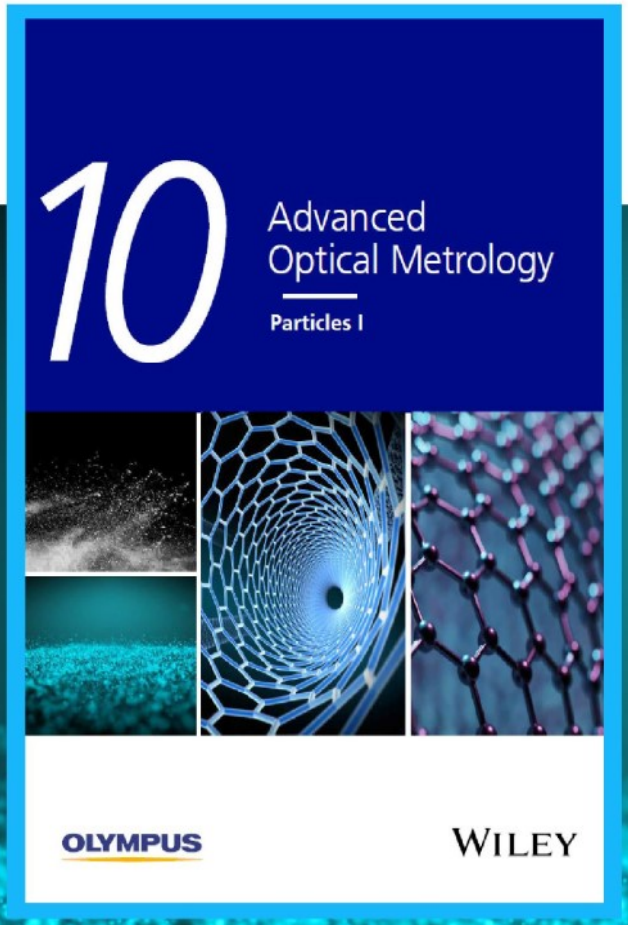
Particles I

Access the latest eBook →



Particles: Unique Properties, Uncountable Applications

Read the latest eBook and better your knowledge with highlights from the recent studies on the design and characterization of micro- and nanoparticles for different application areas.

Access Now

This eBook is sponsored by

OLYMPUS®

WILEY

Data-Driven Materials Innovation and Applications

AUTHOR NAMES, AFFILIATIONS, AND ADDRESSES

Zhuo Wang[1,3#], Zhehao Sun[2#], Hang Yin[2#], Xinghui Liu[4#], Jinlan Wang[5], Haitao Zhao[1], Cheng Heng Pang[3*], Tao Wu[3], Shuzhou Li[6*], Zongyou Yin[2*], Xuefeng Yu[1]*

[1]Materials Interfaces Center, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, Guangdong, PR China.
[2]Research School of Chemistry, The Australian National University, ACT 2601, Australia.
[3]Department of Chemical and Environmental Engineering, The University of Nottingham Ningbo China, Ningbo 315100, PR China.
[4]Department of Chemistry, Sungkyunkwan University (SKKU), 2066 Seoburo, Jangan-Gu, Suwon 16419, Republic of Korea.
[5]School of Physics, Southeast University, Nanjing, 211189, PR China.
[6]School of Materials Science and Engineering, Nanyang Technological University, Singapore, Singapore.
[#]These authors contributed equally to this work.
[*]Corresponding authors:
zongyou.yin@anu.edu.au; lisz@ntu.edu.sg; Chengheng.Pang@nottingham.edu.cn;
ht.zhao@siat.ac.cn

Keywords

machine learning, data-driven, material innovation, material informatics, material application

Abstract

Owing to the rapid developments to improve the accuracy and efficiency of both experimental and computational investigative methodologies, the massive amounts of data generated have led the field of materials science into the fourth paradigm of data-driven scientific research. This transition requires the development of authoritative and up-to-date frameworks for data-driven approaches for material innovation. This review presents a critical discussion on the current advances in the data-driven discovery of materials with a focus on frameworks, machine-learning algorithms, material-specific databases, descriptors, and targeted applications in the field of inorganic materials. Frameworks for rationalizing data-driven material innovation are described, and a critical review of essential sub-disciplines is presented, including (i) advanced data-intensive strategies and machine-learning algorithms; (ii) material databases and related tools and platforms

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/adma.202104113.

for data generation and management; (iii) commonly used molecular descriptors used in data-driven processes. Furthermore, an in-depth discussion on the broad applications of material innovation, such as energy conversion and storage, environmental decontamination, flexible electronics, optoelectronics, superconductors, metallic glasses, and magnetic materials, is provided. Finally, how these sub-disciplines (with insights into the synergy of materials science, computational tools, and mathematics) support data-driven paradigms is outlined, and the opportunities and challenges in data-driven material innovation are highlighted.

Table of Content

1. Introduction

Data-driven innovation has transformed all aspects of our life. It typically involves the invention of novel products and systems based on the knowledge extracted from data by using advanced analysis tools. The adoption of data-driven approaches has led to data-based decision-making innovations in commerce and technology, such as autonomous vehicles, MuZero, and Alphafold (artificial intelligence for mastering games and predicting protein folding, respectively).[1-4] In particular, the massive amounts of data generated by employing both computational and experimental methods, in combination with advanced machine-learning (ML) techniques, have led the field of materials science into the fourth paradigm of scientific research (**Figure 1**).[5] This data-driven paradigm has guided the development of the Material Genome Initiative (MGI), which has resulted in the advancement of experimental tools, computational techniques, and big-data analysis.[6, 7] The transformation from the trial-and-error to the data-driven paradigm requires a combination of authoritative and updated knowledge from the three domains of mathematics and statistics, computer science, and materials science.[8] The advancement and appropriate integration of these three domains will contribute to material data generation and analysis, uncertainty characterization, and efficient exploration of structure-property relationships, providing new knowledge and accelerating the discovery of innovative materials.

Innovative materials are essential and indispensable to breakthroughs in numerous applications, from energy conversion and storage to flexible electronics and optoelectronics.[9-16] For instance, novel photovoltaic materials that are cheap, stable, and environmentally friendly, easy to synthesize, and exhibit a high power conversion efficiency are being investigated.[17] Moreover, researchers are identifying highly active electrocatalysts that are selective towards the reduction of carbon dioxide.[18] The development of effective data-driven approaches is essential to meet the rapidly growing demand for innovative materials with improved and robust performance.[19, 20] A basic data-driven framework involves three fundamental stages: employment of data-intensive

strategies and ML algorithms,[21, 22] development of a comprehensive database and data generation approaches,[23, 24] and construction of descriptors that can link data-intensive and experimental strategies.[25, 26] The main objective is the rapid and efficient discovery of high-performance innovative materials by applying data-driven approaches. To achieve this goal, the fundamental stages of the data-driven framework must be utilized and integrated highlight and the relationships between a material's composition, structure, process, and properties implicit in the data must be examined.

Data-driven approaches for discovering innovative materials have certain advantages: (1) they outperform conventional trial-and-error approaches in terms of efficiency and accuracy;[27-29] (2) they can rapidly learn and extract the complex and implicit inner correlations and knowledge from the massive amounts of material data;[30-33] (3) they can achieve tailored material design based on desired functionalities because of their ability to obtain composition-structure-process-property relations;[27, 34] (4) they use ML models and descriptors to utilize complex features such as electron density and molecular graphs for improving the performance of combinatorial generalization and relational reasoning.[25, 35] Because of these advantages, many data-driven approaches exhibit high accuracy and efficiency in the prediction of properties and the exploration of property relationships.[36] Furthermore, the potential of dynamic and iterative meta-optimization data-driven processes, which represent an active learning loop that incorporates the fundamental stages, has been shown in some recent studies.[18, 37] Thus, the recent advances in data-driven innovative material discovery must be reviewed.

Comprehensive reviews have detailed the applicability of data-driven approaches to energy materials[9, 30, 38] structural materials,[39], polymeric materials,[40] and porous materials,[32] with the help of high-throughput approaches such as density functional theory (DFT) and ML.[5] The applications of ML in synthetic chemistry[41] and the prediction of material properties[29] have also been published. However, the focus of these reviews has typically been on a particular type of

material or ML technique; the interdependence between the fundamental stages, including ML algorithms, material-related databases, key descriptors, and their practical applications, of a data-driven framework for material innovation has not been reviewed. The recent advancements of each fundamental stage have also necessitated the development of the relationship between these stages, such as between ML algorithms for data augmentation and descriptor generation. Thus, a timely review of data-driven material innovation and the emerging broad applications, including in energy conversion and storage, environmental decontamination, flexible electronics, optoelectronics, superconductors, metallic glasses, and magnetic materials, is expected to promote further research and development in academia and industry.

This review presents a summary of the recent advances in data-driven discovery of materials and their innovative applications. First, we introduce the various components of the conceptual framework, including the important stages that guide the data-driven process. Next, we discuss advanced data-intensive strategies and ML algorithms and review material databases and relevant programming tools and platforms used for high-throughput computations. Then, we critically review the descriptors used in the discovery of innovative materials. We present a critical discussion on how data-driven processes are applied to material innovation. Finally, we conclude by providing a perspective on the opportunities and challenges in the field.

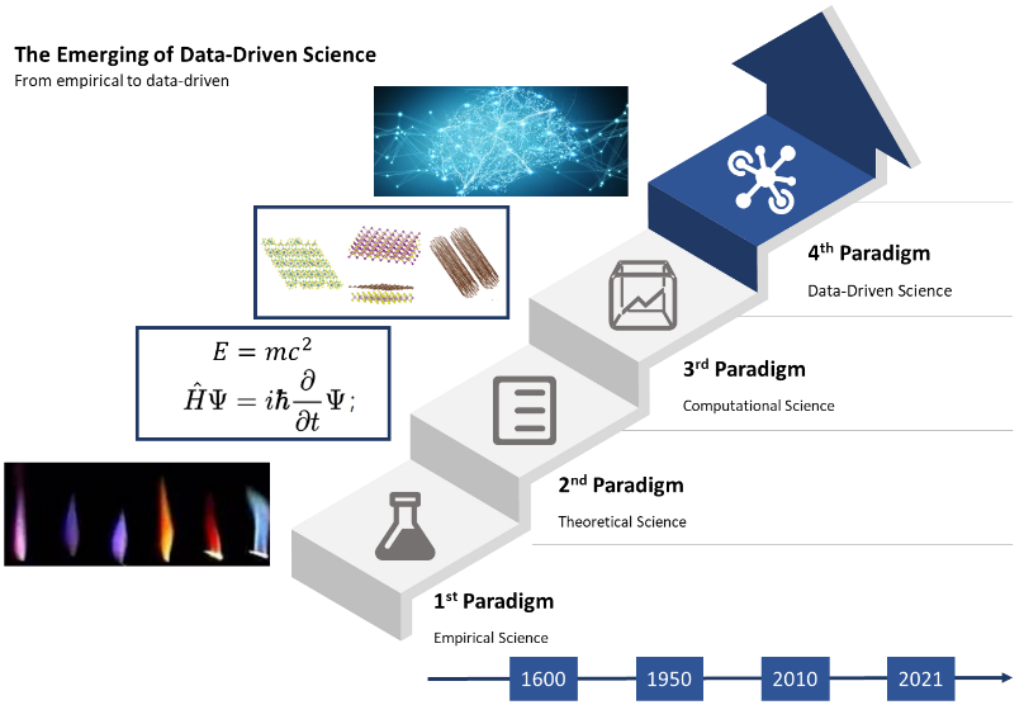


Figure 1. The four paradigms of science evolved along with time, including empirical science, theoretical science, computational science and data-driven science.

2. Frameworks in Data-Driven Innovative Materials Discovery

The development of the data-driven framework for material innovation has been extensively studied by using ML algorithms,[30] material databases,[23, 24, 42] and molecular descriptors.[33, 43] A classical data-driven framework for innovative material discovery typically consists of five fundamental stages: goal identification, data processing, feature engineering, ML and analysis, and application (**Figure 2**).[44] This section describes commonly used frameworks for data-driven processes, including direct design,[27, 45] inverse design,[27, 34] and active learning.[18, 27, 37] Critical stages such as data processing, feature engineering, and ML model training facilitate the utilization and processing of material data and molecular descriptors and the effective implementation of ML algorithms.[46, 47]

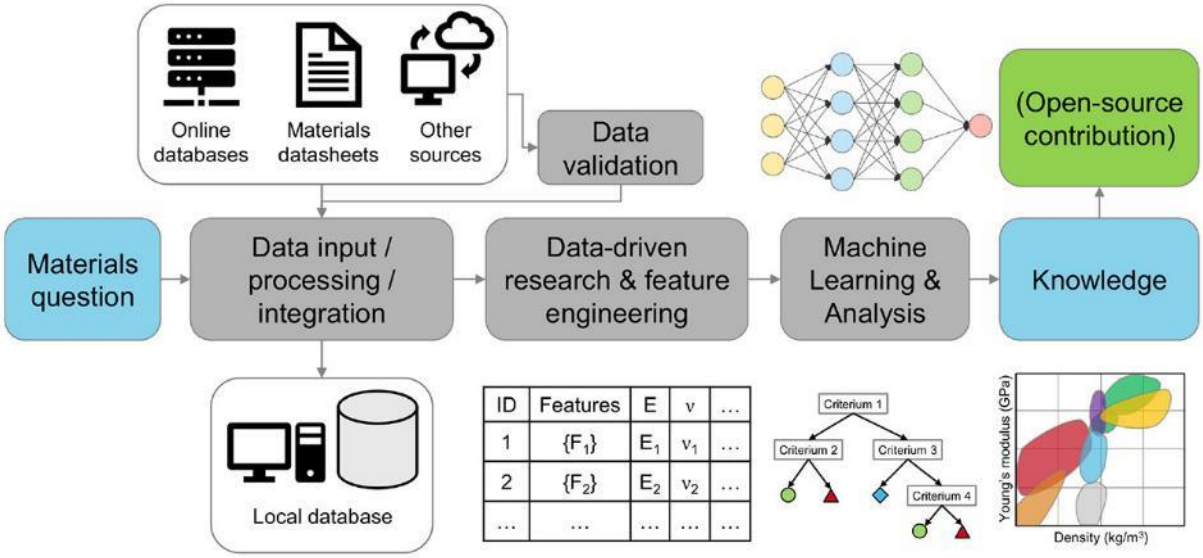


Figure 2. The schematic of a ML workflow in the data-driven innovative material discovery process. Reproduced with permission.[44] Copyright 2020, ACS Publications.

The design and selection of the data-driven framework depend on the application and the material. Although ML can be potent and effective in a data-driven process, it is not the panacea to solve all challenges in materials science.[46] ML models cannot find solutions to questions that are ill-posed or not appropriately expressed. An in-depth and comprehensive understanding of the chemistry phenomena is necessary to accurately describe the question and relate it to a clear goal. The goal of a data-driven process should be specific, measurable, attainable, relevant, and timely.[5] Different ways of defining the goal will lead to varying outcomes of the data-driven process. For example, for the discovery of high-performance photovoltaic materials, Lu et al.[36] employed ML to predict the bandgap of candidate materials, whereas Padula et al.[48] predicted the power conversion efficiency. The nature of the question is also vital for designing the data-driven framework; using a classification model to explore the correlation between target properties and input features or a regression model to distinguish between several categories of materials is difficult. For instance, Jin et al.[17] applied a classification ML model to screen two-dimensional photovoltaic materials with suitable power conversion efficiencies, whereas Sahu et al.[49] employed a regression ML model to predict the power conversion efficiency of candidate photovoltaic materials. Thus, the design of a suitable data-driven framework requires the customization of data processing, feature engineering, and ML model deployment based on the questions being appropriately posed.

2.1. Frameworks for the Overall Data-Driven Process

The data-driven process framework organizes and integrates the fundamental stages of processing data,[50] generating molecular descriptors [33] and deploying the ML model.[44] Such frameworks determine the data flow and the interaction style between the theory and experiments or computations.[27, 34, 38] In this section, we introduced the most commonly employed frameworks to support the discovery of innovative materials including direct design, inverse design and active learning. As illustrated in Figure 3a and 3b, direct and inverse design differ from one another in terms of the direction assumed by predictions between material structure and target functionality.

Active learning (Figure 3c) focuses primarily on data flow in the dynamic iteration loop to improve and accelerate the search and prediction process.[18, 37] Alternative data-driven frameworks have also been reported in light of specific material phenomena to be addressed. For example, regression and classification models could be assembled into a single framework to enable high-throughput materials screening.[31] A transfer learning model could also be integrated into the framework to solve for a small data problem data within the broader the data-driven process.[51]

It is worth noting that the ML techniques in such data-driven frameworks extend far beyond property prediction and pattern recognition.[26, 52, 53] They can be utilized in other fundamental stages to generate features,[54] evaluate feature importance,[10] and visualize data.[18] In both direct and inverse design, the selection of ML algorithms influences the framework architecture.[34]

2.1.1. Direct Design

Direct design is the conventional approach to material discovery and primarily involves measurement and theoretical interpretation of the target property.[27] This trial-and-error approach involves searching for the material demonstrating the targeted functionality within the chemical space, which the prior knowledge can help constrain.[5] Analogous to the structure-property relations derived by data-driven approaches, the direct design approach typically employs the structural features of known materials to predict target properties. Though direct design is widely employed, it presents obstacles to deliberate discovery. For example, as the direct design initiates from a known structure, it is unable to arrive at materials whose structure is not known *a priori* but may possess the desired properties.[27] The case-by-case searching characteristic of direct design is both time- and cost-intensive when extensive structure screening is employed to involve as many materials as possible.[55, 56]

As asserted by Zunger,[27] direct design could be classified into descriptive and predictive approaches. Descriptive direct design employs both modeling and theory to interpret and confirm experimental observations. The predictive direct design, however, can be sub-divided into property prediction for a specific material, or candidate material search in a material space. For example, Jin et al.[17] applied a data-driven predictive direct design framework, screeing 26 out of 187,093 inorganic crystal structures as potential photovoltaic candidates. The blue squares at the bottom of the graph of Figure 3a illustrate known compounds with specified compositions (presented by atom numbers Z_A and Z_B), while question mark-labeled region corresponds to unreported compounds. The upper plot of Figure 3a represents the value of specific material properties as a function of Z_A and Z_B. In a direct-design-based data-driven framework, the materials discovery journey follows the path from the bottom part of the graph to the top part.

2.1.2. Inverse Design

Inverse design can be regarded as the opposite of direct design.[56] In an inverse-design-based data-driven framework, the workflow is initiated in the functional space and terminates in the chemical space.[27] Its objective is to discover tailored materials with desired properties without the exploration of large material space.[56] In the inverse design framework, the target functionality is used as the input to predict the corresponding material structure. Rather than arriving at a unique structure with the desired functionality, the goal is to determine a distribution of probable

structures. For instance, Dudiy et al.[57] employed inverse design in conjunction with specified target properties (e.g. deepest nitrogen level), followed by a search for a desirable material structure.

High-throughput virtual screening (HTVS) is one of the earliest employed methods in inverse design. However, HTVS analysis is generally applied to a smaller number of structures in the course of exploring various functionalities.[27] More recently, generative models, a class of ML method involving the implementation advanced algorithms, including variational autoencoders (VAEs),[58] generative adversarial networks (GANs),[59] recurrent neural network (RNN),[60] and reinforcement learning,[61] are commonly employed in inverse design to determine the molecular structure and the probability distribution both of material elemental parameters and desired target properties (Figure 3b). For example, Jin et al.[62] propose a VAE-based inverse design framework to generate graphs of molecular structure. Inverse design represents an advanced, effective data-driven framework for the discovery of novel materials; open research questions remain, including formulation of the molecular presentation in the inverse design process.[34]

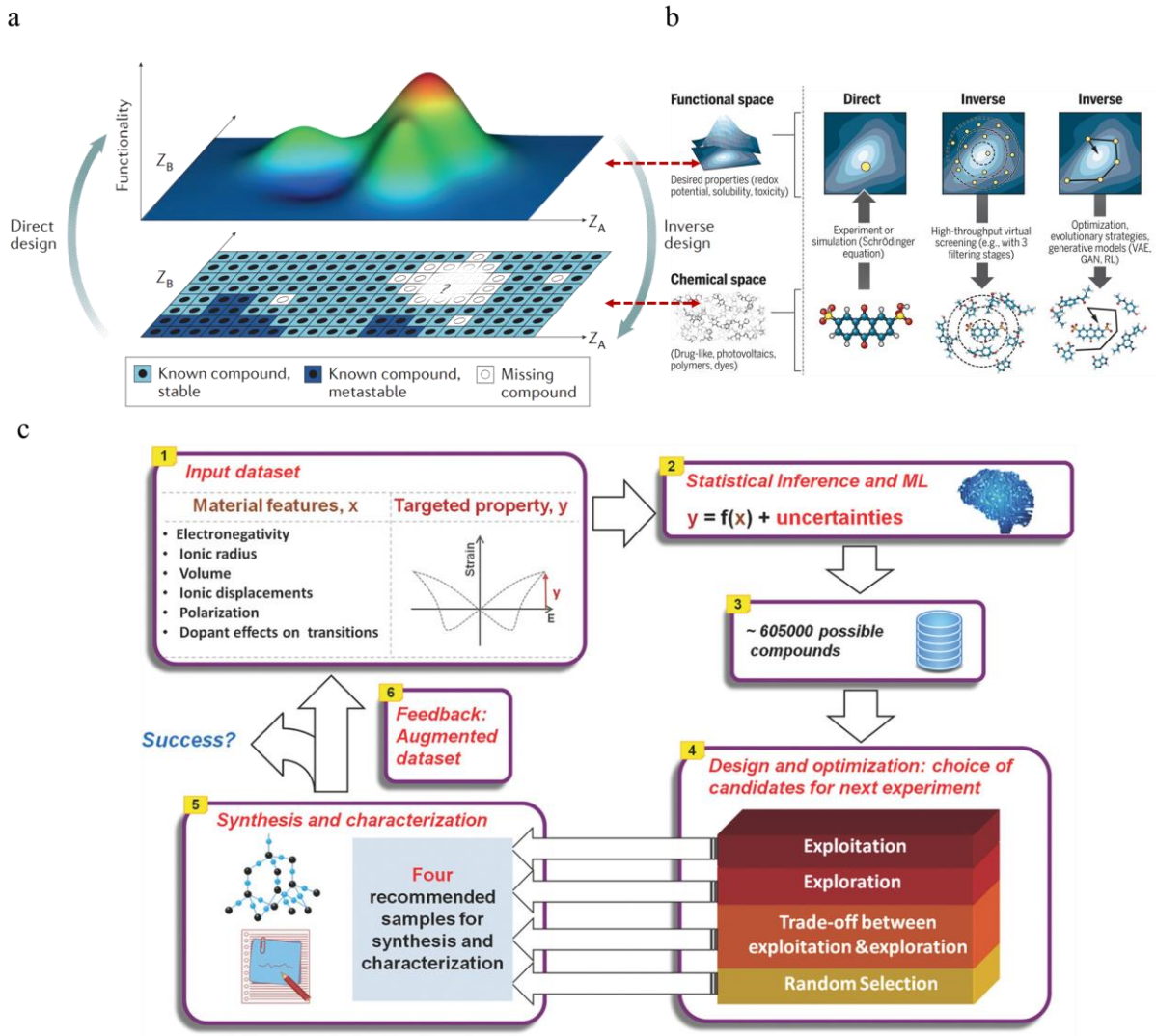


Figure 3. a) Direct and inverse methods for the design and discovery of materials. Reproduced with permission.[27] Copyright 2018, Springer Nature Publications. b) The schematic of direct design and inverse design with different targets in material design and discovery. Reproduced with permission.[34] Copyright 2018, AAAS Publications. c) The active learning framework for the discovery

of materials with high electrostrains. Reproduced with permission.[37] Copyright 2018, Wiley Publications.

2.1.3. Active Learning

The essential idea of active-learning-based data-driven frameworks is to provide high-performance ML models with less training; the machine selects its own training dataset[63] In an active learning framework, the stages of ML training, data processing, and the generation of new training sets are iteratively combined.[18, 37, 64] For instance, Zhong et al.[18] proposed a random-forest-based active ML framework that iteratively trained more than 300 ML models to predict the binding energy of carbon monoxide on the surface of catalyst for the carbon dioxide reduction reaction (CRR). The trained ML model indicated promising adsorption sites during their active learning workflow, which guided the DFT computation for the subsequent iteration. The DFT results evaluated in the latest iteration were combined with the original data to construct a new training dataset, which would yield an updated ML model.

In general, an active learning framework contains an inquiry loop to guide further experiments or computations.[63, 64] Active learning is most applicable when numerous data instances and their labels are easily collected, synthesized or computed to address queries in iterative training.[63] In an active learning framework proposed by Yuan *et al*.,[37] the electrostrain of piezoelectric candidates were iteratively queried. Such active learning frameworks are suitable for dynamic optimization problems and sequential design in innovative material discovery.

2.2. Fundamental Stages in Data-Driven Framework

A complete data-driven material discovery framework involves fundamental stages including raw data processing,[32, 50] feature engineering,[33] and ML model training (Figure 2).[44] In the data processing stage, there are two major steps: data acquisition and data pre-processing.[65] Generally, there are two types of data utilized in a data-driven material discovery process: experimental data and computational data.[5] Both could be either self-generated or queried from existing databases. Relevant, sufficient, consistent and complete data is the foundation of a successful data-driven process.[32] Collected data may contain a number of issues including missing, redundant, abnormal or imbalanced data.[32] Data pre-processing ensures that the ML model performs satisfactorily. Data pre-processing generally consists of four main stages: outlier detection, data complementation, discretization, and normalization.[47] Data may exist in various forms, including numerical values, structure graphs, images, text, or signals. For example, Lee et al.[22] trained a deep learning model to predict potential defects in electron microscopy images with aberration-corrected scanning transmission taken as the model input. Both the quantity and quality of data influence the selection and performance of ML models. For instance, neural network models typically require more data to be reliably implemented.[44] It is critical to acquire material data from reliable sources; commonly used material databases and relevant data management tools are systematically discussed in Section 4.

Feature engineering is the process of constructing the descriptor space, which mainly consists of two steps: the selection or generation of descriptors; construction of the descriptor space.[66] The selection of descriptors depends on the goal of the data-driven process and is characterized by the

greatest extent of human intervention. The target of this step is to identify and extract the most appropriate and critical descriptors from the pre-processed data to construct descriptor space. Problem-specific domain knowledge is essential here, for example, to specify the relevant properties and determine the proper scale length (atomistic, coarse-grained, and global).[32] However, there may be situations in which no suitable descriptor is available, or the basic descriptors are not sufficient to describe the environment or frame the materials with respect specific targets. Thus, an alternative is to generate high-performance descriptors from the original ML training dataset. A good descriptor space is one that is sufficient for the prediction and resolution of the target functional space.[5] Therefore, an in-depth review of molecular descriptors is presented in Section 5 to offer insights on the construction of descriptor space.

ML model training, which follows the construction of the descriptor space, includes model selection, evaluation, and optimization.[67] The implementation of the majority ML algorithms requires the specification of hyperparameters which determine the ML model configuration of ML.[68] Various hyperparameters result in different model formulations; model selection aims at identifying with the appropriate hyperparameter formulation which results in the best model performance. Therefore, hyperparameter tuning is critical to model optimization; it controls the complexity and flexibility of the model to identify the balance between overfitting and underfitting by handling the variance-bias trade-off.[32] More complex models tend to fit training data better but also exhibit a higher variance on the test data, whereas a simpler models (such as regularized linear regression) tends to exhibit a higher bias on the test data. Hyperparameter tuning and model selection can be classified as a meta-optimization task,[68] where validation techniques are employed to evaluate the performance in terms of the ML algorithm objective function.

2.3 Model Performance Evaluation and Uncertainty Quantification

The ultimate goal of the ML model deployment stage is to train the model such that offeres accurate predictions for both test and unseen data; therefore, it becomes essential to effectively assess the performance while characterizing the inherent uncertainty of the model.[32, 69] A review by Morgan and Jacob[69] gives an excellent overview and sample cases of best practices in ML model development, assessment and uncertainty quantification. In this subsection, we will discuss model performance evaluation methods and uncertainty quantification in the context of model deployment, focusing on commonly employed validation techniques and performance evaluation metrics.

2.3.1. Performance Evaluation Techniques

Three techniques are commonly employed for model performance evaluation: holdout,[68] cross-validation (CV),[70] and bootstrap.[32] In most ML deployment processes, the data are divided into training data, validation data, and test data.[69] The holdout approach statically splits the available data for training, validation, and testing at a fixed ratio. Though the holdout approach is straightforward, it may introduce pessimistic bias when the size of the original dataset is small; such splitting further reduces the size while potentially impacting the statistics of the training data. CV represents a continuous, iterative, crossing-over training and validation process that can be regarded as the ensemble of the holdout approach, sampling data without replacement.[68] For a typical *k*-fold

CV process, the dataset is divided equally into *k* parts, one of which is adopted as the validation set; the remaining $k - 1$ parts are combined into a new training subset. When the number of folds is equal to the data points ($k = n$), a special case of CV is manifested (the leave-one-out cross-validation (LOOCV), which, though computationally expensive, is useful when the dataset is small.[32] Sahu et al.[71] applied the LOOCV to 280 data points of small molecule OPV systems to evaluate ML model predictions of power conversion efficiency. Unlike CV, bootstrap samples data with replacement result in only approximately 63.2% of the data points being sampled[72] and potentially a high bias given that the sampled data is not representative of the complete dataset. To correct this bias, Efron[73] has proposed a 0.632(+) bootstrap approach. In general, CV provides a nearly unbiased estimator with high variance, while bootstrap approaches tend to yield estimators with low variance for a small dataset.[73, 74]

2.3.2. Performance Evaluation Metrics

The determination of performance metrics is essential for ML model evaluation and optimization. For regression models, commonly employed metrics are the mean absolute error (MAE), mean square error (MSE), root mean square error (RMSE) and coefficient of dependence (R^2), which are expressed as follows:[5, 29]

$$MAE = \frac{1}{N}\sum_{i=0}^{N}|y_i - \hat{y}_i|\,, \#(2.1)$$

$$MSE = \frac{1}{N}\sum_{i=0}^{N}(y_i - \hat{y}_i)^2\,, RMSE = \sqrt{MSE}, \#(2.2)$$

$$R^2 = 1 - \frac{\sum_{i=0}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{N}(y_i - \bar{y})^2}, \#(2.3)$$

where N refers to the number of sample data points, $y_i, \hat{y}_i, and\, \bar{y}$ represent the actual value, predicted value, and mean value, respectively. The MAE treats the errors equally, whereas larger errors are allocated a higher weight in the MSE and RMSE. The MSE and RMSE are differentiable and commonly used to identify minima optimization processes. R^2 represents the proportion of the variance in true values relative to the predicted values.

The predictivity of classification models can be described by the value of four indicators: true positive (TP), true negative (TN), false positive (FP), and false negatives (FN).[75] Frequently employed evaluation metrics, including Accuracy, Precision, Recall, and F1, can be derived based on the four indicators. Numerous misjudgments resulting in false positives contribute to low precisions, whereas missing of positives correspond to low recalls. A combined metric, called the F1 score, balances these two metrics and is beneficial for cases in which the data is imbalanced.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \#(2.4)$$

$$Precision = \frac{TP}{TP + FP} \#(2.5)$$

$$Recall = \frac{TP}{TP + FN} \#(2.6)$$

$$F_1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \#(2.7)$$

The receiver operating characteristic (ROC) curve and the area under the curve (AUC) are also effective performance metrics in binary classification. The ROC represents the plot of the true positive rate (TPR) versus the false positive rate (FPR), where the formulas for TPR and FPR are presented as follows:

$$TPR = \frac{TP}{TP + FN}, \#(2.8)$$

$$FPR = \frac{FP}{FP + TN}. \#(2.9)$$

A perfect binary classifier would demonstrate an AUC=1; AUC = 0.5 indicates that the binary classifier is no better than random guessing.[30]

2.3.3. Domain of Applicability and Uncertainty Quantification

The reliability and accuracy of the trained models must be evaluated by considering domain applicability and quantifying uncertainties.[69] to the determination of domain applicablity relates to distance metrics between the potential and training data points. Though many methods have been proposed to measure such distances,[76, 77] they are relatively difficult to implement to obtain qualitative guidance on model applicability. All such methods rely upon calculated distance metrics whose validity has not been determined for the particular problem, while also requiring the definition of suitable thresholds.[69]

Predicted value uncertainties are more intuitive and readily quantified to enable the evaluation of model performance. Evaluating error bars is an important tool to support model comparisons, stability estimation and of the reliability of model predictions.[32] Ensemble approaches are commonly employed to quantify uncertainties; a popular methodology involves training the same model via bootstrap or CV, and then treating the ensemble variance as a surrogate for the error bars.[78] An alternative approach involves utilizing the same training data while refitting the model by adjusting the model architecture.[69] A large variance between these predictions in a specific chemical domain indicates that the ML models are still tangling and require additional training data.[79] The two types of ensemble methods can also be combined in random forest decision tree models, for which Morgan and Jacobs provide an in-depth example.[69] The ensemble approaches are more computationally expensive; however, their flexibility enables them to be employed in numerous models.

Prediction uncertainty can also be quantified by distance-based approaches, which are based on the concept that such uncertainties correlate with the distance between the potential corresponding training data points. Hirschfeld et al.[80] employed log-scaled Tanimoto distance[81] and Euclidean distance[82] to quantify the displacement between potential points from training data and predictions of molecular properties, respectively. Bayesian approaches [83] can also automatically quantify uncertainty while potentially avoiding iterations, though this requires the adoption of specific ML models making it less generally applicable.[32, 69]

3. Data-Intensive Strategies and Algorithms for Innovative Materials Discovery

Recent developments in materials science have corresponded to a large amount of accumulated

data from both theoretical and experimental studies.[84] However, how to identify appropriate techniques to process this accumulated data to shed light on implicit correlations and guide the course of future studies remains an open question, impeding the advancement of materials science. To address this, various algorithms have been proposed in previous decades for obtaining solutions to practical data-processing problems in materials science.[85] As the foundation of the data-driven study of materials science, these algorithms were introduced with various intentions. In this section, several algorithms and methods, including supervised learning,[85] unsupervised learning,[86] and deep learning,[87] for materials science study are discussed with relevant examples.

3.1. Supervised Learning: Regression and Classification

Supervised learning is a learning strategy for problems in which both inputs and outputs are given. The goal of supervised learning is to identify the function which best maps inputs to outputs consistent with the given data.[85] The methods of supervised learning can be categorized into regression and classification, including linear regression, logistic regression, support vector machine (SVM), and decision tree.

3.1.1. General Linear Regression Algorithms

Linear regression algorithms are a common component of ML[88] and are widely used to build prediction models which connect input scalars to continuous output values. The most common regularized models of multivariate linear regression are: ridge regression; least absolute shrinkage and selection operator (LASSO).

(Multivariate) Linear Regression

First, we discuss a common form of (multivariate) linear regression, which is the basis for advanced versions.[88] Linear regression is based on the assumption that the relationship between the input data matrix $\boldsymbol{X}$ and dependent variable $\boldsymbol{y}$ is linear:

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i \#(3.1)$$

which can be more succintly expressed as

$$\boldsymbol{y} = \boldsymbol{\beta X} + \boldsymbol{\varepsilon} \#(3.2)$$

where $\boldsymbol{X}$ is the input (or sometimes called the design matrix), which could either be a row matrix or n dimensional matrix. $\boldsymbol{\beta}$ is the vector of regression coefficients, which is usually estimated from the least square method; hence,

$$\hat{\beta} = arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{\beta X}\|_2^2 = arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} - y_i\right)^2 \#(3.3)$$

with $\|\cdot\|_2^2$ denoting the square of the L2 norm and can be expanded as:

$$\|\boldsymbol{y} - \boldsymbol{\beta X}\|_2^2 = (\boldsymbol{y} - \boldsymbol{\beta X})^{\boldsymbol{T}}(\boldsymbol{y} - \boldsymbol{\beta X}) = \mathbf{y}^{\boldsymbol{T}}\mathbf{y} - \mathbf{y}^{\boldsymbol{T}}\boldsymbol{\beta X} - \boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\mathbf{y} + \boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\beta X} \#(3.4)$$

The optimal solution is obtained by taking the partial derivative of the aforementioned expression with respect to $\boldsymbol{\beta}$:

$$\frac{\partial(\|\boldsymbol{y}-\boldsymbol{\beta X}\|_2^2)}{\partial \boldsymbol{\beta}}=\frac{\partial\left(\mathbf{y}^{\boldsymbol{T}}\mathbf{y}-\mathbf{y}^{\boldsymbol{T}}\boldsymbol{\beta X}-\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\mathbf{y}+\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\beta X}\right)}{\partial \boldsymbol{\beta}}=-\mathbf{2}\mathbf{y}^{\boldsymbol{T}}\boldsymbol{X}+\mathbf{2}\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\#(3.5)$$

Setting (3.5) equal to zero yields the optimal $\boldsymbol{\beta}$, such that

$$\widehat{\boldsymbol{\beta}}=\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-\mathbf{1}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}\#(3.6)$$

This approach to linear regression has been widely used in materials science in cases where the predicted values are linearly associated with the input features. For instance, Winkler et al. employed linear regression to construct nano quantitative structure-activity relationship (nano-QSAR) models of the biological effects of nanoparticles.[89] Fernandez et al. also utilized linear regression to explore the electronic properties of graphene.[90, 91] Jinnouchi et al. developed a linear regression model to predict the catalytic activity associated with direct NO decomposition on the surface of RhAu alloy nanoparticles.[92]

Ridge Regression

Introducing a regulizer to the linear regression model prevents overfitting while reducing the overall complexity of the model. In linear regression, when the input matrix $\boldsymbol{X}$ is a singular matrix (the number of features is larger than the number of samples), errors may emerge when calculating $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$. Hence, the ridge regression approach was proposed, which adds small positive quantities $\lambda\boldsymbol{I}$ to $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$,[93] yielding:

$$\widehat{\boldsymbol{\beta}}=\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}+\lambda\boldsymbol{I}\right)^{-\mathbf{1}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}=\ arg\min_{\boldsymbol{\beta}}\left(\|\boldsymbol{y}-\boldsymbol{\beta X}\|_{\mathbf{2}}^{\mathbf{2}}+\lambda\|\boldsymbol{\beta}\|_{\mathbf{2}}^{\mathbf{2}}\right)\#(3.7)$$

The improved performance of ridge regression has enabled the prediction of various material properties. For example, González et al. utilized ridge regression (as one of the three methods) to predict the surface plasmon resonance of perfect and concave Au nanocubes.[94]

LASSO

Another regularized linear regression method is the LASSO. Unlike the ridge regression based on the L_2 norm, the LASSO employs the L_1 norm as a regularizer.[95]

$$\widehat{\boldsymbol{\beta}}=\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}+\lambda\boldsymbol{I}\right)^{-\mathbf{1}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}=\ arg\min_{\boldsymbol{\beta}}\left(\|\boldsymbol{y}-\boldsymbol{\beta X}\|_{\mathbf{2}}^{\mathbf{2}}+\lambda\|\boldsymbol{\beta}\|_{\mathbf{1}}\right)\#(3.8)$$

When the λ is small enough, some of the coefficients will be forced to reduce to 0, making the LASSO sparsity to rapidly filter the input features. For instance, when predicting the bandgap of functionalized MXenes, Singh et al. employed the LASSO to reduce the number of input features from 47 to 15, which significantly enhanced the efficiency of the ML model.[96]

3.1.2. Logistic Regression

The logistic regression is based on logit function $u=\log(\frac{y}{1-y})$ as a link function.

$$\log\frac{\boldsymbol{y}}{\mathbf{1}-\boldsymbol{y}}=\boldsymbol{\beta X}\#(3.9)$$

This is a classical binary classification model to simulate the probability or possibility of a certain event or class. The combination of multiple logistic regression models can achieve the multiclass classification.[97] Instead of being embedded into other algorithms as a classifier, the logistic regression can directly be used as a classifier within a materials study, such as in the study of the relationship between Fermi energy and structural/morphological features of Ag nanoparticles,[98] or as a control group to predict the structure-property relationship of Pt nanoparticle catalysts.[99]

3.1.3. Support Vector Machine (SVM)

The SVM algorithm is widely used in materials science owing to its excellent performance in data pattern recognition and classification.[100] The SVM implements the structural risk minimization principle to the upper limit of the generalization error (eq. (3.10))[101] and demonstrates excellent performance for samples based on high-dimensional data, or when the sample size is small. SVM models effectively overcome the "overlearning" problem.[29] The SVM can also be utilized to solve regression problems by introducing an alternative loss function;[102] the SVM-based regression model is referred to as support vector regression (SVR). Fang et al.[103] combined the genetic algorithm (GA) with SVR to predict the extent of atmospheric corrosion in metals such as steel and zinc. Compared with other algorithms, the (GASVR) hybrid method has exhibited improved predictive performance. Other SVR-based hybrid algorithms have also been utilized for various applications; in Ref.[104] a feature-selection-based two-stage SVR (FSTS-SVR) was utilized to develop a predictive model for the Ge_xSe_{1-x} glass transition temperature. Because of the structural variations at the turning point, a two-stage onset glass transition temperature (T_g) model was constructed based on FSTS-SVR to achieve the highest accuracy. This hybrid method has also show potential as an efficient algorithm for the multistage simulation and prediction of characteristic T_g. Chen et al.[105] proposed the use of an SVM algorithm to predict the exposure temperatures of fire-damaged concrete structures. Their SVM simulation demonstrated that the concrete ultrasonic pulse velocity was the most effective parameter in improving the accuracy of estimations. SVM models have been used to predict various other material properties such as ionic conductivities,[106, 107] glass transition temperatures,[108-110] catalyst active sites[111] and adsorption energies,[112] and various other properties of innovative materials.[113, 114]

$$\left[\frac{1}{n}\sum_{i=1}^{n}\max\left(0, 1-y_i\left(\boldsymbol{\beta}^T x_i - b\right)\right)\right] + \lambda\|\boldsymbol{\beta}\|^2 \#(3.10)$$

3.1.4. Kernel Ridge Regression (KRR)

Kernel ridge regression (KRR), which combines ridge regression and the kernel trick, is a simplified version of SVR[115, 116] but utilizes a different loss function. Although both KRR and SVR are based on L2 regularization, KRR employs a loss function based on the squared error loss while SVR utilizes the epsilon-intensive loss. In addition, fitting a KRR model can yield a closed-form solution and is faster than SVR for medium-sized datasets. However, KRR is slower than SVR when learning a sparse model because its learned model is non-sparse. To transfer ridge regression to KRR, the matrix inverse lemma (eq. (3.11)) was introduced to eq. (3.8), yielding eq. (3.12). After a dual variable, α, is specified (eq. (3.13)), the original primal variable, ***β,*** evolves to eq. (3.14); an updated prediction *y**

can be constructed for new set *x** (eq. (3.15)). KRR has been widely used in materials analysis. For instance, Sheremetyeva et al. applied KRR to study the correlation between the stimulated Raman spectrum and twist angle of twisted bilayer graphene.[117] Singh et al. employed different ML models, including KRR, with multiple input features to estimate the bandgaps of MXenes.[96]

$$(\boldsymbol{P^{-1}} + \boldsymbol{B^T R^{-1} B})^{-1} \boldsymbol{B^T R^{-1}} = \boldsymbol{P B^T} (\boldsymbol{B P B^T} + \boldsymbol{R})^{-1} \#(3.11)$$

$$\boldsymbol{\beta} = \boldsymbol{X^T} \left(\lambda \boldsymbol{I} + \boldsymbol{X X^T}\right)^{-1} \boldsymbol{y} \#(3.12)$$

$$\boldsymbol{\alpha} = (\boldsymbol{K_X} + \lambda \boldsymbol{I_N})^{-1} \boldsymbol{y} \#(3.13)$$

$$\boldsymbol{\beta} = \boldsymbol{X^T \alpha} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{\alpha_i x_i} \#(3.14)$$

$$\boldsymbol{y^*} = \boldsymbol{\beta^T x^*} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{\alpha_i x_i^T x^*} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{\alpha_i k(x^*, x_i)} \#(3.15)$$

3.1.5. Gaussian Process Regression (GPR)

Gaussian process regression (GPR), also know as Kriging, is a non-parameteric model that utilizes Gaussian process priors to perform regression analysis. Instead of directly generating the regression function $f(x)$, GPR generates a distribution of an infinite number of functions $f(x)$. For a given dataset D: (X, Y), let $f(x_i) = y_i$, yielding the vector $f = [f(x_1), f(x_2), f(x_3), \ldots, f(x_n)]$. Defining the set of x_i as X*, and the corresponding prediction value as f*, eq. (3.16) can be constructed based on Bayes' theorem.

$$\boldsymbol{p(f*|f)} = \frac{\boldsymbol{p(f|f*)p(f*)}}{\boldsymbol{p(f)}} = \frac{\boldsymbol{p(f, f*)}}{\boldsymbol{p(f)}} \#(3.16)$$

GPR has gained significant traction in computational materials science, including in the prediction of atomistic properties such as interatomic potentials.[118] For instance, Singh et al. implemented GPR to position the band edges of MXenes, achieving a minimum root-mean-squared error (rmse) of 0.12 eV.[119] Wee et al. developed an electron-phonon averaged GPR (EPA-GPR) method to efficiently estimate and fast-screen the thermoelectric properties of materials for pre-defined applications.[120]

3.1.6. Decision Tree (DT)

Decision trees (Figure 4) are a classic supervised ML algorithm, which have been widely used for classification and regression in the material design process. Decision trees break up a complex decision into a union of several simpler decisions which, when synthesized, form an operable final solution. Through a series of 'yes' or 'no' questions pertaining to the input descriptors, the decision tree can arrive at the internal relationship between descriptors with relative ease, subdividing the discrete function values into classes with a common label.[66] Decision trees can handle interactions between descriptors as well as various classes of input data (such as numbers and text). The three core steps of decision tree learning are feature selection, decision tree generation, and the pruning of decision tree. Training datasets may offer an abundance of features which offer contribute to

varying extents to the final decision. The goal is to identify highly-related features corresponding to improved classification performance. This step is followed by tree generation; originating from the so-called the root node, the information gain at each subsequent node is calculated. The feature corresponding to the largest information gain will be specified as the node feature; the sub-node will subsequently be established with respect to each value of the feature. However, this approach engenders a high risk of overfitting, which can be mitigated by tree pruning to improve performance. The earliest decision tree model was referred to as the iterative dichotomiser 3 (ID3) algorithm,[121] which was developed by utilizing information gain to select features. A decision tree can also be constructed for regression based on ID3 by replacing information gain with standard deviation reduction.[122] The C4.5 decision tree model was the successor of ID3; it utilized the information gain ratio as the criterion of feature selection.[123] Other than algorithms which have been developed based on information theory, models such as classification and regression trees (CART)[124] utilize Gini impurity, which measures the frequency incorrectly labeled selected elements. Decision trees are have been used extensively in materials science, such as in the analysis of the cytotoxicity of nanoparticles,[125] the prediction of the exciton valley polarization landscape of two dimensional (2D) semiconductors,[126] and the synthesis of metal-organic nanocapsules.[127]

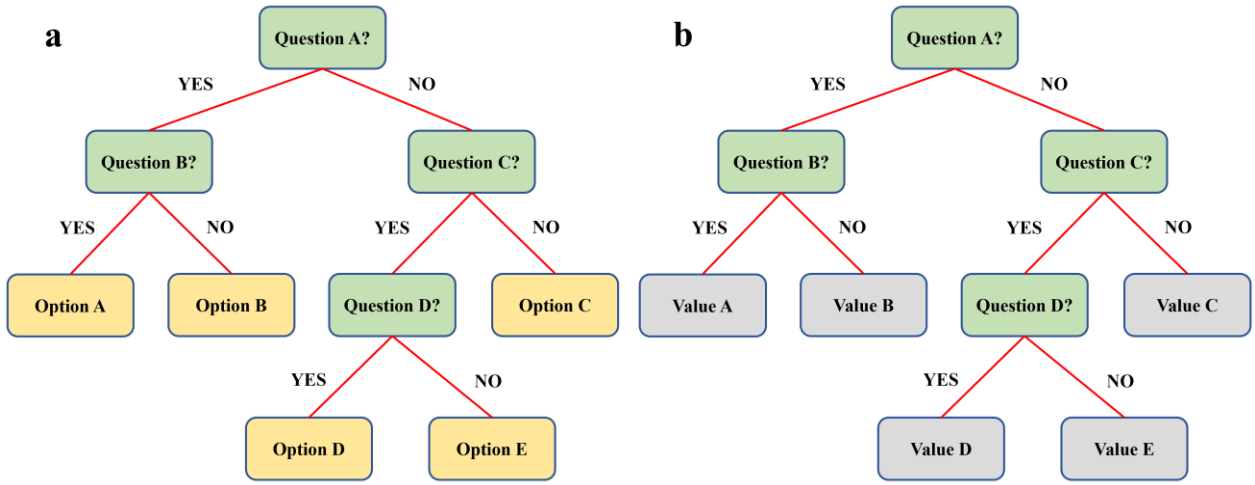


Figure 4. Schematic illustration of decision tree for a) classification and b) regression.

3.1.7. k-Nearest Neighbor (*k*NN)

The *k*-nearest neighbor (*k*NN) algorithm is a non-parametric method for classification and regression.[128] *k*NN is based that an concept that an unlabeled sample can be represented by the nearest *k* labelled samples in feature space. The benefits of *k*NN can be realized without pre-estimation of parameters or training; however, this poses a significant trade-off. The dominance of certain types of samples in the datasets (i.e., an unbalanced distribution of samples) will influence the accuracy of *k*NN. This method is also computationally expensive given that the distances between large numbers of labeled and unlabeled samples are needed. However, *k*NN has found wide application in materials science; for example, Padula et al. applied *k*NN to predict the photovoltaic parameters and efficiency of organic solar cells;[48] Byun et al. employed *k*NN as a basis to build a predictive model for the toxicity of oxide nanomaterials.[129]

3.2. Unsupervised Learning: Clustering and Dimension Reduction

Counter to supervised learning, unsupervised learning focuses on datasets with little or no pre-existing labels.[86] Unsupervised learning can be deconstructed into two primary methods: principal component and cluster analysis. In contexts where the structure-property relationships of materials have not been fully defined, unsupervised learning offers an effective approach to identifying such implicit correlations. Detailed algorithms are be summarized in this section as an overview to unsupervised learning in materials science.

3.2.1. Principal Component Analysis (PCA)

Principal component analysis (PCA) aims to simplify data.[130] PCA can be characterized as an orthogonal linear transformation method that projects data onto a new coordinate system. The first component is yielded according to the eq. (3.17):

$$\boldsymbol{w}_1 = arg \max_{\|w\|=1} \left(\|\boldsymbol{X}\boldsymbol{w}\|_2^2 \right) = arg \max_{\|w\|=1} \left(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} \boldsymbol{w} \right) \#(3.17)$$

where $\boldsymbol{w}_1$ is an array of 1 × m dimensional weights.

Subsequent components can be obtained based on eq. (3.18):

$$\boldsymbol{w}_k = arg \max_{\|w\|=1} \left(\left\| \widehat{\boldsymbol{X}}_k \boldsymbol{w} \right\|_2^2 \right) = arg \max_{\|w\|=1} \left(\boldsymbol{w}^{\mathrm{T}} \widehat{\boldsymbol{X}}_k{}^{\mathrm{T}} \widehat{\boldsymbol{X}}_k \boldsymbol{w} \right) \#(3.18)$$

where

$$\widehat{\boldsymbol{X}}_k = \boldsymbol{X} - \sum_{i=1}^{k-1} \boldsymbol{X} \boldsymbol{w}_i \boldsymbol{w}_i^{\mathrm{T}} \#(3.19)$$

The power of PCA is in extracting the most important information from datasets to reduce the dimensionality of input features and further compress the size of datasets.[131] In exploring the structure-property relationships of materials, a large number of structural features will be initially considered, whereas a relatively small subset of these features contribute meaningfully to the particular material property. The utilization of PCA to reduce the dimensionality of structural features has been demonstrated to significantly enhance the efficiency of investigations of structure-property relationships.[52, 117, 132]

3.2.2. Expectation Maximization (EM)

Expectation maximization (EM) is an iterative strategy to estimate the maximum posterior (MAP) (i.e. maximum likelihood) of parameters in statistical models that depend on latent variables.[133] EM algorithms can be used for data processing; for example, Benammar et al. integrated the EM algorithm with split spectrum processing to develop ultrasonic methods to process signals for the detection of delamination defects in carbon-fiber-reinforced polymer-multilayered composite materials.

3.2.3. k-means Clustering

k-means clustering is another unsupervised learning method that has been widely applied. The

aim of k-means clustering is to assign M data points in N dimensions into k clusters.[134] The clustering process is designed to find the minimum sum of the distance between each data point and its corresponding cluster center. Hence, the selection of k is critical to the success of the k-means clustering algorithm. PCA can be introduced to guide the selection of k by reducing the dimension of features. Since k-means clustering does not rely on the prior knowledge to assign data, it is suitable for the identification of implicit structure-property relationships by clustering unlabeled data. Such methods have been utilized to great effect in various studies: Darr et al. utilized k-means clustering for high throughput data collection and characterization for synthesized nanomaterials;[135] Neumayer et al. employed k-means clustering to group data processed by PCA to support the study of ferroelectric properties of layered $CuInP_2S_6$.[132]

3.2.4. t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-distributed stochastic neighbor embedding (t-SNE) is a type of ML method that is compatible with data visualization[53] by reducing high-dimensional data to two or three dimensions for visualization. t-SNE is a popular method for data analysis in various domains. Given that multiple structural features are investigated in ML-based materials investigation, the most common application of t-SNE is to visualize the high dimensional features as low dimension images.[136] However, t-SNE has also been extensively utilized more substansively, such as for the prediction of nanoparticle structure-property relationships [137] and the exploration of optimal microstructures for targeted properties.[138]

3.3 Deep Learning

Deep learning, as a new branch of ML, has been widely used for various applications, including natural language processing (NLP), computer vision, and data mining.[87] Data can be represented with multiple levels of abstraction based on computational models consisting of multiple processing layers. Various deep learning methods (**Figure 5**) have been effectively applied to the investigation of materials properties. This section presents relevant algorithms.

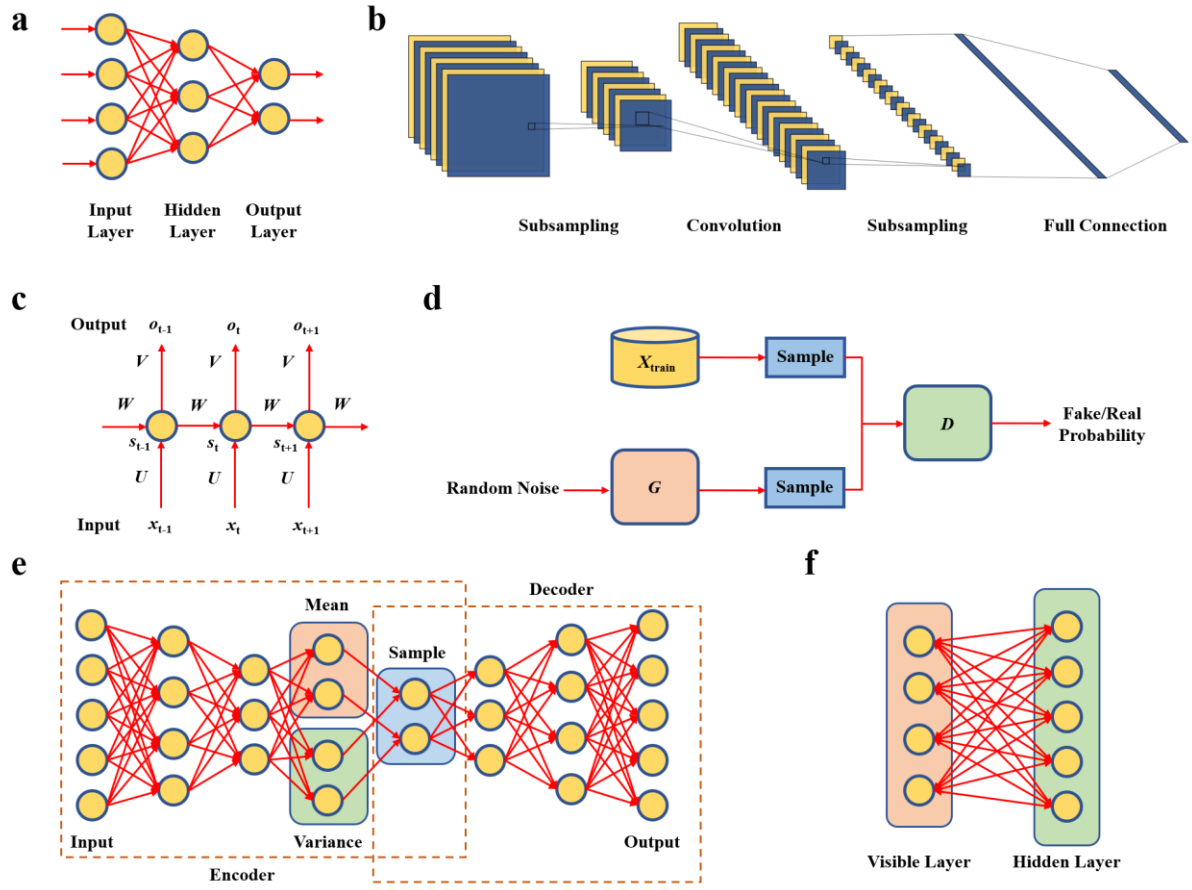


Figure 5. Schematic illustration of basic models of a) ANN, b) CNN, c) RNN, d) GAN, e) VAE, and f) RBM.

3.3.1. Artificial Neural Network (ANN)

As a nonlinear statistical analysis approach, the ANN (Figure 5a) algorithm is capable of self-learning and adaption.[139] The ANN is the most common neural network and is applicable to a wide range of problems. Back propagation (BP), an example of the ANN, has been widely used to predict various material properties: tensile and temperature responses, wastage, elongation, compressive properties, and corrosion properties.[140-142] Chen et al.[143] compared the performance of linear regression and BP-ANN for the prediction of polymer glass transition temperatures, finding that BP-ANN has a much lower average prediction error (17K) than that of linear regression (30K). Because BP-ANN does not require extensive background knowledge of structural properties, it is advantageous in the development of solutions with a specified degree of prediction error tolerance and good generalizability. However, BP-ANNs are characterized by a slow convergence rate and sometimes may be trapped into the local, rather than global, minima. These shortcomings can be overcome by combining ANN with the radial basis function (RBF-ANNs) to enable high convergence rates while avoiding local minima trapping. Gajewski and Sadowski[144] applied RBF-ANN to investigated crack propagation in layered bituminous pavement, ultimately detecting a strong positive correlation between B2 bituminous layer thickness and extent of cracking. ANN algorithms

have also found a place in other applications; for example, such algorithms have been successfully applied to predict the density and viscosity of biofuel compounds.[145] Scott et al.[146] demonstrate the effectiveness of ANN in predicting the oxygen diffusion properties of ceramic materials to support the development of new materials suitable for environmental applications (such as clean energy production and technologies for the reduction of greenhouse gas emissions). ANN has also been applied to accurately predict excited-state energies,[147] melting points,[148] diffusion barriers,[149] and other functional features.[150-152]

3.3.2. Convolutional Neural Network (CNN)

CNNs (Figure 5b) represent another class of deep neural network and are most commonly used for visual imaginary analysis.[153, 154] An advantage of the CNN is weight sharing, indicating its ability to process high dimensional data; another important advantage is automatic feature extraction, corresponding to favorable performance for feature classification. In materials science, CNNs can be directly used to process the images generated using various techniques to enable the analysis of materials structures. For example, Schiøtz et al. applied a CNN to atomic-resolution transmission electron microscopy (TEM) images identify material local atomic structures;[155] Ziatdinov et al. trained a CNN model to analyze images generated from real-time monitoring by scanning transmission electron microscopy (STEM) to identify lattice defects in WS_2 and map its solid-state reactions and transformations.[156] By transforming crystal structures to crystal graphs, CNN can also be used to accelerate materials discovery[157] and predict novel material properties.[158] Zhang et al. have demonstrated a method that uses CNN trained by periodic table attributes to predict a variety of material properties including lattice parameters, enthalpy of formation, and compound stability.[159]

3.3.3. Recurrent Neural Network (RNN)

Recurrent neural networks (RNN, Figure 5c) differs from ANN and CNN, by accounting for temporal sequences.[60] The current status of an RNN cell is influenced not only by the current inputs, but also by its previous status. RNN is usually used for the study of NLP,[160] image generation,[161] etc. RNN has been extensively used to study the kinetics of chemical reactions,[162] which are fundamentally path-dependent. Recently, Shin et al. employed RNN to accelerate the generation of atomic data in traditional *ab initio* molecular dynamics (AIMD).[163] The RNN model trained by AIMD enabled the prediction of atomic velocities and Si atomic positions.

3.3.4. GAN

A GAN (Figure 5d) was first proposed by Goodfellow et al. in 2014.[59] The objective of GAN is to build a generative model, *G* (capture the data distribution), and a discriminative model, *D* (distinguish the sample from either the training set or model *G* with an estimated probability), and find an equilibrium solution between *G* and *D*; here, *G* recovers the training set and *D* is equal to 0.5 for every sample. GAN is well suited to supporting applications in additive manufacturing, which requires a large number of architectural materials. However, traditional materials design methods, such as bioinspiration, the Edisonian approach, theoretical analysis, and topology optimization, are based on prior knowledge possessed by designers. In contrast, Mao et al. presented an experience-free method based on the GAN algorithm to study Hashin-Shtrikman upper bounds on isotropic elasticity to design complex architecture materials.[164] In another study, Hu et al. employed a GAN-

based method to generate novel hypothetical inorganic materials (which are not recorded in existing databases), to enable the inverse design of such materials.[165] Aspuru-Guzik et al. have also demonstrated numerous cases that utilize generative models for inverse materials design.[34, 166-168] They employed a GAN-based method to study the crystal structures of Mg-Mn-O ternary materials, successfully predicting 23 new crystal structures.[169]

3.3.5. VAE

An autoencoder is a type of ANN that consists of an encoder and decoder.[58] Unlike complete and regularized autoencoders (which are discriminative models), VAEs (Figure 5e) represent a class of generative model.[170] Compared with the GAN, VAE models can be trained with greater ease given their more complete mathematical basis. VAE has demonstrated broad applicability to chemistry and materials science, including in the design of small molecules.[171, 172] Batra et al. employed VAE to study the properties/performance of polymers, especially geared towards the discovery of polymers that are robust under extreme conditions (e.g., high temperatures and electric fields).[173] Stein et al. developed a materials image autoencoder based on the VAE to investigate the optical properties of materials, including prediction of spectra from images and vice versa.[174]

3.3.6. Restricted Boltzmann Machine (RBM)

The restricted Boltzmann machine (RBM, Figure 4f) is a generative stochastic ANNs that can learn the probability distribution of input datasets.[175] The RBM has been widely used for reducing the dimensionality of data,[176] classification,[177] feature learning,[178], etc. The RBM is a special topological structure of the Boltzmann machine (BM) originating from statistical physics. Hence, the RBM can be used to solve the problems in quantum physics.[179] Kais et al. reported a hybrid quantum algorithm based on the RBM to accurate characterize the molecular potential energy surfaces of a small molecule system.[180] Recently, Nomura et al. employed RBM to investigate the synthesis of MoS_2 via chemical vapor deposition (CVD), while obtaining insights into metallic 1T- and semiconducting 2H-MoS_2, and the generation of defects during the growth of MoS_2 by employing the CVD method.[181]

3.4. Ensemble Methods

Ensemble learning is characterized by the construction of a high-performance algorithm by combining a collection of weaker models. Rather than developing new algorithms, existing algorithms are combined to achieve improve results. A collection of simple, basic models is selected for ensemble learning. There are two main approaches to assemble such models: boosting (**Figure 6**a)[182] and bagging (bootstrap aggregating, Figure 6b).[183] The primary difference between the two is the approach to assigning vote weights to sub-models. In boosting, elite models are identified through training and testing; higher vote weights are subsequently assigned to the models with better performance. In contrast, the bagging method is much more democratic in that each model has equal vote weight. In general, results obtained by boosting method are characterized by a lower bias, while those obtained by bagging will be characterized by a lower variance. This section presents various boosting and bagging methods in detail.

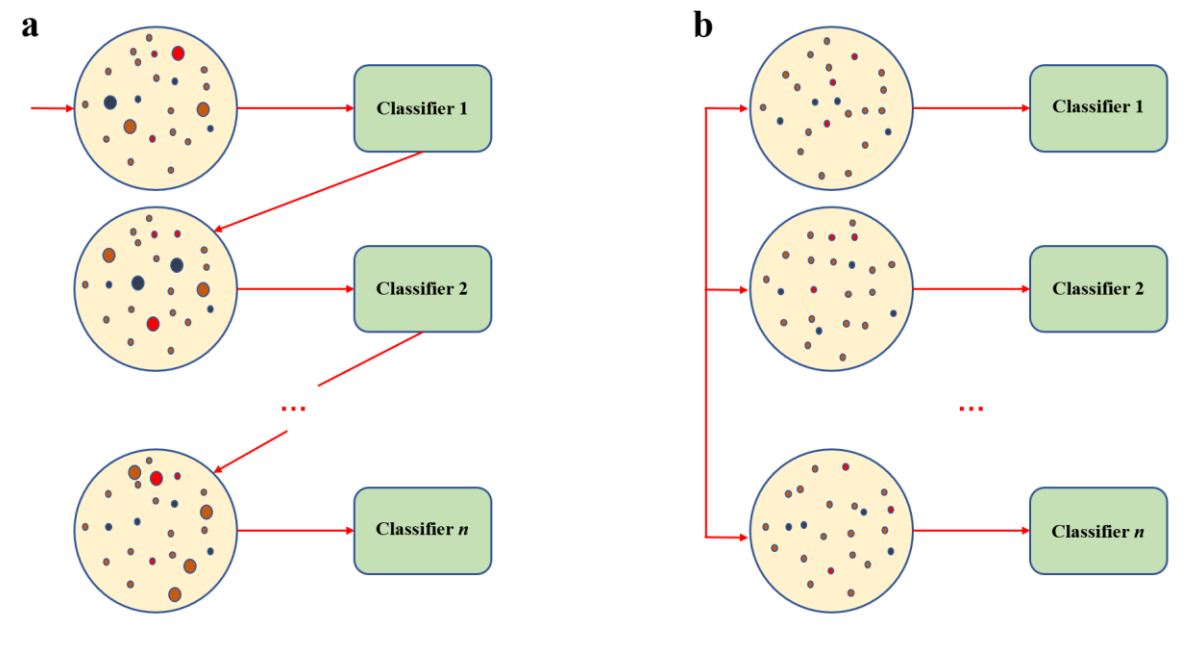


Figure 6. Schematic illustration of general structures of a) Boosting and b) Bagging.

3.4.1. Boosting

Boosting is an algorithm that can be used to reduce the variance in supervised learning,[182] while converting weak learners to strong learners.[184] In boosting, each weak classifier possesses connections to other weak classifiers, collectively yielding a strong classifier. However, the inherent flaw of traditional boosting is that the minimum learning accuracy of a single weak classifier is required as a basis for the improvement mechanism.

AdaBoost

To improve the boosting algorithm, Freund and Schapire developed AdaBoost, which is short for adaptive boosting.[185] The advantage of AdaBoost is that it does not require prior knowledge of weak learners to realize the boosting efficiency. Hence, compared to traditional boosting, AdaBoost is more suitable for practical problems. When training an AdaBoost model, the weight of a sample that is not correctly classified in a current round will be increased in the subsequent round of training, enabling the evolution of a stronger classifier over several iterations. As AdaBoost is easy to operate and resists overfitting, it is liberally used in various contexts. Tonezzer et al. applied AdaBoost to classify different gases detected by a carbon-modified SnO_2 nanowire sensor.[186] Wang et al. implemented AdaBoost to classify carbon nanomaterials based on their TEM images.[136] AdaBoost has also played a key role in the recent rise of artificial chemists. For instance, Abolhasani et al. developed an artificial chemist for the synthesis of quantum dots, in conjunction with the implementation of AdaBoost to enhance performance.[187]

Gradient Boosting

Gradient boosting is also applicable to classification and regression.[188] Gradient boosting utilizes the negative gradient (response) of the cost function of the current model to train the

successive model by iteratively combining weak classifiers, ultimately yielding an optimized model. Gradient boosting has been employed in a number of materials studies. Ma et al. utilized gradient boosting to predict the power conversion efficiency of organic solar cells based on 13 descriptors extracted from the microscopic properties of organic materials.[71] The gradient boosting model demonstrated excellent performance (with a Pearson's coefficient of 0.79). Fazzio et al. implemented gradient boosting to determine the thermodynamic stability of 2D materials,[31] while Wei et al. also implemented a gradient boosting classifier to identify novel 2D photovoltaic materials.[17]

3.4.2. Bagging

Bagging, also referred to as bootstrap aggregating, is another common ensemble method.[183] In contrast with the boosting method, weak classifiers in bagging are individual (not correlated). During the training process, samples are randomly selected and trained for each weak classifier; weak classifiers are then aggregated into a strong classifier. However, the performance of bagging is highly dependent upon the datasets, such that a large bias in the dataset will introduce large bias into the bagging model. The random forest[189] method was developed to address this shortcoming.

Random Forest

The most widely used ensemble learning method is random forest, [189] which is composed of many individual (i.e., not correlated) decision trees to improve prediction accuracy and prevent overfitting. For example, when executing classification tasks, each decision tree in the forest will execute the classification operation for each new input sample. The most classified result will be deemed the final overall result of random forest. The random forest algorithm has been widely utilized in the field of materials science. Zhong et al.[18] trained a random forest model to predict the adsorption energy of CO on the surface of the designed catalysts. Artrith et al.[190] implemented random forest in conjunction with a Gaussian process regression to predict the transition state energy, activity, and selectivity of a bimetallic catalyst for ethanol reforming.

3.5. Intelligent Optimization Algorithms

Intelligent optimization algorithms have undergone significant development over the past 40 years.[191] They also represent an important domain of artificial intelligence research. Detailed algorithms, (**Figure 7**) including the genetic algorithm, particle swarm optimization, and simulated annealing algorithm, will be discussed in this section.

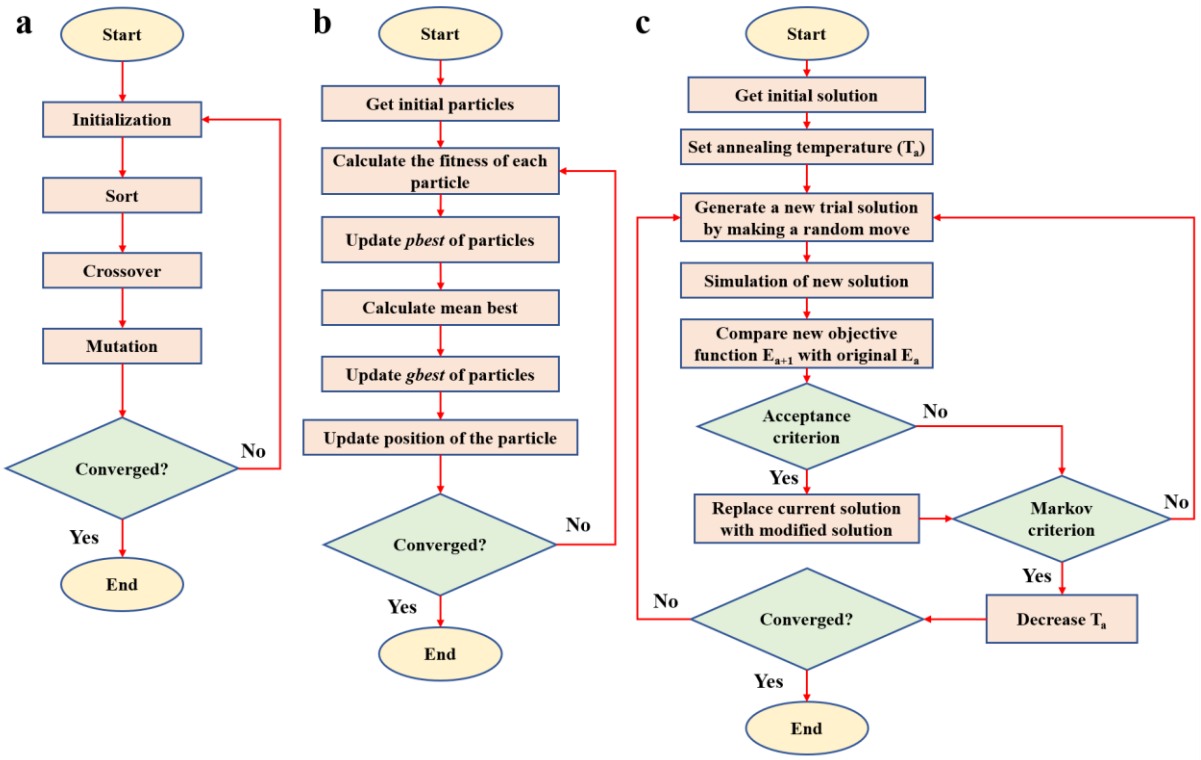


Figure 7. Schematic illustration of flowcharts of a) GA, b) PSO (*p*best refers to best fitness value in history, and *g*best refers to global best value), and c) SAA.

3.5.1. Genetic Algorithm (GA)

The genetic algorithm (GA, Figure 7a) is a widely used optimization method inspired by the process of natural selection.[192] The main function of GA is to identify the globally optimal solution by simulating a process analogous to natural evolution. During the GA process, an initialized population is first randomly generalized first; as the iterations progress, individuals which are more "fit" will be selected to represent their generation; individual genomes are then recombined or mutated to produce the next generation. GA has been utilized to great effect in materials science. Morgan et al. implemented GA to optimize the defect structures in bulk crystalline materials, with the objective of predicting the stable cluster structures in an automated fashion.[193] Fernandez et al. used GA as one of the algorithms to investigate the electronic properties of graphene based on either atomic radial distribution function scores[194] or topological information.[91] Cherukara et al. employed GA to enhance the simulation efficiency for predicting the thermal conductivity of stanene.[195]

3.5.2. Particle Swarm Optimization (PSO)

Particle swarm optimization (PSO, Figure 7b) was developed for the optimization of non-linear functions.[196] PSO is designed to solve problems by utilizing a population of candidate solutions (also referred to as particles) and iteratively moving these particles within the solution space until a globally optimal solution is identified or the iteration limit is reached. PSO has been applied to study various problems in materials science. For instance, Ma et al. developed a PSO based method, coined

Crystal Structure Analysis by Particle Swarm Optimization (CALYPSO), to efficiently study the multidimensional potential energy surfaces of materials.[197] This method was used to predict the structure of single layered, multi-layered, and quasi-2D materials.[198] Lin et al. also developed a PSO-based strategy to control the microstructure formation in Ni-based superalloys during hot forging.[199]

3.5.3. Simulated Annealing Algorithm (SAA)

To solve local optimization problems, the simulated annealing algorithm (SAA, Figure 7c), which includes the Metropolis algorithm and annealing process, was developed by Kirkpatrick et al. in 1983.[200] During the process of searching for an optimum, a worse solution can be accepted based on the probabilistic equation. Therefore, in contrast with the traditional gradient descent, the random process in SAA offers an opportunity to jump out of the local optimum to reach the global optimum. SAA has been employed in various contributions to materials science, for example, Erchiqui combined SAA and GA to optimize the shaping of thermoplastics during the thermoforming process.[201] AlRashidi et al. used SAA to extract and identify the photovoltaic parameters of different types of solar cells.[202] Recently, Major et al. integrated the Monte Carlo method and SAA to predict the cation ordering in different mixed transition metal oxides materials.[203]

3.6. Data-Processing and Data-Mining Methods

Data is fundamental to the data-driven study of materials. Data processing and data mining methods can directly influence the results of the materials study. We detail several data processing and data mining methods, including transfer learning, Bayesian global optimization, and adaptive ML in this section.

3.6.1. Transfer Learning

Transfer learning (TL) is a branch of ML that focuses on the use of pre-existing knowledge/models to solve a new but relevant problem.[204] Due to this nature of TL, it has been used to study materials based on small dataset, which is critical since large dataset for such materials may not always be available. Agrawal et al. leveraged TL in conjunction with large DFT computational datasets, other small DFT datasets, and experimental data to build a reliable predictive model for material formation energies, ultimately achieving a low mean absolute error of 0.07 eV/atom (**Figure 8**a).[205] Yoshida et al. developed a library containing more than 140,000 pre-trained models for various properties based on large datasets; they subsequently used TL in conjunction with this library to predict various material properties.[206] In another case, Reed et al. used TL based on small datasets to screen billions of compositions for potential application as lithium-ion conductors.[207]

3.6.2. Bayesian Optimization

Bayesian optimization (BO) is a sequential decision-making approach to gradient-free global optimization.[208] It is conventionally implemented for computationally expensive functions. BO has been used in materials studies for the determination of physical parameters, experimental design, material discovery, and optimization of atomic structures.[209] For example, Osada et al. employed the BO method to investigate optimal conditions for the growth of Si thin films based on several parameters and their interactions (Figure 8b).[210] After optimization, the growth rate of Si films was twice as high as that prior to optimization.

3.6.3. Adaptive ML

Adaptive ML (AML) refers to ML algorithms in which model parameters can be automatically optimized during the execution of the algorithm. AML covers a wide range of algorithms, including the aforementioned AdaBoost and adaptive SAA. For most practical applications, AIMD offers accurate simulation results; however, its computational expense prohibits extensive application. Ramprasad and Botu implemented AML to identify fingerprints mapping atomic configurations to material properties , thereby accelerating the AIMD simulation.[211] In another study, Xin et al. developed an AML strategy to identify ABO_3-type cubic perovskite-based catalysts (Figure 8c) for highly efficient electrocatalytic oxygen evolution reaction (OER).[212]

3.7. Reinforcement Learning

Reinforcement learning (RL), along with supervised and unsupervised learning, represent the three basic ML categories.[61] The common RL model is based on the Markov decision process, which pursues the best long-term reward (Figure 8d). RL can be used for the optimization of organic synthesis routes[162] and the design of drug molecules.[213] In the realm of materials science, Rho et al. utilized RL as a model to search for the most suitable optical nanomaterials.[214] Whitelam and Tamblyn have also demonstrated the favorable performance of RL in controlling self-assembly, from small molecules to large porous 2D materials.[215]

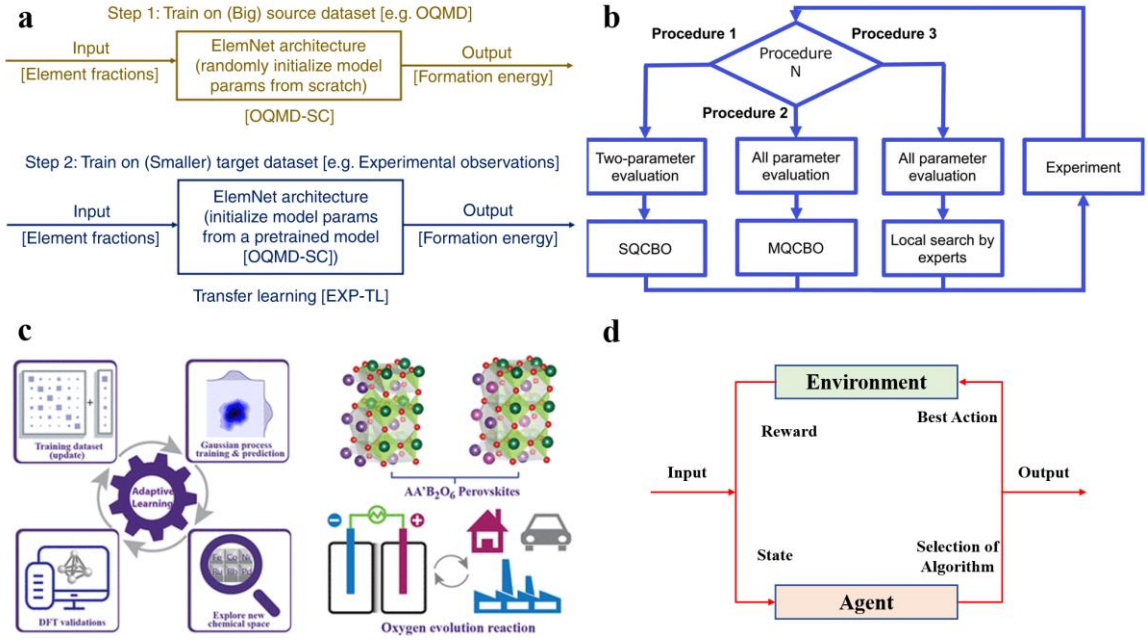


Figure 8. a) TL approach for the study of materials property prediction. OQMD: big DFT-computed source dataset Reproduced with permission.[205] Copyright 2019, Springer Nature Publications. b) Flowchart of BO process in the optimization for epitaxial growth of Si thin films. SQCBO: single quality constraint Bayesian optimization; MQCBO: multiple quality constraint Bayesian optimization. Reproduced with permission.[210] Copyright 2020, Elsevier Publications. c) Schematic illustration of AML in discovery of perovskite electrocatalysts. Reproduced with permission.[212] Copyright 2020, ACS Publications. d) Schematic illustration of RL process.

4. Available Chemical Databases for Innovative Material Discovery

Recent developments in data-centric approaches are expected to dramatically accelerate the progress in materials science because experimental and computational methods generate massive amounts of data, causing increasing complexity.[216] databases pertaining to both computational and experimental materials have been established to serve various specialized activities, rather than for dissemination or to enable contributions from the broader community.[217] The primary challenge in choosing and comparing databases is identifying the specific function that the database uniquely support, while also being able to compare various databases on the same structural basis.[218] **Table 1** lists the properties of dominant databases and their various attributes including data types, materials of focus, number of entries, data source, license, and a simple database descriptor.

Relatively simple analytical tasks pose challenges unique to the data-driven era because we are unable to capture, curate, store, search, share, analyze, and visualize the data in the absence of proper tools.[219] Thus, the identification of large numbers of correlations and patterns complex datasets has necessarily been carried out by high-throughput implementations of ML algorithms for decades to generate predictive and classification models for targeted physical properties. We have summarized representative high throughput tools (pymatgen,[220] qmpy,[221] ASE,[222] and atomate[223]) and workflow management tools (FireWorks,[224] AFLOWπ,[225] matminer,[226] and AiiDA[227, 228]). This class of high-throughput and workflow management tools is generally available in an open-source, Python infrastructure, with data connectivity implemented in RESTful API. These components aid in automating, managing, persisting, sharing, and reproducing the complex workflows associated with modern computational science and all associated data, reducing the cost and enhancing the efficiency of data summarization approaches with respect to the popular "*five V's*": volume, velocity, variety, veracity, and value.[229] Representative databases and the high-throughput management toolkits have been summarized in **Figure 9**. We also introduce the powerful QSTEM[230] tool for quantitative image simulation in electron microscopy.

More specifically, individual databases each solve one specific problem by relaying the specific descriptors which have been extracted from other existing databases. For instance, database formulation may be motivated by the need to synthesize specific materials for a specific application, such as the accelerated discovery of stable lead-free hybrid organic-inorganic perovskites (HOIP)[36], accurate prediction of battery life[231], and various catalysis applications[232]. The potential of data-driven strategies to uncover complex phenomena and design novel, high-performance materials is dependent on the quality and accessibility of databases and high-throughput tools, and which would otherwise not be possible with conventional trial-and-error approaches.

4.1. Databases

The continued advancement of science depends on shared and reproducible data. In the context of both computational and experimental materials science and rational materials design, this entails constructing large (open) databases of materials properties.[216] Several representative databases are presented as follows.

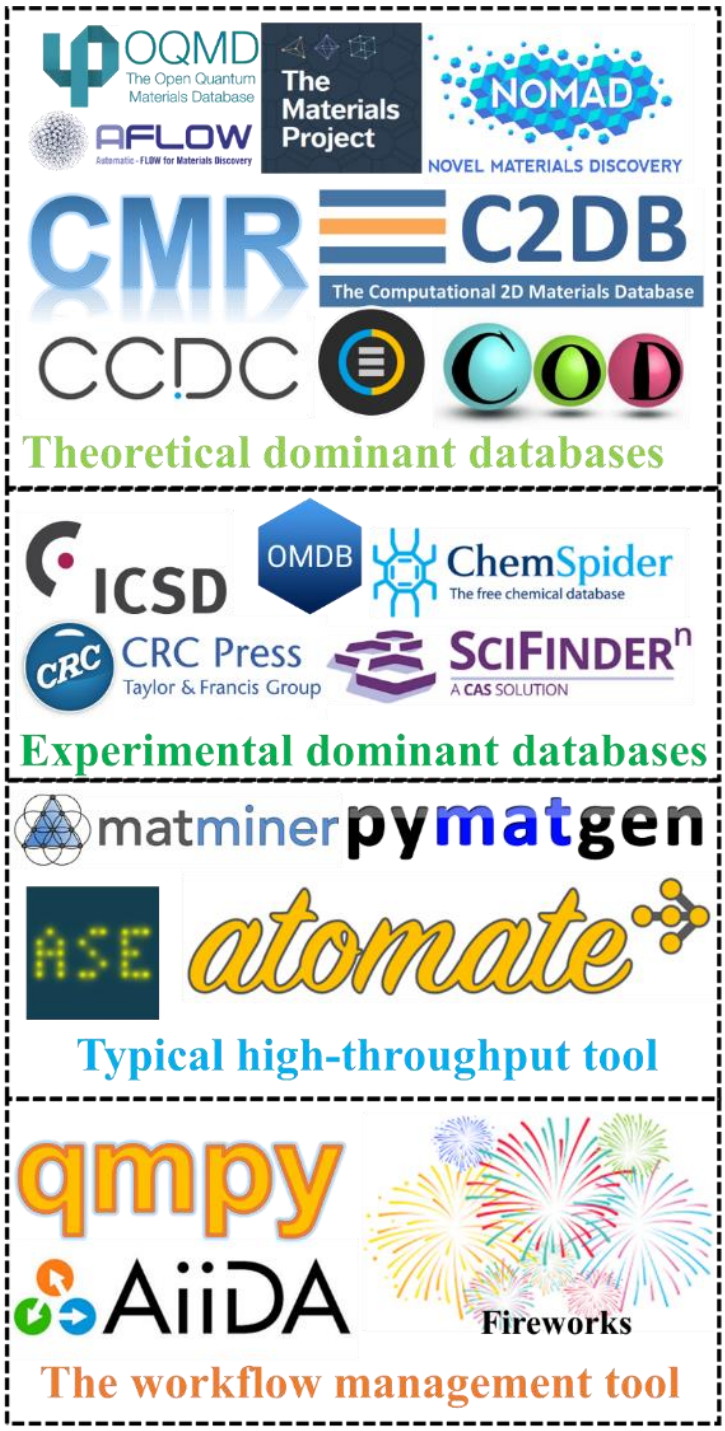


Figure 9. The representative theoretical (experimental) databases and high-throughput packages with management framework.

4.1.1. Computational Databases

Open Quantum Materials Database (OQMD)

The OQMD[233, 234] is a DFT database containing calculated thermodynamic and structural properties of 815,654 materials, developed by Chris Wolverton's group at Northwestern University.

The OQMD contains approximately 300,000 calculated structures, mainly from two sources: ~10% from the Inorganic Crystal Structure Database (ICSD)[235] and ~90% from the iteration of many chemistries for some of simple prototypes. For the crystal structures in the ICSD, ~44,000 structures are calculable, of which the OQMD contains DFT calculations of 32,559 ICSD structures. The remaining calculable ICSD structures are continually being calculated and added to the OQMD. Additionally, 259,511 hypothetical compounds have been generated based on 16 elemental prototypes, 12 binary prototypes with their compositions, and three ternary prototypes with their compositions.[236] [237] [234] Moreover, OQMD provides a qhull algorithm for establishing DFT ground-state phase diagrams at ambient (high) pressure and Grand Canonical Linear Programming (GCLP) to analyze the complex ground state thermodynamics of metal hydrides[238] [239] [240]. The OQMD provides the entirety of the underlying database to be freely downloaded at oqmd.org/download/, in addition to a Representational State Transfer (REST) Application Programming Interface (RESTful API) for programmatic access, which allows scientists and engineers to use simple Hyper Text Transfer Protocol (HTTP) requests to access all living data [218].

For instance, Tiantian Hu et al. used the Wasserstein GAN model in conjunction with the OQMD database to generate novel hypothetical materials (**Figure 10**a).[241] Victor Fung et al. predicted adsorption energies using the density of state data from the OQMD and Materials Project (MP) database combined with CNNs, targeting the accelerated discovery of catalytic materials (Figure 10b).[242] The MP database is introduced in the subsequent section.

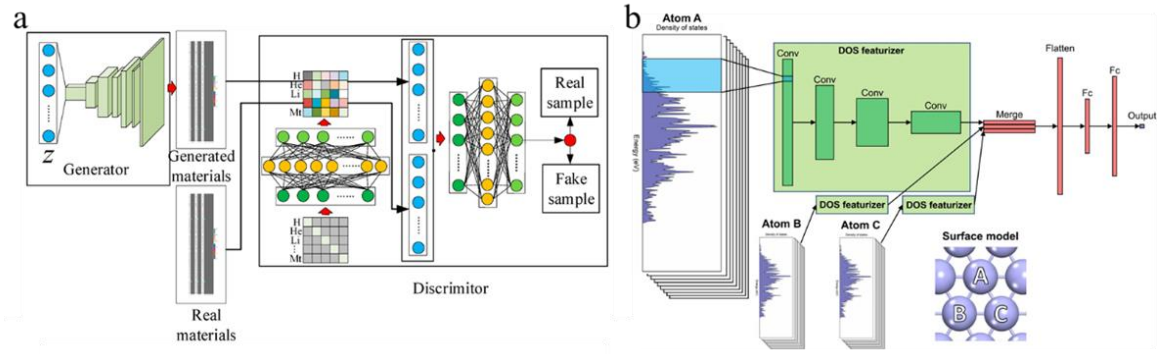


Figure 10. a) The Wasserstein Generative Adversarial Network (WGAN) model using the OQMD database to generate novel hypothetical materials. Reproduced with permission.[241] Copyright 2020, MDPI Publications. b) Using the density of state data from the OQMD and MP database by convolutional neural networks (CNNs) for the accelerated discovery of catalytic materials. Reproduced with permission.[242] Copyright 2021, Springer Nature Publications.

Materials Project (MP)

The Materials Project (MP) provides open web-based access to computed information on known and predicted materials to inspire and design novel materials.[24] Most of the MP data pertain to chemical compounds in the ICSD.[235, 243] A significant challenge is the generation of novel compositions and compounds to perform calculations[24] even though there already exist multiple algorithmic, e.g., Optimization-based, [243-246] and data-driven approaches[247-249] to tackle this problem. For materials included in the MP database, selected properties such as total energies[250], electronic structure[250], thermodynamic equations of state parameters[251], phonons[252],

piezoelectricity[253], elasticity[254], dielectricity[255], and thermoelectricity[256] have been calculated and included. In addition, MP includes apps to visualize phase diagrams[257, 258] and Pourbaix diagrams[259]. Several other convenient applications such as Materials Explorer[253, 254], Battery Explorer[260], Reaction Explorer[257], Structure Predictor[261], Crystal Toolkit[220], Nanoporous Materials Explorer[220], Molecules Explorer[262, 263], Redox Flow Battery Dashboard[264], X-Ray Absorption Spectra (XAS)[265], Interface Reactions[266], and Synthesis Description Explorer[267] have also been included in MP. Both Python Materials Genomics (pymatgen)[220] and FireWorks[224] open-source libraries are available for materials analysis and high-throughput application. Note that all the underlying data for the calculations of ~530,000 nanoporous materials and 130,000 inorganic compounds are accessible via the Materials API[268] based on REST principles.

Although the MP database was originally developed to predict the adsorption energy of the catalytic materials,[242] it has supported many other applications such as the accelerated discovery of stable spinel material[269] and carbon dioxide electrocatalysis[18]. Additionally, the MP and OQMD databases' magnetization properties are nearly comparable.[218] However, the Automatic-FLOWLIB (AFLOW) skews to larger magnetizations compared with MP and OQMD.[218]

Automatic-FLOWLIB (AFLOW)

AFLOW provides a globally available database of 3,312,125 material compounds with over 566,373,375 calculated properties and growing[270]; it is a powerful tool for materials discovery and property predictions using ML, the prototype encyclopedia, and the generation of convex hulls. As a multi-purpose repository, AFLOW comprises of 323,516 electronic structures, 125,496 Bader charges, 6,049 elastic and 6,038 thermal properties. This continuously updated compilation currently contains over 1,724 binary systems with more than 356,343 binary entries, 30,071 ternary systems with more than 2,400,160 ternary entries, and 150,621 quaternary systems with more than 450,567 quaternary entries. For convenience, several apps and documents have been customized for specific applications. For instance, AFLOW-ML contains three functional modules only requiring structural information: the Property Labeled Material Fragments (PLMF[271]) provides the bandgap, energy, modulus, heat capacity etc.; the Molar Fragment Descriptor (MFD[272]) predicts vibrational free energies and entropies; AFLOW Superconductor (ASC)[273] can classify material as superconductors while also estimating the critical temperature. AFLOW-CHULL, powered by the AFLUX Search-API, is a cloud-oriented platform for autonomous phase stability analysis, a valuable tool for guiding synthesis based on high-throughput and even autonomous approaches[274]. AFLOW-AAPL (Automatic Anharmonic Phonon Library) is an efficient and accurate framework for calculating lattice thermal conductivity of solids, which was developed to compute the third-order interatomic force constants and solve the Boltzmann transport equation within the high throughput AFLOW framework.[275]

We introduce the high-throughput first-principle-calculation framework of PAOFLOW and AFLOWπ. The key components of PAOFLOW involve managing sets of calculations to determine band structures, the density of states, complex dielectric constants, diffusive and anomalous spin and charge transport coefficients, etc. using a methodology that generates finite basis Hamiltonians from the projection of first principles plane-wave pseudopotential wavefunctions on atomic orbitals. The critical components of AFLOWπ involve robust data generation, real-time feedback and error control, curation and archival of data, and post-processing tools for analysis and visualization. AFLOWLIB API[276] following REST principles is introduced for the AFLOWLIB.org materials data repositories consortium and provides a powerful tool for accessing a large set of simulated material

properties data. For instance, Valentin Sranev et al. apply the random forest ML strategy as a classification and regression model in conjunction with the AFLOW database and ISCD, providing 35 compounds with critical temperatures above 20 K as experimental candidates.[277]

Novel Material Discovery (NOMAD)

The concept of the NOMAD was developed in 2014, independently and in parallel to the "FAIR Guiding Principles."[278] The Novel Materials Discovery (NOMAD) Laboratory is a user-driven platform for sharing and exploiting computational materials science data.[279] With the NOMAD repository and its code-independent canonicalized NOMAD archive, NOMAD consists of the world's most extensive data collection in this area. Based on a searchable, accessible, interoperable, and reusable data infrastructure, it offers a variety of services, including advanced visualization, NOMAD encyclopedias, and artificial intelligence. Further, the NOMAD CoE established an innovative tool for mining this data to locate structure, correlations, and novel information that would otherwise be difficult to identify through the study of a small database.

Note that usable and clearly defined metadata is a prerequisite for this normalization step to a code-independent format, rendering even the development of the NOMAD Meta Info[280] a significant challenge. In addition, the Open Databases Integration for Materials Design (OPTIMADE) consortium aims to promote materials databases interoperation by developing a standard REST API. Recently, Acosta et al. established the materials map of two-dimensional (2D) honeycomb structures for analyzing and identifying 2D topological insulators based on the NOMAD concept.[280, 281] Unlike the OQMD, MP, AFLOW, and NOMAD databases, the Computational Materials Repository (CMR) has many independent projects that consist of the Atomic Simulation Environment (ASE)[222] dataset, such as the computational 2D materials database (C2DB) [282] (Figure 9), and the detailed information is as follows.

Computational Materials Repository (CMR)

CMR[283] has resulted from a collaboration under the Quantum Materials Informatics Project (www.qmip.org) to establish core technologies for integrated computational materials design[283]. CMR addresses data challenges to enhance the possibility of designing new materials based on quantum physics calculations. CMR provides software infrastructure (such as the Computational 2D Materials Database[282], Bondmin optimization algorithm,[284] and CatApp database[285]) that support the collection, storage, retrieval, analysis, and sharing of data generated by numerous electronic structure simulators. Furthermore, CMR provides some basic functionality for processing large amounts of data, though more software development in this area is being implemented to facilitate large-scale collaboration in the future. We present representative computational and free-of-charge databases in Figure 9. However, the subsequent section introduces a number of reputable databases with historical significance (Figure 9) focusing primarily on the collection of experimental data.

4.1.2. Experimental Databases

ICSD

The ICSD[235] is the world's largest database of fully evaluated and published data containing inorganic crystal structures primarily derived from experimental results. Currently, the ICSD[286] has more than 232,012 entries, including ~2,902 elemental crystal, ~38,506 binary compounds, ~73,048

ternary compounds, and ~73,688 quarternary and quintenary compounds. The database is updated twice a year based on over 80 leading scientific journals and more than 1,400 other scientific journals; data sources have been expanded to include experimental inorganic structures, experimental metal-organic structures, and theoretical inorganic structures.

To be included in the database, the structure must be fully characterized. For instance, atomic coordinates can be determined or derived from known structure types, and the composition must be fully specified. Typical entries include chemical names, formulas, unit cells, space groups, complete atomic parameters (including atomic displacement parameters if available), site occupancy, titles, authors, and literature citations. For published data, many items (such as Wykov sequences, molecular formulas, weights, ANX formulas, and mineral groups) are introduced through expert evaluation or generated by computer programs.

The keyword-based search in the ICSD can be specified in terms of physical properties, analytical methods used, and technical application. Note that the ISCD data has been used to indicate promising novel applications of new ionic conductors, solar cell adsorbers, advanced ceramic materials, nature's missing compounds, and structural relations between the crystalline compounds. In addition, ICSD data have been included in almost all other computational databases, such as OQMD, MP, and AFLOW. Organic and inorganic compounds are two of the main categories of chemical materials. Thus, we introduce the Cambridge Structural Database (CSD) for organic materials.

CSD

The CSD[287] is the world's largest and most comprehensive collection for small-molecule organic and organometallic crystal structures, containing over one million structures from X-ray and neutron diffraction analyses. For comprehensive coverage of single-crystal data, cell parameters and all available data are included even if no coordinates are available. Similarly, powder structures are available from the International Centre for Diffraction Data (ICDD)[288] even though the coordination information is missing. Note that there is a slight overlap between the CSD and the ICSD in the area of molecular inorganics, but that purely inorganic structure is not contained in the CSD.

The CSD database has provides data in two distinct ways. The first is pertains only to structural aggregation and standardization, making it easier to access individual entries. The second is based on further study of data collection and the discovery of new knowledge transcending the results from individual experiments. Python-based API[289] has also been introduced to enable end-users to query CSD using customized script. Accessing data via scripts in conjunction with other packages such as RDKit[290] is very useful for more advanced structural data analysis. For instance, users will be able to use ML more conveniently in conjunction with APIs for solvate prediction, implementing fragment pocket analysis using structural information, and supporting crystal (co-crystal) structure prediction.[291] More detailed insights could be developed as the scale of data increased, having a profound impact across the scientific community with specific consequences for drug discovery and development.[289] However, the ICSD and CSD have paid licenses (as shown in Table 1), affecting a number of institutions or members who cannot access the data. We subsequently introduce the open-access Crystallography Open Database (COD)[42] database, including both organic and inorganic materials.

COD

The COD[42] is the most extensive open-access collection of minerals, metal organics, organometallics, and small organic crystal structures, excluding biomolecules which are otherwise stored in Protein Data Bank. The COD currently contains over 385,000 records and is constantly growing in size and quality. The COD has introduced a new data deposition website that allows the manual and automatic uploading of data and structures to the COD. This automation is greatly facilitated by the introduction of the Crystallographic Interchange Framework (CIF). In addition to web access, the COD provides a RESTful interface which allows the querying of information about COD entries based on specific criteria or the crystal structure file itself. Additionally, SQL (Structure Query Language) is the most powerful mechanism to query these relational databases, providing more functionality than COD web pages and COD RESTful interfaces.

A widely accepted application of the COD is for material identification with the help of the powder diffraction method and search-match procedure. The largest diffractometer vendors (including Bruker, PANalytical, and Rigaku) ship COD collection software that are compatible with their equipment and provide regular updates on the COD website or on their own pages. In bioinformatics and drug design, the COD is used as a source of open data for restraint libraries[292]. Finally, the COD is also used in basic research to support investigations into hydrogen storage, characterization of 2D materials, etc.[293]

4.1.3. Data Infrastructure

Citrination Platform

The Citrination Platform[294] takes an intermediate view on the challenge of materials data infrastructure, driven by the goal to make vast quantities of cross-disciplinary materials data both human-searchable and machine-readable for data mining. In the design of material data infrastructure, the Citrination Platform offers convenient technology for data import, storage, and access. It can be used in various fields such as extracting knowledge through catalysis informatics[295], screening of inorganic materials synthesis parameters[296], and finding novel thermal materials[297].

Materials Data Facility (MDF)

Materials Data Facility (MDF) [298, 299] services are uniquely differentiated to support the publishing, discovery, and access to materials datasets using distributed data publication and discovery models, which are built on and leverage production services provided by Globus, a nonprofit software-as-a-service (SaaS).[298] MDF supports this vision by providing interconnection points that allow producers of material data to dispatch a wide range of results which is discovered and aggregated by data consumers from each independent source. Currently, MDF stores 30 TB of data from simulation and experiment, and also indexes hundreds of datasets contained in external repositories, with millions of individual MDF metadata records created from these datasets to aid fine-grained discovery."

4.2. High-Throughput (HT) Programming Packages and Workflow Management Frameworks

The constant availability of computing power and the sustainable development of advanced computing methods have contributed significantly to recent scientific advances. The data-driven era

for materials science has dramatically impacted novel materials discovery, physical properties prediction, and the underlying patterns of numerous materials. Consequently, mass computational and experimental material databases have been established to serve various specialized purposes rather than sharing and enabling contributions to the materials science community based on "FAIR Guiding Principles."[278] These developments present new challenges posed by the vast amount of computations and data to manage.[219] Selecting the most appropriate database to address specific scientific problems remains the primary challenge. It is critical to identify the key differences between various databases while being cognizant of the ways in which they overlap.[219]

Next-generation exascale supercomputers will exacerbate these challenges, implying that automated and scalable solutions will be essential. For instance, Cronin et al. reported the convergence of multiple synthetic paradigms for a universally programmable chemical synthesis machine[300] and summarized the current process for universal chemical synthesis and discovery using ML[301]. Cooper et al. used a mobile robot to search the most efficient photocatalysts for hydrogen production from water.[302] Thus, it is vital to summarize the high throughput tools and workflow management frameworks that can conveniently handle data obtained from various databases. Figure 9 presents a summary of the representative high throughput tools (pymatgen,[220] qmpy, [221] ASE,[222] and atomate[223]) and workflow management frameworks (FireWorks, [224] AFLOWπ, [225] matminer, [11] and AiiDA[227, 228]).

4.2.1. Programming Packages

Python Materials Genomics (pymatgen)

Pymetgen[220] is a robust, open-source Python library for materials analysis. A major enabler in high-throughput computational materials science efforts is a robust set of software for performing computational initialization (structure generation, required input files, etc.) and post-computational analysis to derive useful material properties from raw computational data. As mentioned in f MP section, the pymetgen library provides a convenient tool for obtaining useful materials data via MP's REST API for structure generation, manipulation, and thermodynamic analysis.

The pymatgen library provides (1) a core Python object for material data representation, (2) a well-tested set of structures and thermodynamic analyses relevant to a number of applications, and (3) targeting researcher needs by establishing an open platform for collaboration and developing a sophisticated analysis of material data obtained from both first-principles calculations and experiments. The overview of a typical workflow for pymatgen is presented in **Figure 11**a. For example, Ceder et al.[303] utilized pymatgen to map the body-centered cubic-like anion framework to solid-state lithium superionic conductors (Figure 11b). Additionally, The grand potential phase diagram was used to identify the domain phase in the $Li_7La_{2.75}Ca_{0.25}Zr_{1.75}Nb_{0.25}O_{12}$ system (Figure 11c).[304] Analogous to the use of pymatgen in conjunction with MP, qmpy has also been developed to support workflows based on data from OQMD.

The OQMD running and maintenance toolkit (qmpy)

The qmpy[221] toolkit stores crystal structure data, automates DFT calculations, handles computational resources, and performs thermodynamic analysis. Moreover, qmpy is a package containing many computational materials science tools, bundled with two executable scripts: *qmpy* and *oqmd*.

qmpy is used to run and maintain the OQMD. The ultimate reference for the searching model is based on 'filter', 'exclude', and 'get' methods. There are many advanced functions such as advanced searching, using *qmpy* to manage a high-throughput calculations, and the ability to create customized Python scripts which takes advantage of qmpy features. For instance, Wolverton et al. use qmpy to develop accurate formation energy comparisons between the DFT and experimental data (Figure 11d).[23] Difference between OQMD and experimental for the fit-none, fit-partial, and fit-all chemical-potential sets have been presented. Specifically, in the fit-none case, the average difference is 0.105 eV/atom, with a MAE of 0.136 eV/atom; the average error is reduced to 0.020 eV/atom with a MAE of 0.096 eV/atom using chemical potentials from the fit-partial set; finally, the average error is 0.002 eV/atom with a MAE of 0.081 eV/atom using the chemical potentials of all elements (fit-all), which is the slightly better fitting than both fit-none and fit-partial chemical potential case.[23]

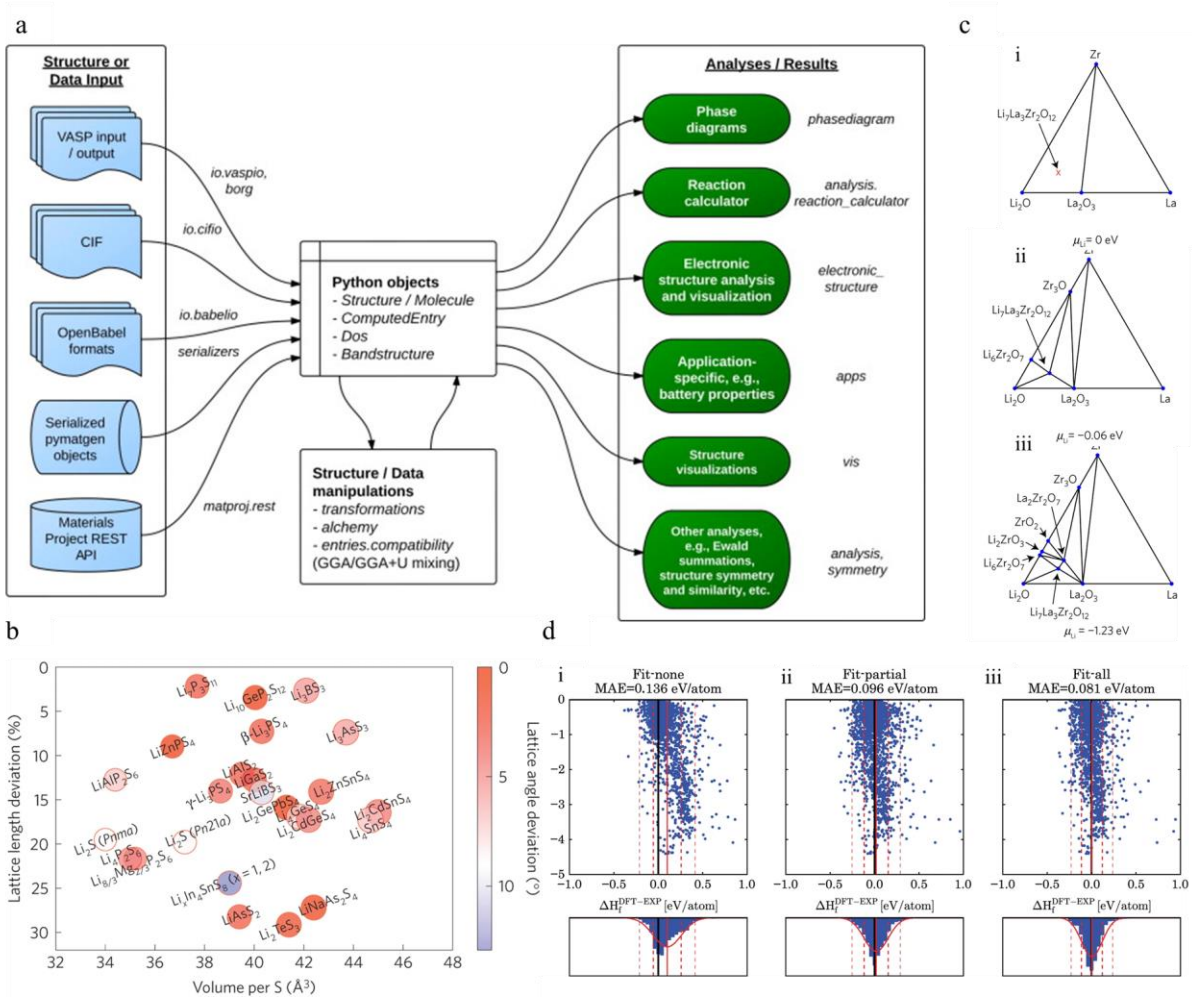


Figure 11. a) overview of a typical workflow for pymatgen. Reproduced with permission.[220] Copyright 2013, Elsevier Publications. b) Mapping the body-centered cubic-like anion framework to solid-state lithium superionic conductor. Reproduced with permission.[303] Copyright 2015, Springer Nature Publications. c) The grand potential phase diagram was used for identifying the domain phase in the $Li_7La_{2.75}Ca_{0.25}Zr_{1.75}Nb_{0.25}O_{12}$ system at potentials (μ_{Li}) of i. 0 eV, ii. -0.06 eV and iii. -1.23 eV, respectively. Reproduced with permission.[304] Copyright 2017, Springer Nature Publications. d) Schematic illustration of comparison between the OQMD and 1670 experimentally measured

formation energies for i. fit-none, ii. fit-partial and iii. fit-all sets of the elemental potential. Reproduced with permission.[23] Copyright 2015, Springer Nature Publications.

Atomic Simulation Environment (ASE)

ASE[222] is a collection of Python modules to set up, control, visualize, and analyze simulations on an atomic and electronic scale. ASE acts as a front-end for atomistic simulation, where both the atomic structures and parameters that control the simulation are easily defined. Simultaneously, the full functionality of the Python language is also available, giving users interactive and detailed control over several interrelated simulations.

To perform many complex atomic-scale simulations, ASE relies on three powerful libraries (Numpy, SciPy, and matplotlib). For the top of the atoms-calculator interface, ASE provides algorithms for various atomistic simulation tasks (30 different atomic-scale codes such as structure optimization, molecular dynamics, the nudged elastic band simulation (**Figure 12**a). For example, Tran and Ulissi identified suitable electrocatalysts for CO_2 reduction and H_2 evolution by using an active learning method through VASP implemented by ASE with FireWorks managed frameworks[224] (Figure 12b).[13] In addition to the pymatgen, qmpy, and ASE (Figure 9), the last representative HT package that we introduce is Atomate; the detailed framework to manage HT automation is as follows.

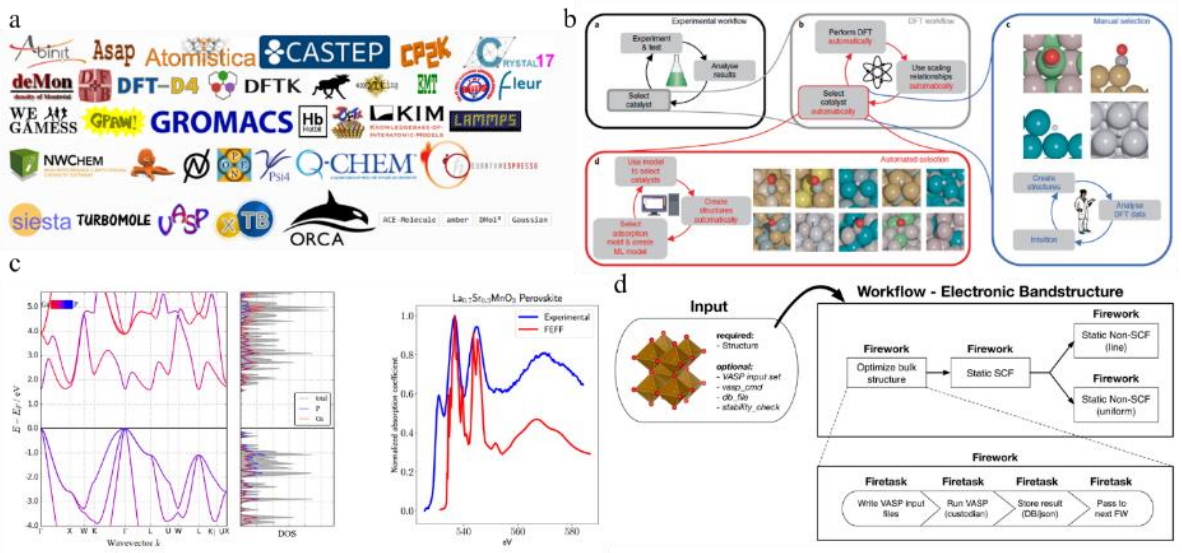


Figure 12. a) The overview of the platform of the support from ASE. b) discover the suitable electrocatalysts for CO_2 reduction and H_2 evolution by using an active learning method through VASP implemented by ASE with FireWorks managed frameworks. Reproduced with permission.[13] Copyright 2018, Springer Nature Publications. c) Example of GaP band structure (left) computed via atomate/VASP and $La_{0.7}Sr_{0.3}MnO_3$ ELNES spectra (right) computed via atomate/FEFF using Atomate. Reproduced with permission.[223] Copyright 2017, Elsevier Publications. d) The basics workflows for band structure using Atomate. Reproduced with permission.[223] Copyright 2017, Elsevier Publications.

Atomate

Atomate[223], an open-source Python framework for the simulation, analysis, and design of materials focuses on the automation and extensibility of conventional software applications which may other otherwise be difficult to use or implement at scale. Atomate offers powerful theory and calculation tools for the analysis and design of novel materials. Atomate makes it possible to perform complex materials science computations using very straightforward statements. The FEFF software integration was recently introduced (Figure 12c), and other computational packages are under development.

Atomate aims to gather knowledge about the computational procedures of different methods of material analysis into easy-to-use workflows and workflow components that can be modified and reconfigured as needed. Workflows currently available in atomate include band structure, bulk modulus, elastic tensors, Raman spectra, permittivity, , and various types of spectral calculations (XAS, EELS). Atomate is built on top of state-of-the-art open-source libraries such as pymatgen, custodian, and FireWorks. Building these libraries not only serves as a friendly and straightforward introduction to computational materials science but is also powerful enough for the most demanding theoretical users who require precise control and large-scale execution. Specification of the crystal structure is all that is required to allow atomate set up a complete workflow to provide properties of interest (Figure 12d); this can be accomplished for a single material, 100 materials, or 100,000 materials. For instance, Wu et al. calculated the band structure and elastic properties for polycrystalline $SnSe_2$ with various amounts of Br dopant using the Atomate package.[305] Zheng et al. automatically generated an ensemble-learned matching of XAS using the Atomate package in conjunction with the workflow management framework of FireWorks (**Figure 13**a).[306]

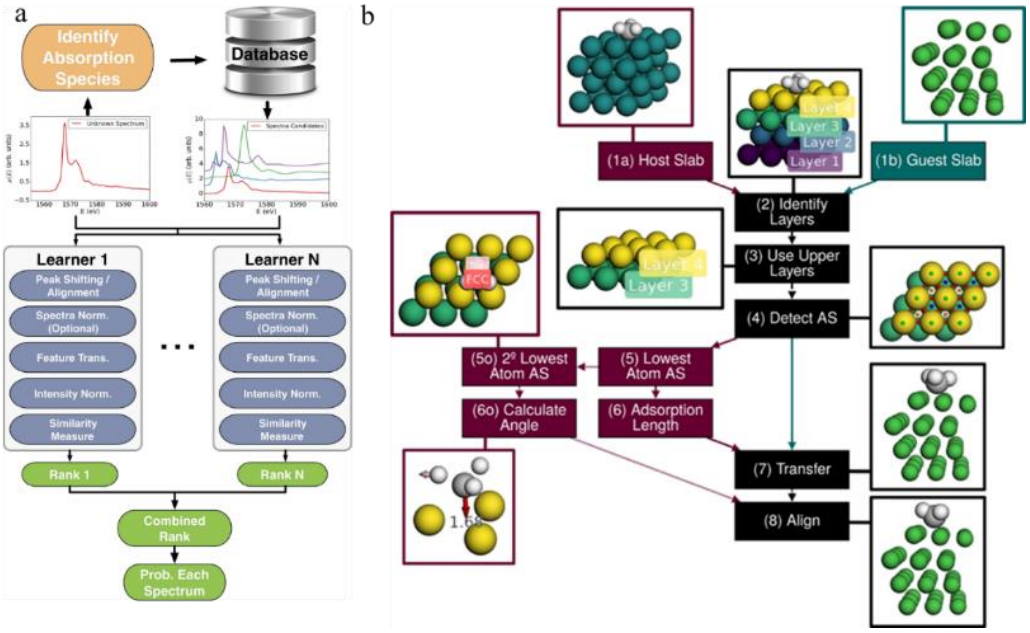


Figure 13. a) Overview of generation and ensemble-learned matching of X-ray absorption spectra using the Atomate package with the workflow management frameworks of FireWorks. Reproduced with permission.[306] Copyright 2018, Springer Nature Publications. b) To making the structure's relationship with heterogeneous catalysis using workflows of FireWorks. Reproduced with permission.[307] Copyright 2021, Wiley Publications.

4.2.2. Workflow Management Frameworks

FireWorks (FWS)

FWS is a ML library for Python that provides a DataFrame implementation compatible with PyTorch Tensors.[224] The user can build a model that references the input by column name consistent with tabular data formats applicable to all data analysis software and languages such as SQL, Stata, Excel, R DataFrames, and Python Pandas. The operation makes it easier to track variables when working with data in this format and integrating these models into existing Pandas-based data science workflows. Fireworks consists of a number of modules designed to work together to facilitate various aspects of deep learning and data processing. As illustrated in Figure 12b and Figure 13a, in addition to working with other codes such as ASE and Atomate, FWS can work in conjunction with VASP workflows to create, track, and stop the work or process. For instance, Sergio et al. utilized this strategy to develop the structure's relationship with investigations in heterogeneous catalysis (Figure 13b).[307] AFLOWπ integrates with the AFLOW, which is also a popular HT framework.

AFLOWπ

AFLOWπ,[225] a minimalist approach to high-throughput *ab initio* calculations, including the generation of tight-binding hamiltonians without any additional input, is easily portable, simple to use, and integrated with the AFLOW.org repositories. AFLOWπ was initially developed for verification and testing purposes but has evolved into a modular software infrastructure that provides an automation workflow for tight-binding Hamiltonian ab initio generation within a projected atomic orbital. The simulations for elastic constant, complex dielectric constant, diffusive transport coefficient, phonon dispersions with Hubbard *U* correction and optic spectra are included. For instance, Emmanuel et al. calculated the band structure (**Figure 14**a) and phono dispersion (Figure 14b) for thermoelectric bulk colusite using AFLOWπ, which is otherwise a challenging process to implement with traditional DFT simulation.[308]

Automated Interactive Infrastructure and Database (AiiDA)

AiiDA[227, 228] is an open-source python infrastructure platform to support and streamline the four core pillars of the ADES model: Automation, Data, Environment, and Sharing (**Figure 15**a). Leveraging the AiiDA Workflow Manager and its plugin ecosystem, developers can access simulation code that scales through the Python API combined with automatic simulation tracking for full reproducibility. As a core principle of AiiDA's design, its focus on data provenance represents a significant departure from the other management systems mentioned earlier.

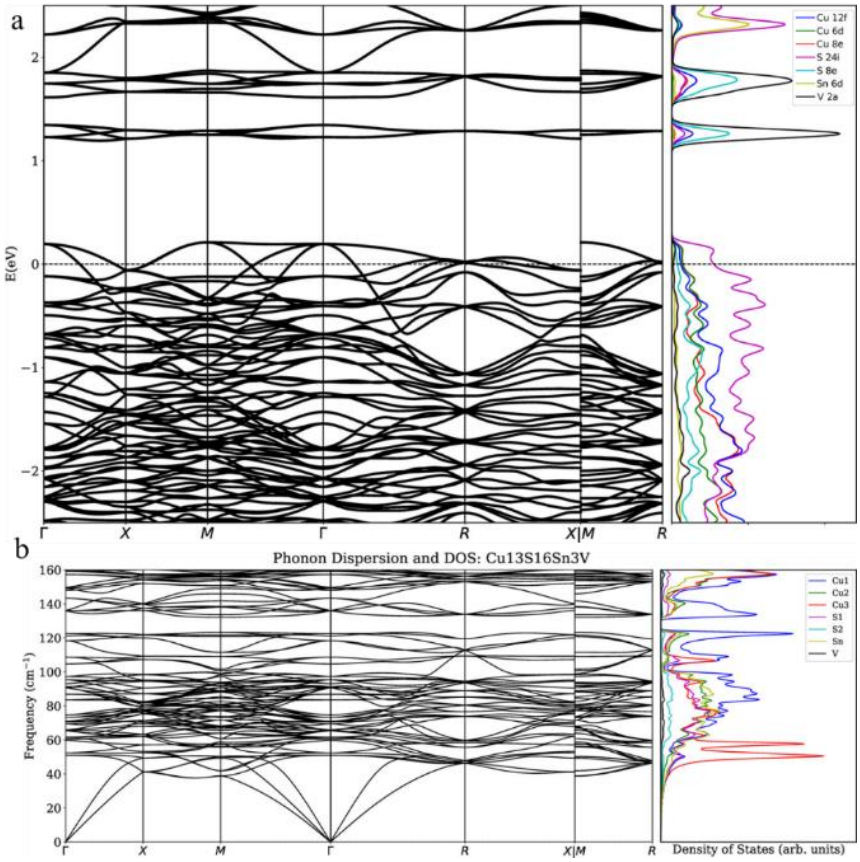


Figure 14. The a) band structure and b) phonon dispersion of thermoelectric bulk colusite using AFLOWπ. Reproduced with permission.[308] Copyright 2018, ACS Publications.

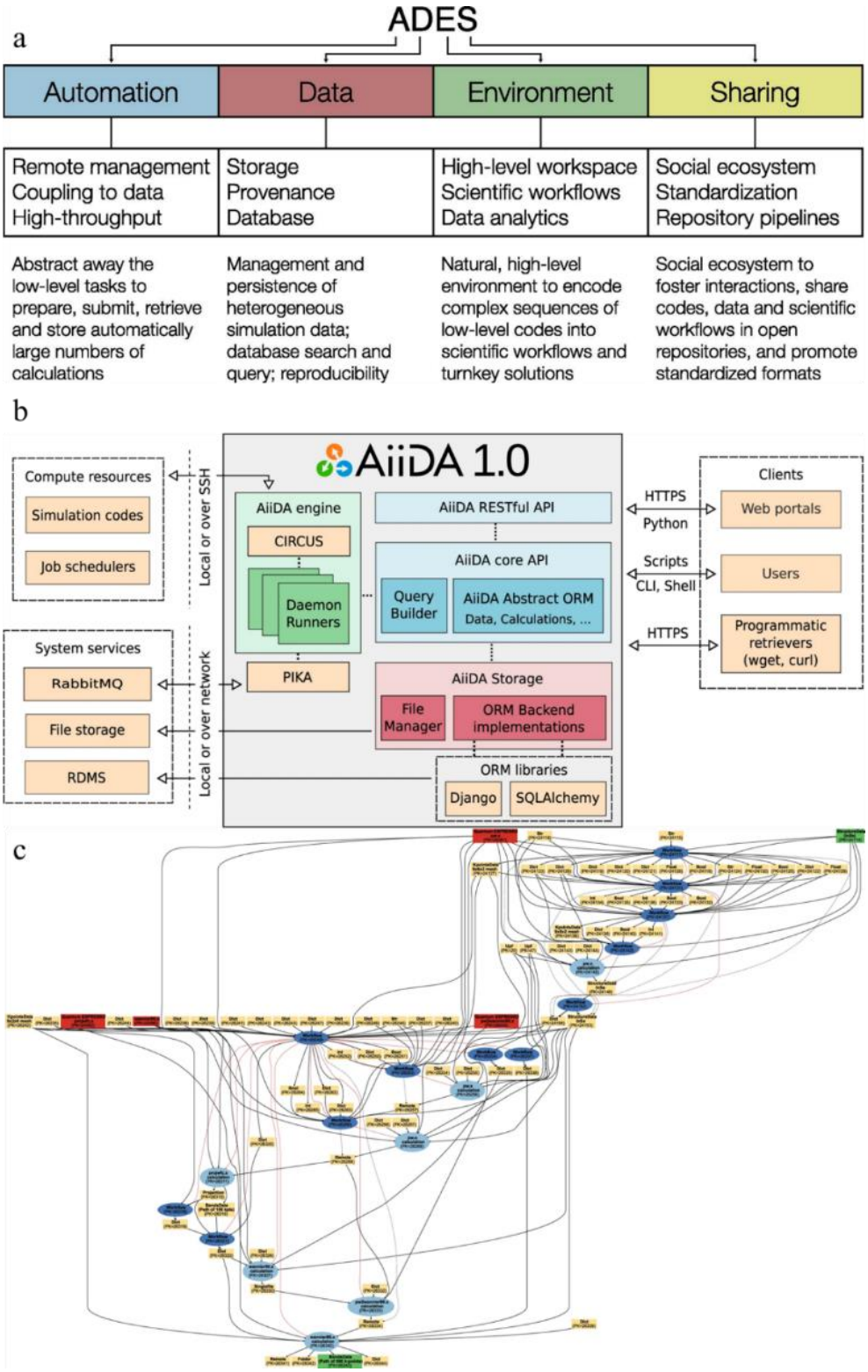


Figure 15. a) the ADES infrastructure in AiiDA. b) Schematic overview of the AiiDA.1.0 c) Provenance graph automatically generated by AiiDA. Reproduced with permission.[309] Copyright 2020, Springer Nature Publications.

AiiDA aims to provide a framework that enables the design and execution of complex high-throughput computational workflows with a fully automated history and built-in support for high-performance computing on remote supercomputers (Figure 15b).[228] Additionally, the main goal of

the AiiDAlab[228] platform is to provide an environment where users with varying expertise can access and perform computational workflows embedded into the AiiDAlab apps. Each user has a separate AiiDAlab account which grants them access to the AiiDAlab instance through a web browser. AiiDA is a flexible tool interoperable with any simulation software due to its plugin system, making computational science more transparent, user-friendly, and ultimately fully reproducible, in full compliance with FAIR principles. For instance, Valerio et al. have used AiiDA to automate the Maximally-localised Wannier functions and synthesize the corresponding Provenance graph (Figure 15c).

Matminer

Matminer[226] is an open-source toolkit for materials data mining based on the Python library, which provides a comprehensive library implementation of feature extraction routines developed by the materials community and features 47 feature classes that can generate thousands of individual descriptors and combine them into mathematical functions. The general workflow and overview of Matminer are shown in Figure 16a.

Matminer works with Panda's data formats to convert complex material attributes into numeric descriptors for data mining functionality (Figure 16b). It can then perform data mining on materials and make various downstream ML libraries and tools available for materials science applications. For example, low-modulus Ti-Nb-Zr alloys were discovered with the aid of the MP database in conjunction with the matminer library (Figure 16c).[310]

ChemML

ChemML[311, 312] is an open machine learning and informatics program suite for analyzing, mining, and modeling chemical and materials data. Specifically, ChemML is developed in the Python 3 programming language and uses a host of data analysis, ML libraries (accessible through the Anaconda distribution), and domain-specific libraries. ChemML allows its users to perform various data science tasks and execute machine learning workflows adapted specifically for the chemical and materials context. In addition, ChemML is designed to facilitate methodological innovation; it is one of the cornerstones of the software ecosystem for data-driven in silico research.[311]

MAterials Simulation Toolkit for Machine Learning (MAST-ML)

MAST-ML[313] is an open-source Python package designed to broaden and accelerate the use of machine learning in materials science research, particularly for non-experts without programming ability. It provides flexible access to the most important algorithms while codifying best-in-class machine learning model development and evaluation practices. MAST-ML provides predefined routines for many input setup, model fitting, and post-analysis tasks, as well as a simple structure for executing a multi-step machine learning model workflow, such as lattice for thermal conductivity[314] and magnesite flotation studies. [315]

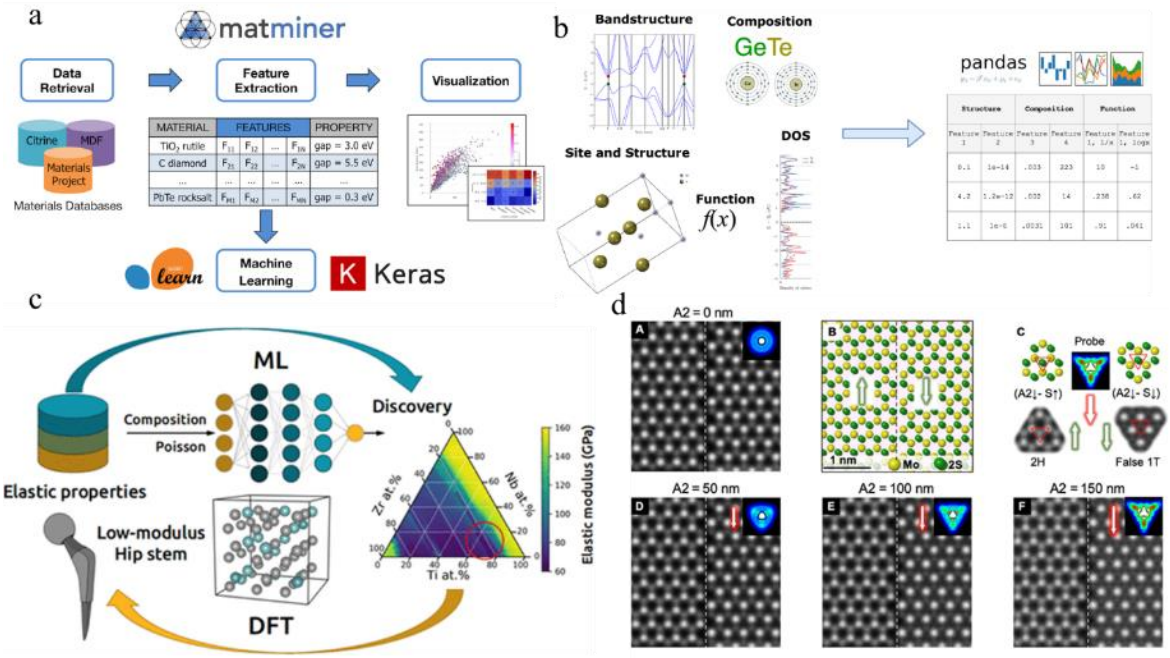


Figure 16. a) The general workflow and overview of matminer. b) Obtaining materials data from various sources into the panda's data format. c) Low-modulus Ti-Nb-Zr alloys were discovered with the aid of the MP database and the matminer library. Reproduced with permission.[310] Copyright 2020, ACS Publications. d) The QSTEM simulation of high-resolution-STEM ADF imaging. Reproduced with permission.[316] Copyright 2020, AAAS Publications.

4.2.3. Simulations

Quantitative TEM/STEM Simulations (QSTEM)

QSTEM[230] is a program for quantitative image simulation in electron microscopy, including TEM, STEM, and CBED image simulations based on the multislice algorithm. Several features of QSTEM are notable. First, QSTEM has the potential to work with arbitrary samples and orientations (such as interfaces, defects, and imperfect crystals and not only low-index zone axes of the single crystal) because of the principle of the multislice algorithm. Second, the atomic scatter coefficient must be accurate to the large angles required for STEM simulations. For instance, Lopatin et al. investigated the correlation of atomic simulation images using QSTEM with HR-STEM ADF images to reveal the false T phase of transition metal dichalcogenides (Figure 16d).[316] Finally, *pyqstem* has been created as an open-source python library based on QSTEM. The *pyqstem* project interfaces with QSTEM code through Python and ASE to provide a single environment for model building, image simulation, and analysis.

Table 1. The database including the name, data type, materials types , simple key description, number of the entries, data sources, and license.

Database	Types	Materials	Descriptor	No. Entries	Data Source	License	Ref
Open Quantum Materials Database (OQMD)	Computational	Inorganic Solids	Multi-purpose repository	~300,000	ICSD, Hypothesis	Free	[221, 233, 234]
Materials Project (MP)	Computational	Inorganic Solids; Nanoporous Materials	Multi-purpose repository	>130,000 ~530,000	ICSD	Free	[24]
Automatic-FLOW (AFLOW)	Computational	Inorganic Solids, Alloys	Multi-purpose repository	3,312,125	ICSD	Free	[317]
Novel Material Discovery (NOMAD)	Computational	Inorganic Solids	Multiple-source repository	--	Literarues	Free	[280]
The Computational Materials Repository (CMR)	Computational	Perovskites, 2D Materials	Multi-purpose (3D and 2D materials)repository	--	OQMD	Free	[283]
Inorganic Crystal Structure Database (ICSD)	Experimental	Inorganic Crystal Structures	Structural Properties	>232.012	Literarues	Non-Free	[235]
Cambridge Structural Database (CSD)	Experimental	Metal Organic Frameworks, Orgaincs Molecure	Organic and Inorganic experimental	>800.239	Literatures, ICDD	Non-Free	[318]
Crystallography Open Database (COD)	Experimental, Computational		Structural Properties	>385,000	Literatures,		[42]
The Computational 2D Materials Database (C2DB)	Computational	2D Materials	Structural, Thermodynamic, Elastic, Electronic, Magnetic, And Optical Properties	~4,000	MP, CMR	Free	[319]
Clean Energy Project (CEP)	Computational	Organic Photovoltaics	Multiple source repository for solar cells	>2,000,000	Literatures, Hypothesis	Free	[320]
Organic Materials Database (OMDB)	Computational	Organic Materials	Electronic Structrue, Density of States	~12,500	COD	Free	[321]
Joint Automated	Computational	2D/Solid Inorganics	Structural,	~40,000	MP, OQMD,	Free	[322]

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/adma.202104113.

Repository For Various Integrated Simulations (JARVIS)-DFT			Thermodynamic, Electronic, Elastic, Properties		AFLOW, Literatures.		
Citrination	Experimental, Computational	Inorganic Solids, Molecules	Multi-purpose repository	--	Literatures,	Free	[217]
Materials Cloud	Experimental, Computational	All Materials	Multiple-source repository	--	ICSD,COD, Literatures	Free	[323]
Alloy Database	Computational	Intermetallics	Structrue, Cohesive Energies	--	ISCD	Free	[324]
CatApp	Computational	Molecules on Surfaces	Reaction/activation Energies	--	--	Free	[285]
Computational Chemistry Comparison and Benchmark DataBase (CCCBDB)	Computational	Atoms, Moleculres	Thermochemical Properties	~2069	--	Free	[325]
Computational Electronic Structure Database (CompES-X)	Computational	Inorganic Solids	Electronic Structure	>100	--	Free	--
Crystalium	Computational	Elemental Solids	Surface, Grain Boundary Energetics	>145	Literatures	Free	[326]
Phonondb	Computational	Inorganic Solids	Phonons, Thermal Properties	--	MP	Free	
TE Design Lab	Computational	Semiconductors	Electronic, Thermoelectric Properties	~2701	Literatures	Free	[327]
AIST Research Information Databases	Experimental	General Materials Data	Substances, Chemical Accidents, Geological Information	--	Literatures	Free	[328]
American Mineralogist Crystal Structure Database	Experimental	Minerals	Structural Properties	2627	Literatures	Free	[329]
ASM Alloy Center Database	Experimental	Alloys	Composition, Structue, Physical Properties	--	Literatures	Non-Free	--
ASM Phase Diagrams	Experimental	Alloys	Thermodynamic	6200	Literatures	Non-	--

			Properties			Free	
CALPHAD databases	Experimental	Alloys	Thermodynamic, Kinetic, And Properties Databases	--	Literatures	Non-Free	--
ChemSpider	Experimental	Chemical Materials	Multiple-Source Repository	99,000,000	Literatures	Free	[330]
CINDAS High-Performance Alloys Database	Experimental	Alloys	Physical Properties	298	Literatures	Non-Free	--
CRC Handbook	Experimental	General Materials Data	Multi-purpose repository	--	--	Non-Free	--
CrystMet	Experimental	Metals	Chemical and physical information	70,000	Literatures	Non-Free	[331]
DOE Hydrogen Storage Materials Database	Experimental	General Materials Data	Hydrogen storage	--	Literatures	Free	--
Granta CES Selector	Experimental	Metals, Polymers, Composites, Medical Materials, Coatings, Aerospace Materials	Multi-purpose repository	>4000	Literatures	Non-Free	--
Handbook of Optical Constants of Solids, Palik	Experimental	General Materials Data	Hard-copy sources	--	Hard-copy sources	Non-Free	[332]
International Glass Database System (INTERGLAD)	Experimental	Glass	Structrues Properties	350,000	--	Non-Free	--
Knovel	Experimental	General Materials Data	Multi-purpose repository	--	Literatures	Non-Free	[333]
Matbase	Experimental	General Materials Data	Transcription Factors and The Corresponding Weight Matrices	--	Literatures	Free	--
MatDat	Experimental	General Materials Data	Physical Properties	>4000	Literatures	Non-Free	--
MatNavi (NIMS)	Experimental	Polymers, Inorganic and Metallic Materials	Multi-purpose repository	--	Literatures	Free	[334]
MatWeb	Experimental	Carbon, Ceramis,	Multi-purpose repository	140,000	Literatures	Free	[335]

		Fluid, Metal, Polymer, Wood and Natural Products					
Mindat	Experimental	Minerals, rocks, Meteorites	Multi-purpose repository	--	Literatures	Free	--
NanoHUB	Experimental	Nanomaterials	Multi-purpose repository	--	Literatures	Free	[336]
NIST Materials Data Repository (DSpace)	Experimental, Computational	General Materials Data	Multi-purpose repository	--	Literatures	Free	--
NIST Interatomic Potentials Repository	Computational	Meatals, Semiconductors, Oxides, and Carbon-containing systems	interatomic potentials	--	Literatures	Free	[337, 338]
NIST Standard Reference Database 3 (NIST SRD 3)	Experimental, Computational	Inorganic Solids	Multi-purpose repository	210,000	Literatures	Non-Free	--
Open Knowledge Database Of Interatomic Models (Open KIM)	Computational	Moleculars	interatomic potential repository	--	--	Free	[339]
Pauling File	Experimental, Computational	Inorganic Solids	Phase-Disgrams, Crystal Structures, Physical Properteis	357,612	Literatures	Non-Free	[340]
Pearson's Crystal Data (PCD)	Experimental	Inorganic Solids	Multi-purpose repository	350,000	Literatures	Non-Free	[341]
Pearson's Handbook: Crystallographic Data	Experimental	Intermetallic phases	Crystallographic Data	--	Hard-copy sources	Non-Free	--
Powder Diffraction File (PDF)	Experimental	Inorganic Solids	Crystallographic Data	--	Literatures	Non-Free	[342]
PubChem	Experimental	Molecures	Multiple-source repository	32,000	Literatures	Free	[343]
Reaxys	Experimental	Chemical data	Multi-purpose repository	>118,000	Literatures, Patents	Non-Free	[344]
SciFinder	Experimental	Chemical data	Multi-purpose repository	47,000,000	Literatures, Patents	Non-Free	[345]

SciGlass	Experimental	Glasses	Multi-purpose repository	360,293	Literatures, Patents	Non-Free	--
SpringerMaterials	Experimental	General Materials Data	Multi-purpose repository	--	Literatures, Patents	Non-Free	--
Total Materia	Experimental	Metallic Materials Data	Multi-purpose repository	350,000	Literatures, Patents	Non-Free	--
UCSB-MRL thermoelectric database	Experimental	Thermoelectric Materials	Thermoelectric Properties	18,000	Literatures	Free	[346]
NRELMatDB	Computational	Inorganic Solids	Quasiparticle Energies, Renewable Energy Application	--	Literatures, Patents	Free	[347]
Metallurgical Thermochemistry, Kubaschewski	Experimental	Thermoelectric Materials	Thermoelectric Properties	--	Hard-copy sources	Non-Free	--
3D Materials Atlas	Experimental	General Materials Data	3D Characterization	--	--	Free	--
Inorganic Material Database (AtomWork)	Experimental	Inorganic Solids, Metals	Material Properties, Phase Diagrams	82,000	Literatures,	Non-Free	--
Mineralogy Database	Experimental	Minerals	Structure Properties, physical and optical Properties	4714	Literatures	Free	[348]
CSD Teaching Database	Experimental	Organic Materials	Structure Properties, physical and optical Properties	>750	CSD	Free	--
Database of Zeolite Structures	Computational	zeolites	Multi-purpose repository	--	Literatures, Hypothesis	Free	[349]
RCSB Protein Data Bank	Experimental	biological macromolecular structures	Multi-purpose repository	>173,005	Literatures,	Free	

5. Key Descriptors Bridging Data Intensive Discoveries and Experimental Strategies for Innovative Materials

The key premise of the ML framework is that learning can be viewed as a reasonable model to explain the observed data.[350] Descriptors are the carriers of information exchange between humans and machines. In the context of materials science, they deliver information about molecular properties to machines in digital form. Key to the efficient use of ML in the field of chemical materials is the "descriptor selection" tool, which takes the entire descriptor set as an input, or combines it into a new reduced, but more reliable, descriptor set through correlation analysis while providing a mapping to a Key Performance Indicator (KPI) fingerprint [50]. In this section, the strategy of transforming material data to ML through descriptors is introduced; descriptors can be divided into five main types: constitutional descriptors [20, 36, 351-359]; geometric descriptors [36, 50, 353-361]; quantum chemistry descriptors[11, 12, 20, 36, 50, 190, 351-371]; electrostatic descriptors [36, 50, 352, 355, 361, 363, 365, 369, 370]; combinational descriptors. These will be elaborated upon in the relevant subsections. Finally, we describe some of the extension packages of descriptors in the field of AI for materials science.

5.1. Information Bridging: from Chemical Structures to ML Models

5.1.1. Descriptor Importance

The selection of descriptors directly determines the feasibility of introducing ML to solve the posed question. When the scientific connection between the descriptor and the actuation mechanism is not clear, the causal relationship of the learned descriptor-attribute relationship is uncertain. Therefore, the reliable prediction, identification, and scientific development of new materials are called into question. Analyzing the problem and defining a suitable descriptor is a meaningful and necessary step.[372]

A number of studies have emphasized the importance of material descriptors in accelerating the calculation of material properties or material design. Ghiringhelli, L. M. et al. [372] detail the required characteristics of a set of descriptors: the calculation of descriptors should not be as intensive as that of KPIs; they uniquely characterize materials and the basic processes which pertain to properties; very different materials should be characterized by very different descriptor values (and vice versa); their size should be as small as possible. Sahu et al.,[71] utilized 13 microscopic properties of organic materials as descriptors to build a PCE prediction model. The results indicated that such descriptors can effectively be applied in the context of promising high-throughput virtual screening of new donor molecules for efficient organic photovoltaics. Implementing descriptors with appropriate features plays an important role in accelerating outcomes of material design, or the study of material characteristics.

5.1.2. Bridging and Transferring Process

Data bridging and transfer processes often introduce uncertainty to ML predictions. The evaluation of this uncertainty indicates whether the required prediction accuracy has been satisfied.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/adma.202104113.

The MGI[24] aims to capture, manage, and utilize material structure/property information on a large scale to enable the rapid, cost-effective, and efficient development of new materials with predictable properties. Although the use of such "genome" methods (to promote attribute prediction, virtual design, and material discovery) is relatively new, the concepts driving the development of materials informatics are firmly grounded in previous lessons learned from the fields of chemoinformatics and bioinformatics.

The management and utilization of material structure/attribute information have increased the significance of cheminformatics to ML; a number of new methods have emerged for information and data conversion. Behler describes some of the ways in which chemoinformatics and ML methods have been adapted for materials science and engineering applications, including methodologies to create, verify, and use material quantitative structure and property relationship (MQSPR) models[373]. Friederich et al.[354] used full autocorrelation (FA) functions to transfer the features of chemical complexes. Combining DFT and ML methods, the obtained predictions of reactivity within large chemical spaces containing thousands of complexes. Affordable descriptors were transferred as functions and demonstrated as fingerprints for each complex by considering a specified product of atomic properties (PiPj) calculated in terms of all atoms. Compound compositions were guided by the properties of atoms i and j (**Figure 17**a). These atomic properties include electronegativity, atomic number, identity topology, and size. Each descriptor is multiplied as a function of Diracδδ to encode the structure and properties of the compound.

The selection of the descriptor, removal of redundant features, and establishment of relationships are crucial to the process of transferring information. As shown in Figure 17b, the prediction strategy integrates input HOIP data with the ML algorithm and DFT calculation [36]. Based on the ML program, an input HOIP dataset is established; each input item is described by a signature that is used to train and test the ML model. Element design analysis is required as a prerequisite to remove redundant features and establish structure-attribute relationships. After the input feature set is fixed, grid search technology and 5-fold CV are utilized to select the best descriptor. The network is subsequently

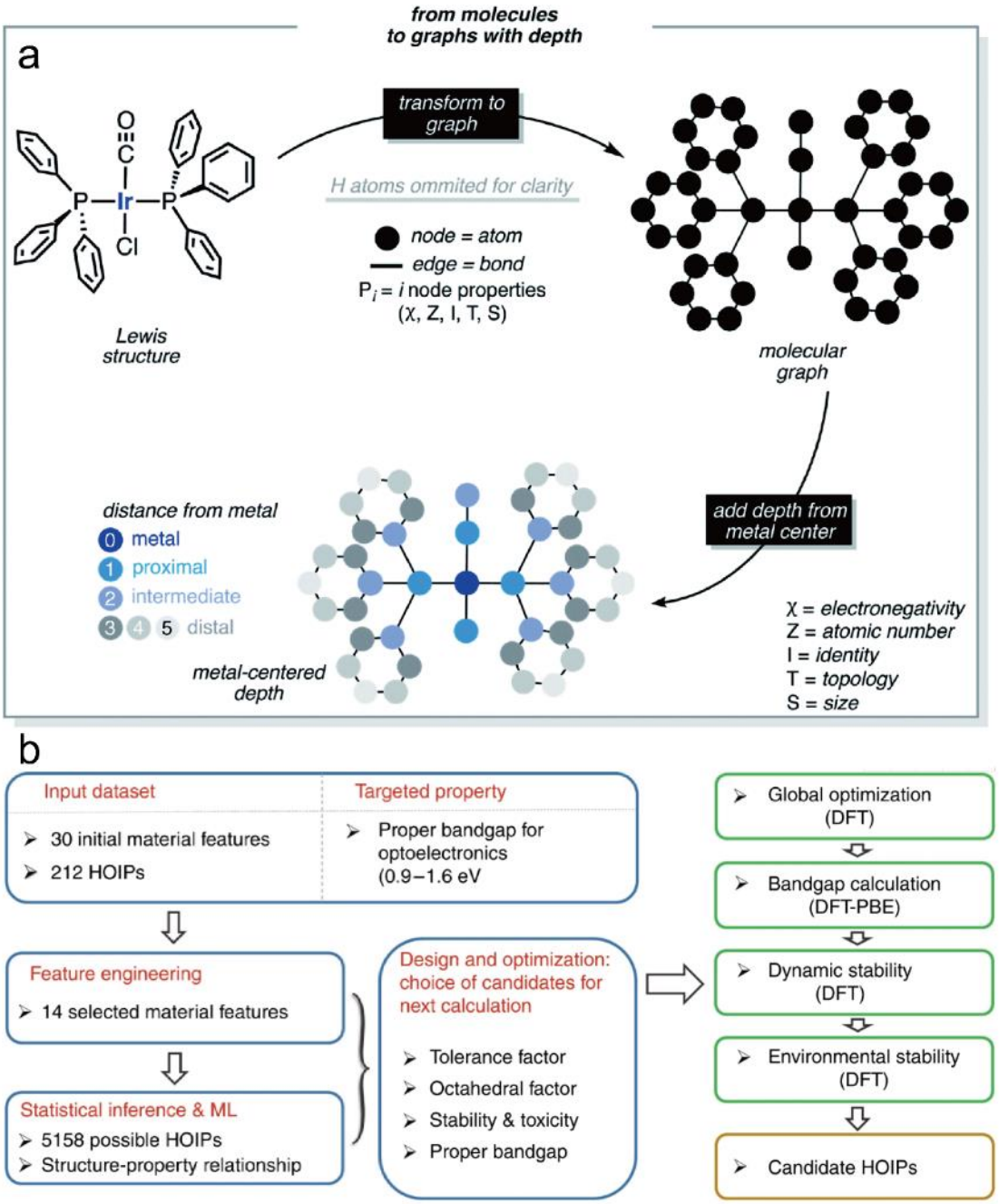


Figure 17. a) Schematic diagram of molecular graph in the calculation of autocorrelation and deltametric functions. Reproduced with permission.[354] Copyright 2020, RSC Publications. b) The schematic diagram of designing lead-free HOIP based on ML combined with DFT. The blue box represents the process of screening through the ML algorithm from the HOIP database. The green box indicates the use of DFT to calculate the electronic performance and stability evaluation of the candidate. Reproduced with permission.[36] Copyright 2018, Springer Nature Publications.

trained to predict the electronic performance and stability of the HOIPs. In this work, the 14 most important descriptors were sorted and selected to collectively describe HOIPs in the chemical space. These descriptors included structural features and elemental properties of A-, B-, and X-site ions. Based on linear correlations for features analysis, redundant or irrelevant features could improve

the accuracy and efficiency of the ML model and achieve accurate predictions based on relatively small training datasets. This work successfully predicted the bandgaps of thousands of HOIPs by using the trained ML model. The evaluation of the bridging and transfer process of characteristic information represented by the descriptor is key to successful ML model predictions. In the process of information transfer, it is also essential to provide more accurate descriptors without losing the original information characteristics. Some descriptors, though assigned a large weight, do not contribute to reliable model predictions (i.e. the phenomenon of over-egging the pudding).

5.1.3. Properties of Ideal Descriptors

Descriptors that can train predictive models to adapt to target attributes are highly desirable. **Figure 18**a presents a representative graphical summary of the workflow of the descriptor design, which is usually applicable throughout the development of a novel strategy. This summary represents a general processing method suitable for any application involving the main dataset, descriptor, training model, etc. Traditional methods rely on chemical intuition to determine the key descriptors for a specific application and develop a relationship which best represents observed material properties. It is more desirable, however, to automate the generation of interesting chemical insights through a rational design approach which does not rely on chemical intuition.

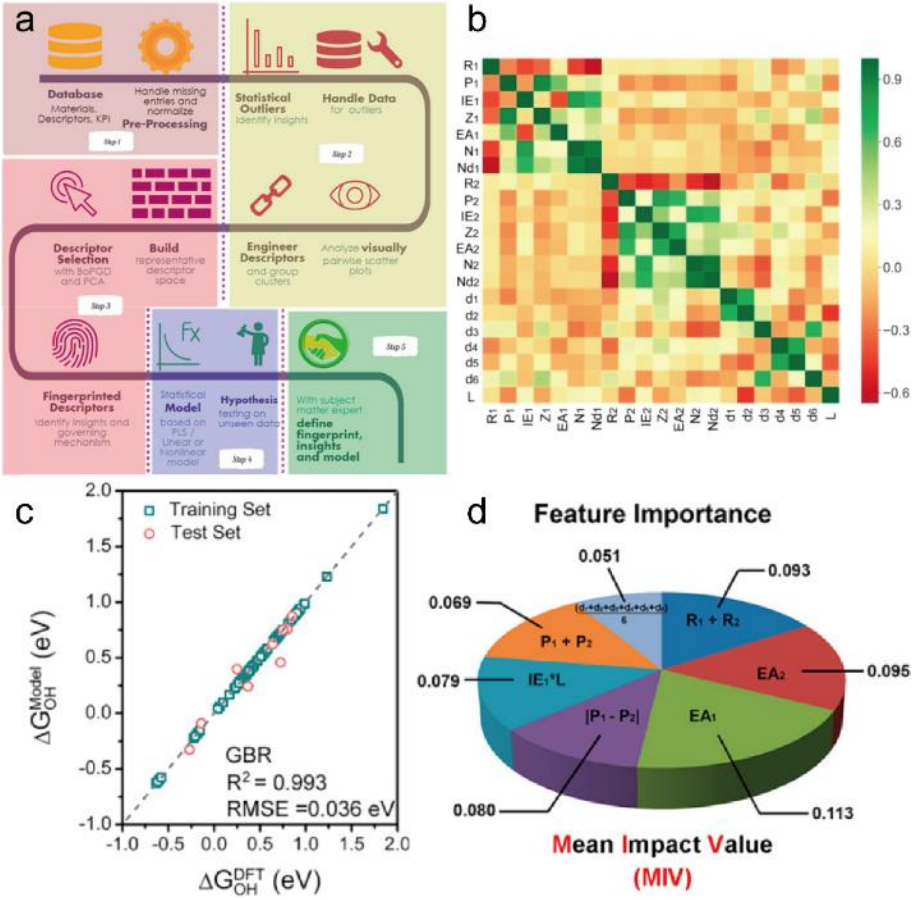


Figure 18. a) The relationship between data, descriptors, and models. Reproduced with permission.[50] Copyright 2017, ACS Publications. It involves the following steps: preprocessing, data analysis, fingerprinting descriptors, statistical model or linear/nonlinear model building and validations, and insights from a subject matter expert. b) Heat map of the Pearson correlation

coefficient matrix among the selected features for DMSCs. c) Comparison of DFT-computed ΔG_{OH*} values with those predicted by GBR algorithm. d) Feature importance based on the Mean Impact Value (MIV). b-d) Reproduced with permission.[374] Copyright 2019, ACS Publications.

Regression fitting, correlation coefficient statistics, dataset partitioning, the establishment of new functions, and other methods have been widely applied to locate and rank ideal descriptors which correspond to the most relevant performance features. Meredig and Wolverton[375] introduced a "cluster ranking model" (CRM) framework to identify unique descriptors that can predict the properties of new dopants. They used the X-means algorithm to cluster various dopants together, followed by regression fitting to rank the descriptors, ultimately utilizing the unique descriptors to model the behavior within each cluster. The existence of clusters in various sample datasets (four dopant clusters were present in this study) improves the effectiveness of the method. Given that all descriptors are ranked by using a regression model, they must necessarily fit to the prediction model of the target attribute. Selected descriptors are those that can best predict the target attributes; they are not necessarily indicative of the phenomenological mechanism. Ward et al.[33] generated an extensible set of attributes that can be used for materials with any number of constituent elements. This set of attributes can broadly capture enough diverse physical/chemical properties of materials to form the basis of accurate predictive models. The group used a total of 145 attribute sets, including stoichiometric attributes, elemental property statistics, electronic structure attributes, and ionic compound attributes. They proved that these attributes are sufficient for describing various properties, while also proposing a novel method to divide the dataset into groups of similar materials to improve prediction accuracy. This work demonstrated the applicability of this novel method to the prediction of various physical properties of crystalline and amorphous materials. Zhu et al.[374] employed DFT calculations, with the assistance of ML, to screen highly efficient dual-metal-site catalysts (DMSCs) for oxygen reduction reaction (ORR). They evaluated the correlation coefficient for selected DMSC features, as shown in Figure 18b. The performance of the ML model can be significantly improved by selecting features that are independent from one another (i.e., not redundant), based on an analysis of linear correlations of several features. The speed at which ML-based approaches can be used to arrive at valuable material property insights, including the identification of descriptors, has significantly improved in recent years. To obtain accurate descriptor relevant to the catalytic activity of DMSC, this work reported the seven characteristics which were deemed most relevant to the catalytic performance of DMSCs in terms of Mean Impact Value (MIV) (Figure 18d). These characteristics include: the electron affinity between two metal atoms; Van der Waals radius; Pauling electronegativity difference; the product of ionization energy and the distance between two metal atoms; the relationship between Pauling electronegativity and atomic distance.

5.2. Categories of Descriptors

In recent years, a large number of articles have demonstrated the importance of material descriptors in accelerating the discovery and design of novel materials. When identifying descriptors which are compatible with ML methods for material discovery, the initial set of descriptors should generally be broad/diverse. Both the choice of fingerprint descriptors and the methods employed to discover/estimate unique mappings are critical, especially when dealing with small datasets. From

the perspective of ML, fingerprint descriptors are a subset (or offspring) of a superset of parent descriptors; they are unique to attributes and materials. The dimensionality or cardinality of the descriptor should be kept as low as possible, while the original descriptor space should be sufficient. This mathematical mapping is also unique to the construction model that maps fingerprint descriptors to attributes or KPIs[50]. The key descriptors used in recent studies for training models in materials science are summarized in **Table 2** and are detailed further in subsequent sections.

Table 2. Key Descriptors used for the model training in material science.

Notation	Description	Class	Ref
	Atomic Number	Constitutional	[351-356]
	Atomic Weight	Constitutional	[353, 356, 357]
	Numbers of and orbital electron	Constitutional	[36, 352, 355]
	Numbers of and valence electron	Constitutional	[352, 353, 355-357] [36]
MN	Mendeleev number	Constitutional	[356, 357]
	Melting Temperature		[356, 357]
	Bond Number	Constitutional	[358]
	Space Group Number	Constitutional	[356, 357]
CN	the number of atoms of that element coordinated	Constitutional	[20, 351, 354, 359]
	Pauling electronegativity	Quantum chemical	[351] [36, 356]
			[190, 352, 354, 355, 357]
	The median monometallic adsorption energy	Quantum chemical	[351]
IC	Ionic Charge	Quantum chemical	[36]
EA	Electron Affinity	Quantum chemical	[36, 50, 352, 355, 361, 362]
IE	Ionization Energy	Quantum chemical	[36, 50, 352, 361-363]
HOMO	The highest occupied molecular orbital	Quantum chemical	[36, 356, 363]
LUMO	The lowest unoccupied molecular orbital	Quantum chemical	[36, 356, 363]
	Bandgap Energy	Quantum chemical	[356, 357, 360, 364]
WF	Work Function	Quantum chemical	[50, 361]
	Binding Energy	Quantum chemical	[20, 50, 190, 363, 365-368]
	Adsorption Energy	Quantum chemical	[11, 20, 50, 190, 353, 355, 358, 359, 365-

			369]
	Local Pauling electronegativity	Quantum chemical	[50, 361]
	Cohesive energy	Quantum chemical	[355]
DOS	Density of states	Quantum chemical	[370, 371]
PDOS	Partial Density of states	Quantum chemical	[358]
	Bader Charge Transfer	Quantum chemical	[355]
	Fermi Energy	Quantum chemical	[50, 370, 371]
	Gibbs Free Energy	Quantum chemical	[12, 353, 355, 358]
G	Surface Energy Density	Quantum chemical	
	Total energy of surface slab obtained	Quantum chemical	[12]
	Bulk energy per atom	Quantum chemical	[12]
H	Over potential	Quantum chemical	[12, 360, 370, 376]
	Current density		[376]
	Activation energy	Quantum chemical	[190, 354, 370]
	Transition-state energy	Quantum chemical	[190, 354, 363, 370]
	Atomic nearest-neighbor distances		[190]
	Optical gap energy	Quantum chemical	[362, 363]
	Width of a band	Electrostatic	[50, 365]
	Centre of a band	Electrostatic	[50, 355, 365, 369, 370]
	Skewness of a band	Electrostatic	[50, 365]
	Kurtosis of a band	Electrostatic	[50, 365]
	Filling of a band	Electrostatic	[50, 365]
	Spatial Extent of -orbitals	Electrostatic	[50, 361]
	Adsorbate-metal coupling matrix element	Electrostatic	[50, 361, 370]
	Metal -metal coupling matrix element	Electrostatic	[370]
	Partial distribution function	Geometric	[359]
	Polarizability	Electrostatic	[36, 363]

First ionization potential	Electrostatic	[36, 352, 355]
Magnetic Moment	Electrostatic	[357]
Bond Length Position	Geometric	[358]
Atomic Identity	Geometric	[354]
Optical Transmittance		[362]
Lattice parameters	Geometric	[355]
Molar Ratio		[355]
Dipole moment	Electrostatic	[363]
Atomic Radius	Geometric	[50, 353, 355, 361]
Rotational angles	Geometric	[360]
Distance between two layers	Geometric	[360]
Bond Length	Geometric	[353, 355, 358, 360]
Bond Angle	Geometric	[353]
Distance to alloy atoms	Geometric	[359]
Estimation for the interatomic distance using Vegard's law	Geometric	[359]
Covalent Radius	Geometric	[354, 356, 357]
Specific Volume	Geometric	[356, 357]
Van der Waals radii	Geometric	[352]
Tolerance Factor	Geometric	[36]
Octahedral Factor	Geometric	[36]
Iron Radii	Geometric	[36]
Sum of the of and orbital radii	Geometric	[36]
Atomic Radius	Geometric	[50, 353, 355, 361]
Cutoff radius		[11, 377]
Bond distance		[11, 354]
Atom pair distance		[11]

5.2.1. Constitutional Descriptors

Constitutional descriptors are the simplest and most commonly used descriptors in materials science. They contain compositional information about materials without their geometric or topological information. Hence, constitutional descriptors are also known as 0D or 1D descriptors. The most widely used constitutional descriptors are the numbers of atoms, bonds, electrons, and rings; molecular weight; and atomic composition indices. Constitutional descriptors are not sensitive to conformational changes and are easily calculated. Constitutional descriptors, which are easy to obtain, often appears as part of combined descriptors. Despite their simplicty, such descriptors can convey essential information, generally in combination with other classes of descriptors.[10, 18]

5.2.2. Geometric Descriptors

Geometric descriptors, also known as 3D descriptors, are molecular representations that convey structural information about the material. Common geometric descriptors include the 3D-Wiener index, gravitational indices, molecular surface area, molecular volume, radial distribution function, and WHIM descriptors. Topographic indices can be regarded as a special subset of geometric descriptors. For instance, Ruck et al.[19] proposed a ML framework that can accurately predict strain, while rationalizing the impact of strain on a Pt core-shell nano-catalyst's oxygen reduction activity. This work predicted the strain coordination on core-shell nanoparticle atoms by applying geometric descriptors to ML, including coordination number, partial distribution function, distance to alloy atoms, and interatomic distance (from Vegard's law). The generalized coordination number under strain was the basis for the linear relationship between the strain and the adsorption energy of *OH and *OOH. The formulation of this descriptor enabled the identification of the most favorable active site on the core-shell nanoparticles. The novel generalized strain coordination number descriptor proposed in this work furnished accurate predictions of strain within 3%. Zhang et al. [11] assess the local structural environment in the vicinity of a selected adsorption site on an amorphous Ni_2P catalyst. The bond distance between each sub-pair from the seven designated atoms was selected as the primary indicator. In this manner, the adsorption energy can be characterized more accurately by first specifying the surface structure attributes. In addition, use of the bond distance within the specified local structure as a feature may implicitly ignore the influence of other chemical environments. In this study, a chemical environment representation method based on the symmetry function of the atomic center is also presented, which is suitable for periodic systems and is independent of bonding properties. The symmetric function transforms Cartesian coordinates into a set of symmetric functions that describe the chemical environment of atoms; this approach has been proven to successfully fit the potential energy surface (PES).

5.2.3. Quantum Chemical Descriptors

Quantum chemical descriptors are widely used in the screening of catalysts, and are commonly sub-divided into energy- and electron-based descriptors. Energy descriptors include common system total energy, electron energy, bandgap, valence band top (VBM), conduction band bottom (CBM), d-band center energy, formation energy, binding energy, Gibbs free energy, highest occupied molecular orbital (HOMO), lowest unoccupied molecular orbital (LUMO), and other descriptors. Electronic descriptors typically include electron affinity (EA), electron density, localization, charge

transfer and distribution, and other descriptors. Standard first-principles calculations based on DFT can be used to generate the aforementioned quantum chemical descriptors. The acquisition of such descriptors is usually combined with DFT calculations (or extracted from existing databases); in this manner, new material properties are discovered through a synergy between quantum chemistry and ML.

DFT-based calculations of metal surface adsorption and reaction are now mature enough to contribute to our understanding of the complexity of the catalytic surface and the adsorbate. Such calculations are also accurate enough to characterize the bonding mechanism, determine the reaction path, and compare different systems relevant to heterogeneous catalysis. Though experimentation is irreplaceable, such calculations may offer a simple (or, in some cases, the only) approach to assessing potential catalyst properties. [378] Peterson et al. [366] compared the binding energy trends of intermediates during the electrochemical reduction of CO_2 and proposed the novel “active volcano" descriptor for the first time. This descriptor effectively describes experimentally observed trends in transition metal catalysts, including offering specific interpretations as to the dominance of copper as an electrocatalyst. This study also proposes a new strategy for discovering catalysts that can operate at reduced overpotentials. Extending classic theoretical calculations introduces avenues for the accelerated discovery of high-performance electrocatalysts. Owing to the significant heterogeneity of exposed active sites and the variations in the crystal structure with composition, the exposed surface may be different from that of normal single metal nanoparticles; this surface heterogeneity must be captured by further DFT calculations. Traditional methods, which are more suited to single-metal catalyst, cannot effectively handle this complexity. This work systematically considered all active sites to address this problem. The number of DFT calculations, though large, is feasible to implement for a small number of composites. Ulissi et al.[20] implemented a neural network potential fitted with DFT to greatly reduce the thousands of DFT calculations which would otherwise have been required to obtain the relaxation adsorption energy of each adsorption site on each surface. However, this method does not consider surface segregation or apparent disorder of crystal components, demonstrating a discrepancy with respect to the real experimental environment. Artrith et al. [190] constructed an ML model to predict the transition state energy from the thermochemical reaction energy (model 1). The descriptors selected for model 1 included geometric descriptors, chemical species, binding energies, and reaction energies. A second ML model (model 2) was then trained to capture the behavior of catalytic activity and selectivity based on all transition state energies. Descriptors selected for model 2 included the results and descriptors from model 1 (shown in **Figure 19**a). Both models could directly predict catalytic activity/selectivity from chemical properties and attributes which can be determined from high-throughput DFT calculations. Integrating a large DFT calculation datasets into the trained ML model (applying a simple linear regression model between experimental catalytic activity and selectivity) the key C-C bond scission reaction step involved in the ethanol reforming reaction was determined. Ma et al.[379] determined the reactivity descriptors that characterize effectiveness of alloy electrocatalysts in selectively converting CO_2 to C_2 species. Based on the reactive descriptor (ie CO adsorption energy), the theoretical limits of the potentials of essential CO_2 electroreduction reaction steps (along the C_1 and C_2 paths) were calculated. Inspired by d-band chemisorption theory, input features pertaining to bimetallic surfaces included characteristics of the d-states distribution. Physical constants related to the host metal were treated as secondary characteristics to better describe the tendency of

chemical bonding on a series of metal surfaces. Having developed a comprehensive descriptor-based catalyst design method based on ML, the study has identified a promising 001-terminated Cu polymetal, which possesses a relatively low overpotential and high efficiency/selectivity for the reduction of CO_2 to C_2. The predicted rate obtained by combining DFT calculations with rate theory was found to be in good agreement with available experimental data pertaining to the formation of various products on several metal electrodes and within the potential range applied by J. et al.[367] A two-parameter is proposed in Figure 19b. This pre-screening tool is composed of various H-atom adsorption energies (at the top site) and CO adsorption energies to identify the most promising CO_2RR catalyst candidates. The tool can also predict whether the electroreduction product is a hydrocarbon/alcohol, H_2, CO, or HCOO--. However, to predict the selectivity of a given product, the activation energy in each fundamental step needs to be calculated to evaluate the relationship between the reaction rate and the applied potential. The insights gained from such calculations can be used to develop standards to identify new and improved catalysts for electrochemical reduction of CO_2. Bai et al. [380] used ML to build a model that can correlate four attributes with hydrogen evolution rates. This was achieved by selecting ionization potential (IP, approximated by the energy of the highest occupied molecular orbital (HOMO)), electron affinity (EA, approximated by the energy of the lowest unoccupied molecular orbital (LUMO)), optical gap, and experimentally measured transmittance as the four descriptors (Figures 19c and d). The model was evaluated by using the LOOCV, indicating that the test data are applicable to copolymers that were not considered during training. The results indicated that the correlation between the HER of the polymer and each of the individual properties is relatively weak supporting the view that the photocatalytic activity is a composite property that cannot be encapsulated by only a few descriptors. The electrochemical reduction of CO_2 often generates a variety of products determined by the reaction conditions and catalyst performance, including some that form valuable chemical substances (such as hydrocarbons and alcohols). Bagger et al. [368] calculated key binding energies for non-coupled intermediates,- ΔE_{H*}, ΔE_{COOH*}, ΔE_{CO*}, and $\Delta E_{CH3O*,}$ to identify the "genes" of CO_2 products. The extensive exploration of quantum chemistry descriptors makes the theory (in conjunction with ML approaches) better than experimental designs, considerably reducing the costs associated with experimentation. The identification of novel quantum chemistry descriptors will gradually increase our understanding of catalytic phenomena.

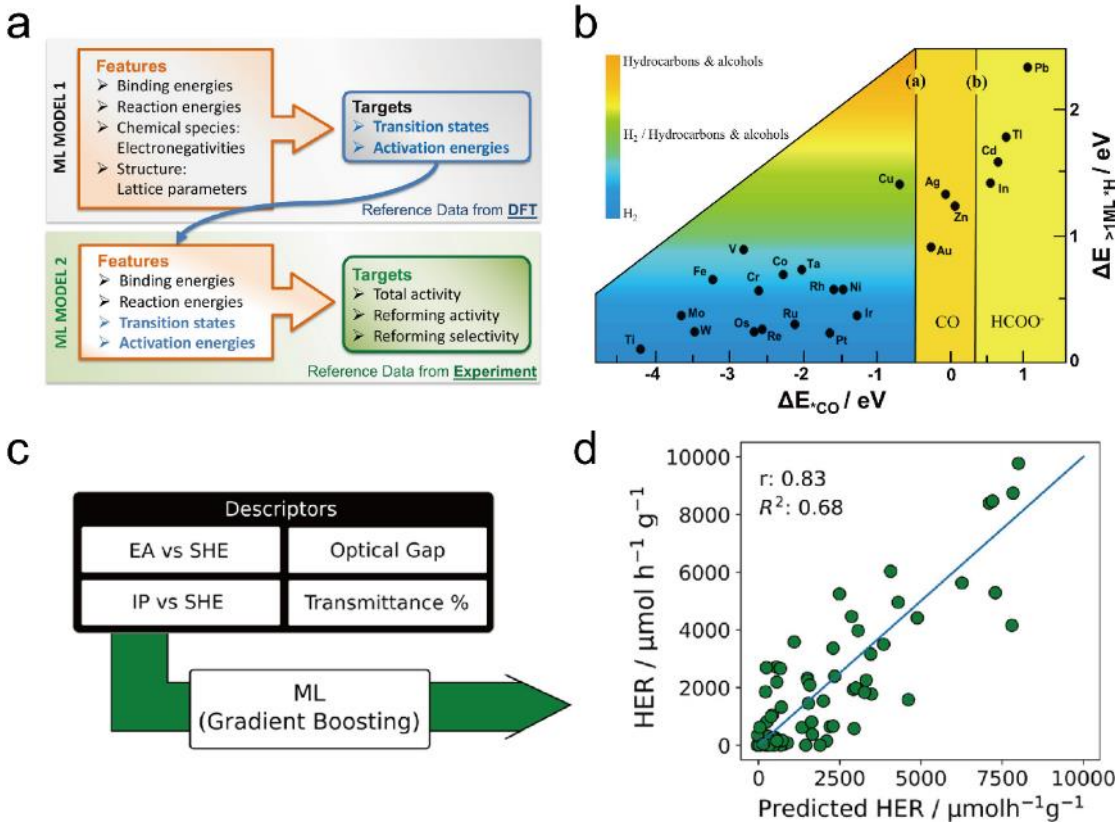


Figure 19. a) Flowchart of the combined ML approach consisting of two ML models. Reproduced with permission.[190] Copyright 2020, ACS Publications. b) Two-parameter descriptor of the electrocatalytic activity of metal electrodes. Reproduced with permission.[367] Copyright 2018, ACS Publications. c) Properties used to train the gradient-boosting model, where IP, EA, and optical gap are calculated, and transmittance is measured experimentally. d) Experimentally observed hydrogen evolution rates vs hydrogen evolution rates predicted using a gradient-boosted trees ML model. The model is evaluated by leave-one-out cross-validation, meaning the data shown are for co-polymers not considered during training. c-d) Reproduced with permission.[380] Copyright 2019, ACS Publications.

5.2.4. Electronic Descriptors

The categories of electronic descriptors often overlap with quantum chemical descriptors; classification of electronic descriptors, however, tend to be more detailed. They primarily include descriptors pertaining to atomic charges in the material, such as charge polarization, positive and negative of charges, number of charges, and electron density. The selection of electronic descriptors often also involves exploring the physical properties metals and alloys, transition metals, and metal atom doping materials. Electronic descriptors are most suited to contexts in which electronic transport influences material properties. Wexler et al. measured the relative importance of various descriptors in describing the HER activity of non-metal-doped $N_{i2}P(0001)$ surfaces [10]. They compiled bond lengths, bond angles, charges, mass numbers, atomic weights, and atomic radii as geometry descriptors; other geometric parameters for pertaining to the DFT-relaxed structure were adopted as structural and charge descriptors. Key to this approach is defining the normalization ability of the descriptor based on ΔGH data. **Figure 20**b demonstrates the top 10 descriptors contained in the

dataset. The first two descriptors are: the selected Ni-Ni bond length (constituent atoms are distinguished by their respective distance from the first doping site; see Figure 20); the average Ni-Ni bond length. Among the 10 descriptors exhibiting the highest correlation, seven are geometric descriptors pertaining to the shape of the Ni3 hollow site. Another important characteristic is the standard deviation of the dopant charge. This study also utilized atomic charge as a descriptor, finding that its correlation with measured properties is relatively poor. This indicates that the electronic partition metric may not be important for the analysis of HER performance. This work highlighted unnecessary/potentially redundant descriptors which do not directly affect the bonding; in addition, this property change is already implicit in the Ni-P bond length descriptor. Sun et al.[112] successfully introduced metal atom Bader charge transfer, metal d-band center, and d-orbital electron number below the Fermi energy as part of the DFT-calculated descriptors to investigate the hydrogen evolution performance of MXene and MBenes, both doped and not doped with single atoms. Electrostatic descriptor can be interpreted as a type of cross-descriptor involving both physical and chemical properties of materials; they are often indispensable to ML-based frameworks to model material properties

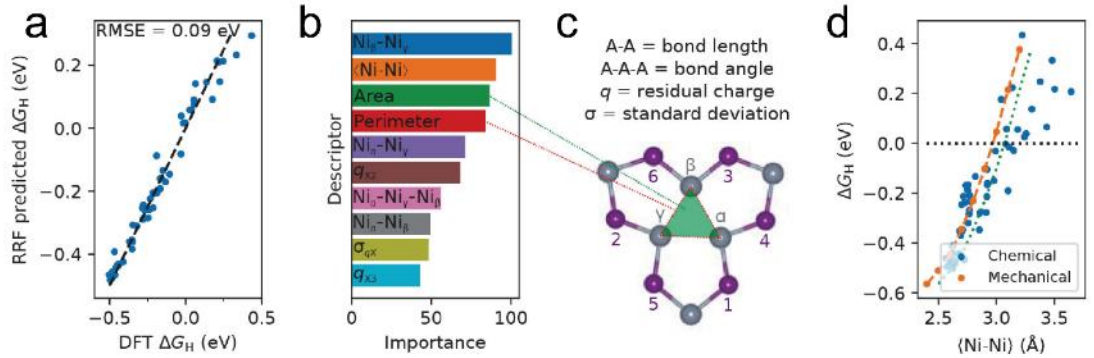


Figure 20. a) ΔGH predicted by RRFs vs DFT. b) Relative importance of descriptors calculated from RRF model. Only the top 10 features are shown. c) Definition of descriptors in b). We label the three Ni atoms α, β, and γ based on their distance from the first doping site. d) Effect of average Ni–Ni bond length on ΔG_H as induced by chemical pressure and mechanical pressure. a-d) Reproduced with permission.[10] Copyright 2018, ACS Publications.

5.2.5. Combinational Descriptors

Recent trends have focused on the selection and development of combinational descriptors. Because the various classes of descriptors are complex and numerous, combined descriptors can often transcend this complexity, demonstrating better expression of material properties than descriptors of single type in of the. Zhong, et al.[18] incorporated DFT data into an ML workflow to predict CO adsorption energies for each adsorption site enumerated in **Figure 21**. DFT data for CO adsorption energies were saved in a database. Each element present in the bulk structure (the list of which originated from the Materials Project database) was described with a vector of four numbers: atomic number (Z), Pauling electronegativity (χ), number of atoms of the element coordinated with the CO molecule (CN) as determined by a cut-off radius of 5 Å and a Voronoi polyhedral angle cutoff tolerance of 0.8, and median monometallic adsorption energy of CO on that element (ΔẼ), as extracted from the database of CO adsorption energies.

The hydrogen evolution performance of MXenes and MBenes (doped or not doped with single atoms) was systematically investigated Sun et al.[112] based on a combined DFT and ML framework. Correlation analysis was employed to reduce the number of descriptors employed. Remaining descriptors (various element-specific attributes, structural energy, and lattice parameters) were obtained with relative ease and uniquely characterize corresponding physical and chemical properties. This work applied the following DFT-calculated descriptors: cohesive energies of MXenes and MBenes; Bader charge transfer of the metal atom, doping atom and C/N/B; d-band center of the metal; d-orbital electron number below Fermi energy; bond length between the metal/doping atom and the nearest metal of the same layer; bond length between the metal/doping atom and the nearest boron; molar ratio of metal; C/N/B lattice parameters. Elemental descriptors comprised the atomic mass, period number, group number, atomic radius, valence electron, electronegativity, electron affinity, first ionization energies of the metal, and doping atom. Ge et al.[381] utilized four combinational descriptors: cosine of the rotation angle, distance between the two secondary parts, ratio of the average bond length, and bandgap of MX_2. These newly generated descriptor PL can encapsulate the electrocatalytic performance of NiOER and are effective for both HER and OER. Friederich, P., et al.[354] utilized full autocorrelation (FA) functions to combine features and overcome descriptor complexity. Combining descriptors will encapsulate a larger number material properties early on in the ML/DFT workflows, improving its accuracy and efficiency and analyzing ever more complex systems.

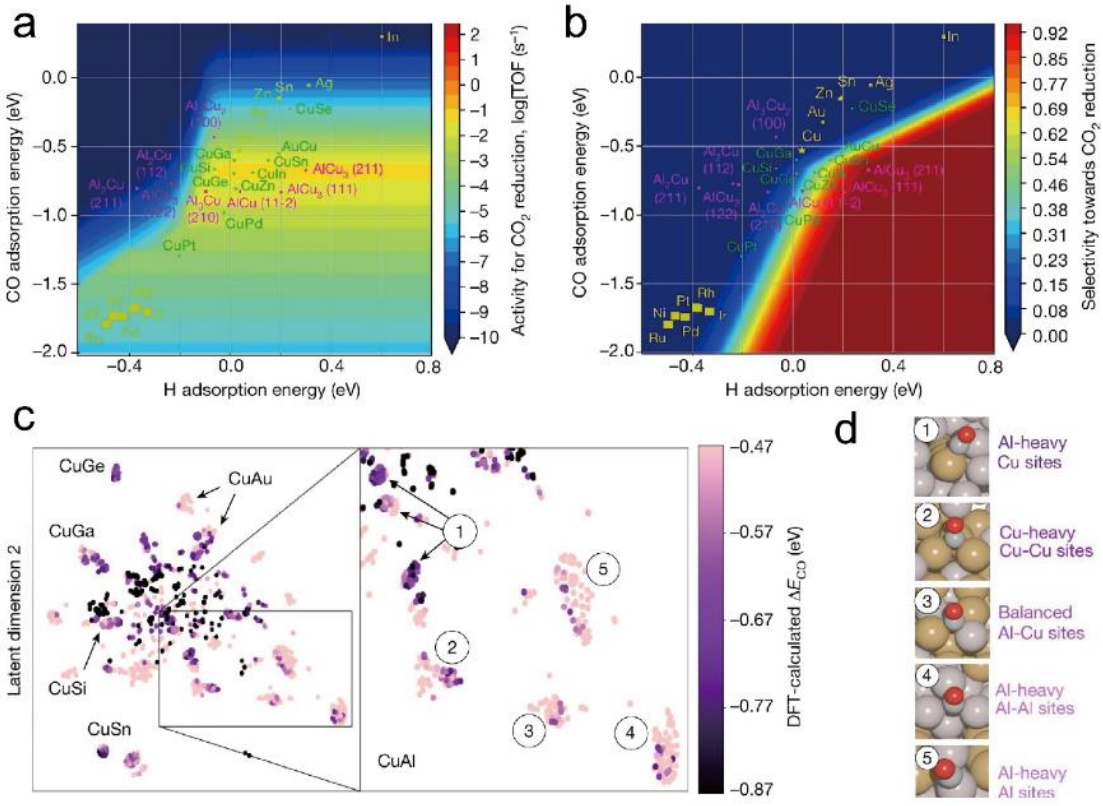


Figure 21. a) two-dimensional activity volcano plot for CO_2 reduction. TOF, turnover frequency. b) A two-dimensional selectivity volcano plot for CO_2 reduction. CO and H adsorption energies in panels a and b were calculated using DFT. Yellow data points are average adsorption energies of monometallics; green data points are average adsorption energies of copper alloys; and magenta

data points are average, low-coverage adsorption energies of Cu-Al surfaces. c) t-SNE19 representation of approximately 4,000 adsorption sites on which we performed DFT calculations with Cu-containing alloys. The Cu-Al clusters are labelled numerically. d) Representative coordination sites for each of the clusters labeled in the t-SNE diagram. Each site archetype is labelled by the stoichiometric balance of the surface, that is, Al-heavy, Cu-heavy or balanced, and the binding site of the surface. a-d) Reproduced with permission.[18] Copyright 2020, Springer Nature Publications.

5.3. Descriptor-Related Tools

Complementary tools have gradually been developed over the past year to support the application of advanced descriptors. In this section, we will introduce recent open-source descriptor-related tools which are implemented with Python infrastructure and API-based frameworks for data sharing. Such tools are able to rapidly implement the complicated process of managing, transmitting, sharing and sending all ML relevant to a particular materials science problem.

5.3.1. Programming Packages and Codes

In this subsection, we summarize the following representative programming packages and codes: Fixed-Size Numeric Descriptor Generator (DScribe),[382] Sure Independence Screening and Sparsifying Operator (SISSO),[26] and LASSO.[383] These packages are widely used in materials science to generate appropriate ML descriptors.

DScribe

The application of ML in materials science is usually hindered by the lack of data conversion processing prior to training the model. Such data is usually converted into a specific descriptor, which is a key step in building an ML model for attribute prediction in materials science. DScribe is a convenient bridge between data and descriptors. DScribe[382] is a software package for ML that provides popular feature transformations ("descriptors") for atomic material simulation. DScribe accelerates the application of ML in atomic property prediction by providing user-friendly, ready-made descriptors. Currently, the DScribe software package contains descriptors that can be represented in vector form and do not depend on any specific learning model. By decoupling descriptor creation from ML models, users can experiment with various descriptor/model combinations in parallel, and can directly apply emerging learning models to existing data. The software package currently contains implementations of the Coulomb matrix, Ewald sum matrix, sine matrix, multi-body tensor representation (MBTR), atomic center symmetry function (ACSF), and atomic position smooth overlap (SOAP). The library is based on the python interface, with computationally intensive routines written in C or C++. Source code, tutorials and documentation can be obtained online. These introductory materials use the following examples to illustrate use cases of the package: the prediction of solid formation energies; the prediction of the atom ion charges in organic molecules.

SISSO

The lack of a reliable method for identifying descriptors is one of the key factors hindering the development of effective materials. The SISSO[26] is a new systematic method for discovering material property descriptors in the framework of dimensionality reduction based on compressed

sensing. The SISSO solves the large and relevant feature space and converges to the best solution based on feature combinations most suited to the target material characteristics. In addition, the SISSO requires only a small amount of training to obtain stable results. This method is based on the quantitative prediction of the ground state enthalpy of octal binary materials (using ab initio data) and is applied in the illustrative example (with experimental data) to predict the classification of binary metals/insulators. In both cases, an accurate predictive model can be generated. The predictive ability of the metal-insulator classification model has been validated on test data and it rediscovers the transition from insulator to metal caused by the available pressure and allows the prediction of immature transition candidates, which has been laid for experimental verification of the foundation. Compared with previous model recognition methods, the SISSO can become an effective tool for automatic material development.

LASSO

The LASSO[383] typically penalizes high weights to avoid the occurrence of overfitting, while adjusting/reducing the coefficients of the regression model to finally generate the most reasonable number of optimal descriptor KPIs. To reduce the computational effort of employing DFT to calculate a large amount of combined data, the LASSO identifies only those physical descriptors that have a significant impact on adsorption performance. Ge et al.[381] generated 257,703 possible descriptors through calculations. Based on these descriptors, the LASSO fits an equation that best describes the linear relationship. This process was repeated 50 times, each instance standardizing the remaining 90% of the training data prior to the LASSO step. By evaluating the predicted error characteristics of all possible descriptors, we considered the rotation angle of the TMDC heterojunction as a key descriptor describing catalytic performance. Four variables were used: cosine of the rotation angle, distance between the two secondary parts, ratio of the average bond length, and bandgap of MX_2. The new generated descriptor PL can effectively capture the electrocatalytic performance of NiOER, and is effective for both HER and OER.

5.3.2. Descriptor-Related Software

In this subsection, representative descriptor-related software (Open-Source Cheminformatics Software (RDKit), Commercial Descriptor Generation Software (Dragon), Open-Source Descriptor Generation Software (PaDEL-Descriptor)[384]) is briefly discussed.

Open-Source Cheminformatics Software (RDKit)

RDKit (RDKit: Open-Source Cheminformatics Software. https://www.rdkit.org. Accessed 07 Aug 2020) is an open-source toolkit for chemoinformatics, based on the 2D and 3D molecular manipulation of compounds, using ML methods for compound descriptor generation, fingerprint generation, compound structure similarity calculation, 2D and 3D molecular display, etc. RDKt is a very powerful open-source chemical information python toolkit. Its core data structure and algorithms are implemented in C++. It enables a large number of 2D/3D calculation operations on chemical molecules to generate molecular descriptors for ML. Many of the latest ML software packages are based on the use of RDKit's open-source tool creation.

Descriptors in RDKit contains properties such as the number of benzene rings, the number of functional groups, and LogP, which correspond to various properties reflected in the structure of the

molecule. It follows that a combined descriptor may also be proposed which can represent all the partial structures of the molecule.

Commercial Descriptor Generation Software (Dragon)

Dragon (https://chm.kode-solutions.net/products_dragon.php) is the most widely used application for molecular descriptor calculation. Its new version, Dragon 7.0, provides an improved user interface, new descriptors, and additional features such as fingerprint calculation and support for disconnected structures. Dragon can evaluate 5270 molecular descriptors, making it compatible with most theoretical methods. The list of descriptors includes: the simplest atom type, functional group, and fragment number; topology and geometric descriptors; three-dimensional descriptors; multiple attribute estimates (such as logP); drugs and lead-like alarms (such as Lipinski's alarm). Dragon has established an easy-to-operate graphical user interface and command line interface, which is very useful for batch processing of large amounts of data. Dragon now also enables the calculation of hash molecular fingerprints, which can completely customize several parameters and generate all molecular fragments used in the fingerprint process. The graphical user interface also includes more advanced tools to analyze the descriptors following data processing (extended univariate statistics, pairwise correlation, principal component analysis) and import user-defined variables (such as available experimental values) to perform the merge set operation. Starting from version 7.0, Dragon allows the calculation of descriptors for molecules with disconnected structures (such as salts, ionic liquids), thereby providing various theoretical methods to extend the descriptor algorithms for such structures.

Open-Source Descriptor Generation Software (PaDEL-Descriptor)

PaDEL-Descriptor[384] is software for calculating molecular descriptors and fingerprints. The software can calculate 797 descriptors (including 663 1D, 2D descriptors and 134 3D descriptors) and 10 fingerprints. These evaluations of these descriptors and fingerprints is based on the Chemistry Development Kit. Descriptors and fingerprints include atomic type electron topological state descriptors, McGowan volume, molecular linear free energy relationship descriptors, ring numbers, counts of chemical substructures identified by Laggner, binary fingerprints, and Klekota Count of chemical substructures recognized by Roth. The PaDEL-Descriptor is developed in Java and consists of both library and interface components. The library component allows easy integration with quantitative structure-activity relationship software to furnish descriptor calculation functions, while the interface component allows it to be used as an independent software. The software implements a Master/Worker framework to speed up the calculation of molecular descriptors by utilizing multiple CPU cores in parallel. Therefore, this tool offers many key advantages relative to other independent software for the calculation of molecular descriptor. It is open source with both a graphical user interface and command line interface. It can run on all major platforms (Windows, Linux, MacOS) and supports more than 90 different molecular file formats.

6. Applications of Data-Driven Innovative Materials

The success of a large number of ML applications in materials science has preliminarily demonstrated the capability of data-driven approaches in the discovery of innovative materials. By

appropriately integrating ML techniques, material databases, and molecular descriptors, material properties can be efficiently and accurately predicted to support the focused design of innovative materials. Such approaches represent a synergy between materials science, computer science, and mathematics. In this section, recent advances in the applications of such synergies to the development of materials for energy conversion and storage,[9-13, 381, 385] environmental decontamination,[14] flexible electronics,[16] optoelectronics,[386] superconductors,[277] metallic glasses,[33] and magnet materials are investigated. The data-driven strategies, ML techniques, and corresponding performance are evaluated with respect to the specific material-focused question being addressed in each application. In addition, cases that employ ML techniques to implement data augmentation and feature generation are also discussed.[26]

An overview of the applications of data-driven, innovative material discovery is represented in **Table 3**. These examples will be discussed in greater detail in the subsequent sections. It is striking that a number of the data-driven techniques described in the previous sections have not yet found application in innovative material discovery. A discussion on such possibilities will also be provided in greater detail, followed by an overall future outlook.

Table 3. Data-driven innovative material applications.

Applications	Materials	Target Properties	ML Model/Algorithms	Data Source	Most Related Descriptors	Type of Canonical Test	Best Performance on Test Data	Ref.
HER	Ni_3P_2(0001) of Ni_2P	Adsorption free energy of H* (ΔG_H)	•Regularized Random Forests (RRFs)	•DFT computation	•Ni-Ni bond length •Ni-Ni-Ni bond angle •Hollow site area •Hollow site perimeter	3-fold CV	ΔG_H: RMSE of 0.09 eV	[10]
	Amorphous Ni_2P	•Frozen adsorption energy (E_{frozen}) •Relax adsorption energy (E_{relax})	•ANN •Gradient boosting DT •GA	•DFT computation	•Bond length •Symmetry functions	Holdout	•E_{frozen}: RMSE of 0.11 eV •E_{relax}: RMSE of 0.10 eV	[11]
OER	IrO2 and IrO3	•Biding free	•Convolutional	•DFT comput	•Atomic structur	Holdout	•Coverage:	[12]

Polymorphs	energy (ΔG) for coverage calculations •Biding free energy (ΔG) for OER calculations	neural network (CNN)	ation •Material Project	es		MAE, RMSE and R^2 of 0.07 eV, 0.10 eV and 0.93, respectively. •OER: MAE, RMSE and R^2 of 0.13 eV, 0.18 eV and 0.8, respectively	
Doped RuO_2 and IrO_2	•Identify new descriptors for calculation of adsorption enthalpy of O*	•SISSO	•DFT computation	•SISSO Features •Width of the d-band •Charge transfer energy •Filling	5-fold CV	•E_{O*}: MAE and RMSE of 0.65eV and 0.18 eV, respectively	[385]

		(E_{O*})			of the d-band •Kurtosis of the d-band			
OWS	Transition Metal Dichalcogenides (TMDC): MoS_2, WS_2, WSe_2, $MoSe_2$, $MoTe_2$, and WTe_2	•HER Overpotentials (η_{HER}) •OER Overpotentials (η_{OER})	•LASSO	•DFT computation	•Cosine of the rotational angle •The distance between two secondary parts •The average mx2 bond length •The bandgap ratio of the two components	Holdout	•η_{HER}: R^2 of 0.80 •η_{OER}: R^2 of 0.83	[381]
PVs	Lead-free hybrid organic-inorganic	•Bandgap	•GBR	•ICSD	•Tolerance factor •Number	5-fold CV	Bandgap: RMSE and R^2	[36]

perovskites				of ionic charges •Octahedral factor •p-orbital electron		of 0.086 eV and 0.97, respectively.	
Small molecule organic photovoltaic materials	•Power conversion efficiency (PCE)	•Linear regression (LR) •k-nearest neighbour (kNN) •Artificial neural networks (ANN) •Random forest (RF) •Gradient boosting regression tree (GBRT)	•Experiment data •DFT computation	•Hole–electron binding energy in donor molecules •The reorganization energy for holes in donor molecules •The unsaturated atom number in the main	10-fold CV	PCE: RMSE and MAPE of 1.07% and 17.1%, respectively.	[71]

				conjugation path of donor molecules •Polarizability of donor molecules			
Organic photovoltaics materials	•PCE	•ANN •kNN •GBRT	•Experiment data •DFT computation	•Hole–electron binding energy in donor molecules •The reorganization energy for holes in donor molecules •The unsaturated atom number	10-fold CV	PCE: RMSE and MAPE of 1.107% and 21.0%, respectively.	[49]

				in the main conjugation path of donor molecules •The number of hetero atoms			
Organic photovoltaics materials	•PCE •Open circuit voltage (V_{OC}) •Short circuit current (J_{SC})	•k-nearest neighbour (k-NN) •Kernel ridge regression (KRR)	•DFT computation •Literature data	•HOMO energy for the donor •LUMO energy for the donor •LUMO energy for the acceptor •The total internal reorganisation	LOOCV	•PCE: RMSE of 1.33% •V_{OC}: RMSE of 0.1037 V •J_{SC}: RMSE of 3.0464 mA/cm^2	[48]

				energy •Daylight fingerprint •Morgan fingerprint			
Metal oxides photovoltaic materials	•V_{OC} •J_{SC} •Internal quantum efficiency (IQE)	•Principal Component Analysis (PCA) •k-NN •Genetic programming	•Literature data •Experiment data	•The thickness of the absorber layer •Thickness of the window layer •Bandgap of abosorb layer •The distance between the cell and the center of deposition plume	LOOCV Holdout	•V_{OC}: R^2 of 0.92 •J_{SC}: R^2 of 0.90 •IQE: R^2 of 0.91	[54]

				• Resistance of the absorber layer • Maximum value of calculated theoretical photocurrent			
Two-dimensional photovoltaic materials	• Applicability in PV applications	• Gradient boosting classifier (GBC), • Support vector machine (SVM) • Random forest classification (RFC), • Ada	• ICSD	• Packing factor (Pf), • Average sublattice neighbour count (SNC), • Mulliken electronegativity maximum and	5-fold CV	Accuracy, recall, precision, and AUC of 1, 1, 1 and 1, respectively.	[17]

		boostin g (Ada), •Linear regressi on •Stochast ic gradient descent classifie r (SGDC) , •Decisio n tree classifie r (DTC).		minimu m value •average atomic volume •Lattice paramet er •Averag e bond ionicity of sublatti ce •Anion framew ork coordin ation			
Kesterite I2-II-IV-V4 quaternary compounds	•Bandga p	•LR •Support vector regressi on (SVR)-linear kernel •Support vector regressi on	•DFT comput ation •MP	•Electro negativi ty •Ionic radius •Row in the periodic table	10-fold CV	•RMSE and R^2 of 0.283 eV and 0.957 for SVR-RBF, respect ively. •Accura	[387]

		(SVR)-radial bias function kernel •Boosted regression tree •RF •Logistic regression				cy of 1 for Logistic regression	
16-atom constructed wurtzite nitrides in an orthorhombic cell	•Bandgap •Band offset	•Linear regression •Support vector regression (SVR)-linear kernel •Support vector regression (SVR)-poly kernel •Support vector	•DFT computation	•Electronegativity •Covalent radius •Valence •First ionization energy	LOOCV	•Bandgap: RMSE of 0.094 eV •Band offset: RMSE of 0.183 eV	[388]

			regression (SVR)-radial kernel •ANN •Deep neural network (DNN)					
ORR	Dual-metal-site catalysts(DMSC)	•Adsorption free energy of OH* (ΔG_{OH})	•Gradient Boosted Regression (GBR)	•DFT computation	•Electron affinity •Electronegativity •Sum of the van der Waals (vdW) radius of the two transition-metal atoms. •Absolute value of the differen	Holdout	•ΔG_{OH}: RMSE and R^2 of 0.036 eV and 0.993, respectively.	[374]

				ce between the two transition-metal atoms •Sum of the Pauling negativity of the two transition-metal atoms.			
Binary and ternary nanocatalysts: PtCu, PtNi, CuNi, PtCuNi	•Energy contribution of atom i. (E_i)	•Neural Network Potential (NNP) with Monte Carlo •Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorith	•DFT computation •Molecular dynamics simulations	•Gaussian descriptor on the symmetry functions of radial (G^2) •Gaussian descriptor on the symmet	Holdout	E_i: RMSE of 0.007 eV.	[389]

		m		ry functions of angular (G^4)			
Bimetallic Pt core-shell nanocatalysts	•Strained coordination number (cn*(*j*))	•KRR	•DFT computation •EMT Calculations	•Coordination number •Generalized coordination number •Partial distribution function •Distance to alloy atoms •Interatomic distance from Vegard's law	Holdout	cn*: MAE of 0.07%	[19]
Titanium alloys: $TiAl_2O_5$	•Kohn–Sham density functio	•Behler−Parrinello neural	•DFT computation	Behler−Parrinello descriptors[373]	Holdout	E_{BPNN}: RMSE of 0.017 eV	[390]

		nal theory energy of $TiAl_2O_5$ structures per atom (E_{BPNN}) •Kohn–Sham density functional theory force of $TiAl_2O_5$ structures (F_{BPNN})	network s (BPNNs)			F_{BPNN}: RMSE of 0.27 eV/A	
CRR	Intermetallics	•CO and H adsorption energy on active sites	•RF regression •PCA •t-SNE	•Materials Project •DFT computation	•Atomic number, •Coordinated number •Electronegativity •Adsorpt	RMSE, MAE and MAD of 0.46 eV, 0.29 eV and 0.17 eV, respecti	[13]

				ion energy		vely.	
Bimetallic: Ni, NiGa, Ni_3Ga, Ni_5Ga_3	•CO adsorption energy on active sites	•Neural Network Potential (NNP)	•Materials Project •DFT computation	•Adsorption energy relative to the unrelaxed slab •The gas-phase CO energy	Holdout	RMSE is about 0.2 eV	[20]
Bimetallic or multimetallic: Cu-Al alloy	•CO adsorption energy on active sites (ΔE_{CO})	•Random Forest (RF) •t-SNE	•Materials Project •DFT computation	•Atomic number, •Coordinated number •Electronegativity •Adsorption energy	5-fold CV	ΔE_{CO}: Both MAE and MAD are 0.1 eV.	[18]
Bimetallic or multimetallic: (100)-terminated Cu	•ΔE_{CO}	•ANN	•DFT computation	•Filling (f) of d-band •Center (ε_d) of	10-fold CV	•ΔE_{CO}: RMSE is 0.13 eV.	[379]

	multimetallic alloys			d-band •Width (w_d) of d-band •Skewness (γ_1) d-band •Kurtosis (γ_2) of d-band, •Local Pauling electronegativity (χ_l)			
	High-entropy alloys CoCuGaNiZn and AgAuCuPdPt	•ΔE_{CO} •ΔE_H	•Gaussian process regression (GPR)	•DFT computation	•CO and H adsorption energy on local atomatic environment	5-fold CV	•ΔE_{CO}: MAE is 0.046 eV. •ΔE_H: MAE is 0.048 eV	[391]
NRR	IrO_2, MoS_2	•Free energy of all possible adsorba	•GPR	•DFT computation	•Surface coverage configurations	Holdout	ΔG: MAE of 0.57 eV	[392]

		te coverages						
	NRR electrocatalytic electrode	•Total current density ($\|i_{total}\|$) •Faradaic efficiency (F.E)	•ANN	•Self-generation	•Overpotential •Electrode morphorlogy •The kinetic predisposition of NRR	3-fold CV	$\|i_{total}\|$: RMSE, and R^2 of 2.6 mA/cm^2 and 1, respectively. *F.E.*: RMSE, and R^2 of 1.676% and 0.998, respectively.	[393]
Thermoelectricity	$Ba(MgX)_2$, (X = P, As, Bi), X_2YZ_6 (X = K, Rb, Y=Pd, Pt, Z = Cl, Br), K_2PtX_2 (X = S, Se), $NbCu_3X_4$ (X = S, Se,	•Types of Seebeck factors (S) •Types of Power factors (σS^2)	•Gradient boosting decision tree (GBDT) •DT •RF •kNN •ANN	•JARVIS-DFT •BoltzTrap calculations	•CFID descriptors •Chemical descriptor •Radial distribution function	3-fold CV Holdout	•Types of S: AUC of 0.96 •Types of σS^2: AUC of 0.82	[394]

Te), Sr_2XYO_6 (X = Ta, Zn, Y=Ga, Mo), $TaCu_3X_4$ (X = S, Se, Te), and XYN (X = Ti, Zr, Y=Cl, Br).				• Angle-distribution up to first neighbourDihedral angle distribution			
100 single crystal inorganic materials	• The lattice thermal conductivity (k_l)	• GPR	• MP	• Bulk modulus • Space group number • Maximum atomic radius • Volume per atom	5-fold CV	log(k_l): RMSE, and R^2 is 0.118 W/mK and 0.927, respectively	[395]
Off-stoichiometric samples (namely, Al23.5+xFe36.5Si40–x) of the	• σS^2	• GPR	• Experiment measurements	• Al/Si Ratio • Temperature	Holdout	σS^2: R^2 is 0.99	[396]

	Al2Fe3Si3 compound							
Piezoelectricity	Pb-free $BaTiO_3$	•Electro strains	•boosting gradient method •	•Experiment measurements	•Electronegativity •Ionic radius, volume •Ionic displacements •Polarization and •Dopant effects on transition	LOO-Bootstrap	RMSE is 0.013%	[37]
	$(Ba_{0.50}Ca_{0.50})TiO_3$-$Ba(Ti_{0.70}Zr_{0.30}Sn_{0.30})O_3$	•Morphotropic phase boundary	•Bayesian learning •SVR radial bias function •SVR linear regression •LR	•Experimental measurements	•Unit cell volume difference •The ratio of average ionic radii •Ionic displacements	LOOCV Bootstrapping	R^2 is 0.8789	[397]

					•The ratio of the effective nuclear charge •Ratio of electronegativities			
Rechargeable Alkali-Ion Battery	Li containing crystalline solids	•Is it a superionic material	•Logistic regression	•MP •ICSD	•The average number of Li neighbors for each Li •The average sublattice bond ionicity •The average anion coordination number in the anion	LOOCV	Accuracy is 90%	[398]

				framework •The average shortest Li–anion and Li-Li distance in angstroms			
Li metal anode	•Shear moduli •Bulk moduli •Elastic constants C_{11} •Elastic constants C_{12} •Elastic constants C_{44}	•Graph convolutional neural network •GBR •KRR	•DFT computation •MP	•Crystal structure •Mass density •Ratio of bond iconicity between Li and sublattice •Sublattice electronegativity	Nested 3-fold CV	•Shear moduli: RMSE is 0.1268 log(GPa) in log of Shear moduli •Bulk moduli: RMSE is 0.1013 log(GPa) in	[399]

				•Volume per atom		log of Shear moduli •C_{11}: R^2 is 0.60 •C_{12}: R^2 is 0.79 •C_{44}: R^2 is 0.60	
Electrolyte solvents	•Coordination energy (E_{coord})	•MLR •LASSO •Exhaustive search with linear regression	•KISHIDA Chemical Database •Experiment measurements	•Ionic radius •NBO charge of O atom •Atomic weight •Bolling point of solvent •HOMO •LUMO	10-fold CV	E_{coord}: CV error is 0.016 eV	[400]
Silicate-based cathodes with the composition of Li–Si–(Mn, Fe, Co)–O.	•Types of crystal system	•ANN •SVM •*k*-NN •RF	•MP	•Formation energy •Energy above the hull •Bandgap •Number	Monte Carlo Validation	Accuracy is 76%	[401]

					of sites			
	Electrode materials for metal-ion batteries	•Electrode voltage	•ANN •SVM •KRR	•MP	•Working ion in the battery •The concentration of the active metal ion in a given compound crystal lattice types •Space group numbers.	10-fold CV	MAE and R^2 is 0.44 V and 0.86, respectively	[402]
Supercapactior	Carbon-based electrodes	•Capacitance	•LR •LASSO •ANN	•Published literature	•pore size, •I_D/I_G, •Specific surface area •N-doping level	Holdout	R^2 is 0.91	[403]

Environmental Deconta mination	TiO2	•Degradation rate	•MLR	•Experiment measurements	•Bond Lipophilicity •Dipole •Bond dipole •Bond molar refractivity	LOOCV	RMSE and R^2 is 0.1073 and 0.9625, respectively	[14]
Flexible Electronics	Ag/polyamic acid (Ag/PAA) composites	•Sheet resistance •Processing time	•ANN	•Experiment measurements	•Concentration of PAA •Concentration of NaBH4, • Reduction time of NaBH4, •The ion exchange time of AgNO3.	Holdout	Relative error is less than 1.96%	[16]
Optoelectronics	2D octahedral	•Bandgap	•GBR •PCA	•DFT comput	•Distorted stacked	10-fold	MAE and R^2 is 0086	[15]

	oxyhalides			ation	octahedral factors	CV	and 0.835, respectively	
Superconductors	12,000+ known superconductors	•Critical temperature (T_C)	•RF	•SuperCon •ICSD	•Stoichiometric descriptors •Elemental property statistics •Electronic structure descriptors •Ionic compound descriptors	Holdout	Classification: Accuracy is 92% Regression: R^2 is 0.88	[277]
	Superconductors in the SuperCon data set	•T_C	•DNN	•SuperCon •COD •Published literatu	•Composition of materials	Holdout	Classification: Accuracy is 95% Regress	[404]

			re				ion: R^2 is 0.92	
Metallic glasses	Ternary amorphous alloys	•Volume n per atom •Fromation energy •Bandgap energy	•RF •LR •Rotation Forest •Reduced-errpr Pruning Tree	•Noneq uilibriu m Phase Diagra m of Ternar y Amorp hous Alloys	•Stoichio metric descript ors •Element al propert y statistic s •Electro nic structur e descript ors •Ionic compou nd descript ors	10-fold CV	MAEs for volume n per atom, fromati on energy, and bandgap energy are 0.653 $Å^3$, 0.0882 eV and 0.0645 eV, respecti vely. R^2 are above 0.91 for the predicti ons of all three properti es.	[33]

	Bulk metallic glasses	•The existence ability in an amorphous state: glass-forming ability (GFA) •The critical casting diameter (D_{max}) •The supercooled liquid range (ΔT_x)	•RF	•Experiment measurements	•Composition of materials	10-fold CV	GFA: Accuracy, FPR, TPR are 89%, 7% and 90%, respectively. D_{max}: R^2 and MAE are 0.8 and 0.21 nm ΔT_x: MAE is 8.8K	[405]
Magnetic materials	Ferromagnetic materials and antiferromagnetic	•Curie temperature •Magnetic ground state	•RF	•Atom Work	•Magpie descriptors •SOAP •Space group number	10-fold CV	Curie temperature: MAE and R^2 are 55 K and 0.91,	[406]

						respectively. Magnetic ground state: Accuracy and F1 score are 87.3% and 91%, respectively.	
Permanent magnets	•Uniaxial magneto-crystalline anisotropy constant (K_1) •The magnetization ($\mu_0 M$) •The relative phase	•SVR •Decision Tree	•Published literature	•Crystal configuration	10-fold CV	K_1: Pearson correlation coefficient and MAE are 0.9 and 3.902, respectively. $\mu_0 M$: Pearson correlation coeffici	[407]

		stability energy (E_f)					ent and MAE are 0.95 and 0.089, respectively. E_f: Pearson correlation coefficient and MAE are 0.95 and 0.043, respectively.
Soft magnetic materials	•Magnetic saturation •Coercivity •Magnetostriction	•GBDT •LR, •SVM, •DTs, •*k*-NN, •RF	•Published literature	•Annealing temperature •Annealing Time •Primary Crystallization Onset •Primary	20-fold CV	R^2 for Magnetic saturation, coercivity and magnetostriction are 0.86, 0.82 and	[408]

Crystallization Peak	0.76, respectively.

6.1. Energy Conversion

The data-driven empowered material discovery process provides novel opportunities to bring about breakthroughs in the energy conversion field. Innovations in materials for energy conversion are both essential and urgent as a basis for solutions to a large number of challenges, from improving energy efficiency to carbon-neutral electricity generation.[9, 36, 381] In this section, we introduce the application of data-driven material design approaches in the field of energy conversion including water-splitting,[10-12] photovoltaics,[17, 36, 71] fuel cells,[374, 389] carbon dioxide reduction,[13, 18] nitrogen reduction,[392] thermoelectricity[394, 395], and pizezoelectricity[37, 397].

6.1.1. Water-Splitting

The most prevalent data-driven application for water-splitting is in facilitating the development of novel catalytic and cathode materials. Water-splitting is one of the most promising methods to produce hydrogen,[409] an ultimate clean energy resource. The water-splitting reaction ($H_2O \rightarrow H_2+1/2O_2$) includes two half-reactions: the hydrogen evolution reaction (HER); the oxygen evolution reaction (OER). In this subsection, applications of ML techniques and related data-intensive strategies for the discovery of high-performance catalyst materials for HER,[10, 11] OER[12, 385] and overall water splitting (OWS)[381] are investigated.

Hydrogen Evolution Reaction (HER)

Data-driven approaches can promote the discovery of high-performance catalysts for efficient HER based on in-depth investigations of the structure-activity relations. Highly active noble-metal-free catalysts play a vital role in high-efficiency HERs. In recent decades, various candidates such as NiP_2[410, 411]and MoS_2[412, 413] have been proposed. For instance, Wexler et al.[10] applied data-driven approaches to study the influence on HER activity of nonmetal dopants on the Ni_3P_2 termination of $Ni_2P(0001)$ surfaces. A regularized random forest (RRFs) ML model with 3-fold CV was applied to predict the adsorption free energy (ΔG_H) of 55 training structures with 0.09 eV RMSE (Figure 22a).

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/adma.202104113.

The binding energy of H to the surface of Ni_3-hollow is too strong for HER; thus, the surface P sites were replaced by nine nonmetal elements, including As, B, C, N, O, S, Se, Si, and Te, to tune the binding energy of H; the number of dopants (n_X) varies from 1 to 6. The impact of the applied mechanical and chemical pressure (via nonmetal doping) on the ΔG_H were also explored (Figure 22b). The descriptor space was constructed with the descriptors compiled from DFT-relaxed structures, including the geometric structure parameters, the length of Ni-Ni bond, the angle of the Ni-Ni-Ni bond, Löwdin charges, elemental data (mass number, atomic weight, and atomic radius), summary statistics (the mean and standard deviation of these descriptors), and other geometric structure parameters (the area and perimeter of the Ni_3 hollow site). Based on the architecture of RRFs, it was found that the most related descriptors is the particular bond length of Ni-Ni, the constituent atoms of which are differentiated by their distance to the first doping site (Figure 22c). Moreover, as shown in Figure 22c, seven out of the 10 of the most essential descriptors are geometrically related to the Ni_3-hollow adsorption site, indicating that the chemical pressure is the key driving force in altering the catalytic activity of HER. To confirm this insight obtained by ML model, further simulations investigating the effect of mechanical pressure executed, the results of which suggested that the optimal average bond length of Ni-Ni lies between 2.97 and 3.07 Å in a Ni_3 motif for electrocatalytic HER to enable ideal thermoneutral H adsorption. Optimal bond length could be set as a criterion for high-throughput screening of binary Ni-nonmetal bulk materials. In addition, the mixed doping of nonmetals could be further explored. Alternatively, the enhancement of the HER electrocatalytic activity through applied chemical pressure could be investigated for other doped and undoped TM phosphides, such as Fe_2P and Co_2P, which have a bulk structure similar to that of Ni_2P. The trained ML model in this study could accurately (0.09 eV RMSE) predict the ΔG_H; the results of this analysis strongly support the idea that the modified local geometry of the Ni_3-hollow adsorption site significantly enhances the HER activity of $Ni_3P_2(0001)$. Further experimental and

theoretical investigation of nonmetal doping on transition metal phosphides could also be conducted based on the insights rendered by ML model to identify high-performance HER catalysts.

The catalytic performance of amorphous Ni_2P has also been studied by Zhang et al.[11]; a data-driven model was applied to facilitate the high-throughput prediction of adsorption energy of H (E_H). A genetic algorithm (GA) was deployed as the global optimization method to search for favorable amorphous configurations. Two ML algorithms (feed-forward neural network (NN) and gradient boosting decision tree (GBDT)) were trained to predict the E_H. However, since common descriptors, for the description of crystal structure materials (such as d-band and electronegativity) are not suitable for amorphous materials, it is challenging to characterize their chemical environment. The authors decomposed this problem by dividing the prediction of E_H into two parts: the binding energy of H at the adsorption site i.e., the frozen adsorption energy (E_{frozen}); the relaxed energy (E_{relax}) produced by H adsorption. The ML model based on the GBDT algorithm used the hand-craft descriptors derived from the pristine surface of Ni_2P. Within a set cutoff radius (R_c) (Figure 22d), the three closest P and Ni atoms (P_{1-3} and Ni_{1-3}) to the Ni adsorption site (Ni_{ads}) for H adsorption were specified. The distance of the bond between each pair of the seven atoms was taken as the primary descriptor. Based on Pearson coefficient correlation analysis, it was suggested that $Ni_{ads}P_3$ possess a high negative inner relation with E_{frozen} (Figure 22e). The prediction of E_{frozen} achieved an RMSE of 0.06 eV and 0.11 eV for the training and testing sets, respectively (Figure 22f). Thus the P content has a significant influence on the reactivity on the Ni_{ads} site. The first model only focuses on the effects of the bond distance, which does not account for the remaining chemical environment. Hence, in the second model, the atom-centered symmetry function method was employed to form descriptors accounting for chemical environment to successfully fit the potential energy surface (PES). An R_c was also set to explore the adsorption site's local chemical environment (Ni_1) (Figure 22g). The bond length of the five closest atoms to Ni_1 represented by the symmetry functions in the R_c- local environment was constructed as the descriptor space. Here, a high-dimension neural

network composed of five sub-neural networks (correspond to the five atoms) was implemented to predict each atom's contribution (E_i) to the E_{frozen} (Figure 22h). For the second model, the RMSE of of E_{frozen} prediction achieve values of 0.05 eV and 0.10 eV for the training and testing sets, respectively. In brief, the NN-based model possesses several advantages. It is more accurate than the first model and can be applied more generally to situations of similar or greater complexity based on the NN's favorable extensibility. With the implementation of the symmetry function method, the NN model made full use of geometric information. The NN model also provided novel insights into the identification of adsorption patterns (hollow-site, bridge, and top) via embedding the high dimension descriptor vector into a low dimension vector. This work generated a ML model that could accurately predict the adsorption energy on a specific site of amorphous NiP_x materials for HER based on local information about the chemical environment. With the employment of the trained ML model, 40 optimal active sites on amorphous Ni_2P were screened out and divided into five main patterns. Rational analysis of the configuration of such patterns enables the control of catalytic activity via modifications of the surface atom configuration.

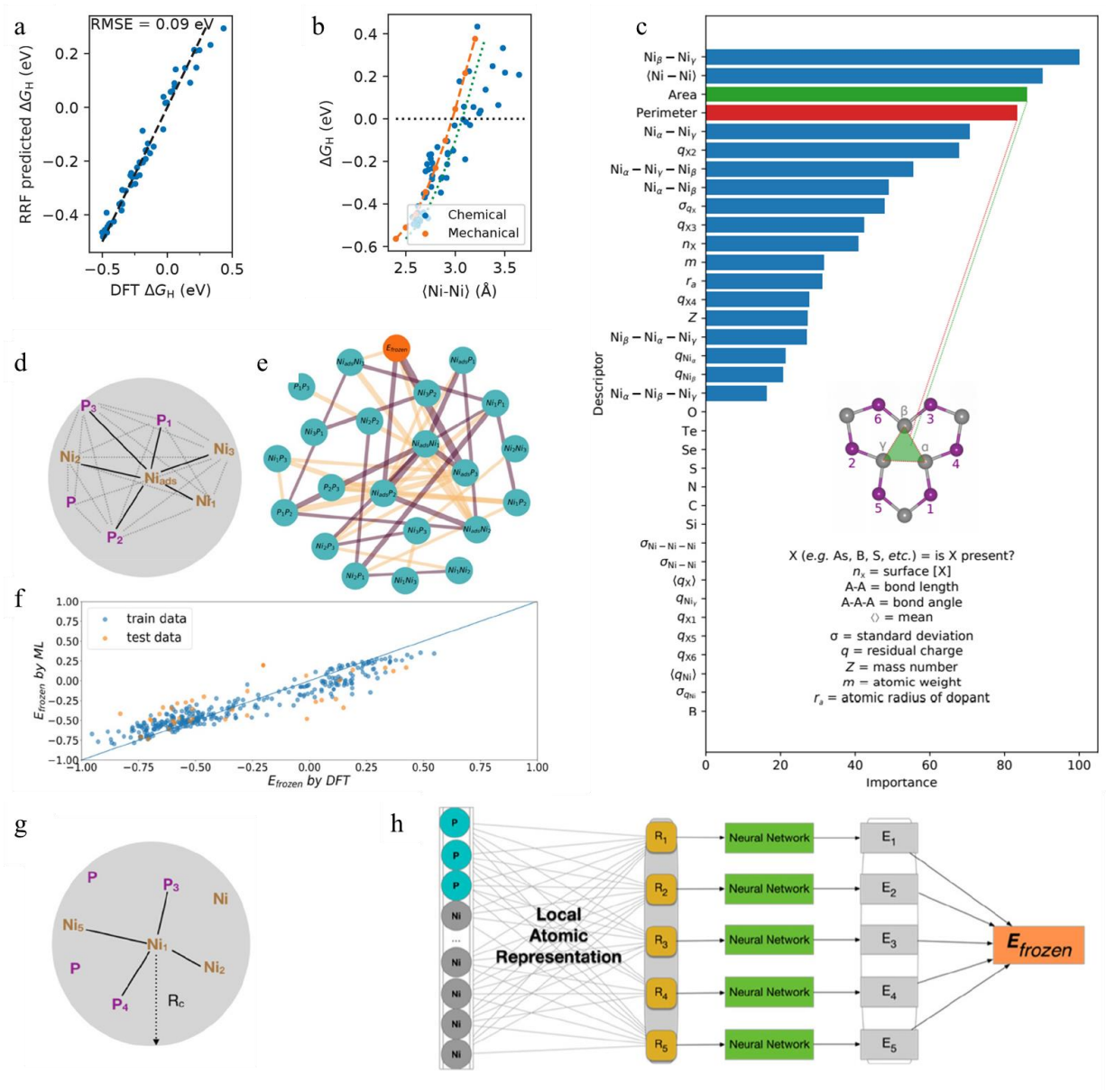


Figure 22. a) The prediction of ΔG_H by the RRFs compare to the DFT computation results. The black dashed line indicates the perfect agreement. b) The variation of ΔG_H under the change of average Ni-Ni bond length induced by chemical and mechanical pressure. c) The importance ranking of descriptors derived from RRF model. a-c) Reproduced with permission.[10] Copyright 2018, ACS Publications. d) The local environment of the adsorption sites (Ni_{ads}). Ni_{1-3} and P_{1-3} denote the first, second and third closest Ni and P atom to Ni_{ads}. The solid line represents the real chemical bonds, and the dashed line represents the distance of atom pairs. e) The graph type visualization of Pearson's correlation, where the thickness of lines represents the importance of the descriptors. The

red colour indicates a positive correlation and green for a negative relationship. f) The performance of GBDT on the prediction of E_{frozen}. g) The local environment of a sphere with a radius R_c centred by the adsorption sites (Ni_1). (H) The scheme of the atomic neural network. d-g) Reproduced with permission.[11] Copyright 2020, ACS Publications.

To conclude, the majority of data-driven HER applications have been implemented to predict the activity of HER catalysts. The two aforementioned studies shed light on the importance of descriptors in the data-driven ML process. Descriptors involved in these processes can be divided into geometric and electronic structure descriptors.[379, 414-417] The electronic structure descriptors are usually computationally-expensive to obtain via the quantum-chemical simulation and are relatively unsuitable for other applications. It is therefore advantageous that geometric descriptors demonstrate higher impact on the adsorption energy of catalyst active sites.

Oxygen Evolution Reaction (OER)

The majority of data-driven applications in OER are focus on the prediction of adsorption enthalpy. The adsorption enthalpies of transition metals (TMs) and their alloys can be successfully described by d-band theory[418] and related scaling relations.[419] However, this simplified correlation does not apply to other types of materials such as the TM oxides and, therefore, a theoretical study of numerous crystal structures, surfaces, and active sites of metal oxide materials has been reported by Back et al. (**Figure 23**a).[12] They systematically investigated the catalytic activity of Ir-containing oxides, one of the state-of-the-art OER catalytic materials. They employed a customized CNN model[158] (with crystal graphs including atomic and bonding information as an input) to predict binding free energies (ΔG). The crystal structures of IrO_2 and IrO_3 were taken from the Materials Project; two adopted structures of TiO_2 polymorphs taken as surrogates for the structure of IrO_2. The authors discovered that, in IrO_2, the predicted overpotentials of all the active sites on the (121)

surface are lower than that on the (110) and (100) surfaces. Additionally, less stable low-index surfaces such as (100) and two unique terminations of the (111) surface possess higher OER activity than (110). The study achieved a test error of 0.10 eV (RMSE) and 0.07 eV (MAE) for 300 coverage calculation training data points (Figure 23b), and 0.18 eV (RMSE) and 0.13 eV (MAE) for OER calculations (Figure 23c). Their DFT results indicated that several sites on the low-index surface possess a higher activity than those on the rutile (110) surface. An increase in the number of active, high-index surfaces could be achieved by decreasing the size of Ir oxide nanoparticle. Moreover, surface Ir atoms with higher oxidation tend to be more active for OER. A graph-based predictive ML model of surface coverage and site activity of IrO_2 and IrO_3 has been presented as a reasonably accurate substitution for DFT computations. Such an integrated ML and DFT framework is a promising approach to reducing computational cost and achieving high-throughput screening of OER catalysts.

Xu et al. [385] employed the compressed sensing method, called the SISSO,[26, 420, 421] to identify suitable activity descriptors for the prediction of OER adsorption enthalpies at doped RuO_2 and IrO_2. The validation set of OER adsorption enthalpies was computed by DFT computations (Quantum ESPRESSO[422] and BEEF-vdW functional[423]) with an uncertainty of approximately 0.5 V in the theoretical overpotentials (Figure 23d). The SISSO descriptors significantly surpass previous single descriptors in terms of both computational cost and accuracy. For primary descriptors, the width of the d-band (W_d) and charge transfer energy (CTE) demonstrate the highest Pearson correlation coefficient (0.744 and 0.734, respectively) with the simulated adsorption enthalpies, indicating that the adsorption enthalpies are unlikely to be characterized by a linear relationship of only one primary descriptor. The primary descriptors are then constructed by employing a series of algebraic/functional operators to the primary descriptor set. The operators are applied to generate descriptor space, where the iteration number (N) is used as a hypermeter and denoted as the rung (Φ_N). After the first iteration of the descriptor construction (Φ_1), the top five one-dimension SISSO

descriptors, such as the difference between the width of the d-band and the centre of the O_{2p}-band (W_d-ε_{O2p}), show a higher correlation coefficient than the best primary descriptors mentioned earlier. The combination of a number of low primary features could produce a descriptor with a higher Pearson correlation coefficient. An even higher correlation coefficient could be achieved by increasing number of iterations employed for the descriptor construction. As presented in Figure 23e, the remaining average uncertainty of SISSO-derived overpotential is approximately 0.2 V. The SISSO model provides explicit algebraic expressions for predicting adsorption enthalpies with the use of highly correlated descriptors. The ML results indicate that Co and Fe would be the ideal dopants to improve the activity of OER reactions, which is an observation that also agrees with the experimental data. The SISSO provides solutions for the generation of high-performance composite descriptors for ML training shows great promise other applications such as the ORR, HER, and CRR, characterized by both higher prediction accuracy and low computation cost.

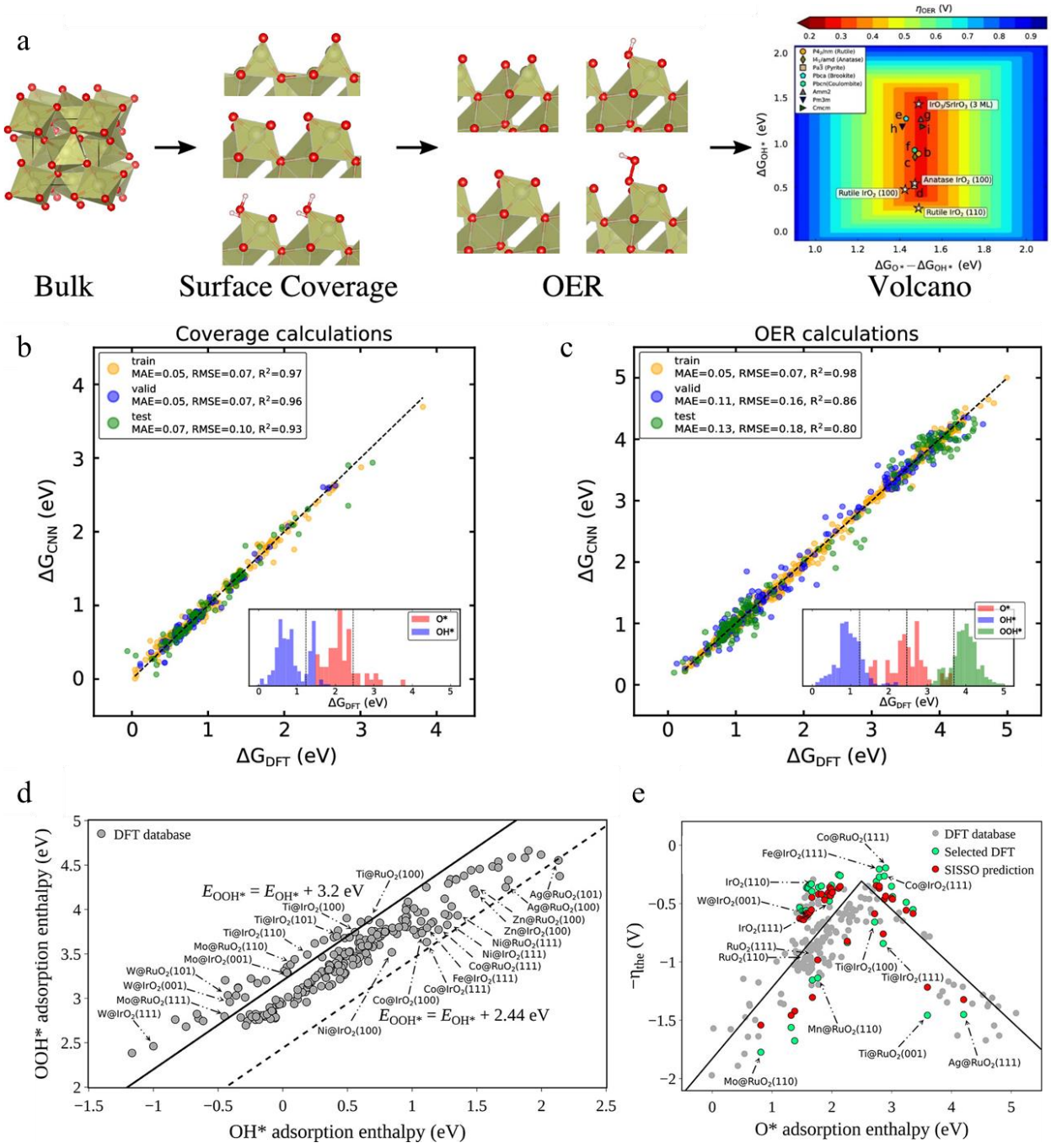


Figure 23. a) Scheme of automated DFT analysis for the performance of catalysis on OER. b) The prediction results of CNN model on the value of ΔG for coverage calculations. c) The prediction results of CNN model on the value of ΔG for OER calculations. a-c) Reproduced with permission.[12] Copyright 2019, ACS Publications. d) The relationship between the adsorption energy of OOH* and the adsorption of OH*. The solid line is based on the author's DFT database, and the dashed line represents the ideal scaling relationships.[424] e) The volcano plot of the overpotential as a function of

the adsorption enthalpies of O*. The green points represent the selected DFT computation results, and the red points are the corresponded SISSO predictions. d-e) Reproduced with permission.[385] Copyright 2020, ACS Publications.

For the OER catalysts, studying the activity of oxide catalysts is more complicated relative to that of transition-metal catalysts due to their abundance of facets and surface coverages. For the accurate predation of adsorption energy on the adsorption sites via data-driven approaches, the selection and generation of proper descriptors is critical. Although individual geometric and electronic primary features show relatively poor correlation, more related descriptors can be constructed by utilizing systematic methods such as the SISSO and symmetry functions.

Overall Water-Splitting

Researchers have focused on the use of ML to identify innovative bifunctional catalysts demonstrating high activity for simultaneous HER and OER reactions which constitute the overall electrocatalytic water splitting process. Ge et al.[381] implement the LASSO regression algorithm with quantum mechanics to generate predictive models to identify innovative structures with superior electrocatalytic performance towards both HER and OER. In this work, two-dimensional (2D) van der Waals heterostructure materials based on transition metal dichalcogenides (TMDCs) were studied (**Figure 24**a and 24b). The authors demonstrated that a significant improvement in performance could be achieved with the combination of two independent TMDCs such that the descriptor space was constructed from the cosine of the rotational angle (θ) (Figure 24c and 24d), the distance (d) between two secondary parts, the average MX_2 (monolayered TMDC) bond length (l) and the bandgap ratio (λ) of the two components. The value of ΔG_{*OOH} was used to compute other free energy variations to describe the reaction performance (Figure 24e and 24f). As the catalytic activities of the MX_2 monolayer towards HER and OER are insufficient, the heterojunction system of

two MX_2 layers has been proposed. The LASSO algorithm combines the four descriptors into a complex LASSO performance descriptor (PL), which exhibits a favorable linear relationship with the overpotentials of HER (η_{HER}) and OER(η_{OER}) (Figure 24g). In this study, the LASSO approach was used to determine performance descriptors combining four variables, and an equation was applied to predict the bifunctional catalytic performance of rotated TMDCs for both HER and OER without resorting to expensive computations or experiments. Among all the computed structures, the most favorable heterojunction system consisted of $MoTe_2$ and WTe_2 monolayers with a 300° rotation angle; η_{HER} and η_{OER} achieved values of 0.03 V and 0.17 V, respectively.

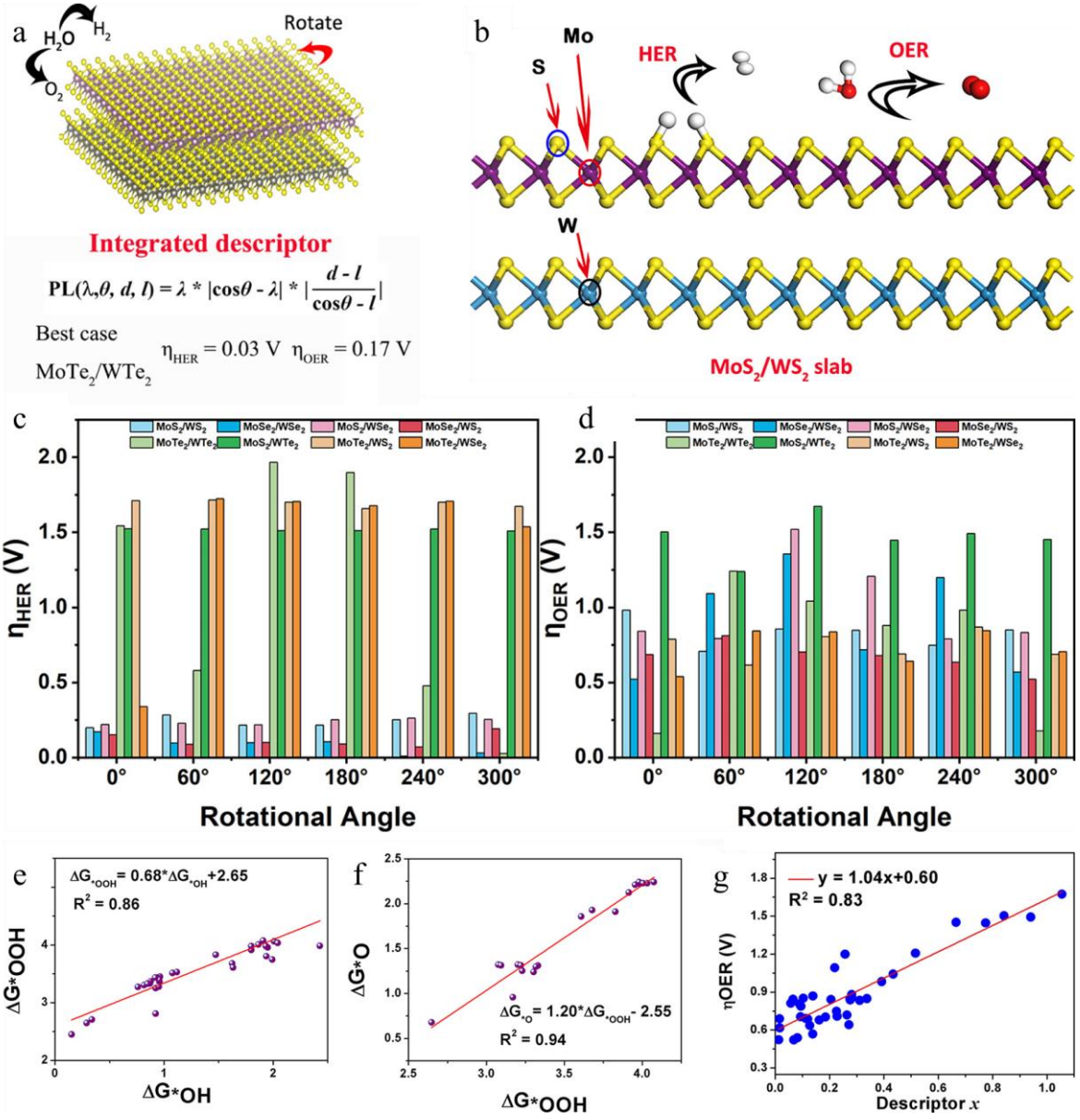


Figure 24. a) The illustration of the bifunctional electrocatalysts for HER and OER. b) Scheme of the heterostructure with 0° rotation angles for the catalyst. The atom of Mo, S and W are represented by purple, yellow and blue, respectively. c) The relationship between the rotation angle (θ) and the HER overpotentials. d) The relationship between the rotation angle and the OER overpotentials. e) The linear correlation relationship between ΔG_{*OH} and ΔG_{*OOH}. f) The linear correlation relationship between ΔG_{OOH*} and ΔG_{O*}. g) The relationship between the OER overpotential and the PL descriptor. a-g) Reproduced with permission.[381] Copyright 2020, ACS Publications.

6.1.2. Photovoltaics (PV)

Data-driven research in the PV field can facilitate a faster and more efficient route for the discovery of candidate materials. Sunlight is the most abundant clean and renewable energy source; in the field of energy conversion, PV effects provide the opportunity to effectively utilize solar energy.[425] The direct conversion of solar energy to electricity via PV cells, or thermal energy in integrated solar energy systems, has become the dominant approach to future green energy generation.[426] It is driven by the pursuit of innovative, high-performance PV materials and the optimization of deposition approaches for solar cell applications. In this sub-section, data-driven methods for the discovery of various PV materials, including perovskite-based,[36, 427] organic,[49, 71] metal oxide,[54] and other novels PVs,[17, 387] are reviewed.

Perovskite-Based PV

Perovskites-based solar cells have improved in efficiency in recent years; numerous data-driven studies have been conducted for the deeper investigation and discovery of PV candidate materials.[427] Lu et al. [36] developed a data-driven framework combining quantum-chemical simulation and ML to discover innovative and novel HOIPs for PV. HOIPs are one of the most favourable material classes for PV and have garnered tremendous interest. The essential characteristics of HOIPs include the high-performance power conversion efficiency (PCE), competitive experimental synthesis cost, and tunable bandgaps.[428-431] The authors specifically focused on discovering the stable Pb-free HOIPs with by employing of six different ML algorithms. As shown in Figure 17b, the data-driven framework consists of four critical processes: the construction of the descriptor space, training, application of the ML model, and validation of thermal and

environmental stability via DFT computations. The material space consisted of 212 HOIPs with the orthorhombic-like crystal structures completely belonging to the perovskites family with ABX_3 stoichiometry, where A^+ denotes monovalent or organic molecular cations, B^{2+} denotes divalent metal cations, X represents c atoms. Thirty initial descriptors pertaining to the 212 HOIPs were processed by the Perdew-Burke-Ernzerh (PBE) functional[432] for further feature engineering. The importance of these descriptors to the target properties, such as bandgap (E_g^{PBE}), was evaluated via the gradient boosting regression (GBR) and 'last-place elimination' methods. The descriptors with less impact on the bandgap were excluded to ultimately arrive at 14 of the most important descriptors. The geometric structure descriptors, tolerance factor (T_f) and octahedral factor (O_f), were ranked as the first and third most important parameters to the bandgaps, respectively; the total number of ionic charges for B-sites (IC_B) was the second most important descriptor (**Figure 25**a). According to Figure 25b, the Pearson correlation coefficients' heat map indicates low inner correlations between pairs of descriptors, suggesting that redundant descriptors have been successfully removed. The optimal descriptor set was employed to train of six different supervised ML regression models: GBR, kernel ridge regression (KRR),[433] decision tree regression (DTR), SVR,[434] Gaussian process regression (GPR), and multilayer perceptron regression. Three metrics were used to evaluate the model performance: R^2, MSE, and the Pearson coefficient (r), which indicates the correlation between the predicted value (E_g^{ML}) and the real value (E_g^{PBE}) of the bandgap. Further, 80% of the HOIP data was adopted as the training set, while the remaining was used for testing based on the hold-out method. The GBR model demonstrated superior values of R^2 and RMSE of 97% and 0.086, respectively. The bandgap of 5,504 different HOIP candidates (32 A-site monovalent or organic molecular cations, 43 B-site divalent cations and 4 X-site halogen anions) were predicted. Further screening was conducted based on geometric stability, toxicity, difficulty of synthesis, and the bandgap value (which would ideally lie in the range of 0.9-1.6 eV for PVs). Finally, six novel orthorhombic HOIPs ($C_2H_5OInBr_3$, $C_2H_5OSnBr_3$, $C_2H_6NSnBr_3$, $C_2H_6NInBr_3$ $NH_3NH_2InBr_3$, and NH_4InBr_3)

were proposed as promising candidates for PVs. Further evaluation of thermal and environmental stability, and electronic properties of the six candidates has also been conducted via first-principle computation, indicating that $C_2H_5OInBr_3$, $C_2H_5OSnBr_3$ and $C_2H_6NSnBr_3$ demonstrate higher stability against water and oxygen. The bandgap values predicted by ML models are in significant agreement with those computed by DFT (RMSE less than 0.1 eV), which proves the superiority of the ML approach. Based on a combination of DFT and ML techniques, an efficient workflow to screen stable, lead-free HOIP candidates with suitable bandgaps is proposed. In contrast with traditional high-throughput screening processes, the recommended approach coincides with a lower computation cost as only the most promising HOIP candidates are characterized with DFT computations.

Organic Materials Based PV

Unlike HOIPs, organic photovoltaics (OPV) materials consist of conjugated molecules or polymers and have the advantages of low price, flexibility, light weight, and transparency.[48-51] Although a subset OPVs have achieved a PCE of 17.3%, [435] compared to other PV materials, most OPV PCE values are relatively low (usually around 10%). The Scharber model[436] is widely used to calculate OPV systems' PCE,[437, 438]though its accuracy is is questionable given that it only takes the frontier orbital energies into consideration.

Sahu et al.[71] constructed a dataset of 280 small molecule OPV systems and discovered that degeneration almost occurs on the frontier molecular orbitals of donor molecules for high-performance unit devices; in such a case, orbitals other than only the HOMO and LUMO should be taken into consideration. They proposed a data-driven framework for the prediction of PCE by training several different models such as linear regression (LR), k-nearest neighbour (kNN), ANNs, random forest (RF), and gradient boosting regression tree (GBRT). Quantum chemical properties used to build descriptor space. Simple topological descriptors are not fully representative of the energy conversion process in an OPV and do not produce favorable results when applied to the conjugated molecules database for solar cell applications.[439] The Pearson correlation coefficients

were calculated to ensure the mutual independence of each feature, and a descriptor with 13 features was constructed. Out of the 280 data points, 30 were adopted as the testing data, while employing LOOCV. With 10-fold CV over 250 data, the GB-based ML model yielded excellent results with a Pearson coefficient of 0.79 and an RMSE of 1.07% towards PCE for the test data, and a Pearson coefficient of 0.76and 1.09% RMSE towards PCE on all data points with LOOCV (Figure 25c). Within the GB and RF models, the hole–electron binding energy in donor molecules (E_{bind}) possessed the highest correlation coefficient to the PCE for both models (Figure 25d). However, as mentioned in the previous discussion, the evaluation of quantum chemical descriptors is usually computationally expensive. Though the predictive ML model in this study has not been employed to propose novel high-performance OPV materials, it has showed substantial potential in the preliminary high-throughput virtual screening of promising candidates and is able to capture the complexity of OPV devices to pinpoint the critical descriptors that affect the PCE. This ML application is helpful in understanding the operating mechanism and rational design of OPVs.

In a data-driven OPV research, Sahu et al.[49] used similar descriptor spaces; the polarizability of donor molecules was replaced by the number of hetero atoms (N_{het}^{D}) to reduce the computational cost; 300 newly reported small-molecular OPVs were investigated, and 250 data points were used as the training set. As illustrated in Figure 25e,

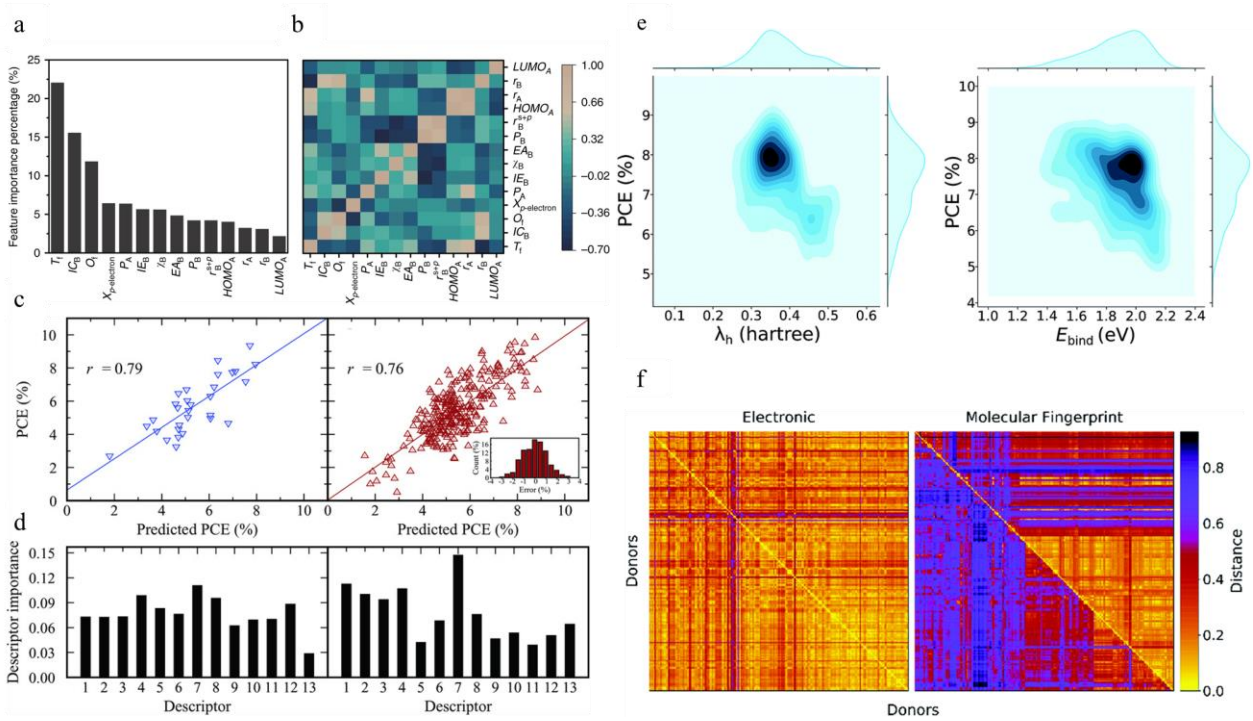


Figure 25. a) The selected descriptors and their importance. b) The Pearson correlation heat map of the selected 14 descriptors. a-b) Reproduced with permission.[36] Copyright 2018, Springer Nature Publications. c)The prediction results of GB model on the PCE versus the experimental PCE on the testing set (left) and the whole data set with LOOCV (right). The inset represents the error's probability density. d)The predicted importance of each descriptor on the GB (left) and RF (right) model. Descriptors are in following order: 1)N_{atom}^{D}, 2)polarizability, 3)the energetic difference of LUMO and LUMO+1, 4)the energetic difference of HOMO and HOMO-1, 5)IP(v), 6)λ_h, 7) E_{bind}, 8)E_{LL}^{DA}, 9)E_{HL}^{DA}, 10)the energy of the electronic transition to a singlet excited state under the largest oscillator strength, 11)the dipole moment change in going from the ground state to the first excited state on donor molecules, 12)the energy of the electronic transition to the lowest-lying triplet state, 13)the energetic difference of LUMO and LUMO+1 on acceptors. c-d) Reproduced with permission.[71] Copyright 2018, Wiley Publications. e) Joint distributions derived from kernel density estimation for vital descriptors. Reproduced with permission.[49] Copyright 2019, RSC Publications. f) The donor distance matrices for the donors in the data set The upper and bottom triangular for both matrices

are the descriptors computed by Daylight fingerprints and Morgan fingerprints, respectively. Reproduced with permission.[48] Copyright 2019, Elsevier Publications.

molecules with large reorganization energy for holes in donor molecules (λ_h) and E_{bind} were screened out effectively. Similar to their previous ML model training strategy and validation method,[71] the trained model based on GBRT and ANN demonstrated that donor units such as benzodithiophene, dithieno-benzodithiophene and naphtho-dithiophene, and acceptor units based on naphthobisthiadiazole and isoindigo possess great potential to construct high efficient materials for OPVs. An ML-based virtual screening of 10,170 molecules was performed; 126 molecules with PCE greater than 8% for both models were proposed. In addition, to predict accurate property values by ML, it is also of great significance to obtain implicit chemical knowledge to boost the design of molecules with excellent photo-physical properties for high-performance OPVs. Padula et al.[48] presented an ML workflow to identify efficient OPV materials based on chemical similarity. They acquired similarity metrics such as Euclidean distance[82] and Tanimoto similarity index[81] (Figure 25f) between donor pairs while considering structural and electronic parameters to verify the domain of applicability. ML models based on kernel ridge regression (KRR) and *k*-nearest neighbors (*k*-NN) algorithms were employed to verify the correlations between those metrics with target properties including PCE, J_{SC}, and V_{OC}. The best-performed KRR model yielded a Pearson's correlation coefficient around 0.7, which lends insights for a highly-efficient and reliable approach to virtually screen OPVs via control of chemical topology and electronic structure.

To sum up, a number of ML-based OPV studies have been conducted to screen out high PCE materials. The hole–electron binding energy in donor molecules (E_{bind}) is highly correlated with the PCE of OPVs. Well-trained ML models can enable the prediction of promising materials for both donors and acceptors as a basis for constructing high-performance OPVs. ML-based approaches are also key to extracting chemical knowledge to develop new OPV design principles.

Metal Oxides PV

Metal oxides (MOs) meet most of the prerequisites of the solar cell materials, including high efficiency of electricity generation from solar power, low cost, stability over long periods, environmental friendliness, and ease of synthesis.[440] The employment of data-driven techniques could further improve the PCE of MO PVs. Yosipof et al.[54] have proposed a methodology (**Figure 26**a) for the application of data mining techniques and an ML algorithm to explore the relationship between the properties and performance of two MO-based solar cell libraries, $TiO_2|Cu_2O$ and $TiO_2|$Cu-O libraries, where Cu-O denotes that the CuO was the metal oxide applicable to the library preparation but multiple oxides were found in the cell. The three PV properties of photocurrent density of the short circuit (J_{SC}), photovoltage of the open circuit (V_{OC}), and the internal quantum efficiency (*IQE*) were used to indicate the performance of the solar power cell. The structure of the photovoltaic cell is shown in Figure 26b. The descriptor space consists of seven experimentally measured material descriptors including the thickness of the window (T_w) and absorber layer (T_a), the ratio between T_a and (T_a+T_w), bandgap of the absorber layer (*BGP*), the distance between the cell and the center of the deposition plume (D_{center}), the resistance of the absorber layer (R_a), and the maximum value of calculated theoretical photocurrent (J_{max}). The principal component analysis (PCA)[441] algorithm was applied to reduce the dimensionality of the descriptor space for the data visualization. Figures 26c and 26d present the value of the first and second principal component (PC), respectively, for each cell as a function of the position within the $TiO_2|$Cu-O library, where outlier cells are circled. Comparing the function plot of the Cu-O and TiO_2 layer thickness with the cell position within the $TiO_2|$Cu-O library (Figures 26e and 26f, respectively), shows that the patterns highly coincide with the PC

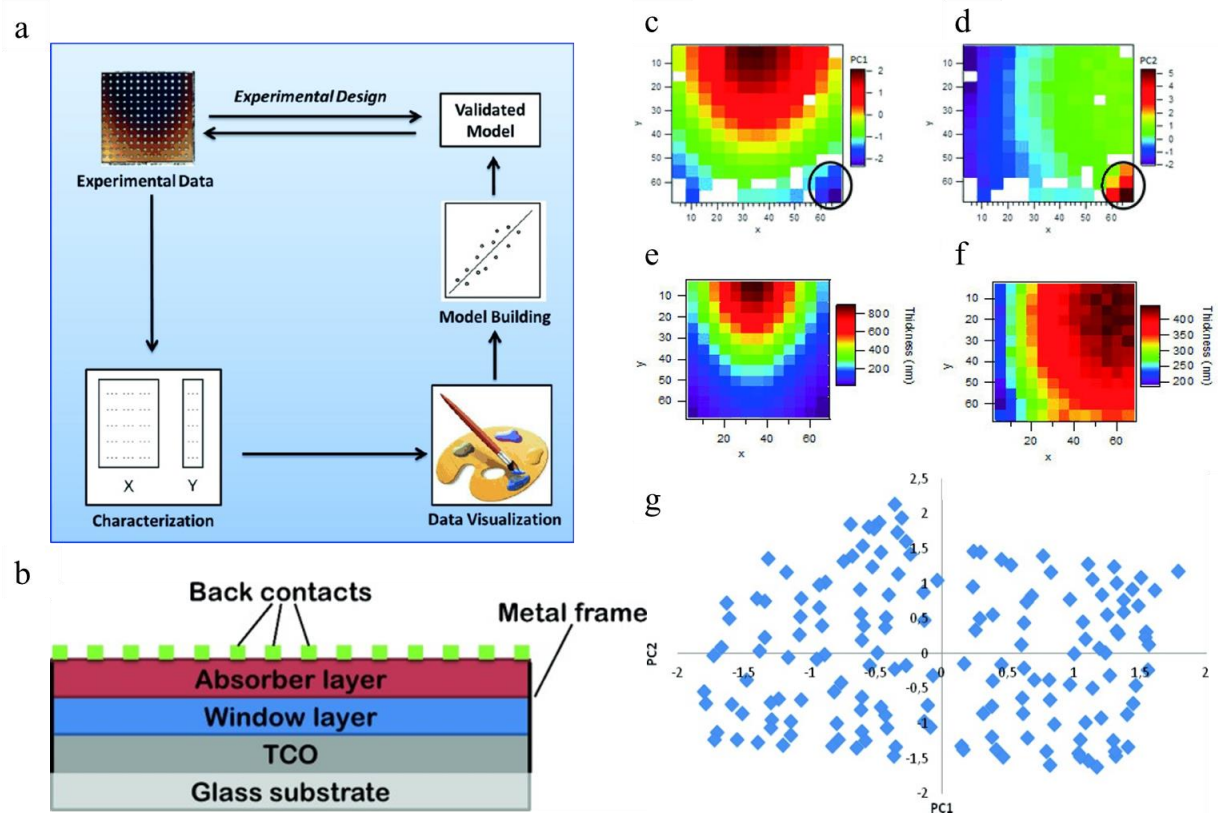


Figure 26. a) The flow diagram of the ML assisted model to explore the relation between properties and performance of two MO-based solar cell libraries. b) The scheme of a combinatorial metal oxide photovoltaic library. c) The value of PC1 and d) PC2 for each cell with the variation of cell position within the TiO2|Cu-O library, where the outlier cells are circled. e) The value of Cu-O and f) TiO_2 layer thickness for each cell with the variation of cell position within the TiO2|Cu-O library. g) Solar cell distribution in PC space. a-g) Reproduced with permission.[54] Copyright 2015, Wiley Publications.

plots. Figure 26g represents the solar cell distribution in the PC space, indicating that the cells are evenly distributed. *k*-NN and genetic programming (GP) algorithms were used for model generation to establish the relationship between the descriptors and activities indicated by J_{SC}, V_{OC}, and *IQE*, where LOOCV and standard CV were applied. The best *k*-NN model yields good prediction statics with an R^2 of 0.92, which suggests that T_a was found to be the only significant predictor of J_{SC} and *IQE*. Both T_a and T_w show comparable contributions to the prediction of V_{OC}. In this study, the PCA

model extracted the implicit chemical information from the data; the generated PC plot is helpful for sample clustering and outlier detection. The obtained correlations were in good agreement with those in a previously published work,[442] indicating the applicability and reliability of the developed ML model.

Other Novel PV

2D materials, with their unique chemical structures, have fascinating and novel properties. Therefore, it is critical to employ a data-driven process to better understand structure-property relationships and determine more promising photovoltaic candidates. Jin et al.[17] ruled out 26 potential 2D photovoltaic (2DPV) candidate materials from 187,093 inorganic crystal structures identified experimentally. The developed framework, based on integrating high-throughput material screening and the ML algorithms, achieved high accuracy and efficiency in identifying 2DPV (**Figure 27**a). The established descriptor space consisted of 19 critical descriptors such as the packing factor (Pf), average sublattice neighbor count (SNC), and Mulliken electronegativity minimum value. The training dataset consisted of 98 experimentally identified PV materials as positive samples and 98 non-PV materials as negative samples. To evaluate the model performance, 5-fold CV was employed, where four essential evaluation metrics, namely, accuracy, recall, precision, and AUC, were selected to examine the models. After the implementation of ML models, the best-performed gradient boosting classifier (GBC) model screened 3,011 out of the 187,093 PV candidates from the ICSD with all four of the performance evaluation metrics being approximately 1. Further screening was performed by considering the layered structure, leading to only 26 2DPV candidates ultimately being kept and further classified into 10 different prototypes based on their space groups (Figure 27b). The electrical and optical properties of these candidates, such as PCE and bandgaps, were investigated by DFT computation. Three out of the 26 2DPV candidates including Sb_2Se_2Te, Sb_2Te_3, and Bi_2Se_3 were found to possess remarkable PCEs of 24.06%, 22.65%, and 15.85%, respectively. With the

analysis of the electronic and geometric structures, the author revealed two crucial factors that contributed to the high PCE: the long-pair s-orbital in the compound enables p-p optical transition (Figure 27c); asymmetric structures induce a built-in electric field (Figure 27d), which prolongs the exciton lifetime. The employed ML classifier efficiently discovered promising 2DPV candidates and revealed implicit structure-property relations.

In another study, the bandgap properties of kesterite I_2-II-IV-VI_4 (Figure 27e) were studied by Weston et al.[387] by integrating first-principle computations with ML techniques. Regression ML models such as LR, SVR with different kernels, and tree-based regressions were trained with 184 I_2-II-IV-VI_4 compounds to predict the magnitude of the bandgap. A logistic regression-based binary classifier was employed to identify whether the bandgap was direct or indirect. The performance of these regression models were determined via 10-fold CV and assessed using the RMSE and R^2. An accuracy metric was employed to evaluate the performance of logistic regression classifier. The radial-bias-kernel-based SVR model outperformed the other models, with a low RMSE of 0.283 eV and R^2 of 0.957 (Figure 27f). The descriptor space was initially constructed with 12 elemental properties for the regression models; further feature engineering was conducted for to improve the performance of the classification model, with a final prediction accuracy of 89%. The well-tuned regression models predicted the bandgap of 1,568 kesterite I_2-II-IV-VI_4 semiconductors, leading to the discovery of 243 materials whose bandgap falls in the optimal range of 1.2–1.8 eV. Cross-checking with material thermodynamic properties in the MP database revealed that 34 of screened materials were synthesizable, 25 of which were at the ground state. Follow-up first-principle calculations further verified that 25 of the 34 potential candidates exhibited the optimal bandgap value.

Huang et al.[388] used an ML model to predict the bandgap and band alignment of nitride-based semiconductors (Figure 27g). Based on the combination of HSE and DFT-PBE functionals, the value of bandgap and band offset towards the wurtzite GaN was accurately computed and then used to train

and test the ML models. Eighteen accessible elemental properties were taken as descriptors and several ML algorithms were employed(such as SVR with different kernels and NN). In the design space, the studied wurtzite nitrides consisted of 16 atoms in an orthorhombic supercell, and thus 68,115 possible structures were taken into account in consideration of all possible cation-nitrogen combinations. 300 out of the 68,115 structures were randomly selected as the training set, and the LOOCV was implemented. The SVR algorithm with radial kernel outperformed other models with a predicted RMSE of 0.298 eV and 0.183 eV in terms of the bandgap and band offset (Figure 27h and 27i), respectively. With further feature engineering, a descriptor space with 26 elemental properties was constructed, which decreased the RMSE of the bandgap prediction by around 0.005 eV. Two trends could be noticed: with the increase of cation types, the band sets tend to increase while the bandgap has an inclination to narrow. The prediction results of known nitrides models were in favorable agreement with corresponding first-principal calculations; a number of material candidates were explored with the potential to support novel discoveries in the domains of ultraviolet LEDs, infrared detectors and solar cell absorbers.

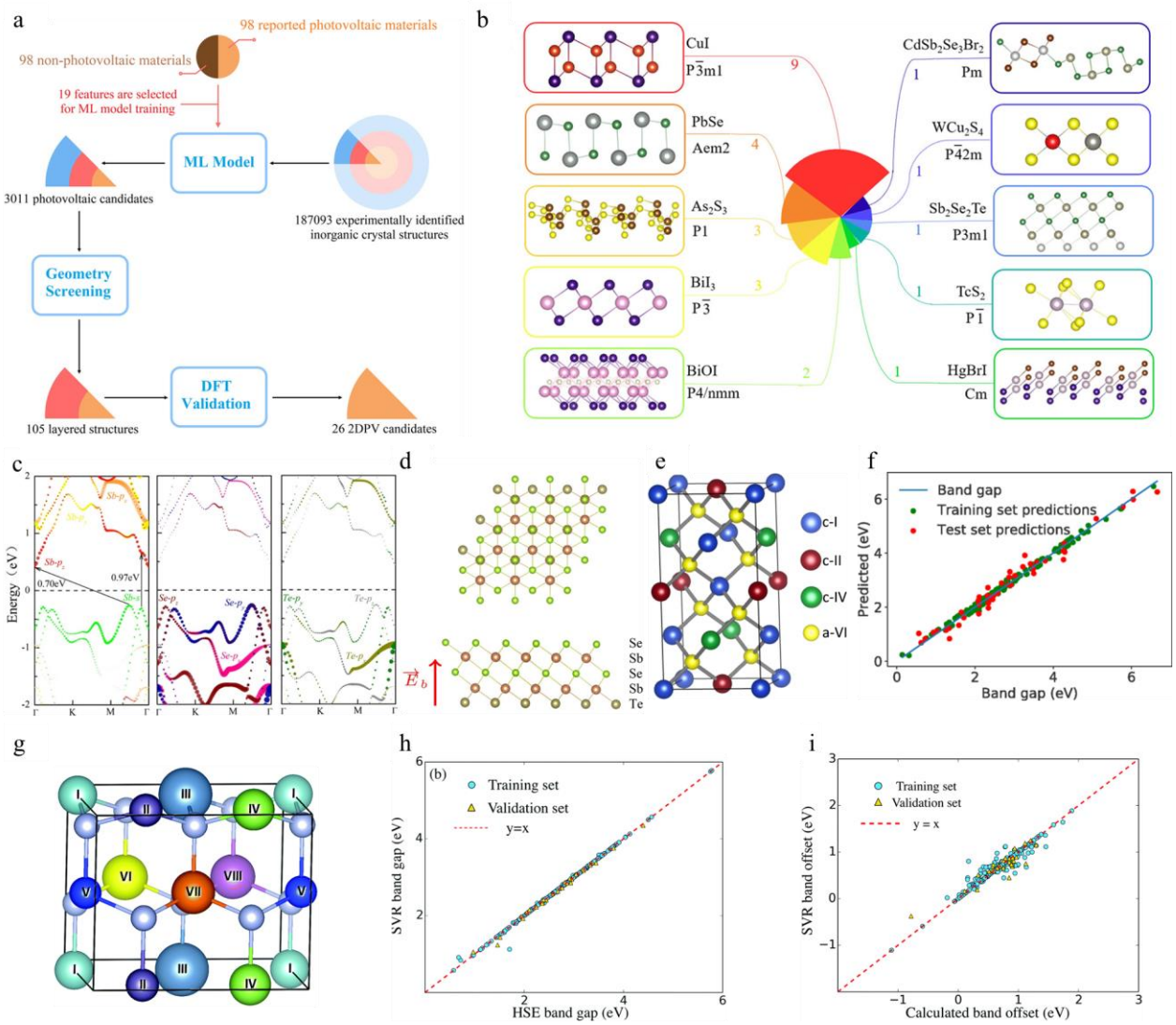


Figure 27. a) The framework of ML assisted material screening of SDPV candidates. b) The 10 structural prototypes of the 26 2DPV candidates. c) Projected band structure of Sb_2Se_2Te. d) Top and side view of the Sb_2Se_2Te cell. a-d) Reproduced with permission.[17] Copyright 2020, ACS Publications. e) The illustration of the structure of Zinc-blende-based kesterite for a I_2-II-IV-VI_4 compound. f) The performance of the radial bias kernel support vector regression model on the prediction of the bandgap e-f) Reproduced with permission.[387] Copyright 2018, AIP Publications. g) The illustration of the nitride structure in the design space and position of 16 ions. The labelled atoms indicate cations and the rest represent nitrogen atoms. h) The performance of the radial bias kernel support vector

regression model on the prediction of the HSE bandgap and i) bandgap offset g-i) Reproduced with permission.[388] Copyright 2019, RSC Publications.

In the ML-based applications of high-performance PV material discovery, several properties such as PCE, bandgap, V_{OC} and J_{SC} are critical to evaluate the performance of the potential PV material; therefore, ML models that can make efficient and accurate predictions are ultimately pursued. It is also vital to explore implicit relations between the material parameter and the target property via data-driven, ML techniques to acquire novel insights for the further development of PVs.

6.1.3. Fuel Cells and Metal-Air Batteries

Using data-driven technology to discover innovative, economical and efficient electrocatalysts has gradually become the focus of oxygen reduction reaction (ORR) research. ORR plays a vital role in chemical-electrical energy conversion in fuel cells and metal-air batteries, which is a promising and indispensable field in the development of renewable energy.[443] Recently, a new frontier ORR catalyst has emerged referred to as dual-metal-site catalysts (DMSCs). By employing ML techniques, Zhu et al. [374]identify the origin of ORR activity and reveal design principles that offer a universal description of the activity in relation to intrinsic properties for graphene-based DMSCs. In this research, they used DFT simulations to screen potential catalyst candidates by considering the two criteria of geometric structure and free energy for the reaction. Each candidate's catalytic performance was quantified based on the theoretical potential of the rate-limiting step (U_L); a value larger than 0.7V was regarded as favorable ORR activity. Their U_L of such DMSCs can only be higher than 0.7V when the rate-limiting step is either the first or fourth electrochemical step. A linear scaling relationship between ΔG_{OOH^*} and ΔG_{OH^*} for the evaluated DMSCs were determined via regression(ΔG_{OOH^*} = 0.92 ΔG_{OH^*}+ 3.01); thus, the trends in ORR activity with the variations in ΔG_{OOH^*} and ΔG_{OH^*} can be plotted (**Figure 28**a). Based on the DFT computations, numerous primary physiochemical parameters were enumerated as possible descriptors for ML training. As the activity of catalysts is essentially dominated by electronic strictures, properties of localized d-orbital and continuum s- and p- orbitals were selected as the primary descriptors. Additionally, considering interactions between two transition-metal atoms, some geometric structure-related properties were set as descriptors. The Person correlation coefficient matrix was used to identify the inner correlation between random descriptor pairs to eliminate redundant descriptors (Figure 18b). With some simple mathematical transformations, the descriptor space was extended and optimized in accordance with the ML model's prediction accuracy. Finally, a gradient boosting regression (GBR) model with an R^2 of 0.993 and RMSE of 0.036 eV was obtained (Figure 18c). The mean impact value (MIV)[444] method was coupled with the trained ML model to evaluate each descriptor's influence on the ORR activity (Figure 18d). The seven most related descriptors are: the electron affinity (EA_1 and

EA_2); the sum of the van der Waals (vdW) radius ($R_1 + R_2$); the absolute value of the difference between and the sum of the Pauling negativity ($|P_1-P_2|$, P_1+P_2) of the two transition-metal atoms; the product ($IE_1 \times L$) of the ionization energy of the first transition-metal atom (IE_1); the distance (L) between the two transition-metal atoms; the average distance between the two transition-metal atoms and the surrounding N atoms ((d1+d2+d3+d4+d5+d6)/6). Among the seven descriptors, five are electronics properties. However, isolated individual descriptors may have their limitations and may not be sufficient to describe the effects of atoms on catalytic performance. In contrast, too many descriptors would lead to the dimensionality curse and disrupt the model's predictive performance. Hence, it is essential to discover and identify new, high-dimensional descriptors which are highly related to the target results and carry the most information. Based on the data generated from DFT computations and microkinetic simulations, the trained ML model can accurately describe the ORR catalytic activity of DMSCs via fundamental parameters with acceptable error.

To study the electrocatalytic performance of more complex, larger structures, traditional DFT calculations are limited due to their large computational expense and time. Researchers have gradually developed new strategies that combine ML with DFT and other computing methods. Kang et al.[389] used Gaussian descriptors[373, 445] to characterize local atomic structure. The authors applied an ML-based framework to explore the thermo-electrochemical properties of ternary nano-electrocatalysts. A model of high-dimensional neural network potentials (NNPs) was trained with the employment of the atomistic ML package (AMP)[446] to describe the interactions between components (Figure 28b). The NNP method was then implemented in conjunction with Monte Carlo (MC) methods and molecular dynamics (MD) simulation to identify the effect of strain originating from surface segregation of selective components at the surface of the catalyst. 13,877 DFT calculated data for PtNi, PtCu, CuNi, and PtCuNi nanoparticles were used for the training sample. The training set of the model system was composed of nanoscale icosahedrons with transition-metal species mixed randomly. To distinguish the local structural environment, Gaussian descriptors on radial (G^2) and angular (G^4) symmetry functions were employed as the main parameters. The RMSE of the NNP model on the prediction of single-atom energy contribution converged to less than 7 meV with the implementation of the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm.[447, 448] The proposed candidate PtCuNi ternary that contains 60% Pt possesses a size of 2.6 nm demonstrates outstanding electrocatalytic ability toward ORR. According to the thermal-

electrochemical stability analysis via MC and MD simulations under the canonical ensemble, the candidate is also consistently more stable than binary nanoparticle and pure Pt. The design principle that emerges from the ML model is that those ternary nanoparticles with 60% Pt composition and icosahedron configurations in which Cu/Ni and Pt assume the core and shell, respectively, possess superior ORR catalysis performance in terms of both activity and stability.

The electrocatalytic performance of the core-shell catalysts is very attractive, but due to their impractical size, there still remain an insufficient number of mechanism studies having been reported. ML could further support the future development and exploration of core-shell catalysts. Rück[19] and his co-workers have further studied strained Pt-based core-shell electrocatalysts. They propose an ML-based framework for the prediction, with site-specific strain precision, to investigate how effect of strain on Pt core-shell nanocatalysts towards the ORR activity. The strained coordination number (cn*(j)), which describes the compressive and tensile strain on atom j with the variation of atomic coordination, was set as the target property of the ML model. The ML model was trained with a kernel ridge regression (KRR) algorithm, which applies a radial basis function (RBF) kernel to test nanoparticles whose structures are optimized for the minimum energy. The effective medium theory (EMT)[449] was used to calculate the structure energy by employing the ASAP calculator in the Atomic Simulation Environment.[450] The EMT-calculated energy was validated by DFT calculations on 1.9 nm sized core-shell nanoparticles. As is shown in Figure 28c, for each core, the ML model was trained with nanoparticle sizes from 1.6 nm to 5.4 nm at 0.2nm intervals. Five descriptors were selected: the coordination number (cn(j)) and generalized coordination number(CN(j)) to describe local-site structure, which has significant impact on the adsorption energy of the intermediates; the partial distribution function (PDF(j, r)); distance to alloy atoms($d_{alloy}(j)$); the interatomic distance from Vegard's law ($d_{veg}(j)$). The MAE of the ML prediction of the strain on single atoms varied from 0.0007 to 0.0159 with respect to different catalyst cores. In this study, the relation established by the ML model indicates that the size of the nanoparticle determines the

optimal strain. The mass activities could be enhanced by weakening compressive strain on PtAg and PtAu of sizes of 2.83 nm or by strengthening compressive strain on PtCu and PtNi of sizes of 1.92 nm.

To summarize, data-driven techniques are primarily implemented to establish the relation between the intrinsic properties and catalytic activity in the field of ORR. Some fundamental factors, including electronegativity, electron affinity and radii of the embedded transition-metal atoms, exhibit a high correlation with the ORR activity of DMSCs. Furthermore, in the design of core-shell ORR nanocatalysts, ML models indicate that the bimetallic material composition, size, and shell thickness of nanoparticles control the mass activity. In addition to catalytic activity, the thermal-electrochemical properties could also be predicted by ML models trained on descriptors generated by symmetric functions.

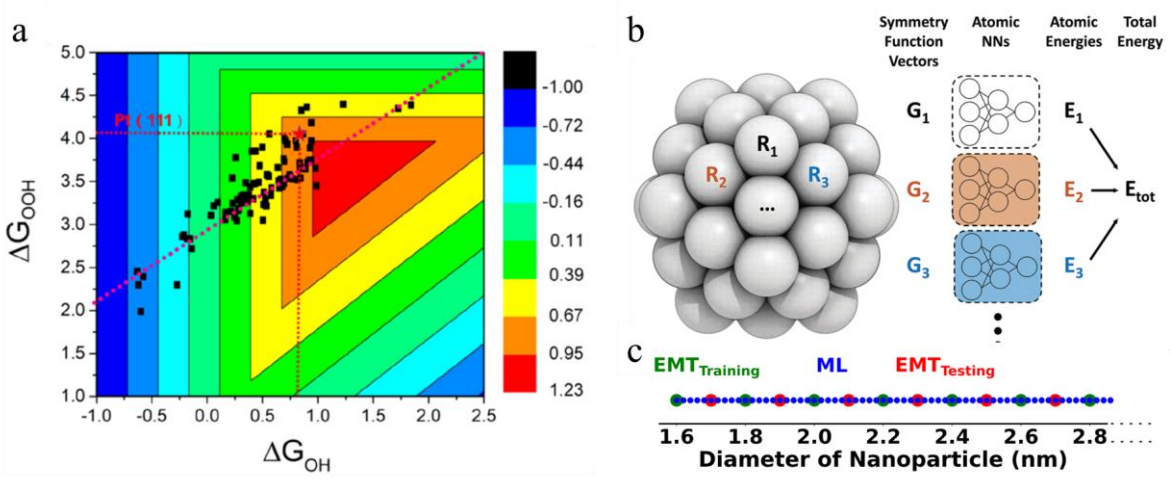


Figure 28. a) The trends plot of ORR activity with the variation of ΔG_{OOH*} and ΔG_{OH*} of DMSCs. Reproduced with permission.[19] Copyright 2020, ACS Publications. b) The scheme of the high dimension NNP method. The symmetry functions are transformed from the Cartesian to represent chemical environments. The NN then predicts the contribution of energy based on the symmetry functions and the total energy is obtained by adding up all of the energy contributions. c) The size of nanoparticles used for training, testing and ML prediction, which are represented in green, red and blue colour, respectively. b-c) Reproduced with permission.[389] Copyright 2018, RSC Publications.

6.1.4. Carbon Dioxide Reduction Reaction

CRR is considered to be a promising, clean, and environmentally friendly strategy to reduce greenhouse gas emissions and resolve the energy crisis; it has been broadly studied to improve reaction efficiency and selectivity.[451, 452] The introduction of ML for accelerating the discovery of CRR catalysts has been widely implemented in this domain, including the prediction of adsorption energies,[18] identification of active sites on the surface of catalysts,[20] optimization of reaction conditions for improving selectivity,[453] carbon dioxide capture ability, and design of catalysts.

Tran and Ulissi [13] employed an active ML model to guide the DFT simulation to identify optimal intermetallic electrocatalysts for CO_2 reduction and H_2 evolution. A workflow was established to screen a chemical space of 1499 candidates across 31 different elements (33% p-block and 50% d-block) of intermetallic materials acquired from the Materials Project. The open-source code pymatgen was implemented, by which 17,507 adsorption surfaces and 1,684,908 adsorption sites were enumerated. The vector employed to represent the environment of the coordination site contained four descriptors: atomic number (Z), Pauling electronegativity (χ) of the element, number of atoms of the element that coordinate (CN) with CO, and crude estimate of the adsorption energy on the site (ΔE) (**Figure 29**a). A framework of continuous, alternating iterations between ML screening and DFT computation was constructed, where the results of DFT simulation were fed back to the ML model, and newly predicted potential adsorption sites with near-optimal values (ΔE_{CO} = -0.67 eV and ΔE_H = -0.27 eV) were sent back for DFT calculations to generate new training data. Figure 29b represents the normalized distribution for the low coverage, DFT computed CO adsorption energies (ΔE_{CO}) of all of the DFT researched surfaces. The low coverage ΔE_{CO} computed by DFT for surface (131) and predicted by the ML model for surface (844) are shown in Figure 29c and 28d, respectively. The RMSE, MAE, and MAD of the active learning model's prediction were 0.46, 0.29 and 0.17 eV, respectively. One reason for this considerable error could be the use of ideal structures rather than relaxed structures for DFT calculation, as it is faster and less computationally expensive, though with the trade-off of the prediction accuracy.

Zhong et al.[18] used an ML model to predict the CO adsorption energies (ΔE_{CO}) on the adsorption sites of copper-containing intermetallic crystals, among which Cu-Al alloy was found to be the most promising electrocatalyst. The ML-predicted CO adsorption energy combined with the volcano scaling relationships[451] revealed the highest number of catalytic adsorption sites, where the CO adsorption value energies were near the optimal value of -0.67 eV (Figure 21a).[13, 18] A similar descriptor space was applied for each element type-coordinate with CO to characterize the first and second neighbouring shell of CO for each active site, with the difference that ΔE is replaced by the median adsorption energy ($\Delta\tilde{E}$) between the pure element and CO, yielded from the prior DFT simulation. The constructed vector space was then sent to an automated ML tool called the Tree-based Pipeline Optimization Tool (TPOT)[454] to implement the random forest regression (RFR) model. By using 19,644 DFT simulated data points of ΔE_{CO} and an extra tree regressor with 5-fold CV, the RFR model demonstrated both a median absolute deviation (MAD) and mean absolute error (MAE) of about 0.1 eV in predicting the ΔE_{CO} on the test data (5% of the whole data size), which is comparable to the accuracy of DFT simulation. The trained ML model was then coupled with the quantum chemical computation framework to construct an active ML system. The ML model predicted the ΔE_{CO} of all the adsorption sites enumerated by the DFT framework from Materials Project (MP); those sites whose predicted ΔE_{CO} was close to -0.6 eV were automatically collected and sent to the next stage. DFT simulations of ΔE_{CO} were subsequently executed for these sites, and the additional yielded data of ΔE_{CO} were then added in the training dataset to iterate a new ML model. The further optimized and improved ML model would identify new promising adsorption sites based on the value of predicted ΔE_{CO}, which could be fed back to the DFT framework to provide new ML training data. Thus an automatically, iteratively and systematically active ML workflow was established and a DFT database of ΔE_{CO} on promising adsorption sites was constructed. In this work, the structures established from MP were managed by Atomic Simulation Environment (ASE); [450] the Python Materials Genomics (pymatgen), which currently powers the MP, was used to enumerate all

the surfaces and adsorption sites. DFT calculations were performed with VASP, while software including Lungi and FireWorks were used to manage the computation framework and workflow. The active ML workflow finally trained more than 300 RFR models, and the guided DFT simulations were ultimately conducted for 4000 different candidates of adsorption sites with a near-optimal value of ΔE_{CO} on the Cu-containing surface quarter of which the majority were on Cu-Al Surfaces (Figure 21c). The integration of the volcano relationship, DFT simulation, and active ML achieved efficient and accurate prediction ideal electrocatalysts for active and selective CO_2 reduction to C_2H_2. Based on the ML results, the author concluded that those Cu-Al alloys that contain higher Cu composition are more promising for CRR. A follow-up experimental validation was performed and the CO_2-to-C_2H_4 performance achieved ~55% PCE under 150 mA cm^{-2} at the cathode side. Although numerous DFT-calculated adsorption energies are required for the training of ML model, this approach reveals the importance of the data-driven and active-ML-guided experimental exploration in overcoming the limitations of the conventional single-component catalysts in CRR.

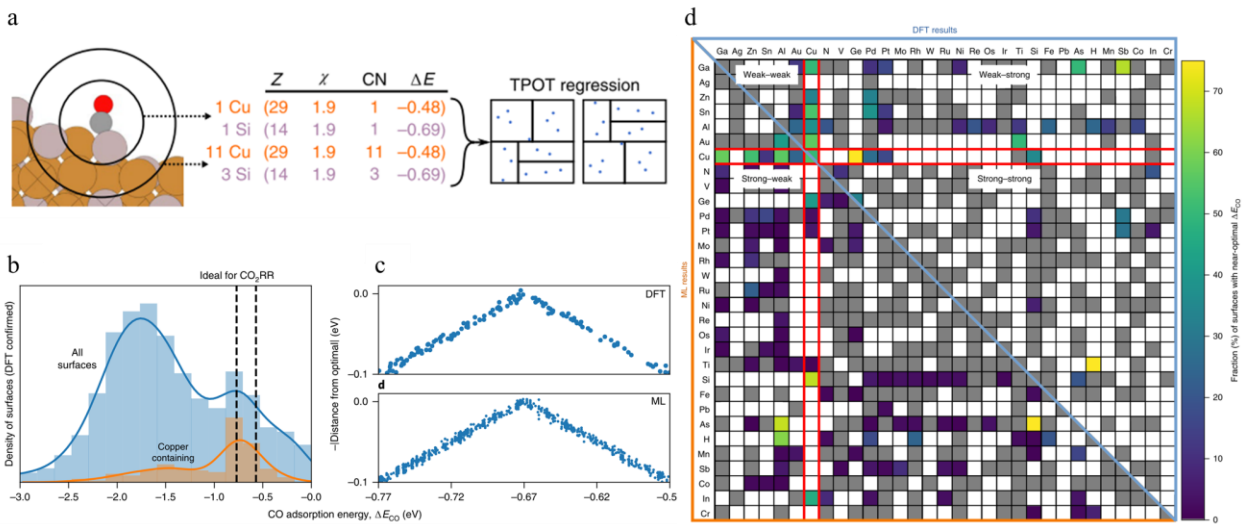


Figure 29. a) The sample of the numerical encoding for the adsorption site. The constructed descriptor space is employed as model input by the Tree-based Pipeline Optimization Tool (TPOT) to predict ΔE_{CO}. b) The normalized distribution of the low coverage, DFT derived ΔE_{CO} for all of the DFT computed surfaces. The sub-distribution for cooper containing surface is marked in orange, and the

black dashed lines indicate the range of for the optimal ΔE_{CO} (-0.67 eV). c) The low coverage ΔE_{CO} computed by DFT for surface (131) and d) predicted by ML model for surface (844). Reproduced with permission.[13] Copyright 2018, Springer Nature Publications.

Pedersen et al.[455] have explored a probabilistic and unbiased method to research high-entropy alloy performance as the electrocatalysts for the reduction of CO_2 and CO. The authors integrated the quantum chemical simulations and ML model to predict the ΔE_{CO} and adsorption energy of hydrogen (ΔE_H) of all the adsorption sites on the surface of the disordered CoCuGaNiZn and AgAuCuPdPt HEAs. The disordered surface consists of different metal atoms that would naturally provide many distinct adsorption sites with each adsorbate's unique adsorption properties, as determined by the site's microstructure. Hence, a Gaussian process regression (GPR) model was established that uses the adsorption energy of CO and H in the local atomic environment around the adsorption sites (computed by DFT) to predict the ΔE_{CO} and ΔE_H. The training data size was ca. 1000, where 5-fold CV was applied with MAEs of 46-64 meV (**Figure 30**a). The predictive model allows the optimization of HEA compositions to increase the probability of catalyzing performance improvement. Every local adsorption site contributes to the HEAs' global catalytic properties; some of the local optimal compositions such as $Co_9Ga_{42}Ni_7Zn_{42}$, $Ga_{83}Ni_{17}$, $Ag_{69}Cu_{31}$, and $Ag_{84}Pd_{16}$/$Au_{84}Pd_{16}$ were predicted. The best five-metal alloy candidates that contain at least of 10% of each elements are $Co_{10}Cu_{10}Ga_{60}Ni_{10}Zn_{10}$ and $Ag_{30}Au_{33}Cu_{17}Pd_{10}Pt_{10}$. A concurrent and independent work published by Nellaiappan et al.[456] have experimentally investigated the CRR performance on the AgAuCuPdPt HEA, where the results are in favorable agreement with the predictions in this work.

Important descriptors for the performance of CRR catalysts are also a necessary means to improve the accuracy of ML. Ma et al.[379] pioneered the use of a feed-forward ANN ML model via open-source PyBrain code to establish a nonlinear correlation between the descriptor vector and the ΔE_{CO}. The descriptor vector consisted of 13 electronic properties which were determined

theoretically, among which characterize the properties of the clean adsorption surface (such as d-states distribution) including the filling (f), center (ε_d), width (W_d), skewness (γ_1) and kurtosis (γ_2) of a d-band, in conjunction with the local Pauling electronegativity (χ_l) determined by delocalized sp-states, were taken as the primary descriptors. The secondary descriptors such as work function (W), atomic radius (r_0), the spatial extent of d-orbitals (r_d), ionization potential (IE), electron affinity (EA), Pauling electronegativity (χ) and the square of adsorbate-metal interatomic d coupling matrix element (V_{ad}^2), were also fed into the ML model. All the input features were standardized to improve the performance of the ANN model, and a 10-fold CV was performed; the ML-predicted adsorption energy of CO was shown to agree well with the DFT simulations, where the average RMSE achieved a value of 0.13 eV (Figure 30b). The outperformed candidate {100}-terminated Cu multimetallic alloys were discovered to have lower overpotentials but potentially higher selectivity towards the reduction of CO_2 to C_2 species. After a perturbation to the input descriptors was performed and the model responses were compared, the importance of the descriptors was examined (Figure 30c). The developed ML model demonstrated a novel methodology for capturing complexity in electrocatalytic CRR and acquiring accurate values of adsorption energies without expensive quantum chemical computations, providing in-depth understanding and strategies for catalysts design.

The majority of applications of data-driven innovation in CRR are for predicting the adsorption energy of CO and H to evaluate the activity and selectivity of the catalyst candidates. The atom environment of the local adsorption site plays a dominant role in the catalyst performance, and descriptors, such as electronegativity and coordination numbers, have high impact on adsorption energy. The exploration of the catalytic performance and material structure by using data-driven techniques provides the possibility of a rational design of high-performance materials to boost the CRR.

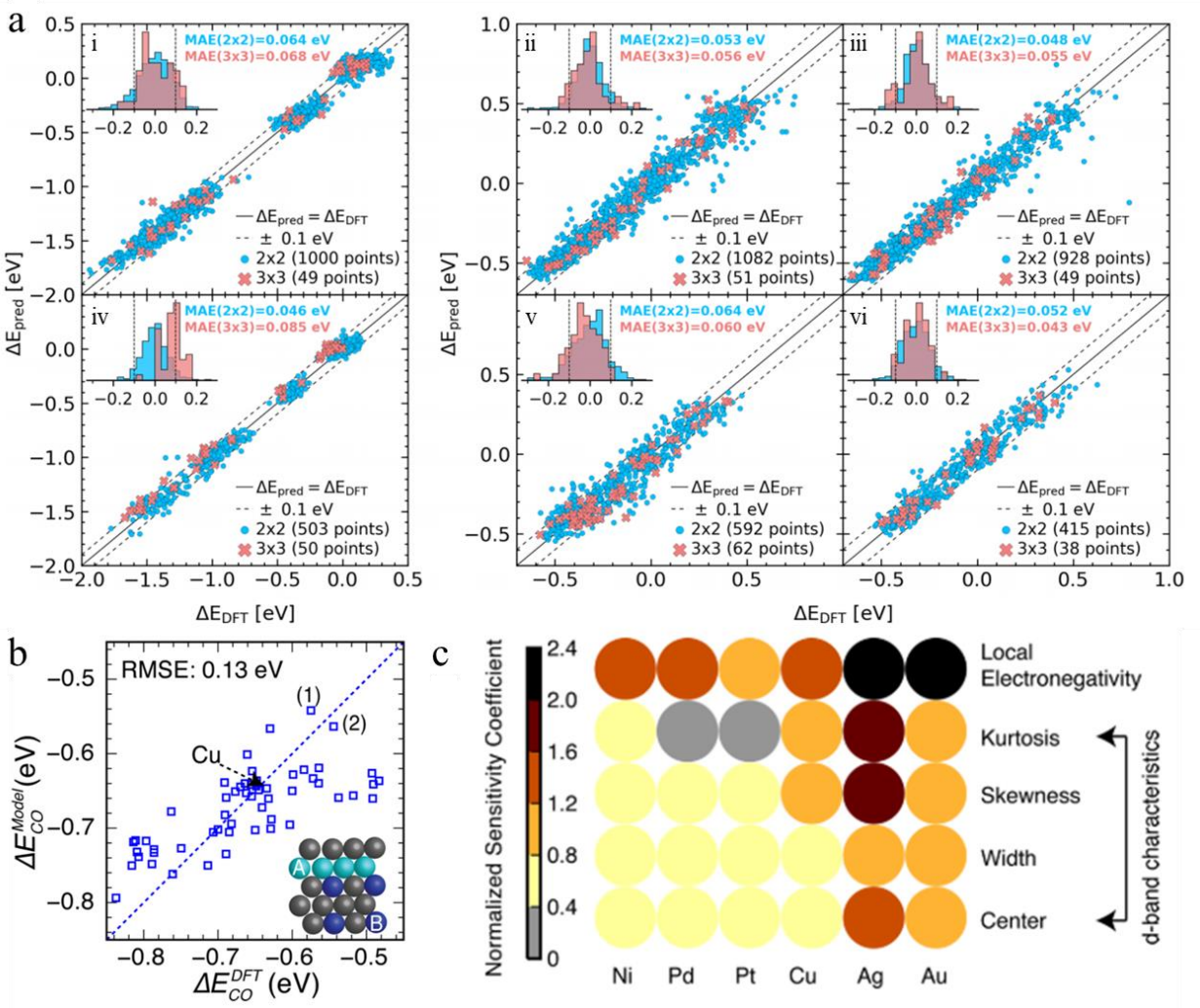


Figure 30. a) The performance of the GBR ML model for adsorption energy prediction. The ML predicted and DFT computed adsorption energies for on-top CO (i,iv), fcc-hollow H (ii,v) and hcp-hollow H (iii,vi) on the CoCuGaNiZn (i-iii) and AgAuCuPdPt (iv-vi) HEAs. Reproduced with permission.[455] Copyright 2020, ACS Publications. b) The performance of the NN ML model for adsorption energy prediction Cu monolayer alloys. c) The nominalized sensitivity coefficient of the d-band descriptors. Reproduced with permission.[379] Copyright 2015, ACS Publications.

6.1.5. Nitrogen Reduction Reaction (NRR)

The ML-assisted discovery of innovative catalysts for the NRR is the core of the alternative techniques to the Haber-Bosch process[457-460] for synthesizing NH_3 and are expected to yield excellent selectivity, activity, and efficiency. The traditional Haber-Bosch ammonia synthesis poses a significant environmental impact given that its reaction conditions are inherently energy-intensive, while primarily using natural gas as the hydrogen precursor .[457] Considerable attention has been paid to identifying noble metal-free catalysts to enable electrocatalytic N_2 fixation.[461-463]

Ulissi et al.[392] employed a Gaussian process regression (GPR) model to predict the free energy of all the possible adsorbate coverages on the target surface ((110) IrO_2 and four site edge of MoS_2 stripe) for NRR, by which the complexity was reduced to generate the corresponding Pourbaix

diagram (**Figure 31**a). Starting with 10 sample points, the GPR model iteratively predicted the free energies of unknown configurations, and the required electronic structure DFT relaxations were significantly reduced from about 90 to 20 points (**Figure 31**b). This study proposed a rational and systematic approach to obtain accurate free-energy diagrams with less computational cost. Hoar et al.[393] trained two feed-forward multilayer perceptron neural networks, accounting for the microkinetic model and electrode geometry, to evaluate the performance of the electrochemical NRR process (Figure 31c). Finite element methods were coupled with NN to guide the interrogation of the relationship between electrode geometry and electrocatalytic performance to extract insights for design principles. The total current density (i_{total}) and Faradaic efficiency (*F.E.*) were taken to measure the performance, whose training data was gathered from their previous work.[464] The wire length (*l*), diameter (*d*), and array periodicity in a square lattice (*p*) were used to define the microwire/nanowire array morphologies. The microkinetic variables for NRR were presented by the equivalent exchange current density of the first electron-transfer and first proton-transfer step. The framework proposed in this research provided a methodology to rapidly optimize wire-array morphology for reported catalysts and explore the effect of electrode geometry on catalytic performance. Four high-performance catalysts were predicted, whose *F.E.* and i_{total} were higher than 90% and 2 mA/cm^2, respectively. Although these results have not been experimentally validated, they indicate the feasibility of rational electrocatalysis design for optimal morphology. The reported insights and extracted knowledge from this framework could be extended to other catalytic applications such as ORR and CRR, which will guide effective morphology optimization.

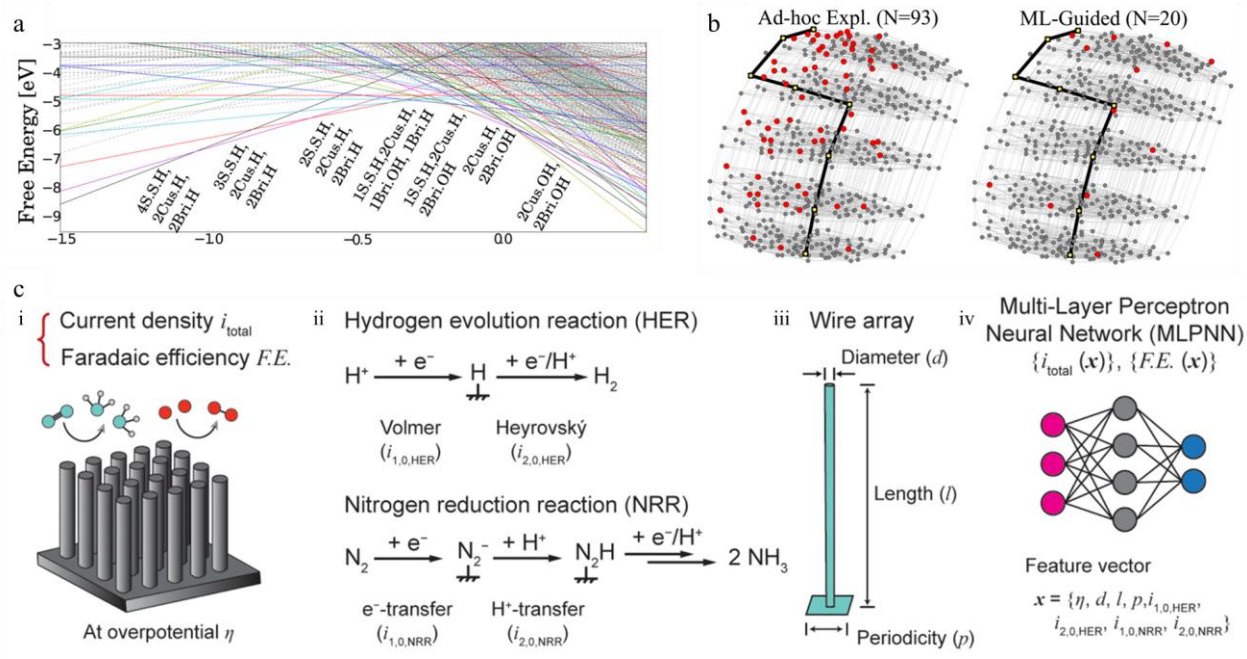


Figure 31. a) The final free energy diagram generated by ML models. b) The visualized network and exploration process of possible surface configurations for IrO_2. a-b) Reproduced with permission.[392] Copyright 2016, ACS Publications. c) The schematic diagram for the microkinetic model with MLPNN, four stages are involved: i. the measurement of efficiency. ii. the microkinetic modelling of NRR and HER. iii. the geometry parameter of the electrode and iv. the feature set and labels of MLPNNs. Reproduced with permission.[393] Copyright 2020, ACS Publications.

6.1.6. Thermoelectricity

The development of thermoelectric applications has long been limited by material performance, which could be addressed by ML techniques to identify high-performance thermoelectric materials. Converting heat into electricity (and vice versa) using thermoelectric materials[465-467] is of significant industrial and technological interest for numerous applications such as electricity regeneration from waste heat,[468] refrigeration[469], and other applications.[470, 471] The performance of a thermoelectric material can be quantified by either a dimensionless figure of merit ($zT=\sigma S^2T/(k_l+k_e)$) or the power factor (σS^2),[472-475] where σ, S, T, k_l and k_e denote the electrical conductivity, Seebeck coefficient, the absolute temperature, the lattice thermal conductivity and electronic part of thermal conductivity, respectively.

Choudhary et al.[394] designed a systematic data-driven framework for high-efficiency identification of thermoelectric materials by combining quantum chemical computation with ML models. With the utilization of the JARVIS-DFT database, 2,932 out of 36,000 three-dimensional (3D) and 148 out of 900 two-dimensional (2D) materials were identified as promising candidates for thermoelectric applications. In the multi-step screening process, material properties such as the Seebeck coefficient, power factors, and bandgaps were computed and selected as the thresholds. In **Figure 32**a, we show the possibility that a compound containing given elements has a high-power factor. The transport properties of materials were calculated by applying the Boltzmann transport equation with the BoltzTrap code implementation.[476, 477] Several ML models, including the GBDT, RF, *k*-NN, and ANN were trained to rapidly screen out thermoelectric materials for further quantum computation validation. A complete chemo-structure descriptor set called the classical force-field inspired descriptors (CFID)[478] was used for the ML model training. Though the CFID provides 1,557 descriptors for each material, the low-variance descriptors were removed and the descriptor space was standardized by using preprocessing techniques such as "VarienceThreshold" and "StandardScaler" in the scikit-learn Python package.[479] With the employment of their data-driven framework, they found that the materials in the family of ZrBrN possess ultra-low lattice thermal conductivity. The database and tool utilized for the evaluation and prediction of thermoelectric performance would significantly promote the discovery and characterization of thermoelectric materials.

The lattice thermal conductivity κ_l determines nonmetal components' ability to conduct heat and serves as a vital design parameter. Chen et al.[395] have developed a GPR model to instantly and accurately predict κ_l of inorganic materials with an experimental 29-dimensional descriptor space that contains 100 inorganic materials. It was discovered that the space group is one of the most critical descriptors that influence predictions. The average factor difference of k_l by the GPR model is 1.36, which is in agreement with reported values from semi-empirical models such as Slack[480] and Debye-Callaway.[481] In another property prediction task, Hou et al.[396] developed a data-driven framework to optimize the σS^2 of the off-stoichiometric $Al_{23.5+x}Fe_{36.5}Si_{40-x}$ (Figure 32c) component of $Al_2Fe_3Si_3$ while controlling Al/Si ratio. The determined optimal ratio x=0.9 yielded a 40% improvement of σS^2 (670 μW/mK^2) compared with that of the sample with x=0. Both of the two ML-based research are helpful for the rational design and initial screening of novel thermoelectric materials with desigred target properties, including κ_l, S, and σS^2, for specific applications. In addition, the explored relations via ML can enable better understanding of heat transport in inorganic materials.

To conclude, the thermal conductivity, power factor, and Seebeck coefficient are the critical properties that determine the conversion efficiency between heat and electricity. The ML-assisted data-driven approaches are capable of accelerating the pre-screening process, enhancing the understanding of performance, and providing insights for material rational design. A Pre-existing thermoelectricity database and data-intensive tools[394] for predicting thermal electricity performance would significantly reduce the time for data acquisition and thermoelectric material characterization.

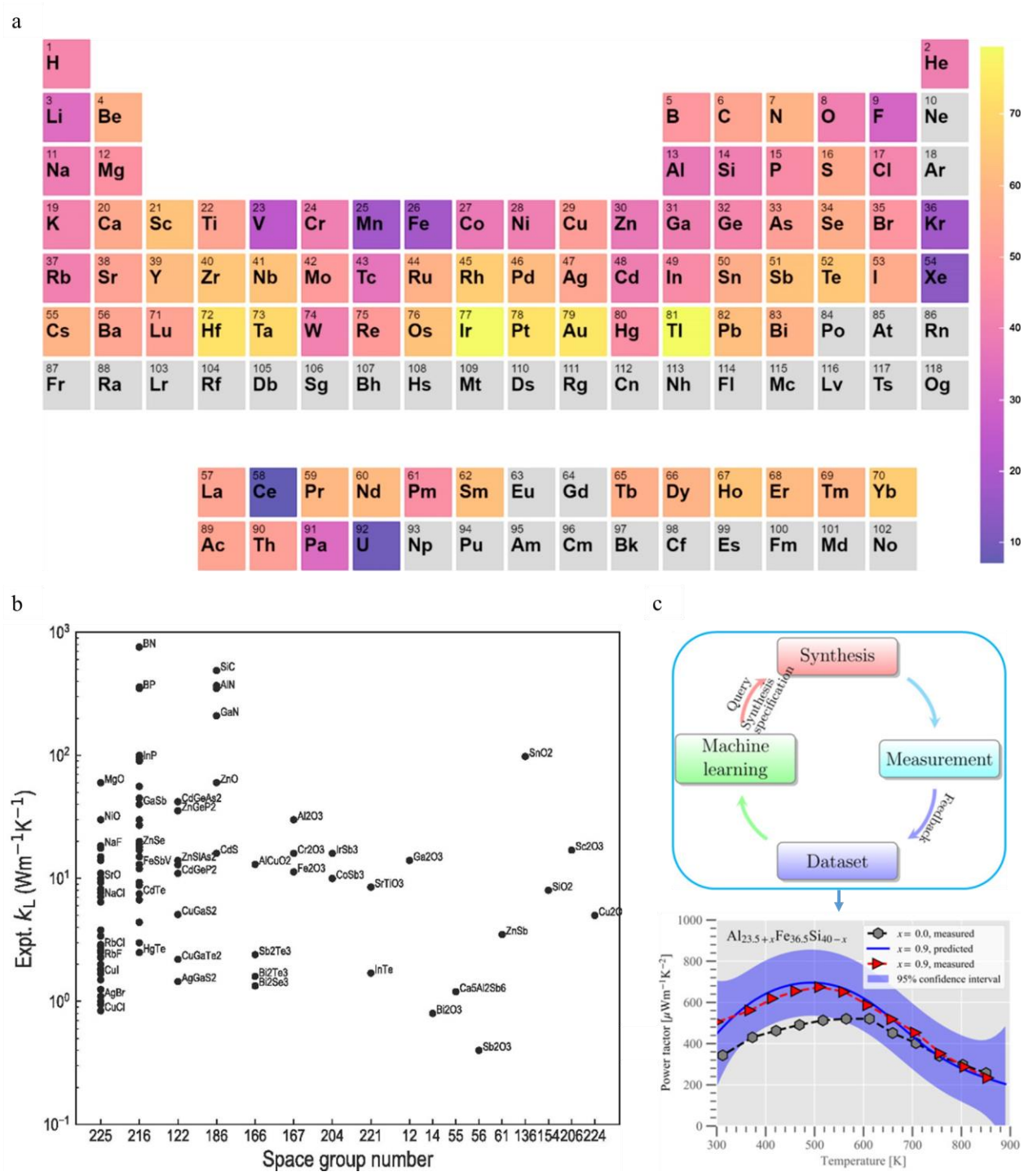


Figure 32. a) The heat map for elements that could form high power factor compounds. Reproduced with permission.[394] Copyright 2020, IOP Publications. b). Relationship between space group and experimental measured κ_l of the selected inorganic compounds.[395] Reproduced with permission.[396] Copyright 2019, Elsevier Publications. c) Framework for the ML-assisted design of $Al_{23.5+x}Fe_{36.5}Si_{40-x}$ toward an optimized power factor (PF) and the relationship exploration between power factor and temperature. Reproduced with permission.[396] Copyright 2019, ACS Publications.

6.1.7. Piezoelectricity

Data-driven processes are necessary to discover novel groups of piezoelectric materials and to support relation exploration between structures and piezoelectricity. The mutual coupling mechanical strain and electrical fields of piezoelectric materials leads to numerous applications such as actuators, transducers and sensors.[482-484] Many materials exhibit piezoelectricity, such as crystalline materials (quartz, langasite), ceramics (lead zirconate titanate), lead-free piezoelectricity (barium titanate)[37] and polymers (polyvinylidene fluoride).[37, 397, 485, 486] Though intensive research on the lead zirconate titanate (PZT) family has been conducted based their large piezoelectric coefficients [487] and vertical morphotropic phase boundary (MPB),[488] they are facing global restrictions attributable to the toxicity of Pb^{2+}; as a result, there is an urgent demand for other families of piezoelectric materials.[489-491]

Yuan et al.[37] applied an active learning framework with the coupling of ML models and optimization algorithms to accelerate the discovery of innovative, lead-free $BaTiO_3$ (BTO)-based piezoelectric materials with high electrostrains. The core concept of the active learning loop (Figure 3c) is that the ML model could be iteratively tuned and optimized according to prior predictions and feedback. This approach balances the trade-off between the fitting of regression and uncertainty in predictions,[492] which yields an optimal strategy for determining criteria to guide the rational design of materials. Pearson correlation coefficients were employed to eliminate highly-correlated descriptor pairs to construct the descriptor space. To enable a greater degree of feature engineering, the boosting gradient method was employed to rank each descriptor's importance to the electrostrains, and the seven highest-ranked descriptors, including electronegativity, ionic radius, volume, ionic displacements, polarization and dopant effects on transition, were used in the construction of the descriptor space. Six different ML models were trained for the prediction and optimization, and the advanced active learning framework indicated that the $(Ba_{0.84}Ca_{0.16})(Ti_{0.90}Zr_{0.07}Sn_{0.03})O_3$ possessed the largest electrostrain of 0.23% in the BTO family. Based on further DFT investigation, the authors indicated that the presence of Sn lead to an improvement in electrostrain. Follow-up experiments yield comparable results to the ML prediction of electrostrains. In this data-driven study, an active learning framework was developed to accurately and efficiently predict material target properties and guide experiments based on the optimal criteria derived from the trade-off between exploitation and exploration.[492] The core idea of the iterative framework is to find the next candidate by employing uncertainty to explore chemical space.[493] This method may have the potential to be applied to other specific applications given that the identified candidate can be experimentally or computationally validated following the recommendation. Similar to other key properties, the MPB has also been researched via data-driven approaches by Xue et al.[397] With the Bayesian learning algorithm's utilization, they predicted,

synthesized, and characterized a solid solution of $(Ba_{0.50}Ca_{0.50})TiO_3$-$Ba(Ti_{0.70}Zr_{0.30}Sn_{0.03})O_3$ with outstanding temperature reliability.

6.2. Energy Storage

In recent years, data-driven innovation has resulted in outstanding breakthroughs in the theory and calculation of the energy storage. Rechargeable alkali-ion batteries possess portability and high energy density, and are advanced energy storage candidates for clean energy and transportation. For studies focusing on alkali-ion batteries, the modification of electrolytes and electrodes has been the goal of extensive development. Data-driven and ML-aided approaches have aroused interest in the design of key parameters for alkali-ion battery such as voltage, capacity, volume, and other electrochemical-related battery performance parameters. In addition, because of their higher power density, long-cycle stability and high safety, supercapacitors are regarded as an alternative (or supplementary) rechargeable battery in applications that require high-power transmission or fast storage of energy.[494] This section will briefly introduce the application of data-driven material innovation in the theoretical design and development of rechargeable alkali-ion batteries and supercapacitors.

6.2.1. Rechargeable Alkali-Ion Battery

This sub-section will discuss the accelerated discovery of potential material candidates for electrolytes and electrodes based on data-driven strategies. As a key component of electrochemical energy storage, rechargeable batteries are extremely vital for various applications, including new energy vehicles, consumer electronics, and aerospace. To meet the growing needs of these applications, larger volumes of rechargeable batteries are being demanded with higher energy density, higher power density, longer cycle life, greater safety, and at an acceptable cost. Thus, it is essential to develop key rechargeable battery materials, including those for electrodes and electrolytes, to improve the performance of rechargeable batteries [4]. Data-driven screening of electrolytes often quickly identifies promising electrolytes through indicators such as chemical and structural stability[398], electronic properties[398], mechanical properties[399], and coordination energy[400]. For electrodes, voltage[402], volume[87] and redox potential[401] are essential for ML to successfully predict and evaluate the performance of electrodes.

Electrolytes

Electrolytes are vital components of rechargeable batteries; it is essential to find high-performance electrolytes in the development of advanced rechargeable batteries.[495, 496] With the significant advance of quantum chemical computations and ML learning techniques, some researchers have applied high-throughput data-driven approaches to discover innovative, next-generation battery electrolytes.[28, 497-501] Sendek et al.[398] have proposed a workflow of large scale computational material screening for solid electrolytes in lithium-ion batteries (**Figure 33**a). The authors first acquired atomistic and electronic structure parameters for 12,831 lithium-containing candidates from the MP database, including the equilibrium atom position, the energy above the convex hull, the bandgaps, and the Gibbs free energy, utilizing the Python package Pymatgen.[220] This was followed by a primary screening stage using four prerequisite criteria: low electronic conductivity, high chemical and structural stability, and low material cost. A logistic regression model

was trained to identify the structures that are most likely to exhibit excellent lithium conduction based on five features including the average number of Li neighbors for each Li, the average sublattice bond ionicity, the average anion coordination number in the anion framework, the average shortest Li–anion and Li–Li distance in angstroms. The training set consisted of 40 crystal structures whose ionic conductivity values were available in the literature. The threshold of superionic conductive behavior was set as 0.1 mS/c;, finally, 21 structures demonstrated potential as high-performance electrolytes, some of which have been experimentally investigated.[502-505] This method is applicable to confirming the ionic conductivity of unreported inorganic materials.

Similarly, Ahmad et al.[399] conducted a high-throughput data-driven search over for solid electrolytes with outstanding dendrite suppression capability of Li on the anode. A crystal graph-based convolutional neural network (CGCNN)[158] was trained to predict the moduli of shear and bulk given a large, available, low noise dataset obtained from low uncertainty first-principle-calculated values. The CGCNN model was trained by only structural descriptors, which bypass first-principles calculations. Additional ML models based on GBR and KRR were also employed to predict the elastic constants of cubic materials (Figure 33d). Those predicted mechanical properties are critical in stabilizing the interface and computationally expensive to obtain via first-principle methods. Those properties were taken as the input of the theoretical framework utilizing the stability parameter[506, 507] to figure out the dendrite initiation on the Li metal anode. The stiffness of the material was found to be positively correlated with the mass density and the ratio of bond iconicity between Li and the sublattice, whereas a negative correlation was obtained with the sublattice electronegativity and volume per atom. Further investigations of thermodynamic stability and electronic conductivity were performed. Additionally, the method proposed by Sendek et al.[398] was employed to confirm the ionic conductivity. Over 20 mechanically anisotropic interfaces and 4 electrolytes including Li_2WS_4-$P\bar{4}2m$, Li_2WS_4-$I\bar{4}2m$, $LiBH_4$-$P\bar{1}$ and LiOH-P_4/nmm were predicted as promising to be employed to suppress dendrite growth. The screened candidates were highly anisotropic and generally soft, which indicate opportunities for acquiring innovative solid electrolytes with both high ionic conductivity and dendrite suppression. The R^2 on the predictions of elastic constants C_{11}, C_{12}, and C_{44} were 0.60, 0.79, and 0.6, respectively; this might be due to the uncertainty inherent in the DFT-calculated values;[508, 509] the use of low uncertainity might improve the model performance. With the ability of data handling and feature generation, the proposed methodology in this study is readily applicable the screening of other inorganic materials for properties of interest.

Existing studies have mainly concentrated on solid electrolytes. Investigations of liquid electrolytes hav barely been reported,[510, 511] mainly because the molecular structure of a liquid system is more flexible, which makes it challenging to extract structural information. Ishikawa et al.[400] integrated a data-driven method with quantum chemistry computations to predict the coordination energy (E_{coord})[512, 513] of alkali group metal ions (Li, Na, K, Rb, and Cs) in battery electrolyte solvents. The E_{coord} is closely related to ion transfer at the interface of electrolyte/electrode, which is first obtained by quantum chemical computations. The calculated

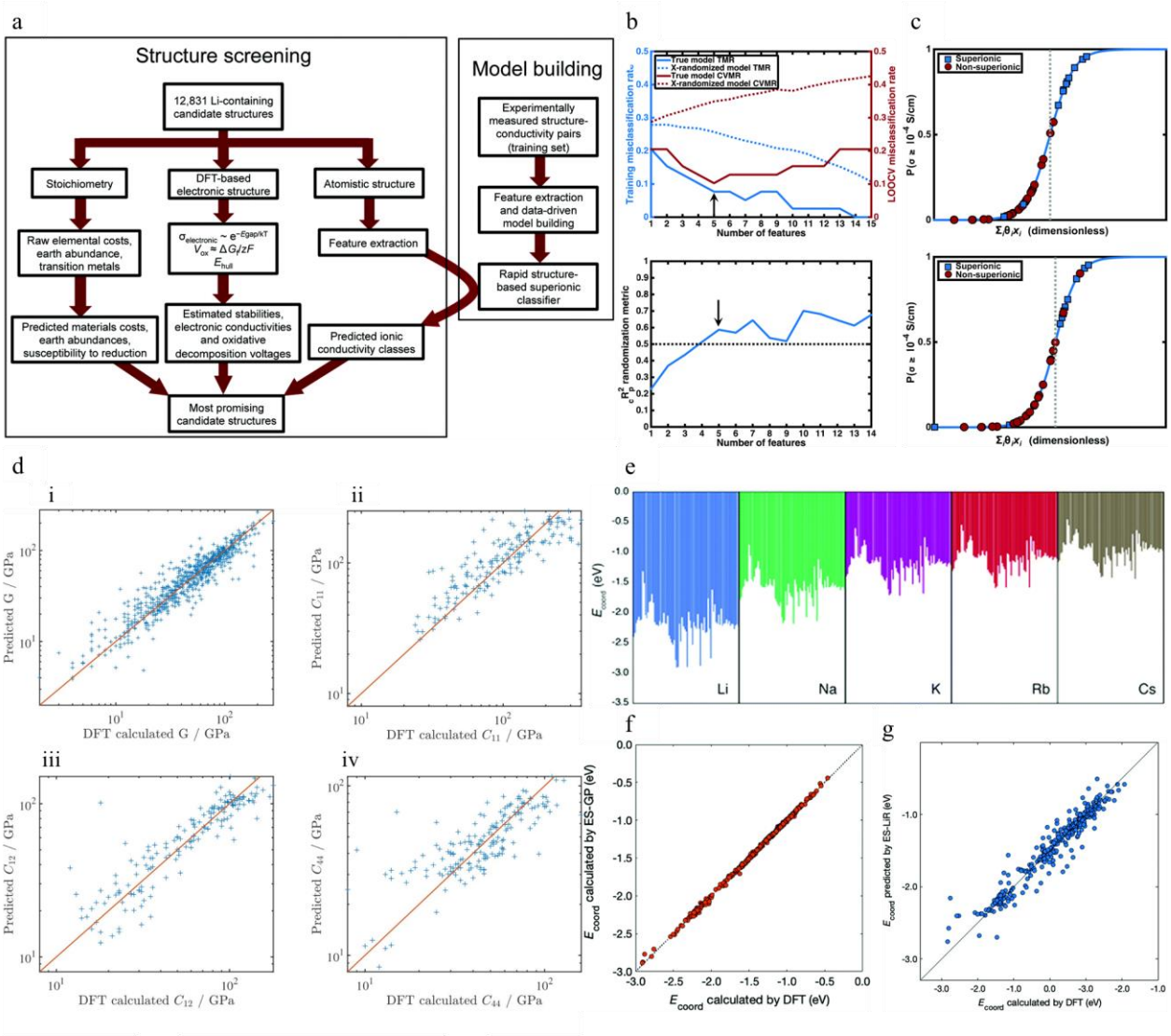


Figure 33. a) Flow diagram of the ML assisted material screening process for Li-contained candidates. b) (top) The training misclassification rate (TMR) and cross-validation misclassification rate (CVMR) via LOOCV. The dashed lines in the top diagram describe the mean value of the performed *X*-randomization analysis which is applied to ensure the model is not built on chance correlation. (bottom) The performance of ML models compare with chance correlations, the black dashed line indicates the threshold. c) The performance of the training data using logistical regression with LOOCV. a-c) Reproduced with permission.[398] Copyright 2017, RSC Publications. d) The comparison diagram of elastic properties between the ML predicted and DFT computed value: (1) shear modulus and elastic constants (2) C_{11} (3) C_{12} and (4) C_{44}. Reproduced with permission.[399] Copyright 2018, ACS Publications. e) E_{coord} of 70 solvents and the five alkali metal ions. f) The performance of the ES-GP model for the prediction of E_{coord}. g) The performance of the ES-LiR model for the prediction of E_{coord}. Reproduced with permission.[400] Copyright 2019, RSC Publications.

E_{coord} for 5 alkali ions is shown in Figure 33e. Three ML regression methods, namely, MLR, LASSO, and exhaustive search with linear regression (ES-LiR),[514-516] were implemented to identify the relationship between E_{coord} and selected descriptors. The descriptor space consists of both ion and

solvent properties, such as the ions' atomic weight and boiling point of the solvents. The results revealed that the most critical descriptors are the ionic radius and the oxygen atom's charge connected to the metal ion. The ES-LiR model yielded a CV error[514-516] of 0.127 eV for the prediction accuracy of E_{coord} (Figure 33f). By implementing the exhaustive search with Gaussian process (ES-GP) (Figure 33g), a further improvement of the prediction accuracy with a CV error of 0.016 eV was achieved. This study demonstrated that the integrated data-driven techniques and quantum chemistry calculations can accurately predict E_{coord} of any alkali metal ion coordination. The trained ML model could be employed to search for battery electrolyte materials, where several descriptors including ionic radius and NBO charge of the O atom are identified as critical in developing next-generation post-Li batteries.

Electrodes

Accelerating the discovery of suitable materials for high-power, safe, and stable electrodes is essential for developing improved rechargeable batteries. Because of the development of first-principles computations, the properties of unknown electrode materials can be obtained to support the research of complex phenomena.[517-520] Nevertheless, the advancement of ML techniques can enable more efficient discovery of innovative materials to identify the complex, implicit correlations between crystal structure and various properties of electrode materials such as voltage, capacity, ionic and electronic mobility, stability, redox potential, and volume changes in the battery.[517-521]

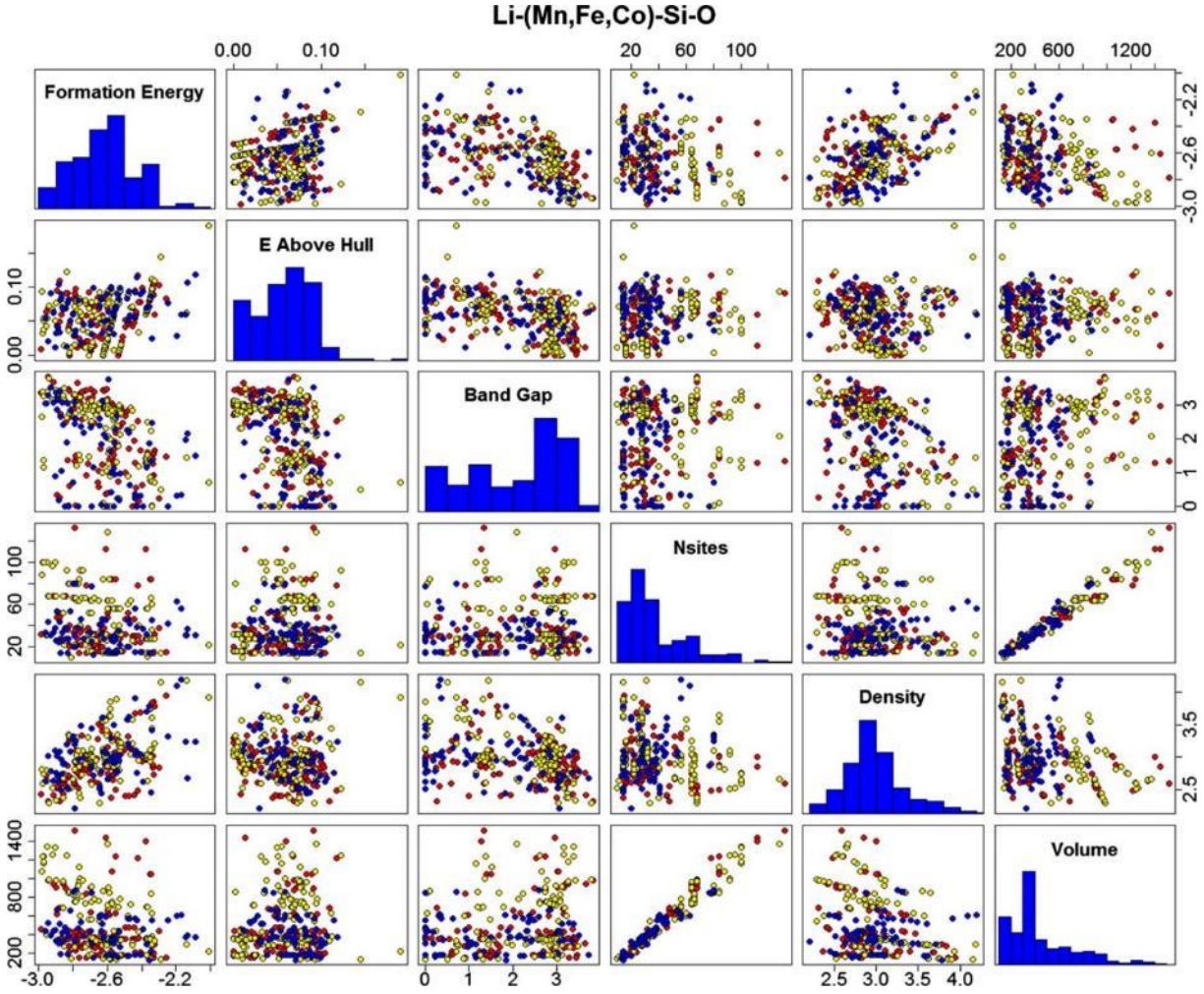


Figure 34. The plot of different properties pairs of Li–(Mn, Fe, Co)–Si–O cathodes according to the extracted data from MP database. The red, yellow and blue dots indicate the monoclinic, orthorhombic and triclinic crystal systems, respectively. Reproduced with permission.[401] Copyright 2016, Elsevier Publications.

Five ML classification models, including ANN, SVM, k-NN, RF, and extremely randomized trees (ERT) were implemented by Shandiz et al.[401] to categorize the crystal systems of silicate-based cathodes with the composition of Li–Si–(Mn, Fe, Co)–O into three major types: monoclinic; triclinic; orthorhombic. The training dataset contained 339 cathode material data points obtained from MP,[24, 522] with 5 descriptors including formation energy (E_f), energy above hull (E_H), bandgap (E_g), number of sites (N_s), and volume of unit cell (V_{uc}) (**Figure 34**). The prediction results indicated that the ensemble methods (RF and ERT) gave the highest accuracy of over 75% under Monte Carlo validation,[523] where the N_s and V_{uc} were dominant in determining the crystal system type. More recently, Joshi et al.[402] developed an ML-based tool to predict the voltage of electrode materials in metal-ion batteries. A total of 3,977 samples were collected from the MP database, where 237 features, such as the elemental properties of their constituents[33] and properties of chemical compounds,[372] were initially added to the descriptor vector. A PCA[524, 525] model was then performed to reduce the dimensionality of the descriptor vector to 80. The deep neural network (**Figure 35**a)[87], SVM[526], and KRR[116] model yield an R^2 value of 0.84, 0.86, and 0.86, respectively, on the prediction of voltage, therefore offering an alternative way to generate voltage profile diagrams instead of DFT methods[527] (Figure 35b). Additionally, nearly 5,000 electrode material candidates were proposed for Na- and Ki-ion batteries via these ML models, some of which were comparable with published experimental and DFT values.[528-530] Further improvement of the model performance could be implemented via the employment of different algorithms, more data, and novel ways of characterizing intercalation reactions.

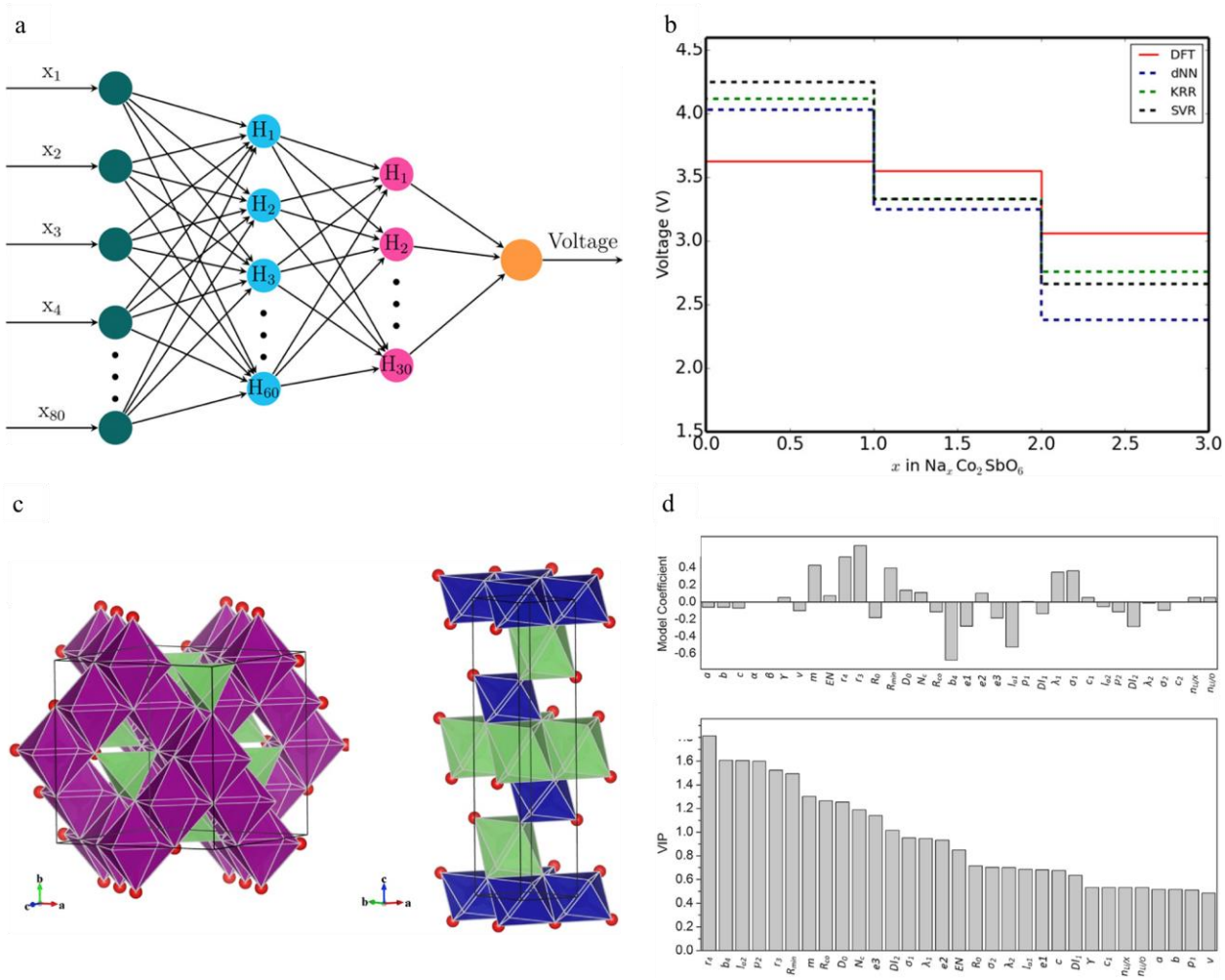


Figure 35. a) Schematic diagram of the neural network employed in this study. x_i represents the input of the NN and H_i represents the nodes in the hidden layers. b) The obtained voltage profile diagram from several ML models and DFT computation for $Na_xCo_2SbO_6$. a-b) Reproduced with permission.[402] Copyright 2019, ACS Publications. c) The scheme of crystal structures for (left) spinel LiX_2O_4 and (right) layered $LiXO_2$. d) (top) The model coefficient plot and (bottom) variable importance plot of the independent variables for the modeling PLS. c-d) Reproduced with permission.[531] Copyright 2017, Elsevier Publications.

Small volume changes of cathodes are critical for extending the cycle life of batteries.[532] Wang et al.[531] reported a methodology integrating first-principles calculations and partial least square (PLS) regression to formulate the quantitative structure-activity relationship (QSAR) of the volume change for cathode materials in Li-ion batteries. The scheme of crystal structures of the material is shown in Figure 35c. A total of 34 descriptors in five types, including element, crystal structure, composition, local distortion and electronic level, were selected to acquire the QSAR formulation (Figure 35d). It was found that the radius of X^{4+} ion and the octahedron descriptors of X contributed the most to cathode volume change. The established QSAR could be applied to a broader range of real or simulated materials. It is still challenging to design the low-strain cathode with the determined optimal combination of the descriptors, which might be realized via codoping at various atomic sites.

Data-driven innovation has emerged as a significant driver of material discovery and fundamental knowledge exploration in rechargeable alkali-ion batteries. This is typically accompanied by the integration of first-principles computation and ML techniques, which reveal implicit structure-property correlations and accelerate the high-throughput screening of electrolyte and electrode materials.[399] Although some of the trained ML models discussed earlier might have relatively weak prediction accuracy,[399] which might be caused by the uncertainty of DFT computation,[508] they still reflect the correct trends in target properties with with respect to material parameters. The selection of descriptors is of great significance to model performance. In some cases, geometric attributes and electronic properties can sufficiently describe the material and are relatively (computationally) cheap to obtain.[398, 399, 531]

6.2.2. Supercapacitors

Data-driven based sophisticated systems could promote the discovery of electrode materials to further enhance the performance of supercapacitors.[403] As a class of advanced energy storage, supercapacitors enjoy long circle life and high power density .[533] The carbon-based electrode, a critical component of a supercapacitor, is most widely used due to its extraordinary chemical and physical properties.[533, 534] For instance, Zhu et al.[403] adopt several ML models, such as the LR, LASSO, and ANN (**Figure 36**a), to predict the capacitance of the carbon-based electrodes. Specifically, 681 training data points were obtained from the literature with five selected supercapacitor descriptors: specific surface area,[533, 534] voltage window, calculated pore size,[535] N-doping level,[536] and intensity ratio of the D-band to G-band (I_D/I_G) (Figure 36b).[537] The ANN model exhibited the best accuracy with an R^2 of 0.91 and adaptability with respect to the capacitance prediction (Figure 36c), revealing the potential of the ML model (especially of ANN) to accelerate and assist the innovation of electrode materials in the domain of supercapacitors.

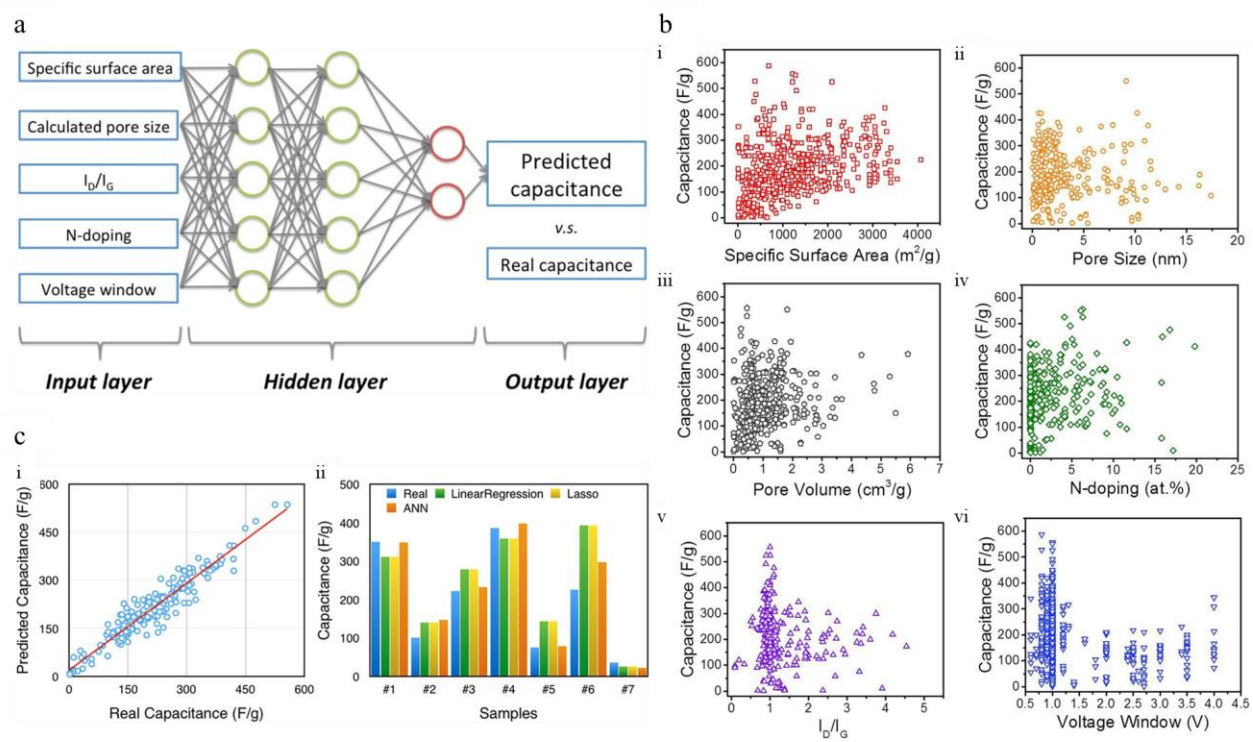


Figure 36. a) The relationship between capacitance and i. specific surface area, ii. pore size, iii. Pore volume, iv. N-doping, v. I_D/I_G, and vi. voltage window. b) The schematic of the ANN architecture used in this study. c) The i. performance and ii. results of different ML models of the capacitances prediction. Reproduced with permission.[403] Copyright 2018, Elsevier Publications.

6.3. Environmental Decontamination

Advanced oxidation processes (AOPs) are essential for treating industrial wastewater and decontaminating dyes, which are an emerging concern.[538] Data-driven techniques on AOPs have focused on experimental design, reaction condition optimization, process modeling, and oxidation performance prediction.[539-545] Notably, compared to the conventional AOPs such as Fenton's process, radiolysis, ozonation, and ultrasonic process, the photoelectrochemical oxidation process is an innovative and promising AOP technique for the degradation of water pollutants.[546-551] Dondapati et al.[14] established a quantitative structure-property relationship (QSPR) (**Figure 37**a) to predict the degradation rate of phenolic pollutants on modified nanoporous titanium oxide electrode (TiO_2) in advanced photoelectrochemical (PEC) oxidation. The multiple linear regression (MLR) model was employed to elucidate the QSPR, and the best predictive model achieved an R^2 and RMSE of 0.9625 and 0.1073, respectively, under LOOCV. The descriptors used in this study were mainly collected from PaDEL[43] or calculated by Gaussian 16W.[552] The determined QSPR revealed the effects of important physicochemical and electronic properties during the PEC oxidation process. An increased hydrophobic nature accompanying the meta position substituent tended to have a high degradation rate. Additionally, the topological descriptors related to molecule shape and connectivity were vital for improving the degradation rate of dyes. Further, the radial distribution function (RDF)[553] descriptors can reflect electronic properties and molecular density. Those descriptors were also computed for Rhodamine B (Figure 37b), which was taken as the model pollutant for the degradation simulation. It was found that the determined dominant descriptors in phenolic degradation were also critical in the degradation of Rhodamine B. Though the training data size was relatively insufficient for ML employment, this study still reveals the importance of the electronic, hydrophobic and topological properties that significantly influence the degradation of organic pollutants in the AOP process. High-throughput experiments[554] or simulations might increase the training data size for the enhancement of data-driven innovation in the domain of AOP.

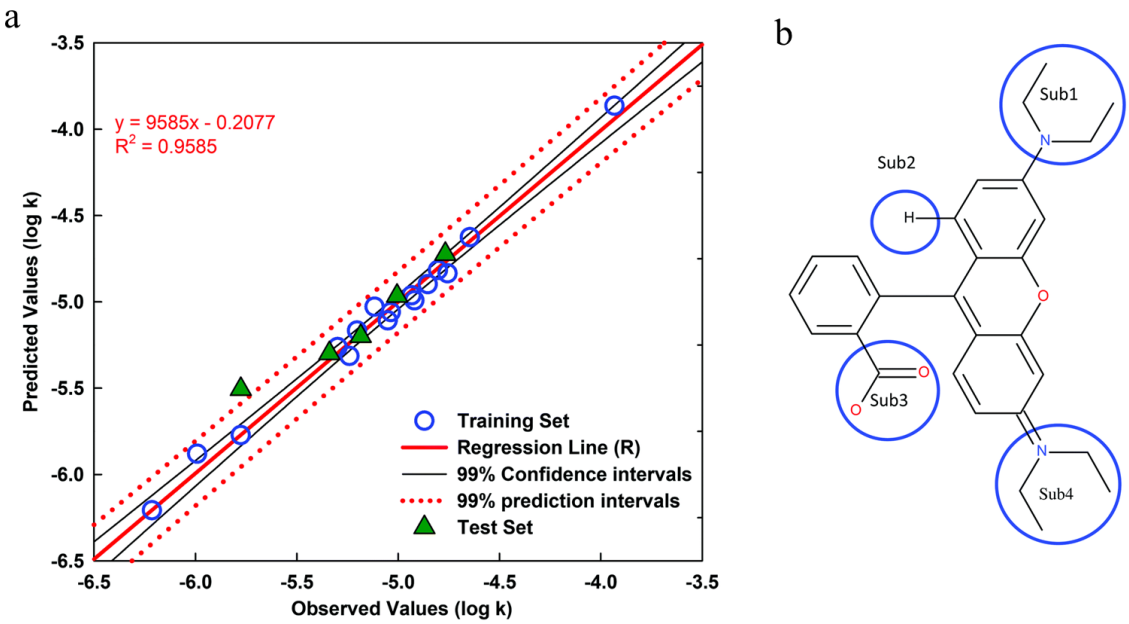


Figure 37. a) The performance of the MLR models on capacitances prediction. b) The structure diagram of Rhodamine B with the breakdown of substituents (numbered) used for the generation of computational descriptors. Reproduced with permission.[14] Copyright 2020, RSC Publications.

6.4. Flexible Electronics

The fabrication of polymer-based flexible materials[555-558] is generally contingent on numerous variables such as temperature, humidity, concentration, and processes parameters.[555-557] The integration of data-driven technologies could effectively enhance the functional design, innovative synthesis, characterization, and process optimization in this domain.[16] Zhang et al.[16] employed a differential evolution (DE)-based backpropagation (BP) neural network to determine the electrical properties of Ag/poly amic acid (Ag/PAA) composites with respect to growth conditions (**Figure 38**). Based on the orthogonal analysis, four independent variables were taken as descriptors: the concentrations of PAA and $NaBH_4$, reduction time of $NaBH_4$, and ion exchange time of $AgNO_3$. The final sheet resistance of the Ag/PAA composite film and processing time were set as output parameters. A dataset of 1,077 samples was used for training, while 49 samples are used for validation, resulting in a high accuracy ML model with a prediction error of less than 1.96%. This study proposed a data-driven framework to explore and optimize reaction conditions to boost the material and device design efficiencies.

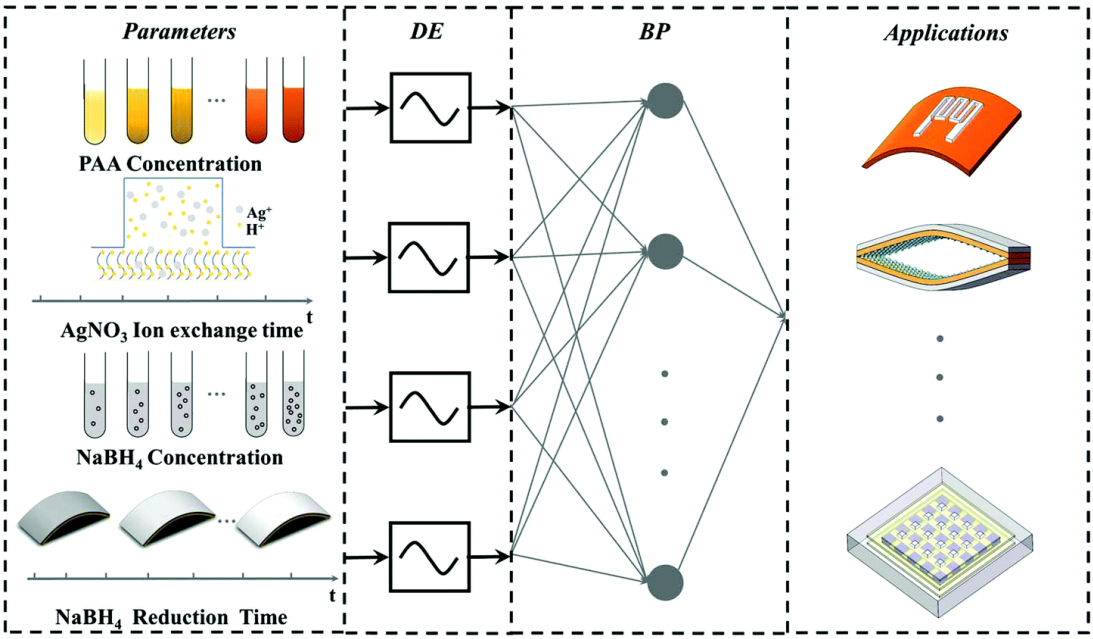


Figure 38. The workflow of the ML-assisted Ag/PAA composites optimization process by the DE-BP model. Reproduced with permission.[16] Copyright 2020, RSC Publications.

6.5. Optoelectronics

The data-driven process powered by ML enables the disentangling of complicated photochemical reactions involved in complex system consisting of multicomponent materials, accelerating the progress of fundamental understanding and rational material design.[386] Conventional trial-and-error methods inhibit the high-throughput screening of novel optoelectronic materials. Innovative 2D materials have received much attention for potential optoelectronic applications.[559, 560] Ma et al.[15] reported a workflow (**Figure 39**a) integrating high-throughput first principle calculations and ML techniques to predict 2D octahedral oxyhalides with enhanced optoelectronic properties. Specifically, through high-throughput quantum chemical computations, the training set was composed of the geometric and electronic descriptors of 300 different octahedral oxyhalides. The implementation of PCA[52] enabled importance evaluation and reduction of the number of descriptors (Figure 39b and 38c), where the proposed distorted stacked octahedral factors exhibit superiority in describing the geometric pattern of the inequitable atoms and critical influence on the bandgap. The high-performance GBR model with the a MSE of 0.086 and R^2 of 0.835 under 10-fold CV was then employed to accelerate the screening of 5000 2D candidates to excavate potential optoelectronic materials. Several 2D optoelectronic octahedral oxyhalides such as $Bi_2Se_2Br_2$, Bi_2Se_2BrI, and $Bi_2Se_2I_2$, were regarded as promising candidates due to their high electron motilities, mild bandgaps, and absorbance coefficients. This study successfully took advantage of data-driven innovation to screen suitable optoelectric candidates with the suitalbe target property values, indicating the effectiveness of a ML model trained on geometric and electronic descriptors. This study also indicated that the selection of appropriate descriptors is significant to improving the performance of ML models.

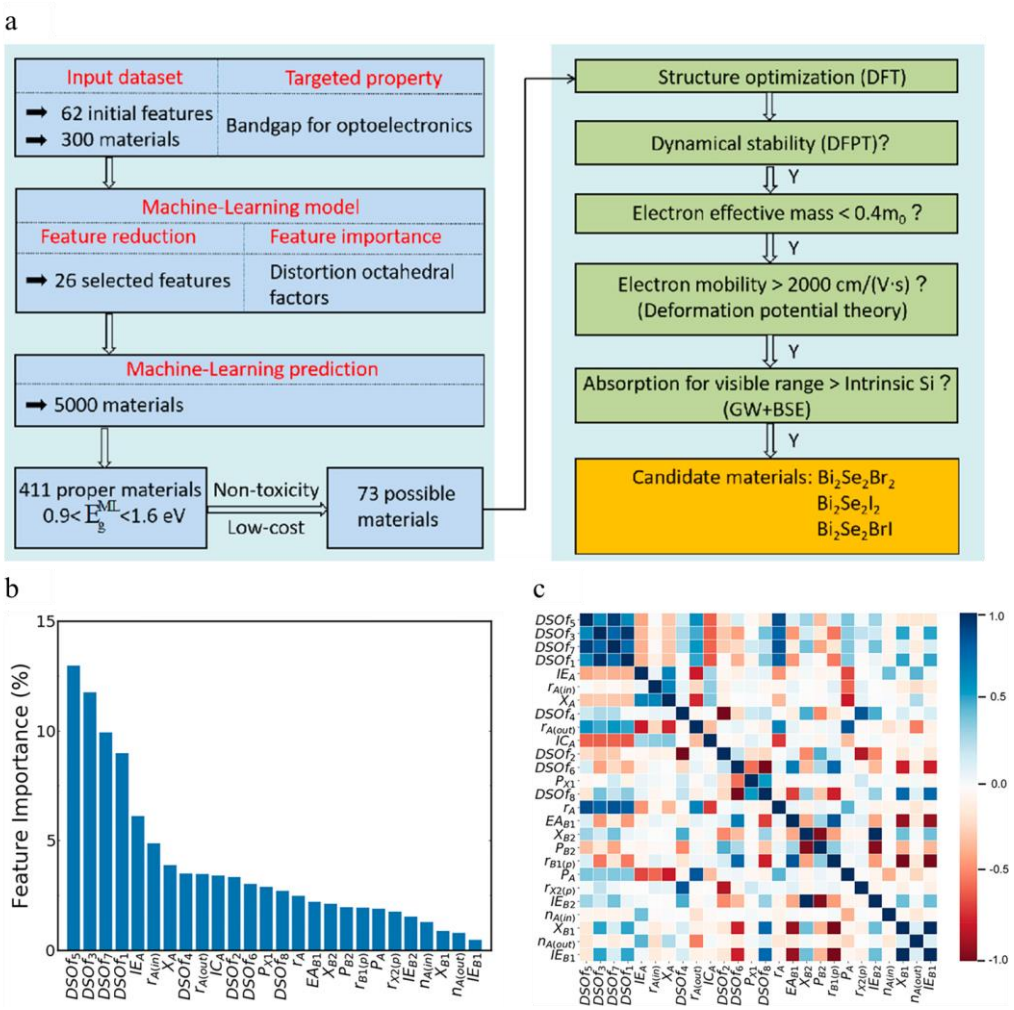


Figure 39. a) The workflow of the property prediction and high-throughput computation. b) The importance of the selected 26 features. c) The inner correlation of the 26 selected features. Reproduced with permission.[15] Copyright 2019, RSC Publications.

6.6. Superconductors

In the field of superconductors, with the integration of ML techniques and the development of relevant databases, several successful predictions have been made indicating the effectiveness of ML techniques in this domain.[561] Stanev et al.[277] proposed a framework combining several ML models to identify appropriate superconductor candidates among 110,000 different compositions in the ICSD, the superconductivity of a large part of which has not been experimentally tested. A classification model was trained to identify whether the critical temperature (T_C) was higher than 10 K. A separate regression model was employed in predicting T_C to enable better understanding of the material properties that determine the T_C. The training data were collected from the SuperCon database, including the chemical composition and T_C. Part of the descriptors used in this study was computed via the Material Agnostic for Informatics and Exploration (Magpie),[33] including elemental property statistics and electronic structure attributes. With the employment of data from AFLOW Online Repositories, additional calculated material properties like DOS and electronic entropy per atom were collected. During the model training stage, it was found that the size of the dataset and

the number of descriptors were critical to the model's performance (**Figure 40**a and 39b). The trained classifier achieved an accuracy of 92% in identifying materials with T_C higher than 10 K and the regressor achieved an R^2 of 0.88 in predicting the T_C for low-T_C, cuprate and iron-based compounds. The framework integrating the classification and regression model reported 35 compounds with a higher than 20 K critical temperature (T_C) (Figure 40d and 39e) for further experimental validation. Part of the reported compounds possess similar chemical and structural properties with cuprate superconductors, indicating that the ML framework can identify the hidden patterns of the training dataset. Additionally, most of the highlighted compounds share a standard peculiar electronic band structure; the energy of the highest occupied electronic state is immediately above the one or more flat or near-flat bands. The related prominent peak of the density of states (DOS) (Figure 40c) can cause significant electronic instability, as one possible approach to achieve high-temperature superconductivity.[562, 563] Further, Konno et al.[404] developed a deep learning model (Figure 40f) with an R^2 of 0.92 in predicting the T_C using only compositions of compounds. Moreover, by utilizing descriptors derived from electronic band structure, Isayev et al.[564] introduced an innovative fingerprint approach that could quickly identify materials, such as superconductors, semiconductors, metals, topological insulators, piezoelectric, and mapping properties bandgap with T_C (Figure 40g and 39h). In thes ML-based superconducting material discovery processes, the critical temperature is one of the most important properties to be investigated, where the structural and electronic property information are highly correlative descriptors

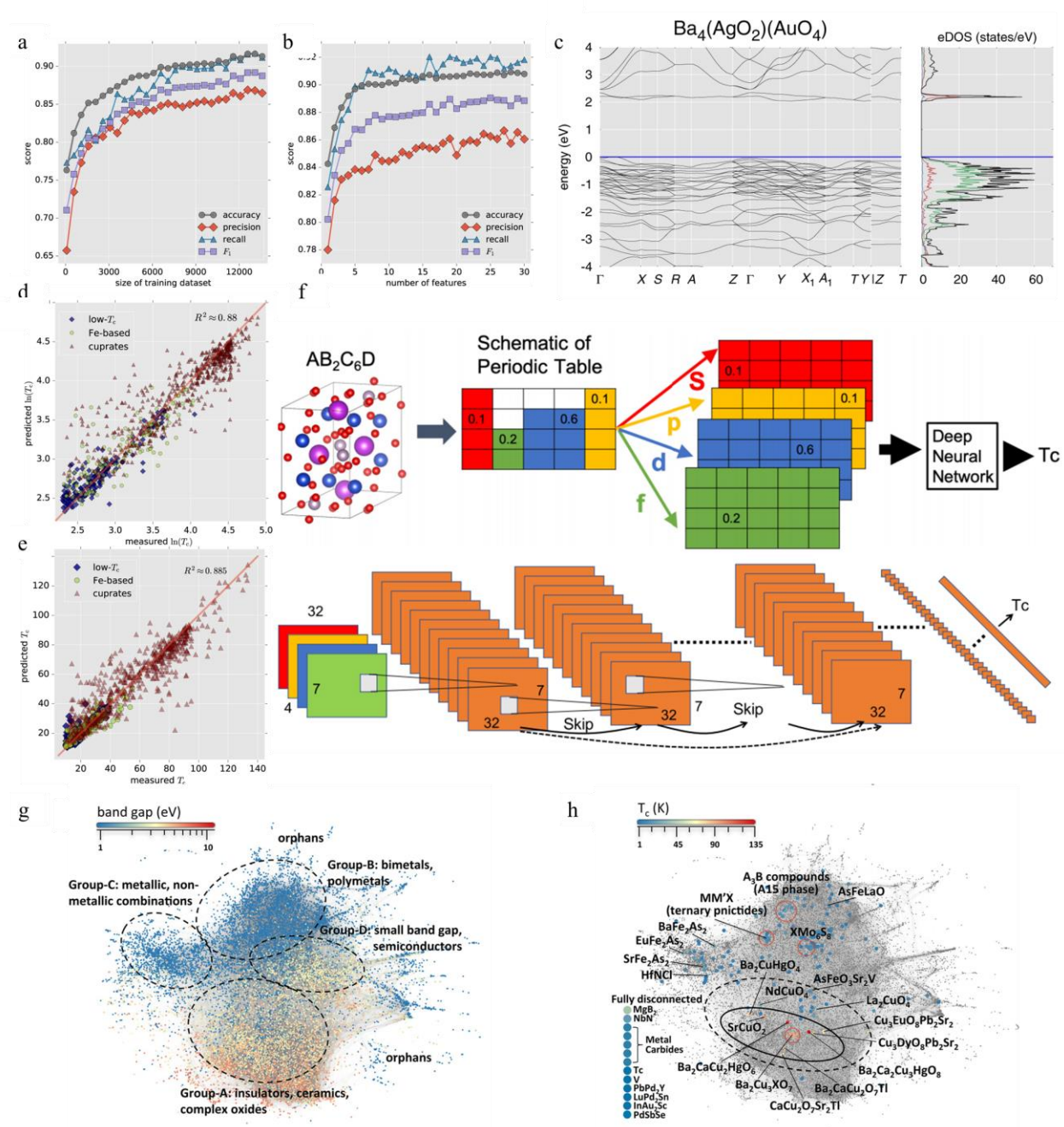


Figure 40. Accuracy, precision, recall, and F1 score as a function of the a) size of the training set with a fixed test set and b) the number of predictors. c) The DOS of $Ba_4(AgO_2)(AuO_4)$, where the ML model identified as potential candidate materials with a critical temperature (T_C) of more than 20K. d) The comparison of predicted and measured ln(T_C) and e) T_C for the general ML model. a-e) Reproduced with permission.[277] Copyright 2018, Springer Nature Publications. f) (Top) the representation of the ML workflow. (Bottom) The schematic of the deep neural network. Reproduced with permission.[404] Copyright 2021, AIP Publications. g) The mapping of the material, bandgap, and h) T_C. Reproduced with permission.[564] Copyright 2015, ACS Publications.

6.7. Metallic Glasses

Recently, data-driven researche has been conducted to discover metallic glass and investigate the glass-forming ability.[33, 565-568] Although metallic glass has various unique properties such as soft magnetism and high wear resistance, metallic glass systems usually possess a special composition based on empirical rules and extensive experimentation.[569] The investigation of metallic glass formation, structure, and properties has attracted much attention due to its fundamental scientific importance and potential for further applications.[570, 571] For instance, by using reported experimental data to establish an ML model, Ward et al. [33] proposed a data-driven framework to accelerate the discovery of novel alloys and predict glass formation ability. The training set was collected from 'Nonequilibrium Phase Diagram of Ternary Amorphous Alloys',[572] which comes from hundreds of ternary phase diagrams containing the potential for glass formation based on many experiments conducted at thousands of compositions. The authors selected 5,396 distinct compositions with an amorphous ribbon forming ability and evaluated them by melt spinning. For a single composition, if at least one measurement indicated the potential to form an utterly amorphous sample, it was assumed that it may form a metallic glass. With the implementation of the described screening process, 70.8% of the entries in the training set were found to be consistent with metallic glasses. A total of 145 descriptors in four categories, namely, the stoichiometric descriptors, elemental property descriptors, electronic structure descriptors,[573] and ionic compound descriptors, was used to construct the descriptor space. A random forest classifier[189] was applied to classify the material into two classes with respect to the calculated possibilities for glass formation. Those materials whose predicted glass formation probability was higher than 50% are considered positive predictors of glass formation, while others are considered negative. For the ML model's testing process, the glass formation ability of the Al-Ni-Zr ternary system was well-matched with the literature data (**Figure 41**a and 40b), indicating the ML model could precisely pinpoint the desired ideal compositions in yet-unassessed alloy systems. The data-driven framework (Figure 41c) designed by Ward et al[405] could accurately predict the critical properties of candidate bulk metallic glasses, incluing the existence of an amorphous state, the critical casting diameter (D_{max}), and the supercooled liquid range (ΔT_x) (Figure 41e). The only input dataset for the ML model was the materials' compositions, which consisted of over 8000 metallic glasses experiments. The trained ML models were implemented to optimize the properties of existing commercial alloys and discover novel compositions for forming metallic glasses with enhanced properties. These two cases present data-driven frameworks for the design and identification of innovative bull metallic glasses. Both of the two applied ML-based workflow employed the stoichiometric descriptors to discover innovative bull metallic glasses.

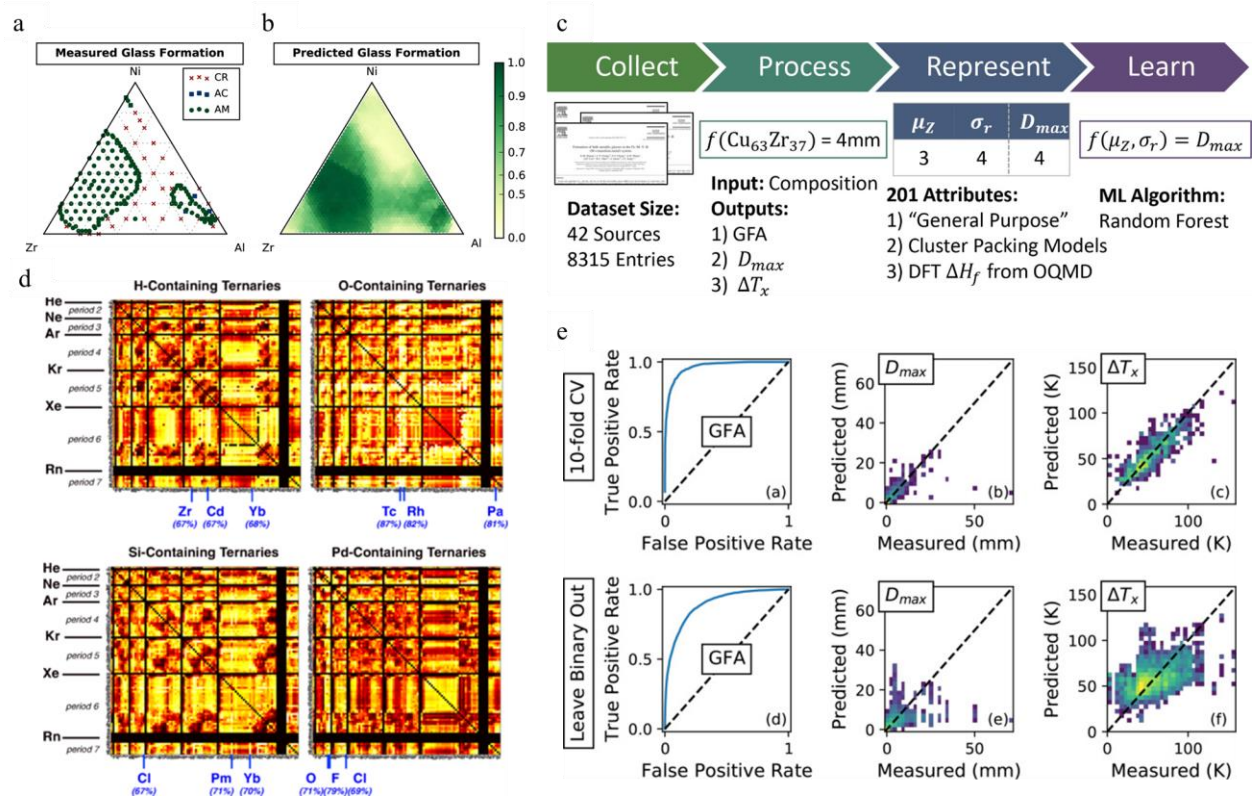


Figure 41. a) The experimental measurements and b) ML prediction of metallic glass-forming ability. Reproduced with permission.[33] Copyright 2016, Springer Nature Publications. c) The schematic of the construction of a data-driven framework for the prediction of metallic glass properties. Reproduced with permission.[405] Copyright 2018, Elsevier Publications. d) The heat map of ternary compositions' stability ranking. Reproduced with permission.[573] Copyright 2014, AIP Publications. e) The performance of ML models on the prediction of glass formation ability, critical casting diameter, and critical temperature. Reproduced with permission.[405] Copyright 2018, Elsevier Publications.

6.8. Magnetic Materials

Data-driven approaches have been employed to predict the ordering temperature and classify different types of magnets.[406, 574] Ferromagnetic materials (FM) have a broad spectrum of applications in spintronic, including logic, memory, and sensing, while the emerging antiferromagnetic (AFM) has also attracted intense attention.[575] To predict the Curie temperature (TC) of the FM and FM/AFM classification model, Long and co-workers [406] performed an RF model (Figure 42a and 41b) according to TC and magnetic ground state. The 1,749 FM and 1,056 AFM compounds were collected from the experimental AtomWork database,[576], consisting of structures and properties of magnetic materials. To present the structure information, such as the coordination and the distance between atoms, chemical descriptors were computed by the Material Agnostic for Informatics and Exploration (Magpie);[33] the smooth overlap of atomic position (SOAP) descriptor was selected.[577] The volume of the unit cell and the space group number were also taken as structural descriptors. Finally, 139 chemical and 26 structural descriptors adopted each compound. Similar, Möller et al.[407] utilized an ML model to identify the optimal chemical composition of novel permanent magnets. Specifically, a kernel-based SVR algorithm was applied to establish a data-

driven framework for accurately predicting properties such as uniaxial magneto (K_1), the magnetization ($\mu_0 M$), and the relative phase stability energy (E_f) (Figure 42c). Soft magnetic materials are widely employed in plenty of electromagnetic distribution, generation, and conversion devices such as converters, transformers, inductors, generators, and sensors[578-580] due to the advantage of being able to rapidly switch their magnetic polarization under a relatively small magnetic field. For example, Wang et al.[408] utilized data-driven approaches (Figure 42d) to boost novel soft magnetic materials' design and discovery.

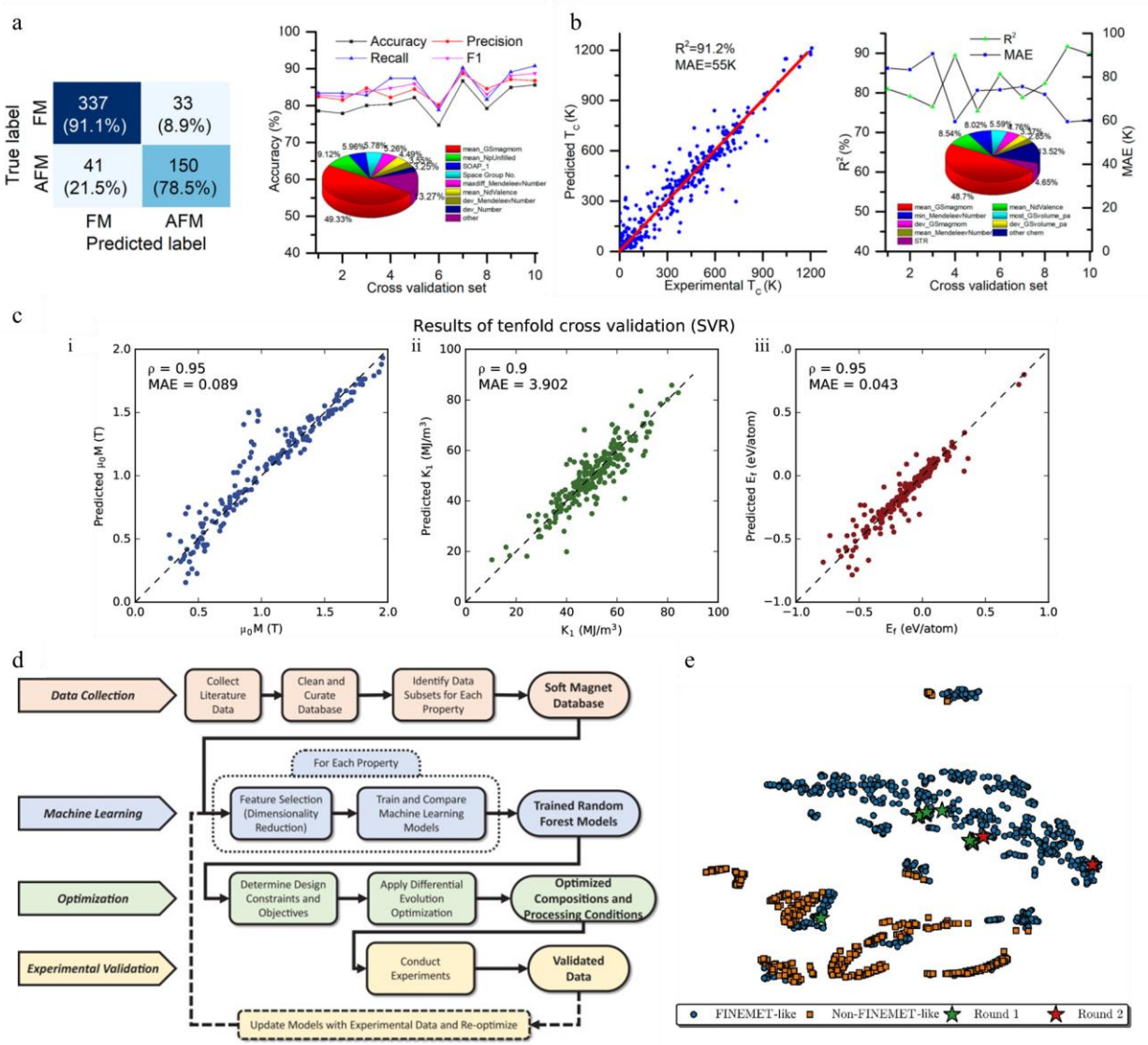


Figure 42. a) The model performance and feature importance evaluation of the classification model and b) regression model. Reproduced with permission.[406] Copyright 2021, Taylor and Francis Ltd. Publications. c) The 10-fold cross-validation of the SVR models for the prediction of (i) magnetization $\mu_0 M$ (ii) K_1, and (iii) the relative phase stability energy E_f. Reproduced with permission.[407] Copyright 2018, Elsevier Publications. d) The workflow of the material design process based on the integration of ML and conventional experiments. e) The t-SNE visualization to compare the optimized alloys predicted by ML to the alloys published literature. Reproduced with permission.[408] Copyright 2020, Elsevier Publications.

Their data-driven framework involves establishing the experimental database, employing ML models, identifying the trends of magnetic properties, and predicting/ optimizing the next-generation design of soft magnetic materials. They propose the target material system, FINEMET-type soft magnetic nanocrystalline alloys,[581] with two classes of properties: extrinsic and intrinsic properties.[582, 583] The Pearson correlation coefficient is combined with the GBDT algorithm and applied to simplify the model, while avoiding a significant loss of information. Five different ML models, LR, SVM, DTs, *k*-NN, and RF[189], are implemented to accurately predict properties, including magnetic saturation (B_S), coercivity (H_C), and magnetostriction (λ). A stochastic optimization model is applied for the exploration and optimization of those target properties. Based on these predictive models, several optimized soft magnetic materials specified for composition and heat treatment conditions were predicted, prepared, and validated (Figure 42e), exhibiting excellent agreement between experiments and predictions, which verified the reliability of the established data-driven model.

6.9. Materials Thermodynamic Stability Prediction

To accelerate material design and discovery with the assistance of ML, thermodynamic stability of materials is considered as the first essential requirement, such as for the discovery of perovskites,[584] two-dimensional materials,[67] and alloys[585]. Then, the further screening of the properties would meet different needsspecific to the application. The driving force for discovering new materials still relies on the DFT calculations, and the synthesis of new materials often faces more challenges in experiments. Researchers performed the data-driven design of materials through theoretical computing methods, which integrate the databases, ML algorithm, and high-throughput DFT calculations. Unexplored materials still need to be discovered and designed through data-driven strategies, such as alloys,[585] heterostructures,[586] 2D materials[587] and even doped materials.[10] Data-driven studies of thermodynamic stability have important implications for the identification of novel materials, either as detailed complex features or simpler scalar features.

For the descriptors of thermodynamic properties, the current focus is mainly on the geometric descriptors related to the material's crystal structure. Goodall et al.[588] used only the stoichiometry of the material as a descriptor to train the machine and employed automatic learning to improve the descriptor to solve the problem of predicting material properties in the absence of a known crystal structure. Choudhary et al.[589] used deep transfer learning to employ large DFT calculation datasets (such as the Open Quantum Materials Database (OQMD)[590]) together with other smaller DFT calculation datasets and reasonable experimental results to establish accurate prediction models for the formation energy of the material. Bartel et al.[591] tested the stability predictions of seven ML models using the material project database and the DFT calculations of 85,014 unique chemical components. The error generated by the ML model is a fundamental error such that the ML model does not have duplicate beneficial error elimination, which hinders the accurate prediction of material stability. Therefore, the accuracy of the existing ML model to predict formation energies is dependent on the accuracy of the DFT.

Usually, thermodynamic stability is defined as the total energy and the energy above the convex hull. Schleder et al.[592] investigated the thermodynamic stability of 2D materials in a computational 2D materials database (C2DB) with the assistance of the SISSO. The descriptor of the

training machine did not have clear information about the position of the atom, and the stability of the 2D material in the database can only be accurately classified only relying on the prototype structure. The structure in which former "chemical intuition" was abandoned was once again discovered by Faber et. al.[593] The development of ML for the study of energy formation consisting of all possible elpasolites was successfully used for the discovery of stable as well as unconventional chemistries. Legrain et al.[272] only used descriptors based on the chemical formula of 300 compounds as training sets to train an ML model. The prediction of F_{vib} and S_{vib} descriptors based on the chemical composition for small training sets is outperformed by some of the more detailed descriptors for more extensive training sets. Gibbs Energy determines the equilibrium conditions of chemical reactions and material stability. For inorganic compounds, G is critical for predicting the synthesizability and stability of materials, especially for thermoelectric materials, photothermal materials, fuel cells, and other applications which concern the temperature-dependent stability of materials. Bartel et al.[594] accurately predicted Gibbs energy by applying the SISSO approach and adopted the temperature, atomic mass, and (calculated) atomic volume of materials as primary descriptors to train a model. Although thermodynamic stability is the key criterion for high-throughput computational screening of materials to predict the possibility of synthesis of specific material, the interaction of thermodynamics with several other measures, such as kinetics and non-equilibrium process conditions, have a more significant impact on the synthesizability of materials.

7. Conclusion and Perspectives

Data-driven material innovation shows excellent potential for the rational design and discovery of materials in terms of efficiency, accuracy, and intelligence. In a data-driven material innovation process, the data are the foundation, ML algorithm is the core, descriptor transfers the information, and framework integrates these disciplines to implement innovative applications. In this review, the recent advances in data-driven innovation of materials science are elaborated. First, several data-driven frameworks, along with direct design, inverse design, and active learning, are discussed based on the flow of data and information in the data-driven process. Then, the frequently employed ML algorithms and the relevant data-processing strategies are reviewed in terms of information extraction from data. The chemical databases that store and manage material data and the related digital tools are systematically discussed. Furthermore, the molecular descriptors that carry the chemical information in the data-driven process are introduced. Finally, a critical discussion on how the data-driven approach is applied for various materials is provided. The development of novel and intelligent algorithms, the capability of computational and experimental material databases to generate and store data, and the design and validation of accurate and efficient descriptors have many outcomes. Their synergistic integration is promising and effective for innovative material discovery.

Although considerable progress has been made over the last several decades, the research direction in the field of materials science is shifting into a novel paradigm of data-driven science. Here, we present certain challenges and perspectives with the objective to understand the research and development in the relevant fields.

(i) In addition to the establishment of structure-property relations and material discovery, data-driven techniques could be employed in materials science in the form of autonomous laboratories

for chemical synthesis.[67] The ML-enabled self-guided experimentation that integrates automation experimental platforms and artificial intelligence is becoming the next-generation facility.[302, 595] In particular, ML techniques play a vital role in determining the variations in material properties based on the changes in macroscopic parameters, including reaction conditions and operation parameters, enabling improved process-property relation fitting.[596, 597] The automated laboratories empowered by ML have the potential to substantially boost the material discovery process with the integration of automated platforms[300] or robotics [302], which fully embrace the vision of autonomous laboratories. With the implementation of data-driven strategies, the traditional experiments are expected to be performed without the supervision of humans.[67] For example, Burger et al.[302] developed a mobile robot that automatically conducted 688 experiments for searching photocatalysts by using a Bayesian search algorithm. Angelone et al.[300] proposed the "Chemputer," a universally programmable chemical synthesis machine that can perform 17 different reactions using one platform architecture. The combination of data-driven techniques and automated laboratories is expected to significantly boost innovative material discovery and provide more opportunities in material synthesis with high productivity and quality.

(ii) Data-driven innovation can accelerate material discovery. However, ML techniques are not panaceas that can solve all problems in material discovery without domain knowledge.[67] Implementing a complete data-driven process with critical stages, such as data-preprocessing,[50] descriptor generation,[33] ML-model deployment,[44] uncertainty quantification, and domain applicability, is laborious.[44, 69] Performing ML-centered data-driven research can be challenging for material scientists with limited background in computations.[44] Therefore, achieving best practice of ML employment is a significant stage in the data-driven paradigm of materials science, in which a systematic methodology or ecosystem that unifies the material science community with a consistent interface may guarantee reliability and reproducibility of the trained ML models.[69] Several efforts have been made to develop general ML-centric frameworks for a broad range of materials. For example, Ward et al.[33] proposed a general-purpose ML methodology for predicting properties of inorganic materials. Wang et al.[44] reported broad guidelines on ML-model deployment, domain applicability, and model persistence. Moreover, the automation of the ML workflow and related tools, such as ChemML[311, 312], MAST-ML[313], TPOT,[18, 454] and automatminer,[598] are receiving attention. General guidelines and tools for data-driven techniques are necessary and should be investigated to reduce the difficulties for chemistry and material scientists. We envision that such a general systematic infrastructure can benefit material scientists in efficiently constructing automated and accurate ML workflows to solve their specific material problems; this is critical and helpful for the wider applications of ML in the data-driven innovation of broader materials.

(iii) The implementation of data-driven material innovation is based on frameworks. Multiple material properties may have to be predicted, and therefore, novel data-driven frameworks are essential. Multitask prediction is suitable for the scenario where multiple properties are excepted to be predicted. ML models such as the SISSO and atoms-in-molecules neural network have been reported for implementation in the prediction of several related properties simultaneously.[420, 599] The form of material data varies in terms of numerical values, images, texts, and graphs, and one research project could have several data sources and data modalities. In the ML community, multimodal ML[600] and transfer learning are applied to process and relate information from multiple

modalities and fidelities.[601] The transfer-learning model can combine both low- and high-fidelity data to make high-fidelity predictions at low computation costs.

(iv) Algorithms are the core of the data-driven approach to material innovation. Researchers with different backgrounds, such as in chemistry, materials science, and computer science, tend to choose different strategies to show the novelty of their study. Researchers in the field of chemistry and materials science tend to employ existing algorithms and models to solve their problems, whereas computer scientists focus on improving algorithms to enhance the performance. Although the combination of common algorithms and models can provide good solutions in the current studies on chemistry and materials science, it is not as revolutionary as the development of new algorithms. To introduce existing algorithms and their improved version from computer science should be the preferred strategy to study chemistry and materials science in the future, and fast and surrogate algorithms receive the most attention. For example, graph networks that are based on the structure of materials and support relational reasoning and combinatorial generalization are promising for high-fidelity learning. In addition to learning from material structures directly, algorithms for analyzing experimental characterization data, such as XAS and STEM, are critical and could significantly promote the utilization of high-modality data in the future. General algorithms still cannot fulfill the demand of materials science, because most of the algorithms are mathematics–based, whereas the materials science problems usually have hidden physical laws, indicating that various algorithms can be used in parallel to solve the same problem and the algorithm with the best performance can be selected.[67] Therefore, the rational design and selection of algorithms that are specific for materials science can help scientists have a better understanding of materials science problems and save computing resources.

(v) Descriptors are the bridge in a data-driven process for transferring information between humans and machines, and new combinational descriptors for ML must be further developed and evaluated. Many descriptors with high relevance have been developed. However, these descriptors are not universal and often solve the same type of or similar prediction problems. At present, the interaction between experimental and simulated descriptors is also incompatible. A possible solution to the compatibility problem is the use of active learning, robotics-assisted high-throughput experiments, and artificial intelligence approaches, which will enable the development of new combinational descriptors. As the critical input for a data-driven workflow, desired descriptors should ensure uniqueness and carry as much related information as possible. Although the current insufficiency of database descriptors and the inaccessibility of descriptors for small-sample datasets hinder the development of descriptors, an issue that will be improved with the data blowing out in materials science. To match the advancement of the algorithms and databases mentioned above, descriptors which could economically be acquired from database or computed and precisely present complex nature are desired to be developed. Further, it is critical to develop methodology and theory to derive outperformed descriptors in various approaches and integrate multiple modalities of data to present materials more accurately and efficiently for implementing advanced algorithms, which is one of the most challenging and demanding tasks in data-driven material innovation in the future.

(vi) Data is the foundation of data-driven material innovation, and therefore, reliable and sufficient data sources are critical. Although data-driven material innovation is widely and extensively explored by the scientific community and several databases have been established and

used in materials science, challenges in data acquisition still remain: (1) representative experimental databases (such as the ICSD and SciFinder) still needs a license for access;[345, 602] (2) no unified application programming interface is available for connecting different databases and other software, although the python-based RESTful API has been widely adopted;[233, 234, 270] (3) the primary challenge in choosing and comparing databases is identifying the specific function for the databases' difference and determining the equivalency for the same structure in various databases.[218]

Databases can be used to solve one specific problem by relaying the specific descriptors that are extracted from the selected databases using the appropriate high-throughput tools and workflow management frameworks. We can use this strategy to explore new materials, new structures, and new properties by utilizing meaningful information and patterns. The strategy can also be employed to synthesize the specific materials for the specific applications mentioned in previous sections. The data-driven strategies can be used to uncover complexities and design novel materials with excellent properties based on the powerful and accessible databases, high-throughput tools, and workflow management frameworks. For the modern materials science community, data-driven strategies show considerable potential for the future. The development of the databases and related tools is impossible by using the trial-and-error methods in traditional approaches. Moreover, advanced algorithms in materials science are generally restricted because of the lack of sufficiently diverse and extensive databases. In addition to the data generated by performing simulations and computations, high-throughput experiments conducted by using modular robotic systems[300], mobile robots,[302] and automated platforms for programmable material screening and synthesis could be an alternative and critical approach for database construction in the future.

In summary, data-driven research is expected to rapidly expand and progress in the future, thereby accelerating material innovations to bridge the gap between science and technology and facilitating a rapid development of emerging advanced materials.

Acknowledgments

We acknowledge the support from the ANU Futures Scheme (Q4601024). We also gratefully express gratitude to all parties which have contributed towards the success of this project, both financially and technically, especially the S&T Innovation 2025 Major Special Programme (grant number 2018B10022) funded by the Ningbo Science and Technology Bureau, China, as well as the Provincial Key Laboratory Programme (2020E10018) funded by the Zhejiang Provincial Department of Science and Technology. We also appreciate the support from the Functional Materials Interfaces Genome (FIG) project.

Conflict of Interest

The authors declare no conflict of interest.

Abbreviations

Terms	Explanations
ANN	Artificial Neural Network
CRR	Carbon Reduction Reaction
CV	Cross Validation
DFT	Density Functional Theory
DNN	Deep Neural Network
DT	Decision Tree
GBDT	Gradient Boosting Decision Tree
GBR	Gradient Boosting Regression
GBRT	Gradient Boosting Regression Tree
GPR	Gaussian Process Regression
HER	Hydrogen Evolution Reaction
HOIP	Hybrid Organic-Inorganic Perovskites
ICSD	Inorganic Crystal Structure Database
k-NN	*k*-Nearest Neighbor
KPI	Key Performance Indicator
KRR	Kernel Ridge Regression
LASSO	Least Absolute Shrinkage and Selection Operator
LOOCV	Leave One Out Cross Validation
LR	Linear Regression
MAE	Mean Absolute Error
ML	Machine Learning
MLR	Multi-Linear Regression
MP	Materials Project
MSE	Mean Square Error
NNP	Neural Network Potential
NRR	Nitrogen Reduction Reaction
OER	Oxygen Evolution Reaction
OPV	Organic Photovoltaics
OQMD	Open Quantum Materials Database
ORR	Oxygen Reduction Reaction
PCA	Principal Component Analysis
PV	Photovoltaics
QSAR	Quantitative Structure Activity Relationship
QSPR	Quantitative Structure Property Relationship

RF	Random Forest
RMSE	Root Mean Square Error
SISSO	Sure Independent Screening and Sparsifying Operator
SVM	Support Vector Machine
SVR	Support Vector Regression
AAPL	Automatic Anharmonic Phonon Library
ADES	Automation, Data, Environment, And Sharing
AFM	Antiferromagnetic
AiiDA	Automated Interactive Infrastructure And Database
AIMD	*Ab Initio* Molecular Dynamics
AML	Adaptive Machine Learning
ANN	Artificial Neural Network
AOP	Advanced Oxidation Processes
API	Application Programming Interface
ASE	Atomic Simulation Environment
AUC	The Area Under The Curve
BO	Bayesian Optimization
BP	Back Propagation
C2DB	Computational 2D Materials Database
CART	Classification And Regression Tree
CBM	Conduction Band Bottom
CGCNN	Crystal Graph-Based Convolutional Neural Network
CMR	Computational Materials Repository
CNN	Convolutional Neural Network
COD	Crystallography Open Database
CRM	Cluster Ranking Model
CRR	Carbon Reduction Reaction
CSD	Cambridge Structural Database
CV	Cross Validation
CVD	Chemical Vapor Deposition
DE	Differential Evolution
DFT	Density Functional Theory
DMSC	Dual-Metal-Site Catalysts
DNN	Deep Neural Network
DOS	Density of States
DT	Decision Tree
EA	Electron Affinity

EELS	Electron Energy Loss
EM	Expectation Maximization
FM	Ferromagnetic
FN	False Negative
FP	False Positive
FPR	False Positive Rate
FWS	Fireworks
GA	Genetic Algorithm
GAN	Generative Adversarial Networks
GBDT	Gradient Boosting Decision Tree
GBR	Gradient Boosting Regression
GBRT	Gradient Boosting Regression Tree
GCLP	Grand Canonical Linear Programming
GPR	Gaussian Process Regression
HER	Hydrogen Evolution Reaction
HOIP	Hybrid Organic-Inorganic Perovskites
HT	High Throughput
HTTP	Hyper Text Transfer Protocol
HTVS	High-Throughput Virtual Screening
ICSD	Inorganic Crystal Structure Database
ID3	Iterative Dichotomiser 3
IP	Ionization Potential
k-NN	*K*-Nearest Neighbor
KPI	Key Performance Indicator
KRR	Kernel Ridge Regression
LASSO	Least Absolute Shrinkage And Selection Operator
LOOCV	Leave One Out Cross Validation
LR	Linear Regression
MAE	Mean Absolute Error
MAP	Maximum Posteriori
MAST-ML	Materials Simulation Toolkit For Machine Learning
MDF	Materials Data Facility
MGI	Material Genome Initiative
MIV	Mean Impact Value
ML	Machine Learning
MLPNN	Multi-Layer Perceptron Neural Network
MLR	Multi-Linear Regression

MP	Materials Project
MPB	Morphotropic Phase Boundary
MQCBO	Multiple Quality Constraint Bayesian Optimization
MSE	Mean Square Error
NLP	Natural Language Processing
NNP	Neural Network Potential
NOMAD	Novel Material Discovery
NRR	Nitrogen Reduction Reaction
OER	Oxygen Evolution Reaction
OPTIMADE	Open Databases Integration For Materials Design
OPV	Organic Photovoltaics
OQMD	Open Quantum Materials Database
ORR	Oxygen Reduction Reaction
PCA	Principal Component Analysis
PCE	Power Conversion Efficiency
PES	Potential Energy Surface
PF	Power Factor
PSO	Particle Swarm Optimization
PV	Photovoltaics
QSAR	Quantitative Structure Activity Relationship
QSPR	Quantitative Structure Property Relationship
RBF	Radial Basis Function
RBM	Restricted Boltzmann Machine
REST	Representational State Transfer
RF	Random Forest
RL	Reinforcement Learning
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
SAA	Simulated Annealing Algorithm
SISSO	Sure Independent Screening And Sparsifying Operator
SQCBO	Single Quality Constraint Bayesian Optimization
STEM	Scanning Transmission Electron Microscopy
SVM	Support Vector Machine
SVR	Support Vector Regression
TEM	Transmission Electron Microscopy
TL	Transfer Learning

TN	True Negative
TP	True Positive
TPR	True Positive Rate
t-SNE	t-Distributed Stochastic Neighbor Embedding
VAE	Variational Autoencoders
VASP	Vienna Ab-Initio Simulation Package
VBM	Valence Band Top
WGAN	Wasserstein Generative Adversarial Network
XAS	X-Ray Absorption Spectra

Reference

[1]A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu, D. Hassabis, Nature 2020, 577, 706.

[2]D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, D. Hassabis, Science 2018, 362, 1140.

[3]O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M. Kroiss, I. Danihelka, A. Huang, L. Sifre, T. Cai, J. P. Agapiou, M. Jaderberg, A. S. Vezhnevets, R. Leblond, T. Pohlen, V. Dalibard, D. Budden, Y. Sulsky, J. Molloy, T. L. Paine, C. Gulcehre, Z. Wang, T. Pfaff, Y. Wu, R. Ring, D. Yogatama, D. Wünsch, K. McKinney, O. Smith, T. Schaul, T. Lillicrap, K. Kavukcuoglu, D. Hassabis, C. Apps, D. Silver, Nature 2019, 575, 350.

[4]J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, T. Lillicrap, D. Silver, Nature 2020, 588, 604.

[5]G. R. Schleder, A. C. M. Padilha, C. M. Acosta, M. Costa, A. Fazzio, Journal of Physics: Materials 2019, 2, 032001.

[6]J. J. de Pablo, B. Jones, C. L. Kovacs, V. Ozolins, A. P. Ramirez, Current Opinion in Solid State and Materials Science 2014, 18, 99.

[7]M. L. Green, C. L. Choi, J. R. Hattrick-Simpers, A. M. Joshi, I. Takeuchi, S. C. Barron, E. Campo, T. Chiang, S. Empedocles, J. M. Gregoire, A. G. Kusne, J. Martin, A. Mehta, K. Persson, Z. Trautt, J. Van Duren, A. Zakutayev, Applied Physics Reviews 2017, 4, 011105.

[8]B. Sun, M. Fernandez, A. S. Barnard, Nanoscale Horizons 2016, 1, 89.

[9]A. Chen, X. Zhang, Z. Zhou, InfoMat 2020, 2, 553.

[10]R. B. Wexler, J. M. P. Martirez, A. M. Rappe, Journal of the American Chemical Society 2018, 140, 4678.

[11]J. Zhang, P. Hu, H. Wang, The Journal of Physical Chemistry C 2020, 124, 10483.

[12]S. Back, K. Tran, Z. W. Ulissi, ACS Catalysis 2019, 9, 7651.

[13]K. Tran, Z. W. Ulissi, Nature Catalysis 2018, 1, 696.

[14]J. S. Dondapati, A. Chen, Physical Chemistry Chemical Physics 2020, 22, 8878.

[15]X.-Y. Ma, J. P. Lewis, Q.-B. Yan, G. Su, The Journal of Physical Chemistry Letters 2019, 10, 6734.

[16]M. Zhang, J. Li, L. Kang, N. Zhang, C. Huang, Y. He, M. Hu, X. Zhou, J. Zhang, Nanoscale 2020, 12, 3988.

[17]H. Jin, H. Zhang, J. Li, T. Wang, L. Wan, H. Guo, Y. Wei, The Journal of Physical Chemistry Letters 2020, 11, 3075.

[18]M. Zhong, K. Tran, Y. Min, C. Wang, Z. Wang, C.-T. Dinh, P. De Luna, Z. Yu, A. S. Rasouli, P. Brodersen, S. Sun, O. Voznyy, C.-S. Tan, M. Askerka, F. Che, M. Liu, A. Seifitokaldani, Y. Pang, S.-C. Lo, A. Ip, Z. Ulissi, E. H. Sargent, Nature 2020, 581, 178.

[19]M. Rück, B. Garlyyev, F. Mayr, A. S. Bandarenka, A. Gagliardi, The Journal of Physical Chemistry Letters 2020, 11, 1773.

[20]Z. W. Ulissi, M. T. Tang, J. Xiao, X. Liu, D. A. Torelli, M. Karamad, K. Cummins, C. Hahn, N. S. Lewis, T. F. Jaramillo, K. Chan, J. K. Nørskov, ACS Catalysis 2017, 7, 6600.

[21]D. C. Elton, Z. Boukouvalas, M. D. Fuge, P. W. Chung, Molecular Systems Design & Engineering 2019, 4, 828.

[22]C.-H. Lee, A. Khan, D. Luo, T. P. Santos, C. Shi, B. E. Janicek, S. Kang, W. Zhu, N. A. Sobh, A. Schleife, B. K. Clark, P. Y. Huang, Nano Letters 2020, 20, 3369.

[23]S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, C. Wolverton, npj Computational Materials 2015, 1, 15010.

[24]A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, Apl Materials 2013, 1, 011002.

[25]C. Chen, W. Ye, Y. Zuo, C. Zheng, S. P. Ong, Chemistry of Materials 2019, 31, 3564.

[26]R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, L. M. Ghiringhelli, Physical Review Materials 2018, 2, 083802.

[27]A. Zunger, Nature Reviews Chemistry 2018, 2, 0121.

[28]G. Hautier, A. Jain, S. P. Ong, Journal of Materials Science 2012, 47, 7317.

[29]Y. Liu, T. Zhao, W. Ju, S. Shi, Journal of Materiomics 2017, 3, 159.

[30]C. Chen, Y. Zuo, W. Ye, X. Li, Z. Deng, S. P. Ong, Advanced Energy Materials 2020, 10, 1903242.

[31]G. R. Schleder, C. M. Acosta, A. Fazzio, ACS Applied Materials & Interfaces 2020, 12, 20149.

[32]K. M. Jablonka, D. Ongari, S. M. Moosavi, B. Smit, Chemical Reviews 2020, 120, 8066.

[33]L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, npj Computational Materials 2016, 2, 16028.

[34]B. Sanchez-Lengeling, A. Aspuru-Guzik, Science 2018, 361, 360.

[35]D. P. Tabor, L. M. Roch, S. K. Saikin, C. Kreisbeck, D. Sheberla, J. H. Montoya, S. Dwaraknath, M. Aykol, C. Ortiz, H. Tribukait, C. Amador-Bedolla, C. J. Brabec, B. Maruyama, K. A. Persson, A. Aspuru-Guzik, Nature Reviews Materials 2018, 3, 5.

[36]S. Lu, Q. Zhou, Y. Ouyang, Y. Guo, Q. Li, J. Wang, Nature Communications 2018, 9, 3405.

[37]R. Yuan, Z. Liu, P. V. Balachandran, D. Xue, Y. Zhou, X. Ding, J. Sun, D. Xue, T. Lookman, Advanced Materials 2018, 30, 1702884.

[38]G. H. Gu, J. Noh, I. Kim, Y. Jung, Journal of Materials Chemistry A 2019, 7, 17096.

[39]T. D. Sparks, S. K. Kauwe, M. E. Parry, A. M. Tehrani, J. Brgoch, Annual Review of Materials Research 2020, 50, 27.

[40]M. M. Cencer, J. S. Moore, R. S. Assary, Polymer International 2021, n/a.

[41]F. Strieth-Kalthoff, F. Sandfort, M. H. S. Segler, F. Glorius, Chemical Society Reviews 2020, 49, 6154.

[42]S. Grazulis, D. Chateigner, R. T. Downs, A. F. Yokochi, M. Quiros, L. Lutterotti, E. Manakova, J. Butkus, P. Moeck, A. Le Bail, J Appl Crystallogr 2009, 42, 726.

[43]C. W. Yap, Journal of Computational Chemistry 2011, 32, 1466.

[44]A. Y.-T. Wang, R. J. Murdock, S. K. Kauwe, A. O. Oliynyk, A. Gurlo, J. Brgoch, K. A. Persson, T. D. Sparks, Chemistry of Materials 2020, 32, 4954.

[45]B.-E. Liu, W. Yu, Chinese Journal of Polymer Science 2020, 38, 908.

[46]A. S. Barnard, Matter 2020, 3, 22.

[47]S. B. Kotsiantis;, D. Kanellopoulos;, P. E. Pintelas, Zenodo 2007.

[48]D. Padula, J. D. Simpson, A. Troisi, Materials Horizons 2019, 6, 343.

[49]H. Sahu, F. Yang, X. Ye, J. Ma, W. Fang, H. Ma, Journal of Materials Chemistry A 2019, 7, 17480.

[50]P. Pankajakshan, S. Sanyal, O. E. de Noord, I. Bhattacharya, A. Bhattacharyya, U. Waghmare, Chemistry of Materials 2017, 29, 4190.

[51]N. C. Frey, D. Akinwande, D. Jariwala, V. B. Shenoy, ACS Nano 2020, 14, 13406.

[52]P. M. Shenai, Z. Xu, Y. Zhao, in *Principal component analysis-engineering applications*, IntechOpen, 2012, 25.

[53]L. v. d. Maaten, G. Hinton, Journal of Machine Learning Research 2008, 9, 2579.

[54]A. Yosipof, O. E. Nahum, A. Y. Anderson, H.-N. Barad, A. Zaban, H. Senderowitz, Molecular Informatics 2015, 34, 367.

[55]T. Weymuth, M. Reiher, International Journal of Quantum Chemistry 2014, 114, 823.

[56]J. G. Freeze, H. R. Kelly, V. S. Batista, Chemical Reviews 2019, 119, 6595.

[57]S. V. Dudiy, A. Zunger, Physical Review Letters 2006, 97, 046401.

[58]M. A. Kramer, AIChE Journal 1991, 37, 233.

[59]I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Advances in Neural Information Processing Systems 2014, 27, 2672.

[60]O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, N. A. Mohamed, H. Arshad, Heliyon 2018, 4.

[61]R. S. Sutton, A. G. Barto, *Reinforcement learning: An introduction*, MIT press, 2018.

[62]W. Jin, R. Barzilay, T. Jaakkola, in *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80 (Eds: D. Jennifer, K. Andreas), PMLR, Proceedings of Machine Learning Research 2018, 2323.

[63]B. Settles, Synthesis Lectures on Artificial Intelligence and Machine Learning 2012, 6, 1.

[64]J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, A. E. Roitberg, The Journal of Chemical Physics 2018, 148, 241733.

[65]J. Cai, X. Chu, K. Xu, H. Li, J. Wei, Nanoscale Advances 2020, 2, 3115.

[66]J. Wei, X. Chu, X.-Y. Sun, K. Xu, H.-X. Deng, J. Chen, Z. Wei, M. Lei, InfoMat 2019, 1, 338.

[67]H. Yin, Z. Sun, Z. Wang, D. Tang, C. H. Pang, X. Yu, A. S. Barnard, H. Zhao, Z. Yin, Cell Reports Physical Science 2021, 2, 100482.

[68]S. Raschka, CoRR 2018, abs/1811.12808.

[69]D. Morgan, R. Jacobs, Annual Review of Materials Research 2020, 50, 71.

[70]D. M. Hawkins, S. C. Basak, D. Mills, Journal of Chemical Information and Computer Sciences 2003, 43, 579.

[71]H. Sahu, W. Rao, A. Troisi, H. Ma, Advanced Energy Materials 2018, 8, 1801032.

[72]B. Efron, R. Tibshirani, Statistical Science 1986, 1, 54.

[73]B. Efron, R. Tibshirani, Journal of the American Statistical Association 1997, 92, 548.

[74]J.-H. Kim, Computational Statistics & Data Analysis 2009, 53, 3735.

[75]J. Lever, M. Krzywinski, N. Altman, Nature Methods 2016, 13, 603.

[76]F. Sahigara, K. Mansouri, D. Ballabio, A. Mauri, V. Consonni, R. Todeschini, Molecules 2012, 17.

[77]A. Schwaighofer, T. Schroeter, S. Mika, G. Blanchard, Combinatorial Chemistry & High Throughput Screening 2009, 12, 453.

[78]A. A. Peterson, R. Christensen, A. Khorshidi, Physical Chemistry Chemical Physics 2017, 19, 10978.

[79]J. Behler, Journal of Physics: Condensed Matter 2014, 26, 183001.

[80]L. Hirschfeld, K. Swanson, K. Yang, R. Barzilay, C. W. Coley, Journal of Chemical Information and Modeling 2020, 60, 3770.

[81]D. Bajusz, A. Rácz, K. Héberger, Journal of Cheminformatics 2015, 7, 20.

[82]J. P. Janet, C. Duan, T. Yang, A. Nandy, H. J. Kulik, Chemical Science 2019, 10, 7913.

[83]R. C. Smith, *Uncertainty Quantification: Theory, Implementation, and Applications*, Society for Industrial and Applied Mathematics, 2013.

[84]S. R. Kalidindi, M. De Graef, Annual Review of Materials Research 2015, 45, 171.

[85]M. I. Jordan, T. M. Mitchell, Science 2015, 349, 255.

[86]G. Hinton, T. J. Sejnowski, The MIT Press, 1999.

[87]Y. LeCun, Y. Bengio, G. Hinton, Nature 2015, 521, 436.

[88]G. A. Seber, A. J. Lee, *Linear regression analysis*, Vol. 329, John Wiley & Sons, 2012.

[89]V. C. Epa, F. R. Burden, C. Tassa, R. Weissleder, S. Shaw, D. A. Winkler, Nano Letters 2012, 12, 5808.

[90]M. Fernandez, H. Shi, A. S. Barnard, Carbon 2016, 103, 142.

[91]M. Fernandez, J. I. Abreu, H. Shi, A. S. Barnard, ACS Combinatorial Science 2016, 18, 661.

[92]R. Jinnouchi, R. Asahi, The Journal of Physical Chemistry Letters 2017, 8, 4279.

[93]A. E. Hoerl, R. W. Kennard, Technometrics 1970, 12, 55.

[94]J. A. Arzola-Flores, A. L. González, The Journal of Physical Chemistry C 2020, 124, 25447.

[95]H. Zou, Journal of the American Statistical Association 2006, 101, 1418.

[96]A. C. Rajan, A. Mishra, S. Satsangi, R. Vaish, H. Mizuseki, K.-R. Lee, A. K. Singh, Chemistry of Materials 2018, 30, 4031.

[97]D. W. Hosmer Jr, S. Lemeshow, R. X. Sturdivant, *Applied logistic regression*, Vol. 398, John Wiley & Sons, 2013.

[98]B. Sun, M. Fernandez, A. S. Barnard, Journal of Chemical Information and Modeling 2017, 57, 2413.

[99]A. J. Parker, G. Opletal, A. S. Barnard, Journal of Applied Physics 2020, 128, 014301.

[100]C. Cortes, V. Vapnik, Machine Learning 1995, 20, 273.

[101]S. R. Gunn, ISIS technical report 1998, 14, 5.

[102]K. R. Müller, A. J. Smola, G. Rätsch, B. Schölkopf, J. Kohlmorgen, V. Vapnik, "Predicting time series with support vector machines", presented at *Artificial Neural Networks — ICANN'97*, Berlin, Heidelberg, 1997//, 1997.

[103]S. F. Fang, M. P. Wang, W. H. Qi, F. Zheng, Computational Materials Science 2008, 44, 647.

[104]Y. Liu, J. Wu, G. Yang, T. Zhao, S. Shi, Science Bulletin 2019, 64, 1195.

[105]B.-T. Chen, T.-P. Chang, J.-Y. Shih, J.-J. Wang, Computational Materials Science 2009, 44, 913.

[106]K. Fujimura, A. Seko, Y. Koyama, A. Kuwabara, I. Kishida, K. Shitara, C. A. J. Fisher, H. Moriwake, I. Tanaka, Advanced Energy Materials 2013, 3, 980.

[107]L. Xu, L. Wencong, P. Chunrong, S. Qiang, G. Jin, Computational Materials Science 2009, 46, 860.

[108]J. F. Pei, C. Z. Cai, Y. M. Zhu, Journal of Theoretical and Computational Chemistry 2012, 12, 1350002.

[109]F. Gharagheizi, P. Ilani-Kashkouli, A. H. Mohammadi, Chemical Engineering Science 2012, 81, 91.

[110]A. Alzghoul, A. Alhalaweh, D. Mahlin, C. A. S. Bergström, Journal of Chemical Information and Modeling 2014, 54, 3396.

[111]J. Hu, X. Cao, X. Zhao, W. Chen, G.-p. Lu, Y. Dan, Z. Chen, Frontiers in Chemistry 2019, 7, 444.

[112]X. Sun, J. Zheng, Y. Gao, C. Qiu, Y. Yan, Z. Yao, S. Deng, J. Wang, Applied Surface Science 2020, 526, 146522.

[113]K. Yu, Y. Cheng, Talanta 2007, 71, 676.

[114]C. Helma, T. Cramer, S. Kramer, L. De Raedt, Journal of Chemical Information and Computer Sciences 2004, 44, 1402.

[115]N. Cristianini, J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, Cambridge 2000.

[116]V. Vovk, in *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, (Eds: B. Schölkopf, Z. Luo, V. Vovk), Springer Berlin Heidelberg, Berlin, Heidelberg 2013, 105.

[117]N. Sheremetyeva, M. Lamparski, C. Daniels, B. Van Troeye, V. Meunier, Carbon 2020, 169, 455.

[118]V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, G. Csányi, Chemical Reviews 2021, 121, 10073.

[119]A. Mishra, S. Satsangi, A. C. Rajan, H. Mizuseki, K.-R. Lee, A. K. Singh, The Journal of Physical Chemistry Letters 2019, 10, 780.

[120]D. Wee, J. Kim, S. Bang, G. Samsonidze, B. Kozinsky, Physical Review Materials 2019, 3, 033803.

[121]J. R. Quinlan, Machine Learning 1986, 1, 81.

[122]W.-Y. Loh, WIREs Data Mining and Knowledge Discovery 2011, 1, 14.

[123]J. R. Quinlan, *C4. 5: programs for machine learning*, Elsevier, 2014.

[124]R. J. Lewis, "An introduction to classification and regression tree (CART) analysis", 2000.

[125]H. I. Labouta, N. Asgarian, K. Rinker, D. T. Cramb, ACS Nano 2019, 13, 1583.

[126]K. Tanaka, K. Hachiya, W. Zhang, K. Matsuda, Y. Miyauchi, ACS Nano 2019, 13, 12687.

[127]Y. Xie, C. Zhang, X. Hu, C. Zhang, S. P. Kelley, J. L. Atwood, J. Lin, Journal of the American Chemical Society 2020, 142, 1475.

[128]N. S. Altman, The American Statistician 1992, 46, 175.

[129]J.-S. Choi, M. K. Ha, T. X. Trinh, T. H. Yoon, H.-G. Byun, Scientific Reports 2018, 8, 6110.

[130]S. Wold, K. Esbensen, P. Geladi, Chemometrics and Intelligent Laboratory Systems 1987, 2, 37.

[131]H. Abdi, L. J. Williams, WIREs Computational Statistics 2010, 2, 433.

[132]S. M. Neumayer, M. A. Susner, M. McGuire, S. T. Pantelides, S. Kalnaus, P. Maksymovych, N. Balke, IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control 2020, 1.

[133]A. P. Dempster, N. M. Laird, D. B. Rubin, Journal of the Royal Statistical Society: Series B (Methodological) 1977, 39, 1.

[134]J. A. Hartigan, M. A. Wong, Journal of the Royal Statistical Society. Series C (Applied Statistics) 1979, 28, 100.

[135]R. Quesada-Cabrera, X. Weng, G. Hyett, R. J. H. Clark, X. Z. Wang, J. A. Darr, ACS Combinatorial Science 2013, 15, 458.

[136]Q. Luo, E. A. Holm, C. Wang, Nanoscale Advances 2021.

[137]A. S. Barnard, G. Opletal, Nanoscale 2019, 11, 23165.

[138]Z.-L. Wang, T. Ogawa, Y. Adachi, Advanced Theory and Simulations 2020, 3, 2000040.

[139]S.-C. Wang, in *Interdisciplinary Computing in Java Programming*, (Ed: S.-C. Wang), Springer US, Boston, MA 2003, 81.

[140]Z. Guo, S. Malinov, W. Sha, Computational Materials Science 2005, 32, 1.

[141]F. Altun, Ö. Kişi, K. Aydin, Computational Materials Science 2008, 42, 259.

[142]İ. B. Topçu, M. Sarıdemir, Computational Materials Science 2007, 41, 117.

[143]X. Chen, L. Sztandera, H. M. Cartwright, International Journal of Intelligent Systems 2008, 23, 22.

[144]J. Gajewski, T. Sadowski, Computational Materials Science 2014, 82, 114.

[145]D. A. Saldana, L. Starck, P. Mougin, B. Rousseau, N. Ferrando, B. Creton, Energy & Fuels 2012, 26, 2416.

[146]D. J. Scott, P. V. Coveney, J. A. Kilner, J. C. H. Rossiny, N. M. N. Alford, Journal of the European Ceramic Society 2007, 27, 4425.

[147]F. Häse, S. Valleau, E. Pyzer-Knapp, A. Aspuru-Guzik, Chemical Science 2016, 7, 5139.

[148]M. Salahinejad, T. C. Le, D. A. Winkler, Journal of Chemical Information and Modeling 2013, 53, 223.

[149]H. Wu, A. Lorenson, B. Anderson, L. Witteman, H. Wu, B. Meredig, D. Morgan, Computational Materials Science 2017, 134, 160.

[150]R. E. Raj, B. S. S. Daniel, Computational Materials Science 2008, 43, 767.

[151]S. Sivasankaran, R. Narayanasamy, T. Ramesh, M. Prabhakar, Computational Materials Science 2009, 47, 46.

[152]P. Cavaliere, Computational Materials Science 2007, 38, 722.

[153]Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, Neural Computation 1989, 1, 541.

[154]Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Proceedings of the IEEE 1998, 86, 2278.

[155]J. Madsen, P. Liu, J. Kling, J. B. Wagner, T. W. Hansen, O. Winther, J. Schiøtz, Advanced Theory and Simulations 2018, 1, 1800037.

[156]A. Maksov, O. Dyck, K. Wang, K. Xiao, D. B. Geohegan, B. G. Sumpter, R. K. Vasudevan, S. Jesse, S. V. Kalinin, M. Ziatdinov, npj Computational Materials 2019, 5, 12.

[157]C. W. Park, C. Wolverton, Physical Review Materials 2020, 4, 063801.

[158]T. Xie, J. C. Grossman, Physical Review Letters 2018, 120, 145301.

[159]X. Zheng, P. Zheng, R.-Z. Zhang, Chemical Science 2018, 9, 8426.

[160]T. Mikolov, S. Kombrink, L. Burget, J. Černocký, S. Khudanpur, "Extensions of recurrent neural network language model", presented at *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 22-27 May 2011, 2011.

[161]G. Karol, D. Ivo, G. Alex, R. Danilo, W. Daan, PMLR, 2015, 1462.

[162]Z. Zhou, X. Li, R. N. Zare, ACS Central Science 2017, 3, 1337.

[163]J. Wang, C. Li, S. Shin, H. Qi, The Journal of Physical Chemistry C 2020, 124, 14838.

[164]Y. Mao, Q. He, X. Zhao, Science Advances 2020, 6, eaaz4169.

[165]Y. Dan, Y. Zhao, X. Li, S. Li, M. Hu, J. Hu, npj Computational Materials 2020, 6, 84.

[166]J. Noh, J. Kim, H. S. Stein, B. Sanchez-Lengeling, J. M. Gregoire, A. Aspuru-Guzik, Y. Jung, Matter 2019, 1, 1370.

[167]C. Shen, M. Krenn, S. Eppel, A. Aspuru-Guzik, Machine Learning: Science and Technology 2021, 2, 03LT02.

[168]Z. Yao, B. Sánchez-Lengeling, N. S. Bobbitt, B. J. Bucior, S. G. H. Kumar, S. P. Collins, T. Burns, T. K. Woo, O. K. Farha, R. Q. Snurr, A. Aspuru-Guzik, Nature Machine Intelligence 2021, 3, 76.

[169]S. Kim, J. Noh, G. H. Gu, A. Aspuru-Guzik, Y. Jung, ACS Central Science 2020, 6, 1412.

[170]D. P. Kingma, M. Welling, in *ICLR*, 2014.

[171]J. Lim, S. Ryu, J. W. Kim, W. Y. Kim, Journal of Cheminformatics 2018, 10, 31.

[172]Q. Liu, M. Allamanis, M. Brockschmidt, A. Gaunt, Advances in Neural Information Processing Systems 2018, 31, 7795.

[173]R. Batra, H. Dai, T. D. Huan, L. Chen, C. Kim, W. R. Gutekunst, L. Song, R. Ramprasad, Chemistry of Materials 2020, 32, 10489.

[174]H. S. Stein, D. Guevarra, P. F. Newhouse, E. Soedarmadji, J. M. Gregoire, Chemical Science 2019, 10, 47.

[175]P. Smolensky, in *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations*, MIT Press, 1986, 194.

[176]G. E. Hinton, R. R. Salakhutdinov, Science 2006, 313, 504.

[177]H. Larochelle, Y. Bengio, in *Proceedings of the 25th International Conference on Machine Learning*, Association for Computing Machinery, Helsinki, Finland 2008, 536.

[178]C. Adam, N. Andrew, L. Honglak, PMLR, 2011, 215.

[179]R. G. Melko, G. Carleo, J. Carrasquilla, J. I. Cirac, Nature Physics 2019, 15, 887.

[180]R. Xia, S. Kais, Nature Communications 2018, 9, 4195.

[181]J. Liu, A. Mohan, R. K. Kalia, A. Nakano, K.-i. Nomura, P. Vashishta, K.-T. Yao, Computational Materials Science 2020, 173, 109429.

[182]R. E. Schapire, Machine Learning 1990, 5, 197.

[183]L. Breiman, Machine Learning 1996, 24, 123.

[184]Z.-H. Zhou, *Ensemble methods: foundations and algorithms*, CRC press, 2012.

[185]Y. Freund, R. E. Schapire, Journal of Computer and System Sciences 1997, 55, 119.

[186]M. Tonezzer, S. C. Izidoro, J. P. A. Moraes, L. T. T. Dang, Frontiers in Materials 2019, 6.

[187]R. W. Epps, M. S. Bowen, A. A. Volk, K. Abdel-Latif, S. Han, K. G. Reyes, A. Amassian, M. Abolhasani, Advanced Materials 2020, 32, 2001626.

[188]J. H. Friedman, The Annals of Statistics 2001, 29, 1189.

[189]L. Breiman, Machine Learning 2001, 45, 5.

[190]N. Artrith, Z. Lin, J. G. Chen, ACS Catalysis 2020, 10, 9438.

[191]F. Tao, Y. Laili, L. Zhang, in *Configurable Intelligent Optimization Algorithm: Design and Practice in Manufacturing*, (Eds: F. Tao, L. Zhang, Y. Laili), Springer International Publishing, Cham 2015, 3.

[192]D. Whitley, Statistics and Computing 1994, 4, 65.

[193]A. Kaczmarowski, S. Yang, I. Szlufarska, D. Morgan, Computational Materials Science 2015, 98, 234.

[194]M. Fernandez, H. Shi, A. S. Barnard, Journal of Chemical Information and Modeling 2015, 55, 2500.

[195]M. J. Cherukara, B. Narayanan, A. Kinaci, K. Sasikumar, S. K. Gray, M. K. Y. Chan, S. K. R. S. Sankaranarayanan, The Journal of Physical Chemistry Letters 2016, 7, 3752.

[196]J. Kennedy, R. Eberhart, "Particle swarm optimization", presented at *Proceedings of ICNN'95 - International Conference on Neural Networks*, 27 Nov.-1 Dec. 1995, 1995.

[197]Y. Wang, J. Lv, L. Zhu, Y. Ma, Physical Review B 2010, 82, 094116.

[198]Y. Wang, M. Miao, J. Lv, L. Zhu, K. Yin, H. Liu, Y. Ma, The Journal of Chemical Physics 2012, 137, 224108.

[199]D.-D. Chen, Y.-C. Lin, X.-M. Chen, Advances in Manufacturing 2019, 7, 238.

[200]S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi, Science 1983, 220, 671.

[201]F. Erchiqui, Applied Thermal Engineering 2018, 128, 1263.

[202]K. M. El-Naggar, M. R. AlRashidi, M. F. AlHajri, A. K. Al-Othman, Solar Energy 2012, 86, 266.

[203]S. Kunnikuruvan, A. Chakraborty, D. T. Major, The Journal of Physical Chemistry C 2020, 124, 27366.

[204]S. J. Pan, Q. Yang, IEEE Transactions on Knowledge and Data Engineering 2010, 22, 1345.

[205]D. Jha, K. Choudhary, F. Tavazza, W.-k. Liao, A. Choudhary, C. Campbell, A. Agrawal, Nature Communications 2019, 10, 5316.

[206]H. Yamada, C. Liu, S. Wu, Y. Koyama, S. Ju, J. Shiomi, J. Morikawa, R. Yoshida, ACS Central Science 2019, 5, 1717.

[207]E. D. Cubuk, A. D. Sendek, E. J. Reed, The Journal of Chemical Physics 2019, 150, 214701.

[208]J. Snoek, H. Larochelle, R. P. Adams, Advances in Neural Information Processing Systems 2012, 25, 2951.

[209]Z. Hou, K. Tsuda, in *Machine Learning Meets Quantum Physics*, (Eds: K. T. Schütt, S. Chmiela, O. A. von Lilienfeld, A. Tkatchenko, K. Tsuda, K.-R. Müller), Springer International Publishing, Cham 2020, 413.

[210]K. Osada, K. Kutsukake, J. Yamamoto, S. Yamashita, T. Kodera, Y. Nagai, T. Horikawa, K. Matsui, I. Takeuchi, T. Ujihara, Materials Today Communications 2020, 25, 101538.

[211]V. Botu, R. Ramprasad, International Journal of Quantum Chemistry 2015, 115, 1074.

[212]Z. Li, L. E. K. Achenie, H. Xin, ACS Catalysis 2020, 10, 4377.

[213]M. Popova, O. Isayev, A. Tropsha, Science Advances 2018, 4, eaap7885.

[214]I. Sajedian, T. Badloe, J. Rho, arXiv preprint arXiv:1810.10964 2018.

[215]S. Whitelam, I. Tamblyn, Physical Review E 2020, 101, 052604.

[216]C. Draxl, M. Scheffler, in *Handbook of Materials Modeling: Methods: Theory and Modeling*, (Eds: W. Andreoni, S. Yip), Springer International Publishing, Cham 2020, 49.

[217]J. Hill, A. Mannodi-Kanakkithodi, R. Ramprasad, B. Meredig, in *Computational Materials System Design*, (Eds: D. Shin, J. Saal), Springer International Publishing, Cham 2018, 193.

[218]V. I. Hegde, C. K. Borg, Z. del Rosario, Y. Kim, M. Hutchinson, E. Antono, J. Ling, P. Saxe, J. E. Saal, B. Meredig, arXiv preprint arXiv:2007.01988 2020.

[219]L. Zhou, S. Pan, J. Wang, A. V. Vasilakos, Neurocomputing 2017, 237, 350.

[220]S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, G. Ceder, Computational Materials Science 2013, 68, 314.

[221]S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, C. Wolverton, npj Computational Materials 2015, 1.

[222]A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, Journal of Physics: Condensed Matter 2017, 29, 273002.

[223]K. Mathew, J. H. Montoya, A. Faghaninia, S. Dwarakanath, M. Aykol, H. M. Tang, I. H. Chu, T. Smidt, B. Bocklund, M. Horton, J. Dagdelen, B. Wood, Z. K. Liu, J. Neaton, S. P. Ong, K. Persson, A. Jain, Computational Materials Science 2017, 139, 140.

[224]A. Jain, S. P. Ong, W. Chen, B. Medasani, X. H. Qu, M. Kocher, M. Brafman, G. Petretto, G. M. Rignanese, G. Hautier, D. Gunter, K. A. Persson, Concurrency and Computation-Practice & Experience 2015, 27, 5037.

[225]A. R. Supka, T. E. Lyons, L. Liyanage, P. D'Amico, R. Al Rahal Al Orabi, S. Mahatara, P. Gopal, C. Toher, D. Ceresoli, A. Calzolari, S. Curtarolo, M. B. Nardelli, M. Fornari, Computational Materials Science 2017, 136, 76.

[226]L. Ward, A. Dunn, A. Faghaninia, N. E. R. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. M. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K. A. Persson, G. J. Snyder, I. Foster, A. Jain, Computational Materials Science 2018, 152, 60.

[227]A. V. Yakutovich, K. Eimre, O. Schütt, L. Talirz, C. S. Adorf, C. W. Andersen, E. Ditler, D. Du, D. Passerone, B. Smit, N. Marzari, G. Pizzi, C. A. Pignedoli, Computational Materials Science 2021, 188, 110165.

[228]S. P. Huber, S. Zoupanos, M. Uhrin, L. Talirz, L. Kahle, R. Hauselmann, D. Gresch, T. Muller, A. V. Yakutovich, C. W. Andersen, F. F. Ramirez, C. S. Adorf, F. Gargiulo, S. Kumbhar, E. Passaro, C. Johnston, A. Merkys, A. Cepellotti, N. Mounet, N. Marzari, B. Kozinsky, G. Pizzi, Sci Data 2020, 7, 300.

[229]T. L. Nguyen, "A Framework for Five Big V's of Big Data and Organizational Culture in Firms", presented at *2018 IEEE International Conference on Big Data (Big Data)*, 10-13 Dec. 2018, 2018.

[230]C. T. Koch, *Determination of core structure periodicity and point defect density along dislocations*, 2002.

[231]K. A. Severson, P. M. Attia, N. Jin, N. Perkins, B. Jiang, Z. Yang, M. H. Chen, M. Aykol, P. K. Herring, D. Fraggedakis, M. Z. Bazant, S. J. Harris, W. C. Chueh, R. D. Braatz, Nature Energy 2019, 4, 383.

[232]J. R. Kitchin, Nature Catalysis 2018, 1, 230.

[233]J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, C. Wolverton, Jom 2013, 65, 1501.

[234]S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Ruhl, C. Wolverton, Npj Computational Materials 2015, 1, 1.

[235]A. Belsky, M. Hellenbrandt, V. L. Karen, P. Luksch, Acta Crystallographica Section B: Structural Science 2002, 58, 364.

[236]A. A. Emery, J. E. Saal, S. Kirklin, V. I. Hegde, C. Wolverton, Chemistry of Materials 2016, 28, 5621.

[237]D. S. Wang, M. Amsler, V. I. Hegde, J. E. Saal, A. Issa, B. C. Zhou, X. Q. Zeng, C. Wolverton, Acta Materialia 2018, 158, 65.

[238]A. R. Akbarzadeh, V. Ozoliņš, C. Wolverton, Advanced Materials 2007, 19, 3233.

[239]V. I. Hegde, M. Aykol, S. Kirklin, C. Wolverton, Science Advances 2020, 6, eaay5606.

[240]M. Amsler, V. I. Hegde, S. D. Jacobsen, C. Wolverton, Physical Review X 2018, 8, 041021.

[241]T. Hu, H. Song, T. Jiang, S. Li, Symmetry 2020, 12.

[242]V. Fung, G. Hu, P. Ganesh, B. G. Sumpter, Nature Communications 2021, 12, 88.

[243]G. Bergerhoff, R. Hundt, R. Sievers, I. D. Brown, Journal of Chemical Information and Computer Sciences 1983, 23, 66.

[244]S. V. Dudiy, A. Zunger, Phys Rev Lett 2006, 97, 046401.

[245]A. R. Oganov, C. W. Glass, J Chem Phys 2006, 124, 244704.

[246]M. d'Avezac, J. W. Luo, T. Chanier, A. Zunger, Phys Rev Lett 2012, 108, 027401.

[247]G. Hautier, C. Fischer, V. Ehrlacher, A. Jain, G. Ceder, Inorganic chemistry 2011, 50, 656.

[248]G. Hautier, C. C. Fischer, A. Jain, T. Mueller, G. Ceder, Chemistry of Materials 2010, 22, 3762.

[249]C. C. Fischer, K. J. Tibbetts, D. Morgan, G. Ceder, Nat Mater 2006, 5, 641.

[250]A. Jain, G. Hautier, C. J. Moore, S. P. Ong, C. C. Fischer, T. Mueller, K. A. Persson, G. Ceder, Computational Materials Science 2011, 50, 2295.

[251]K. Latimer, S. Dwaraknath, K. Mathew, D. Winston, K. A. Persson, Npj Computational Materials 2018, 4, 1.

[252]G. Petretto, S. Dwaraknath, P. C. M. H, D. Winston, M. Giantomassi, M. J. van Setten, X. Gonze, K. A. Persson, G. Hautier, G. M. Rignanese, Sci Data 2018, 5, 180065.

[253]M. de Jong, W. Chen, H. Geerlings, M. Asta, K. A. Persson, Sci Data 2015, 2, 150053.

[254]M. de Jong, W. Chen, T. Angsten, A. Jain, R. Notestine, A. Gamst, M. Sluiter, C. Krishna Ande, S. van der Zwaag, J. J. Plata, C. Toher, S. Curtarolo, G. Ceder, K. A. Persson, M. Asta, Sci Data 2015, 2, 150009.

[255]I. Petousis, D. Mrdjenovich, E. Ballouz, M. Liu, D. Winston, W. Chen, T. Graf, T. D. Schladt, K. A. Persson, F. B. Prinz, Sci Data 2017, 4, 160134.

[256]W. Chen, J.-H. Pöhls, G. Hautier, D. Broberg, S. Bajaj, U. Aydemir, Z. M. Gibbs, H. Zhu, M. Asta, G. J. Snyder, Journal of Materials Chemistry C 2016, 4, 4414.

[257]A. Jain, G. Hautier, S. P. Ong, C. J. Moore, C. C. Fischer, K. A. Persson, G. Ceder, Physical Review B 2011, 84, 045115.

[258]S. P. Ong, L. Wang, B. Kang, G. Ceder, Chemistry of Materials 2008, 20, 1798.

[259]K. A. Persson, B. Waldwick, P. Lazic, G. Ceder, Physical Review B 2012, 85, 235438.

[260]F. Zhou, M. Cococcioni, C. A. Marianetti, D. Morgan, G. Ceder, Physical Review B 2004, 70, 235121.

[261]G. Hautier, C. Fischer, V. Ehrlacher, A. Jain, G. Ceder, Inorg Chem 2011, 50, 656.

[262]X. H. Qu, A. Jain, N. N. Rajput, L. Cheng, Y. Zhang, S. P. Ong, M. Brafman, E. Maginn, L. A. Curtiss, K. A. Persson, Computational Materials Science 2015, 103, 56.

[263]L. Cheng, R. S. Assary, X. Qu, A. Jain, S. P. Ong, N. N. Rajput, K. Persson, L. A. Curtiss, J Phys Chem Lett 2015, 6, 283.

[264]R. Dmello, J. D. Milshtein, F. R. Brushett, K. C. Smith, Journal of Power Sources 2016, 330, 261.

[265]K. Mathew, C. Zheng, D. Winston, C. Chen, A. Dozier, J. J. Rehr, S. P. Ong, K. A. Persson, Sci Data 2018, 5, 180151.

[266]W. D. Richards, L. J. Miara, Y. Wang, J. C. Kim, G. Ceder, Chemistry of Materials 2016, 28, 266.

[267]E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder, E. Olivetti, Chemistry of Materials 2017, 29, 9436.

[268]S. P. Ong, S. Cholia, A. Jain, M. Brafman, D. Gunter, G. Ceder, K. A. Persson, Computational Materials Science 2015, 97, 209.

[269]Z. Wang, H. Zhang, J. Li, Nano Energy 2021, 81, 105665.

[270]S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. W. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, O. Levy, Computational Materials Science 2012, 58, 227.

[271]O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, A. Tropsha, Nat Commun 2017, 8, 15679.

[272]F. Legrain, J. Carrete, A. van Roekeghem, S. Curtarolo, N. Mingo, Chemistry of Materials 2017, 29, 6220.

[273]V. Stanev, C. Oses, A. G. Kusne, E. Rodriguez, J. Paglione, S. Curtarolo, I. Takeuchi, Npj Computational Materials 2018, 4, 1.

[274]C. Oses, E. Gossett, D. Hicks, F. Rose, M. J. Mehl, E. Perim, I. Takeuchi, S. Sanvito, M. Scheffler, Y. Lederer, O. Levy, C. Toher, S. Curtarolo, J Chem Inf Model 2018, 58, 2477.

[275]J. J. Plata, P. Nath, D. Usanmaz, J. Carrete, C. Toher, M. de Jong, M. Asta, M. Fornari, M. B. Nardelli, S. Curtarolo, npj Computational Materials 2017, 3, 45.

[276]R. H. Taylor, F. Rose, C. Toher, O. Levy, K. Yang, M. B. Nardelli, S. Curtarolo, Computational Materials Science 2014, 93, 178.

[277]V. Stanev, C. Oses, A. G. Kusne, E. Rodriguez, J. Paglione, S. Curtarolo, I. Takeuchi, npj Computational Materials 2018, 4, 29.

[278]M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. t Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S. A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, Sci Data 2016, 3, 160018.

[279]C. Draxl, M. Scheffler, Journal of Physics: Materials 2019, 2, 036001.

[280]C. Draxl, M. Scheffler, MRS Bulletin 2018, 43, 676.

[281]C. M. Acosta, R. Ouyang, A. Fazzio, M. Scheffler, L. M. Ghiringhelli, C. Carbogno, arXiv preprint arXiv:1805.10950 2018.

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

[283]D. D. Landis, J. S. Hummelshoj, S. Nestorov, J. Greeley, M. Dulak, T. Bligaard, J. K. Norskov, K. W. Jacobsen, Computing in Science & Engineering 2012, 14, 51.

[284]E. Garijo Del Rio, S. Kaappa, J. A. Garrido Torres, T. Bligaard, K. W. Jacobsen, J Chem Phys 2020, 153, 234116.

[285]J. S. Hummelshoj, F. Abild-Pedersen, F. Studt, T. Bligaard, J. K. Norskov, Angew Chem Int Ed Engl 2012, 51, 272.

[286]D. Zagorac, H. Muller, S. Ruehl, J. Zagorac, S. Rehme, J Appl Crystallogr 2019, 52, 918.

[287]C. R. Groom, I. J. Bruno, M. P. Lightfoot, S. C. Ward, Acta Crystallogr B Struct Sci Cryst Eng Mater 2016, 72, 171.

[288]S. N. Kabekkodu, J. Faber, T. Fawcett, Acta Crystallogr B 2002, 58, 333.

[289]J. C. Cole, S. Wiggin, F. Stanzione, Struct Dyn 2019, 6, 054301.

[290]C. W. Coley, W. H. Green, K. F. Jensen, J Chem Inf Model 2019, 59, 2529.

[291]L. E. Connor, A. D. Vassileiou, G. W. Halbert, B. F. Johnston, I. D. H. Oswald, CrystEngComm 2019, 21, 4465.

[292]F. Long, R. A. Nicholls, P. Emsley, S. Graaeulis, A. Merkys, A. Vaitkus, G. N. Murshudov, Acta Crystallogr D Struct Biol 2017, 73, 112.

[293]S. Gražulis, A. Merkys, A. Vaitkus, Handbook of Materials Modeling: Methods: Theory and Modeling 2020, 1863.

[294]J. O'Mara, B. Meredig, K. Michel, Jom-Us 2016, 68, 2031.

[295]A. J. Medford, M. R. Kunz, S. M. Ewing, T. Borders, R. Fushimi, ACS Catalysis 2018, 8, 7403.

[296]E. Kim, K. Huang, S. Jegelka, E. Olivetti, npj Computational Materials 2017, 3, 53.

[297]H. Zhang, K. Hippalgaonkar, T. Buonassisi, O. M. Løvvik, E. Sagvolden, D. Ding, arXiv preprint arXiv:1901.05801 2019.

[298]B. Blaiszik, L. Ward, M. Schwarting, J. Gaff, R. Chard, D. Pike, K. Chard, I. Foster, MRS Communications 2019, 9, 1125.

[299]B. Blaiszik, K. Chard, J. Pruyne, R. Ananthakrishnan, S. Tuecke, I. Foster, Jom-Us 2016, 68, 2045.

[300]D. Angelone, A. J. S. Hammer, S. Rohrbach, S. Krambeck, J. M. Granda, J. Wolf, S. Zalesskiy, G. Chisholm, L. Cronin, Nature Chemistry 2021, 13, 63.

[301]P. S. Gromski, J. M. Granda, L. Cronin, Trends in Chemistry 2020, 2, 4.

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

[303]Y. Wang, W. D. Richards, S. P. Ong, L. J. Miara, J. C. Kim, Y. Mo, G. Ceder, Nature Materials 2015, 14, 1026.

[304]X. Han, Y. Gong, K. Fu, X. He, G. T. Hitz, J. Dai, A. Pearse, B. Liu, H. Wang, G. Rubloff, Y. Mo, V. Thangadurai, E. D. Wachsman, L. Hu, Nature Materials 2017, 16, 572.

[305]Y. Wu, W. Li, A. Faghaninia, Z. Chen, J. Li, X. Zhang, B. Gao, S. Lin, B. Zhou, A. Jain, Y. Pei, Materials Today Physics 2017, 3, 127.

[306]C. Zheng, K. Mathew, C. Chen, Y. Chen, H. Tang, A. Dozier, J. J. Kas, F. D. Vila, J. J. Rehr, L. F. J. Piper, K. A. Persson, S. P. Ong, npj Computational Materials 2018, 4, 12.

[307]S. Pablo-García, M. Álvarez-Moreno, N. López, International Journal of Quantum Chemistry 2021, 121, e26382.

[308]C. Bourgès, Y. Bouyrie, A. R. Supka, R. Al Rahal Al Orabi, P. Lemoine, O. I. Lebedev, M. Ohta, K. Suekuni, V. Nassif, V. Hardy, R. Daou, Y. Miyazaki, M. Fornari, E. Guilmeau, Journal of the American Chemical Society 2018, 140, 2186.

[309]V. Vitale, G. Pizzi, A. Marrazzo, J. R. Yates, N. Marzari, A. A. Mostofi, npj Computational Materials 2020, 6, 66.

[310]C. A. F. Salvador, B. F. Zornio, C. R. Miranda, ACS Applied Materials & Interfaces 2020, 12, 56850.

[311]M. Haghighatlari, G. Vishwakarma, D. Altarawy, R. Subramanian, B. U. Kota, A. Sonpal, S. Setlur, J. Hachmann, Wiley Interdisciplinary Reviews: Computational Molecular Science 2020, 10, e1458.

[312]M. Haghighatlari, J. Hachmann, Current Opinion in Chemical Engineering 2019, 23, 51.

[313]R. Jacobs, T. Mayeshiba, B. Afflerbach, L. Miles, M. Williams, M. Turner, R. Finkel, D. Morgan, Comp Mater Sci 2020, 176, 109544.

[314]C. Loftis, K. Yuan, Y. Zhao, M. Hu, J. Hu, The Journal of Physical Chemistry A 2021, 125, 435.

[315]Y. Fu, B. Yang, Y. Ma, Q. Sun, J. Yao, W. Fu, W. Yin, Powder Technology 2020, 376, 486.

[316]S. Lopatin, A. Aljarb, V. Roddatis, T. Meyer, Y. Wan, J.-H. Fu, M. Hedhili, Y. Han, L.-J. Li, V. Tung, Science Advances 2020, 6, eabb8431.

[317]S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. Hart, S. Sanvito, M. Buongiorno-Nardelli, Computational Materials Science 2012, 58, 227.

[318]C. R. Groom, I. J. Bruno, M. P. Lightfoot, S. C. Ward, Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials 2016, 72, 171.

[319]S. Haastrup, M. Strange, M. Pandey, T. Deilmann, P. S. Schmidt, N. F. Hinsche, M. N. Gjerding, D. Torelli, P. M. Larsen, A. C. Riis-Jensen, 2D Materials 2018, 5, 042002.

[320]J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sanchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway, A. Aspuru-Guzik, Journal of Physical Chemistry Letters 2011, 2, 2241.

[321]S. S. Borysov, R. M. Geilhufe, A. V. Balatsky, PloS one 2017, 12, e0171501.

[322]K. Choudhary, K. F. Garrity, A. C. E. Reid, B. DeCost, A. J. Biacchi, A. H. R. Walker, Z. Trautt, J. Hattrick-Simpers, A. G. Kusne, A. Centrone, A. Davydov, J. Jiang, R. Pachter, G. Cheon, E. Reed, A. Agrawal, X. F. Qian, V. Sharma, H. L. Zhuang, S. V. Kalinin, B. G. Sumpter, G. Pilania, P. Acar, S. Mandal, K. Haule, D. Vanderbilt, K. Rabe, F. Tavazza, Npj Computational Materials 2020, 6.

[323]N. Mounet, M. Gibertini, P. Schwaller, D. Campi, A. Merkys, A. Marrazzo, T. Sohier, I. E. Castelli, A. Cepellotti, G. Pizzi, Nature nanotechnology 2018, 13, 246.

[324]M. Widom, M. Mihalkovic, Journal of materials research 2005, 20, 237.

[325]R. D. Johnson III, 1999.

[326]R. Tran, Z. Xu, B. Radhakrishnan, D. Winston, W. Sun, K. A. Persson, S. P. Ong, Scientific Data 2016, 3, 160080.

[327]P. Gorai, D. Gao, B. Ortiz, S. Miller, S. A. Barnett, T. Mason, Q. Lv, V. Stevanović, E. S. Toberer, Computational Materials Science 2016, 112, 368.

[328]M. Kouchi, M. Mochimaru, H16PRO287 2005.

[329]R. T. Downs, M. Hall-Wallace, American Mineralogist 2003, 88, 247.

[330]H. E. Pence, A. Williams, ACS Publications, 2010.

[331]P. S. White, J. R. Rodgers, Y. Le Page, Acta Crystallographica Section B: Structural Science 2002, 58, 343.

[332]E. D. Palik, *Handbook of optical constants of solids*, Vol. 3, Academic press, 1998.

[333]E. Kress-Rogers, C. Brimelow, *Knovel solvents-a properties database*, ChemTec Publishing, 2000.

[334]T. Ogata, M. Yamazaki, 2012.

[335]L. MatWeb, Material property data, Data base of materials data sheets 1996.

[336]G. Klimeck, M. McLennan, S. P. Brophy, G. B. Adams III, M. S. Lundstrom, Computing in Science & Engineering 2008, 10, 17.

[337]C. A. Becker, F. Tavazza, Z. T. Trautt, R. A. B. de Macedo, Current Opinion in Solid State and Materials Science 2013, 17, 277.

[338]L. M. Hale, Z. T. Trautt, C. A. Becker, Modelling and Simulation in Materials Science and Engineering 2018, 26, 055003.

[339]E. B. Tadmor, R. S. Elliott, J. P. Sethna, R. E. Miller, C. A. Becker, Jom 2011, 63, 17.

[340]P. Villars, M. Berndt, K. Brandenburg, K. Cenzual, J. Daams, F. Hulliger, T. Massalski, H. Okamoto, K. Osaki, A. Prince, Journal of Alloys and Compounds 2004, 367, 293.

[341]P. Villars, K. Cenzual, Crystal Structure Database for Inorganic Compounds (Materials Park (OH): ASM International, 2012) 2009.

[342]J. Faber, T. Fawcett, Acta Crystallographica Section B: Structural Science 2002, 58, 325.

[343]S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, Nucleic acids research 2019, 47, D1102.

[344]J. Goodman, ACS Publications, 2009.

[345]S. W. Gabrielson, Journal of the Medical Library Association: JMLA 2018, 106, 588.

[346]M. W. Gaultois, T. D. Sparks, C. K. Borg, R. Seshadri, W. D. Bonificio, D. R. Clarke, Chemistry of Materials 2013, 25, 2911.

[347]V. Stevanović, S. Lany, X. Zhang, A. Zunger, Physical Review B 2012, 85, 115104.

[348]D. Barthelmy, http://webmineral. com/ 2007.

[349]C. Baerlocher, http://www. iza-structure. org/databases/ 2008.

[350]Z. Ghahramani, Nature 2015, 521, 452.

[351]M. Zhong, K. Tran, Y. Min, C. Wang, Z. Wang, C.-T. Dinh, P. De Luna, Z. Yu, A. S. Rasouli, P. Brodersen, Nature 2020, 581, 178.

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

[353]R. B. Wexler, J. M. P. Martirez, A. M. Rappe, J Am Chem Soc 2018, 140, 4678.

[354]P. Friederich, G. dos Passos Gomes, R. De Bin, A. Aspuru-Guzik, D. Balcells, Chemical Science 2020, 11, 4584.

[355]X. Sun, J. Zheng, Y. Gao, C. Qiu, Y. Yan, Z. Yao, S. Deng, J. Wang, Applied Surface Science 2020, 526.

[356]D. W. Davies, K. T. Butler, A. Walsh, Chemistry of Materials 2019, 31, 7221.

[357]L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, Npj Computational Materials 2016, 2, 16028.

[358]J. Hu, X. Cao, X. Zhao, W. Chen, G. P. Lu, Y. Dan, Z. Chen, Front Chem 2019, 7, 444.

[359]M. Ruck, B. Garlyyev, F. Mayr, A. S. Bandarenka, A. Gagliardi, J Phys Chem Lett 2020, 11, 1773.

[360]L. Ge, H. Yuan, Y. Min, L. Li, S. Chen, L. Xu, W. A. Goddard, 3rd, J Phys Chem Lett 2020, 11, 869.

[361]X. Ma, Z. Li, L. E. Achenie, H. Xin, J Phys Chem Lett 2015, 6, 3528.

[362]Y. Bai, L. Wilbraham, B. J. Slater, M. A. Zwijnenburg, R. S. Sprick, A. I. Cooper, J Am Chem Soc 2019, 141, 9063.

[363]H. Sahu, W. Rao, A. Troisi, H. Ma, Advanced Energy Materials 2018, 8.

[364]M. Fathinia, A. Khataee, S. Aber, A. Naseri, Applied Catalysis B: Environmental 2016, 184, 270.

[365]X. Ma, Z. Li, L. E. Achenie, H. Xin, The journal of physical chemistry letters 2015, 6, 3528.

[366]A. A. Peterson, J. K. Nørskov, The Journal of Physical Chemistry Letters 2012, 3, 251.

[367]J. Hussain, H. Jónsson, E. Skúlason, ACS Catalysis 2018, 8, 5240.

[368]A. Bagger, W. Ju, A. S. Varela, P. Strasser, J. Rossmeisl, ChemPhysChem 2017, 18, 3266.

[369]J. Kang, S. H. Noh, J. Hwang, H. Chun, H. Kim, B. Han, Phys Chem Chem Phys 2018, 20, 24539.

[370]B. Hammer, J. K. Nørskov, Advances in catalysis 2000, 45, 71.

[371]H. Masood, C. Y. Toe, W. Y. Teoh, V. Sethu, R. Amal, ACS Catalysis 2019, 9, 11774.

[372]L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, M. Scheffler, Physical Review Letters 2015, 114, 105503.

[373]J. Behler, The Journal of Chemical Physics 2011, 134, 074106.

[374]X. Zhu, J. Yan, M. Gu, T. Liu, Y. Dai, Y. Gu, Y. Li, The Journal of Physical Chemistry Letters 2019, 10, 7760.

[375]B. Meredig, C. Wolverton, Chemistry of Materials 2014, 26, 1985.

[376]B. B. Hoar, S. Lu, C. Liu, J Phys Chem Lett 2020, 11, 4625.

[377]M. O. J. Jäger, E. V. Morooka, F. Federici Canova, L. Himanen, A. S. Foster, npj Computational Materials 2018, 4.

[378]B. Hammer, J. K. Nørskov, in *Advances in Catalysis*, Vol. 45, Academic Press, 2000, 71.

[379]X. Ma, Z. Li, L. E. K. Achenie, H. Xin, The Journal of Physical Chemistry Letters 2015, 6, 3528.

[380]Y. Bai, L. Wilbraham, B. J. Slater, M. A. Zwijnenburg, R. S. Sprick, A. I. Cooper, Journal of the American Chemical Society 2019, 141, 9063.

[381]L. Ge, H. Yuan, Y. Min, L. Li, S. Chen, L. Xu, W. A. Goddard, The Journal of Physical Chemistry Letters 2020, 11, 869.

[382]L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, A. S. Foster, Computer Physics Communications 2020, 247, 106949.

[383]R. Tibshirani, Journal of the Royal Statistical Society Series B-Statistical Methodology 2011, 73, 273.

[384]C. W. Yap, Journal of computational chemistry 2011, 32, 1466.

[385]W. Xu, M. Andersen, K. Reuter, ACS Catalysis 2020, 734.

[386]A. Saeki, Polymer Journal 2020, 52, 1307.

[387]L. Weston, C. Stampfl, Physical Review Materials 2018, 2, 085407.

[388]Y. Huang, C. Yu, W. Chen, Y. Liu, C. Li, C. Niu, F. Wang, Y. Jia, Journal of Materials Chemistry C 2019, 7, 3238.

[389]J. Kang, S. H. Noh, J. Hwang, H. Chun, H. Kim, B. Han, Physical Chemistry Chemical Physics 2018, 20, 24539.

[390]M. C. Groenenboom, R. M. Anderson, J. A. Wollmershauser, D. J. Horton, S. A. Policastro, J. A. Keith, The Journal of Physical Chemistry C 2020, 124, 15171.

[391]T. A. A. Batchelor, J. K. Pedersen, S. H. Winther, I. E. Castelli, K. W. Jacobsen, J. Rossmeisl, Joule 2019, 3, 834.

[392]Z. W. Ulissi, A. R. Singh, C. Tsai, J. K. Nørskov, The Journal of Physical Chemistry Letters 2016, 7, 3931.

[393]B. B. Hoar, S. Lu, C. Liu, The Journal of Physical Chemistry Letters 2020, 11, 4625.

[394]K. Choudhary, K. F. Garrity, F. Tavazza, Journal of Physics: Condensed Matter 2020, 32, 475501.

[395]L. Chen, H. Tran, R. Batra, C. Kim, R. Ramprasad, Computational Materials Science 2019, 170, 109155.

[396]Z. Hou, Y. Takagiwa, Y. Shinohara, Y. Xu, K. Tsuda, ACS Applied Materials & Interfaces 2019, 11, 11545.

[397]D. Xue, P. V. Balachandran, R. Yuan, T. Hu, X. Qian, E. R. Dougherty, T. Lookman, Proceedings of the National Academy of Sciences 2016, 113, 13301.

[398]A. D. Sendek, Q. Yang, E. D. Cubuk, K.-A. N. Duerloo, Y. Cui, E. J. Reed, Energy & Environmental Science 2017, 10, 306.

[399]Z. Ahmad, T. Xie, C. Maheshwari, J. C. Grossman, V. Viswanathan, ACS Central Science 2018, 4, 996.

[400]A. Ishikawa, K. Sodeyama, Y. Igarashi, T. Nakayama, Y. Tateyama, M. Okada, Physical Chemistry Chemical Physics 2019, 21, 26399.

[401]M. Attarian Shandiz, R. Gauvin, Computational Materials Science 2016, 117, 270.

[402]R. P. Joshi, J. Eickholt, L. Li, M. Fornari, V. Barone, J. E. Peralta, ACS Applied Materials & Interfaces 2019, 11, 18494.

[403]S. Zhu, J. Li, L. Ma, C. He, E. Liu, F. He, C. Shi, N. Zhao, Materials Letters 2018, 233, 294.

[404]T. Konno, H. Kurokawa, F. Nabeshima, Y. Sakishita, R. Ogawa, I. Hosako, A. Maeda, Physical Review B 2021, 103, 014509.

[405]L. Ward, S. C. O'Keeffe, J. Stevick, G. R. Jelbert, M. Aykol, C. Wolverton, Acta Materialia 2018, 159, 102.

[406]T. Long, N. M. Fortunato, Y. Zhang, O. Gutfleisch, H. Zhang, Materials Research Letters 2021, 9, 169.

[407]J. J. Möller, W. Körner, G. Krugel, D. F. Urban, C. Elsässer, Acta Materialia 2018, 153, 53.

[408]Y. Wang, Y. Tian, T. Kirk, O. Laris, J. H. Ross, R. D. Noebe, V. Keylin, R. Arróyave, Acta Materialia 2020, 194, 144.

[409]J. Joy, J. Mathew, S. C. George, International Journal of Hydrogen Energy 2018, 43, 4804.

[410]E. J. Popczun, J. R. McKone, C. G. Read, A. J. Biacchi, A. M. Wiltrout, N. S. Lewis, R. E. Schaak, Journal of the American Chemical Society 2013, 135, 9267.

[411]P. Liu, J. A. Rodriguez, Journal of the American Chemical Society 2005, 127, 14871.

[412]B. Hinnemann, P. G. Moses, J. Bonde, K. P. Jørgensen, J. H. Nielsen, S. Horch, I. Chorkendorff, J. K. Nørskov, Journal of the American Chemical Society 2005, 127, 5308.

[413]T. F. Jaramillo, K. P. Jørgensen, J. Bonde, J. H. Nielsen, S. Horch, I. Chorkendorff, Science 2007, 317, 100.

[414]Z. Li, X. Ma, H. Xin, Catalysis Today 2017, 280, 232.

[415]Z. Li, S. Wang, W. S. Chin, L. E. Achenie, H. Xin, Journal of Materials Chemistry A 2017, 5, 24131.

[416]Z. Li, N. Omidvar, W. S. Chin, E. Robb, A. Morris, L. Achenie, H. Xin, The Journal of Physical Chemistry A 2018, 122, 4571.

[417]H. Xin, A. Holewinski, N. Schweitzer, E. Nikolla, S. Linic, Topics in Catalysis 2012, 55, 376.

[418]B. Hammer, J. K. Nørskov, Surface Science 1995, 343, 211.

[419]F. Abild-Pedersen, J. Greeley, F. Studt, J. Rossmeisl, T. R. Munter, P. G. Moses, E. Skúlason, T. Bligaard, J. K. Nørskov, Physical Review Letters 2007, 99, 016105.

[420]R. Ouyang, E. Ahmetcik, C. Carbogno, M. Scheffler, L. M. Ghiringhelli, Journal of Physics: Materials 2019, 2, 024002.

[421]M. Andersen, S. V. Levchenko, M. Scheffler, K. Reuter, ACS Catalysis 2019, 9, 2752.

[422]P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. Dal Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari, R. M. Wentzcovitch, Journal of Physics: Condensed Matter 2009, 21, 395502.

[423]J. Wellendorff, K. T. Lundgaard, A. Møgelhøj, V. Petzold, D. D. Landis, J. K. Nørskov, T. Bligaard, K. W. Jacobsen, Physical Review B 2012, 85, 235149.

[424]I. C. Man, H.-Y. Su, F. Calle-Vallejo, H. A. Hansen, J. I. Martínez, N. G. Inoglu, J. Kitchin, T. F. Jaramillo, J. K. Nørskov, J. Rossmeisl, ChemCatChem 2011, 3, 1159.

[425]G. J. Hedley, A. Ruseckas, I. D. W. Samuel, Chemical Reviews 2017, 117, 796.

[426]D. Ginley, M. A. Green, R. Collins, MRS Bulletin 2008, 33, 355.

[427]M. L. Agiorgousis, Y.-Y. Sun, D.-H. Choe, D. West, S. Zhang, Advanced Theory and Simulations 2019, 2, 1800173.

[428]J. M. Ball, A. Petrozza, Nature Energy 2016, 1, 16149.

[429]F. Li, C. Ma, H. Wang, W. Hu, W. Yu, A. D. Sheikh, T. Wu, Nature Communications 2015, 6, 8238.

[430]J. Huang, Y. Yuan, Y. Shao, Y. Yan, Nature Reviews Materials 2017, 2, 17042.

[431]W. Li, Z. Wang, F. Deschler, S. Gao, R. H. Friend, A. K. Cheetham, Nature Reviews Materials 2017, 2, 16099.

[432]J. P. Perdew, K. Burke, M. Ernzerhof, Physical Review Letters 1996, 77, 3865.

[433]Y. Zhuo, A. Mansouri Tehrani, J. Brgoch, The Journal of Physical Chemistry Letters 2018, 9, 1668.

[434]A. J. Smola, B. Schölkopf, Statistics and Computing 2004, 14, 199.

[435]L. Meng, Y. Zhang, X. Wan, C. Li, X. Zhang, Y. Wang, X. Ke, Z. Xiao, L. Ding, R. Xia, H.-L. Yip, Y. Cao, Y. Chen, Science 2018, 361, 1094.

[436]M. C. Scharber, D. Mühlbacher, M. Koppe, P. Denk, C. Waldauf, A. J. Heeger, C. J. Brabec, Advanced Materials 2006, 18, 789.

[437]Y. Imamura, M. Tashiro, M. Katouda, M. Hada, The Journal of Physical Chemistry C 2017, 121, 28275.

[438]C. Zanlorenzi, L. Akcelrud, Journal of Polymer Science Part B: Polymer Physics 2017, 55, 919.

[439]V. Venkatraman, B. K. Alsberg, Dyes and Pigments 2015, 114, 69.

[440]S. Rühle, H. N. Barad, Y. Bouhadana, D. A. Keller, A. Ginsburg, K. Shimanovich, K. Majhi, R. Lovrincic, A. Y. Anderson, A. Zaban, Physical Chemistry Chemical Physics 2014, 16, 7066.

[441]I. T. Jolliffe, J. Cadima, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 2016, 374, 20150202.

[442]A. Y. Anderson, Y. Bouhadana, H.-N. Barad, B. Kupfer, E. Rosh-Hodesh, H. Aviv, Y. R. Tischler, S. Rühle, A. Zaban, ACS Combinatorial Science 2014, 16, 53.

[443]A. Kulkarni, S. Siahrostami, A. Patel, J. K. Nørskov, Chemical Reviews 2018, 118, 2302.

[444]J.-L. Jiang, X. Su, H. Zhang, X.-H. Zhang, Y.-J. Yuan, Chemical Biology & Drug Design 2013, 81, 650.

[445]J. Behler, M. Parrinello, Physical Review Letters 2007, 98, 146401.

[446]A. Khorshidi, A. A. Peterson, Computer Physics Communications 2016, 207, 310.

[447]C. G. Broyden, IMA Journal of Applied Mathematics 1970, 6, 76.

[448]R. Fletcher, The Computer Journal 1970, 13, 317.

[449]K. W. Jacobsen, P. Stoltze, J. K. Nørskov, Surface Science 1996, 366, 394.

[450]A. Hjorth Larsen, J. Jørgen Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. Bjerre Jensen, J. Kermode, J. R. Kitchin, E. Leonhard Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. Bergmann Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, K. W. Jacobsen, Journal of Physics: Condensed Matter 2017, 29, 273002.

[451]X. Liu, J. Xiao, H. Peng, X. Hong, K. Chan, J. K. Nørskov, Nature Communications 2017, 8, 15438.

[452]Q. Lu, J. Rosen, Y. Zhou, G. S. Hutchings, Y. C. Kimmel, J. G. Chen, F. Jiao, Nature Communications 2014, 5, 3242.

[453]M. Siebert, G. Krennrich, M. Seibicke, A. F. Siegle, O. Trapp, Chemical Science 2019, 10, 10466.

[454]R. S. Olson, R. J. Urbanowicz, P. C. Andrews, N. A. Lavender, L. C. Kidd, J. H. Moore, "Automating Biomedical Data Science Through Tree-Based Pipeline Optimization", presented at *Applications of Evolutionary Computation*, Cham, 2016, 2016.

[455]J. K. Pedersen, T. A. A. Batchelor, A. Bagger, J. Rossmeisl, ACS Catalysis 2020, 10, 2169.

[456]M. K. D. P. K. G. S. Nellaiappan S; Kumar N; Kumar R; Parui A, A. K.; Sharma, S.; Tiwary, C. S.; Biswas, K., ChemRxiv. Cambridge: Cambridge Open Engage 2019.

[457]B. H. R. Suryanto, H.-L. Du, D. Wang, J. Chen, A. N. Simonov, D. R. MacFarlane, Nature Catalysis 2019, 2, 290.

[458]J. Deng, J. A. Iñiguez, C. Liu, Joule 2018, 2, 846.

[459]S. Z. Andersen, V. Čolić, S. Yang, J. A. Schwalbe, A. C. Nielander, J. M. McEnaney, K. Enemark-Rasmussen, J. G. Baker, A. R. Singh, B. A. Rohr, M. J. Statt, S. J. Blair, S. Mezzavilla, J. Kibsgaard, P. C. K. Vesborg, M. Cargnello, S. F. Bent, T. F. Jaramillo, I. E. L. Stephens, J. K. Nørskov, I. Chorkendorff, Nature 2019, 570, 504.

[460]S. L. Foster, S. I. P. Bakovic, R. D. Duda, S. Maheshwari, R. D. Milton, S. D. Minteer, M. J. Janik, J. N. Renner, L. F. Greenlee, Nature Catalysis 2018, 1, 490.

[461]X. Cui, C. Tang, Q. Zhang, Advanced Energy Materials 2018, 8, 1800369.

[462]X. Zhu, S. Mou, Q. Peng, Q. Liu, Y. Luo, G. Chen, S. Gao, X. Sun, Journal of Materials Chemistry A 2020, 8, 1545.

[463]C. Guo, J. Ran, A. Vasileff, S.-Z. Qiao, Energy & Environmental Science 2018, 11, 45.

[464]S. Lu, D. H. Lee, C. Liu, Small Methods 2019, 3, 1800332.

[465]G. Chen, M. S. Dresselhaus, G. Dresselhaus, J. P. Fleurial, T. Caillat, International Materials Reviews 2003, 48, 45.

[466]H. Wang, Y. Pei, A. D. LaLonde, G. Jeffery Snyder, in *Thermoelectric Nanomaterials: Materials Design and Applications*, (Eds: K. Koumoto, T. Mori), Springer Berlin Heidelberg, Berlin, Heidelberg 2013, 3.

[467]M. Zebarjadi, K. Esfarjani, M. S. Dresselhaus, Z. F. Ren, G. Chen, Energy & Environmental Science 2012, 5, 5147.

[468]L. E. Bell, Science 2008, 321, 1457.

[469]H. J. Goldsmid, in *Thermoelectric Refrigeration*, (Ed: H. J. Goldsmid), Springer US, Boston, MA 1964, 210.

[470]J. Yang, T. Caillat, MRS Bulletin 2006, 31, 224.

[471]S. B. Riffat, X. Ma, Applied Thermal Engineering 2003, 23, 913.

[472]D. Narducci, Applied Physics Letters 2011, 99, 102104.

[473]M. W. Gaultois, T. D. Sparks, C. K. H. Borg, R. Seshadri, W. D. Bonificio, D. R. Clarke, Chemistry of Materials 2013, 25, 2911.

[474]N. Neophytou, H. Karamitaheri, H. Kosina, Journal of Computational Electronics 2013, 12, 611.

[475]D. Beretta, N. Neophytou, J. M. Hodges, M. G. Kanatzidis, D. Narducci, M. Martin- Gonzalez, M. Beekman, B. Balke, G. Cerretti, W. Tremel, A. Zevalkink, A. I. Hofmann, C. Müller, B. Dörling, M. Campoy-Quiles, M. Caironi, Materials Science and Engineering: R: Reports 2019, 138, 100501.

[476]G. K. H. Madsen, D. J. Singh, Computer Physics Communications 2006, 175, 67.

[477]G. K. H. Madsen, J. Carrete, M. J. Verstraete, Computer Physics Communications 2018, 231, 140.

[478]K. Choudhary, B. DeCost, F. Tavazza, Physical Review Materials 2018, 2, 083801.

[479]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Journal of machine Learning research 2011, 12, 2825.

[480]Y. Zhang, C. Ling, npj Computational Materials 2018, 4, 25.

[481]S. A. Miller, P. Gorai, B. R. Ortiz, A. Goyal, D. Gao, S. A. Barnett, T. O. Mason, G. J. Snyder, Q. Lv, V. Stevanović, E. S. Toberer, Chemistry of Materials 2017, 29, 2494.

[482]E. Sun, W. Cao, Progress in Materials Science 2014, 65, 124.

[483]H. Zhang, C. Groh, Q. Zhang, W. Jo, K. G. Webber, J. Rödel, Advanced Electronic Materials 2015, 1, 1500018.

[484]X. Liu, X. Tan, Advanced Materials 2016, 28, 574.

[485]D. Xue, P. V. Balachandran, H. Wu, R. Yuan, Y. Zhou, X. Ding, J. Sun, T. Lookman, Applied Physics Letters 2017, 111, 032907.

[486]W. Liu, X. Ren, Physical Review Letters 2009, 103, 257602.

[487]R. Guo, L. E. Cross, S. E. Park, B. Noheda, D. E. Cox, G. Shirane, Physical Review Letters 2000, 84, 5423.

[488]A. Bouzid, E. M. Bourim, M. Gabbay, G. Fantozzi, Journal of the European Ceramic Society 2005, 25, 3213.

[489]T. R. Shrout, S. J. Zhang, Journal of Electroceramics 2007, 19, 113.

[490]J. Rödel, K. G. Webber, R. Dittmer, W. Jo, M. Kimura, D. Damjanovic, Journal of the European Ceramic Society 2015, 35, 1659.

[491]J. Rödel, W. Jo, K. T. P. Seifert, E.-M. Anton, T. Granzow, D. Damjanovic, Journal of the American Ceramic Society 2009, 92, 1153.

[492]D. R. Jones, M. Schonlau, W. J. Welch, Journal of Global Optimization 1998, 13, 455.

[493]P. V. Balachandran, D. Xue, J. Theiler, J. Hogden, T. Lookman, Scientific Reports 2016, 6, 19660.

[494]Y. Wang, Y. Song, Y. Xia, Chemical Society Reviews 2016, 45, 5925.

[495]K. Xu, Chemical Reviews 2004, 104, 4303.

[496]M. D. Bhatt, C. O'Dwyer, Physical Chemistry Chemical Physics 2015, 17, 4799.

[497]L. Cheng, R. S. Assary, X. Qu, A. Jain, S. P. Ong, N. N. Rajput, K. Persson, L. A. Curtiss, The Journal of Physical Chemistry Letters 2015, 6, 283.

[498]M. D. Halls, K. Tasaki, Journal of Power Sources 2010, 195, 1472.

[499]S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, O. Levy, Nature Materials 2013, 12, 191.

[500]M. Korth, Physical Chemistry Chemical Physics 2014, 16, 7919.

[501]T. Husch, N. D. Yilmazer, A. Balducci, M. Korth, Physical Chemistry Chemical Physics 2015, 17, 3394.

[502]H. Wada, M. Menetrier, A. Levasseur, P. Hagenmuller, Materials Research Bulletin 1983, 18, 189.

[503]Y. Tomita, H. Matsushita, K. Kobayashi, Y. Maeda, K. Yamada, Solid State Ionics 2008, 179, 867.

[504]K. Yamada, K. Kumano, T. Okuda, Solid State Ionics 2006, 177, 1691.

[505]R. Court-Castagnet, C. Kaps, C. Cros, P. Hagenmuller, Solid State Ionics 1993, 61, 327.

[506]Z. Ahmad, V. Viswanathan, Physical Review Letters 2017, 119, 056003.

[507]Z. Ahmad, V. Viswanathan, Physical Review Materials 2017, 1, 055403.

[508]Z. Ahmad, V. Viswanathan, Physical Review B 2016, 94, 064105.

[509]Z. Deng, Z. Wang, I.-H. Chu, J. Luo, S. P. Ong, Journal of The Electrochemical Society 2015, 163, A67.

[510]X. Chen, X. Shen, B. Li, H.-J. Peng, X.-B. Cheng, B.-Q. Li, X.-Q. Zhang, J.-Q. Huang, Q. Zhang, Angewandte Chemie International Edition 2018, 57, 734.

[511]X. Chen, H.-R. Li, X. Shen, Q. Zhang, Angewandte Chemie International Edition 2018, 57, 16643.

[512]M. Okoshi, Y. Yamada, A. Yamada, H. Nakai, Journal of The Electrochemical Society 2013, 160, A2160.

[513]M. Okoshi, Y. Yamada, S. Komaba, A. Yamada, H. Nakai, Journal of The Electrochemical Society 2016, 164, A54.

[514]K. Sodeyama, Y. Igarashi, T. Nakayama, Y. Tateyama, M. Okada, Physical Chemistry Chemical Physics 2018, 20, 22585.

[515]Y. Igarashi, K. Nagata, T. Kuwatani, T. Omori, Y. Nakanishi-Ohno, M. Okada, Journal of Physics: Conference Series 2016, 699, 012001.

[516]Y. Igarashi, H. Takenaka, Y. Nakanishi-Ohno, M. Uemura, S. Ikeda, M. Okada, Journal of the Physical Society of Japan 2018, 87, 044802.

[517]Y. S. Meng, M. E. Arroyo-de Dompablo, Energy & Environmental Science 2009, 2, 589.

[518]M. Nishijima, T. Ootani, Y. Kamimura, T. Sueki, S. Esaki, S. Murai, K. Fujita, K. Tanaka, K. Ohira, Y. Koyama, I. Tanaka, Nature Communications 2014, 5, 4553.

[519]Y. S. Meng, M. E. Arroyo-de Dompablo, Accounts of Chemical Research 2013, 46, 1171.

[520]L.-M. Yan, J.-M. Su, C. Sun, B.-H. Yue, Advances in Manufacturing 2014, 2, 358.

[521]Y. Liu, B. Guo, X. Zou, Y. Li, S. Shi, Energy Storage Materials 2020, 31, 434.

[522]A. Jain, G. Hautier, C. J. Moore, S. Ping Ong, C. C. Fischer, T. Mueller, K. A. Persson, G. Ceder, Computational Materials Science 2011, 50, 2295.

[523]Q.-S. Xu, Y.-Z. Liang, Y.-P. Du, Journal of Chemometrics 2004, 18, 112.

[524]K. Pearson, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 1901, 2, 559.

[525]I. Jolliffe, in *International Encyclopedia of Statistical Science*, (Ed: M. Lovric), Springer Berlin Heidelberg, Berlin, Heidelberg 2011, 1094.

[526]W. S. Noble, Nature Biotechnology 2006, 24, 1565.

[527]X. Zhang, Z. Zhang, S. Yao, A. Chen, X. Zhao, Z. Zhou, npj Computational Materials 2018, 4, 13.

[528]S. P. Ong, V. L. Chevrier, G. Hautier, A. Jain, C. Moore, S. Kim, X. Ma, G. Ceder, Energy & Environmental Science 2011, 4, 3680.

[529]J. Billaud, R. J. Clément, A. R. Armstrong, J. Canales-Vázquez, P. Rozier, C. P. Grey, P. G. Bruce, Journal of the American Chemical Society 2014, 136, 17243.

[530]U. Nisar, R. A. Shakoor, R. Essehli, R. Amin, B. Orayech, Z. Ahmad, P. R. Kumar, R. Kahraman, S. Al-Qaradawi, A. Soliman, Electrochimica Acta 2018, 292, 98.

[531]X. Wang, R. Xiao, H. Li, L. Chen, Journal of Materiomics 2017, 3, 178.

[532]Z. Chen, L. Christensen, J. R. Dahn, Electrochemistry Communications 2003, 5, 919.

[533]P. Simon, Y. Gogotsi, Nature Materials 2008, 7, 845.

[534]L. L. Zhang, X. S. Zhao, Chemical Society Reviews 2009, 38, 2520.

[535]J. Chmiola, G. Yushin, Y. Gogotsi, C. Portet, P. Simon, P. L. Taberna, Science 2006, 313, 1760.

[536]A. Chen, Y. Yu, T. Xing, R. Wang, Y. Li, Y. Li, Materials Letters 2015, 157, 30.

[537]S. Zhu, J. Li, C. He, N. Zhao, E. Liu, C. Shi, M. Zhang, Journal of Materials Chemistry A 2015, 3, 22266.

[538]A. Giwa, A. Yusuf, H. A. Balogun, N. S. Sambudi, M. R. Bilad, I. Adeyemi, S. Chakraborty, S. Curcio, Process Safety and Environmental Protection 2021, 146, 220.

[539]P. R. Zonouz, A. Niaei, A. Tarjomannejad, Journal of the Taiwan Institute of Chemical Engineers 2016, 65, 276.

[540]B. Vaferi, M. Bahmani, P. Keshavarz, D. Mowla, Journal of Environmental Chemical Engineering 2014, 2, 1252.

[541]A. Aleboyeh, M. B. Kasiri, M. E. Olya, H. Aleboyeh, Dyes and Pigments 2008, 77, 288.

[542]A. R. Soleymani, J. Saien, H. Bayat, Chemical Engineering Journal 2011, 170, 29.

[543]S. T. Bararpour, M. R. Feylizadeh, A. Delparish, M. Qanbarzadeh, M. Raeiszadeh, M. Feilizadeh, Journal of Cleaner Production 2018, 176, 1154.

[544]A. R. Amani-Ghadim, M. S. S. Dorraji, Applied Catalysis B: Environmental 2015, 163, 539.

[545]L. Jing, B. Chen, D. Wen, J. Zheng, B. Zhang, Journal of Environmental Management 2017, 203, 182.

[546]H. Park, H.-i. Kim, G.-h. Moon, W. Choi, Energy & Environmental Science 2016, 9, 411.

[547]A. Dayan, R. Mor Yosef, J. Risphon, E. Tuval, G. Fleminger, The Journal of Physical Chemistry A 2019, 123, 9456.

[548]V. Augugliaro, G. Camera-Roda, V. Loddo, G. Palmisano, L. Palmisano, J. Soria, S. Yurdakal, The Journal of Physical Chemistry Letters 2015, 6, 1968.

[549]L. Zeng, X. Li, S. Fan, Z. Yin, M. Zhang, J. Mu, M. Qin, T. Lian, M. Tadé, S. Liu, Electrochimica Acta 2019, 295, 810.

[550]C. Zhai, M. Zhu, Y. Lu, F. Ren, C. Wang, Y. Du, P. Yang, Physical Chemistry Chemical Physics 2014, 16, 14800.

[551]Y. Di, C. Ma, Y. Fu, X. Dong, X. Liu, H. Ma, ACS Applied Materials & Interfaces 2021, 13, 8405.

[552]R. Parveen, T. R. Cundari, J. M. Younker, G. Rodriguez, L. McCullough, ACS Catalysis 2019, 9, 9339.

[553]J. D. L. Dutra, J. W. Ferreira, M. O. Rodrigues, R. O. Freire, The Journal of Physical Chemistry A 2013, 117, 14095.

[554]L. Buglioni, F. Raymenants, A. Slattery, S. D. A. Zondag, T. Noël, Chemical Reviews 2021.

[555]J. Zhang, T. Zhou, L. Wen, ACS Applied Materials & Interfaces 2017, 9, 8996.

[556]M. C. McAlpine, H. Ahmad, D. Wang, J. R. Heath, Nature Materials 2007, 6, 379.

[557]K. A. Homan, J. Chen, A. Schiano, M. Mohamed, K. A. Willets, S. Murugesan, K. J. Stevenson, S. Emelianov, Advanced Functional Materials 2011, 21, 1673.

[558]T. J. Dennes, J. Schwartz, ACS Applied Materials & Interfaces 2009, 1, 2119.

[559]S. Das Sarma, S. Adam, E. H. Hwang, E. Rossi, Reviews of Modern Physics 2011, 83, 407.

[560]R. Frisenda, E. Navarro-Moratalla, P. Gant, D. Pérez De Lara, P. Jarillo-Herrero, R. V. Gorbachev, A. Castellanos-Gomez, Chemical Society Reviews 2018, 47, 53.

[561]I. I. Mazin, Nature 2010, 464, 183.

[562]N. B. Kopnin, T. T. Heikkilä, G. E. Volovik, Physical Review B 2011, 83, 220503.

[563]S. Peotta, P. Törmä, Nature Communications 2015, 6, 8944.

[564]O. Isayev, D. Fourches, E. N. Muratov, C. Oses, K. Rasch, A. Tropsha, S. Curtarolo, Chemistry of Materials 2015, 27, 735.

[565]Y. T. Sun, H. Y. Bai, M. Z. Li, W. H. Wang, The Journal of Physical Chemistry Letters 2017, 8, 3434.

[566]M. K. Tripathi, S. Ganguly, P. Dey, P. P. Chattopadhyay, Computational Materials Science 2016, 118, 56.

[567]L.-M. Wang, Y. Tian, R. Liu, W. Wang, Applied Physics Letters 2012, 100, 261913.

[568]M. K. Tripathi, P. P. Chattopadhyay, S. Ganguly, Computational Materials Science 2015, 107, 79.

[569]H. Tin Kam, IEEE Transactions on Pattern Analysis and Machine Intelligence 1998, 20, 832.

[570]W. L. Johnson, Progress in Materials Science 1986, 30, 81.

[571]A. L. Greer, Science 1995, 267, 1947.

[572]Y. Kawazoe, T. Masumoto, A. P. Tsai, J. Z. Yu, T. Aihara Jr, Springer-Verlag Berlin Heidelberg.

[573]B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, C. Wolverton, Physical Review B 2014, 89, 094104.

[574]S. Sanvito, C. Oses, J. Xue, A. Tiwari, M. Zic, T. Archer, P. Tozman, M. Venkatesan, M. Coey, S. Curtarolo, Science Advances 2017, 3, e1602241.

[575]T. Jungwirth, J. Sinova, A. Manchon, X. Marti, J. Wunderlich, C. Felser, Nature Physics 2018, 14, 200.

[576]Y. Xu, M. Yamazaki, P. Villars, Japanese Journal of Applied Physics 2011, 50, 11RH02.

[577]A. P. Bartók, R. Kondor, G. Csányi, Physical Review B 2013, 87, 184115.

[578]G. E. Fish, Proceedings of the IEEE 1990, 78, 947.

[579]J. A. Bas, J. A. Calero, M. J. Dougan, Journal of Magnetism and Magnetic Materials 2003, 254-255, 391.

[580]A. M. Leary, P. R. Ohodnicki, M. E. McHenry, JOM 2012, 64, 772.

[581]Y. Yoshizawa, S. Oguma, K. Yamauchi, Journal of Applied Physics 1988, 64, 6044.

[582]M. A. Willard, M. Daniil, in *Handbook of Magnetic Materials*, Vol. 21 (Ed: K. H. J. Buschow), Elsevier, 2013, 173.

[583]G. Herzer, Acta Materialia 2013, 61, 718.

[584]J. M. Howard, E. M. Tennyson, B. R. Neves, M. S. Leite, Joule 2019, 3, 325.

[585]G. L. Hart, T. Mueller, C. Toher, S. Curtarolo, Nature Reviews Materials 2021, 6, 730.

[586]C. B. Wahl, M. Aykol, J. H. Swisher, J. H. Montoya, S. K. Suram, C. A. Mirkin, Science advances 2021, 7, eabj5505.

[587]M. He, L. Zhang, Computational Materials Science 2021, 196, 110578.

[588]R. E. Goodall, A. A. Lee, Nature communications 2020, 11, 1.

[589]D. Jha, K. Choudhary, F. Tavazza, W.-k. Liao, A. Choudhary, C. Campbell, A. Agrawal, Nature communications 2019, 10, 1.

[590]S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, C. Wolverton, npj Computational Materials 2015, 1, 1.

[591]C. J. Bartel, A. Trewartha, Q. Wang, A. Dunn, A. Jain, G. Ceder, npj Computational Materials 2020, 6, 1.

[592]G. R. Schleder, C. M. Acosta, A. Fazzio, ACS applied materials & interfaces 2019, 12, 20149.

[593]F. A. Faber, A. Lindmaa, O. A. Von Lilienfeld, R. Armiento, Physical review letters 2016, 117, 135502.

[594]C. J. Bartel, S. L. Millican, A. M. Deml, J. R. Rumptz, W. Tumas, A. W. Weimer, S. Lany, V. Stevanović, C. B. Musgrave, A. M. Holder, Nature communications 2018, 9, 1.

[595]F. Häse, L. M. Roch, A. Aspuru-Guzik, Trends in Chemistry 2019, 1, 282.

[596]V. Kumar, A. Kumar, D. Chhabra, P. Shukla, Bioresource Technology 2019, 271, 274.

[597]P. Sakiewicz, K. Piotrowski, J. Ober, J. Karwot, Renewable and Sustainable Energy Reviews 2020, 124, 109784.

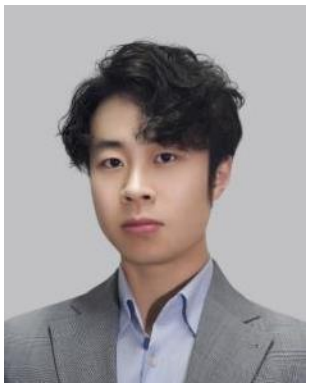
[598]A. Dunn, Q. Wang, A. Ganose, D. Dopp, A. Jain, npj Computational Materials 2020, 6, 138.

[599]R. Zubatyuk, J. S. Smith, J. Leszczynski, O. Isayev, Science Advances 2019, 5, eaav6490.

[600]T. Baltrušaitis, C. Ahuja, L. Morency, IEEE Transactions on Pattern Analysis and Machine Intelligence 2019, 41, 423.

[601]T. Zubatiuk, O. Isayev, Accounts of Chemical Research 2021, 54, 1575.

[602]A. Belsky, M. Hellenbrandt, V. L. Karen, P. Luksch, Acta Crystallographica Section B 2002, 58, 364.



Zhuo Wang is currently a Ph.D. student at the department of chemical and environmental engineering, University of Nottingham, Ningbo, China. He received his M.Sc. in Scientific Computing from the University College London in 2018 and B.Eng. (Hons) in Chemical Engineering from the University of Nottingham in 2017. His research interests include material informatics, machine learning, energy materials, and carbon neutrality.

Zhehao Sun received his B.Eng. degree in Thermal Engineering from the Dalian University of Technology in 2017 and obtained his M. Eng in Energy and Environmental Engineering from the Dalian University of Technology with Prof. Dawei Tang in 2020. He is currently a Ph.D. candidate in the Research School of Chemistry, Australian National University. His research interests include functional nanomaterials, first-principles calculations, machine learning, and solid-state physics.

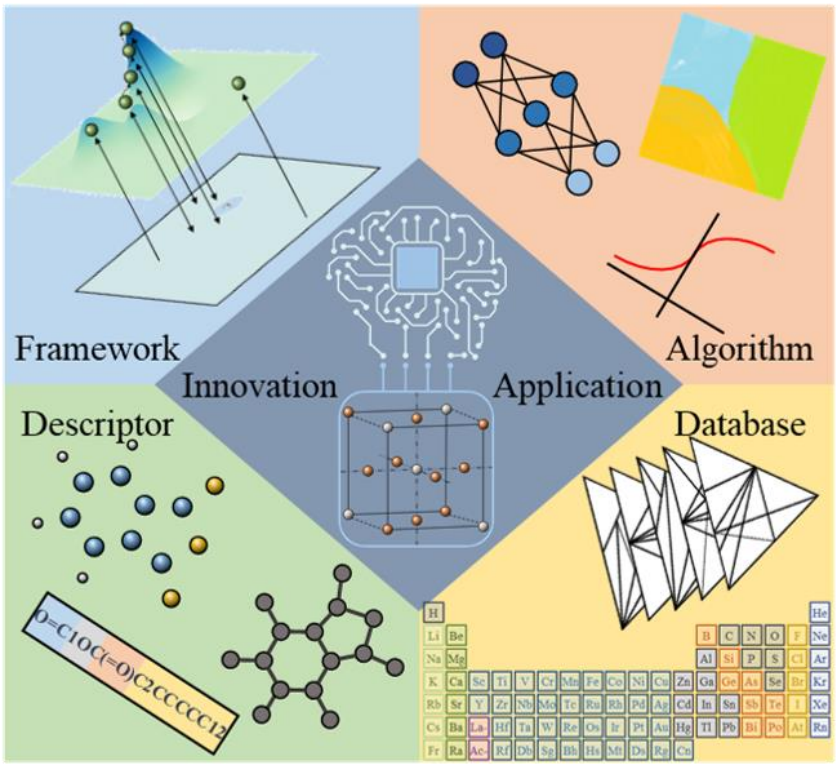
Dr. Haitao Zhao is an Associate Professor at Materials Interfaces Center, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. He completed his doctoral degree in chemical and environmental engineering from the University of Nottingham. Dr. Zhao has then conducted postdoctoral research at Zhejiang University and Massachusetts Institute of Technology. He is conducting interdisciplinary research on development of functional materials automation platform based on data mining, high-throughput DFT calculations, machine learning, robot chemist or in situ characterization techniques for accelerating materials discovery.

Professor Cheng Heng PANG is a Professor of Chemical Engineering and Advanced Materials at the University of Nottingham Ningbo China. He is the Director of Ningbo Key Laboratory on Clean Energy Conversion Technologies. Prof. Pang obtained his PhD from the University of Nottingham UK, and graduated from Nottingham Malaysia with a First Class degree in Chemical Engineering and top in his class. He now works at the interphase, and integration, between science and engineering with particular interests on deriving clean energy and advanced functional materials from renewable sources. Prof. Pang has received several awards including the Ten Outstanding Young Malaysians Award.

Professor Tao Wu is the Vice Provost for China Beacons Institute and the Dean of the Faculty of Science and Engineering at the University of Nottingham Ningbo China. He is also the Director of the Zhejiang Provincial Key Laboratory for Carbonaceous Wastes Processing and Process Intensification Research. Prof. Wu received his Ph.D. degree in chemical engineering from the University of Nottingham UK and has over 25 years of experience in clean energy conversion technologies and

new materials for environmental applications. His research covers a wide range from blue-sky research to proof of concept and patent development leading to commercialization.

Associate Professor Zongyou Yin obtained his BS and MS degrees at Jilin University in China, and completed his Ph.D. at Nanyang Technological University (NTU) in Singapore. Then, he started his postdoc careers at NTU/Singapore, IMRE/Singapore, followed by MIT and then Harvard University. Dr Yin started his own research group at Australian National University (ANU) from 2017. His group's research is interdisciplinary, encompassing the chemistry and physics of nano-to-atomic materials, fundamental relationship among materials-structures-devices, and synergistic integration of multi-functions towards systems for energy and wearable applications. He has been selected as one of the world's most highly-cited researchers in 2021 by Clarivate.



This review aims to provide a timely and critical discussion on the recent advances, strategies, insights, and challenges of data-driven-based innovations and applications in material science. Essential sub-disciplines, including framework, machine learning algorithms, available chemical databases, commonly used key descriptors, and innovations and applications based on their synergy, are reviewed.