Hu *et al. J Mater Inf* 2023;3:18
**DOI:** 10.20517/jmi.2023.23

**Journal of Materials Informatics**

**Review**

**Open Access**

Check for updates

# Recent advances and applications of machine learning in electrocatalysis

**You Hu[1], Junhua Chen[1], Zheng Wei[2], Qiu He[1,*], Yan Zhao[1,3,*]**

[1]College of Materials Science and Engineering, Sichuan University, Chengdu 610065, Sichuan, China.
[2]International School of Materials Science and Engineering, Wuhan University of Technology, Wuhan 430070, Hubei, China.
[3]The Institute of Technological Sciences, Wuhan University, Wuhan 430072, Hubei, China.

[*]**Correspondence to:** Prof. Yan Zhao, Dr. Qiu He, College of Materials Science and Engineering, Sichuan University, 24 Yihuan Road, Chengdu 610065, Sichuan, China. E-mail: yanzhao@scu.edu.cn; hq5220@scu.edu.cn

**How to cite this article:** Hu Y, Chen J, Wei Z, He Q, Zhao Y. Recent advances and applications of machine learning in electrocatalysis. *J Mater Inf* 2023;3:18. https://dx.doi.org/10.20517/jmi.2023.23

**Received:** 8 Jun 2023 **First Decision:** 10 Jul 2023 **Revised:** 31 Jul 2023 **Accepted:** 28 Aug 2023 **Published:** 31 Aug 2023

**Academic Editor:** Xingjun Liu **Copy Editor:** Dong-Li Li **Production Editor:** Dong-Li Li

## Abstract

Electrocatalysis plays an important role in the production of clean energy and pollution control. Researchers have made great efforts to explore efficient, stable, and inexpensive electrocatalysts. However, traditional trial and error experiments and theoretical calculations require a significant amount of time and resources, which limits the development speed of electrocatalysts. Fortunately, the rapid development of machine learning (ML) has brought new solutions to scientific problems and new paradigms to the development of electrocatalysts. The combination of ML with experimental and theoretical calculations has propelled significant advancements in electrocatalysis research, particularly in the areas of materials screening, performance prediction, and catalysis theory development. In this review, we present a comprehensive overview of the workflow and cutting-edge techniques of ML in the field of electrocatalysis. In addition, we discuss the diverse applications of ML in predicting performance, guiding synthesis, and exploring the theory of catalysis. Finally, we conclude the review with the challenges of ML in electrocatalysis.

**Keywords:** Machine learning, electrocatalysis, performance prediction

## INTRODUCTION

The concept of electrocatalysis was originally a branch of electrochemistry, and after nearly a century of

development, it has become a multidisciplinary subject, including chemistry, solid-state physics, materials science, and other fields. Currently, electrocatalysis is widely used in important technological fields such as energy conversion and storage, environmental pollution control, and the synthesis of green materials. On the other hand, with the depletion of fossil fuels and the increasing environmental pollution caused by their consumption, finding sustainable and clean energy sources to pursue energy transformation and development has become one of the primary goals of scientific research. Therefore, electrocatalysis has received significant attention due to its critical role in these studies[1,2].

The factors that influence electrochemical reactions are multifaceted, with catalysts being the core among them. In addition, the development of inexpensive, efficient, and durable catalysts for specific reactions has always been the primary task of electrocatalysis research. However, traditional empirical experimental research methods suffer from the drawbacks of being time-consuming, costly, and inefficient[3,4]. Theoretical models and generalized paradigms, represented by thermodynamic laws, have laid the theoretical foundation for material research, making it no longer purely empirical. However, with the deepening of scientific research, the theoretical models become increasingly complex and difficult to solve practical problems[5]. By the mid-20th century, with the rapid development of supercomputers and various theoretical calculation methods, including the density functional theory (DFT)[6,7] and molecular dynamics (MD)[8], the physics-based simulation became an important tool for guiding material design[9,10]. However, these methods still face problems such as insufficient consideration of experimental conditions, hypothetical structures without thermodynamic stability, and high computational costs[11].

Although the above-mentioned three paradigms have inherent limitations, they are still the mainstream research methods in various scientific fields to the present day[5]. The application of these paradigms has generated a substantial volume of data. Recently, with the advancement of the Materials Genome Project[12] and the rapid development of artificial intelligence (AI) technology, the combination of big data and AI has emerged as the "fourth paradigm of science"[13]. Machine learning (ML) is a pivotal subfield of AI, which leverages diverse algorithms to construct models that uncover latent relationships in historical data. These models can then be utilized for data classification and prediction[14-16]. For example, with enough data of high quality, generative models in ML can be used to predict the closest material to the target material without having to blindly explore the vast chemical space[17]. Moreover, ML can also assist in the interpretation of complex experimental data and provide insights into the underlying mechanisms of material performance. Therefore, ML has been applied to many aspects of materials research, including guiding synthesis, assisting characterization, discovering novel material, and developing theoretical methods[18]. In this paper, we focus on the application of ML in electrocatalysis research. Figure 1 demonstrates that the development of ML-assisted electrocatalysis research is relatively recent and has garnered significantly increasing attention since 2019.

This review introduces ML, summarizes the latest progress of ML in the discovery and optimization of electrochemical catalysts, and discusses the challenges in this field. We provide a more comprehensive summary of specific approaches to ML-accelerated electrocatalysis research compared to the published reviews[19-22], in addition to introducing some new techniques that can help streamline the ML process. We believe that this review can provide researchers in related fields with a clearer understanding of ML-accelerated electrocatalysis research.

## ML WORKFLOW

Although Samuel[23] and Mitchell[24] have proposed successive definitions of ML, these definitions are currently not strictly recognized. Simply put, ML is an algorithm that can learn from data and improve

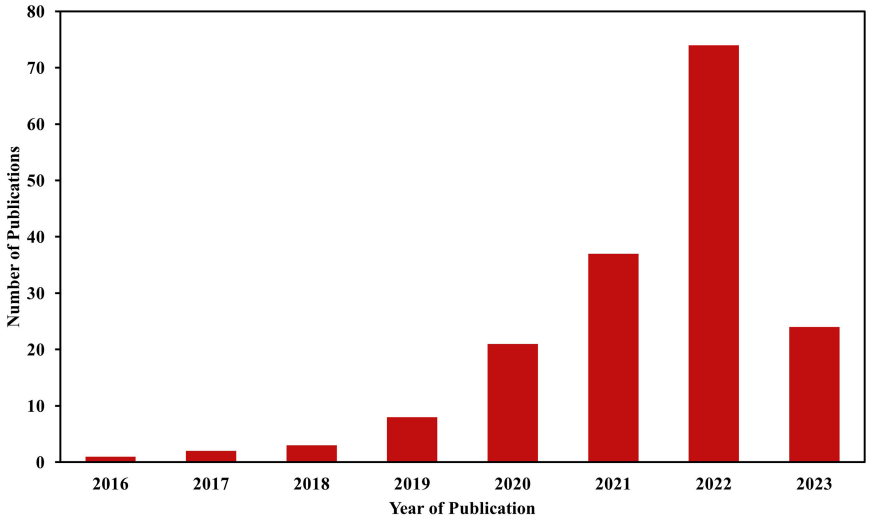


**Figure 1.** Statistics on publications combining electrocatalysis and ML from 2016 to 2023 that were gathered by conducting a search query with "electrocatalysis" and "machine learning" as keywords in the subject field on the Web of Science website. The data was accessed on July 29, 2023.

performance for a specific task. ML algorithms can predict functional relationships without explicit instructions, provide a mapping between inputs and corresponding outputs, or only provide relationships between inputs[25]. In theory, as long as the training data is sufficient and reliable, the computer can summarize the potential rules.

As shown in Figure 2, the ML process mainly consists of data collection, pre-processing, feature engineering, algorithm selection, model training, and model evaluation. Many of these processes are general techniques in the field of ML and are not unique to electrocatalysis and materials science. Therefore, this section is mostly a conceptual introduction to these processes, and the technical details can be obtained in specialized ML papers and books. Given that supervised learning is widely employed in the materials domain, it is naturally the primary focus of this review.

### Data collection

In ML research, data is the foundation upon which models are built, trained, and tested. The quantity and quality of data are crucial factors that determine the efficacy of a ML model. The data sources include material databases, experiments, theoretical calculations, and published literature. The development of material databases originated in the 1880s[26]. To date, various types of material databases have been established[21,27,28]. Table 1 summarizes some of the major databases in materials science. Databases have the advantage of providing different types of data (such as crystal structures, thermodynamic properties, and phase diagrams) on a wide range of materials quickly. However, the completeness of the recorded information in these databases, particularly for experimental databases, may be insufficient, and the lack of certain experimental conditions can hinder the user's comprehensive understanding of the material. Additionally, discrepancies may arise between data generated by various publications, experimental methods, and conditions. In contrast, literature sources typically provide detailed experimental methods and procedures, but data collection through literature is time-consuming and inefficient. The use of ML-based text extraction methods can effectively improve data collection efficiency[29,30], but the reliability of the paper still needs to be carefully evaluated. Generating new material characteristic data through experiments or theoretical calculations is also an important data collection method. This method can maximize the

**Table 1. List of commonly used databases for structures and property information of materials**

| Name | Brief information | Data source | URL |
|---|---|---|---|
| Materials Project (MP) | Properties of known and predicted materials | Calculation using standard calculation scheme | https://materialsproject.org/ |
| Open Quantum Materials Database (QQMD) | Thermodynamic and structural properties | DFT calculation | http://oqmd.org/ |
| AFLOWLIB | The database has millions of materials and can predict new crystal structures | High throughput calculation | http://aflowlib.org/ |
| ICSD | Inorganic Crystal Structure Database | Published structures | https://icsd.products.fiz-karlsruhe.de/en |
| Organic Materials Database | Electronic structure database for 3D organic crystals | Calculation | https://omdb.mathub.io/ |
| ZINC | 2D and 3D structures of commercially available molecules | Calculation | https://zinc15.docking.org/ |
| NREL Materials Database | Properties of materials for renewable energy applications (photovoltaics, materials for photoelectrochemical water splitting, thermoelectrics) | Calculation | https://materials.nrel.gov/ |
| Non-linear Optical Materials Database | Chemical formula, space group, and calculated band gap refractive index of the material | DFT calculation | http://nlo.hbu.cn |

DFT: Density functional theory.

control of variables (experimental methods and conditions or calculation methods). However, it is time-consuming, laborious, and expensive. It is worth noting that researchers are often reluctant to record or publish "failed" experimental data, but such data is also valuable for ML[31,32]. When training ML models, the inclusion of both successful and failed experimental data within the dataset can enhance the identification of the key determinants of material properties.

## Pre-processing

The pre-processing of datasets typically includes several steps, such as data cleaning, feature scaling, and dataset splitting. Data cleaning is designed to remove "dirty data" from a dataset, which includes duplicates, missing values, noise, inconsistencies, redundancies, and outliers in the database[33,34]. Young *et al.* confirmed in their research that there is a significant error rate in databases containing structural information, while even small errors in structural representation can result in substantial predictive inaccuracies[35]. Therefore, it is crucial to identify and address these problems during the data pre-processing stage in order to ensure the validity and reliability of the subsequent analysis[36-42]. Feature scaling, also known as data normalization, has two main purposes. Firstly, it maps the initial data range to a fixed interval to avoid large differences in the value range of different features. Secondly, feature scaling removes data dimensions and makes different features comparable to each other. It can accelerate the convergence speed of gradient descent algorithms[43]. Data splitting is an essential procedure to divide the original data into different sets, namely, the training set for training the model and the test set for evaluating the quality of the model[44]. Sometimes, it is also necessary to set aside validation sets for model tuning[45].

## Feature engineering

Material data cannot be directly recognized by a computer and needs to be encoded into computer-recognizable descriptors. As shown in Figure 3[46,47], the descriptors are obtained using different encoding methods. There are four representative methods for encoding crystal solids: structural diagrams, coulomb matrices, topological descriptors, and diffraction fingerprints[48-50]. Feature coding relies heavily on the expertise of the researcher, and manual coding also tends to lead to incompatibility and low interpretability of the model.

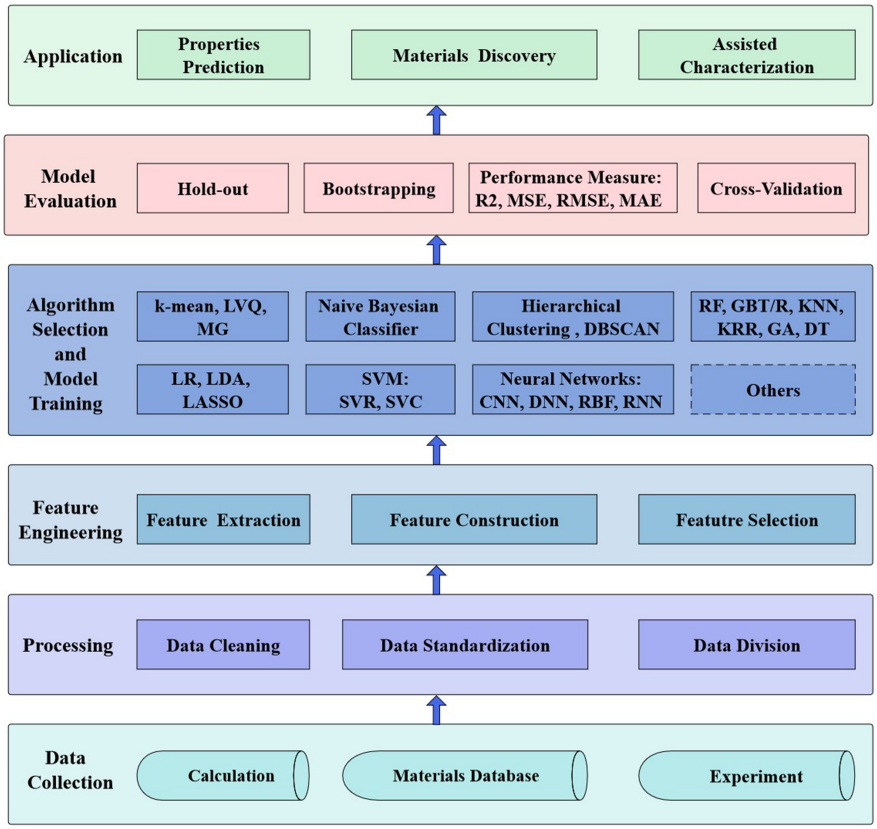


**Figure 2.** The workflow of ML. CNN: Convolutional neural network; DBSCAN: density-based spatial clustering of applications with noise; DNN: deep learning neural networks; DT: decision tree; GBR: gradient boosting regression; GBT: gradient boosting tree; KNN: k-nearest neighbor; KRR: kernel ridge regression; LASSO: least absolute shrinkage and selection operator; LDA: linear discriminant analysis; LR: linear regression; LVQ: learning vector quantization; MAE: mean absolute error; MG: mixture-of-Gaussian; ML: machine learning; MSE: mean square error; PCA: principal component analysis; RMSE: root mean square error; RNN: recurrent neural network; R2: R-square; SVC: support vector classification; SVM: support vector machines; SVR: support vector regression.

With the development of ML techniques, it is expected to automate the coding of atomic structures[51,52]. In particular, crystal graphical representations have attracted attention in recent years. In 2017, Isayev *et al.* published seminal results in which they proposed a descriptor called property-labeled material fragment (PLMF) [Figure 3B] for constructing a generalized property prediction model for inorganic crystalline materials[47]. One year later, Xie *et al.* developed a crystal graph convolutional neural network (CGCNN) framework, which can learn material properties from atomic connectivity in crystals, providing a generic and interpretable representation of materials[53]. The model can provide an approximate accuracy to DFT in the prediction of properties such as formation energy, band gap, and shear modulus. Since then, graphical representations of materials have been rapidly developed, and various graph network models have been proposed[54-56]. However, current graphical representations are more applicable to systems containing only rigid bonds. This is because the presence of flexible bonds causes small changes in the spacing of the atoms, making it impossible to determine the nearest atoms[50].

In addition to structural coding, the choice of descriptors is critical. Owing to the diversity of data types, a large number of descriptors are usually generated from the collected data. However, not all descriptors have utility value for specific problems. The selection of suitable descriptors for model training is of paramount importance in addressing specific scientific problems. An appropriate descriptor set can speed up model

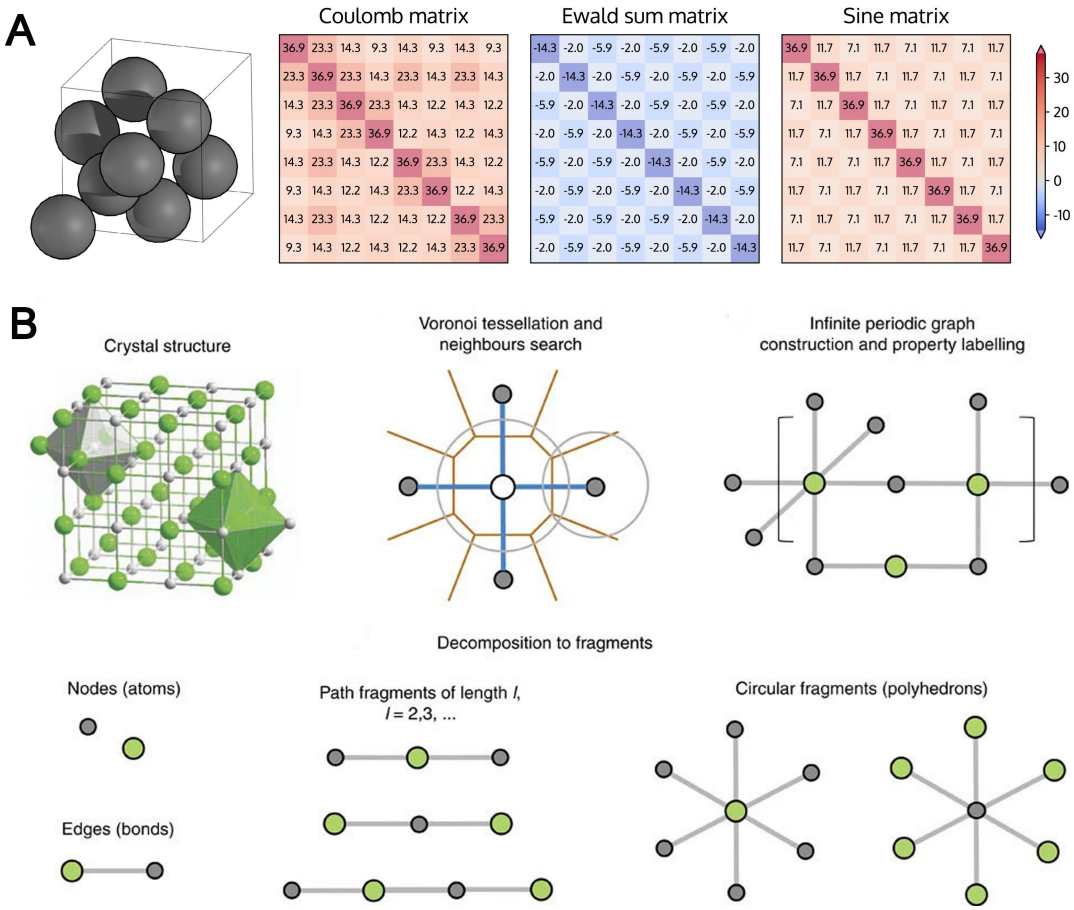


**Figure 3.** (A) Illustration of the Coulomb matrix, Ewald sum matrix, and sine matrix for a periodic diamond structure[46]. Copyright 2020, Elsevier; (B) Schematic representing the construction of the property-labeled materials fragments (PLMF)[47]. Copyright 2017, Springer.

training and improve model quality[57]. In contrast, an overabundance of descriptors can lead to overfitting and "The Curse of Dimensionality"[58], whereas an insufficient number of descriptors can result in inadequate expression of material properties and poor performance of the trained model. A previous reference[21] summarized some general rules that descriptor sets should follow.

### Algorithm selection and model training

Table 2 lists some of the commonly used ML algorithms and representative examples of their use in materials research[59-72]. Detailed descriptions of these algorithms are widely available, but the difficulty lies in choosing the most appropriate algorithm for a given task. On this issue, some generalized rules are outlined in a previous reference[21]. However, these rules are based on simplifying assumptions. While they can expedite the process of finding the most suitable algorithm, they do not offer a one-size-fits-all solution. Following these rules may yield multiple suitable algorithms or, in some cases, none. To address this challenge, researchers have developed meta-learning, also known as "learning to learn"[73]. It involves acquiring knowledge by learning from meta-data (algorithm configurations, parameter settings, other measurable properties, *etc.*) of previous similar tasks and transferring it to new tasks to identify the best algorithm and hyperparameter combination for the given problem[74-80]. Meta-learning has found applications in the pharmaceutical field[81,82] and energy materials design. For instance, in 2021, Sun *et al.*

**Table 2. List of commonly used ML algorithms for materials research**

| Algorithm model | Application |
|---|---|
| ANN | Material design[59] |
| CNN | Binding energies prediction[60] |
| Clustering | Spectral analysis[61] |
| GPR | Adsorption energy prediction[62] |
| Generative models | New material discovery[63,64] |
| Gradient Boosting Algorithm | Materials screening, discovery, and property prediction[65,66] |
| KRR | Molecular orbital energy prediction[67] |
| RF | Determine the importance of descriptors[68,69] |
| SVM | Catalytic activity prediction and simplification of DFT calculations[70,71] |
| SISSO | Descriptor selection[72] |

ANN: Artificial neural network; CNN: convolutional neural network; GPR: gaussian process regression; KRR: kernel ridge regression; ML: machine learning; RF: random forest; SISSO: sure independence screening and sparsifying operator; SVM: support vector machines.

developed a meta-learning model that collectively predicts the adsorption capacity of various materials under different pressures and temperatures[83].

### Model evaluation and selection

The metric that quantifies the error of the model on the training set is known as the training error. However, this metric solely reflects the ability of the model to fit the training set and falls short in assessing its performance on the target problem. Our focus lies in understanding the error of the model on unseen data, referred to as the generalization error. To accurately evaluate the generalization error, it is essential to assess the model performance using a separate test set. In supervised learning, commonly employed model evaluation methods include hold-out, bootstrapping, and cross-validation[84]. Regression models commonly employ evaluation indicators such as the coefficient of determination (R2), mean square error (MSE), root MSE (RMSE), and mean absolute error (MAE)[85]. Classification models incorporate precision, recall, accuracy, and F1 score[86,87]. The choice of evaluation methods and indicators depends on the availability of the specific data and the objectives of the task[84].

## ACCELERATING ELECTROCATALYST RESEARCH USING ML

Accelerating electrocatalyst research using ML is a promising approach in materials science. There are two main approaches to accelerate the study of electrocatalysts through the utilization of ML. The first approach entails the utilization of ML models to prognosticate material properties, explore the current material space, and conduct a screening of potential electrocatalysts that satisfy requisite criteria. These predictions are subsequently validated through either experimental or computational means, thereby reducing the need for trial and error and minimizing the associated expenses. The second approach facilitates the optimization of existing catalysts and the discovery of new catalysts by providing valuable insights that inform the synthesis and theoretical calculations of new catalysts.

### Prediction of electrocatalyst performance

*Activity and selectivity*

In 1920, the French chemist Paul Sabatier proposed that the adsorption of reactants on a catalyst should be neither too weak nor too strong. Weak adsorption impedes the occurrence of significant reactions, while strong adsorption results in the formation of stable intermediate products that cover the catalyst surface, impeding the sustainability of reactions[88]. In 2003, Nørskov *et al.* used DFT calculations to demonstrate that the adsorption energy of an intermediate can be a descriptor of catalytic activity and moderate

adsorption energy generally contributes to a better catalytic activity[89-91]. However, adsorption energies cannot be accurately measured experimentally, and DFT can only calculate the adsorption energy of a small number of active sites on the catalyst surface. This limits the development of catalyst design based on this descriptor. The development of AI has overcome this limitation. In recent years, ML has become a popular method for catalyst design. Specifically, ML has been applied to predict the adsorption energy of reaction intermediates on various catalysts and, in turn, predict the catalytic activity and selectivity of catalysts.

Alloys are common electrochemical catalysts. In the search for $CO_2$ electrocatalysts, Zhong *et al.*[92] screened 244 different copper-containing intermetallic compounds from the Materials Project[93], and they listed 12,229 surfaces and 228,969 adsorption sites. They used DFT to calculate the adsorption energy of certain sites, and based on these data, a ML model was trained using a random forest (RF) algorithm. This ML model was then used to predict the adsorption energy of CO on various adsorption sites. Combining the predicted values with the volcano plot relationship[94], the best active sites were identified. The optimal sites were then simulated by DFT, and the data obtained was fed back to the ML model for training. In this way, an automated search framework was established to search for surfaces and adsorption sites with CO adsorption energy close to the optimal value. The framework conducted approximately 4,000 DFT simulations in total and generated a set of candidate materials for experimental testing. Experimental results indicated that Cu-Al had the best activity and selectivity for $CO_2$ reduction. Park *et al.*[95] used a CGCNN model[53] to predict the binding energy of *COOH on gold-silver nanostructures. The CGCNN model exhibited a MAE of 0.024 eV for the *COOH binding energy prediction on the test set. They further demonstrated a stable configuration of the *COOH intermediate on the $Au_1Ag_1$ surface, in which C is bonded to Au and O is bonded to Ag.

High-entropy alloys (HEAs) were discovered in 2004 and have recently emerged as discovery platforms for catalytic materials[96,97], demonstrating excellent catalytic performance in existing reports[98-100]. However, the large number of possible active sites and the vast chemical space make it difficult to comprehensively study them using traditional methods[101]. The integration of ML has transformed the traditional research strategy, enabling the comprehensive studies of HEAs. Batchelor *et al.* conducted a study on oxygen reduction reaction (ORR), wherein they calculated the adsorption energy of *OH and *O on 871 and 998 different 2 × 2 unit cells, respectively[102]. Notably, each of these unit cells was characterized by a distinct set of random effective binding sites. Based on these data, they trained a model using ordinary least squares algorithms to predict the entire span of available adsorption energy on the IrPdPtRhRu surface of HEAs. The model was tested on a set of 3 × 4 non-symmetric unit cell surface sites. The root-mean-square deviation (RMSD) of the adsorption energies of *OH and *O were 0.063 and 0.076 eV, respectively. Pedersen *et al.* proposed a method for discovering selective and active catalysts for the reduction of $CO_2$ and CO on HEAs[103]. By combining DFT with Gaussian process regression (GPR), the CO and H adsorption energies of all sites on the (111) surfaces of disordered CoCuGaNiZn and AgAuCuPdPt HEAs were predicted. This allowed for the optimization of the HEA composition, which, in turn, increased the probability of the sites with weak H adsorption to suppress the formation of molecular hydrogen. Simultaneously, it enhanced the likelihood of sites with strong CO adsorption to promote CO reduction. A selectivity-activity plot was drawn using predicted adsorption energies [Figure 4], which describes how the selectivity of $CO_2$/CO reduction reactions ($CO_2$RR/CORR) and the activity of CORR are expected to change as the composition of HEAs varies.

In recent years, single-atom catalysts (SACs) have shown excellent performance in various catalytic reactions and have become the forefront of catalysis research. ML has been used to predict material properties in the design process of SACs, which reduces the number of DFT calculations and thus lowers

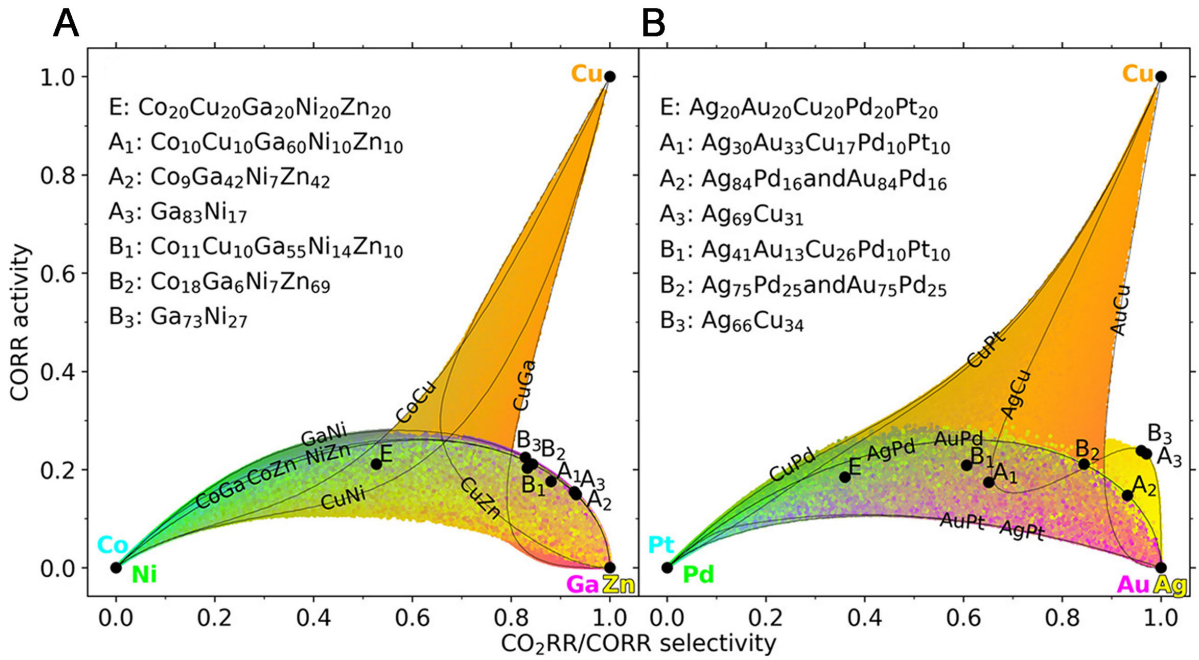


**Figure 4.** Plots of CORR activity varying with $CO_2$RR/CORR selectivity achieved by CoCuGaNiZn (A) and AgAuCuPdPt (B)[103]. Copyright 2020, American Chemical Society. $CO_2$RR/CORR: $CO_2$/CO reduction reactions.

the cost. In 2020, Zafari *et al.* used a deep learning neural network (DNN) to predict effective electrocatalysts for nitrogen reduction reaction (NRR) in boron (B)-doped graphene-based SACs[65]. The DNN model is shown in Figure 5, and Figure 6 illustrates the relation between the loss function, optimizer, layers, input data, and targets. The output of the DNN was used to identify qualified candidate samples for NRR, which were defined as having a probability of being an effective catalyst greater than 0.5. Multiple ML methods were used to predict the adsorption and free energies of some intermediates during the NRR reaction process. Among these models, the light gradient boosting machine (LGBM) showed the best prediction accuracy (RMSE = 0.11 eV). In 2022, Sun *et al.* used GPR to predict the selectivity of syngas in the process of $CO_2$ reduction over the surfaces of graphdiyne (GDY)-based SACs from the perspective of adsorption energy[104]. Considering the influence of the acidity and basicity of the medium, four strategies were employed to determine the selectivity of $H_2$ and CO. Distinct selectivity was obtained through different comparison strategies, indicating that flexible control of the syngas composition must rely on a comprehensive exploration of thermodynamic adsorption and electron regulation[104].

Perovskite-type oxides are catalysts that offer several advantages, including high efficiency, low cost, and environmental friendliness. However, the complex substitution of multiple elements in these catalysts makes traditional research methods inefficient. Wang *et al.* proposed a surface center-environment feature model and developed a ML approach based on this model to predict the adsorption free energies and overpotentials of reactive intermediates (HO*, O*, and HO*) on chalcogenide oxide surfaces[105]. Their strategy has proven effective in the targeted selection of chalcogenide catalysts with desired properties, and there is potential for extending the surface center-environment model to other catalyst types in the future to broaden its applicability.

*Stability*

In catalyst design, thermodynamic stability is a crucial factor and is often quantitatively described using

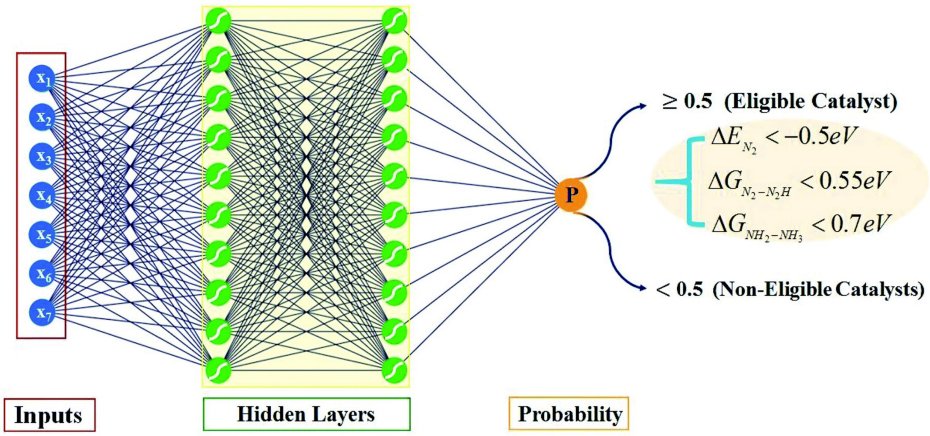


**Figure 5.** ANN (10 neurons in each hidden layer) architecture[65]. Copyright 2020, Royal Society of Chemistry. ANN: Artificial neural network.

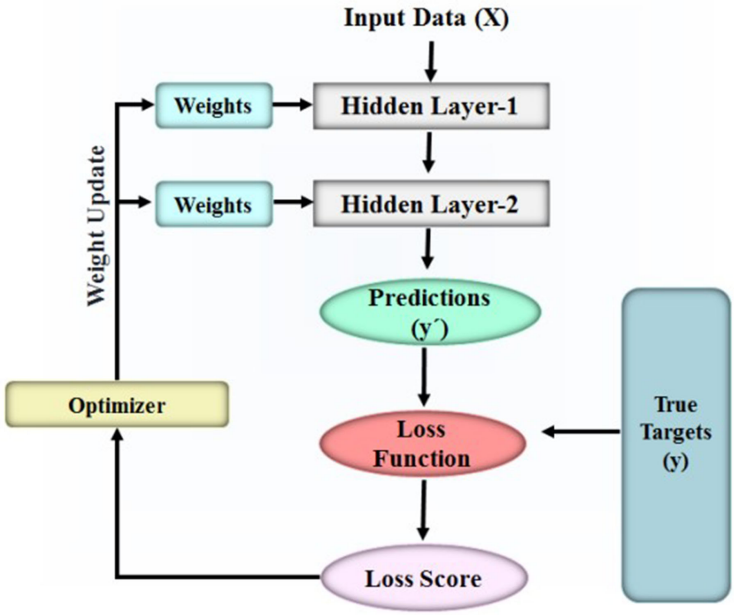


**Figure 6.** Relation between the loss function, optimizer, layers, input data, and targets[65]. Copyright 2020, Royal Society of Chemistry.

formation energy. In 2015, Faber *et al.* proposed a set of crystal structure feature vectors that can be used via ML models to predict solid-state formation energy[106]. Initially, the Coulomb matrix representation was developed for organic molecules, while the Ewald sum matrix (extended Coulomb matrix) and sine matrix were proposed for periodic systems. A dataset of 3,938 crystal structures was extracted from the Materials Project, with 3,000 of them constituting a training dataset for a kernel ridge regression (KRR) model to predict crystal formation energy and stability. Two years later, Seko *et al.* demonstrated a method to generate a set of composite descriptors from simple elemental and structural representations for predicting compound formation energy[107]. This model achieved a prediction error of 0.041 eV/atom. Schmidt *et al.* constructed a dataset of approximately 250,000 cubic perovskite systems using DFT calculations[108]. This dataset was used to train and test a range of ML algorithms [ridge regression, RF, extremely randomized

trees, and neural networks (NN)] for predicting inorganic solid-state energies. After conducting an average of more than 20 training sessions and tests, the results indicated that the extremely randomized trees had the highest prediction accuracy (MAE = 123.1 ± 0.8 meV/atom). Ward *et al.* mapped the enthalpy of generation calculated by DFT to a set of two types of attributes (composition-dependent attributes of elemental properties and attributes derived from the Voronoi tessellation of the crystal structure of the compound)[109]. A decision tree model was tested on a dataset of 435,000 formation energies from the Open Quantum Materials Database (OQMD). It achieved an average absolute error of 80 meV/atom in predicting formation enthalpy.

In addition to using formation energies to describe structural stability, the design of sub-stable surface structures can also be achieved by searching for the minimum energy path during transformations between different surface structures. In 2000, Henkelman *et al.* proposed a modification of the nudged elastic band method (NEB) for finding the minimum energy path based on DFT computations[110]. This method is more reliable than classical force field-based dynamics methods, but it is computationally intensive and challenging to apply to complex structures[111]. The development of ML overcomes these limitations. In 2018, Kolsbjerg *et al.* demonstrated that approximate structural relaxation with a NN enables orders of magnitude faster global optimization using an evolutionary algorithm within a DFT framework[112]. This significant increase in computational speed makes it possible to filter out the best energy paths from hundreds of kinetic paths. In 2021, Yoon *et al.* proposed a deep reinforcement learning (DRL) environment called CatGym for predicting thermal surface reconstruction pathways and their associated kinetic barriers in crystalline solids under reaction conditions[113]. For a given catalyst surface, the DRL agent iteratively adjusts the positions of atoms and learns strategies for generating kinetic pathways to nearby local minima with different surface compositions resulting from surface segregation. The reconstruction pathway to the global minimum surface configuration generated by the DRL agent agrees well with the minimum energy path calculated using NEB.

All of the above strategies evaluate structural stability from an energy perspective, and there are other strategies. In 2016, Ulissi *et al.* developed a strategy to efficiently generate surface Pourbaix maps using a Gaussian regression process based on a small amount of conformational free energy calculated by DFT[114]. Such surface phase maps can not only show the most stable surface structure as a function of pH and potential but also help to understand surface chemistry. They generated a Pourbaix map [Figure 7] of the $IrO_2$ (110) surface using only 20 electronic structure relaxations, whereas about 90 are required using typical search methods. And the same efficiency was obtained on the $MoS_2$ surface. In 2021, Vulcu *et al.* investigated the stability and surface changes of the electrodes by comparing Raman spectra recorded before and after electrochemical treatment[115]. However, due to the great similarity between the data generated by the analysis and the spectra, ML algorithms were used for discrimination. Five modeling approaches [the decision trees, the discriminant analysis, support vector machines (SVM), k-nearest neighbors (KNN), and ensemble classifiers] were used in this research. The findings demonstrated that sulfur-doped reduced graphene oxide (S-RGO-Pt) has higher molar stability in alkaline media.

### Quantitative structure-property relationship

The activity of electrocatalysts is not simply dominated by a few properties but is the result of the interaction and mutual limitation of multiple features and properties. Therefore, it is important to reveal the structure-property relationship for the rational design of electrocatalysts. Quantitative structure-property relationship (QSPR) has been widely used in materials research fields[116-119], but its application in the field of electrocatalysis has only recently shown some promising advancements. Parker *et al.* used non-linear and non-parametric extra trees classifier to classify 1,300 Pt nanoparticles into disordered and ordered structures based on the degree of surface disorder and growth rate[120]. Subsequently, non-linear and non-parametric

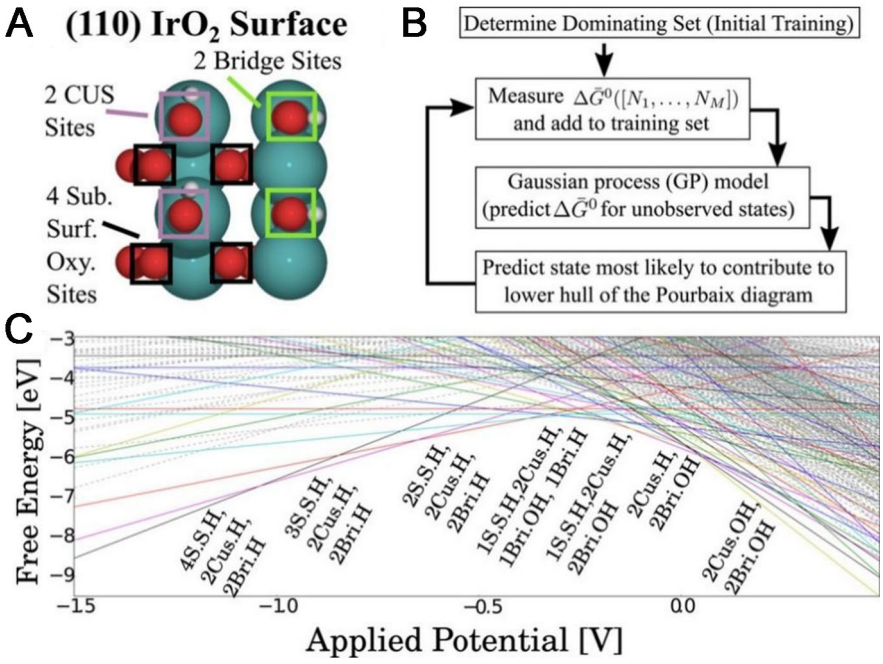


**Figure 7.** Demonstration of Pourbaix diagram construction for an $IrO_2$ surface. (A) Illustration of three types of adsorption sites considered for a 2 × 2 $IrO_2$ slab; (B) Algorithm for Pourbaix diagram construction using a ML model to guide simulation choice; (C) Final Pourbaix diagram, with the states forming the lower hull labeled. Dashed lines are predicted states of unmeasured configurations[114]. Copyright 2016, American Chemical Society. ML: Machine learning.

extra trees regressors were used to investigate the relationship between the structural properties of the two types of particles and the ORR, hydrogen oxidation reaction, and hydrogen evolution reaction (HER). The results show that small particles of disordered materials perform better for hydrogen precipitation reactions and hydrogen oxidation reactions. In addition, for ordered structures, increasing (111) surface area would promote ORR, while increasing (110) surface area would enhance hydrogen evolution and hydrogen oxidation reactions. Esterhuizen *et al.* used an interpretable ML model, the generalized additivity model, to quantify and explain the relationship between the geometry of the adsorption site and the strength of chemisorption[121]. Through several case studies, they explained the relationship between the basic electronic, geometrical, and compositional features of Rh, Pd, Ag, Ir, Pt, and Au alloys and the chemisorption strengths, coordination metals, and strains of O, S, OH, and Cl adsorbates. Based on the available feature shapes, three key features of the adsorption sites were identified as affecting the chemisorption strength on the metal alloy phases: the strain in the surface layer, the number of d-electrons in the ligand metal, and the size of the ligand atom.

The mapping between material synthesis, material characteristics, and performance is illustrated in Figure 8A. The synthesis conditions of electrocatalysts affect their structure and, thus, performance, while simple QSPR does not consider the synthesis conditions. Based on QSPR, Ebikade *et al.* developed a data-driven quantitative synthesis-structure-property relationships ($QS^2PRs$) method to enhance the performance of nitrogen-doped carbon (NDC) for hydrogen precipitation reactions[122]. Figure 8B outlines the active learning algorithm based on Kriging methods that were used to construct a predictive model. The NDC synthesis process was used as the objective function, with the synthesis conditions being the input function and the total N content being the response to be optimized. Combined with other ML tools, the optimal pyrolysis conditions for the preparation of NDC can be effectively determined, as well as the

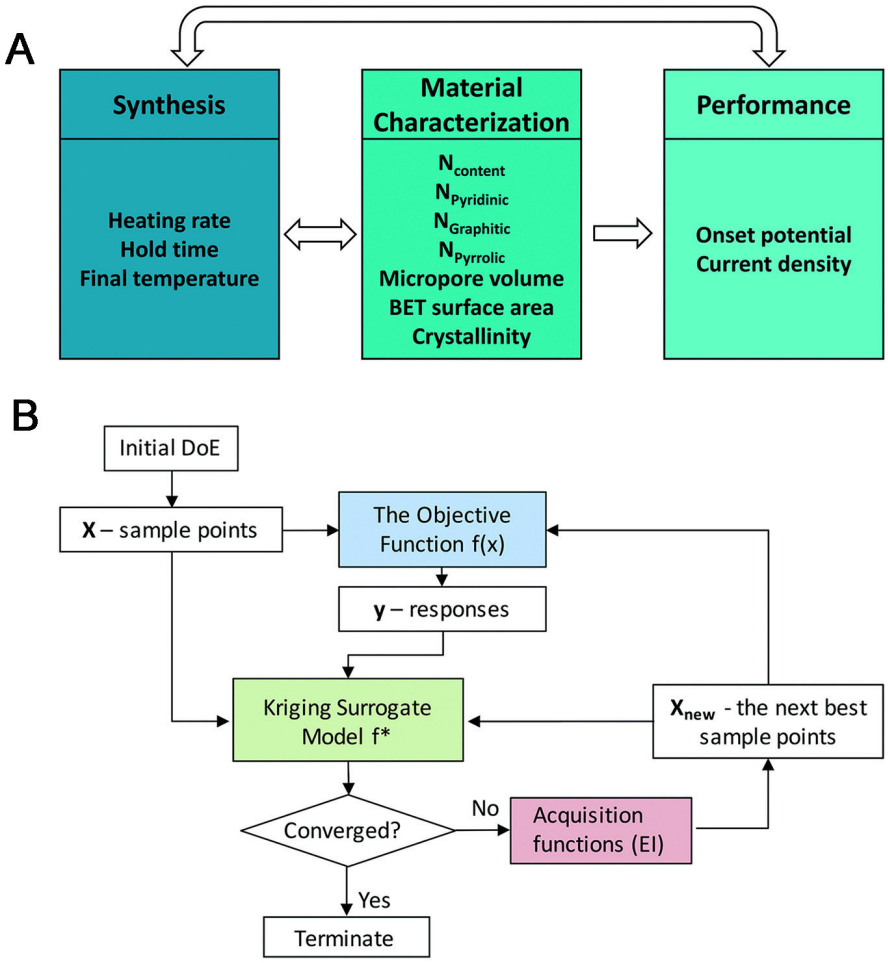


**Figure 8.** (A) Mapping between synthesis conditions, material characterization, and performance; (B) Kriging-based active learning algorithm[122]. Copyright 2020, Royal Society of Chemistry.

electrochemical properties of resulting NDC catalytic materials.

### Descriptor identification

Finding important parameters that determine the catalytic performance of materials has been a focus of research in the field of electrocatalysis. Over the past few decades, several descriptors have been developed to reveal the structure-performance relationship, including descriptors for adsorption energy of reaction intermediates, electron descriptors represented by d-band centers, structural descriptors, and universal descriptors[123]. These descriptors have provided important guidance for the development of electrocatalysts but still have some limitations, such as being difficult to measure and having poor universality. In recent years, ML has become a new, fast, and effective tool for descriptor development or key parameter identification[124-128].

Wexler *et al.* combined DFT and ML to study the activity of $Ni_2P$ for the HER[68]. They used a regularized RF algorithm to discover the relative importance of structural and charge descriptors and found that the Ni-Ni bond length was the most important descriptor for HER activity. This finding sheds light on the mechanism of dopant-induced changes in the reactivity of $Ni_2P$. Jäger *et al.* established complex descriptors

to accurately and reasonably predict adsorption energies[129]. They investigated the smooth overlap of atomic positions, many-body tensor representation, and atomic central symmetry function in predicting the hydrogen adsorption free energy ($\Delta G_H$) of 91 $MoS_2$ clusters and 24 copper-gold clusters. After a comparative analysis, the smooth overlap of atomic positions descriptor was used to explain the adsorption energy. In addition, it was concluded that merging data from different nanoclusters could significantly reduce the need for fitting potential energy surfaces.

Weng *et al*. used symbolic regression (SR) to guide the design of novel oxide perovskite catalysts for oxygen evolution reaction (OER) [Figure 9][130]. A descriptor, μ/t, was identified from $4.32 \times 10^7$ candidates, which has high accuracy and low complexity. The μ and t represent the octahedral factor and tolerance factor, respectively. This accelerated the discovery of new high-performance oxide perovskite catalysts for OER. Fung *et al*. studied the descriptors of the catalytic activity of nitrogen-doped graphene-based SACs for HER by constructing the correlation between the d-state center and $\Delta G_H$[131]. Notably, $\Delta G_H$ is a widely studied descriptor for the interaction between molecules and metal surfaces in HER[132,133]. However, the computed results showed a relatively weak correlation between the d-state center and $\Delta G_H$ ($R^2 = 0.66$). Other descriptors were also studied, such as the formation energy of single-atom positions, the number of filled and unfilled d-states near the Fermi level, and atomic properties of single atoms, ionization potential, electronegativity, number of d-electrons, covalent radius, and Zunger d-orbital radius. In addition, as shown in Figure 10, the performance of several commonly used ML models for predicting $\Delta G_H$ is compared, including KRR, RF, NN, and sure independence screening and sparsifying operator (SISSO).

Compared with metal catalysts, metal oxide catalysts have more localized and complex electronic structures. This causes the lack of suitable activity descriptors to replace expensive DFT calculations in predicting the catalytic activity of metal oxides. Xu *et al*. demonstrated the use of a compressed sensing method (SISSO) to identify the algebraic expressions of surface-derived features as descriptors[134]. Subsequently, they utilized the primary electronic and geometric features to predict the adsorption enthalpies of intermediates on doped $RuO_2$ and $IrO_2$ electrocatalysts in OER. The results showed that none of the primary features was uniquely important, and the descriptor was significantly superior to previously emphasized single descriptors in terms of accuracy and computational cost. Andersen *et al*. explored the possibility of using the SISSO method to identify low-dimensional descriptors[135]. These descriptors are used to predict the enthalpies of adsorption on various active sites of metals and oxides. Zafari *et al*. used two-dimensional (2D) transition metal borides (MBene), defect-engineered materials, and p-conjugated polymers (2DCP)-supported SACs to promote $N_2$ reduction to $NH_3$ while suppressing HER[136]. By building a ML model (LGBM) based on the dataset, a new NRR descriptor combining a bond orientation parameter (BOP) and simple element features was proposed. Linear feature correlation analysis showed that N-N bond length was highly correlated with catalytic activity. This indicated that activation of $N_2$ was crucial for the high performance of the catalyst. In 2022, using DFT, ML, and a cross-validation scheme, Wan *et al*. selected the best performing RF regression model (with an RMSE of 0.24 V/0.23 V for ORR/OER) from models constructed by five different supervised ML algorithms[137]. This model was used to characterize the easily accessible physical and chemical properties of carbon-nitride-related SACs with respect to the ORR/OER overpotential. Three promising oxygen electrocatalysts with higher activity than noble metals were identified, including RhPc, Co-N-C, and Rh-$C_4N_3$. Further model analysis determined the number of electrons in the d orbitals of the metal active centers as the most effective descriptor. The study successfully predicted the overpotentials of ORR and OER on carbon-nitride-related SACs and demonstrated the superiority of the ML model over traditional experimental approaches and theoretical models.

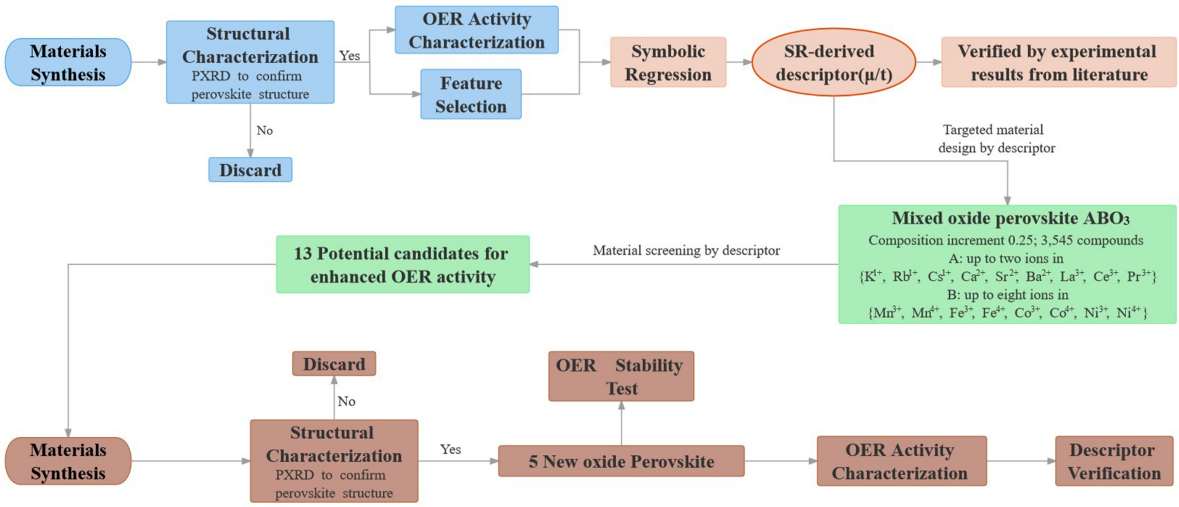


**Figure 9.** Workflow diagram. It contains four major parts: dataset generation (blue), SR (red), materials design and screening (green), and experimental verification (brown)[130]. Copyright 2020, Springer. OER: Oxygen evolution reaction; SR: symbolic regression.

## ML interatomic potential

The potential-energy surface (PES) is defined as a function of the potential energy of the resulting atomic configuration if atomic coordinates are provided[138]. The complexity of PESs varies depending on the chemical system described. PESs may depend on only a few coordinates or may be highly complex high-dimensional functions. Theoretically, PESs can be obtained by solving the Schrödinger equation for the chemical system, which is the most accurate method. Despite its accuracy, the exact solution of the Schrödinger equation for practical systems is currently not available. Even the approximate solution of the Schrödinger equation is limited by the computational cost and is difficult to use for systems with large time and length scales, such as the most widely used DFT[7,139].

To address the difficulties of PES calculations, researchers have developed an alternative to PES-interatomic potential models. These models parameterize the interactions between atoms in a relatively simple functional form and are widely used in materials science[140]. MD simulations aided by the use of interatomic potential models enable access to larger time and length scales and enhance the ability to simulate chemical systems with atomic numbers up to hundreds of thousands[141]. Initially, the potential functions were mainly constructed manually, but now they are mainly constructed by ML. In recent years, many ML models for potential or force field prediction have been published. These include various NN potentials (NNPs)[142-147], graph networks[148,149], Gaussian approximation potentials (GAP)[150,151], SVM[152], moment tensor potentials (MTP)[153], gradient-domain ML (GDML)[154] and many more. ML interatomic potentials have emerged as valuable tools for materials research[19], but their application to electrocatalysts is limited, with few studies reported so far.

Artrith *et al.* combined first-principles calculations with large-scale Monte Carlo simulations, assisted by an NNP, to study the equilibrium surface structure and composition of bimetallic Au/Cu nanoparticles[155]. To ensure the accuracy of NN, up to 3,915 Au/Cu nanoparticles (with a size of 6 nm) were extensively sampled under different chemical potentials and synthesis conditions. They demonstrated that NNPs based on first principles provide a promising approach to accurately investigate the relationship between solvent, surface composition and morphology, surface electronic structure, and catalytic activity in systems consisting of thousands of atoms. Chen *et al.* used local ML potentials (MLPs) to obtain structural descriptors and

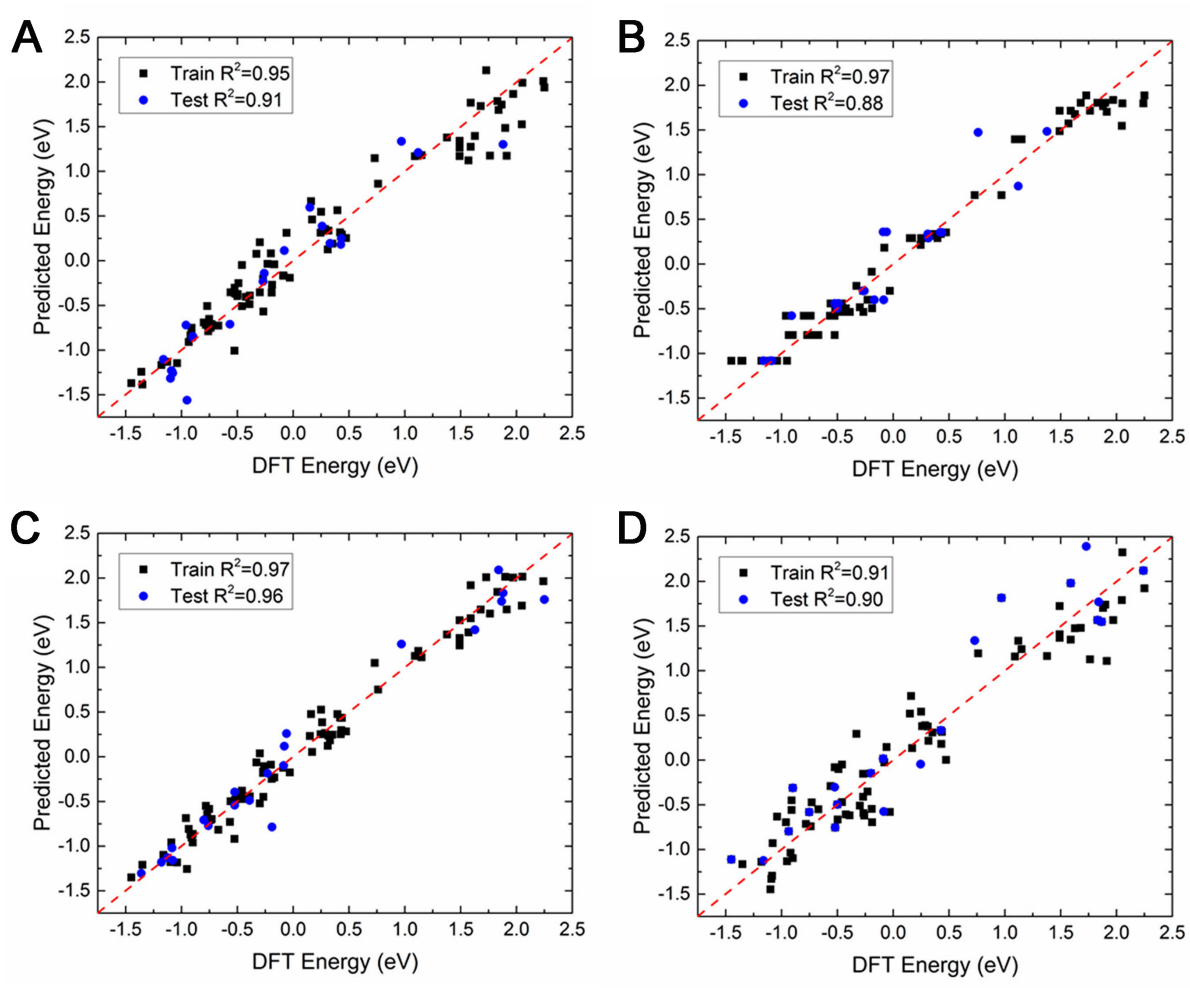


**Figure 10.** Comparison of DFT calculated versus ML predicted $\Delta G_H$ using (A)KRR; (B) RF regression; (C) NN regression; and (D) SISSO regression[131]. Copyright 2020 American Chemical Society. DFT: Density functional theory; KRR: kernel ridge regression; NN: neural network; ML: machine learning; RF: random forest; SISSO: sure independence screening and sparsifying operator.

achieved local structure optimization by combining simple physical properties with graph convolutional NN[156]. Subsequently, they selected 43 high-performance alloys from 2,973 candidates as potential electrocatalysts for hydrogen precipitation reactions. Some of the 43 alloys have been validated in experiments. Li *et al.* combined the quantum mechanical path integral-based rate theory of cyclic polymer MD with an NNP of first-nature principle accuracy to calculate the surface reaction rate[157]. They applied this approach to the example of NO desorption on a Pd (111) surface. The results indicated that the resonance approximation and neglect of lattice motion in the conventional transition state theory can respectively overestimate and underestimate the entropy change during desorption. These lead to opposite errors in the rate constant prediction, thereby resulting in a situation where the errors cancel out. After taking into account the anharmonicity and lattice motion, the study correctly revealed the surface entropy change during the desorption process, which is usually neglected due to the apparent local structural changes.

ML interatomic potentials have gained rapid momentum in recent years, and a large number of reported examples have demonstrated their potential value. However, they currently face several challenges. The first is the generation of reference data. Constructing MLPs requires the generation of extensive reference

datasets using electronic structure calculations, which need to be performed at a highly converged level[148]. This process is very demanding and time-consuming, which makes empirical force fields orders of magnitude faster than ML models. Reducing the size of the reference dataset is a current endeavor[158]. The second challenge is the poor transferability of ML models due to the high-dimensional feature space, which is inherent to high-dimensional fitting functions and is known as "The Curse of Dimensionality"[143]. It means that when confronted with different material systems, old ML models may lead to serious failures, necessitating the training of new models from scratch. To address this problem, it would be beneficial to develop more automated database generation methods and potential training methods[143].

## CHALLENGES AND PROSPECTS

Significant progress has been made in utilizing ML to accelerate the optimization and discovery of electrocatalysts. However, there are still some challenges that need to be addressed in order to fully realize the potential of ML in this field.

First of all, in terms of data, ML requires a large and reliable dataset to ensure its quality of learning. Currently, there are problems such as inadequate data acquisition efficiency, a large amount of published data not being included in databases, and important experimental data not being recorded in the literature. For example, factors such as the shape of the reactor and stirring speed can affect the catalytic performance[50] but may not always be reported in experimental data. Additionally, researchers are often unwilling to publish "failure data" that can be used for ML[32]. In addition to the size and comprehensiveness of the data, the quality of the data should also be considered, as different data sources can cause some errors.

Secondly, in terms of workflow, while ML modeling can theoretically be completed with limited professional knowledge, the success of the model currently depends heavily on the experience of researchers. This is because the properties of materials are affected by various physical and chemical factors and process conditions. A large number of influencing factors make redundant features difficult to avoid, thereby leading to dimensional catastrophes[58]. These issues can result in poor prediction performance and high model complexity. To address these challenges, it is important to select appropriate descriptors, which requires a thorough understanding of catalysis theory. Moreover, selecting the algorithm is also difficult, as there is no single algorithm suitable for all problems. Many researchers choose multiple algorithms during modeling and use the test set to select the best performing algorithm. This undoubtedly increases workload. To solve this problem, promoting collaboration between scientists in different fields (mathematics, computer science, materials science, and catalysis science) would be an effective way.

Thirdly, the interpretability of the model is an issue. The conventional ML models are difficult to formalize and are, therefore, regarded as "black boxes". As a result, it is difficult to extract scientific knowledge that can be applied to general situations from ML models. Developing interpretable ML models is an effective solution to this issue, and there have been some related reports and research efforts in this area[133,159-162].

The above issues are some of the specific challenges currently faced in accelerating electrocatalyst development using ML. In addition, there are also problems, such as poor model generalization and difficulty in surpassing DFT calculations.

The previously mentioned problems are indeed challenging, but they do not address the fundamental aspects of chemical science discovery. It is important to acknowledge that while ML has accelerated specific research tasks, it has not yet fully influenced the field of electrocatalysis as a whole. This is primarily due to the lack of a systematic and standardized data-driven approach, which is essential for accelerating scientific

discovery at its core. For a more comprehensive discussion on this topic, constructive comments can be found in a recent review[163].

Overall, ML has the potential to have a significant impact on the future of scientific research in this area, as problems continue to be solved and a standardized system is established.

## DECLARATIONS

### Authors' contributions

Manuscript draft: Hu Y, Chen J
Proposed the conception and design of this review: Zhao Y, He Q, Wei Z
Collected references and provided revision: Hu Y, Chen J, Wei Z, He Q, Zhao Y
Provided supervision and acquired funding: Zhao Y

### Availability of data and materials

Not applicable.

### Financial support and sponsorship

We acknowledge the financial support from the National Natural Science Foundation of China (Grant No. 22273096) and the Fundamental Research Funds for Central Universities (20826041G4185).

### Conflicts of interest

All authors declared that there are no conflicts of interest.

### Ethical approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.



## REFERENCES

1. Seh ZW, Kibsgaard J, Dickens CF, Chorkendorff I, Nørskov JK, Jaramillo TF. Combining theory and experiment in electrocatalysis: insights into materials design. *Science* 2017;355:eaad4998. DOI PubMed
2. Brockway PE, Owen A, Brand-correa LI, Hardt L. Estimation of global final-stage energy-return-on-investment for fossil fuels with comparison to renewable energy sources. *Nat Energy* 2019;4:612-21. DOI
3. Reymond JL. The chemical space project. *Acc Chem Res* 2015;48:722-30. DOI PubMed
4. Walsh A. The quest for new functionality. *Nat Chem* 2015;7:274-5. DOI PubMed
5. Agrawal A, Choudhary A. Perspective: materials informatics and big data: realization of the "fourth paradigm" of science in materials science. *APL Mater* 2016;4:053208. DOI
6. Rajagopal AK, Callaway J. Inhomogeneous electron gas. *Phys Rev B* 1973;7:1912-9. DOI
7. Kohn W, Sham LJ. Self-consistent equations including exchange and correlation effects. *Phys Rev* 1965;140:A1133-8. DOI
8. Abraham FF. Computational statistical mechanics methodology, applications and supercomputing. *Adv Phys* 1986;35:1-111. DOI
9. Yao N, Chen X, Fu ZH, Zhang Q. Applying classical, *Ab Initio*, and machine-learning molecular dynamics simulations to the liquid electrolyte for rechargeable batteries. *Chem Rev* 2022;122:10970-1021. DOI PubMed
10. 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 PubMed
11. Li J, Lim K, Yang H, et al. AI applications through the whole life cycle of material discovery. *Matter* 2020;3:393-432. DOI
12. Feldman K, Agnew SR. The materials genome initiative at the national science foundation: a status report after the first year of funded research. *JOM* 2014;66:336-44. DOI

13. Tolle KM, Tansley DSW, Hey AJG. The fourth paradigm: data-intensive scientific discovery [point of view]. *Proc IEEE* 2011;99:1334-7. DOI
14. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science* 2015;349:255-60. DOI PubMed
15. Ghahramani Z. Probabilistic machine learning and artificial intelligence. *Nature* 2015;521:452-9. DOI PubMed
16. Muratov EN, Bajorath J, Sheridan RP, et al. Correction: QSAR without borders. *Chem Soc Rev* 2020;49:3525-64. DOI PubMed
17. Lyngby P, Thygesen KS. Data-driven discovery of 2D materials by deep generative models. *npj Comput Mater* 2022;8:232. DOI
18. Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. *Nature* 2018;559:547-55. DOI PubMed
19. Steinmann SN, Wang Q, Seh ZW. How machine learning can accelerate electrocatalysis discovery and optimization. *Mater Horiz* 2023;10:393-406. DOI PubMed
20. Chen L, Zhang X, Chen A, Yao S, Hu X, Zhou Z. Targeted design of advanced electrocatalysts by machine learning. *Chin J Catal* 2022;43:11-32. DOI
21. Mai H, Le TC, Chen D, Winkler DA, Caruso RA. Machine learning for electrocatalyst and photocatalyst design and discovery. *Chem Rev* 2022;122:13478-515. DOI
22. Zhang X, Tian Y, Chen L, Hu X, Zhou Z. Machine learning: a new paradigm in computational electrocatalysis. *J Phys Chem Lett* 2022;13:7920-30. DOI
23. Samuel AL. Some studies in machine learning using the game of checkers. *IBM J Res Dev* 2000;44:206-26. DOI
24. Mitchell T, Buchanan B, Dejong G, Dietterich T, Rosenbloom P, Waibel A. Machine learning. *Annu Rev Comput Sci* 1990;4:417-33. DOI
25. Keith JA, Vassilev-Galindo V, Cheng B, et al. Combining machine learning and computational chemistry for predictive insights into chemical systems. *Chem Rev* 2021;121:9816-72. DOI PubMed PMC
26. Luckenbach R. The beilstein handbook of organic chemistry: the first hundred years. *J Chem Inf Comput Sci* 1981;21:82-3. DOI
27. Xu Y. Accomplishment and challenge of materials database toward big data. *Chin Phys B* 2018;27:118901. DOI
28. Wei Z, He Q, Zhao Y. Machine learning for battery research. *J Power Sources* 2022;549:232125. DOI
29. Tshitoyan V, Dagdelen J, Weston L, et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* 2019;571:95-8. DOI PubMed
30. Kim E, Huang K, Saunders A, Mccallum A, Ceder G, Olivetti E. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem Mater* 2017;29:9436-44. DOI
31. Coley CW, Green WH, Jensen KF. Machine learning in computer-aided synthesis planning. *Acc Chem Res* 2018;51:1281-9. DOI PubMed
32. Raccuglia P, Elbert KC, Adler PDF, et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* 2016;533:73-6. DOI PubMed
33. Rahm E, Do HH. Data cleaning: problems and current approaches. *IEEE Data Eng Bull* 2000;23:3-13. Available from: https://www.betterevaluation.org/sites/default/files/data_cleaning.pdf. [Last accessed on 29 Aug 2023]
34. Artrith N, Butler KT, Coudert FX, et al. Best practices in machine learning for chemistry. *Nat Chem* 2021;13:505-8. DOI PubMed
35. Young D, Martin T, Venkatapathy R, Harten P. Are the chemical structures in your QSAR correct? *QSAR Comb Sci* 2008;27:1337-45. DOI
36. Chu X, Ilyas IF, Krishnan S, Wang J. Data cleaning: overview and emerging challenges. Proceedings of the 2016 International Conference on Management of Data. ACM; 2016. p. 2201-6. DOI
37. Qin SJ, Chiang LH. Advances and opportunities in machine learning for process data analytics. *Comput Chem Eng* 2019;126:465-73. DOI
38. Ndung'u RN. Data preparation for machine learning modelling. *Int J Comput Appl Technol Res* 2022;11:231-5. DOI
39. Hautamaki V, Karkkainen I, Franti P. Outlier detection using k-nearest neighbour graph. In: Proceedings of the 17th International Conference on Pattern Recognition, 2004; 2004 Aug 26; Cambridge, UK. IEEE; 2004. p. 430-3. DOI
40. Muller E, Assent I, Steinhausen U, Seidl T. OutRank: ranking outliers in high dimensional data. In: 2008 IEEE 24th International Conference on Data Engineering Workshop; 2008 Apr 07-12; Cancun, Mexico. IEEE; 2008. p. 600-3. DOI
41. Do K, Tran T, Phung D, Venkatesh S. Outlier detection on mixed-type data: an energy-based approach. In: Li J, Li X, Wang S, Li J, Sheng Q, editors. Advanced Data Mining and Applications. Cham: Springer; 2016. p. 111-25. DOI
42. Tang B, He H. A local density-based approach for outlier detection. *Neurocomputing* 2017;241:171-80. DOI
43. Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Francis B, David B, editors. Proceedings of the 32nd International Conference on Machine Learning. PMLR; 2015. p. 448-56. Available from: https://proceedings.mlr.press/v37/ioffe15.html. [Last accessed on 29 Aug 2023].
44. Xu Y, Goodacre R. On Splitting Training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *J Anal Test* 2018;2:249-62. DOI PubMed PMC
45. Joseph VR, Vakayil A. SPlit: an optimal method for data splitting. *Technometrics* 2022;64:166-76. DOI
46. Himanen L, Jäger MOJ, Morooka EV, et al. DScribe: library of descriptors for machine learning in materials science. *Comput Phys Commun* 2020;247:106949. DOI
47. Isayev O, Oses C, Toher C, Gossett E, Curtarolo S, Tropsha A. Universal fragment descriptors for predicting properties of inorganic crystals. *Nat Commun* 2017;8:15679. DOI PubMed PMC

48. Rupp M, Tkatchenko A, Müller KR, von Lilienfeld OA. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys Rev Lett* 2012;108:058301. DOI PubMed
49. Carhart RE, Smith DH, Venkataraghavan R. Atom pairs as molecular features in structure-activity studies: definition and applications. *J Chem Inf Comput Sci* 1985;25:64-73. DOI
50. Li S, Liu Y, Chen D, Jiang Y, Nie Z, Pan F. Encoding the atomic structure for machine learning in materials science. *WIREs Comput Mol Sci* 2022;12:e1558. DOI
51. Mao J, Jain AK. Artificial neural networks for feature extraction and multivariate data projection. *IEEE Trans Neural Netw* 1995;6:296-317. DOI PubMed
52. Schleider L, Pasiliao EL, Qiang Z, Zheng QP. A study of feature representation via neural network feature extraction and weighted distance for clustering. *J Comb Optim* 2022;44:3083-105. DOI
53. Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys Rev Lett* 2018;120:145301. DOI PubMed
54. Chen C, Ye W, Zuo Y, Zheng C, Ong SP. Graph networks as a universal machine learning framework for molecules and crystals. *Chem Mater* 2019;31:3564-72. DOI
55. Louis SY, Zhao Y, Nasiri A, et al. Graph convolutional neural networks with global attention for improved materials property prediction. *Phys Chem Chem Phys* 2020;22:18141-8. DOI
56. Choudhary K, Decost B. Atomistic line graph neural network for improved materials property predictions. *npj Comput Mater* 2021;7:185. DOI
57. 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. DOI
58. Bellman RE. Adaptive control processes: a guided tour. Princeton University Press; 1961. DOI
59. Kim B, Lee S, Kim J. Inverse design of porous materials using artificial neural networks. *Sci Adv* 2020;6:eaax9324. DOI PubMed PMC
60. Back S, Yoon J, Tian N, Zhong W, Tran K, Ulissi ZW. Convolutional neural network of atomic surface structures to predict binding energies for high-throughput screening of catalysts. *J Phys Chem Lett* 2019;10:4401-8. DOI PubMed
61. Timoshenko J, Frenkel AI. "Inverting" X-ray absorption spectra of catalysts by machine learning in search for activity descriptors. *ACS Catal* 2019;9:10192-211. DOI
62. Li Z, Achenie LEK, Xin H. An adaptive machine learning strategy for accelerating discovery of perovskite electrocatalysts. *ACS Catal* 2020;10:4377-84. DOI
63. Sanchez-Lengeling B, Aspuru-Guzik A. Inverse molecular design using machine learning: generative models for matter engineering. *Science* 2018;361:360-5. DOI PubMed
64. Song Y, Siriwardane EMD, Zhao Y, Hu J. Computational discovery of new 2D materials using deep learning generative models. *ACS Appl Mater Interfaces* 2021;13:53303-13. DOI PubMed
65. 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
66. Davies DW, Butler KT, Walsh A. Data-driven discovery of photoactive quaternary oxides using first-principles machine learning. *Chem Mater* 2019;31:7221-30. DOI
67. Stuke A, Todorović M, Rupp M, et al. Chemical diversity in molecular orbital energy predictions with kernel ridge regression. *J Chem Phys* 2019;150:204121. DOI PubMed
68. Wexler RB, Martirez JMP, Rappe AM. Chemical pressure-driven enhancement of the hydrogen evolving activity of $Ni_2P$ from nonmetal surface doping interpreted via machine learning. *J Am Chem Soc* 2018;140:4678-83. DOI PubMed
69. Panapitiya G, Avendaño-Franco G, Ren P, Wen X, Li Y, Lewis JP. Machine-learning prediction of CO adsorption in thiolated, Ag-alloyed Au nanoclusters. *J Am Chem Soc* 2018;140:17508-14. DOI PubMed
70. Baghban A, Habibzadeh S, Zokaee Ashtiani F. Bandgaps of noble and transition metal/ZIF-8 electro/catalysts: a computational study†. *RSC Adv* 2020;10:22929-38. DOI PubMed PMC
71. Mageed AK. Modeling photocatalytic hydrogen production from ethanol over copper oxide nanoparticles: a comparative analysis of various machine learning techniques. *Biomass Conv Bioref* 2023;13:3319-27. DOI
72. Ouyang R, Curtarolo S, Ahmetcik E, Scheffler M, Ghiringhelli LM. SISSO: a compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Phys Rev Mater* 2018;2:083802. DOI
73. Smith-miles KA. Cross-disciplinary perspectives on meta-learning for algorithm selection. *ACM Comput Surv* 2009;41:1-25. DOI
74. Cohen-Shapira N, Rokach L. TRIO: task-agnostic dataset representation optimized for automatic algorithm selection. In: 2021 IEEE International Conference on Data Mining (ICDM); 2021 Dec 07-10; Auckland, New Zealand. IEEE; 2021. p. 81-90. DOI
75. Shahoud S, Winter M, Khalloof H, Duepmeier C, Hagenmeyer V. An extended meta learning approach for automating model selection in big data environments using microservice and container virtualizationz technologies. *Int Thin* 2021;16:100432. DOI
76. Dyrmishi S, Elshawi R, Sakr S. A decision support framework for autoML systems: a meta-learning approach. In: 2019 International Conference on Data Mining Workshops (ICDMW); 2019 Nov 08-11; Beijing, China. IEEE; 2019. p. 97-106. DOI
77. Maher M, Sakr S. cSmartML: a meta learning-based framework for automated selection and hyperparameter tuning for machine learning algorithms. Available from: https://openproceedings.org/2019/conf/edbt/EDBT19_paper_235.pdf. [Last accessed on 18 Oct 2023]

78. Dias LV, Miranda PBC, Nascimento ACA, Cordeiro FR, Mello RF, Prudêncio RBC. ImageDataset2Vec: an image dataset embedding for algorithm selection. *Expert Syst Appl* 2021;180:115053. DOI
79. Chale M, Bastian ND, Weir J. Algorithm selection framework for cyber attack detection. In: Proceedings of the 2nd ACM Workshop on Wireless Security and Machine Learning. 2020. p. 37-42. DOI
80. Elrahman AA, El Helw M, Elshawi R, Sakr S. D-SmartML: a distributed automated machine learning framework. In: 2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS); 2020 Nov 29 - Dec 01; Singapore. IEEE; 2020. p. 1215-8. DOI
81. Ma J, Fong SH, Luo Y, et al. Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. *Nat Cancer* 2021;2:233-44. DOI PubMed PMC
82. Olier I, Sadawi N, Bickerton GR, et al. Meta-QSAR: a large-scale application of meta-learning to drug design and discovery. *Mach Learn* 2018;107:285-311. DOI PubMed PMC
83. Sun Y, DeJaco RF, Li Z, et al. Fingerprinting diverse nanoporous materials for optimal hydrogen storage conditions using meta-learning. *Sci Adv* 2021;7:eabg3983. DOI PubMed PMC
84. Raschka S. Model evaluation, model selection, and algorithm selection in machine learning. arXiv. [Preprint.] November 11, 2020 [accessed 2023 August 29]. Available from: https://arxiv.org/abs/1811.12808.
85. Chicco D, Warrens MJ, Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput Sci* 2021;7:e623. DOI PubMed PMC
86. Hossin M, Sulaiman MN. A review on evaluation metrics for data classification evaluations. *Intl J Data Min Knowl Manag Process* 2015;5:1. DOI
87. Dener M, Al S, Orman A. STLGBM-DDS: an efficient data balanced DoS detection system for wireless sensor networks on big data environment. *IEEE Access* 2022;10:92931-45. DOI
88. Che M. Nobel prize in chemistry 1912 to sabatier: organic chemistry or catalysis? *Catalysis Today* 2013;218-19:162-71. DOI
89. Logadottir A, Rod TH, Nørskov JK, Hammer B, Dahl S, Jacobsen CJH. The brønsted-evans-polanyi relation and the volcano plot for ammonia synthesis over transition metal catalysts. *J Catal* 2001;197:229-31. DOI
90. Nørskov JK, Bligaard T, Logadottir A, et al. Universality in heterogeneous catalysis. *J Catal* 2002;209:275-8. DOI
91. Mao Y, Chen J, Wang H, Hu P. Catalyst screening: refinement of the origin of the volcano curve and its implication in heterogeneous catalysis. *Chin J Catal* 2015;36:1596-605. DOI
92. Zhong M, Tran K, Min Y, et al. Accelerated discovery of $CO_2$ electrocatalysts using active machine learning. *Nature* 2020;581:178-83. DOI PubMed
93. Jain A, Ong SP, Hautier G, et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL Mater* 2013;1:011002. DOI
94. Liu X, Xiao J, Peng H, Hong X, Chan K, Nørskov JK. Understanding trends in electrochemical carbon dioxide reduction rates. *Nat Commun* 2017;8:15438. DOI PubMed PMC
95. Park JW, Choi W, Noh J, et al. Bimetallic gold-silver nanostructures drive low overpotentials for electrochemical carbon dioxide reduction. *ACS Appl Mater Interfaces* 2022;14:6604-14. DOI PubMed
96. Yeh JW, Chen SK, Lin SJ, et al. Nanostructured high-entropy alloys with multiple principal elements: novel alloy design concepts and outcomes[†]. *Adv Eng Mater* 2004;6:299-303. DOI
97. Cantor B, Chang ITH, Knight P, Vincent AJB. Microstructural development in equiatomic multicomponent alloys. *Mater Sci Eng A* 2004;375-7:213-8. DOI
98. Mori K, Hashimoto N, Kamiuchi N, Yoshida H, Kobayashi H, Yamashita H. Hydrogen spillover-driven synthesis of high-entropy alloy nanoparticles as a robust catalyst for $CO_2$ hydrogenation. *Nat Commun* 2021;12:3884. DOI PubMed PMC
99. Wang D, Chen Z, Huang YC, et al. Tailoring lattice strain in ultra-fine high-entropy alloys for active and stable methanol oxidation. *Sci Chin Mater* 2021;64:2454-66. DOI
100. Li H, Han Y, Zhao H, et al. Fast site-to-site electron transfer of high-entropy alloy nanocatalyst driving redox electrocatalysis. *Nat Commun* 2020;11:5437. DOI PubMed PMC
101. Chen ZW, Chen L, Gariepy Z, Yao X, Singh CV. High-throughput and machine-learning accelerated design of high entropy alloy catalysts. *Trends Chem* 2022;4:577-9. DOI
102. Batchelor TA, Pedersen JK, Winther SH, Castelli IE, Jacobsen KW, Rossmeisl J. High-entropy alloys as a discovery platform for electrocatalysis. *Joule* 2019;3:834-45. DOI
103. Pedersen JK, Batchelor TAA, Bagger A, Rossmeisl J. High-entropy alloys as catalysts for the $CO_2$ and CO reduction reactions. *ACS Catal* 2020;10:2169-76. DOI
104. Sun M, Huang B. Flexible modulations on selectivity of syngas formation via $CO_2$ reduction on atomic catalysts. *Nano Energy* 2022;99:107382. DOI
105. Wang X, Xiao B, Li Y, et al. First-principles based machine learning study of oxygen evolution reactions of perovskite oxides using a surface center-environment feature model. *Appl Surf Sci* 2020;531:147323. DOI
106. Faber F, Lindmaa A, von Lilienfeld OA, Armiento R. Crystal structure representations for machine learning models of formation energies. *Int J Quantum Chem* 2015;115:1094-101. DOI
107. Seko A, Hayashi H, Nakayama K, Takahashi A, Tanaka I. Representation of compounds for machine-learning prediction of physical properties. *Phys Rev B* 2017;95:144110. DOI

108. 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
109. Ward L, Liu R, Krishna A, et al. Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Phys Rev B* 2017;96:024104. DOI
110. Henkelman G, Uberuaga BP, Jónsson H. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *J Chem Physs* 2000;113:9901-4. DOI
111. Li H, Jiao Y, Davey K, Qiao SZ. Data-driven machine learning for understanding surface structures of heterogeneous catalysts. *Angew Chem Int Ed Engl* 2023;62:e202216383. DOI PubMed
112. Kolsbjerg EL, Peterson AA, Hammer B. Neural-network-enhanced evolutionary algorithm applied to supported metal nanoparticles. *Phys Rev B* 2018;97:195424. DOI
113. Yoon J, Cao Z, Raju RK, et al. Deep reinforcement learning for predicting kinetic pathways to surface reconstruction in a ternary alloy. *Mach Learn Sci Technol* 2021;2:045018. DOI
114. Ulissi ZW, Singh AR, Tsai C, Nørskov JK. Automated discovery and construction of surface phase diagrams using machine learning. *J Phys Chem Lett* 2016;7:3931-5. DOI PubMed
115. Vulcu A, Radu T, Porav AS, Berghian-grosan C. Low-platinum catalyst based on sulfur doped graphene for methanol oxidation in alkaline media. *Mater Today Energy* 2021;19:100588. DOI
116. Thanikaivelan P, Subramanian V, Raghava Rao J, Nair BU. Application of quantum chemical descriptor in quantitative structure activity and structure property relationship. *Chem Phys Lett* 2000;323:59-70. DOI
117. Dearden JC, Cronin MTD, Kaiser KLE. How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). *SAR QSAR Environ Res* 2009;20:241-66. DOI
118. Wu W, Xu H, Wang Z, et al. PINK1-parkin-mediated mitophagy protects mitochondrial integrity and prevents metabolic stress-induced endothelial injury. *PLoS One* 2015;10:e0132499. DOI PubMed PMC
119. Safder U, Nam K, Kim D, Shahlaei M, Yoo C. Quantitative structure-property relationship (QSPR) models for predicting the physicochemical properties of polychlorinated biphenyls (PCBs) using deep belief network. *Ecotoxicol Environ Saf* 2018;162:17-28. DOI PubMed
120. Parker AJ, Opletal G, Barnard AS. Classification of platinum nanoparticle catalysts using machine learning. *J Appl Phys* 2020;128:014301. DOI
121. Esterhuizen JA, Goldsmith BR, Linic S. Theory-guided machine learning finds geometric structure-property relationships for chemisorption on subsurface alloys. *Chem* 2020;6:3100-17. DOI
122. Ebikade EO, Wang Y, Samulewicz N, Hasa B, Vlachos D. Active learning-driven quantitative synthesis-structure-property relations for improving performance and revealing active sites of nitrogen-doped carbon for the hydrogen evolution reaction†. *React Chem Eng* 2020;5:2134-47. DOI
123. Wang B, Zhang F. Main descriptors to correlate structures with the performances of electrocatalysts. *Angew Chem Int Ed Engl* 2022;61:e202111026. DOI PubMed
124. Ghiringhelli LM, Vybiral J, Levchenko SV, Draxl C, Scheffler M. Big data of materials science: critical role of the descriptor. *Phys Rev Lett* 2015;114:105503. DOI PubMed
125. De S, Bartók AP, Csányi G, Ceriotti M. Comparing molecules and solids across structural and alchemical space†. *Phys Chem Chem Phys* 2016;18:13754-69. DOI PubMed
126. Hinuma Y, Mine S, Toyao T, Kamachi T, Shimizu KI. Factors determining surface oxygen vacancy formation energy in ternary spinel structure oxides with zinc†. *Phys Chem Chem Phys* 2021;23:23768-77. DOI PubMed
127. Liu X, Zhang Y, Wang W, et al. Transition metal and N doping on AlP monolayers for bifunctional oxygen electrocatalysts: density functional theory study assisted by machine learning description. *ACS Appl Mater Interfaces* 2022;14:1249-59. DOI PubMed
128. Liu X, Liu T, Xiao W, et al. Strain engineering in single-atom catalysts: $GaPS_4$ for bifunctional oxygen reduction and evolution†. *Inorg Chem Front* 2022;9:4272-80. DOI
129. Jäger MOJ, Morooka EV, Federici Canova F, Himanen L, Foster AS. Machine learning hydrogen adsorption on nanoclusters through structural descriptors. *npj Comput Mater* 2018;4:37. DOI
130. Weng B, Song Z, Zhu R, et al. Simple descriptor derived from symbolic regression accelerating the discovery of new perovskite catalysts. *Nat Commun* 2020;11:3513. DOI PubMed PMC
131. Fung V, Hu G, Wu Z, Jiang D. Descriptors for hydrogen evolution on single atom catalysts in nitrogen-doped graphene. *J Phys Chem C* 2020;124:19571-8. DOI
132. Hammer B, Nørskov JK. Electronic factors determining the reactivity of metal surfaces. *Surf Sci* 1995;343:211-20. DOI
133. Nørskov JK, Bligaard T, Rossmeisl J, Christensen CH. Towards the computational design of solid catalysts. *Nat Chem* 2009;1:37-46. DOI PubMed
134. Xu W, Andersen M, Reuter K. Data-driven descriptor engineering and refined scaling relations for predicting transition metal oxide reactivity. *ACS Catal* 2021;11:734-42. DOI
135. Andersen M, Reuter K. Adsorption enthalpies for catalysis modeling through machine-learned descriptors. *Acc Chem Res* 2021;54:2741-9. DOI PubMed
136. Zafari M, Nissimagoudar AS, Umer M, Lee G, Kim KS. First principles and machine learning based superior catalytic activities and selectivities for $N_2$ reduction in MBenes, defective 2D materials and 2D π-conjugated polymer-supported single atom catalysts†. *J*

*Mater Chem A* 2021;9:9203-13. DOI

137. Wan X, Yu W, Niu H, Wang X, Zhang Z, Guo Y. Revealing the oxygen reduction/evolution reaction activity origin of carbon-nitride-related single-atom catalysts: quantum chemistry in artificial intelligence. *Chem Eng J* 2022;440:135946. DOI
138. Behler J. Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. *Phys Chem Chem Phys* 2011;13:17930-55. DOI
139. Hohenberg P, Kohn W. Inhomogeneous electron gas. *Phys Rev* 1964;136:B864. DOI
140. Deringer VL, Caro MA, Csányi G. Machine learning interatomic potentials as emerging tools for materials science. *Adv Mater* 2019;31:1902765. DOI PubMed
141. Vink RLC, Barkema GT, van der Weg WF. Raman spectra and structure of amorphous Si. *Phys Rev B* 2001;63:115210. DOI
142. Artrith N, Urban A. An implementation of artificial neural-network potentials for atomistic materials simulations: performance for $TiO_2$. *Comput Mater Sci* 2016;114:135-50. DOI
143. Zhang Y, Hu C, Jiang B. Embedded atom neural network potentials: efficient and accurate machine learning with a physically inspired representation. *J Phys Chem Lett* 2019;10:4962-7. DOI
144. Schütt KT, Unke OT, Gastegger M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. arXiv. [Preprint.] June 7, 2021 [accessed 2023 August 29]. Available from: https://arxiv.org/abs/2102.03150.
145. Behler J. Four generations of high-dimensional neural network potentials. *Chem Rev* 2021;121:10037-72. DOI PubMed
146. Singraber A, Behler J, Dellago C. Library-based *LAMMPS* implementation of high-dimensional neural network potentials. *J Chem Theory Comput* 2019;15:1827-40. DOI PubMed
147. Kocer E, Ko TW, Behler J. Neural network potentials: a concise overview of methods. *Annu Rev Phys Chem* 2022;73:163-86. DOI PubMed
148. Zitnick CL, Das A, Kolluru A, et al. Spherical channels for modeling atomic interactions. *Adv Neural Inf Process Syst* 2022;35:8054-67. Available from: https://proceedings.neurips.cc/paper_files/paper/2022/file/3501bea1ac61fedbaaff2f88e5fa9447-Paper-Conference.pdf. [Last accessed on 29 Aug 2023]
149. Batzner S, Musaelian A, Sun L, et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat Commun* 2022;13:2453. DOI PubMed PMC
150. Byggmästar J, Nordlund K, Djurabekova F. Gaussian approximation potentials for body-centered-cubic transition metals. *Phys Rev Mater* 2020;4:093802. DOI
151. Bartók AP, Payne MC, Kondor R, Csányi G. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys Rev Lett* 2010;104:136403. DOI PubMed
152. Balabin RM, Lomakina EI. Support vector machine regression (LS-SVM) - an alternative to artificial neural networks (ANNs) for the analysis of quantum chemistry data? *Phys Chem Chem Phys* 2011;13:11710-8. DOI PubMed
153. Shapeev AV. Moment tensor potentials: a class of systematically improvable interatomic potentials. *Multiscale Model Sim* 2016;14:1153-73. DOI
154. Sauceda HE, Gastegger M, Chmiela S, Müller KR, Tkatchenko A. Molecular force fields with gradient-domain machine learning (GDML): comparison and synergies with classical force fields. *J Chem Phys* 2020;153:124109. DOI PubMed
155. Artrith N, Kolpak AM. Understanding the composition and activity of electrocatalytic nanoalloys in aqueous solvents: a combination of DFT and accurate neural network potentials. *Nano Lett* 2014;14:2670-6. DOI PubMed
156. Chen L, Tian Y, Hu X, et al. A universal machine learning framework for electrocatalyst innovation: a case study of discovering alloys for hydrogen evolution reaction. *Adv Funct Mater* 2022;32:2208418. DOI
157. Li C, Li Y, Jiang B. First-principles surface reaction rates by ring polymer molecular dynamics and neural network potential: role of anharmonicity and lattice motion. *Chem Sci* 2023;14:5087-98. DOI PubMed PMC
158. Behler J. Erratum: "Perspective: machine learning potentials for atomistic simulations" [J. Chem. Phys. 145, 170901 (2016)]. *J Chem Phys* 2016;145:170901. DOI PubMed
159. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining explanations: an overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA); 2018 Oct 1-3; Turin, Italy. IEEE; 2019. p. 80-9. DOI
160. Iwasaki Y, Sawada R, Stanev V, et al. Identification of advanced spin-driven thermoelectric materials via interpretable machine learning. *npj Comput Mater* 2019;5:103. DOI
161. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. *Entropy* 2020;23:18. DOI PubMed PMC
162. Allen AEA, Tkatchenko A. Machine learning of material properties: predictive and interpretable multilinear models. *Sci Adv* 2022;8:eabm7185. DOI PubMed PMC
163. Yano J, Gaffney KJ, Gregoire J, et al. The case for data science in experimental chemistry: examples and recommendations. *Nat Rev Chem* 2022;6:357-70. DOI