.1.2. Photovoltaic materials

With regard to photovoltaic materials, several critical challenges need to be addressed. First, searching next generation of features, which can perform accurate prediction and easily accessible, is worth further studying. Second, it is recommended to combine domain expert knowledge with feature engineering and modeling process to improve the effectiveness of the developed ML model. At the same time, validation experiments should be carried out to confirm the analysis results of ML model, such as the prediction candidate with high-performance. Currently, only few studies have validated their predicted materials with experiments. Model validation should be considered as a necessary step in the future. Another major issue that exists in the field of data-driven solar materials science is data scarcity. It is believed that the problem of small datasets can be alleviated by the latest developed AI technologies, such as text mining and image recognition. Furthermore, by combining data extracted from DFT calculations and experiments, data sources for photovoltaic materials can be enriched. From the perspective of algorithm selection, ANN and GA were considered as the two most commonly used ML algorithms for solar batteries [162]. In addition to property prediction and advanced materials screening, ML technology can be applied to optimize device structures and fabrication processes of solar batteries to promote the industrialization process of photovoltaic materials. Moreover, although there are large number of research articles related to data-driven photovoltaic materials science and engineering, relevant comprehensive review papers in this field are limited. Therefore, more systematic review papers should be published to point out the future directions as well as pave way for the development of advanced photovoltaic materials.

6.1.3. Catalytic materials

The use of ML technology in the development of new catalytic materials is still in the early stages of exploration and is majorly driven by experience. The possible reason is that the catalytic process usually involves multi-dimensional and multi-scale chemical reactions, which is a complex and dynamic process. In addition, the experimental conditions for catalytic reactions reported in the literature are usually too broad, and some specific experimental details are deliberately hidden. Moreover, the method and format of reporting data, especially grey data, during the experiment are not uniform. The above behaviors have caused difficulties in database establishment and parameter selection, and in turn hindered the rapid development of advanced catalytic materials. To promote the development of catalysis informatics, future research may consider the following perspectives. First, catalytic materials scientists should pay attention to combining ML technology with the existing physical and chemical models of catalytic reactions, which has the potential to improve the overall performance of ML models. Second, the exploration of unknown catalytic materials as well as reaction mechanism can be accelerated by integrating automated technologies such as intelligent robots, and optimization algorithms such as Bayesian algorithm and genetic algorithm. Third, data used for ML modeling should be collected from various sources, such as online open-source databases, computational data sets, and laboratory experimental data.

6.1.4. CO_2 capture materials

The challenges and perspectives of applying data-driven science to the development of CO_2 capture materials can be highlighted from following prospects. First, the development and adaptation of ML algorithms for different CO_2 capture materials systems is worthy of further study. It is noteworthy that RF algorithm has been widely used in the design and discovery of new CO_2 capture materials, thereby showing great practical application potential. In addition, algorithms with optimized functions can be used to screen CO_2 capture materials, such as GA and GBRT. Second, future research should focus on providing design rules to guide the development of new CO_2 capture materials, such as reverse design based on feature engineering. Third, an automatic integrated system for the development of CO_2 capture materials based on intelligent robot technology, DFT calculations, and experimental studies should be developed. Therefore, the predicted materials can be synthesized to further optimize the ML model.

In addition to the specific challenges and future perspectives for each energy material mentioned above, general perspectives that are applicable to all these materials to promote the development of data-driven energy materials science and engineering are discussed in the following section.

6.2. Perspectives for data-driven energy materials sciences

6.2.1. Improvement and standardization of data infrastructure

ML algorithms are basically extracted knowledge from previous data sources which are commonly derived from computational or experimental results. In this case, a large amount of training data can help the ML model to achieve higher accuracy. However, a critical issue during the application of data-driven materials science is data scarcity, especially for the data sources derived from experimental studies. The reason is that material scientists report their data in different formats, hence posing challenges in the unified collection of data. In addition, the grey data or failed data in the experiment was deliberately hidden. To solve the problem of data sparsity, some promising directions can be further explored in the future.

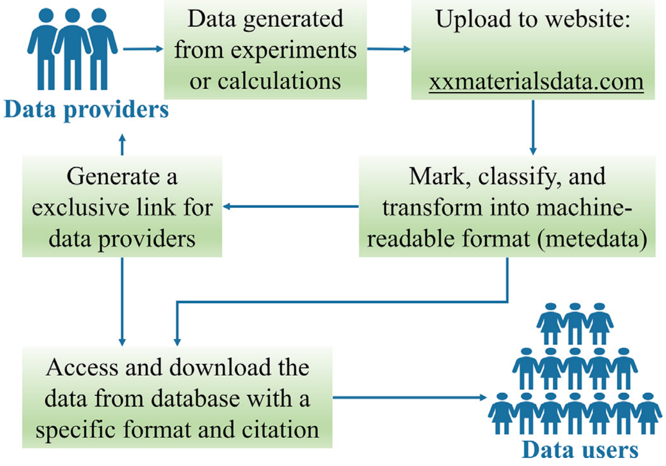

Data providers
Data generated from experiments or calculations
Upload to website: xxmaterialsdata.com
Mark, classify, and transform into machine-readable format (metedata)
Generate a exclusive link for data providers
Access and download the data from database with a specific format and citation
Data users

Fig. 21. The future data infrastructures in data-driven materials science and engineering.

First, the conventional data format reported or published in the materials community should be changed to facilitate the direct and easy collection of data from publications or literature. As shown in **Fig. 21**, researchers and scientists can upload data generated from experiments or calculations to an online open database website. The online database will then mark the data with the provider information and classify the data into a specific subset based on the intelligent recommendation from the system. The data provider can also select the subset manually. Afterwards, the online system will transform the uploaded data into a general machine-readable format. Meanwhile, an exclusive link will also be generated for the data provider. Thus, data users can access and download the data from online database with a specific format and citation. Moreover, for the purpose of accelerating the development of advanced energy materials, data providers can as well add such link as one of the supporting information in their publications. Second, the natural language processing technology, which has been successfully applied to text and image recognition, can be introduced to help materials scientists mine large-scale data from existing literature. For instance, using natural language processing techniques, Kim et al. [221] developed an entity recognition model to connect scientific literature to inorganic synthesis insights. Third, materials engineers should consider exploring data fusion in the next few decades, that is, integrating multiple data sources to generate more consistent and accurate information than any single data source could provide. For instance, Ward et al. [222] improved the accuracy of ML model by 30 % through adding the NIST data into original training data. Hence, more application of this technique should be largely embraced. Fourth, researchers in the materials community should also pay more attention to the results of failed experiments that are usually overlooked. Scientists from Harvard University has launched a project for ML assisted materials discovery using failed experiments [76]. The results show that the application of grey data can pave way for materials development. However, the number of meaningful attempts in this topic is still limited. Therefore, more intensive research in related fields should be carried out. Furthermore, more online opening databases (such as The Materials Projects and NIST), tools and software (such as Jupyter Notebook and GitHub) need to be developed to promote the development of data-driven materials science.

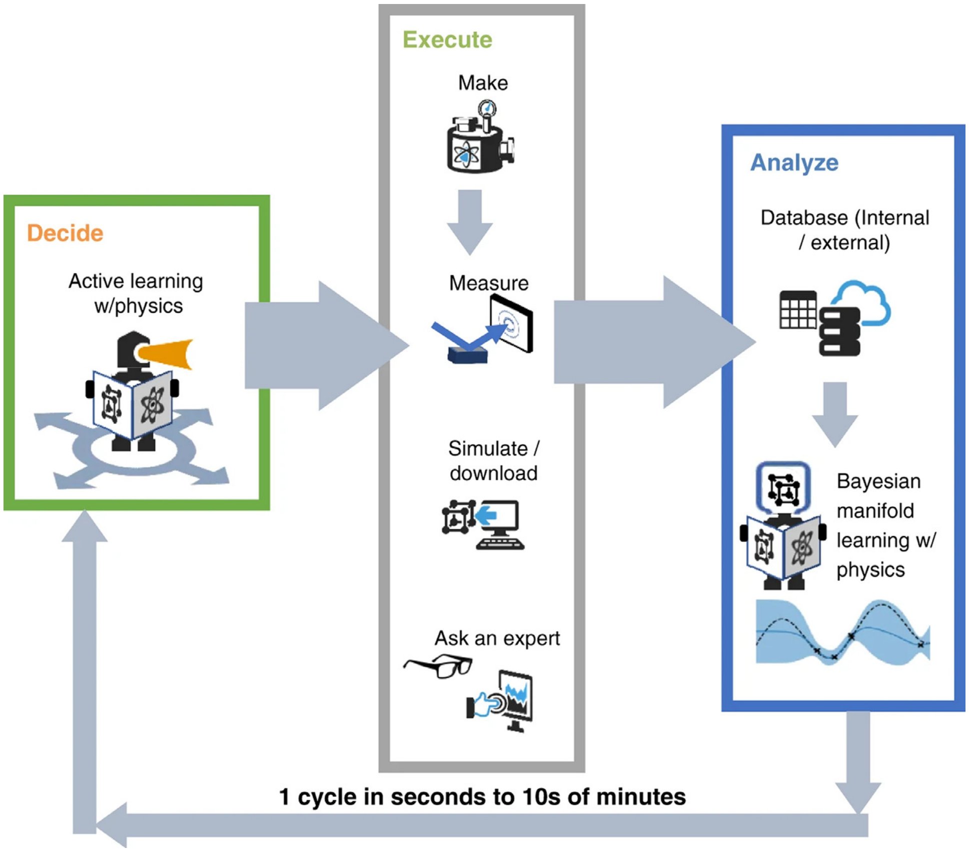

Decide
Active learning w/physics
Execute
Make
Measure
Simulate / download
Ask an expert
Analyze
Database (Internal / external)
Bayesian manifold learning w/ physics
1 cycle in seconds to 10s of minutes

Fig. 22. Closed-loop autonomous materials exploration and optimization. Reproduced with permission [223]. Copyright 2020, Springer Nature Limited.

6.2.2. *Automatic closed-loop optimization framework and model visualization*

Even though ML technology has widely proven to be a useful tool in data-driven materials science, no doubt that there are still some challenges that need to be addressed. For example, in the ML modeling process, the selection of descriptors and the setting of parameters are largely dependent on manual decision-making. Meanwhile, the value of each parameter cannot be automatically updated based on the results of the previous round, thereby resulting in increased time cost. Furthermore, ML models constructed by certain algorithms such as neural networks are difficult to interpret because these models are usually not visible. In this regard, relevant keys and perspectives to successful applications of ML are therefore highlighted as follows.

First, the closed-loop optimization framework of ML algorithms should be developed to accelerate the process of materials discovery. As shown in Fig. 22, the main objective of this approach is to develop a closed-loop iterative process which can formulate hypotheses about manufacturing materials with given structures and properties. The automatic framework would therefore be able to plan and perform experiments, as well as interpret the results. The knowledge extracted from previous round can thereafter be applied to design the next round of experimental exploration as well as simulation through combing the Bayesian active learning. This method has its own disadvantages such as autonomous optimization, that is, the optimal candidate obtained by the optimization process may be a local optimal solution (the ideal solution should be the global optimal solution). In this case, other promising candidates will be unconsciously excluded. Herein, we suggest that the closed-loop optimization framework should be used in conjunction with algorithms supporting global optimization (for example, evolutionary algorithm) to avoid the trap of local optimal solutions. In addition to the automatic framework mentioned above, the application of deep learning neural networks in materials science is also recommended. The reason is that deep learning has a strong nonlinear fitting ability, which can simulate the complex relationship between various features and reveal the material synthesis mechanism. Third, the ML model needs to be more visible. Although the application of ML can facilitate the prediction of materials properties and screen potential candidates, the interpretability of ML model still deserves further exploration. Once the ML model can be clearly explained, the relationship between various parameters and material properties would be easily identified, thereby further promoting the development of energy materials.

6.2.3. *Intelligent robot self-driving laboratory and predictive material synthesis*

The data generated from experiments have significant impacts on data-driven material science. However, due to the low success rate and time-consumption, data scarcity is a common problem in databases that is largely composed of experimental data. Thus, more efforts should be taken to produce large amounts of data samples. With the rapid development of intelligent robots and 3-D printing technology, the concept of self-driving laboratory by robots is gradually coming into fruition. Recently, Andrew Cooper's team at the University of Liverpool has developed an intelligent robot that can operate autonomously over eight days, perform 688 experiments within and identify photocatalyst mixtures that are six times more active [224]. It is foreseeable that such AI chemists will become powerful assistants for materials scientists. Another major challenge hindering the development of materials genomics is the manufacture of materials for ML prediction. The possible reason may be that the structure or component of the predicted material is difficult to synthesize. Moreover, there are currently no predictive theories about which compounds can be synthesized and how they can be synthesized. Thus, research on exploring theories to synthesize ML predicted materials will be a promising direction.

6.2.4. Interdisciplinary communications and supportive policies

Since all scientists have their own expertise and terminology, developing a common language among different disciplines plays a key role in computational materials science. On this basis, the cooperation and joint efforts among computer scientists, chemists, physicists, and materials engineers could promote and hasten the development of new materials. One feasible suggestion is for universities to organize seminars and summer schools, and also develop courses that bridge these areas, such as the International Summer School-Deep Materials: Perspectives on Data-Driven Materials Research, hosted by the Italian national enterprise in nanoscience and nanotechnology. More importantly, the supportive policies and initiatives from the government, research institutions, and universities can further accelerate the development of data-driven materials science and engineering. For example, to achieve a carbon-neutral Europe by 2050, the BATTERY 2030+ Roadmap was proposed to invent the batteries of the future. We hope there will be more projects to speed up scientific discoveries in data-driven materials science.

7. Concluding remarks

The latest progress in data-driven materials science and engineering shows that the application of ML technology can greatly facilitate the discovery, design, development, and deployment of advanced energy materials. In this paper, we first present the roadmap to carbon neutrality to illustrate the importance and necessity of developing novel energy materials. Second, a comprehensive review of fundamental ML tutorials is provided, including open-source materials databases, feature engineering, detailed introduction of typical ML algorithms, and effectiveness analysis of ML model. Afterwards, the recent progress in data-driven materials science and engineering including alkaline ion battery materials, photovoltaic materials, catalytic materials, and CO_2 capture materials, are introduced and discussed. This consists of performance-prediction, screening of potential candidates, and closed-loop optimization of properties for energy materials. Moreover, the keys to successful ML applications and remaining challenges are highlighted, such as improvement and standardization of data infrastructure, ML techniques (automatic closed-loop optimization and model visualization), experimental exploration (self-driving laboratory by robots), interdisciplinary communications and supporting policies. We further highlight the future potentials of automatic closed-loop optimization techniques as well as the application of AI robots. We believe the state-of-the-arts in the application of ML within materials community summarized in this paper would pave way for the development of high-performing energy materials.

Declaration of Competing Interest

There are no conflicts to declare.

Acknowledgment

This work was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project no. 15222018).

References

[1] Jin D, Ocone R, Jiao K, Xuan J. Energy and AI. Energy AI 2020;1:100002, . doi:10.1016/j.egyai.2020.100002.

[2] Perera ATD, Nik VM, Chen D, Scartezzini JL, Hong T. Quantifying the impacts of climate change and extreme climate events on energy systems. Nat Energy 2020;5:150–9. doi:10.1038/s41560-020-0558-0.

[3] Gao P, Chen Z, Gong Y, Zhang R, Liu H, Tang P, et al. The role of cation vacancies in electrode materials for enhanced electrochemical energy storage: synthesis, advanced characterization, and fundamentals. Adv Energy Mater 2020;10:1–25. doi:10.1002/aenm.201903780.

[4] Jurasz J, Canales FA, Kies A, Guezgouz M, Beluco A. A review on the complementarity of renewable energy sources: concept, metrics, application and future research directions. Sol Energy 2020;195:703–24. doi:10.1016/j.solener.2019.11.087.

[5] Dusastre V, Martiradonna L. Materials for sustainable energy. Nat Mater 2016;16:15. doi:10.1038/nmat4838.

[6] Kang Y, Li L, Li B. Recent progress on discovery and properties prediction of energy materials: Simple machine learning meets complex quantum chemistry. J Energy Chem 2021;54:72–88. doi:10.1016/j.jechem.2020.05.044.

[7] Kang P, Liu Z, Abou-Rachid H, Guo H. Machine-learning assisted screening of energetic materials. J Phys Chem A 2020;124:5341–51. doi:10.1021/acs.jpca.0c02647.

[8] Neugebauer J, Hickel T. Density functional theory in materials science. Wiley Interdiscip Rev Comput Mol Sci 2013;3:438–48. doi:10.1002/wcms.1125.

[9] Himanen L, Geurts A, Foster AS, Rinke P. Data-driven materials science: status, challenges, and perspectives. Adv Sci 2019;6. doi:10.1002/advs.201900808.

[10] Chen A, Zhang X, Zhou Z. Machine learning: Accelerating materials development for energy storage and conversion. InfoMat 2020;2:553–76. doi:10.1002/inf2.12094.

[11] Zhou T, Song Z, Sundmacher K. Big Data creates new opportunities for materials research: a review on methods and applications of machine learning for materials design. Engineering 2019;5:1017–26. doi:10.1016/j.eng.2019.02.011.

[12] Wang AYT, Murdock RJ, Kauwe SK, Oliynyk AO, Gurlo A, Brgoch J, et al. Machine learning for materials scientists: an introductory guide toward best practices. Chem Mater 2020;32:4954–65. doi:10.1021/acs.chemmater.0c01907.

[13] Suh C, Fare C, Warren JA, Pyzer-Knapp EO. Evolving the materials genome: how machine learning is fueling the next generation of materials discovery. Annu Rev Mater Res 2020;50:1–25. doi:10.1146/annurev-matsci-082019-105100.

[14] Das S, Pegu H, Sahu K, Nayak AK, Ramakrishna S, Datta D, et al. Machine learning in materials modeling - fundamentals and the opportunities in 2D materials. INC 2020. doi:10.1016/b978-0-12-818475-2.00019-2.

[15] Jablonka KM, Ongari D, Moosavi SM, Smit B. Big-Data science in porous materials: materials genomics and machine learning. Chem Rev 2020;120:8066–129. doi:10.1021/acs.chemrev.0c00004.

[16] Barnett JW, Bilchak CR, Wang Y, Benicewicz BC, Murdock LA, Bereau T, et al. Designing exceptional gas-separation polymer membranes using machine learning. Sci Adv 2020;6:1–8. doi:10.1126/sciadv.aaz4301.

[17] Gayon-Lombardo A, Mosser L, Brandon NP, Cooper SJ. Pores for thought: generative adversarial networks for stochastic reconstruction of 3D multi-phase electrode microstructures with periodic boundaries. Npj Comput Mater 2020;6:1–11. doi:10.1038/s41524-020-0340-7.

[18] Cai J, Chu X, Xu K, Li H, Wei J. Machine learning-driven new material discovery. Nanoscale Adv 2020;2:3115–30. doi:10.1039/d0na00388c.

[19] Wei J, Chu X, Sun X, Xu K, Deng H, Chen J, et al. Machine learning in materials science. InfoMat 2019;1:338–58. doi:10.1002/inf2.12028.

[20] Fanourgakis GS, Gkagkas K, Tylianakis E, Froudakis GE. A universal machine learning algorithm for large-scale screening of materials. J Am Chem Soc 2020;142:3814–22. doi:10.1021/jacs.9b11084.

[21] Stocker S, Csányi G, Reuter K, Margraf JT. Machine learning in chemical reaction space. Nat Commun 2020;11:1–11. doi:10.1038/s41467-020-19267-x.

[22] Kang P, Liu Z, Abou-Rachid H, Guo H. Machine-learning assisted screening of energetic materials. J Phys Chem A 2020;124:5341–51. doi:10.1021/acs.jpca.0c02647.

[23] Iwasaki Y, Ishida M, Shirane M. Predicting material properties by integrating high-throughput experiments, high-throughput ab-initio calculations, and machine learning. Sci Technol Adv Mater 2020;21:25–8. doi:10.1080/14686996.2019.1707111.

[24] Chibani S, Coudert FX. Machine learning approaches for the prediction of materials properties. APL Mater 2020;8. doi:10.1063/5.0018384.

[25] Tang B, Lu Y, Zhou J, Chouhan T, Wang H, Golani P, et al. Machine learning-guided synthesis of advanced inorganic materials. Mater Today 2020;41:72–80. doi:10.1016/j.mattod.2020.06.010.

[26] Zhou Q, Lu S, Wu Y, Wang J. Property-oriented material design based on a data-driven machine learning technique. J Phys Chem Lett 2020;11:3920–7. doi:10.1021/acs.jpclett.0c00665.

[27] Moosavi SM, Jablonka KM, Smit B. The role of machine learning in the understanding and design of materials. J Am Chem Soc 2020. doi:10.1021/jacs.0c09105.

[28] Barrett DH, Haruna A. Artificial intelligence and machine learning for targeted energy storage solutions. Curr Opin Electrochem 2020;21:160–6. doi:10.1016/j.coelec.2020.02.002.

[29] Hong Y, Hou B, Jiang H, Zhang J. Machine learning and artificial neural network accelerated computational discoveries in materials science. Wiley Interdiscip Rev Comput Mol Sci 2020;10:1–21. doi:10.1002/wcms.1450.

[30] Wang H, Ji Y, Li Y. Simulation and design of energy materials accelerated by machine learning. Wiley Interdiscip Rev Comput Mol Sci 2020;10:1–18. doi:10.1002/wcms.1421.

[31] Shimizu R, Kobayashi S, Watanabe Y, Ando Y, Hitosugi T. Autonomous materials synthesis by machine learning and robotics. APL Mater 2020;8:2–8. doi:10.1063/5.0020370.

[32] Meredig B. Five high-impact research areas in machine learning for materials science. Chem Mater 2019;31:9579–81. doi:10.1021/acs.chemmater.9b04078.

[33] Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nature 2018;559:547–55. doi:10.1038/s41586-018-0337-2.

[34] Lu W, Xiao R, Yang J, Li H, Zhang W. Data mining-aided materials discovery and optimization. J Mater 2017;3:191–201. doi:10.1016/j.jmat.2017.08.003.

[35] Wang AYT, Murdock RJ, Kauwe SK, Oliynyk AO, Gurlo A, Brgoch J, et al. Machine learning for materials scientists: an introductory guide toward best practices. Chem Mater 2020;32:4954–65. doi:10.1021/acs.chemmater.0c01907.

[36] Liu Y, Zhao T, Ju W, Shi S, Shi S, Shi S. Materials discovery and design using machine learning. J Mater 2017;3:159–77. doi:10.1016/j.jmat.2017.08.002.

[37] Gu GH, Noh J, Kim I, Jung Y. Machine learning for renewable energy materials. J Mater Chem A 2019;7:17096–117. doi:10.1039/c9ta02356a.

[38] Li J, Lim K, Yang H, Ren Z, Raghavan S, Chen PY, et al. AI applications through the whole life cycle of material discovery. Matter 2020;3:393–432. doi:10.1016/j.matt.2020.06.011.

[39] Chen C, Zuo Y, Ye W, Li X, Deng Z, Ong SP. A critical review of machine learning of energy materials. Adv Energy Mater 2020;10:1903242.

[40] Correa-Baena JP, Hippalgaonkar K, van Duren J, Jaffer S, Chandrasekhar VR, Stevanovic V, et al. Accelerating materials development via automation, machine learning, and high-performance computing. Joule 2018;2:1410–20. doi:10.1016/j.joule.2018.05.009.

[41] Crippa M, Guizzardi D, Muntean M, Schaaf E, Solazzo E, Monforti-Ferrario F, et al. Fossil CO2 Emissions of All World Countries 2020 Report; 2020. doi:102760/56420.

[42] International Renewable Energy Agency. IRENA. Global Energy Transformation. Global Energy Transformation, 2019; 2019.

[43] IEA. India 2020 policy energy review. www.IEA.org 2017:1–304.

[44] Bureau E. Environment Bureau 2017.

[45] Edström K, Dominko R, Fichtner M, Otuszewski T, Perraud S, Punckt C, et al. BATTERY 2030+. Inventing the sustainable batteries of the future. Res Needs Future Act 2020:83.

[46] Gosens J, Kåberger T, Wang Y. China's next renewable energy revolution: goals and mechanisms in the 13th five year plan for energy. Energy Sci Eng 2017;5:141–55. doi:10.1002/ese3.161.

[47] Erickson LE, Brase G, Erickson LE, Brase G. Paris agreement on climate change. Reducing Greenh Gas Emiss Improv Air Qual 2019:11–22. doi:10.1201/9781351116589-2.

[48] Javed MS, Ma T, Jurasz J, Amin MY. Solar and wind power generation systems with pumped hydro storage: review and future perspectives. Renew Energy 2020;148:176–92. doi:10.1016/j.renene.2019.11.157.

[49] Barsali S, Ciambellotti A, Giglioli R, Paganucci F, Pasini G. Hybrid power plant for energy storage and peak shaving by liquefied oxygen and natural gas. Appl Energy 2018;228:33–41. doi:10.1016/j.apenergy.2018.06.042.

[50] Levasseur A, Mercier-Blais S, Prairie YT, Tremblay A, Turpin C. Improving the accuracy of electricity carbon footprint: estimation of hydroelectric reservoir greenhouse gas emissions. Renew Sustain Energy Rev 2021;136:110433, . doi:10.1016/j.rser.2020.110433.

[51] Choi E, Ha J, Hahm D, Kim M. A review of multihazard risk assessment: progress, potential, and challenges in the application to nuclear power plants. Int J Disaster Risk Reduct 2020:101933. doi:10.1016/j.ijdrr.2020.101933.

[52] Fuhrman J., Clarens A.F., McJeon H., Patel P., Doney S.C., Shobe W.M., et al. China's 2060 carbon neutrality goal will require up to 2.5 GtCO2/year of negative emissions technology deployment 2020:1–11.

[53] Esan OC, Shi X, Pan Z, Huo X, An L, Zhao TS. Modeling and simulation of flow batteries. Adv Energy Mater 2020;10:1–42. doi:10.1002/aenm.202000758.

[54] Pan ZF, An L, Wen CY. Recent advances in fuel cells based propulsion systems for unmanned aerial vehicles. Appl Energy 2019;240:473–85. doi:10.1016/j.apenergy.2019.02.079.

[55] Cano ZP, Banham D, Ye S, Hintennach A, Lu J, Fowler M, et al. Batteries and fuel cells for emerging electric vehicle markets. Nat Energy 2018;3:279–89. doi:10.1038/s41560-018-0108-1.

[56] Attia PM, Grover A, Jin N, Severson KA, Markov TM, Liao YH, et al. Closed-loop optimization of fast-charging protocols for batteries with machine learning. Nature 2020;578:397–402. doi:10.1038/s41586-020-1994-5.

[57] Ng M-F, Zhao J, Yan Q, Conduit GJ, Seh ZW. Predicting the state of charge and health of batteries using data-driven machine learning. Nat Mach Intell 2020;2:161–70. doi:10.1038/s42256-020-0156-7.

[58] Liu Y, Guo B, Zou X, Li Y, Shi S. Machine learning assisted materials design and discovery for rechargeable batteries. Energy Storage Mater 2020;31:434–50. doi:10.1016/j.ensm.2020.06.033.

[59] Wu QX, Pan ZF, An L. Recent advances in alkali-doped polybenzimidazole membranes for fuel cell applications. Renew Sustain Energy Rev 2018;89:168–83. doi:10.1016/j.rser.2018.03.024.

[60] Pan Z, Bi Y, An L. A cost-effective and chemically stable electrode binder for alkaline-acid direct ethylene glycol fuel cells. Appl Energy 2020;258:114060, . doi:10.1016/j.apenergy.2019.114060.

[61] Ummary S. Renewable energy to fuels through utilization of EnergyDense liquids (REFUEL) program overview 2016:1–16.

[62] Li F, Thevenon A, Rosas-Hernández A, Wang Z, Li Y, Gabardo CM, et al. Molecular tuning of CO2-to-ethylene conversion. Nature 2020;577:509–13. doi:10.1038/s41586-019-1782-2.

[63] Bogdanov D, Farfan J, Sadovskaia K, Aghahosseini A, Child M, Gulagi A, et al. Radical transformation pathway towards sustainable electricity via evolutionary steps. Nat Commun 2019;10:1–16. doi:10.1038/s41467-019-08855-1.

[64] Wei J, De Luna P, Bengio Y, Aspuru-Guzik A, Sargent E. Use machine learning to find energy materials. Nature 2017;552:23–5. doi:10.1038/d41586-017-07820-6.

[65] Yang X, Luo Z, Huang Z, Zhao Y, Xue Z, Wang Y, et al. Development status and prospects of artificial intelligence in the field of energy conversion materials. Front Energy Res 2020;8:1–12. doi:10.3389/fenrg.2020.00167.

[66] Reyes KG, Maruyama B. The machine learning revolution in materials? MRS Bull 2019;44:530–7. doi:10.1557/mrs.2019.153.

[67] Bzdok D, Krzywinski M, Altman N. Points of significance: machine learning: supervised methods. Nat Methods 2018;15:5–6. doi:10.1038/nmeth.4551.

[68] Wang Y, Seo B, Wang B, Zamel N, Jiao K, Adroher XC. Fundamentals, materials, and machine learning of polymer electrolyte membrane fuel cell technology. Energy AI 2020;1:100014, . doi:10.1016/j.egyai.2020.100014.

[69] Chen C, Zuo Y, Ye W, Li X, Deng Z, Ong SP. A critical review of machine learning of energy materials. Adv Energy Mater 2020;10:1–36. doi:10.1002/aenm.201903242.

[70] Nosengo N. The material code. Nature 2016;533:22–5. doi:10.1038/533022a.

[71] de Pablo JJ, Jackson NE, Webb MA, Chen LQ, Moore JE, Morgan D, et al. New frontiers for the materials genome initiative. Npj Comput Mater 2019;5:1–23. doi:10.1038/s41524-019-0173-4.

[72] Toher C, Oses C, Hicks D, Gossett E, Rose F, Nath P, et al. The AFLOW fleet for materials discovery. ArXiv 2017:1–14. doi:10.1007/978-3-319-44677-6_63.

[73] Kirklin S, Saal JE, Meredig B, Thompson A, Doak JW, Aykol M, et al. The open quantum materials database (OQMD): Assessing the accuracy of DFT formation energies. Npj Comput Mater 2015;1. doi:10.1038/npjcompumats.2015.10.

[74] Talirz L, Kumbhar S, Passaro E, Yakutovich AV, Granata V, Gargiulo F, et al. Materials cloud, a platform for open computational science. Sci Data 2020;7:1–12. doi:10.1038/s41597-020-00637-5.

[75] Xiong S, Wang L. Research progress and development trends of materials genome technology. Adv Mater Sci Eng 2020;2020. doi:10.1155/2020/5903457.

[76] Raccuglia P, Elbert KC, Adler PDF, Falk C, Wenny MB, Mollo A, et al. Machine-learning-assisted materials discovery using failed experiments. Nature 2016;533:73–6. doi:10.1038/nature17439.

[77] Materials Genome Engineering Databases, https://www.mgedata.cn/help. [Accessed December 2020].

[78] Tong Q, Gao P, Liu H, Xie Y, Lv J, Wang Y, et al. Combining machine learning potential and structure prediction for accelerated materials design and discovery. J Phys Chem Lett 2020;11:8710–20. doi:10.1021/acs.jpclett.0c02357.

[79] Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: a new perspective. Neurocomputing 2018;300:70–9. doi:10.1016/j.neucom.2017.11.077.

[80] Khalid S, Khalil T, Nasreen S. A survey of feature selection and feature extraction techniques in machine learning. In: Proceedings of the Science and Information (SAI) Conference, 2014; 2014. p. 372–8. doi:10.1109/SAI.2014.6918213.

[81] Schmidt J, Marques MRG, Botti S, Marques MAL. Recent advances and applications of machine learning in solid-state materials science. Npj Comput Mater 2019;5. doi:10.1038/s41524-019-0221-0.

[82] Mangal A, Holm EA. A comparative study of feature selection methods for stress hotspot classification in materials. Integr Mater Manuf Innov 2018;7:87–95. doi:10.1007/s40192-018-0109-8.

[83] Duangsoithong R, Windeatt T. Correlation-based and causal feature selection analysis for ensemble classifiers. Lect Notes Comput Sci Incl Subser Lect Notes Artif Intell Lect Notes Bioinform 2010;5998 LNAI:25–36. doi:10.1007/978-3-642-12159-3_3.

[84] Sanchez-Pinto LN, Venable LR, Fahrenbach J, Churpek MM. Comparison of variable selection methods for clinical predictive modeling. Int J Med Inform 2018;116:10–17. doi:10.1016/j.ijmedinf.2018.05.006.

[85] Naseriparsa M, Bidgoli A-M, Varaee T. A hybrid feature selection method to improve performance of a group of classification algorithms. Int J Comput Appl 2013;69:28–35. doi:10.5120/12065-8172.

[86] Schmidhuber J. Deep Learning in neural networks: an overview. Neural Netw 2015;61:85–117. doi:10.1016/j.neunet.2014.09.003.

[87] Sha W, Guo Y, Yuan Q, Tang S, Zhang X, Lu S, et al. Artificial intelligence to power the future of materials science and engineering. Adv Intell Syst 2020;2:1900143. doi:10.1002/aisy.201900143.

[88] Isayev O, Fourches D, Muratov EN, Oses C, Rasch K, Tropsha A, et al. Materials cartography: Representing and mining materials space using structural and electronic fingerprints. Chem Mater 2015;27:735–43. doi:10.1021/cm503507h.

[89] Warren JA. The materials genome initiative and artificial intelligence. MRS Bull 2018;43:452–7. doi:10.1557/mrs.2018.122.

[90] Patel P, Ong SP. Artificial intelligence is aiding the search for energy materials. MRS Bull 2019;44:162–3. doi:10.1557/mrs.2019.51.

[91] Isayev O, Oses C, Toher C, Gossett E, Curtarolo S, Tropsha A. Universal fragment descriptors for predicting properties of inorganic crystals. Nat Commun 2017;8:1–12. doi:10.1038/ncomms15679.

[92] Ball P. Using artificial intelligence to accelerate materials development. MRS Bull 2019;44:335–43. doi:10.1557/mrs.2019.113.

[93] Celebi M.E., Aydin K. Unsupervised learning algorithms. 2016. doi: 10.1007/978-3-319-24211-8.

[94] Khatib MEl, de Jong WA. ML4Chem: a machine learning package for chemistry and materials science. ChemRxiv 2020. doi:10.26434/chemrxiv.11952516.v1.

[95] Gomes CP, Selman B, Gregoire JM. Artificial intelligence for materials discovery. MRS Bull 2019;44:538–44. doi:10.1557/mrs.2019.158.

[96] Meredig B, Antono E, Church C, Hutchinson M, Ling J, Paradiso S, et al. Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery. Mol Syst Des Eng 2018;3:819–25. doi:10.1039/c8me00012c.

[97] Samadi SH, Ghobadian B, Nosrati M. Prediction of higher heating value of biomass materials based on proximate analysis using gradient boosted regression trees method. Energy Sources A Recover Util Environ Eff 2021;43:672–81. doi:10.1080/15567036.2019.1630521.

[98] Sodeyama K, Igarashi Y, Nakayama T, Tateyama Y, Okada M. Liquid electrolyte informatics using an exhaustive search with linear regression. Phys Chem Chem Phys 2018;20:22585–91. doi:10.1039/c7cp08280k.

[99] Wang HY, Zhu R, Ma P. Optimal subsampling for large sample logistic regression. J Am Stat Assoc 2018;113:829–44. doi:10.1080/01621459.2017.1292914.

[100] JIANG F, GUAN Z, LI Z, WANG X. A method of predicting visual detectability of low-velocity impact damage in composite structures based on logistic regression model. Chin J Aeronaut 2021;34:296–308. doi:10.1016/j.cja.2020.10.006.

[101] Sendek AD, Yang Q, Cubuk ED, Duerloo KAN, Cui Y, Reed EJ. Holistic computational structure screening of more than 12 000 candidates for solid lithium-ion conductor materials. Energy Environ Sci 2017;10:306–20. doi:10.1039/c6ee02697d.

[102] Wee D, Kim J, Bang S, Samsonidze G, Kozinsky B. Quantification of uncertainties in thermoelectric properties of materials from a first-principles prediction method: an approach based on Gaussian process regression. Phys Rev Mater 2019;3:1–9. doi:10.1103/PhysRevMaterials.3.033803.

[103] Bassman L, Rajak P, Kalia RK, Nakano A, Sha F, Sun J, et al. Active learning for accelerated design of layered materials. Npj Comput Mater 2018;4:1–9. doi:10.1038/s41524-018-0129-0.

[104] Noack MM, Doerk GS, Li R, Streit JK, Vaia RA, Yager KG, et al. Autonomous materials discovery driven by Gaussian process regression with inhomogeneous measurement noise and anisotropic kernels. Sci Rep 2020;10:1–16. doi:10.1038/s41598-020-74394-1.

[105] Schulz E, Speekenbrink M, Krause A. A tutorial on Gaussian process regression: modelling, exploring, and exploiting functions. J Math Psychol 2018;85:1–16. doi:10.1016/j.jmp.2018.03.001.

[106] Okamoto Y, Kubo Y. Ab initio calculations of the redox potentials of additives for lithium-ion batteries and their prediction through machine learning. ACS Omega 2018;3:7868–74. doi:10.1021/acsomega.8b00576.

[107] Hellström M, Behler J. Neural network potentials in materials modeling. Handb Mater Model 2020:661–80. doi:10.1007/978-3-319-44677-6_56.

[108] Kim B, Lee S, Kim J. Inverse design in porous materials using artificial neural networks. ChemRxiv 2019:1–8. doi:10.26434/chemrxiv.7987475.v1.

[109] Bhadeshia HKDH. Neural networks in materials science. Encycl Mater Sci Technol 2008;39:1–5. doi:10.1016/b978-008043152-9.02201-6.

[110] Sha W, Edwards KL. The use of artificial neural networks in materials science based research. Mater Des 2007;28:1747–52. doi:10.1016/j.matdes.2007.02.009.

[111] Sainath TN, Vinyals O, Senior A, Sak H. Convolutional, long short-term memory, fully connected deep neural networks. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP); 2015. p. 4580–4. doi:10.1109/ICASSP.2015.7178838.

[112] Jalem R, Kimura M, Nakayama M, Kasuga T. Informatics-aided density functional theory study on the li ion transport of tavorite-type LiMTO4F (M3+-T5+, M2+-T6+). J Chem Inf Model 2015;55:1158–68. doi:10.1021/ci500752n.

[113] Sulzmann JN, Fürnkranz J, Hüllermeier E. On pairwise naive bayes classifiers. Lect Notes Comput Sci Incl Subser Lect Notes Artif Intell Lect Notes Bioinform 2007;4701 LNAI:371–81. doi:10.1007/978-3-540-74958-5_35.

[114] Nakayama M, Kanamori K, Nakano K, Jalem R, Takeuchi I, Yamasaki H. Data-driven materials exploration for li-ion conductive ceramics by exhaustive and informatics-aided computations. Chem Rec 2019;19:771–8. doi:10.1002/tcr.201800129.

[115] Lu WC, Ji XB, Li MJ, Liu L, Yue BH, Zhang LM. Using support vector machine for materials design. Adv Manuf 2013;1:151–9. doi:10.1007/s40436-013-0025-2.

[116] Kireeva N, Pervov VS. Materials space of solid-state electrolytes: unraveling chemical composition-structure-ionic conductivity relationships in garnet-type metal oxides using cheminformatics virtual screening approaches. Phys Chem Chem Phys 2017;19:20904–18. doi:10.1039/c7cp00518k.

[117] Balachandran PV, Theiler J, Rondinelli JM, Lookman T. Materials prediction via classification learning. Sci Rep 2015;5:1–16. doi:10.1038/srep13285.

[118] Li Y. Predicting materials properties and behavior using classification and regression trees. Mater Sci Eng A 2006;433:261–8. doi:10.1016/j.msea.2006.06.100.

[119] Shi Z, Yang W, Deng X, Cai C, Yan Y, Liang H, et al. Machine-learning-assisted high-throughput computational screening of high performance metal-organic frameworks. Mol Syst Des Eng 2020;5:725–42. doi:10.1039/d0me00005a.

[120] Lever J, Krzywinski M, Altman N. Points of significance: principal component analysis. Nat Methods 2017;14:641–2. doi:10.1038/nmeth.4346.

[121] Garg A, Yun L, Gao L, Putungan DB. Development of recycling strategy for large stacked systems: experimental and machine learning approach to form reuse battery packs for secondary applications. J Clean Prod 2020;275:124152, . doi:10.1016/j.jclepro.2020.124152.

[122] Jang B, Kim M, Harerimana G, Kim JW. Q-learning algorithms: a comprehensive classification and applications. IEEE Access 2019;7:133653–67. doi:10.1109/ACCESS.2019.2941229.

[123] Liu H, Cheng J, Dong H, Feng J, Pang B, Tian Z, et al. Screening stable and metastable ABO3 perovskites using machine learning and the materials project. Comput Mater Sci 2020;177:109614. doi:10.1016/j.commatsci.2020.109614.

[124] Suzuki K, Suzuki K, Suzuki K, Ohura K, Seko A, Seko A, et al. Fast material search of lithium ion conducting oxides using a recommender system. J Mater Chem A 2020;8:11582–8. doi:10.1039/d0ta02556a.

[125] Palkovits SA. Primer about machine learning in catalysis – a tutorial with code. ChemCatChem 2020;12:3995–4008. doi:10.1002/cctc.202000234.

[126] Maleki F, Muthukrishnan N, Ovens K, Reinhold C, Forghani R. Machine learning algorithm validation: from essentials to advanced applications and implications for regulatory certification and deployment. Neuroimaging Clin N Am 2020;30:433–45. doi:10.1016/j.nic.2020.08.004.

[127] Sutton C, Boley M, Ghiringhelli LM, Rupp M, Vreeken J, Scheffler M. Identifying domains of applicability of machine learning models for materials science. Nat Commun 2020;11:1–9. doi:10.1038/s41467-020-17112-9.

[128] Huang JS, Liew JX, Ademiloye AS, Liew KM. Artificial intelligence in materials modeling and design. Arch Comput Methods Eng 2020. doi:10.1007/s11831-020-09506-1.

[129] Vasudevan RK, Choudhary K, Mehta A, Smith R, Kusne G, Tavazza F, et al. Materials science in the artificial intelligence age: high-throughput library generation, machine learning, and a pathway from correlations to the underpinning physics. MRS Commun 2019;9:821–38. doi:10.1557/mrc.2019.95.

[130] Pankajakshan P, Sanyal S, De Noord OE, Bhattacharya I, Bhattacharyya A, Waghmare U. Machine learning and statistical analysis for materials science: stability and transferability of fingerprint descriptors and chemical insights. Chem Mater 2017;29:4190–201. doi:10.1021/acs.chemmater.6b04229.

[131] Kauwe SK, Rhone TD, Sparks TD. Data-driven studies of li-ion-battery materials. Crystals 2019;9:1–9. doi:10.3390/cryst9010054.

[132] Li G, Huang B, Pan Z, Su X, Shao Z, An L. Advances in three-dimensional graphene-based materials: configurations, preparation and application in secondary metal (Li, Na, K, Mg, Al)-ion batteries. Energy Environ Sci 2019;12:2030–53. doi:10.1039/c8ee03014f.

[133] Zhou M, Gallegos A, Liu K, Dai S, Wu J. Insights from machine learning of carbon electrodes for electric double layer capacitors. Carbon N Y 2020;157:147–52. doi:10.1016/j.carbon.2019.08.090.

[134] Aykol M, Herring P, Anapolsky A. Machine learning for continuous innovation in battery technologies. Nat Rev Mater 2020;5:725–7. doi:10.1038/s41578-020-0216-y.

[135] Deringer VL. Modelling and understanding battery materials with machine-learning-driven atomistic simulations. J Phys Energy 2020;2:041003, . doi:10.1088/2515-7655/abb011.

[136] Marom R, Amalraj SF, Leifer N, Jacob D, Aurbach D. A review of advanced and practical lithium battery materials. J Mater Chem 2011;21:9938–54. doi:10.1039/c0jm04225k.

[137] Liu K, Wei Z, Yang Z, Li K. Mass load prediction for lithium-ion battery electrode clean production: a machine learning approach. J Clean Prod 2020:125159, . doi:10.1016/j.jclepro.2020.125159.

[138] Van Der Ven A, Deng Z, Banerjee S, Ong SP. Rechargeable alkali-ion battery materials: theory and computation. Chem Rev 2020;120:6977–7019. doi:10.1021/acs.chemrev.9b00601.

[139] Van Duong M, Van Tran M, Garg A, Van Nguyen H, Huynh TTK, Phung Le ML. Machine learning approach in exploring the electrolyte additives effect on cycling performance of LiNi0.5Mn1.5O4 cathode and graphite anode-based lithium-ion cell. Int J Energy Res 2020;4:1–12. doi:10.1002/er.6074.

[140] Kim S, Jinich A, Aspuru-Guzik A. MultiDK: a multiple descriptor multiple kernel approach for molecular discovery and its application to organic flow battery electrolytes. J Chem Inf Model 2017;57:657–68. doi:10.1021/acs.jcim.6b00332.

[141] Ishikawa A, Sodeyama K, Igarashi Y, Nakayama T, Tateyama Y, Okada M. Machine learning prediction of coordination energies for alkali group elements in battery electrolyte solvents. Phys Chem Chem Phys 2019;21:26399–405. doi:10.1039/c9cp03679b.

[142] Dave A, Mitchell J, Kandasamy K, Burke S, Paria B, Poczos B, et al. Autonomous discovery of battery electrolytes with robotic experimentation and machine-learning. ArXiv 2019. doi:10.1016/j.xcrp.2020.100264.

[143] Wang C, Aoyagi K, Wisesa P, Mueller T. Lithium ion conduction in cathode coating materials from on-the-fly machine learning. Chem Mater 2020;32:3741–52. doi:10.1021/acs.chemmater.9b04663.

[144] Deng Z, Zhu Z, Chu IH, Ong SP. Data-driven first-principles methods for the study and design of alkali superionic conductors. Chem Mater 2017;29:281–8. doi:10.1021/acs.chemmater.6b02648.

[145] Zhang Y, He X, Chen Z, Bai Q, Nolan AM, Roberts CA, et al. Unsupervised discovery of solid-state lithium ion conductors. Nat Commun 2019;10:1–7. doi:10.1038/s41467-019-13214-1.

[146] Wang Y, Xie T, France-Lanord A, Berkley A, Johnson JA, Shao-Horn Y, et al. Toward designing highly conductive polymer electrolytes by machine learning assisted coarse-grained molecular dynamics. Chem Mater 2020;32:4144–51. doi:10.1021/acs.chemmater.9b04830.

[147] Houchins G, Viswanathan V. An accurate machine-learning calculator for optimization of Li-ion battery cathodes. J Chem Phys 2020;153. doi:10.1063/5.0015872.

[148] Attarian Shandiz M, Gauvin R. Application of machine learning methods for the prediction of crystal system of cathode materials in lithium-ion batteries. Comput Mater Sci 2016;117:270–8. doi:10.1016/j.commatsci.2016.02.021.

[149] Wang X, Xiao R, Li H, Chen L. Quantitative structure-property relationship study of cathode volume changes in lithium ion batteries using ab-initio and partial least squares analysis. J Mater 2017;3:178–83. doi:10.1016/j.jmat.2017.02.002.

[150] Allam O, Cho BW, Kim KC, Jang SS. Application of DFT-based machine learning for developing molecular electrode materials in Li-ion batteries. RSC Adv 2018;8:39414–20. doi:10.1039/c8ra07112h.

[151] Takagishi Y, Yamanaka T, Yamaue T. Machine learning approaches for designing mesoscale structure of li-ion battery electrodes. Batteries 2019;5. doi:10.3390/batteries5030054.

[152] Joshi RP, Eickholt J, Li L, Fornari M, Barone V, Peralta JE. Machine learning the voltage of electrode materials in metal-ion batteries. ACS Appl Mater Interfaces 2019;11:18494–503. doi:10.1021/acsami.9b04933.

[153] Jiang Z, Li J, Yang Y, Mu L, Wei C, Yu X, et al. Machine-learning-revealed statistics of the particle-carbon/binder detachment in lithium-ion battery cathodes. Nat Commun 2020;11. doi:10.1038/s41467-020-16233-5.

[154] Turetskyy A, Thiede S, Thomitzek M, von Drachenfels N, Pape T, Herrmann C. Toward data-driven applications in lithium-ion battery cell manufacturing. Energy Technol 2020;8:1–11. doi:10.1002/ente.201900136.

[155] Kilic A, Odabaşı Ç, Yildirim R, Eroglu D. Assessment of critical materials and cell design factors for high performance lithium-sulfur batteries using machine learning. Chem Eng J 2020;390. doi:10.1016/j.cej.2020.124117.

[156] Lee S, Kim Y. Li-ion battery electrode health diagnostics using machine learning. In: Proceedings of the American Control Conference; 2020. p. 1137–42. doi:10.23919/ACC45564.2020.9147633.

[157] Zhang L, He M, Shao S. Machine learning for halide perovskite materials. Nano Energy 2020;78. doi:10.1016/j.nanoen.2020.105380.

[158] Odabaşı Ç, Yıldırım R. Machine learning analysis on stability of perovskite solar cells. Sol Energy Mater Sol Cells 2020;205. doi:10.1016/j.solmat.2019.110284.

[159] Sutherland BR. Solar materials find their band gap. Joule 2020;4:984–5. doi:10.1016/j.joule.2020.05.001.

[160] Feng H-J, Wu K, Deng Z-Y. Predicting inorganic photovoltaic materials with efficiencies >26% via structure-relevant machine learning and density functional calculations. Cell Reports Phys Sci 2020;1:100179, . doi:10.1016/j.xcrp.2020.100179.

[161] Yılmaz B, Yıldırım R. Critical review of machine learning applications in perovskite solar research. Nano Energy 2021;80:105546, . doi:10.1016/j.nanoen.2020.105546.

[162] Li F, Peng X, Wang Z, Zhou Y, Wu Y, Jiang M, et al. Machine learning (ML)-assisted design and fabrication for solar cells. Energy Environ Mater 2019;2:280–91. doi:10.1002/eem2.12049.

[163] Howard JM, Tennyson EM, Neves BRA, Leite MS. Machine learning for perovskites' reap-rest-recovery cycle. Joule 2019;3:325–37. doi:10.1016/j.joule.2018.11.010.

[164] Saidi WA, Shadid W, Castelli IE. Machine-learning structural and electronic properties of metal halide perovskites using a hierarchical convolutional neural network. Npj Comput Mater 2020;6:1–7. doi:10.1038/s41524-020-0307-8.

[165] Choudhary K, Bercx M, Jiang J, Pachter R, Lamoen D, Tavazza F. Accelerated discovery of efficient solar cell materials using quantum and machine-learning methods. Chem Mater 2019;31:5900–8. doi:10.1021/acs.chemmater.9b02166.

[166] Sun Q, Yin WJ. Thermodynamic stability trend of cubic perovskites. J Am Chem Soc 2017;139:14905–8. doi:10.1021/jacs.7b09379.

[167] Pilania G, Balachandran PV, Kim C, Lookman T. Finding new perovskite halides via machine learning. Front Mater 2016;3:1–7. doi:10.3389/fmats.2016.00019.

[168] Im J, Lee S, Ko TW, Kim HW, Hyon YK, Chang H. Identifying Pb-free perovskites for solar cells by machine learning. Npj Comput Mater 2019;5:1–8. doi:10.1038/s41524-019-0177-0.

[169] Jin H, Zhang H, Li J, Wang T, Wan L, Guo H, et al. Discovery of novel two-dimensional photovoltaic materials accelerated by machine learning. J Phys Chem Lett 2020;11:3075–81. doi:10.1021/acs.jpclett.0c00721.

[170] Lu S, Zhou Q, Ouyang Y, Guo Y, Li Q, Wang J. Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. Nat Commun 2018;9:1–8. doi:10.1038/s41467-018-05761-w.

[171] Schmidt J, Shi J, Borlido P, Chen L, Botti S, Marques MAL. Predicting the thermodynamic stability of solids combining density functional theory and machine learning. Chem Mater 2017;29:5090–103. doi:10.1021/acs.chemmater.7b00156.

[172] Takahashi K, Takahashi L, Miyazato I, Tanaka Y. Searching for hidden perovskite materials for photovoltaic systems by combining data science and first principle calculations. ACS Photon 2018;5:771–5. doi:10.1021/acsphotonics.7b01479.

[173] Pilania G, Gubernatis JE, Lookman T. Multi-fidelity machine learning models for accurate bandgap predictions of solids. Comput Mater Sci 2017;129:156–63. doi:10.1016/j.commatsci.2016.12.004.

[174] Min K, Cho E. Accelerated discovery of potential ferroelectric perovskite: via active learning. J Mater Chem C 2020;8:7866–72. doi:10.1039/d0tc00985g.

[175] Liu S, Yuan J, Deng W, Luo M, Xie Y, Liang Q, et al. High-efficiency organic solar cells with low non-radiative recombination loss and low energetic disorder. Nat Photonics 2020;14:300–5. doi:10.1038/s41566-019-0573-5.

[176] Mahmood A, Wang J-L. Machine learning for high performance organic solar cells: current scenario and future prospects. Energy Environ Sci 2021. doi:10.1039/d0ee02838j.

[177] Meftahi N, Klymenko M, Christofferson AJ, Bach U, Winkler DA, Russo SP. Machine learning property prediction for organic photovoltaic devices. Npj Comput Mater 2020;6:1–8. doi:10.1038/s41524-020-00429-w.

[178] Sun W, Li M, Li Y, Wu Z, Sun Y, Lu S, et al. The use of deep learning to fast evaluate organic photovoltaic materials. Adv Theory Simul 2019;2:1–9. doi:10.1002/adts.201800116.

[179] Paul A, Jha D, Al-Bahrani R, Liao WK, Choudhary A, Agrawal A. Transfer learning using ensemble neural networks for organic solar cell screening. ArXiv 2019:1–8.

[180] Nagasawa S, Al-Naamani E, Saeki A. Computer-aided screening of conjugated polymers for organic solar cell: classification by random forest. J Phys Chem Lett 2018;9:2639–46. doi:10.1021/acs.jpclett.8b00635.

[181] Padula D, Simpson JD, Troisi A. Combining electronic and structural features in machine learning models to predict organic solar cells properties. Mater Horiz 2019;6:343–9. doi:10.1039/c8mh01135d.

[182] Sahu H, Yang F, Ye X, Ma J, Fang W, Ma H. Designing promising molecules for organic solar cells: via machine learning assisted virtual screening. J Mater Chem A 2019;7:17480–8. doi:10.1039/c9ta04097h.

[183] Saal JE, Oliynyk AO, Meredig B. Machine learning in materials discovery: confirmed predictions and their underlying approaches. Annu Rev Mater Res 2020;50:49–69. doi:10.1146/annurev-matsci-090319-010954.

[184] Sun W, Zheng Y, Yang K, Zhang Q, Shah AA, Wu Z, et al. Machine learning–assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials. Sci Adv 2019;5:1–9. doi:10.1126/sciadv.aay4275.

[185] Wu Y, Guo J, Sun R, Min J. Machine learning for accelerating the discovery of high-performance donor/acceptor pairs in non-fullerene organic solar cells. Npj Comput Mater 2020;6:1–8. doi:10.1038/s41524-020-00388-2.

[186] Marchenko EI, Fateev SA, Petrov AA, Korolev VV, Mitrofanov A, Petrov AV, et al. Database of two-dimensional hybrid perovskite materials: open-access collection of crystal structures, band gaps, and atomic partial charges predicted by machine learning. Chem Mater 2020;32:7383–8. doi:10.1021/acs.chemmater.0c02290.

[187] Singh S, Pareek M, Changotra A, Banerjee S, Bhaskararao B, Balamurugan P, et al. A unified machine-learning protocol for asymmetric catalysis as a proof of concept demonstration using asymmetric hydrogenation. Proc Natl Acad Sci USA 2020;117:1339–45. doi:10.1073/pnas.1916392117.

[188] Orupattur NV, Mushrif SH, Prasad V. Catalytic materials and chemistry development using a synergistic combination of machine learning and ab initio methods. Comput Mater Sci 2020;174:109474, . doi:10.1016/j.commatsci.2019.109474.

[189] Erdem Günay M, Yıldırım R. Recent advances in knowledge discovery for heterogeneous catalysis using machine learning. Catal Rev Sci Eng 2020;00:1–45. doi:10.1080/01614940.2020.1770402.

[190] Artrith N, Lin Z, Chen JG. Predicting the activity and selectivity of bimetallic metal catalysts for ethanol reforming using machine learning. ACS Catal 2020;10:9438–44. doi:10.1021/acscatal.0c02089.

[191] Goldsmith BR, Esterhuizen J, Liu JX, Bartel CJ, Sutton C. Machine learning for heterogeneous catalyst design and discovery. AIChE J 2018;64:2311–23. doi:10.1002/aic.16198.

[192] Yang W, Fidelis TT, Sun WH. Machine learning in catalysis, from proposal to practicing. ACS Omega 2020;5:83–8. doi:10.1021/acsomega.9b03673.

[193] Li Z, Achenie LEK, Xin H. An adaptive machine learning strategy for accelerating discovery of perovskite electrocatalysts. ACS Catal 2020;10:4377–84. doi:10.1021/acscatal.9b05248.

[194] Li XT, Chen L, Wei GF, Shang C, Liu ZP. Sharp increase in catalytic selectivity in acetylene semihydrogenation on pd achieved by a machine learning simulation-guided experiment. ACS Catal 2020;10:9694–705. doi:10.1021/acscatal.0c02158.

[195] Choi S, Kim Y, Kim JW, Kim Z, Kim WY. Feasibility of activation energy prediction of gas-phase reactions by machine learning. Chem A Eur J 2018;24:12354–8. doi:10.1002/chem.201800345.

[196] Toyao T, Suzuki K, Kikuchi S, Takakusagi S, Shimizu KI, Takigawa I. Toward effective utilization of methane: machine learning prediction of adsorption energies on metal alloys. J Phys Chem C 2018;122:8315–26. doi:10.1021/acs.jpcc.7b12670.

[197] Ma X, Li Z, Achenie LEK, Xin H. Machine-learning-augmented chemisorption model for CO2 electroreduction catalyst screening. J Phys Chem Lett 2015;6:3528–33. doi:10.1021/acs.jpclett.5b01660.

[198] Chen Y, Huang Y, Cheng T, Goddard WA. Identifying active sites for CO2 reduction on dealloyed gold surfaces by combining machine learning with multiscale simulations. J Am Chem Soc 2019;141:11651–7. doi:10.1021/jacs.9b04956.

[199] Meyer B, Sawatlon B, Heinen S, Von Lilienfeld OA, Corminboeuf C. Machine learning meets volcano plots: computational discovery of cross-coupling catalysts. Chem Sci 2018;9:7069–77. doi:10.1039/c8sc01949e.

[200] McCullough K, Williams T, Mingle K, Jamshidi P, Lauterbach J. High-throughput experimentation meets artificial intelligence: a new pathway to catalyst discovery. Phys Chem Chem Phys 2020;22:11174–96. doi:10.1039/d0cp00972e.

[201] Rück M, Garlyyev B, Mayr F, Bandarenka AS, Gagliardi A. Oxygen reduction activities of strained platinum core-shell electrocatalysts predicted by machine learning. J Phys Chem Lett 2020;11:1773–80. doi:10.1021/acs.jpclett.0c00214.

[202] Ge L, Yuan H, Min Y, Li L, Chen S, Xu L, et al. Predicted optimal bifunctional electrocatalysts for the hydrogen evolution reaction and the oxygen evolution reaction using chalcogenide heterostructures based on machine learning analysis of in silico quantum mechanics based high throughput screening. J Phys Chem Lett 2020;11:869–76. doi:10.1021/acs.jpclett.9b03875.

[203] Zheng J, Sun X, Qiu C, Yan Y, Yao Z, Deng S, et al. High-throughput screening of hydrogen evolution reaction catalysts in MXene materials. J Phys Chem C 2020;124:13695–705. doi:10.1021/acs.jpcc.0c02265.

[204] Sun M, Dougherty AW, Huang B, Li Y, Yan CH. Accelerating atomic catalyst discovery by theoretical calculations-machine learning strategy. Adv Energy Mater 2020;10:1–10. doi:10.1002/aenm.201903949.

[205] Yang Z, Gao W, Jiang Q. A machine learning scheme for the catalytic activity of alloys with intrinsic descriptors. J Mater Chem A 2020;8:17507–15. doi:10.1039/d0ta06203k.

[206] Zhong M, Tran K, Min Y, Wang C, Wang Z, Dinh CT, et al. Accelerated discovery of CO2 electrocatalysts using active machine learning. Nature 2020;581:178–83. doi:10.1038/s41586-020-2242-8.

[207] Li G, Yu Y, Pan Z, An L. Two-dimensional layered SnO2 nanosheets for ambient ammonia synthesis. ACS Appl Energy Mater 2020;3:6735–42. doi:10.1021/acsaem.0c00858.

[208] Zafari M, Kumar D, Umer M, Kim KS. Machine learning-based high throughput screening for nitrogen fixation on boron-doped single atom catalysts. J Mater Chem A 2020;8:5209–16. doi:10.1039/c9ta12608b.

[209] Chen A, Zhang X, Chen L, Yao S, Zhou Z. A machine learning model on simple features for CO2 reduction electrocatalysts. J Phys Chem C 2020;124:22471–8. doi:10.1021/acs.jpcc.0c05964.

[210] Sun X, Zheng J, Gao Y, Qiu C, Yan Y, Yao Z, et al. Machine-learning-accelerated screening of hydrogen evolution catalysts in MBenes materials. Appl Surf Sci 2020;526:146522, . doi:10.1016/j.apsusc.2020.146522.

[211] Dasgupta A, Gao Y, Broderick SR, Pitman EB, Rajan K. Machine learning-aided identification of single atom alloy catalysts. J Phys Chem C 2020;124:14158–66. doi:10.1021/acs.jpcc.0c01492.

[212] Melisande Fischer J, Hunter M, Hankel M, Searles DJ, Parker AJ, Barnard AS. Accurate prediction of binding energies for two-dimensional catalytic materials using machine learning. ChemCatChem 2020;12:5109–20. doi:10.1002/cctc.202000536.

[213] Smith A, Keane A, Dumesic JA, Huber GW, Zavala VM. A machine learning framework for the analysis and prediction of catalytic activity from experimental data. Appl Catal B Environ 2020;263:118257, . doi:10.1016/j.apcatb.2019.118257.

[214] Toyao T, Maeno Z, Takakusagi S, Kamachi T, Takigawa I, Shimizu KI. Machine learning for catalysis informatics: recent applications and prospects. ACS Catal 2020;10:2260–97. doi:10.1021/acscatal.9b04186.

[215] Deng X, Yang W, Li S, Liang H, Shi Z, Qiao Z. Large-scale screening and machine learning to predict the computation-ready, experimental metal-organic frameworks for co2 capture from air. Appl Sci 2020;10. doi:10.3390/app10020569.

[216] Anderson R, Rodgers J, Argueta E, Biong A, Gómez-Gualdrón DA. Role of pore chemistry and topology in the CO2 capture capabilities of MOFs: from molecular simulation to machine learning. Chem Mater 2018;30:6325–37. doi:10.1021/acs.chemmater.8b02257.

[217] He Y, Cubuk ED, Allendorf MD, Reed EJ. Metallic metal-organic frameworks predicted by the combination of machine learning methods and ab initio calculations. J Phys Chem Lett 2018;9:4562–9. doi:10.1021/acs.jpclett.8b01707.

[218] Fernandez M, Boyd PG, Daff TD, Aghaji MZ, Woo TK. Rapid and accurate machine learning recognition of high performing metal organic frameworks for CO2 capture. J Phys Chem Lett 2014;5:3056–60. doi:10.1021/jz501331m.

[219] Zhu X, Tsang DCW, Wang L, Su Z, Hou D, Li L, et al. Machine learning exploration of the critical factors for CO2 adsorption capacity on porous carbon materials at different pressures. J Clean Prod 2020;273:122915, . doi:10.1016/j.jclepro.2020.122915.

[220] Chong S, Lee S, Kim B, Kim J. Applications of machine learning in metal-organic frameworks. Coord Chem Rev 2020;423:213487, . doi:10.1016/j.ccr.2020.213487.

[221] Kim E, Huang K, Kononova O, Ceder G, Olivetti E. Distilling a materials synthesis ontology. Matter 2019;1:8–12. doi:10.1016/j.matt.2019.05.011.

[222] Botu V, Batra R, Chapman J, Ramprasad R. Machine learning force fields: construction, validation, and outlook. J Phys Chem C 2017;121:511–22. doi:10.1021/acs.jpcc.6b10908.

[223] Kusne AG, Yu H, Wu C, Zhang H, Hattrick-Simpers J, DeCost B, et al. On-the-fly closed-loop materials discovery via Bayesian active learning. Nat Commun **11**, 2020, 1–11, https://doi.org/10.1038/s41467-020-19597-w.

[224] Burger B, Maffettone PM, Gusev VV, Aitchison CM, Bai Y, Wang X, et al. A mobile robotic chemist. Nature 2020;583:237–41. doi:10.1038/s41586-020-2442-2.