CHEMICAL REVIEWS

pubs.acs.org/CR

Review

Machine Learning for Electrocatalyst and Photocatalyst Design and Discovery

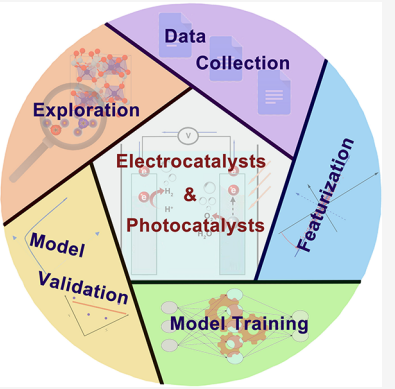
Haoxin Mai, Tu C. Le, Dehong Chen,* David A. Winkler,* and Rachel A. Caruso*

Cite This: https://doi.org/10.1021/acs.chemrev.2c00061

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Electrocatalysts and photocatalysts are key to a sustainable future, generating clean fuels, reducing the impact of global warming, and providing solutions to environmental pollution. Improved processes for catalyst design and a better understanding of electro/photocatalytic processes are essential for improving catalyst effectiveness. Recent advances in data science and artificial intelligence have great potential to accelerate electrocatalysis and photocatalysis research, particularly the rapid exploration of large materials chemistry spaces through machine learning. Here a comprehensive introduction to, and critical review of, machine learning techniques used in electrocatalysis and photocatalysis research are provided. Sources of electro/photocatalyst data and current approaches to representing these materials by mathematical features are described, the most commonly used machine learning methods summarized, and the quality and utility of electro/photocatalyst models evaluated. Illustrations of how machine learning models are applied to novel electro/photocatalyst discovery and used to elucidate electrocatalytic or photocatalytic reaction mechanisms are provided. The review offers a guide for materials scientists on the selection of machine learning methods for electrocatalysis and photocatalysis research. The application of machine learning to catalysis science represents a paradigm shift in the way advanced, next-generation catalysts will be designed and synthesized.



CONTENTS

Received: January 21, 2022

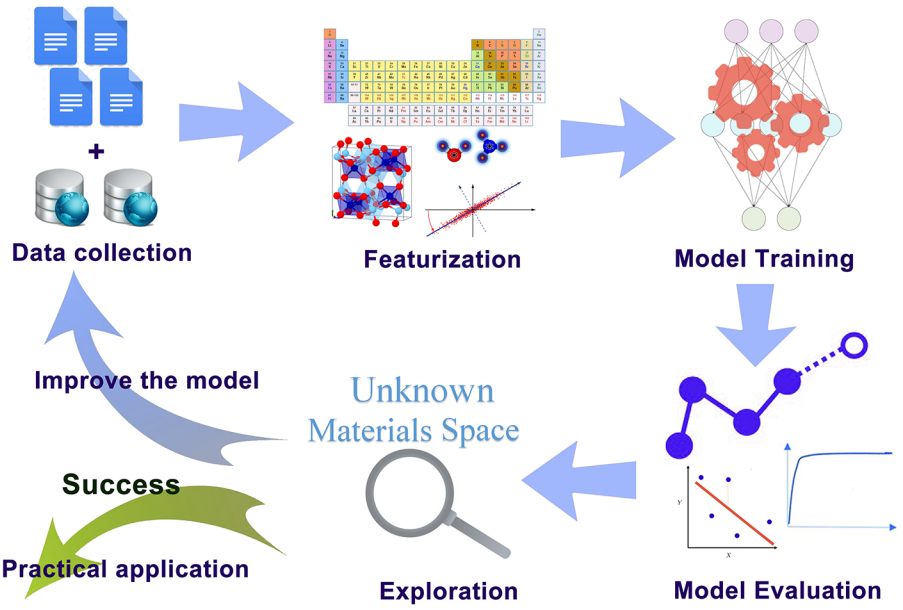


Figure 1. Workflow for applying ML techniques in the development of catalytic materials. It starts with data collection, featurization, model training, model evaluation, and eventually exploring the unknown materials space by using the ML models. The results of exploration will feed into the original data set for model improvement, and the materials with good properties will proceed to practical application.

1. INTRODUCTION

The development of globally sustainable energy systems and environmental protection are two critical challenges for this century, as population and industrialization expand.[1] In particular, the generation of high-purity fuels from clean sources is a major research and development focus.[2−4] Producing chemical fuels from earth-abundant resources such as water, carbon dioxide, and nitrogen requires next-generation electrocatalysts and photocatalysts.[5−10] Removal of pollutants, bacteria, and heavy metals from water using cheap, efficient processes is also essential for maintaining an inhabitable planet, where photocatalytic and electrocatalytic reactions play important roles.[10−15] Clearly, the discovery of higher efficiency, nontoxic electrocatalysts and photocatalysts synthesized from abundant elements is critical to realizing this future. However, the number of catalysts that can be synthesized in principle is vast, much too large to be explored through laboratory experimentation alone.[16−18] To accelerate catalyst discovery, first-principles computational chemistry, especially density functional theory (DFT), is being increasingly employed.[19,20] In particular, the recent development of high throughput DFT calculations enables in silico catalyst screening and exploration of unknown material spaces.[21−25] However, because of the complexity of catalytic materials and the breadth of their applications, computationally expensive and time-consuming DFT calculations are still unable to explore the whole catalyst composition space. Catalyst researchers are putting intensive effort into understanding structure−property relationships, interpreting data patterns, and extracting knowledge from high volume, multidimensional material databases.[26] More effective strategies are therefore needed to tackle these complex issues.

Machine learning (ML) is a suite of data-driven computational methods that are ideally suited for modeling and interpreting vast amounts of complex multidimensional data.[27] ML is a subset of artificial intelligence (AI) that uses a range of algorithms to build models based on patterns in data that can be used to make predictions.[27−29] ML algorithms have recently been shown to accelerate discovery and the optimization of functional materials for real world applications such as catalysts, photovoltaics, optoelectronic materials, and thermoelectrics.[26,30−36] ML algorithms complement expensive, time- and resource-intensive first-principles electronic structure calculations to provide insight into complex, multidimensional structure−property relations.[31,37−46] ML techniques can interpret and model a wide range of data and are useful for de novo discovery of important physical and chemical principles.[47−51] The application of ML to materials science will result in a step-change in theory and methods for the discovery of next-generation functional materials.

Here we review recent progress (2016 onward) in the application of ML to the discovery, development, and design of electrocatalysts and photocatalysts. Excellent reviews exist that summarize the impact of ML in general catalysis informatics for a wide range of catalytic processes or for one or two specific reactions (e.g., the electrocatalytic hydrogen evolution reaction (HER)).[35,52−55] Recent, comprehensive, critical reviews of cutting-edge ML algorithms and their applications to electrocatalysts and photocatalysts are lacking. To address this deficiency, we summarize ML methods commonly used for electrocatalyst and photocatalyst discovery, critically review the literature in this field, and provide a perspective into the future evolution of AI and ML technology and its impact on electro/photocatalyst research and development. This review consists of four sections. Section 1 introduces ML and its application to materials science. In Section 2, the process of implementing ML methods for electro/photocatalytic systems is described and a concise summary of ML algorithms and their advantages and limitations is provided. Section 3 reviews the literature on ML-based electro/photocatalyst design, discovery, and development, and exemplifies this method for materials, such as metals and oxides. Section 4 provides conclusions and a perspective on current challenges and future developments in ML and ML-based catalyst design. This review aims to guide materials scientists in the selection of ML methods for electro/

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

Table 1. List of Popular Structure and Property Databases Used in Materials Science

Name	Information provided	Data source	URL
AFLOWLIB	Structure and properties	High throughput calculation	http://aflowlib.org/
ChemSpider	Chemical structure	Calculation and experiments	http://www.chemspider.com
CMR	Infrastructure supporting the collection, storage, retrieval, analysis, and sharing of data produced by many electronic-structure simulators	Calculation	https://cmr.fysik.dtu.dk/
COD	Structures of organic, inorganic, metal−organics compounds and minerals, excluding biopolymers	Published and unpublished structures	http://crystallography.net/cod/
CSD	Small molecule crystal structures	Experiments	https://www.ccdc.cam.ac.uk/
Harvard Clean Energy Project	Properties of organic solar compounds	High throughput calculation	http://web.archive.org/web/20130627085223/cepdb.molecularspace.org/
HTEM	Properties of thin films	Experiments	https://htem.nrel.gov/
ICSD	Inorganic Crystal Structure Database	Published structures	https://icsd.fiz-karlsruhe.de
Khazana	Structure and property, tools to design materials by learning from the data	Calculation	https://khazana.gatech.edu/
Materials Project	Properties of known and predicted materials	Calculation using standard calculation scheme	https://materialsproject.org/
MatNavi	Crystal structures, electronic structures, properties, and phase diagrams for polymers and inorganic materials	Calculation and experiments	https://mits.nims.go.jp/
MatWeb	Data sheets of thermoplastic and thermoset polymers, metals, ceramics, and other engineering materials	Experiments	http://matweb.com/
NOMAD	User-driven platform for sharing and exploiting computational materials science data	Calculation	https://nomad-coe.eu/
NREL Materials Database	Properties of materials for renewable energy applications (photovoltaics, materials for photoelectrochemical water splitting, thermoelectrics)	Calculation	https://materials.nrel.gov/
Organic Materials Database	Electronic structure database for 3D organic crystals	Calculation	https://omdb.mathub.io/
Open Quantum Materials Database	Thermodynamic and structural properties	DFT calculation	http://oqmd.org/
ZINC	2D and 3D structures of commercially available molecules	Calculation	https://zinc15.docking.org/

photocatalyst design, allowing them to accelerate catalyst discovery for a sustainable future.

2. MACHINE LEARNING MODELING

In principle, machine learning algorithms are universal approximators, able to model any complex relationship in physical systems, given sufficient data. This approach is quite different to traditional computational methods such as DFT, which require the laws of quantum mechanics and chemistry to be explicitly coded into the software. In contrast, ML algorithms learn the underlying rules and patterns in a data set to generate a model capable of making predictions about new molecules or materials (objects).[46,56,57] The basic steps in constructing a ML model are the following (Figure 1): collect data to form a training data set; generate and select relevant mathematical descriptors that encode the properties of the materials; choose a suitable algorithm and build the model; and evaluate the quality and predictive power of the model. These steps are described in detail in the following sections.

2.1. Data Collection

The key challenge for material informatics is obtaining sufficient, reliable materials data. This data set must contain well-defined input variables for the model (microscopic structures, properties, or synthesis conditions of the material) and macroscopic properties of interest that the model is trained to predict.[16,57] There are three main sources of materials data: structure and property databases; materials properties acquired from experimental measurements or computation; and the literature base. Table 1 summarizes databases frequently used for materials informatics studies. Despite strict quality checks, errors and bias from human perception and measurement can exist in these databases.[58] Moreover, variability may be introduced when experiments are carried out under different conditions (in different laboratories) or the properties are characterized by different methods (e.g., nanoparticle sizes determined by electron microscopy or dynamic light scattering).

Another important data source is the literature.[59] The near exponential rise in numbers of papers published provides a huge resource materials scientists can access to build data sets. The structures and properties of the materials contained in these papers, synthesis and processing conditions, and other provenance data are extremely valuable for elucidating materials structure−property relationships. The literature often contains analyses of the connections and relationships between the relevant properties. This deeper knowledge, in contrast to the purely quantitative property and structure data in databases, can give materials scientists deeper insights into structure−property relationships and identify the most useful descriptors used to represent the materials.[60] Traditionally, publication databases are searched by keywords and the relevant information extracted manually; this is inefficient because of the number of publications. Chemical information is being mined from the literature using ML-based text extraction methods (natural language processing).[60−64] These methods are robust for batch extraction of chemical information from publications. However, the reliability of the published data from literature must be carefully assessed. Data from different sources, especially for catalysts, may not be compatible because the different ways catalytic reactions are

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

Table 2. Typical Descriptors Used in a Variety of Material Systems for Various Purposes

Materials	Target	Descriptors	Ref
Intermetallics	Prediction possible compositions	Atomic fingerprints	126
Intermetallics	Electrochemical CO_2 reduction	d-Band features, electronic descriptors (electronegativity, work function, ionization potential, electron affinity, etc.)	88
Intermetallics	Ethanol desorption	Physical descriptors	127
Intermetallics	CO adsorption energy	Geometric fingerprint to represent the local region	128, 129
Intermetallics	Acquire structure information	EXAFS	130
Oxide electrocatalysts	OER activity	Covalency, electron occupancy, d electron number, charge-transfer energy, occupancy of e_g orbitals, metal-oxide-metal bond angles and tolerance factor	131
Oxide electrocatalysts	OER activity	Tolerance factor and the octahedral factor	132
Oxide electrocatalysts	OER activity	Elemental and structural descriptors	133
SACs in N-doped graphene	H_2 adsorption free energy	Covalent radius, d-states center, electronegativity, the number of occupied d states and Bader charge	134
Graphdyine-based SACs	H_2 adsorption free energy	Atomic descriptors, active sites and redox energy barrier	135
Co SACs in N-doped graphene	HER activity	EXAFS	136
SACs in N-doped graphene	HER/ORR/OER limiting potential	The electron numbers of the d orbital, the oxide formation enthalpy of the single atom, Pauling electronegativity of the center metal atom, the sum of Pauling electronegativity of surrounding atoms, and the average of the pK_a values of the surrounding atoms (hydride formation enthalpy of the single atom for HER model)	137
Transition metal SACs in g-C_3N_4	ORR/OER activity	First ionization energy and charge transfer of transition metal atoms	138
Fe SACs in zeolites	ORR activity	Experimental descriptors	139
Transition metal SACs in B-doped graphene	N_2 adsorption energy	Coulomb matrix, electronic descriptors, elemental descriptors	140
Double layer transition metal dichalcogenides	HER/OER overpotentials	Rotational angle, bond length, distance between layers	141
MXenes (general formula $M_{n+1}X_n$ (M, element from group IIIB to VIB; X, C or N; n, 1−3))	HER activity	The bond length of oxygen and surface metal atoms, the distance between the nearest neighbor O atoms, the ionization energy difference, the average affinity energy of the alloy elements, and the valence electrons of X	142
MBenes (element X in MXenes is replaced by boron)	H_2 adsorption free energy	Elemental descriptors, structural energies and lattice parameters	143
Photocatalytic polymer	H_2 evolution	Electron affinity, ionization potential, bandgap, transmittance	144
Photocatalytic polymer	Bandgap and band structure	The frontier orbitals of the donor and acceptor components in the backbone	145
Photocatalytic oxide (perovskites)	Bandgap and HER activity	Elemental and electronic descriptors, tolerance factor	146
Photocatalytic oxide (TiO_2)	Parameter optimization for pollutant degradation	Experimental parameters	147, 148

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

measured and because some potentially relevant factors may be neglected (e.g., shape of reactor, stirring speed, different solvent). Close cooperation among materials scientists, catalyst scientists, and computer scientists on the use of a standard format for data recording and evaluation of the reliability of published data is very important for generating high-quality ML models.

Materials data sets can also be built from experimental and simulation data, where all the data are acquired under similar conditions. Typically, high throughput experiments generate large quantities of data for a specific system.[21,22,24,25,65−68] Conspicuously, all experiments contribute to data set construction, as failed experiments contain valuable information identifying the molecular determinants of success and failure.[69] A web-accessible public database (http://darkreactions.haverford.edu) has been developed to extract data from laboratory notebooks and ongoing experimental data.[70] These data from failed experiments have been beneficial for the design of chemical syntheses.

High throughput computation is also a viable way to accumulate materials data.[22,71−75] Using thermodynamic and electronic structure methods, data sets of virtual materials or local properties of materials can be generated that are useful for materials analysis and discovery using ML tools.[30,74] ML can predict the results of electronic structure calculations when trained on previous calculation results, effectively leveraging these expensive and resource-intensive first principle calculations much more broadly across materials space.[76,77] High throughput calculations accelerate the growth of computed materials databases such as AFLOWLIB and Materials Project. These databases also provide open-source platforms for researchers to implement high throughput calculations.[78,79]

2.2. Featurization

After data collection, materials must be transformed into an appropriate mathematical form used to train ML models.[80] This involves describing key attributes of molecules, compounds, clusters on surfaces of compounds, materials etc., by a series of numbers (vectors or tensors).[80,81] These numbers are materials attributes called descriptors, features, or fingerprints in the literature. The process of generating these mathematical representations of material objects is called featurization or feature engineering and it is of critical importance to the quality and interpretability of the models trained using them. This step required the most human intuition and intervention but recent development in automatic featurization of molecular objects using specific types of deep neural networks promises paradigm shifts in this process.[82−84] Featurization consists of two steps: descriptor generation and descriptor selection.

2.2.1. Descriptor Generation. Effective descriptors must discriminate between objects in the data space and encode attributes that are relevant to the property being modeled and predicted.[31,34] Despite the large number of descriptors that can be generated for molecules and materials, the choice of these descriptors is strongly dependent on the scientific problem to be solved.[31] Although descriptor generation is context dependent, there are still some common rules. First, the descriptor set must provide unique information, such as structure, composition, and physicochemical properties of the materials to ensure they are distinct for each material. Second, the number of descriptors used should not be very large to avoid overfitting models and compromising their ability to predict properties of new data. Redundant or correlated descriptors should be removed to minimize the size of the descriptor set to avoid overfitting, especially in electro/photocatalyst research where the data sets tend to be small. It has been shown that as additional irrelevant descriptors are added to descriptor sets, the quality of a partial least-squares model, for example, is degraded.[85] Redundant descriptors include those having high correlation with other descriptors in the set, those whose values do not change much across the data set (low variance), and those that have very low correlation with the property being modeled. To avoid overfitting, the number of fitted variables in a model should be much less than half of the number of the training examples.[16,86]

In general, descriptors can be divided into different classes depending on the information they encode. As well as encoding structural, compositional, and physicochemical properties of materials, they can also include provenance information on how the material was synthesized or subsequently processed. Experimentally derived descriptors may include temperature, pH value, pressure, reaction time, and the amount of the reactants. Compositional descriptors relate to the composition of each component in complex materials. Topographical descriptors pertain to the texture of a material (pore size, volume, surface area, topological shape), while topological descriptors encode the connectivity of atoms in materials. Some descriptors can be readily extracted from physicochemical handbooks or material databases. For example, atomic descriptors refer to the atomic attributes like atomic radii, masses, the number of protons and valence electrons, etc., physical descriptors are mainly related to the macroscopic attributes of materials such as melting point, boiling point, density and solubility, and some electronic descriptors encode microscopic attributes (electronegativity, ionization energy, etc.).[87] Many types of descriptors can be obtained by calculation, such as structural descriptors and some electronic descriptors (bandgap, dipole moment, highest occupied molecular orbital (HOMO), lowest occupied molecular orbital (LUMO), d-band characteristics, etc.).[88,89] Spectra-based descriptors are derived from experimental measurements (e.g., X-ray diffraction, extended X-ray absorption fine structure (EXAFS), etc.).[90] The acquisition of these descriptors may be relatively costly and time-consuming. Clearly, computational descriptors are preferred because they do not require experiments to derive them. Apart from cost and time considerations, more importantly, they are the only way to predict properties of materials not yet synthesized.

Table 2 summarizes typical descriptors used in a variety of electrocatalysis and photocatalysis models. Although the relevance of descriptors is assessed by the performance of the model, experience and intuition are useful in descriptor selection. For example, experimental descriptors would be used in the case of synthesis optimization,[91] while compositional descriptors would be better suited to describe the materials undergoing composition optimization.[92] When texture−property relationships or structure−property relationships are interpreted, topographical and topological descriptors, atomic descriptors, and electronic descriptors would be used collectively.[87,93−96] For instance, electronegativity and ion radii are commonly used as descriptors for different perovskite catalysts,[97−99] as these features play pivotal roles in Pauling's Rule and tolerance factor that are useful in understanding and predicting the stability of perovskites.[100,101]

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

Descriptors must be chosen with care to ensure they are relevant to the property or process being modeled. For example, electronic and atomic descriptors are not able to differentiate between allotropes, isomers, conformers, and polymorphs that have very similar descriptors for each form. Techniques such as radial distribution functions (density distributions as a function of distance) and Voronoi tessellations (partition the crystal lattice into subregions close to each of a given set of objects) are useful in encoding the crystal lattice.[102,103] The Coulomb matrix is a global descriptor mimicking the electrostatic interaction between the nuclei and is widely used in structure featurization,[104] especially in adsorbate research.[89] Structures of molecules can be described by the simplified molecular-input line-entry system (SMILES),[105] a string of characters composed of letters and symbols that can be converted into numbers using the ASCII values of the characters.[35,106] More refined implementations have been developed based on SMILES. For example, PaDEL and RDKIt codes describe the molecular structure by an array of real numbers.[107,108] Molecular fingerprints encode the presence or absence of particular substructures/patterns in molecules (these differ in various implementations, e.g., MACCS, FP2, daylight, and hybridization) using 1s and 0s.[109−112] More generally, so-called one-hot descriptors denote the presence or absence of any feature or attribute in a molecule, or process.[113]

Generation of structural descriptors can also start from the local environment of each atom and extend to the crystal level. Atom-Centred Symmetry Functions (ACSFs) descriptors account for the local environment of atoms by using multiple two-body and three-body symmetry functions.[114] The Smooth Overlap of Atomic Positions (SOAP) likewise uses a local expansion of a Gaussian smoothed atomic density with orthonormal functions, based on spherical harmonics and radial basis functions, to encode regions of atomic geometries.[115−117] Note that both ACSFs and SOAP are local descriptors that do not represent the entire structure, but information from multiple local sites can be simply averaged or combined by kernel functions.[116] Frequently, d-band center descriptors and p-band center descriptors are used to describe the local activities of metal and anion sites of catalysts, respectively.[118] Moreover, structural descriptors can be built from the molecular graph, in which the nodes correspond to atoms and edges to bonds in crystal and molecular frameworks. These graph-based structural descriptors are widely used in materials systems because they are consistent with conventional representations of a unit cell or a molecule. Models trained on graph-based descriptors generally have good predictive performance for materials properties such as formation energy and bandgap.[88,119,120] They are simple to calculate, store, and interpret as only atomic symbols required, but their ability to model complex interactions within materials is limited.[121] Recently, chemical and physical properties were introduced into graph-based descriptors by the property-labeled materials fragments (PMF) technique.[122] Models built using PMF fragments exhibited excellent predictive ability for many important properties.

We stress that descriptors for catalysis research should not be restricted to one type. Multiple families of descriptors are required to capture all attributes of catalysts and reactions, leading to robust ML models (Table 2).[87,97,123,124] Wang et al. combined elemental descriptors and structural descriptors of oxide perovskite surfaces to build a surface center environment model to predict the adsorption energies of the intermediates (HO*, O*, HOO*) and their overpotentials. They found that the adsorption energies and overpotentials would be low when the B-site was occupied by a series of specific elements and accordingly, new oxide perovskites with good catalytic performance for the oxygen evolution reaction (OER) could be designed.[87] Xu et al. noted that catalytic activity of single-atom catalysts was highly correlated with the coordination number and the electronegativity of the metal center and electronegativity of the nearest atoms.[125] They therefore created a universal descriptor based on these descriptors, with which highly active single-atom catalysts lacking precious metals were designed. More recently, the use of state-of-the-art ML techniques, such as symbolic learning, have been shown to generate sets of descriptors that allow materials scientists to create more accurate ML models and uncover deep feature−property relationships.[395−397]

2.2.2. Descriptor Selection. Because there are a myriad of ways to construct descriptors for a material based on chemical and physical knowledge, it is crucial to choose the subset of the most appropriate descriptors for modeling. This is because using too many descriptors increases model complexity, resulting in overfitting and compromising the predictivity of models, while optimally sparse subsets of descriptors generate models with the best predictivity and make interpretation of models easier. There are two main strategies to select descriptors, down-selection and dimensional reduction. In down selection (also called sparse feature selection), a large number of possible material descriptors are reduced to a smaller number using a range of statistical methods. In regression models, an L_1 regularization term is introduced to penalize terms with a lower relevance to the model by shrinking them to zero, this is called the least absolute shrinkage and selection operator, LASSO (this will be discussed further in Section 2.3.1).[149] More recently, the sure independence screening and sparsifying operator (SISSO) has proven to tackle immense and correlated feature spaces to converge on an optimal combination of features relevant to the materials property of interest.[150] Tree algorithms, such as random forest (RF), are also commonly used to determine importance of each descriptor after training.[151,152] Wexler et al. used a regularized (model complexity penalizing) RF model to study hydrogen evolution activity of Ni_2P.[151] Twenty nine descriptors were input to the model, and the relative importance of each was estimated by calculating its normalized ability to separate the data based on the free energy for hydrogen adsorption. The ten most important descriptors were selected from the descriptor set. This identified the Ni−Ni bond length, the geometry of the Ni_3 sites, and the dopant charge as having the most significant effects on the hydrogen adsorption ability. Although down-selection methods are effective in reducing the number of descriptors, they can sometimes discard useful materials information. Some down-selection algorithms may display instability, especially if some descriptors are categorical rather than continuous, depending on how well the hyperparameters have been chosen.[153]

Another strategy for optimizing descriptors is dimensional reduction. The principle of this strategy is to project the descriptors from a high dimensional space into a lower dimensional space. The new descriptors are generated by linear combinations of the original descriptors, with principal component analysis (PCA) being the most common.[154] In brief, PCA determines the principal components, or a series of

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

orthogonal vectors, as new descriptors, on which the data point variances are maximum. The PCA algorithm speeds up the calculation of ML models (after dimensional reduction) and handles the curse of dimensionality.[155] Data point independence is largely retained after dimensional reduction; however, some useful information from data points can be lost in PCA, resulting in poorer model performance. As mentioned, because PCA retains all original descriptors, models trained on principal components lose performance when many of the original descriptors are uninformative (noise). PCA is commonly employed in modeling of materials properties where a large number of features have been generated, such as heterogeneous catalysts, photovoltaics, and supramolecular materials.[89,99,156,157] Given that PCA only builds linear projections but structure−property relationships are often nonlinear, additional learning algorithms have been developed to perform nonlinear dimension reduction. Kernel PCA uses standard PCA to perform nonlinear dimensional reduction by using a nonlinear transformation to project the data points into a higher dimension feature space where standard PCA can be used.[158] Manifold learning algorithms generate low-dimensional representations that retain neighborhood relationships in the original high-dimensional data, allowing visualization and modeling (Figure 2).[159,160] The t-distributed stochastic neighbor embedding (t-SNE) algorithm has been used extensively for the visualization of high-dimensional data in two or three dimensions.[120,161,162] However, the large time complexity of this algorithm, and variations in data visualizations that are highly dependent on the selection of hyperparameters, limit its applications. An improved algorithm, uniform manifold approximation and projection (UMAP) is an efficient dimensional reduction technique for data visualization and nonlinear dimension reduction.[163] Compared to t-SNE, UMAP largely reduces the operation time and preserves the global structure of the data, thus it is widely used in catalyst screening and structure−property analysis.[164−166]

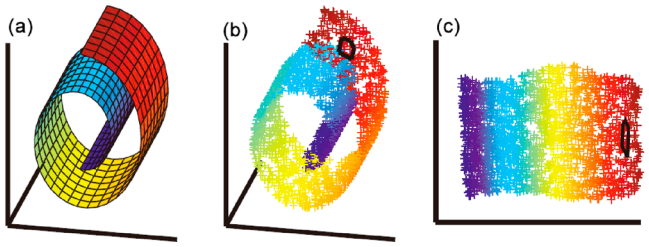


Figure 2. Schematic of manifold learning. (a) Two-dimensional manifold in three-dimensional space. (b) Three-dimensional data distributed on a two-dimensional manifold. (c) The unfolded manifold. Black outlines in panels b and c show the neighborhood of a single point. Reprinted with permission from ref 159. Copyright 2000, American Association for the Advancement of Science.

2.3. Algorithm Selection and Model Training

Once the optimum feature subset is selected, it can be used to train ML models using a wide variety of linear and nonlinear methods. Usually, data sets are partitioned into training sets used to generate models, and test sets not used in training that allow the predictive ability of models to be assessed (vide infra).[57] ML can be conducted in three ways depending on the intended use of the results and the structure of the training data: supervised learning; unsupervised learning; and semi-supervised learning.[31,46,56] Supervised learning models data with independent variables (descriptors) and dependent variables (labels, e.g., material properties). Regression and classification models aim to optimize the adjustable parameters in the model (regression coefficients or model weights) during training to minimize the prediction or classification errors in the training and test sets. The ML model learns the relationship between the input features and the output properties and subsequently predicts the properties of new materials not used to train the model. Such supervised ML methods are frequently used in materials science in place of resource-intensive physics-based calculations or time-consuming and expensive experiments to accelerate materials design and discovery.[43,45,46,73,151] When the training data consist solely of descriptors (unlabeled data), unsupervised learning methods are used to cluster the data and identify trends and patterns therein.[167,168] Outliers in the original data set can also be detected by unsupervised learning methods with a suitable choice of descriptors. If most data are unlabeled (e.g., only a few genes being annotated in a genomic data set), semi-supervised learning can label the unknown data and modify the model according to the knowledge gained from the labeled data set.[169] Currently, semisupervised learning is popular in peptide and gene identification and material synthesis design.[169−171]

In this section, we summarize some popular ML algorithms used in the design, discovery, and development of electro/photocatalysts (Figure 3). More comprehensive descriptions of these algorithms can be found in excellent textbooks and reviews.[31,34,172−174]

2.3.1. Multiple Linear Regression. Multiple linear regression is the simplest algorithm for ML modeling.[175] Because of their ease of interpretation and versatility for both large and small data sets, linear regression models have found wide application in materials science for descriptor selection and identifying feature importance in applications such as catalyst design and the prediction of reaction yield.[176−179]

This algorithm assumes the relationship between descriptors and the output values (properties) is linear. This can be written as

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{w} \tag{1}$$

where $\boldsymbol{X}$ is an $n \times m$ feature matrix (m descriptors and n data points), $\hat{y}$ is the vector of the predicted values of the property, respectively, and $\boldsymbol{w}$ is the weight vector.

To determine the best fitting model, we need to find $\boldsymbol{w}$ to minimize the loss function $\mathrm{J}(\mathbf{w})$:

$$\boldsymbol{w}^* = argmin_{\boldsymbol{w}}\mathrm{J}(\mathbf{w}) = argmin_{\boldsymbol{w}}(\mathbf{y} - \boldsymbol{X}\boldsymbol{w})^T(\mathbf{y} - \boldsymbol{X}\boldsymbol{w}) \tag{2}$$

where $argmin_{\boldsymbol{w}}$ denotes the value of $\boldsymbol{w}$ when the output of $\mathrm{J}(\mathbf{w})$ is minimized, $\mathbf{y}$ is the true values of the property, and $\boldsymbol{w}^*$ denotes the optimized $\boldsymbol{w}$.

Gradient descent is often used to obtain the fitted weights in ML algorithms during model training. Given a random starting point, this algorithm is iteratively moved along the direction of the steepest descent, obtained from the negative gradient of the function, until a local minimum is reached. Mathematically this is represented by

$$\tilde{\boldsymbol{w}}^{(i+1)} = \tilde{\boldsymbol{w}}^{(i)} - \alpha\nabla\mathrm{J}(\tilde{\boldsymbol{w}}^{(i)}) \tag{3}$$

where $\tilde{\boldsymbol{w}}^{(i)}$ is the weight vector calculated after the i^{th} iteration, $\nabla\mathrm{J}(\tilde{\boldsymbol{w}}^{(i)})$ is the gradient at the current spot, and α is the learning rate, which controls the size of each step. A high

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

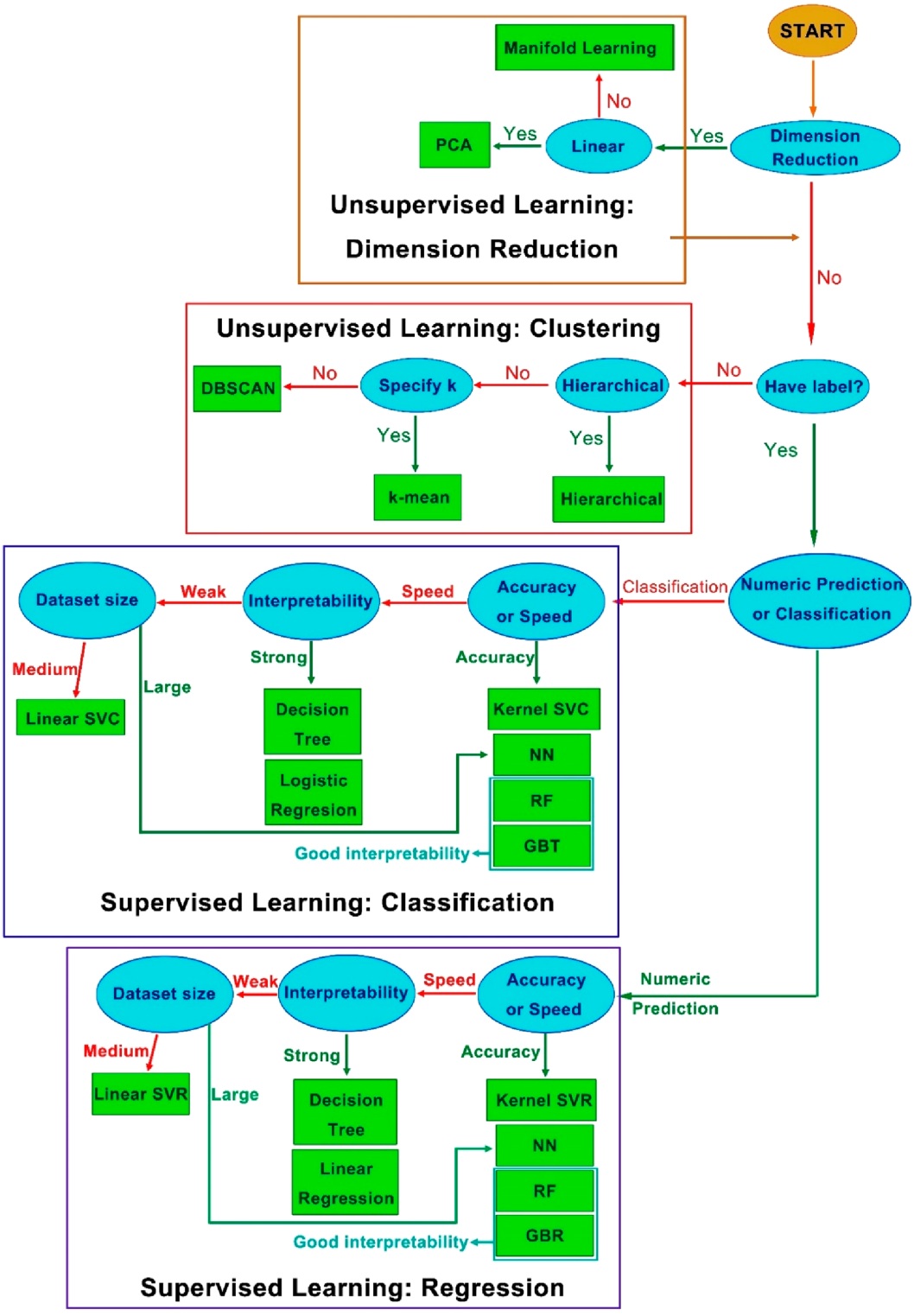


Figure 3. Overview of the popular ML algorithms in materials science and their selection (only exemplary techniques are noted in the figure). The selection of ML algorithms depends on the structure of the data set and type of target. Dimension reduction is applied if the number of features for each data entry is too large. Unsupervised learning is employed if the data are unlabeled, and supervised learning if it is labeled. Regression algorithms model continuous data, while classification algorithms model class data. Most algorithms can build either regression or classification models. Abbreviations in this figure: NN (neural networks), GBT (gradient boosting tree), GBT (gradient boosting regression), SVC (support vector classification), SVR (support vector regression), DBSCAN (density-based spatial clustering of applications with noise). The *k* in *k*-mean denotes the number of clusters preset by human operator.

learning rate may overshoot the minimum of the loss function as the gradient is different in each spot, while the calculation may become very time-consuming if the learning rate is too low. Generally, the learning rate can be relatively high at the start of the iteration when the weights are far from optimal, then decrease gradually during the iteration to obtain more

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

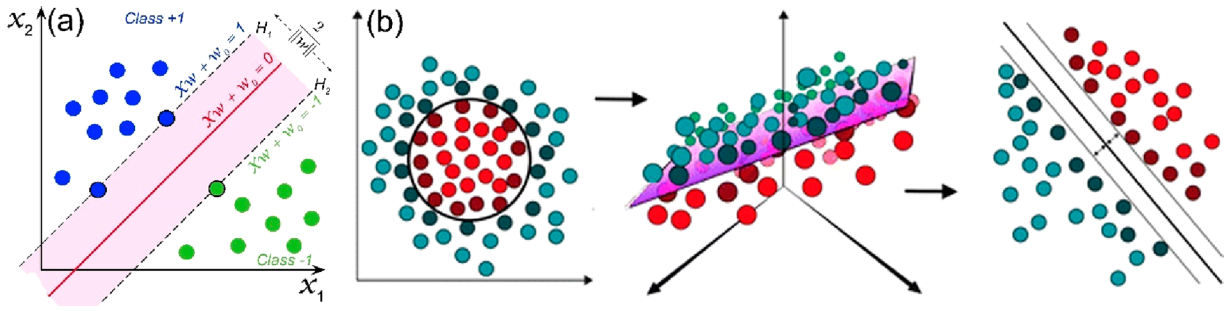


Figure 4. Representation of SVM algorithms. (a) Demonstration of a linearly separable data set. The blue points (class +1) and green points (class −1) are divided by a hyperplane (redline), and the points residing at H1 and H2 are support vectors. This figure was adapted from ref 186. (b) Demonstration of a linearly inseparable data set. Nevertheless, these data points are separable when they are transformed to a high-dimensional space. Reprinted with permission from ref 185. Copyright 2015, American Chemical Society.

precise estimates of the weights. After sufficient iterations, the output of the loss function will reach a minimum value, indicating that the regression model is optimized.

When the output error distributions do not follow a normal distribution (e.g., the values of the modeled property have a skewed distribution or are discrete), a generalized linear model may be used. Here, a link function is provided to connect the linear predictor ($\boldsymbol{Xw}$) to the mean value of the property whose error is in line with a distribution model. For example, the introduction of the sigmoid function to represent the mean value of the discrete property enables the linear predictors to give a discrete outcome (e.g., binary outcome):

$$\ln(\overline{\mathbf{y}}/1 - \overline{\mathbf{y}}) = \boldsymbol{Xw} \tag{4}$$

where $\overline{y}$ is the mean value vector of outputs. This modeling process is called logistic regression.

In multiple linear regression models, the least-squares method is unbiased only when the numbers of the data points, n, are far more than numbers of the weights, m. When $m \approx n$ or $m > n$, the regression model suffers from overfitting. Hence, regularization is imposed on the model to mitigate issues caused by multicollinearity (i.e., highly correlated variables) and to remove low relevance features. The main regularization methods are ridge regression, LASSO regression, and elastic net.

Ridge regression is a simple method to alleviate the issues of multicollinearity by L_2 regularization (penalty is equal to the sum of the square of the weights). It uses a ridge parameter, λ, to leverage bias and variance: the variance of the model will be large as λ approaches zero, while bias is more of an issue with a large λ. An optimized λ can be found by observing the ridge trace plot.[180] Moreover, although some of the weights can be shrunk to small values, ridge regression cannot eliminate weights.

In contrast to ridge regression, LASSO regression performs L_1 regularization (the penalty added in the loss function is related to the sum of the absolute value of the weights). Because the L_1 penalty is a linear term, some of the weights can be eliminated in the construction of the gradient of the loss function, thereby resulting in a sparse model with fewer weights. A larger penalty produces a sparser model with fewer weights and increased bias, while a small penalty produces large variance and low bias. Compared with ridge regression, LASSO regression models are more interpretable, and sparse feature selection is achieved. Ge et al. used LASSO to study electrocatalytic nitrogen reduction, and discovered a new descriptor able to link bond length, bond angle and atomic numbers with nitrogen reduction activity.[181] Nevertheless, LASSO regression is not suitable for group selection, as it tends to arbitrarily select only one weight from a highly correlated group of features and eliminate the rest. Accordingly, the model may be unstable and information inherent in the correlated terms may be discarded.

Elastic net was developed to combine the advantages of ridge regression and LASSO regression by integrating the L_1 penalty and L_2 penalty into the loss function. By tuning the LASSO parameter (λ_1) and ridge parameter (λ_2), this method can group-select weights with a high correlation. Elastic net regression therefore outperforms LASSO regression, especially in the cases where there are larger numbers of descriptors than data points or some of the descriptors are strongly correlated. Despite the good performance, the computational expense of elastic net is much higher than those of ridge and LASSO regressions because of the need for complex cross-validation, and the flexibility of choosing elastic parameters may induce overfitting.

We emphasize that despite the similarity, the distinction between linear regression as a statistical method and as a ML algorithm really depends on whether the purpose is inference or prediction.[182] Linear regression used for prediction is a type of ML algorithm.

In many cases, multiple linear regression is a fast and simple method. It can perform very well even on small data sets when there are linear relations between features and properties, and the performance of these models depends significantly on the linearity of the feature−property relationship. When the feature−property relationship is nonlinear, multiple linear regressions can also be used by transforming descriptors into a higher dimensional feature space where the feature−property relationship is linear, and kernel functions are used to simplify the calculations.

2.3.2. Support Vector Machines. As regression is an ill-posed problem in statistics,[183] which is sensitive to outliers, unstable, prone to overfitting, and linear regression models are unsuited to modeling nonlinear relationships, more robust and general algorithms are needed. The support vector machine (SVM) is a popular ML algorithm that maps the original data points onto a high-dimensional space in which the relationship is linear. SVM can be used to train a classifier (support vector classification, SVC) or a regressor (support vector regression, SVR). For classification models, an optimal hyperplane is found that separates the data points into two classes. Figure 4a shows a simple example, in which a linearly separable data set

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

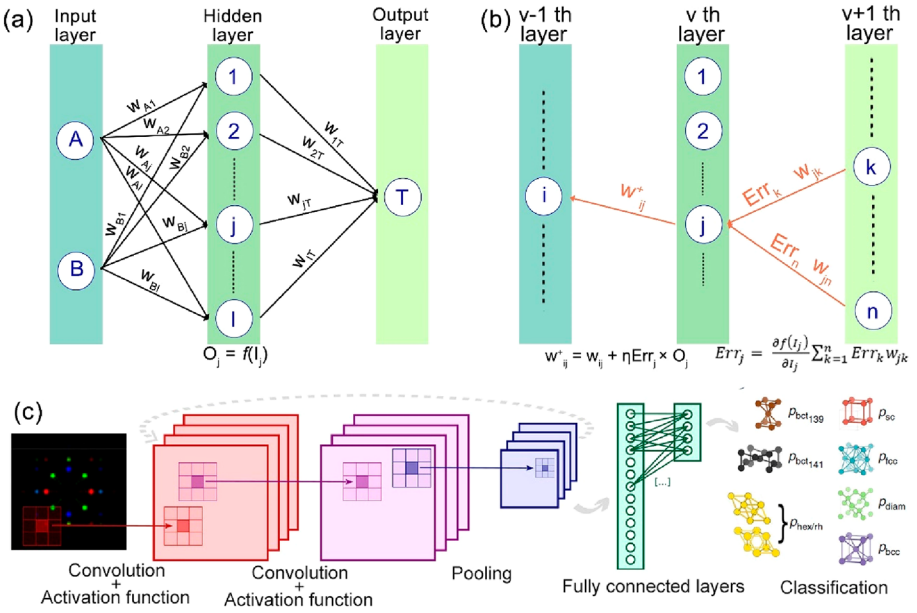


Figure 5. Representations of various neural networks. (a) Scheme of the feed forward of a neural network. (b) Scheme of the error back-propagation of a neural network. (c) Schematic representation of the convolutional neural network used for crystal classification. Reprinted with permission from ref 220. Copyright 2019, Springer Nature.

consisting of n data points described by two descriptors is divided into two classes by an optimal hyperplane positioned to maximize its distance from the closest data points. In many cases when the data points are not linearly separable, SVC can provide nonlinear solutions by transforming the data points to a high dimensional space where linear relations can be established (Figure 4b). Kernel functions are used to simplify the calculations in high dimensional space. To avoid the overfitting that can occur with SVM, a Bayesian implementation of the algorithm, the Relevance Vector Machine (RVM), automatically generates the optimum number of support vectors for a given data set and kernel function and provides better performance than SVM.[184,185]

The principles of SVR are distinct from SVC, albeit related. Instead of searching a hyperplane, where the distance to the closest data points is maximized, the distance between the SVR hyperplane and most data points is no more than the tolerance margin, ξ, with ξ being determined by the required accuracy of the model. Both SVC and SVR can be used for catalytic activity prediction and simplification of DFT calculations.[87,187−190] For example, Baghban et al. shortened the DFT calculation time for determining the bandgap of zeolitic imidazolate frameworks (ZIFs) using an SVM model, providing a simple approach for the design of ZIF-based electrocatalysts.[191]

The SVM, a nonlinear algorithm, provides a global solution to the feature−property relation, shows good generation ability, and is insensitive to outliers. Nevertheless, training of the SVM model is very slow on a large data set, and there is no general rule for the selection of kernel functions. Distinct from the linear algorithms, the transparency of the SVM model is low, and the SVM classifiers are only available for the binary classification, which restricts their applications.

2.3.3. Decision Trees. Decision tree (DT) models generate a tree-like model that classifies the output values by imposing sets of rules on the input features. It is composed of internal nodes, representing tests on a descriptor, branches which are the outcomes of the tests, and leaf nodes that hold class labels or continuous values. Tree models with discrete outputs are called classification trees, and those with continuous outputs regression trees. DT models can discover new knowledge de novo and handle multidimensional data. They are fast and simple to train, can generate high accuracy, and are easier to interpret than some ML methods. However, DT models are prone to overfitting and can be biased and unstable. Small changes in the training data may cause a dramatic change in the structure of the optimal tree. This issue is common to most regression methods as regression is inherently an ill-posed problem.[183] Regularization converts ill-posed problems into well-posed problems. For DTs, this is achieved by pruning the size of the tree. In some cases, it can be difficult for DT models to learn very complex relationships. Despite these disadvantages, DT is still considered one of the most versatile algorithms because of their excellent interpretability. Consequently, DTs have been used in many areas of materials science, such as optimization of catalysts, compound classification, spectra analysis, and toxic prediction.[179,192,193]

2.3.4. Neural Networks. The neural network (NN) is a bioinspired algorithm where numerous connected units send signals to other units, similar to the function of neurons in the brain.[194,195] The NN consists of a series of connected input/output units in which each connection can transmit a weighted signal to other units (Figure 5a).[196] During the learning phase, the NN adjusts the weights and bias on each connection to reduce the errors between the predicted values and the target values until the errors are acceptable or no improvement is obtained. NNs can have long training times and work best with large training data sets. Although NNs are considered difficult to interpret, they have high noise tolerance, can model any complex relationships given enough data, and can make predictions for both continuous and discrete data.[197] As a result, NNs are useful in providing solutions to many problems, not only in electro/photocatalytic science[152,198−200] but also in molecular dynamics simulations,[37,46,80,201−203] synthesis route design,[204−206] pattern recognition,[196,207−209]

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

sequence recognition,[210−212] medical diagnosis,[213−216] and machine translation.[217−219]

The most popular NN algorithm is the backpropagation NN[221] in which, during training, the error between the output value and the target value is propagated back through the network to adjust the weights and bias in the different layers (Figure 5b). It comprises one input layer, one output layer, and one or a few hidden layers, each of which contains one or more nodes or neurons. Figure 5a depicts a typical three-layer backpropagation NN, in which the inputs corresponding to the descriptors of a training data set are fed through to the hidden layer where they are processed by a transfer function and the outputs passed to the output layer to generate predictions for the input data point. The number of input nodes is equal to the descriptors, while one output unit is commonly used for binary classification or continuous value prediction. Multiple output units can also be used to represent different classes for a multiclass problem or multiple continuous properties for multiobjective models. There is no rule about the optimal number of hidden layers to be used, and therefore the design of a backpropagation NN requires trial and variation. However, Bayesian regularization of neural networks can automatically optimize the architecture and number of adjustable weights to yield robust and predictive models of a wide range of properties.[222,223]

Deep learning is a class of ML that extracts the knowledge via a multiple layer NN.[224] The deep learning neural networks (DNNs) have several hidden layers and many nodes per layer, with which features or patterns are progressively abstracted from inputs layer by layer. As DNNs do not rely on human decision (e.g., selections of polynomial degree or kernel function), they can learn relationships of arbitrary complexity given sufficient training data. Their main advantage over "shallow" (3-layer) neural networks is in the representation of data.[225] Instead of manual feature engineering according to domain knowledge, deep learning enables models to learn the underlying structure of the input data and extract useful features, by which the input format becomes more flexible (e.g., images, videos, sounds), and model performance can be improved.[208] Deep learning has become a very useful tool in materials exploration, catalytic reactivity prediction, and experimental condition optimization.[226−229] Kim et al. applied DNNs to model 31 713 known zeolites using atomic information and local methane adsorption energy as inputs. The model could identify novel zeolites with improved adsorption properties.[228] Moreover, this model could be extended to more complex porous materials, such as metal organic frameworks and covalent organic frameworks, provided these materials lay near the domain of applicability of the model. Ding et al. studied the electrocatalytic oxygen reduction activity of ZIF-based materials with DNN as well as another 8 ML models.[230] The summarized results showed pyrolysis time was a decisive factor in the electrocatalytic reactions, in which a moderate pyrolysis time could lead to the best electrocatalytic performance because of the balance of Zn−N evaporation, Fe−N formation, and graphitic N conversion. Yang et al. employed DNN to optimize the experimental parameters for photocatalytic treatment of oily wastewater.[231] Because of the high accuracy of the model prediction, optimal operating parameters could be acquired promptly from past data, thereby saving time and cost on trial-and-error experiments.

Convolutional neural networks (CNNs) are a type of DNN mostly used in image analysis.[232] A CNN consists of several convolutional and pooling layers, with a fully connected layer. Figure 5c illustrates the principle of the CNN. When an image is the input, the convolutional layer extracts features from the image by scanning with one or more filters and generates a feature map by convolution. An activation function is applied to the feature map to introduce nonlinearity and to simulate higher level features. After a few convolution and activation cycles, the pooling layer is used to reduce the dimensions of the data before further convolution. Finally, data is transmitted into the fully connected layer, a multilayer NN, and values or class labels are predicted. For materials science, CNN models are often trained on images of materials features.[220,233,234] Back et al. successfully predicted the CO and H binding energies on the pure metals, alloys and intermetallic surfaces using a CNN, thus expediting high throughput catalyst discovery.[235] Wang et al. used a CNN to map the crystal structures to the Heyd−Scuseria−Ernzerhof (HSE06) bandgaps, and identified 33 materials from 1503 candidates with bandgaps less than 0.9 eV, which were considered as promising electrodes and electrocatalytic materials.[236] In addition, a CNN is capable of determining physical insights of electro/photocatalysis. Wang et al. developed a theory-infused neural network consisting of a CNN module and a d-band theory module to predict the adsorption energy of OH* on {111}-terminated intermetallics.[229] Using graphic features to represent the information on atom, bonding, and neighbor environment on a metal surface as input, a CNN model was constructed to determine parameters for the module built on d-band theory of chemisorption, and the adsorption energy was then calculated by the d-band theory module. The combination of ML and chemisorption theory illustrated the nature of chemical interactions between adsorbates and metal surfaces, which is a critical step in electro/photocatalysis. This strategy could also be extended to other adsorbates (e.g., O*, N*) on metal compounds with strongly correlated d electrons. More examples of the application of deep learning in electro/photocatalysis are given in Section 3.

Despite the broad applicability of NNs, the significant training times and the need for large amounts of training data can limit their applicability in specific materials design, discovery, and development areas. Often, training set sizes are small so model bias and overfitting occur. Therefore, NNs may be combined with other methods such as ensemble learning, active learning, or evolutionary methods. The high complexity of the NN models, which limits their interpretability, is another potential shortcoming of NN models for material exploration and optimization. Informative descriptors should be selected to simplify the models and improve the interpretability, while new techniques, such as class activation maps,[237] can be used to elucidate the black box of NN models.[238−243]

2.3.5. Ensemble Learning. Ensemble learning combines multiple learning algorithms to create a strong overall learner with better performance than the component algorithms (Figure 6).[244,245] There are two main types of ensemble learning, bagging and boosting. Given a training set D of size n, bagging generates m new subsets D_i (i = 1, 2, ..., m) of n' size ($n' < n$) by sampling from D with replacement. Model M_i is learned from the subset D_i, and thus m models will be created. In classification, the prediction of each M_i counts as one vote and the overall model assigns the class by majority vote. In regression, the average value of each prediction will be provided by the ensemble model. Bagged models often have

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

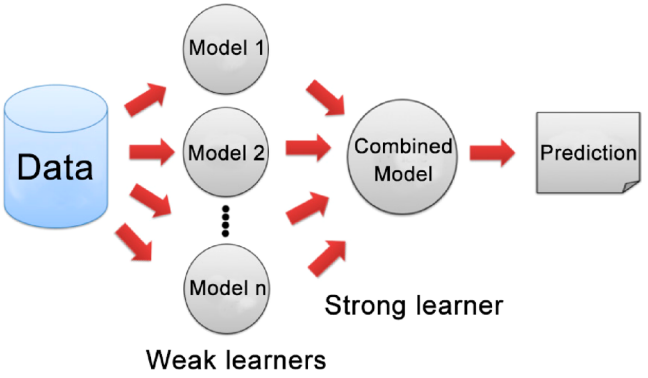


Figure 6. Schematic representation of ensemble learning. Multiple ML models (models 1 to *n*, weak learners) are combined as an ensemble model to obtain better predictions (strong learner).

significantly higher accuracy than a single model trained on the data set, and bagged models can minimize overfitting and noise effects. One of the most frequently used bagging ensemble models is RF, which combines results from a series of decision trees. RF has significant advantages, such as robustness to outliers and errors, high efficiency, good performance, and provides estimates of the variable importance. As a result, the RF model is robust for predictions of catalysts properties such as electrocatalytic activities, gas adsorption free energy, and photocatalytic properties, and uncovering the impactful factors for these properties.[87,152,246,247] For example, Ying et al. built an RF model to study the bifunctional oxygen evolution/reduction reaction activity for single-atom catalysts on C_2N, and identified that the outer electron number and the oxide formation enthalpy of single-atom catalysts were two of the most important factors determining their activity.[248] This work indicated that RF could not only save calculation time and cost but also provide deeper insight into the electrocatalytic activity of single-atom catalysts. Despite the good transparency, the complexity of RF can be high especially when the number of decision trees is large, and thus the training time will be significantly extended.

While bagging treats all training data points equally and simply combines the outputs of each individual model (weak learner), boosting assigns weights to each data point and each weak learner. After a model M_i is trained, the weights of the data points exhibiting larger errors will be increased and those having smaller errors decreased. This emphasizes poorly predicted data points with weights of each weak learner being estimated by the errors as well. AdaBoost is one of the popular boosting algorithms.[249] Compared to bagging methods, AdaBoost may overfit the data because it focuses more on poorly predicted data points,[250] but it may achieve a higher degree of precision than bagging methods. Instead of focusing on the data points having large prediction errors, gradient boosting uses the residual error between the predicted value from the current model and the true value.[251] Like the gradient descent algorithm discussed in Section 2.3.1, gradient boosting minimizes the loss function by gradually imposing a new model into the current model that reduces the loss function to the largest extent. This process is iterated until an optimal loss value is found. Because of the high accuracy and versatility, gradient boosting algorithms, such as gradient boosting tree (GBT) and gradient boosting regression (GBR), are widely used in materials screening, discovery, and property prediction.[140,252−254]

2.3.6. Clustering. Clustering is an unsupervised learning method that partitions a set of data into subsets in which all members are similar to each other. Unlike classification, the clustered data are unlabeled, and their similarity is determined by mathematical descriptors. Clustering is useful for discovering previously unknown patterns in data.

Several clustering algorithms are commonly used in materials science. The best known is *k*-means clustering that partitions data into *k* clusters (*k* is preset by a human operator) based on the Euclidean distance between the data points in the feature space (Figure 7a). It is computationally efficient and can be applied to data in continuous *n*-dimensional space, but it is sensitive to noise and outliers. Hierarchical clustering presents the data points as a hierarchy (Figure 7b). Although there is no limit to cluster numbers in this algorithm, which hierarchy is adopted as the ultimate clustering metric still relies on human input. The results are also sensitive to noise. To overcome noise sensitivity and discover data clusters of arbitrary shape, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was developed. Here, rather than using distances between data points, DBSCAN uses the similarity of the density around the data points, and thus it can easily distinguish noise from information. It is suitable for clustering of both convex and nonconvex data sets (Figure 7c). However, DBSCAN is more complex than *k*-means and hierarchical clustering, requiring longer computation time to cluster large data sets. Clustering algorithms have been applied to spectral analysis and catalyst discovery in catalysis research.[255−257]

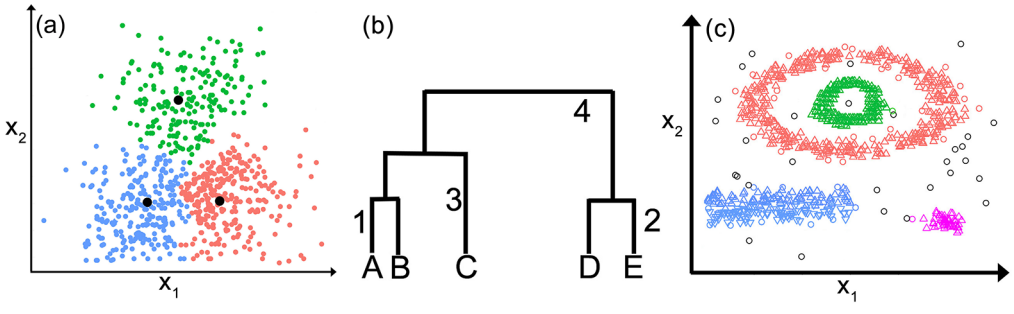


Figure 7. Schematic representation of some cluster algorithms. (a) *k*-means clustering; the data points are partitioned into three clusters (presented as green, blue, and pink dots). The black dots denote the means of these clusters. (b) Hierarchical clustering, in which five data points are partitioned hierarchically. (c) DBSCAN, the data points that are successfully clustered are represented by triangles with various color, the noise data are represented by black circle, and the uncertain data are represented by circles with the color of the cluster that they may be grouped in.

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

Table 3. Machine Learning Frameworks and Tools

Name	Description	URL
Caret	A set of functions that attempt to streamline the process for creating ML models in R	https://topepo.github.io/caret/
COMBO	Highly scalable common Bayesian optimization library, written in Python	https://github.com/tsudalab/combo
DeepChem	Open-source library for the use of deep learning in chemistry, written in Python	https://deepchem.io/
Deeplearning4j	Generates deep neural nets from various shallow nets, written in Java	https://deeplearning4j.konduit.ai/
H2O.ai	ML platform written in Java, and compatible with Python and R	https://www.h2o.ai/
Keras	A high-level neural network API (Application programming interface), written in Python and capable of running on top of TensorFlow, CNTK, or Theano	https://keras.io/
Mlpack	A fast, flexible machine learning library, written in C++	https://mlpack.org/
NSGA-II	A fast sorting and elite multiobjective genetic algorithm	http://www.mathworks.com/matlabcentral/fileexchange/10429-nsga-ii-amulti-objective-optimization-algorithm
Pytorch	An open-source machine learning framework, having Python and C++ interfaces, developed by Facebook	https://pytorch.org/
Scikit-learn	A free software machine learning library for Python, featuring various classification, regression, and clustering algorithms	https://scikit-learn.org/stable/
TensorFlow	A free and open-source software library for ML, developed by Google	https://www.tensorflow.org/
TensorMol	A package of neural network models for chemistry	https://github.com/jparkhill/TensorMol
Weka	Free software containing a collection of visualization tools and algorithms for data analysis and predictive modeling, written in Java	https://www.cs.waikato.ac.nz/ml/weka/

2.3.7. Other Algorithms. *2.3.7.1. k-Nearest Neighbor (KNN).* KNN is widely used in pattern recognition because of its simplicity. Given a training set with n descriptors and an unknown sample, KNN identifies k nearest neighbors of the unknown sample in the n-dimensional space. The parameter k is defined by the users, and the distance between each pair of the unknown and training set data points is calculated by the Euclidean distance in n-dimensional space. This algorithm is fast because it only stores the training set rather than modeling it, so new data can be seamlessly added to the model. Nevertheless, KNN has low efficiency when the training set is very large, or the data points are in very high-dimensional spaces. It is also sensitive to noise, missing values, and outliers, so deep preprocessing of the data is essential. In addition, all the descriptors must be normalized or standardized to balance the contributions of each. Because of its simplicity and interpretability, KNN is used in many areas of materials science, such as the prediction of molecular atomization energies and catalyst screening.[92,258]

2.3.7.2. Kernel Ridge Regression (KRR). KRR is an extension of ridge regression.[259] It introduces a kernel function in ridge regression to tackle the nonlinearity. KRR is similar to SVR, except that KRR uses the squared error penalty, whereas the loss function of SVR is the ε-insensitive loss. Fitting a KRR model is faster than fitting an SVR model on a medium-sized data set (<1000 samples), but it is slower than SVR in fitting and predicting when the data set is large.[260] Currently, KRR models are used in materials ML modeling, e.g., analyzing crystal structures, optimizing catalysts, and learning the quantum chemical properties of molecules.[261−263]

2.3.7.3. Genetic Algorithms. Genetic algorithms are inspired by natural evolution.[264] Materials are represented as genomes (encoded as a bits string) that contain information on composition, structure, synthesis conditions, etc. Given an initial pool of materials chosen randomly or by prior experience that have their fitness assessed, the fittest members of the pool are mutated using genetic operators such as crossover and mutation to form the next generation (progeny) of materials. Fitness functions are application dependent, e.g., catalytic activity, bioactivity, or quantum efficiency. Experiments are conducted to identify the fitness of materials pools in each iteration of the genetic optimization algorithm. Less fit materials are discarded. The cycle continues until properties are optimized or no further improvement occurs. This process can be slow and expensive. ML models can be trained on the data and used as a surrogate fitness function to select materials for the next iteration. Such in silico fitness functions can accelerate the optimization process, analogous to the Baldwin Effect in evolutionary biology (describing how learning in organisms can accelerate evolution).[265,266] The use of ML models as fitness functions to replace the need for some experiments in the evolutionary design of materials can reduce the number of experiments required and allow efforts to focus on the material genomes with the highest performance. Genetic algorithms have been applied to a wide range of materials for discovery, property prediction, and optimization.[231,267−274] More details of the principle of this algorithm and the corresponding applications can be found in the recent comprehensive review by Le et al.[264]

2.3.7.4. Active Learning. Active learning (sometimes called adaptive experimental design) is an iterative type of supervised learning well suited to large data sets whose output values (target property values) are scarce or expensive to obtain.[275] A model (M_0) is first trained on the subset with labeled data. A small quantity of unlabeled data with high predicted uncertainty or high diversity is provided by M_0, and then experiments or calculations are used subsequently to find the output values (discrete values in classification, or continuous values in regression) for these data. Subsequently, a new model (M_1) is trained by using all the labeled data. This process can be iterated until all the data are labeled or the prediction errors are acceptably low. The aim of active learning is to achieve high accuracy by using as few labeled data as possible. For example, a Bayesian NN model built on a small training set labeled by accurate DFT calculations accurately predicted the bandgap of ∼2.2 million 2D heterostructures.[276] Active learning is robust in electro/photocatalyst discovery and optimization, and electro/photocatalytic mechanism research with limited data.[129,148,277,278]

ML models can be constructed by a range of accessible packages and frameworks. Table 3 summarizes commonly used ML platforms and packages written in various programming languages. These are accessible for interested readers to apply to their projects.

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

2.4. Model Validation

Models must be validated by determining how accurately they predict the properties of the training set and, more importantly, unknown data. The coefficient of determination (r^2), root-mean-square error (RMSE), and mean average error (MAE), are common metrics used to assess model quality. For a data set of size n, with measured properties $\langle y_1, y_2, ..., y_n\rangle$, and property average $\overline{y}$, and predicted values $\langle \hat{y}_1, \hat{y}_2, ..., \hat{y}_n\rangle$, these and other metrics are calculated by

$$r^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \quad (5)$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \quad (6)$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} \quad (7)$$

Good models have a r^2 value close to 1 and RMSE and MAE values close to 0. The r^2 value depends on the size of the training set and number of parameters in the model, while the RMSE or MSE values do not. MAE is less affected by large outliers than RMSE. Therefore, the use of RMSE or MAE to assess model predictivity is preferred.[279,280]

For classification models, precision, recall, and accuracy metrics are widely used:

$$\mathrm{precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (8)$$

$$\mathrm{recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} = \frac{\mathrm{TP}}{\mathrm{P}} \quad (9)$$

$$\mathrm{accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \quad (10)$$

where TP refers to true positive, which means the positive cases are correctly labeled by the classifier; TN refers to true negative, the negative cases which are correctly labeled; FP to the false positive, the positive cases which are incorrectly labeled; FN to the false negative, the negative cases which are incorrectly labeled; and P is the number of positive cases. Precision and recall can be combined as a single metric, the F1 score:

$$\mathrm{F1} = \frac{2 \times \mathrm{precison} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \quad (11)$$

The F1-score and related G-mean (geometric mean of recall and precision) are common in the literature. They are more useful in cases where the classes are unbalanced in the data set (many more data points of one class than the other) because accuracy is misleading. For a good classifier, the precision, recall, accuracy, F1 score, and G-mean should all be close to 1.

The causes of poor model metrics include use of low relevance descriptors; underfitting because of too simple a model (e.g., linear vs nonlinear model); overfitting because of too many descriptors; or data sets in which some features are not well represented because of their small size or high diversity and the model cannot learn the features of the data properly. Underfitting can be avoided by comparing both linear and nonlinear ML methods, descriptors should be studied for their relevance (e.g., using correlation), and those not well represented in the data set excluded. If possible, the size of the data set should be expanded, in some cases by imputation where data are missing.

When models are too complex or have too many descriptors, they may suffer from overfitting. Overfitted models make very good predictions of the training data but poor predictions for new data. Overfitting can be detected by use of training and test sets. The original data set is partitioned into two sets, a training set used to construct the model, and a test set for model validation. Typically, 60−90% of data are allocated to the training set and the remainder to the test set.[281] Because the training set and the test set are completely independent, this method provides a least optimistic estimate of model predictive power than cross-validation of bootstrapping methods.[282] Models are considered robust if the training and test set metrics are good and similar. This method is simple and practical, but it requires a sufficiently large data set for training and test set portioning, otherwise the results may be problematic.

Cross-validation is a good alternative to evaluate the model accuracy for small data sets.[258] Here, the original data set is portioned into k subsets, with training and testing performed k times. In each iteration, one set forms a test set while the remaining sets are used for training. This is called k-fold cross-validation. Unlike the train-and-test method, each data point will be used $k - 1$ times for training and once for testing in the cross validation. The "leave-one-out" (LOO) method is a special case of cross-validation where one data point is left out in turn for testing, while the other data are used for training. LOO is one of the commonly used cross-validation methods.[146,166,283,284] The estimate given by LOO is repeatable, but LOO is time-consuming and computationally expensive. To reduce this computation overhead, a 10-fold cross-validation can be used. Note that cross-validation gives a more optimistic estimate of the prediction ability of the model than the train-and-test method because the training and test set are not independent.

Overfitting can be minimized by increasing the size of the data set, reducing the number of descriptors a priori, reducing the model complexity with the same number of descriptors, or by adding regularization terms to remove the descriptors of low relevance to the property being modeled.

The primary steps in constructing ML models to solve the electro/photocatalysis problems have been summarized. We stress that data quality plays a key role in ML model construction. Therefore, data reliability must be carefully evaluated, especially when the data are from different sources. This requires close collaboration among data scientists, catalyst scientists, and materials scientists to build a standard format for electro/photocatalysis data recording. Featurization ensures the number of descriptors is just sufficient for describing the input data, as excessive features may result in model overfitting. Despite a variety of descriptors being developed, the measurement of some descriptors (such as d-band center and intermediate adsorption free energy), is very difficult in catalytic reactions, thus simple and accessible descriptors with interpretable feature−property relationships are preferred. ML model selection depends on the problems to be solved and data structure. In other words, the quality, quantity and diversity of training data, as well as the relevance of descriptors determine which ML model to be used. Model predictivity and transparency should be balanced, and model predictivity must be evaluated before used. Models with good interpretability are

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

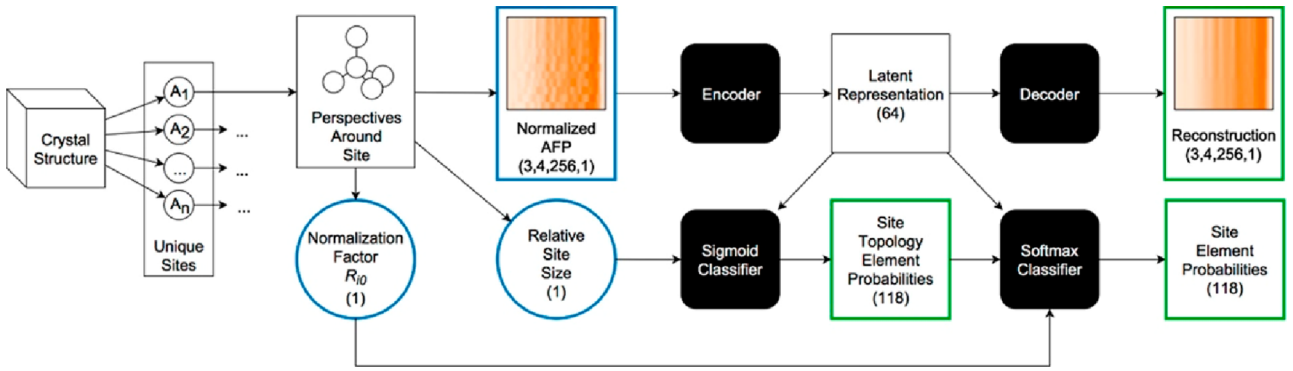


Figure 8. Schematic of the DNN architecture to predict elements from crystallographic sites. The numbers correspond to the dimensionality of each layer. Reprinted with permission from ref 126. Copyright 2018, American Chemical Society.

recommended in the investigation of electro/photocatalysis science, as they can accelerate catalyst discovery, simulate catalytic mechanisms over larger time scale, enhance characterization, and illustrate underlying structure−property relationships.

3. APPLICATION OF MACHINE LEARNING TO ELECTROCATALYSTS AND PHOTOCATALYSTS

Catalysis is a complicated, multidimensional, dynamic process. Consequently, the development of catalytic materials relies mostly on empirical, trial-and-error methods, which require significant time and effort to find the best solutions for multicomponent catalysts under a variety of reaction conditions. In recent years, ML techniques have revolutionized the development and discovery of electrocatalytic and photocatalytic materials.[35,285−288] In this section, the discovery or development of electrocatalysts and photocatalysts using ML techniques are reviewed, and examples of applications in intermetallics, oxides, single-atom catalysts, carbon-based materials, and photocatalytic polymers are provided.

3.1. Electrocatalysis

Electrochemical reactions involve two half reactions occurring at the anode (e.g., OER) and cathode (e.g., HER, the oxygen reduction reaction (ORR), carbon dioxide reduction, nitrogen reduction), respectively, powered by electrons.[5,6] Active catalysts are required to improve the efficiency and minimize the overpotential in these reactions. Among these reactions, the HER and ORR are widely applied in the production of clean energy. The HER is a typical example of a two-electron transfer reaction, and the hydrogen adsorption free energy ΔG_H is key for determining activity. The ORR involves a four-electron transfer to reduce oxygen to water, or a two-electron and two-proton pathway to form hydrogen peroxide. Active electrocatalysts for these reactions possess a ΔG_H close to zero, while having moderate binding energies for the reaction intermediates (H*, OH*, O*, etc.). Platinum is the best performing electrocatalyst for the HER and ORR because of its thermoneutral ΔG_H. However, Ooka et al. reported that thermoneutrality did not yield maximum electrocatalytic activity.[289] They fitted a microkinetic rate equation to the experimental results via a genetic algorithm and found that the binding energy of H on a polycrystalline Pt surface of 0.094 eV, gave maximum catalytic activity. They proposed that a positive ΔG_H was preferred in the HER because of the electrical driving force in this reaction. Despite the high activity, the scarcity and high cost of Pt limit its commercial use. More catalysts, including Pt-free intermetallics, oxides, single-atom catalysts, chalcogenides, and MXenes, have been developed to replace Pt.

3.1.1. Intermetallics. Bimetallic or intermetallic materials (combinations of two or more types of (metallic) atoms) are increasingly used in catalysis, gas adsorption, electronics, and other engineering applications. They have diverse properties that stem from their very wide range of compositions and structures, and the synergistic effects of mixed metals.[290−294] It is important to understand the composition−structure relationships in this class of materials in order to improve the design of intermetallic materials. However, it is not yet feasible to design bespoke intermetallics using a combination of evolutionary algorithms and quantum-mechanical methods as the computation cost and time is too high, given the complexity of these systems.[295,296] Here, ML methods play an important role in predicting intermetallic structures and properties.[297] For example, Ryan et al. developed a DNN model based on known crystal structures to predict the possible compositions and structures of unknown intermetallics (Figure 8).[126] Normalized atomic fingerprints were used as descriptors, providing the model with information on structural topologies instead of crystal structures. Because training of the DNN model is best done with positive and negative examples, the model was trained to predict the probability of each element being located at a specific site in structural templates. They trained the model on 51 723 known crystal structure templates with binary and ternary compositions then applied it to the Mn−Ge system. Four compositions with high probability structures were identified, plus a further 36 in the Li−Mn−Ge ternary system. This study demonstrated how models can effectively explore bimetallic and intermetallic materials and elucidate composition−structure relationships without prior knowledge (other than the training structures) or phase diagrams of the system.

Atomistic design of intermetallic materials plays an essential role in improving the performance of metallic electrocatalysts. In theory, optimal catalysts should possess sufficient active sites that strongly bind reactants and break the relevant bonds to form intermediates or products that are weakly bound and therefore removed readily from the surface of catalysts.[298] Optimization of the geometry and composition of the active sites is key to catalyst design.[299] In electrochemical reactions, optimization of intermetallic electrocatalysts by experiment or high throughput screening using quantum chemical calculations is difficult because of large composition and structure spaces. To accelerate these calculations and allow fast

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

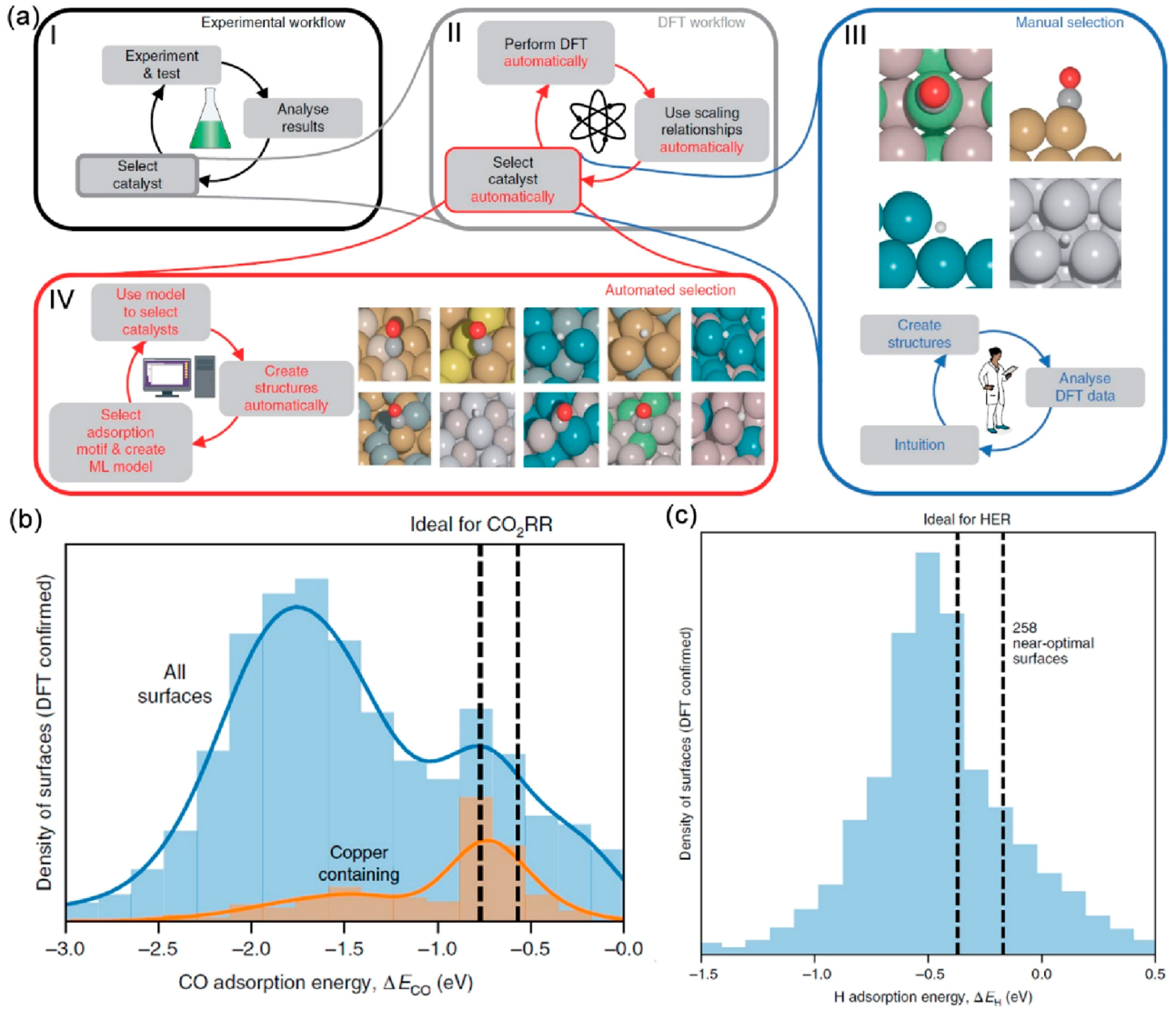


Figure 9. (a) Schematic of workflow for automating catalyst discovery. The experimental workflow for searching catalysts (I) is accelerated by a DFT workflow for screening the catalysts ab initio (II). (III) Scientific intuition is required to select candidates for DFT screenings in conventional workflows. (IV) Workflow uses ML to select candidates systematically and automatically. (b) Normalized distribution of the low-coverage, DFT-calculated CO adsorption energies of all the DFT-analyzed surfaces. The orange bars show the subdistribution for copper. Dashed lines indicate the 0.1 eV range around the optimal ΔE_{CO} value of −0.67 eV. (c) Normalized distribution of low coverage ΔE_H values calculated by DFT workflow. Dashed lines indicate the 0.1 eV range around the optimal ΔE_H value of −0.27 eV. Reprinted with permission from ref 129. Copyright 2018, Springer Nature.

prediction of the activities of intermetallics, Ma et al. used a DNN to predict the CO adsorption energy, the primary metric for catalyst activity in the electrochemical reduction of CO_2.[88] They used the {100}-terminated Cu intermetallic system. To describe the features of members of this system, they used the d-band parameters of the metal sites as primary descriptors, and physical constants for the host metal (spatial extent of metal d-orbitals, square of the adsorbate−metal interatomic d coupling matrix element, work function, atomic radius, ionization potential, electron affinity, and Pauling electronegativity) as secondary descriptors. The DNN model was trained to predict the CO adsorption energy. The RMSE of the model predictions compared to DFT calculations was 0.13 eV, indicating that the ML model was accurate enough to identify active {100}-terminated Cu intermetallics in CO_2 electroreduction. This model was then extended to {111}-terminated bimetallics to predict the adsorption energy of CO and OH, closely related to methanol electro-oxidation, the major oxidation reaction in a methanol fuel cell.[300] This model was trained on the alloy data set with eight types of surface, and was subsequently used to predict the CO and OH adsorption energy on six new prototypes of bimetallic nanoalloys. Again, this model successfully identified most of the known bimetallic catalysts for methanol electro-oxidation and provided some insight into electronic structure−activity relationships. The d-band filling had a larger influence on the adsorption of CO than OH, probably because of the unoccupied $2\pi^*$ molecular orbitals in the CO adsorbates that were immediately above the Fermi level for hybridization. The sp-band properties affected the OH adsorbates more than the CO adsorbates because of Pauli repulsion between the interacting nonorthogonal orbitals. Accordingly, it was suggested that some 3d metals could be doped into the bimetallics to reduce the overpotentials. Toyao et al. employed an extreme randomized trees model to study

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

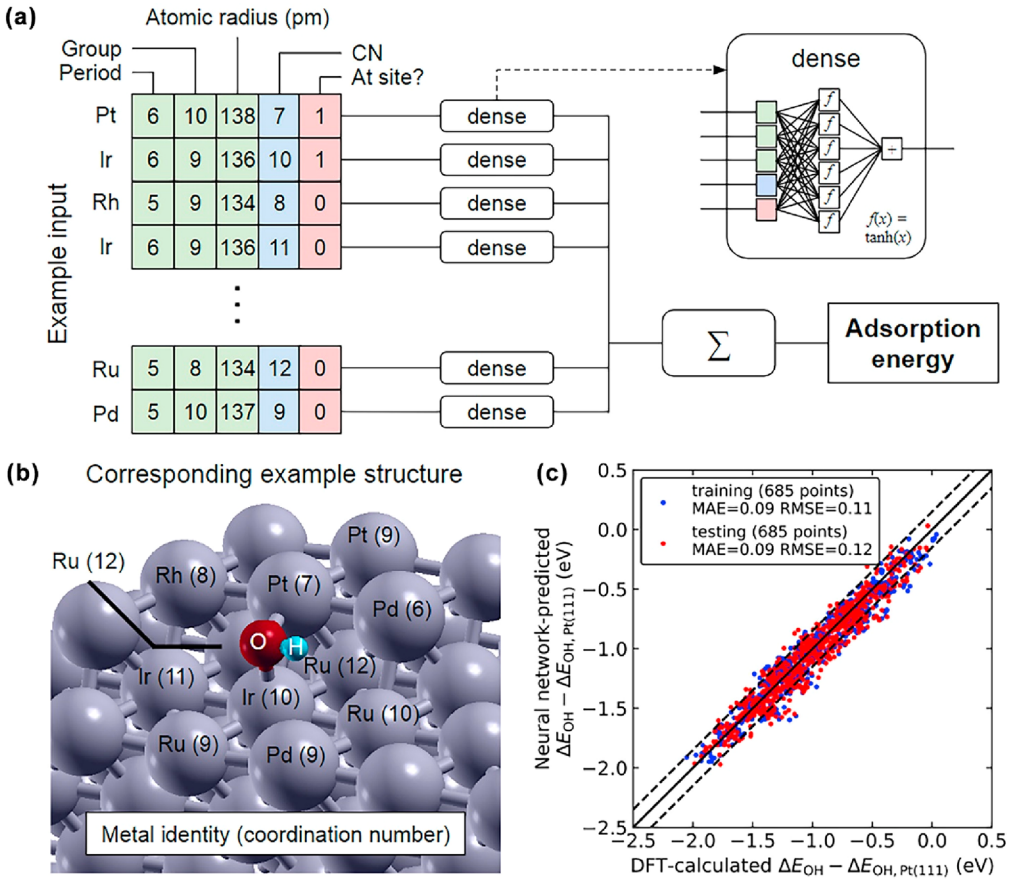


Figure 10. (a) Schematic of the NN model for the adsorption energy of OH* on IrPdPtRhRu. (b) Structure corresponding to Figure 10a, where the active site and its nearest neighbors are labeled. (c) DFT calculated adsorption energies versus NN model predicted adsorption energies. Reprinted with permission from ref 310. Copyright 2020, Cell Press.

the catalytic activity of {111}-terminated Cu bimetallics in the CH_4 formation from CO_2 and H_2.[301] Twelve simple physical descriptors from the periodic table and handbooks of chemistry and physics were used to train a model that predicted the adsorption energies of CH_3, CH_2, CH, C, and H on Cu-based alloys. Interestingly, the adsorption energies of the CH_4-related species were mainly influenced by the periodic table group, surface energies, and melting points of the doped metals. The level of prediction accuracy achieved suggested that this model would be useful for fast screening of large libraries of bimetallic materials for CH_4 conversion.

Haider et al. examined the effects of A_3B bimetallics and copper-based single-atom alloys on ethanol decomposition and the nonoxidative dehydrogenation reaction using a GBR model.[127] The model was trained on 151 A_3B bimetallic alloys using 27 descriptors that included physical properties of both the metals in the alloy, and predicted the oxygen and carbon binding energies. It was observed that in the AA terminated A_3B alloys, the surface energy of dopant B was the most important descriptor, followed by ionization energy, electronegativity, density, and heat of fusion of the dopant. In AB terminated intermetallics, the surface energies of both the dopant B and the matrix A had very similar importance. Although further studies were necessary to illustrate the dynamics of adsorption on AA terminated and AB terminated A_3B alloys, it was rationalized that surface energy, ionization energy, and electronegativity were most important for binding to transition metal surfaces. Surface energy is related to the activity at the surface, and ionization energy and electronegativity are indicative of electron transfer between the surface metal atoms and the adsorbates. For copper-based single-atom alloys, 12 descriptors were used to capture relevant properties. The periodic table single-atom group was also important in predicting the oxygen binding energy.

A challenge with using ML models for intermetallics is that their structures present many facets and termination sites, giving many adsorption modes. Model simplifications in which only a few facets (e.g., {111} and {100}) and terminations are considered bias conclusions on the activity of the intermetallic catalysts. To address this issue, ML was used to extend local information from DFT to fully span the intermetallic surface. Ulissi et al. reported a ML approach that accounted for all active sites on the Ni−Ga bimetallics used for CO_2 reduction.[128] Instead of restricting the reaction to a few exposed facets, they used a geometric fingerprint to represent the local region around each atom. A NN model was trained on thousands of surface atoms labeled by DFT-calculated CO adsorption energies, allowing the CO adsorption energy for each atom to be predicted according to its local environment. The model discovered a new type of active site: active Ni atoms surrounded by surface Ga atoms. This work was extended by Tran et al. using an active learning algorithm to screen a space of 1499 intermetallic combinations of 31 different elements for CO_2 reduction and H_2 evolution.[129] All surfaces with adsorption sites were encompassed by this space (Figure 9). Intermetallic adsorption sites were described by a

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

vector of four numbers for each element: atomic number; Pauling electronegativity; number of atoms coordinated with the adsorbate; and the average adsorption energies on pure metal surfaces. The model was trained on 200 surfaces with DFT-calculated adsorption energies to predict adsorption energies of the CO and H adsorbates. The model then predicted properties of the next 200 surfaces. Those with predicted adsorption energies close to the theoretically optimal value were selected and validated by DFT. The model was then updated by these additional 200 surfaces whose adsorption energy values would be replaced by the DFT results if they existed and subsequently used to predict a further 200 surfaces. This process was repeated until the entire search space was modeled. This approach yielded 130 candidate surfaces across 54 intermetallics for CO_2 reduction and 258 candidate surfaces across 102 intermetallics for H_2 evolution (Figure 9). Besides a significant reduction in DFT computation time, this approach also allowed analysis of intermetallic surfaces and discovery of intermetallic catalysts without expert intuition and is transferrable to other systems.

A synergistic combination of DFT and ML was used to discover efficient catalysts for various electrocatalytic systems. Zhong et al. combined DFT and RF to model CO adsorption on the surface of Cu-containing electrocatalysts, discovering a novel Cu−Al electrocatalyst that can selectively reduce CO_2 to ethylene with the Faradaic efficiency as high as 80%.[302] Their computations showed that the multiple sites and surface orientation of the Cu−Al catalysts made significant contributions to the efficiency and selectivity of CO_2 reduction. Kim et al. developed a slab graph CNN model to analyze 3040 intermetallic surfaces and predict the binding energies of five key intermediates in nitrogen reduction (H, N_2, N_2H, NH, NH_2).[303] Four novel catalysts, $V_3Ir(111)$, $Tc_3Hf(111)$, $V_3Ni(111)$, and $Tc_3Ta(111)$, were identified as efficient catalysts for nitrogen reduction. These studies exemplify the contributions of computation and ML in guiding the exploration of intermetallic electrocatalysts.

In the intermetallics family, high-entropy alloys (HEAs) are also promising catalysts.[304−306] HEAs consist of multiple metals (five or more) in near-equiatomic proportions, in which the increase of configurational entropy overcomes the compound formation enthalpy, thereby stabilizing the alloys. Because of their tunability, HEAs have emerged as efficient electrocatalysts for the HER, OER, and ORR, but the large configurational and structural degrees of freedom in HEAs make the structure−property relations too complex to be disentangled by experiments or DFT calculations. Thus, ML is best suited to study and optimize HEAs.[307] Pedersen et al. demonstrated a fast, accurate prediction of CO and H adsorption energies on various sites of the {111} facets of disordered CuCoGaNiZn and AgAuCuPtPd by combining DFT and ML. Hence the efficiency of the CO_2 reduction reaction could be improved and the formation of the side product, H_2, was largely suppressed by composition engineering of HEAs.[308] To study the effects of the adsorption energy of intermediates on the ORR, Batchelor et al. used DFT to calculate the adsorption energy of OH* and O* on 871 and 998 different 2 × 2 unit cells of IrPdPtRhRu HEAs. They trained a linear ML model on these data that could predict the adsorption energy of these two radicals on the (111) surface of IrPdPtRhRu.[309] To provide deeper insight into the structure−property relationships, Lu et al. developed a NN model to leverage DFT calculations, and studied the adsorption energy of OH* on IrPdPtRhRu with random element distributions and 12 unique coordination environments (Figure 10).[310] Using only elemental descriptors and coordination numbers as input, the NN model showed good agreement with the experimental results, and revealed that the OH* adsorption energy was linearly related to the coordination number of the nearest neighbors. These studies indicated that ML techniques are powerful tools that can discover catalysts in vast chemical spaces. ML models not only provide accurate predictions with less computation effort but capture some underlying physical and chemical principles behind the structure−property relationships.

ML provides robust models of additional properties of intermetallics. Rück et al. applied a KRR model to study the effects of strain on the activity of bimetallic particles with a core−shell structure. This analysis determined that the optimal strain depended on the particle size.[311] Increasing the compressive strain on 1.92 nm Pt@Cu and Pt@Ni particles, or decreasing the compressive strain on 2.83 nm Pt@Ag and Pt@Au particles, improved activity. In another study, Timoshenko et al. used a NN model to extract information on composition, structure, and atomic dynamics of bimetallic materials from extended X-ray absorption fine structure spectroscopy.[130] The method identified asymmetric and non-Gaussian distributions of bond length, small distorted metal nanoparticles, and thermal distortion of materials at high temperature and allowed the contributions from the distant atoms to be analyzed. It constituted a powerful tool for a broad range of in situ investigations of intermetallic nanomaterials, especially in catalysis.

3.1.2. Electrocatalytic Oxides. The overall effectiveness of electrocatalytic water splitting is determined mainly by the slow OER at the anode.[5,6] Therefore, an efficient OER catalyst is critical for boosting the hydrogen evolution efficiency. Many transition metal oxides, because of their interesting structure−activity relationships, abundance, and cost-effectiveness, have attracted substantial attention as electrocatalysts.[312−316] Because the materials space of oxides is essentially infinite, ML techniques are very suitable for fast screening of potential oxide electrocatalysts from materials databases and the literature.[53,54,317]

Hong et al. investigated 101 OER activities of 51 oxide perovskites using experimental measurements from the literature and identified relevant descriptors for screening oxide electrocatalysts.[131] They found five groups of descriptors, two of which, covalency and electron occupancy, played pivotal roles in the OER activity of catalysts. Among the investigated descriptors, the number of d electrons and charge-transfer energy were identified as being the most important for OER activity, while the occupancy of e_g orbitals, metal−oxide−metal bond angles, and tolerance factor were of secondary importance. They further demonstrated that multiple descriptors were essential to accurately predict OER activities. Weng et al. also generated effective descriptors using symbolic regression (which creates an optimal mathematic expression to describe a set of data based on a genetic algorithm).[132] They identified the combination of the tolerance factor (t) and the octahedral factor (μ), μ/t, as a primary descriptor for predicting the OER activity of oxide perovskites quantitatively. It was suggested that this new descriptor captured the relationship between structural stability and OER activities of perovskites (i.e., high OER activity resulted from low stability), thus accelerating discovery of highly efficient oxide perovskites

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

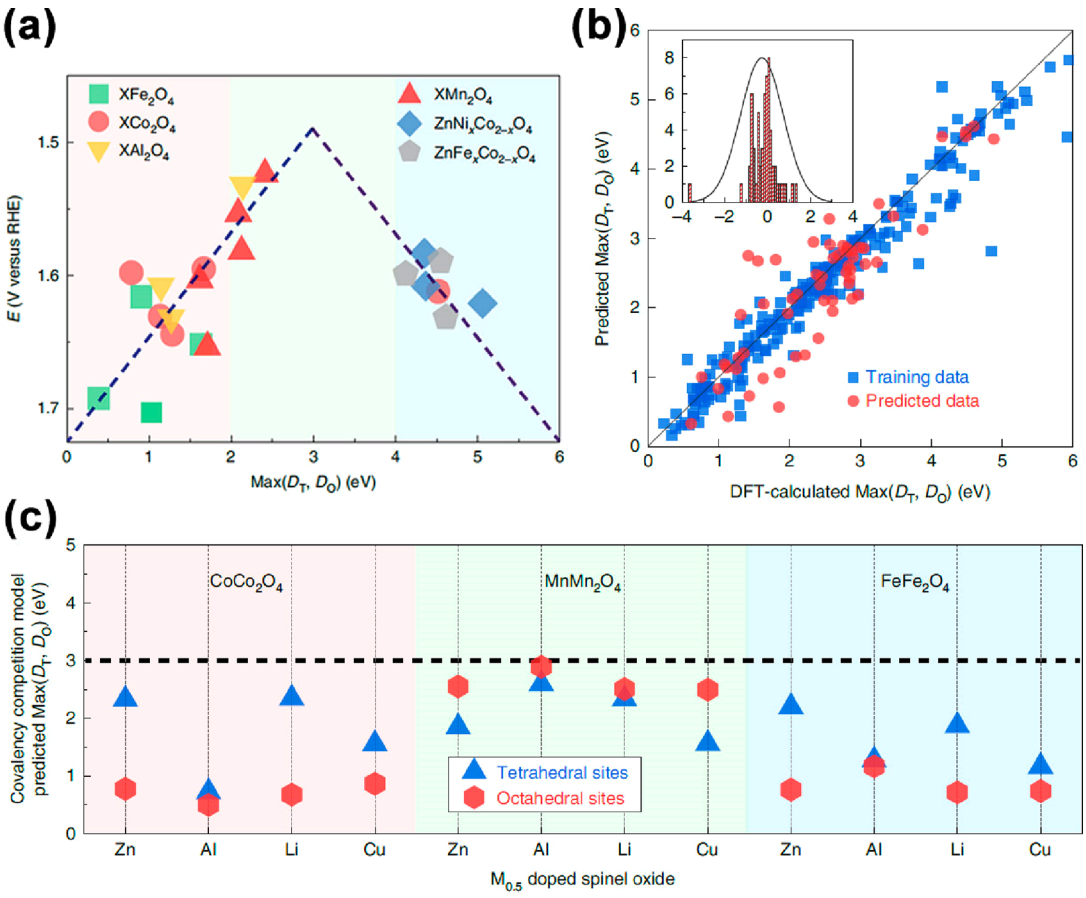


Figure 11. (a) Reaction activity observed in experiments versus the calculated Max(D_T, D_O). (b) DFT-calculated Max(D_T, D_O) versus RF-model-predicted Max(D_T, D_O). The *x*-axis in the inset shows the deviation between the RF-model-predicted and DFT-calculated Max(D_T, D_O) values, and the *y*-axis is the counts of the dots. (c) RF-model-predicted Max(D_T, D_O) values of $CoCo_2O_4$, $MnMn_2O_4$, and $FeFe_2O_4$ substituted by $M_{0.5}$ (M = Zn, Al, Li, and Cu). Reprinted with permission from ref 133. Copyright 2020, Springer Nature.

for OER. Xu et al. used SISSO to identify key features that predicted the adsorption enthalpies of OH* and OOH* on the surfaces of IrO_2 and RuO_2 doped with transition metal ions.[318] They first selected the 24 least correlated descriptors for the geometric, electronic, and atomic features of five low-index facets of various doped IrO_2 and RuO_2. A huge feature set was constructed by SISSO by applying operators to these descriptors. Descriptors with the highest correlation with adsorption enthalpies were selected, and a sparsifying operator was applied to find the best parsimonious solution. Co and Fe were suggested as the best dopants for promoting the activities of IrO_2 and RuO_2 in the OER.

More precise models resulted from a better understanding of the structure of electrocatalytic oxides. Sun et al. studied a series of spinel oxides (AB_2O_4 where A^{2+} is located at the tetrahedral sites and B^{3+} at the octahedral sites), and revealed that the stability of the asymmetric A−O−B backbone was essential for the OER.[133] A−O−B in which bond cleavage occurs more readily under OER conditions was beneficial to the formation of the OER intermediates. Conversely, a strongly polarized A−O or B−O bond resulted in exposed metals lacking unpaired electrons after bond breakage, hindering the OER. Therefore, the spinel oxides with moderate D_T (the distance between the d-band center of A and O p band center) or D_O (the distance between the d-band center of B and O p-band center) were identified as potentially highly active electrocatalysts (Figure 11). A RF model was trained on >300 spinel oxides described by elemental and structural features to predict the D_T and D_O of the spinel oxides. As the model performed well on the test set, it was used to screen other spinel oxides containing abundant elements. One spinel oxide, $[Mn][Al_{0.25}Mn_{0.75}]_2O_4$, was identified as promising and was subsequently confirmed to have superior OER activity.

ML techniques are also utilized to optimize experimental parameters on electrocatalytic degradation. Yu et al. built a NN model to find the optimal experimental conditions for the electrocatalytic degradation of sulfamethazine on an IrO_2−RuO_2 electrode.[319] The model illustrated that reaction time played the most important role in the sulfamethazine removal, while pH value strongly affected the efficiency of reducing chemical oxygen demand. However, it should be noted that their NN model was trained on only 91 data points. Despite the high accuracy reported, the domain of applicability of this model should be further investigated.

3.1.3. Single-Atom Catalysts. General supported metal catalysts consist of metal particles with a broad size distribution and multiple active sites dispersed on a substrate. The heterogeneity attenuates the efficiency and selectivity of the catalysts. To take full advantage of the active sites on these catalysts, the metal particle size must be reduced and controlled. Accordingly, when each metal atom is completely isolated and singly dispersed on supports, single-atom catalysts (SACs) are achieved, which is the ideal morphology for promotion of catalytic activity and selectivity.[320−322] Single-

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

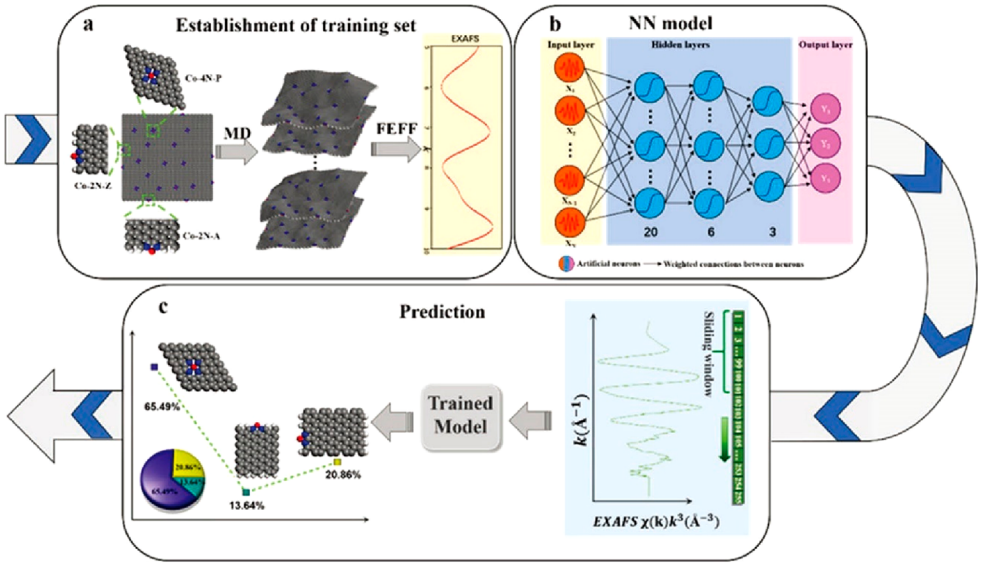


Figure 12. Schematic ML strategy to interpret the EXAFS. (a) Building data set from MD-EXAFS calculation. (b) Structure of the NN model. (c) Prediction of local structural proportion from the experimental EXAFS measurement. Reprinted with permission from ref 136. Copyright 2021, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim.

metal atoms can be anchored to metal oxides, metals, graphene, and porous materials such as metal−organic frameworks and zeolites. Many SACs have high reactivity and selectivity in the HER and OER. ML has been applied to accelerate the development of SACs and to elucidate feature−activity relationships in conjunction with DFT calculations.[323]

To develop novel and efficient SACs for the HER, Fung et al. used SISSO methods to study the transition metal SACs in N-doped graphene and to select the most relevant descriptors for ΔG_H.[134] Covalent radius, d-state centers, electronegativity, the number of occupied d states, and Bader charge were identified as relevant. This knowledge was used to design improved HER catalysts by modifying the graphene supports. Sun et al. investigated the HER processes in graphdiyne-based SACs by combining DFT and ML.[135] Simple atomic descriptors, active sites, and redox energy barriers were employed as features to predict ΔG_H via a tree-based bagging algorithm. As the ML predictions were highly consistent with DFT calculations, fast screening of HER catalyst candidates for experimental synthesis via ML techniques was feasible. To fully understand the relationship between SAC structure and catalytic activity, and to realize the potential of SACs, synchrotron spectroscopy was used to examine the SACs structure and its relationship to activity. Liu et al. developed a NN-based model to extract structural information for Co SACs in N-doped graphene from the experimental EXAFS (Figure 12).[136] The most active material consisted of 13.64% Co atoms located at the armchair sites and 20.86% Co atoms at the zigzag sites of graphene, consistent with DFT calculations. This revealed that edge sites (armchair and zigzag) were responsible for the HER activity of Co SACs. As this Co SAC had the highest HER performance of all reported transition metal SACs, with better HER performance at high current density than commercial Pt/C, more Co-based high-efficiency SACs may be accessible by manipulating the location of Co atoms in graphene.

SACs also play crucial roles in the ORR and OER. Lin et al. investigated the connection between electronic and atomic descriptors of SACs in N-doped graphene and their limiting potentials toward the HER/ORR/OER.[137] Limiting potential is the potential at which electrochemical reactions commence, which is the most straightforward standard to assess the activity of electrocatalysts. Random forest models for the ORR, OER, and HER were trained on 104 graphene-supported SACs whose limiting potentials were calculated by DFT. The limiting potentials predicted by the models were in good agreement with the DFT calculations, and were used to discover additional SACs, Ir@pyridine-N_3C_1, Ir@pyridine-N_2C_2, and one HER SAC, Ni@pyridine-N_3C_1. These novel SACs were predicted to have better activity than commonly used catalysts. Niu et al. built a GBR model to capture the relationship between ORR/OER activity and structural and atomic features of transition metal SACs in g-C_3N_4.[138] Two important features, the first ionization energy and charge transfer of transition metal atoms, were identified that elucidated the properties of SACs. In addition to SAC screening, ML methods are robust for synthesis optimization. Karim et al. synthesized 36 Fe SACs in zeolites by manipulating the experimental parameters.[139] GBR was employed to construct a ML model to link synthesis conditions with ORR activity and, accordingly, the best synthesis conditions were found. A Fe SAC with a significant increase in ORR activity compared to the original data set was obtained.

The formation of ammonia from nitrogen reduction is thermodynamically unfavorable, because the high thermodynamic barrier for cleaving the N−N triple bond energy (940.95 kJ mol^{-1}) must be overcome.[5] Although electrocatalytic nitrogen reduction is a promising approach to break the N−N triple bond, it is a complex system involving six-electron transfers, and it is difficult for DFT to identify all intermediates and suggest catalysts with high selectivity for this reaction. Therefore, ML techniques can elucidate the reaction mechanism, reveal a feature−activity relationship, and screen catalysts based on electrocatalyst data sets. Zarafi et al. used a DNN model to predict the N_2 adsorption energy and the hydrogenation energies of transition metal SACs in B-doped graphene.[140] It was found that the determining step in this system was the reduction of N_2 to N_2H, and, accordingly, three

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

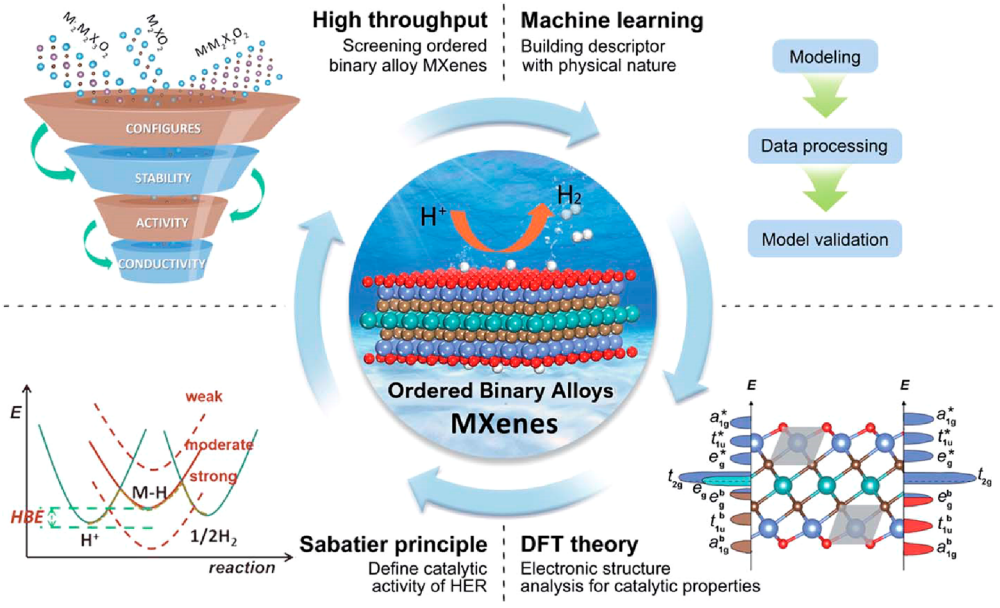


Figure 13. Schematic 2D MXene ordered binary alloy HER catalytic material discovery. Reprinted with permission from ref 142. Copyright 2020, Royal Society of Chemistry.

catalysts with low predicted hydrogenation energies were selected. These catalysts were further demonstrated to be more selective for nitrogen reduction than the HER. This study indicated that ML had great potential for screening catalysts for nitrogen reduction.

3.1.4. Other Electrocatalysts. 2D materials possess large surface area−volume ratios and superior catalytic properties that make them useful for energy storage and conversion. Transition metal dichalcogenides (TMDCs) are typical 2D materials used as HER catalysts because of their excellent electron transfer efficiency.[5] Ge et al. studied the electrocatalytic performance of the heterojunction of two single-layer MX_2 (M = Mo and W; X = S, Se, Te).[141] LASSO was employed to train ML models with rotational angle, bond length, distance between layers, and the ratio of the bandgaps of two MX_2 as descriptors and HER/OER overpotentials as targets. According to the model, a $MoTe_2/WTe_2$ heterojunction with a rotation of 300° was identified as the best electrocatalyst for the HER in this system.

MXenes are another type of 2D material.[324] They conform to the general formula $M_{n+1}X_n$, where M is an element from group IIIB to VIB, X is C or N, and n is typically 1−3. Because of their abundant surface charges, surface functionalization is facile for MXenes. An almost infinite number of MXenes can be generated by combining transition metals, carbon/nitrogen atoms, and a large variety of surface functional groups. This provides a rich palette of potential new materials for optical, electronic, catalytic, and energy applications. ML methods are essential to explore this vast materials landscape. Wang et al. used a series of ML algorithms to map the intrinsic features of 2D MXene ordered binary alloys to their hydrogen adsorption energy (Figure 13).[142] An Adaboost model showed that the HER performance could be predicted by five descriptors: the bond length of oxygen and surface metal atoms, the distance between the nearest neighbor O atoms, the ionization energy difference and the average affinity energy of the alloy elements, and the valence electrons of X. Zheng et al. used a RF model to screen 299 MXenes and identified a set of Os_2B- and S-terminated materials as potential HER catalysts in which the S functional groups are critical for HER performance.[325] These results showed the promise of ensemble learning for discovering highly efficient MXene catalysts.

When element X in MXenes is replaced by boron, a different type of 2D material, MBenes, is generated. Experiments and theoretical computations found that M_2B_1- and M_2B_2-type MBenes are promising catalysts in the HER because of their metallic electrical conductivity. To efficiently identify useful MBenes as HER catalysts, Sun et al. first built a data set of 110 pristine MBenes and 70 single-atom-doped MBenes and used the DFT-calculated ΔG_H as the model target.[143] Easily accessible descriptors (elemental descriptors, structural energies, and lattice parameters) were used to describe the compounds, and four ML algorithms (LASSO, KRR, SVR, and RF) were used to train models. The LASSO regression showed that Bader charge transfer of the surface metal was the most important descriptor affecting ΔG_H, while the d-band center of the surface metal was also important. The SVR had the best performance, and Co_2B_2 and Mn-doped Co_2B_2 MBenes were discovered with predicted superior HER catalyst performance.

3.2. Photocatalysis

Photocatalysis converts solar energy into chemical energy.[326] Useful applications include photocatalytic water splitting, water pollutant degradation, and CO_2 reduction.[8,9,326−328] In contrast to electrochemical reactions, photochemical reactions are driven by photons. The photocatalyst absorbs photons with an energy greater than the bandgap of the photocatalyst; electron−hole pairs are generated as electrons move to the conduction band (CB), leaving holes in the valence band (VB). These photogenerated electrons and holes will then either recombine or move through the photocatalyst to the surface. Once at the surface they can initiate redox reactions, provided the band energies of the photocatalyst are well matched to the reactants. For example, the electrons and holes can trigger the H^+ reduction (hydrogen evolution) and H_2O oxidation (oxygen evolution), respectively, or generate radicals (e.g., $^{\bullet}OH$, $O_2^{\bullet-}$) to oxidize organic pollutants.[9,326,327] Photocatalysts are promising materials for generating sustain-

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

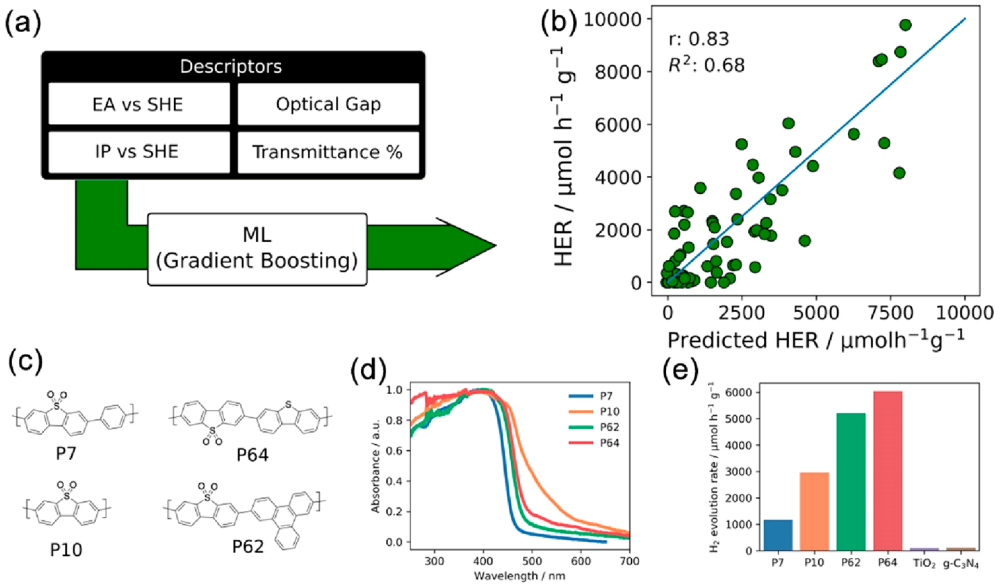


Figure 14. (a) Properties used to train the GBR model. (b) Experimentally observed HER vs HER predicted using a GBR model. (c) Structures of previously reported polymer photocatalysts (P7 and P10) and the best-performing polymer photocatalysts predicted by the ML model (P64 and P62). (d) UV−vis spectra of P7, P10, P62, and P64. (e) H_2 evolution rates of P7, P10, P62, and P64. Reprinted with permission from ref 144. Copyright 2019, American Chemical Society.

able fuels from solar energy, mitigating increases in greenhouse gases, and purifying polluted water. Those with high efficiency, low cost, and chemical stability are in high demand. ML methods have been used to explore large potential photocatalyst spaces, particularly with regard to the specificity of photocatalytic reactions and the complexity of reaction kinetics and thermodynamics. They aim to accelerate photocatalyst screening and optimization of photocatalytic reaction parameters and promise considerable cost reduction for catalyst design and discovery.[52,329,330]

3.2.1. Photocatalytic Organics for Water Splitting. Water splitting is one of the most important applications of photocatalysis. Although most photocatalysts are inorganic, certain organics have shown photocatalytic activity. Among the 100 million unique chemical compounds documented in the PubChem database, a special subset, the conjugated polymers, have attracted considerable attention.[331] Conjugated polymers contain π-conjugated backbones, ionic-functionalized alkyl side chains, and counterions. The π-electrons in the backbone enhance charge transfer and endow these polymers with excellent optical properties, while the ionic-functionality enables them to disperse into polar solvents, such as water. Therefore, a range of conjugated polymers have emerged as photocatalysts,[332−334] but how the structural components (backbone, side chain, ionic group, and counterions) of these polymers affect the photocatalytic properties is still unclear. To study the structure−property relationship and help design better photocatalysts, Zwijnenburg et al. used a GBR model to search 6354 conjugated copolymers to identify new photocatalysts with high sacrificial hydrogen evolution rates (Figure 14).[144] The copolymers were generated by linking 706 dibromo monomers with 9 diboronic acids or esters. The training set of copolymers was generated from 127 dibromides coupled with dibenzo[*b*,*d*]thiophene sulfone characterized for their water-splitting capacities. Each copolymer was encoded by four descriptors: electron affinity (EA, approximated by the energy of the LUMO); ionization potential (IP, approximated by the energy of the HOMO); optical bandgap; and transmittance (a measure of how well the polymer disperses in the reaction medium). The study found that active catalysts required a combination of negative electron affinities, more positive ionization energies, larger bandgaps, and solvent dispersibility. Wan et al. used a RF algorithm to study the bandgap, HOMO, and LUMO of 1296 anionic conjugated polymers, as these properties are important for photocatalytic activity.[145] They deduced that the frontier orbitals of the donor and acceptor components in the backbone play critical roles in determining the bandgap and the position of the frontier orbitals of these polymers. Rational criteria therefore could be proposed for the photocatalyst design. Li et al. probed the relationship between polymer structure and HER activity using ML techniques.[166] Using *k*-means clustering they found that most polymers with high activities shared a common structural feature, the presence of at least one aryl carbonyl moiety. However, the photocatalytic activity of polymers possessing this moiety may differ. Additional structure−property relationships were discovered using a series of supervised learning methods, and exciton electron affinity, electron affinity, exciton binding energy, optical gap, and singlet−triplet energy gap were identified as the most relevant features. These ML models are potentially useful for fast screening of photocatalysts.

3.2.2. Photocatalytic Oxides for Water Splitting. TiO_2 is one of the most widely used photocatalysts because of its abundance, chemical stability, and high photocatalytic activity.[8] Understanding the mechanism by which water molecules dissociate on a TiO_2 surface is of critical importance for TiO_2-based photocatalysts design. Selloni et al. studied this process using a ML-assisted molecular dynamics simulation.[335] Instead of conventional DFT calculations, a DNN model was built to predict the atomic energy for the bulk anatase TiO_2, bulk water, and the anatase {101}−water interface. DFT-calculated properties of small clusters of atoms in these three systems were used to train a DNN model. This model predicted the atomic energy of several other atoms in these clusters. The

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

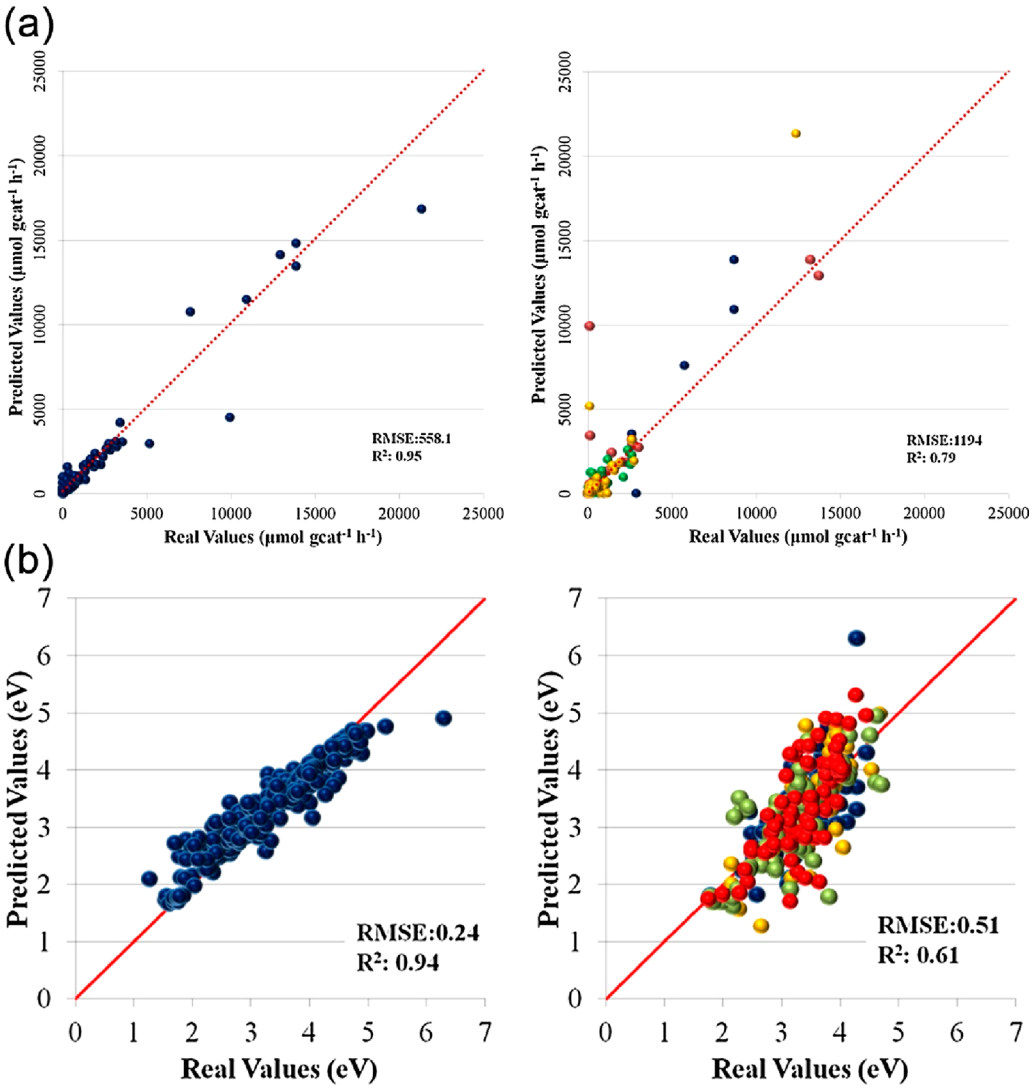


Figure 15. Predicted versus (a) actual hydrogen production and (b) actual bandgap for RF model for visible light data (left) training (right) testing. Reprinted with permission from ref 338. Copyright 2019, Elsevier.

predicted atomic energies were included in the molecular dynamics model, from which atoms with the largest maximum deviation of atomic forces were recalculated by DFT, and the DNN model was updated with these new data. This active learning process was repeated until the mean absolute deviation of atomic forces in a 100 ps simulation was below a given threshold, hence reducing the computational cost and allowing the simulation to extend from 40 ps (DFT calculation) to 2.5 ns. The dominant molecular adsorption mechanism on a defect-free anatase {101}−water interface was elucidated under ambient conditions. Detailed mechanisms of proton transfer at the interface, free energy barriers, and hydroxyl lifetimes were also estimated. A deeper understanding of the dynamics of water dissociation on a TiO_2 surface generated by this approach will assist the synthesis of improved TiO_2-based photocatalysts. In addition to water dissociation efficiency, light absorption efficiency is another key metric for photocatalysts. Zhang et al. clarified the correlation between the bandgap of doped TiO_2 and lattice parameters and surface area using a ML approach. As these structural and morphological parameters were highly correlated with bandgap, the reported ML model could be used to guide the synthesis of doped and undoped TiO_2 with tunable bandgaps. When used for photocatalytic water splitting, experimental parameters can also have pivotal effects on the hydrogen evolution rate. Thus, Iliuta et al. optimized the parameters of hydrogen evolution of TiO_2 photocatalysts in glycerol solution by a NN model and a genetic algorithm.[336] Here, 321 μmol of H_2 was obtained from the photocatalytic reaction in 4 h under the optimal conditions suggested by the NN model.

Oxide perovskites are also effective photocatalysts.[10,328] Their structure-dependent optical and electronic properties suggest that the photocatalytic activity of perovskites can be readily tailored. For example, it is well-documented that anion doping, forming oxynitrides or oxysulfides, can significantly reduce their bandgap and improve the light absorption efficiency.[8] However, anion doping also increases the structural complexity of oxide perovskites, making the prediction of physical properties via conventional quantum chemical calculations intractable. Therefore, ML was used to accelerate DFT calculations. Kaneko et al. studied the anion ordering in $BaNbO_2N$ by modeling DFT calculation results with ridge regression.[337] The ML model reproduced the local anion ordering generated by DFT calculations with high fidelity and

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

identified that, in the most stable supercells, each Nb atom bound to two N atoms and all Nb−N chains were in a cis-conformation. This model described anion ordering in a larger supercell, and enabled prediction of many important electronic properties, such as bandgap.

Surprisingly, few ML modeling studies of perovskite photocatalyst design have been reported in the literature. One limitation is the paucity of experimental or computational data available for model training. Thus, most models are generated from limited training data and may be prone to overfitting and have small domains of applicability. Another limitation is the lack of unified standards for characterizing photocatalysts in different laboratories. Yildirim et al. data-mined 540 case studies from 151 published papers on photocatalytic water splitting by oxide perovskites.[338] They found that $SrTiO_3$ was the most frequently studied titanate; about half of the perovskites are doped at the A-site. Sol−gel was the most popular synthesis method, and 80% of the studies contained cocatalysts. The data mining further indicated that hydrogen production was affected by multiple factors, in accordance with the previous results from ML models. RF models were trained on these data to predict the H_2 evolution rate and bandgap, giving modest but acceptable accuracy (Figure 15). Tao et al. used ML models to screen perovskites with visible light activity.[146] A GBR model for bandgap prediction was trained on 124 reported perovskites, and a NN model for hydrogen evolution rate prediction was trained on 76 reported perovskites. These models were used to screen a material space of >30 000 perovskites with unknown properties. Fourteen oxide perovskites with a bandgap in the range 2.2−2.4 eV and high hydrogen evolution rate were identified. However, the domain of applicability of these models was limited by the small size of the training set. To develop more accurate ML models, there is an urgent need to standardize semiconductor photocatalysis measurements, and the research community should validate experimental results through collaboration, thus making more data available in the literature that can be used to train improved perovskite ML models.

3.2.3. Photocatalytic Oxides for Pollutant Degradation. ML approaches have also been used to study the kinetics of photocatalytic pollutant degradation and to optimize the parameters for product synthesis and photocatalytic reactions.[339−347] Jiang et al. trained a NN model to decode the factors affecting the degradation of 78 organic pollutants by a TiO_2 photocatalyst.[348] The ultraviolet intensity, TiO_2 dosage, the concentration of organic pollutants in water, pH value and temperature, and the structure of organics (described by molecular fingerprint) were used to train the model, using 446 data points from the literature. They concluded that the structure of organic pollutants had the primary influence on degradation rate. Fathinia et al. studied the photocatalytic degradation of phenazopyridine hydrochloride by TiO_2 using a NN model.[147] Experimental parameters were used as descriptors, and the model was trained, validated, and tested on data sets of 128, 28, and 28 samples, respectively. A NN model with one hidden layer (5-14-1) was trained to predict the removal efficiency of phenazopyridine hydrochloride. The prediction accuracy of the model was high, similar to that of a kinetic model, suggesting that the reaction mechanism used in the kinetic model was correct. However, this model may be overtrained as the number of weights (84) is close to the number of data points. With the help of ML, Pellegrino et al. studied the effects of experimental parameters on TiO_2 nanoparticle sizes and aspect ratio in hydrothermal synthesis. According to the NN models, the length and the aspect ratio of the TiO_2 particles could be precisely controlled in the range from 20 to 140 nm and 1.4 to 6, respectively.[349] Yang et al. also attempted to use NN to build models for the photocatalytic performance of TiO_2 nanoparticles, but their results may also be overfitted as the training set was very small.[350] Such ML approaches may be extended to the design and analysis of more complex systems, such as ZnO, CuO, Fe_2O_3, $CoBi_2O_4$, and some high entropy oxides.[190,351−355] ML techniques are also used to study the photodegradation of doped oxides. In a study of the photocatalytic degradation of methylene blue by the N,S-codoped Fe_2O_3, a genetic algorithm was used to find optimal initial weights and biases for a NN model.[352] The study identified the importance of the amount of nanoparticles, concentration of dye, pH value, and the incident light wavelength on dye removal efficiency. Gao et al. used a NN-based active learning approach to screen the most active C/N-doped (001) TiO_2 for pollutant degradation.[148] In this synthesis, titanium carbide and titanium nitride served as the Ti, C, and N sources and hydrogen fluoride (HF) was the capping agent. The first NN model was built to link the synthesis parameters to photocatalytic performance based on the results of the experiments in the first round. Experiments were then carried out according to the best synthesis parameters given by the model, and the experimental photocatalytic activities of the pristine TiO_2 were added into the training set for the second NN model. This process was repeated until all the experimental results were accommodated by the model. Consequently, optimized C/N-doped TiO_2 was synthesized that showed good degradation performance toward three organic pollutants.

As well as catalysts composed of one metal oxide, ML methods are also useful for modeling and predicting the performance of catalysts containing two or more oxides. The formation of a heterojunction can potentially extend the electron−hole pair lifetime and modify the charge transfer during the reactions.[356] Badiei et al. reported a NN modeling approach to study the reaction rate constant of the photocatalytic degradation of acid fuchsin by a TiO_2/NiO heterojunction with various ions mixed in the reaction solution.[357] The NN model successfully ranked the effect of the ions on the reaction rate constant, recapitulating experimental results. Thus, this model can guide the optimization of experimental conditions in photocatalytic pollutant degradation. Kassahun et al. employed NN models to optimize the synthesis conditions of the N-doped TiO_2/SiO_2 heterojunction to get an optimum ciprofloxacin degradation efficiency under visible light.[358] Similar NN approaches were extended to other systems such as S-doped g-C_3N_4/ZnO, g-C_3N_4/Ce-doped ZnO/Ti, and a CdS/$CaFe_2O_4$ Z-scheme system.[359−361] These studies showed that NNs had considerable promise for optimization of the synthesis of photocatalysts containing two materials, but that some of these models might be overfitted as the training sets were small.

We have summarized the applications of ML techniques to electro/photocatalysis. ML techniques have been useful in leveraging expensive DFT calculations and for material screening, revealing feature−property relationships, and optimizing experimental parameters. SVM, NN, RF, and gradient boosting models appear to be the most frequently used models in electro/photocatalysis research; however, there

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

are no specific guides for model selection. The selection of ML models will rely on data quality and diversity, data set size, descriptor relevance, and the property to be modeled. Despite many successful applications, the use of ML in kinetic investigations, particularly in photocatalysis, is still rare. For example, proton coupled electron transfer (PCET) is desirable in electro/photocatalysis, because the thermodynamic barriers to the reaction can be largely reduced.[362−364] To study the PCET mechanisms, DFT calculations have been applied to understand the PCET mechanisms in benzimidazole-phenol (BIP) constructs (model molecules used to study PCET mechanism).[365,366] Although the calculations were consistent with experimental results on electrochemical time scales, indicating a concerted PCET mechanism, the PCET process became asynchronous on the ultrafast time scale, especially when BIP derivatives were attached to a photo absorber such as porphyrin.[367] To illustrate the fundamental PCET mechanisms on an ultrafast time scale, a NN-assisted first-principles molecular dynamics method was used to identify the key molecular vibrational modes for proton transfer.[368] The understanding of the PCET mechanisms contributed to the design of PCET systems in catalysis by providing guides for tuning the electron/proton transfer delay time and the multiproton transfer delay times, which could promote the preference of proton transfer through one pathway rather than the others, enhancing catalytic activity and selectivity.[364] However, there are still very few reports of ML models used to illustrate the electron transfer mechanisms in electro/photocatalytic systems, probably because of the lack of in situ technologies to acquire sufficient data. Another field in photocatalysis that ML techniques have seldom been applied to is the kinetic study of sacrificial photocatalytic systems. Sacrificial reagents play important roles in photocatalytic redox reactions, where they can scavenge one type of charge (e.g., holes) and improve the half reaction driven by the other type of charge (e.g., electrons).[8] Sacrificial reagents are frequently employed in photocatalytic water splitting, as the reduction and oxidation of water involving four electrons have a large thermodynamic barrier. By adding hole scavengers, the H_2 production is improved, and the back reactions are suppressed. Because of the intermediates formed in the oxidation of sacrificial reagents, however, the water reduction may be affected, resulting in low hydrogen production.[369] Therefore, it is critical to understand the mechanisms of sacrificial reagent reactions and discover the most useful sacrificial reagents for diverse photocatalytic reactions. Although the capture of the intermediates in sacrificial reagent reactions is difficult, we expect that ML techniques can consolidate the large amount of experimental and computational data in this field to provide mechanistic insights and identify effective sacrificial reagents for the target photocatalytic reaction. We suggest that close cooperation between data scientists and catalyst scientists will be critical for the development of ML models that can provide deeper insight into the electro/photocatalytic kinetics.

4. CONCLUSIONS AND PERSPECTIVE

4.1. Summary of the use of ML in Electro/Photocatalysis Studies

AI and ML techniques are powerful tools for the discovery of next-generation electrocatalysts and photocatalysts, investigation of the mechanism of catalysis, and understanding of relationships between features and catalytic activities. Given sufficiently large high-quality data sets, ML can swiftly identify patterns in the data and generate underlying structure−property relationships, significantly decreasing the need for laboratory experiment and computation time and resources, while providing comparable accuracy to first-principles calculations. ML approaches can usefully leverage high throughput experiments and calculations, allowing new and larger areas of materials space to be probed. This allows expensive or scarce resources to be concentrated on the most promising candidates, thereby accelerating discovery and optimization of catalysts and elucidating chemical and physical mechanisms responsible for their properties.

4.2. Pitfalls of Data-Driven Electro/Photocatalyst Design and Discovery

Despite a plethora of successful applications in electro/photocatalysis, there are common issues in ML modeling that can severely compromise model performance. As a data-driven approach, the performance of ML models depends heavily on the quantity, diversity, and quality of training data. In most of the research summarized in section 3, training data are collected from materials databases, literature, and high throughput experiments or calculations. However, data comparability must be carefully scrutinized prior to model training. For example, because the HER is sensitive to experimental conditions (e.g., reaction solution, cocatalysts, and the particle size and morphology of the catalysts), only HER data generated under very similar experimental conditions should be used to train ML models describing the relationship between material structures and the HER. Therefore, although the amount of published electro/photocatalytic data rises exponentially every year, only a small fraction can be applied for ML model training. Likewise, experimental conditions or calculation parameters should be consistent when the electro/photocatalytic data are generated from high throughput experiments or calculations.

Descriptors play essential roles in ML modeling. Descriptors must contain enough information related to the target property and have low correlations with each other. ML model performance will be compromised when descriptors contain insufficient information, and models will become increasingly hard to interpret when descriptors are correlated or have highly nonlinear relationships to the target property. ML algorithms, such as RF and LASSO can be used to select informative descriptors from larger pools of candidate descriptors. New descriptors can also be created by the linear combination of the initial descriptors (PCA) or by nonlinear processes (t-SNE, UMAP), and the models therefore can be simplified. Note that models with composite descriptors may be difficult to interpret, and there is a trade-off between model transparency and complexity.

Selection of a ML algorithm to train a model is determined by the data structure and the target property. In section 2.3, we summarized the most popular ML algorithms for electro/photocatalysis, including their strengths and drawbacks, giving materials scientists options to select depending on the suitability of the algorithms for their projects. We stress that overfitting must be detected before applying a model for material discovery and optimization. Overfitting may result from the number of model parameters exceeding the amount of training data, and therefore it is not recommended to train a model with a large number of weights (e.g., NN) on a small data set (e.g., < 100). To avoid overfitting, reducing the

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

number of descriptors (via down-selecting or dimension reduction), expanding the training set, regularization, or including physical prior knowledge is effective. As a useful rule, overfitting is likely to be avoided when the model variables number is less than half of the number of the training data. In section 2.4, common methods for overfitting control were summarized. Cross-validation is effective for small data sets, while validating model predictivity by an independent test set is the most reliable method. In addition, it is essential to be aware of the domain of applicability of models when using them to screen for new materials in unknown material space. The further outside the model domain, the higher the prediction uncertainty and lower prediction confidence. Some dimension reduction techniques, such as t-SNE and UMAP can be used to illustrate the domain of applicability of models. Applying Bayesian regularization to NNs, in which the weights are replaced by the distribution of weights, can also estimate the uncertainties of model predictions.

Models may sometimes have good performance for most data but perform poorly on a small number of materials (outliers). It is unwise to exclude outliers without careful analysis.[16,370,371] In some cases, outliers may be the result of experimental error, identifiable when relevant measurements are repeated. Outliers may also be the result of features that have low occurrences in the training data. Therefore, the structure of the outliers, the descriptors, and the model parameters must be carefully analyzed, as useful information may be extracted therefrom.

4.3. Challenges and Opportunities of Machine Learning in Electro/Photocatalytic Applications

Although significant progress has been made in the application of ML to automated discovery and development of electrocatalysts and photocatalysts, numerous challenges remain. First, the construction of accurate and broadly applicable ML models relies on large, high-quality data sets, but acquiring these data sets can be difficult or expensive. Published data from the literature could be an important data source for ML model construction. However, catalytic reactions are not only strongly dependent on experimental parameters such as temperature, concentration, pressure, and flow rate but are also sensitive to other conditions that might be overlooked (e.g., shape of reactor, stirring speed). This can render different published results incomparable. More rigorous recording of all relevant metadata in experiments will overcome this issue as will the use of positive control materials in experiments. The reliability of the published data is another issue. For example, while sacrificial reagents are widely used in photocatalytic water splitting to capture holes and thus improve the hydrogen generation rate, it was found that hydrogen could be generated from sacrificial reagent oxidation.[372] Das et al. revealed that some solvents are extremely sensitive to UV light, degrading into CO, CH_4, and C_2H_4, causing biased results in photocatalytic CO_2 reduction.[373] Even under identical conditions, some experimental results are difficult to repeat by other laboratories.[374] As such, critical analysis of the published data by catalyst experts prior to adding them to training sets is important. In addition, ML models can be biased because only successful results tend to be reported, while low performing or inactive materials also provide useful information to ML models. Instead of using published data, widespread adoption of high throughput synthesis and characterization methods should alleviate these issues in the future. Also, collaborations between data scientists, materials scientists and catalysts scientists will lead to new ontologies and markup languages for recording and publishing data, and more reliable data for model training.

Second, despite some successful structure−electro/photoactivity modeling examples, kinetic studies of electro/photocatalytic reactions by ML modeling are rarely performed. We exemplified this using a ML-molecular dynamic study of PCET on an ultrafast time frame.[368] However, the application of ML to the investigation of other important kinetic processes, such as single or multielectron oxidation/reduction, charge transfer between solid−solid or solid−liquid interfaces, separation, and recombination of carriers, is still rare. This is probably because of the difficulty of probing the excited state of the molecules via computation and experiments, resulting in insufficient data for ML model training. Recently, some cutting edge experimental technologies, such as Kelvin probe force microscopy, femtosecond resolved transient absorption spectroscopy, and time-resolved transient photoluminescence, have been developed to study carrier mobility.[375,376] Time-dependent DFT has been used to study the dynamics of electrons at excited states in photocatalysts,[377] and ML techniques have found a role in accelerating and increasing the accuracy of the time-dependent DFT calculation.[378,379] It is expected that ML will contribute further to the electro/photocatalytic mechanisms when it is combined with the in situ investigation data and molecular dynamic simulations.

Third, as experimental or first-principles computational data will continue to be scarce, it is important to develop techniques for generating reliable ML models from smaller data sets. Cutting-edge solutions, such as reinforcement, active, and transfer learning, have been developed to address the learning problem with limited data. More recently, meta learning is proposed to be another promising solution to the issues of small data sets. Meta learning generates models that can adapt and generalize to new tasks and new environments through the learning experiences from other models.[380] In other words, a meta learning model will first learn relationships from the other relevant models trained on large data sets, and use this prior knowledge to solve the problems with smaller data sets. As an example, stacking-based meta learning have been used to learn the predictions from various base models, on which accurate models are constructed to predict the bandgap and hydrogen evolution activity of oxide photocatalysts with a wide range of structures.[381] Transformational ML is another meta learning algorithm promising to be the solution of small data set.[382] Unlike stacking that learns knowledge from multiple base models, transformational ML extracts knowledge across a set of related tasks. Currently, transformational ML is robust in drug discovery, and we expect that it will find its application in electro/photocatalysts development (e.g., a mapping of structure to catalytic activities for different molecules). It is anticipated that meta learning can accelerate the research of electro/photocatalytic materials with insufficient data. Ultimately, there is a limit to how much information a ML model can extract from limited amounts of data.

Fourthly, it is challenging to describe the complex catalytic mechanisms via simple and interpretable descriptors. As descriptors not only bridge the gaps between features of catalysts and the dynamic events in catalysis but also have a major impact on ML model quality, feature selection from a pool of potential descriptors is essential. The most popular strategies are selecting descriptors based on chemical and

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

physical intuition, and down-selecting descriptors by feature selection algorithms (e.g., PCA, LASSO, multiple linear regression with expectation maximization (MLREM)[383]) or from the pruning capabilities of algorithms themselves (e.g., tree models, regularization). New techniques, such as symbolic learning, generative NNs,[384] and SISSO, have also been employed to generate informative descriptors from the potential descriptor pool. Note that computed descriptors are more useful than those acquired by experimental measurement because of the time and cost issues with experiments, and the possibility of generating descriptors for materials not yet synthesized.[176] The development of better computable, efficient and interpretable materials descriptors is an important research need. The use of deep learning algorithms such as convolutional neural nets, generative adversarial networks (GANs), and encoder−decoder networks will allow the generation of useful latent descriptors from simple representations of molecules and materials. They will also allow ML models to be used to directly suggest molecules or materials with superior properties based on a trained model.

In summary, electrocatalysts and photocatalysts play key roles in renewable energy generation and environmental remediation. We have reviewed how ML techniques have been applied to research on electro/photocatalysts and are transforming the landscape of computation and experiment. Apart from accelerating material screening and understanding structure−property relationships, further cooperation among data scientists, computer technologists, and catalyst researchers will expedite the evolution of catalyst design. Although current experimental and computational efforts have not produced the optimal prototype for electro/photocatalysts, ML techniques have great potential for text mining, automatic extraction of scientific knowledge from literature, and identifying new types of functional materials for particular applications before their discovery.[63] Seminal work by Yildirim et al. was reviewed in Section 3.2.2.[338] Although the information from the literature was still collected manually and their models were simple, we suggest that advanced text mining techniques will identify potential prototypes with high electro/photoactivity based on the massive body of available scientific literature. Additionally, generative models such as variational autoencoders (VAEs) and GANs, allow inverse design, i.e., property to structure design.[50] These models can suggest areas in material space where there is a high possibility of finding target materials given a series of properties. Generative models have been used to design functional polymers,[385] drug-like molecules,[386] porous materials,[387] and photoanode materials,[388] exemplifying their potential to design electro/photocatalysts. Recently, Noh et al. developed an inverse design approach to discover vanadium oxides via using VAE models.[389] As many vanadium oxides can potentially be used as electro/photocatalysts, we postulate this approach will be relevant for designing electro/photocatalytic vanadium oxides when more features relevant to electro/photocatalysis, such as bandgap, band edge, and V−O bond strength, are added into the model. Agarwal et al. reported a data-driven method to discover new 2D solar water splitting catalysts by a conditional VAE model.[330] Despite only bandgap and band edge restrictions being applied, a useful in silico design was achieved, and a series of new 2D materials were identified as candidate photocatalysts. More important properties, such as the synthesizability of the materials and kinetic conditions (e.g., surface adsorption and desorption), could be applied to the modeling process to enhance its utility and accuracy. Deeper synergies among ML, robotics, synthesis and characterization, and the development of autonomous laboratories, where synthesis and characterization are automated and ML models optimize the experimental parameters according to the previous experimental results then propose the next experiments, will become more common.[26,390−393] Burger et al. demonstrated an excellent example of developing water splitting photocatalysts using an autonomous laboratory, in which the autonomous robot worked 1000 times faster than human researchers and identified photocatalyst mixtures 6 times more active than the initial components.[394] It is likely that by applying more advanced ML algorithms and integrating more catalytic information and data, such autonomous laboratories will not only improve the efficiency of catalyst discovery but also allow the design of the entire catalytic system (e.g., choices of cocatalysts, solvent, and sacrificial reagents). With the development of theory, experiment, and technology, we believe that ML will surmount current challenges and generate more accurate models with larger and more useful domains of applicability for catalysts exploration and design. The evolution of ML techniques will in turn accelerate the development of theory, experiment, and technologies for materials science.

AUTHOR INFORMATION

Corresponding Authors

Dehong Chen − *Applied Chemistry and Environmental Science, School of Science, STEM College, RMIT University, Melbourne, Victoria 3001, Australia;* orcid.org/0000-0003-2867-7155; Email: dehong.chen@rmit.edu.au

David A. Winkler − *Monash Institute of Pharmaceutical Sciences, Monash University, Parkville, Victoria 3052, Australia; Biochemistry and Chemistry, La Trobe University, Bundoora, Victoria 3042, Australia; School of Pharmacy, University of Nottingham, Nottingham NG7 2RD, United Kingdom;* orcid.org/0000-0002-7301-6076; Email: david.winkler@monash.edu

Rachel A. Caruso − *Applied Chemistry and Environmental Science, School of Science, STEM College, RMIT University, Melbourne, Victoria 3001, Australia;* orcid.org/0000-0003-4922-2256; Email: rachel.caruso@rmit.edu.au

Authors

Haoxin Mai − *Applied Chemistry and Environmental Science, School of Science, STEM College, RMIT University, Melbourne, Victoria 3001, Australia*

Tu C. Le − *School of Engineering, STEM College, RMIT University, Melbourne, Victoria 3001, Australia;* orcid.org/0000-0003-3552-8211

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.chemrev.2c00061

Notes

The authors declare no competing financial interest.

Biographies

Haoxin Mai received his M.Sc. degree in Information Technology and Computer Science from the University of Technology Sydney, and Ph.D. degree in Materials Chemistry from the Research School of Chemistry at the Australian National University in 2019. He has worked at the Royal Melbourne Institute of Technology (RMIT) University, Australia, as a research assistant since 2019. His research

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

interests include perovskite photocatalysis and photoluminescence, ferroelectric thin films, controllable synthesis of inorganic colloid nanocrystals, and machine learning.

Tu Le is a lecturer at the School of Engineering, STEM College, RMIT University. Prior to joining RMIT University in 2017, she worked at the Commonwealth Scientific and Industrial Research Organisation. She completed her Ph.D. at Swinburne University of Technology in 2010. She is interested in material design and development using machine learning algorithms for a broad range of applications such as sustainable energy generation and storage, sensors, and therapeutics.

Dehong Chen is a research fellow at Applied Chemistry & Environmental Science, RMIT University, Australia. He received his Ph.D. degree in Chemistry from Fudan University in 2006. Before joining RMIT University, he conducted research work at the University of Melbourne and the Australian National University. He currently has interests in the design and synthesis of diverse functional materials for a range of applications including clean energy utilization, energy storage, environmental remediation, and sensing.

David Winkler is a professor at La Trobe Institute for Molecular Science at La Trobe University, a visiting professor at the University of Nottingham, and a Professor of Medicinal Chemistry, Monash Institute of Pharmaceutical Sciences, Monash University. His research on applying computational chemistry, AI, and machine learning methods to the design of drugs, agrochemicals, nanomaterials, and biomaterials has led to over 200 journal articles and book chapters as well as 25 patents. He has won prestigious awards including the CSIRO Medal for Business Excellence, RACI's Adrien Albert award, and the ACS Herman Skolnik award. He is ranked 227th of 81 000 medicinal chemists, and 999th of 520 000 chemists worldwide (Mendeley 2019).

Rachel Caruso is a professor in the Applied Chemistry and Environmental Science discipline at RMIT University. She studied at The University of Melbourne before leading research teams at the Max Planck Institute of Colloids and Interfaces, The University of Melbourne, the Commonwealth Scientific and Industrial Research Organisation, and RMIT University. Her expertise in materials chemistry has been applied to investigating structural control of materials on the nanoscale, with a focus on porous metal oxides, perovskites, and carbon-based materials with potential for photocatalysis, adsorbents, and energy conversion and storage applications.

ACKNOWLEDGMENTS

The Australian Research Council is acknowledged for support through Discovery Projects DP180103815 and DP220100945.

REFERENCES

(1) Tachibana, Y.; Vayssieres, L.; Durrant, J. R. Artificial photosynthesis for solar water-splitting. *Nat. Photonics* **2012**, *6*, 511−518.
(2) Chu, S.; Majumdar, A. Opportunities and challenges for a sustainable energy future. *Nature* **2012**, *488*, 294−303.
(3) Kittner, N.; Lill, F.; Kammen, D. M. Energy storage deployment and innovation for the clean energy transition. *Nat. Energy* **2017**, *2*, 17125.
(4) Clarke, C. J.; Tu, W.-C.; Levers, O.; Bröhl, A.; Hallett, J. P. Green and Sustainable Solvents in Chemical Processes. *Chem. Rev.* **2018**, *118*, 747−800.
(5) Seh, Z. W.; Kibsgaard, J.; Dickens, C. F.; Chorkendorff, I.; Nørskov, J. K.; Jaramillo, T. F. Combining theory and experiment in electrocatalysis: Insights into materials design. *Science* **2017**, *355*, No. eaad4998.
(6) Suen, N.-T.; Hung, S.-F.; Quan, Q.; Zhang, N.; Xu, Y.-J.; Chen, H. M. Electrocatalysis for the oxygen evolution reaction: recent development and future perspectives. *Chem. Soc. Rev.* **2017**, *46*, 337−365.
(7) Roger, I.; Shipman, M. A.; Symes, M. D. Earth-abundant catalysts for electrochemical and photoelectrochemical water splitting. *Nat. Rev. Chem.* **2017**, *1*, 0003.
(8) Wang, Q.; Domen, K. Particulate Photocatalysts for Light-Driven Water Splitting: Mechanisms, Challenges, and Design Strategies. *Chem. Rev.* **2020**, *120*, 919−985.
(9) Chen, S.; Takata, T.; Domen, K. Particulate photocatalysts for overall water splitting. *Nat. Rev. Mater.* **2017**, *2*, 17050.
(10) Mai, H.; Chen, D.; Tachibana, Y.; Suzuki, H.; Abe, R.; Caruso, R. A. Developing sustainable, high-performance perovskites in photocatalysis: design strategies and applications. *Chem. Soc. Rev.* **2021**, *50*, 13692−13729.
(11) Low, J.; Yu, J.; Jaroniec, M.; Wageh, S.; Al-Ghamdi, A. A. Heterojunction Photocatalysts. *Adv. Mater.* **2017**, *29*, 1601694.
(12) Liu, X.; Iocozzia, J.; Wang, Y.; Cui, X.; Chen, Y.; Zhao, S.; Li, Z.; Lin, Z. Noble metal-metal oxide nanohybrids with tailored nanostructures for efficient solar energy conversion, photocatalysis and environmental remediation. *Energy Environ. Sci.* **2017**, *10*, 402−434.
(13) Wei, H.; McMaster, W. A.; Tan, J. Z. Y.; Chen, D.; Caruso, R. A. Tricomponent brookite/anatase TiO2/g-C3N4 heterojunction in mesoporous hollow microspheres for enhanced visible-light photocatalysis. *J. Mater. Chem. A* **2018**, *6*, 7236−7245.
(14) Singh, N.; Goldsmith, B. R. Role of Electrocatalysis in the Remediation of Water Pollutants. *ACS Catal.* **2020**, *10*, 3365−3371.
(15) Rong, F.; Lu, Q.; Mai, H.; Chen, D.; Caruso, R. A. Hierarchically Porous WO3/CdWO4 Fiber-in-Tube Nanostructures Featuring Readily Accessible Active Sites and Enhanced Photocatalytic Effectiveness for Antibiotic Degradation in Water. *ACS Appl. Mater. Interfaces* **2021**, *13*, 21138−21148.
(16) Le, T.; Epa, V. C.; Burden, F. R.; Winkler, D. A. Quantitative Structure-Property Relationship Modeling of Diverse Materials Properties. *Chem. Rev.* **2012**, *112*, 2889−2919.
(17) Reymond, J. L. The Chemical Space Project. *Acc. Chem. Res.* **2015**, *48*, 722−730.
(18) Walsh, A. The quest for new functionality. *Nat. Chem.* **2015**, *7*, 274−275.
(19) Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev. B* **1964**, *136*, B864.
(20) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, A1133.
(21) Potyrailo, R.; Rajan, K.; Stoewe, K.; Takeuchi, I.; Chisholm, B.; Lam, H. Combinatorial and High-Throughput Screening of Materials Libraries: Review of State of the Art. *ACS Comb. Sci.* **2011**, *13*, 579−633.
(22) Curtarolo, S.; Hart, G. L. W.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O. The high-throughput highway to computational materials design. *Nat. Mater.* **2013**, *12*, 191−201.
(23) Lejaeghere, K.; Bihlmayer, G.; Bjorkman, T.; Blaha, P.; Blugel, S.; Blum, V.; Caliste, D.; Castelli, I. E.; Clark, S. J.; Dal Corso, A.; et al. Reproducibility in density functional theory calculations of solids. *Science* **2016**, *351*, aad3000.
(24) Yan, Q.; Yu, J.; Suram, S. K.; Zhou, L.; Shinde, A.; Newhouse, P. F.; Chen, W.; Li, G.; Persson, K. A.; Gregoire, J. M.; et al. Solar fuels photoanode materials discovery by integrating high-throughput theory and experiment. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 3040−3043.
(25) Qiu, B.; Xing, M.; Zhang, J. Recent advances in three-dimensional graphene based materials for catalysis applications. *Chem. Soc. Rev.* **2018**, *47*, 2165−2216.
(26) Tabor, D. P.; Roch, L. M.; Saikin, S. K.; Kreisbeck, C.; Sheberla, D.; Montoya, J. H.; Dwaraknath, S.; Aykol, M.; Ortiz, C.; Tribukait, H.; et al. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nat. Rev. Mater.* **2018**, *3*, 5−20.

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

(27) Jordan, M. I.; Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255−260.
(28) Ghahramani, Z. Probabilistic machine learning and artificial intelligence. *Nature* **2015**, *521*, 452−459.
(29) Muratov, E. N.; Bajorath, J.; Sheridan, R. P.; Tetko, I. V.; Filimonov, D.; Poroikov, V.; Oprea, T. I.; Baskin, I. I.; Varnek, A.; Roitberg, A.; et al. QSAR without borders. *Chem. Soc. Rev.* **2020**, *49*, 3525−3564.
(30) Gomez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Chae, H. S.; Einzinger, M.; Ha, D. G.; Wu, T.; et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **2016**, *15*, 1120−1127.
(31) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547−555.
(32) Himanen, L.; Geurts, A.; Foster, A. S.; Rinke, P. Data-Driven Materials Science: Status, Challenges, and Perspectives. *Adv. Sci.* **2019**, *6*, 1900808.
(33) Gu, G. H.; Noh, J.; Kim, I.; Jung, Y. Machine learning for renewable energy materials. *J. Mater. Chem. A* **2019**, *7*, 17096−17117.
(34) Chen, C.; Zuo, Y. X.; Ye, W. K.; Li, X. G.; Deng, Z.; Ong, S. P. A Critical Review of Machine Learning of Energy Materials. *Adv. Energy Mater.* **2020**, *10*, 1903242.
(35) Toyao, T.; Maeno, Z.; Takakusagi, S.; Kamachi, T.; Takigawa, I.; Shimizu, K. Machine Learning for Catalysis Informatics: Recent Applications and Prospects. *ACS Catal.* **2020**, *10*, 2260−2297.
(36) Jablonka, K. M.; Ongari, D.; Moosavi, S. M.; Smit, B. Big-Data Science in Porous Materials: Materials Genomics and Machine Learning. *Chem. Rev.* **2020**, *120*, 8066−8129.
(37) Behler, J. First Principles Neural Network Potentials for Reactive Simulations of Large Molecular and Condensed Systems. *Angew. Chem., Int. Ed.* **2017**, *56*, 12828−12840.
(38) Ulissi, Z. W.; Medford, A. J.; Bligaard, T.; Nørskov, J. K. To address surface reaction network complexity using scaling relations machine learning and DFT calculations. *Nat. Commun.* **2017**, *8*, 14621.
(39) Mosquera, M. A.; Fu, B.; Kohlstedt, K. L.; Schatz, G. C.; Ratner, M. A. Wave Functions, Density Functionals, and Artificial Intelligence for Materials and Energy Research: Future Prospects and Challenges. *ACS Energy Lett.* **2018**, *3*, 155−162.
(40) O'Connor, N. J.; Jonayat, A. S. M.; Janik, M. J.; Senftle, T. P. Interaction trends between single metal atoms and oxide supports identified with density functional theory and statistical learning. *Nat. Catal.* **2018**, *1*, 531−539.
(41) Chmiela, S.; Sauceda, H. E.; Müller, K.-R.; Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* **2018**, *9*, 3887.
(42) Zhang, Y.; Mesaros, A.; Fujita, K.; Edkins, S. D.; Hamidian, M. H.; Ch'ng, K.; Eisaki, H.; Uchida, S.; Davis, J. C. S.; Khatami, E.; et al. Machine learning in electronic-quantum-matter imaging experiments. *Nature* **2019**, *570*, 484−490.
(43) Deringer, V. L.; Caro, M. A.; Csanyi, G. Machine Learning Interatomic Potentials as Emerging Tools for Materials Science. *Adv. Mater.* **2019**, *31*, 1902765.
(44) Schütt, K. T.; Gastegger, M.; Tkatchenko, A.; Müller, K. R.; Maurer, R. J. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nat. Commun.* **2019**, *10*, 5024.
(45) Pun, G. P. P.; Batra, R.; Ramprasad, R.; Mishin, Y. Physically informed artificial neural networks for atomistic modeling of materials. *Nat. Commun.* **2019**, *10*, 2339.
(46) Westermayr, J.; Marquetand, P. Machine Learning for Electronically Excited States of Molecules. *Chem. Rev.* **2021**, *121*, 9873−9926.
(47) Belianinov, A.; Ievlev, A. V.; Lorenz, M.; Borodinov, N.; Doughty, B.; Kalinin, S. V.; Fernandez, F. M.; Ovchinnikova, O. S. Correlated Materials Characterization via Multimodal Chemical and Functional Imaging. *ACS Nano* **2018**, *12*, 11798−11818.
(48) Rogers, D. M.; Jasim, S. B.; Dyer, N. T.; Auvray, F.; Refregiers, M.; Hirst, J. D. Electronic Circular Dichroism Spectroscopy of Proteins. *Chem.* **2019**, *5*, 2751−2774.
(49) Lansford, J. L.; Vlachos, D. G. Infrared spectroscopy data- and physics-driven machine learning for characterizing surface microstructure of complex materials. *Nat. Commun.* **2020**, *11*, 1513.
(50) Batra, R.; Song, L.; Ramprasad, R. Emerging materials intelligence ecosystems propelled by machine learning. *Nat. Rev. Mater.* **2021**, *6*, 655−678.
(51) Meuwly, M. Machine Learning for Chemical Reactions. *Chem. Rev.* **2021**, *121*, 10218−10239.
(52) Masood, H.; Toe, C. Y.; Teoh, W. Y.; Sethu, V.; Amal, R. Machine Learning for Accelerated Discovery of Solar Photocatalysts. *ACS Catal.* **2019**, *9*, 11774−11787.
(53) Palkovits, R.; Palkovits, S. Using Artificial Intelligence To Forecast Water Oxidation Catalysts. *ACS Catal.* **2019**, *9*, 8383−8387.
(54) Liu, J.; Liu, H.; Chen, H.; Du, X.; Zhang, B.; Hong, Z.; Sun, S.; Wang, W. Progress and Challenges Toward the Rational Design of Oxygen Electrocatalysts Based on a Descriptor Approach. *Adv. Sci.* **2020**, *7*, 1901614.
(55) Wang, M.; Zhu, H. Machine Learning for Transition-Metal-Based Hydrogen Generation Electrocatalysts. *ACS Catal.* **2021**, *11*, 3930−3937.
(56) Keith, J. A.; Vassilev-Galindo, V.; Cheng, B.; Chmiela, S.; Gastegger, M.; Müller, K.-R.; Tkatchenko, A. Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems. *Chem. Rev.* **2021**, *121*, 9816−9872.
(57) Huang, B.; von Lilienfeld, O. A. Ab Initio Machine Learning in Chemical Compound Space. *Chem. Rev.* **2021**, *121*, 10001−10036.
(58) Jia, X. W.; Lynch, A.; Huang, Y. H.; Danielson, M.; Lang'at, I.; Milder, A.; Ruby, A. E.; Wang, H.; Friedler, S. A.; Norquist, A. J.; et al. Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. *Nature* **2019**, *573*, 251−255.
(59) Kalidindi, S. R.; De Graef, M. Materials Data Science: Current Status and Future Outlook. *Ann. Rev. Mater. Res.* **2015**, *45*, 171−193.
(60) Krallinger, M.; Rabal, O.; Lourenco, A.; Oyarzabal, J.; Valencia, A. Information Retrieval and Text Mining Technologies for Chemistry. *Chem. Rev.* **2017**, *117*, 7673−7761.
(61) Hawizy, L.; Jessop, D. M.; Adams, N.; Murray-Rust, P. ChemicalTagger: A tool for semantic text-mining in chemistry. *J. Cheminform.* **2011**, *3*, 17.
(62) Swain, M. C.; Cole, J. M. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *J. Chem. Inf. Model.* **2016**, *56*, 1894−1904.
(63) Tshitoyan, V.; Dagdelen, J.; Weston, L.; Dunn, A.; Rong, Z. Q.; Kononova, O.; Persson, K. A.; Ceder, G.; Jain, A. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **2019**, *571*, 95−98.
(64) Kim, E.; Huang, K.; Saunders, A.; McCallum, A.; Ceder, G.; Olivetti, E. Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning. *Chem. Mater.* **2017**, *29*, 9436−9444.
(65) Zakutayev, A.; Wunder, N.; Schwarting, M.; Perkins, J. D.; White, R.; Munch, K.; Tumas, W.; Phillips, C. An open experimental database for exploring inorganic materials. *Sci. Data* **2018**, *5*, 180053.
(66) He, J.; Dettelbach, K. E.; Salvatore, D. A.; Li, T.; Berlinguette, C. P. High-Throughput Synthesis of Mixed-Metal Electrocatalysts for CO2 Reduction. *Angew. Chem., Int. Ed.* **2017**, *56*, 6068−6072.
(67) Tung, V. C.; Allen, M. J.; Yang, Y.; Kaner, R. B. High-throughput solution processing of large-scale graphene. *Nat. Nanotechnol.* **2009**, *4*, 25−29.
(68) Jeong, S.; Park, J.; Pathania, D.; Castro, C. M.; Weissleder, R.; Lee, H. Integrated Magneto-Electrochemical Sensor for Exosome Analysis. *ACS Nano* **2016**, *10*, 1802−1809.
(69) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, *51*, 1281−1289.
(70) Raccuglia, P.; Elbert, K. C.; Adler, P. D. F.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J.

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

Machine-learning-assisted materials discovery using failed experiments. *Nature* **2016**, *533*, 73−76.
(71) van Roekeghem, A.; Carrete, J.; Oses, C.; Curtarolo, S.; Mingo, N. High-Throughput Computation of Thermal Conductivity of High-Temperature Solid Phases: The Case of Oxide and Fluoride Perovskites. *Phys. Rev. X* **2016**, *6*, 041061.
(72) Greeley, J.; Jaramillo, T. F.; Bonde, J.; Chorkendorff, I. B.; Norskov, J. K. Computational high-throughput screening of electrocatalytic materials for hydrogen evolution. *Nat. Mater.* **2006**, *5*, 909−913.
(73) Sendek, A. D.; Yang, Q.; Cubuk, E. D.; Duerloo, K. A. N.; Cui, Y.; Reed, E. J. Holistic computational structure screening of more than 12 000 candidates for solid lithium-ion conductor materials. *Energy Environ. Sci.* **2017**, *10*, 306−320.
(74) Zhan, C.; Sun, W. W.; Xie, Y.; Jiang, D. E.; Kent, P. R. C. Computational Discovery and Design of MXenes for Energy Applications: Status, Successes, and Opportunities. *ACS Appl. Mater. Interfaces* **2019**, *11*, 24885−24905.
(75) Li, P.; Zhu, J.; Handoko, A. D.; Zhang, R.; Wang, H.; Legut, D.; Wen, X.; Fu, Z.; Seh, Z. W.; Zhang, Q. High-throughput theoretical optimization of the hydrogen evolution reaction on MXenes by transition metal modification. *J. Mater. Chem. A* **2018**, *6*, 4271−4278.
(76) Gubaev, K.; Podryabinkin, E. V.; Hart, G. L. W.; Shapeev, A. V. Accelerating high-throughput searches for new alloys with active learning of interatomic potentials. *Comput. Mater. Sci.* **2019**, *156*, 148−156.
(77) Kim, K.; Ward, L.; He, J. G.; Krishna, A.; Agrawal, A.; Wolverton, C. Machine-learning-accelerated high-throughput materials screening: Discovery of novel quaternary Hensler compounds. *Phys. Rev. Mater.* **2018**, *2*, 123801.
(78) Curtarolo, S.; Setyawan, W.; Hart, G. L. W.; Jahnatek, M.; Chepulskii, R. V.; Taylor, R. H.; Wang, S. D.; Xue, J. K.; Yang, K. S.; Levy, O.; et al. AFLOW: An automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.* **2012**, *58*, 218−226.
(79) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **2013**, *1*, 011002.
(80) Musil, F.; Grisafi, A.; Bartók, A. P.; Ortner, C.; Csányi, G.; Ceriotti, M. Physics-Inspired Structural Representations for Molecules and Materials. *Chem. Rev.* **2021**, *121*, 9759−9815.
(81) Langer, M. F.; Goeßmann, A.; Rupp, M. Representations of molecules and materials for interpolation of quantum-mechanical simulations via machine learning. *Npj Comput. Mater.* **2022**, *8*, 41.
(82) Jianchang, M.; Jain, A.K. Artificial neural networks for feature extraction and multivariate data projection. *IEEE trans. neural netw.* **1995**, *6*, 296−317.
(83) Wang, Z.; Wang, X.; Yang, W.; Xiao, Y.; Liu, Y.; Chen, L. yNet: a multi-input convolutional network for ultra-fast simulation of field evolvement. *arxiv* **2020**, *2012*, 10575.
(84) Schleider, L.; Pasiliao, E. L.; Qiang, Z.; Zheng, Q. P., A study of feature representation via neural network feature extraction and weighted distance for clustering. *J. Comb. Optim.* **2022**. DOI: 10.1007/s10878-022-00849-y
(85) Blum, A. L.; Langley, P. Selection of relevant features and examples in machine learning. *Artif. Intell.* **1997**, *97*, 245−271.
(86) Hua, J.; Xiong, Z.; Lowey, J.; Suh, E.; Dougherty, E. R. Optimal number of features as a function of sample size for various classification rules. *Bioinformatics* **2005**, *21*, 1509−1515.
(87) Wang, X.; Xiao, B.; Li, Y.; Tang, Y.; Liu, F.; Chen, J.; Liu, Y. First-principles based machine learning study of oxygen evolution reactions of perovskite oxides using a surface center-environment feature model. *Appl. Surf. Sci.* **2020**, *531*, 147323.
(88) Ma, X.; Li, Z.; Achenie, L. E. K.; Xin, H. Machine-Learning-Augmented Chemisorption Model for CO_2 Electroreduction Catalyst Screening. *J. Phys. Chem. Lett.* **2015**, *6*, 3528−3533.
(89) Liu, F.; Yang, S.; Medford, A. J. Scalable approach to high coverages on oxides via iterative training of a machine-learning algorithm. *ChemCatChem.* **2020**, *12*, 4317−4330.
(90) Vicente, R. A.; Neckel, I. T.; Sankaranarayanan, S. K. R. S.; Solla-Gullon, J.; Fernández, P. S. Bragg Coherent Diffraction Imaging for In Situ Studies in Electrocatalysis. *ACS Nano* **2021**, *15*, 6129−6146.
(91) Govind Rajan, A.; Martirez, J. M. P.; Carter, E. A. Why Do We Use the Materials and Operating Conditions We Use for Heterogeneous (Photo)Electrochemical Water Splitting? *ACS Catal.* **2020**, *10*, 11177−11234.
(92) Yang, W.; Fidelis, T. T.; Sun, W.-H. Machine Learning in Catalysis, From Proposal to Practicing. *Acs Omega* **2020**, *5*, 83−88.
(93) Faber, F. A.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R. Machine Learning Energies of 2 Million Elpasolite (ABC(2)D(6)) Crystals. *Phys. Rev. Lett.* **2016**, *117*, 135502.
(94) Bostanabad, R.; Zhang, Y. C.; Li, X. L.; Kearney, T.; Brinson, L. C.; Apley, D. W.; Liu, W. K.; Chen, W. Computational microstructure characterization and reconstruction: Review of the state-of-the-art techniques. *Prog. Mater. Sci.* **2018**, *95*, 1−41.
(95) Jha, D.; Ward, L.; Paul, A.; Liao, W. K.; Choudhary, A.; Wolverton, C.; Agrawal, A. ElemNet: Deep Learning the Chemistry of Materials From Only Elemental Composition. *Sci. Rep.* **2018**, *8*, 17593.
(96) Dean, J.; Taylor, M. G.; Mpourmpakis, G. Unfolding adsorption on metal nanoparticles: Connecting stability with catalysis. *Sci. Adv.* **2019**, *5*, No. eaax5101.
(97) Oliynyk, A. O.; Adutwum, L. A.; Rudyk, B. W.; Pisavadia, H.; Lotfi, S.; Hlukhyy, V.; Harynuk, J. J.; Mar, A.; Brgoch, J. Disentangling Structural Confusion through Machine Learning: Structure Prediction and Polymorphism of Equiatomic Ternary Phases ABC. *J. Am. Chem. Soc.* **2017**, *139*, 17870−17881.
(98) Filip, M. R.; Giustino, F. The geometric blueprint of perovskites. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*, 5397−5402.
(99) Li, Z.; Achenie, L. E. K.; Xin, H. An Adaptive Machine Learning Strategy for Accelerating Discovery of Perovskite Electrocatalysts. *ACS Catal.* **2020**, *10*, 4377−4384.
(100) Pauling, L. The principles determining the structure of complex ionic crystals. *J. Am. Chem. Soc.* **1929**, *51*, 1010−1026.
(101) Goldschmidt, V. M. Die Gesetze der Krystallochemie. *Naturwissenschaften* **1926**, *14*, 477−485.
(102) Schutt, K. T.; Glawe, H.; Brockherde, F.; Sanna, A.; Muller, K. R.; Gross, E. K. U. How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Phys. Rev. B* **2014**, *89*, 205118.
(103) Ward, L.; Liu, R. Q.; Krishna, A.; Hegde, V. I.; Agrawal, A.; Choudhary, A.; Wolverton, C. Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Phys. Rev. B* **2017**, *96*, 024104.
(104) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.
(105) Weininger, D. SMILES, a Chemical Language and Information-System.1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.
(106) Häse, F.; Roch, L. M.; Friederich, P.; Aspuru-Guzik, A. Designing and understanding light-harvesting devices with machine learning. *Nat. Commun.* **2020**, *11*, 4587.
(107) Yap, C. W. PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466−1474.
(108) Elton, D. C.; Boukouvalas, Z.; Butrico, M. S.; Fuge, M. D.; Chung, P. W. Applying machine learning techniques to predict the properties of energetic materials. *Sci. Rep.* **2018**, *8*, 9059.
(109) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273−1280.

(110) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminform.* **2011**, *3*, 33.
(111) Steinbeck, C.; Han, Y. Q.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An open-source Java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493−500.
(112) O'Boyle, N. M.; Sayle, R. A. Comparing structural fingerprints using a literature-based similarity benchmark. *J. Cheminform.* **2016**, *8*, 36.
(113) Meftahi, N.; Klymenko, M.; Christofferson, A. J.; Bach, U.; Winkler, D. A.; Russo, S. P. Machine learning property prediction for organic photovoltaic devices. *Npj Comput. Mater.* **2020**, *6*, 166.
(114) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.
(115) Bartok, A. P.; Kondor, R.; Csanyi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, 184115.
(116) Bartok, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R.; Csanyi, G.; Ceriotti, M. Machine learning unifies the modeling of materials and molecules. *Sci. Adv.* **2017**, *3*, No. e1701816.
(117) Jäger, M. O. J.; Morooka, E. V.; Federici Canova, F.; Himanen, L.; Foster, A. S. Machine Learning Hydrogen Adsorption on Nanoclusters through Structural Descriptors. *Npj Comput. Mater.* **2018**, *4*, 37.
(118) Lee, Y. L.; Kleis, J.; Rossmeisl, J.; Shao-Horn, Y.; Morgan, D. Prediction of solid oxide fuel cell cathode activity with first-principles descriptors. *Energy Environ. Sci.* **2011**, *4*, 3966−3970.
(119) Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **2018**, *120*, 145301.
(120) Chen, C.; Ye, W. K.; Zuo, Y. X.; Zheng, C.; Ong, S. P. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.* **2019**, *31*, 3564−3572.
(121) Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. P. Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. *J. Comput-Aided. Mol. Des.* **2005**, *19*, 693−703.
(122) Isayev, O.; Oses, C.; Toher, C.; Gossett, E.; Curtarolo, S.; Tropsha, A. Universal fragment descriptors for predicting properties of inorganic crystals. *Nat. Commun.* **2017**, *8*, 15679.
(123) Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *Npj Comput. Mater.* **2016**, *2*, 16028.
(124) Andersen, M.; Reuter, K. Adsorption Enthalpies for Catalysis Modeling through Machine-Learned Descriptors. *Acc. Chem. Res.* **2021**, *54*, 2741−2749.
(125) Xu, H.; Cheng, D.; Cao, D.; Zeng, X. C. A universal principle for a rational design of single-atom electrocatalysts. *Nat. Catal.* **2018**, *1*, 339−348.
(126) Ryan, K.; Lengyel, J.; Shatruk, M. Crystal Structure Prediction via Deep Learning. *J. Am. Chem. Soc.* **2018**, *140*, 10158−10168.
(127) Saxena, S.; Khan, T. S.; Jalid, F.; Ramteke, M.; Haider, M. A. In silico high throughput screening of bimetallic and single atom alloys using machine learning and ab initio microkinetic modelling. *J. Mater. Chem. A* **2020**, *8*, 107−123.
(128) Ulissi, Z. W.; Tang, M. T.; Xiao, J.; Liu, X.; Torelli, D. A.; Karamad, M.; Cummins, K.; Hahn, C.; Lewis, N. S.; Jaramillo, T. F.; et al. Machine-Learning Methods Enable Exhaustive Searches for Active Bimetallic Facets and Reveal Active Site Motifs for CO_2 Reduction. *ACS Catal.* **2017**, *7*, 6600−6608.
(129) Tran, K.; Ulissi, Z. W. Active learning across intermetallics to guide discovery of electrocatalysts for CO_2 reduction and H_2 evolution. *Nat. Catal.* **2018**, *1*, 696−703.
(130) Timoshenko, J.; Wrasman, C. J.; Luneau, M.; Shirman, T.; Cargnello, M.; Bare, S. R.; Aizenberg, J.; Friend, C. M.; Frenkel, A. I. Probing Atomic Distributions in Mono- and Bimetallic Nanoparticles by Supervised Machine Learning. *Nano Lett.* **2019**, *19*, 520−529.
(131) Hong, W. T.; Welsch, R. E.; Shao-Horn, Y. Descriptors of Oxygen-Evolution Activity for Oxides: A Statistical Evaluation. *J. Phys. Chem. C* **2016**, *120*, 78−86.
(132) Weng, B.; Song, Z.; Zhu, R.; Yan, Q.; Sun, Q.; Grice, C. G.; Yan, Y.; Yin, W.-J. Simple Descriptor Derived from Symbolic Regression Accelerating the Discovery of New Perovskite Catalysts. *Nat. Commun.* **2020**, *11*, 3513.
(133) Sun, Y.; Liao, H.; Wang, J.; Chen, B.; Sun, S.; Ong, S. J. H.; Xi, S.; Diao, C.; Du, Y.; Wang, J.-O.; et al. Covalency competition dominates the water oxidation structure-activity relationship on spinel oxides. *Nat. Catal.* **2020**, *3*, 554−563.
(134) Fung, V.; Hu, G.; Wu, Z.; Jiang, D.-e. Descriptors for Hydrogen Evolution on Single Atom Catalysts in Nitrogen-Doped Graphene. *J. Phys. Chem. C* **2020**, *124*, 19571−19578.
(135) Sun, M.; Dougherty, A. W.; Huang, B.; Li, Y.; Yan, C.-H. Accelerating Atomic Catalyst Discovery by Theoretical Calculations-Machine Learning Strategy. *Adv. Energy Mater.* **2020**, *10*, 1903949.
(136) Liu, X.; Zheng, L.; Han, C.; Zong, H.; Yang, G.; Lin, S.; Kumar, A.; Jadhav, A. R.; Tran, N. Q.; Hwang, Y.; et al. Identifying the Activity Origin of a Cobalt Single-Atom Catalyst for Hydrogen Evolution Using Supervised Learning. *Adv. Funct. Mater.* **2021**, *31*, 2100547.
(137) Lin, S.; Xu, H.; Wang, Y.; Zeng, X. C.; Chen, Z. Directly predicting limiting potentials from easily obtainable physical properties of graphene-supported single-atom electrocatalysts by machine learning. *J. Mater. Chem. A* **2020**, *8*, 5663−5670.
(138) Niu, H.; Wan, X.; Wang, X.; Shao, C.; Robertson, J.; Zhang, Z.; Guo, Y. Single-Atom Rhodium on Defective g-C3N4: A Promising Bifunctional Oxygen Electrocatalyst. *ACS Sustain. Chem. Eng.* **2021**, *9*, 3590−3599.
(139) Karim, M. R.; Ferrandon, M.; Medina, S.; Sture, E.; Kariuki, N.; Myers, D. J.; Holby, E. F.; Zelenay, P.; Ahmed, T. Coupling High-Throughput Experiments and Regression Algorithms to Optimize PGM-Free ORR Electrocatalyst Synthesis. *ACS Appl. Energy Mater.* **2020**, *3*, 9083−9088.
(140) Zafari, M.; Kumar, D.; Umer, M.; Kim, K. S. Machine learning-based high throughput screening for nitrogen fixation on boron-doped single atom catalysts. *J. Mater. Chem. A* **2020**, *8*, 5209−5216.
(141) Ge, L.; Yuan, H.; Min, Y.; Li, L.; Chen, S.; Xu, L.; Goddard, W. A. Predicted Optimal Bifunctional Electrocatalysts for the Hydrogen Evolution Reaction and the Oxygen Evolution Reaction Using Chalcogenide Heterostructures Based on Machine Learning Analysis of in Silico Quantum Mechanics Based High Throughput Screening. *J. Phys. Chem. Lett.* **2020**, *11*, 869−876.
(142) Wang, X.; Wang, C.; Ci, S.; Ma, Y.; Liu, T.; Gao, L.; Qian, P.; Ji, C.; Su, Y. Accelerating 2D MXene catalyst discovery for the hydrogen evolution reaction by computer-driven workflow and an ensemble learning strategy. *J. Mater. Chem. A* **2020**, *8*, 23488−23497.
(143) Sun, X.; Zheng, J.; Gao, Y.; Qiu, C.; Yan, Y.; Yao, Z.; Deng, S.; Wang, J. Machine-learning-accelerated screening of hydrogen evolution catalysts in MBenes materials. *Appl. Surf. Sci.* **2020**, *526*, 146522.
(144) Bai, Y.; Wilbraham, L.; Slater, B. J.; Zwijnenburg, M. A.; Sprick, R. S.; Cooper, A. I. Accelerated Discovery of Organic Polymer Photocatalysts for Hydrogen Evolution from Water through the Integration of Experiment and Theory. *J. Am. Chem. Soc.* **2019**, *141*, 9063−9071.
(145) Wan, Y.; Ramirez, F.; Zhang, X.; Nguyen, T.-Q.; Bazan, G. C.; Lu, G. Data driven discovery of conjugated polyelectrolytes for optoelectronic and photocatalytic applications. *Npj Comput. Mater.* **2021**, *7*, 69.
(146) Tao, Q.; Lu, T.; Sheng, Y.; Li, L.; Lu, W.; Li, M. Machine learning aided design of perovskite oxide materials for photocatalytic water splitting. *J. Energy Chem.* **2021**, *60*, 351−359.
(147) Fathinia, M.; Khataee, A.; Aber, S.; Naseri, A. Development of kinetic models for photocatalytic ozonation of phenazopyridine on TiO_2 nanoparticles thin film in a mixed semi-batch photoreactor. *Appl. Catal. B Environ.* **2016**, *184*, 270−284.

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

(148) Gao, B.; Sun, M.; Ding, Z.; Liu, W. Machine learning-optimized synthesis of doped TiO2 with improved photocatalytic performance: A multi-step workflow supported by designed wet-lab experiments. *J. Alloys Compd.* **2021**, *881*, 160534.
(149) O'Connor, N. J.; Jonayat, A. S. M.; Janik, M. J.; Senftle, T. P. Interaction trends between single metal atoms and oxide supports identified with density functional theory and statistical learning. *Nat. Catal.* **2018**, *1*, 531−539.
(150) Ouyang, R.; Curtarolo, S.; Ahmetcik, E.; Scheffler, M.; Ghiringhelli, L. M. SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Phys. Rev. Mater.* **2018**, *2*, 083802.
(151) Wexler, R. B.; Martirez, J. M. P.; Rappe, A. M. Chemical Pressure-Driven Enhancement of the Hydrogen Evolving Activity of Ni_2P from Nonmetal Surface Doping Interpreted via Machine Learning. *J. Am. Chem. Soc.* **2018**, *140*, 4678−4683.
(152) Panapitiya, G.; Avendano-Franco, G.; Ren, P. J.; Wen, X. D.; Li, Y. W.; Lewis, J. P. Machine-Learning Prediction of CO Adsorption in Thiolated, Ag-Alloyed Au Nanoclusters. *J. Am. Chem. Soc.* **2018**, *140*, 17508−17514.
(153) Wagner, N.; Rondinelli, J. M. Theory-guided Machine learning in Materials science. *Front. Mater.* **2016**, *3*. DOI: 10.3389/fmats.2016.00028.
(154) Song, F.; Guo, Z.; Mei, D. Feature Selection Using Principal Component Analysis. In *2010 International Conference on System Science, Engineering Design and Manufacturing Informatization*, Yichang, China, November 12−14, 2010; ICSEM, 2010; pp 27−30.
(155) Guo, Q.; Wu, W.; Massart, D. L.; Boucon, C.; de Jong, S. Feature selection in principal component analysis of analytical data. *Chemom. Intell. Lab. Syst.* **2002**, *61*, 123−132.
(156) Garcia-Muelas, R.; Lopez, N. Statistical learning goes beyond the d-band model providing the thermochemistry of adsorbates on transition metals. *Nat. Commun.* **2019**, *10*, 4687.
(157) Yu, Z.; Huang, W. Accelerating Optimizing the Design of Carbon-based Electrocatalyst Via Machine Learning. *Electroanalysis* **2021**, *33*, 599−607.
(158) Scholkopf, B.; Smola, A.; Muller, K.-R. Kernel Principal Component Analysis. *Adv. Kernel Method. - Support Vector Learning* **1999**, 327−333.
(159) Roweis, S. T.; Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. *Science* **2000**, *290*, 2323−2326.
(160) Tenenbaum, J. B.; de Silva, V.; Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science* **2000**, *290*, 2319−2323.
(161) van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.
(162) Jäger, M. O. J.; Ranawat, Y. S.; Canova, F. F.; Morooka, E. V.; Foster, A. S. Efficient Machine-Learning-Aided Screening of Hydrogen Adsorption on Bimetallic Nanoclusters. *ACS Comb. Sci.* **2020**, *22*, 768−781.
(163) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* **2020**, 1802.03426.
(164) Rosen, A. S.; Iyer, S. M.; Ray, D.; Yao, Z.; Aspuru-Guzik, A.; Gagliardi, L.; Notestein, J. M.; Snurr, R. Q. Machine learning the quantum-chemical properties of metal-organic frameworks for accelerated materials discovery. *Matter* **2021**, *4*, 1578−1597.
(165) Back, S.; Na, J.; Ulissi, Z. W. Efficient Discovery of Active, Selective, and Stable Catalysts for Electrochemical H2O2 Synthesis through Active Motif Screening. *ACS Catal.* **2021**, *11*, 2483−2491.
(166) Li, X.; Maffettone, P. M.; Che, Y.; Liu, T.; Chen, L.; Cooper, A. I. Combining machine learning and high-throughput experimentation to discover photocatalytically active organic molecules. *Chem. Sci.* **2021**, *12*, 10742−10754.
(167) Rodriguez-Nieva, J. F.; Scheurer, M. S. Identifying topological order through unsupervised machine learning. *Nat. Phys.* **2019**, *15*, 790−795.
(168) Xie, T.; France-Lanord, A.; Wang, Y. M.; Shao-Horn, Y.; Grossman, J. C. Graph dynamical networks for unsupervised learning of atomic scale dynamics in materials. *Nat. Commun.* **2019**, *10*, 2667.
(169) Libbrecht, M. W.; Noble, W. S. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **2015**, *16*, 321−332.
(170) Kall, L.; Canterbury, J. D.; Weston, J.; Noble, W. S.; MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Meth.* **2007**, *4*, 923−925.
(171) Huo, H. Y.; Rong, Z. Q.; Kononova, O.; Sun, W. H.; Botari, T.; He, T. J.; Tshitoyan, V.; Ceder, G. Semi-supervised machine-learning classification of materials synthesis procedures. *Npj Comput. Mater.* **2019**, *5*, 62.
(172) Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine learning in materials informatics: recent applications and prospects. *Npj Comput. Mater.* **2017**, *3*, 54.
(173) Katritzky, A. R.; Kuanar, M.; Slavov, S.; Hall, C. D.; Karelson, M.; Kahn, I.; Dobchev, D. A. Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction. *Chem. Rev.* **2010**, *110*, 5714−5789.
(174) Bishop, C. *Pattern Recognition and Machine Learning*; Springer-Verlag: New York: 2006.
(175) Brereton, R. G. *Applied Chemometrics for Scientists*; John Wiley & Sons, Ltd.: Chichester, U.K., 2007.
(176) Ghiringhelli, L. M.; Vybiral, J.; Levchenko, S. V.; Draxl, C.; Scheffler, M. Big Data of Materials Science: Critical Role of the Descriptor. *Phys. Rev. Lett.* **2015**, *114*, 105503.
(177) Simon-Vidal, L.; Garcia-Calvo, O.; Oteo, U.; Arrasate, S.; Lete, E.; Sotomayor, N.; Gonzalez-Diaz, H. Perturbation-Theory and Machine Learning (PTML) Model for High-Throughput Screening of Parham Reactions: Experimental and Theoretical Studies. *J. Chem. Inf. Model.* **2018**, *58*, 1384−1396.
(178) Dhayalan, V.; Gadekar, S. C.; Alassad, Z.; Milo, A. Unravelling mechanistic features of organocatalysis with in situ modifications at the secondary sphere. *Nat. Chem.* **2019**, *11*, 543−551.
(179) Singh, S.; Pareek, M.; Changotra, A.; Banerjee, S.; Bhaskararao, B.; Balamurugan, P.; Sunoj, R. B. A unified machine-learning protocol for asymmetric catalysis as a proof of concept demonstration using asymmetric hydrogenation. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117*, 1339−1345.
(180) Hoerl, A. E.; Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **1970**, *12*, 55−67.
(181) Ge, L.; Xu, W.; Chen, C.; Tang, C.; Xu, L.; Chen, Z. Rational Prediction of Single Metal Atom Supported on Two-Dimensional Metal Diborides for Electrocatalytic N2 Reduction Reaction with Integrated Descriptor. *J. Phys. Chem. Lett.* **2020**, *11*, 5241−5247.
(182) Bzdok, D.; Altman, N.; Krzywinski, M. Statistics versus machine learning. *Nat. Meth.* **2018**, *15*, 233−234.
(183) Tikhonov, A. N.; Arsenin, V. Y. *Solutions of Ill-Posed Problems*; Winston: Washington, DC, 1977.
(184) Tipping, M. E. Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* **2001**, *1*, 211−244.
(185) Burden, F. R.; Winkler, D. A. Relevance Vector Machines: Sparse Classification Methods for QSAR. *J. Chem. Inf. Model.* **2015**, *55*, 1529−1534.
(186) Support vector machine. https://en.wikipedia.org/wiki/Support_vector_machine.
(187) Landrum, G. A.; Penzotti, J. E.; Putta, S. Machine-learning models for combinatorial catalyst discovery. *Meas. Sci. Technol.* **2005**, *16*, 270−277.
(188) Anbari, E.; Adib, H.; Iranshahi, D. Experimental investigation and development of a SVM model for hydrogenation reaction of carbon monoxide in presence of Co-Mo/Al_2O_3 catalyst. *Chem. Eng. J.* **2015**, *276*, 213−221.
(189) Rajan, A. C.; Mishra, A.; Satsangi, S.; Vaish, R.; Mizuseki, H.; Lee, K. R.; Singh, A. K. Machine-Learning-Assisted Accurate Band Gap Predictions of Functionalized MXene. *Chem. Mater.* **2018**, *30*, 4031−4038.

(190) Mageed, A. K. Modeling photocatalytic hydrogen production from ethanol over copper oxide nanoparticles: a comparative analysis of various machine learning techniques. *Biomass Convers. Biorefinery* **2021**, s13399-021-01388-y.
(191) Baghban, A.; Habibzadeh, S.; Zokaee Ashtiani, F. Bandgaps of noble and transition metal/ZIF-8 electro/catalysts: a computational study. *RSC Adv.* **2020**, *10*, 22929−22938.
(192) Medasani, B.; Gamst, A.; Ding, H.; Chen, W.; Persson, K. A.; Asta, M.; Canning, A.; Haranczyk, M. Predicting defect behavior in B2 intermetallics by merging ab initio modeling and machine learning. *Npj Comput. Mater.* **2016**, *2*, 1.
(193) Balfer, J.; Bajorath, J.; Vogt, M. Compound Classification Using the scikit-learn Library. In *Tutorials in Chemoinformatics*; Varnek, A., Ed.; John Wiley & Sons Ltd., 2017; pp 223−239.
(194) Oosthuizen, G. D. Machine learning: a mathematical framework for neural network, symbolic and genetics-based learning. In *Proceedings of the Third International Conference on Genetic Algorithms*; Schaffer, J. D., Ed.; Morgan Kaufmann Publishers Inc.: San Francisco, CA, 1989; pp 385−390.
(195) Behler, J. Four Generations of High-Dimensional Neural Network Potentials. *Chem. Rev.* **2021**, *121*, 10037−10072.
(196) Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85−117.
(197) Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Netw.* **1991**, *4*, 251−257.
(198) Hoar, B. B.; Lu, S.; Liu, C. Machine-Learning-Enabled Exploration of Morphology Influence on Wire-Array Electrodes for Electrochemical Nitrogen Fixation. *J. Phys. Chem. Lett.* **2020**, *11*, 4625−4630.
(199) Guo, Z.; Liu, P.; Liu, J.; Du, F.; Jiang, L. Neural Network Inspired Design of Highly Active and Durable N-Doped Carbon Interconnected Molybdenum Phosphide for Hydrogen Evolution Reaction. *ACS Appl. Energy Mater.* **2018**, *1*, 5437−5445.
(200) Lee, J.; Jinnouchi, R. Machine Learning-Based Screening of Highly Stable and Active Ternary Pt Alloys for Oxygen Reduction Reaction. *J. Phys. Chem. C* **2021**, *125*, 16963−16974.
(201) Bartok, A. P.; Kermode, J.; Bernstein, N.; Csanyi, G. Machine Learning a General-Purpose Interatomic Potential for Silicon. *Phys. Rev. X* **2018**, *8*, 041048.
(202) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **2017**, *8*, 3192−3203.
(203) Cheng, D.; Zhao, Z.-J.; Zhang, G.; Yang, P.; Li, L.; Gao, H.; Liu, S.; Chang, X.; Chen, S.; Wang, T.; et al. The nature of active sites for carbon dioxide electroreduction over oxide-derived copper catalysts. *Nat. Commun.* **2021**, *12*, 395.
(204) Medford, A. J.; Kunz, M. R.; Ewing, S. M.; Borders, T.; Fushimi, R. Extracting Knowledge from Data through Catalysis Informatics. *ACS Catal.* **2018**, *8*, 7403−7429.
(205) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604−610.
(206) Walters, W. P.; Barzilay, R. Applications of Deep Learning in Molecule Generation and Molecular Property Prediction. *Acc. Chem. Res.* **2021**, *54*, 263−270.
(207) Mennel, L.; Symonowicz, J.; Wachter, S.; Polyushkin, D. K.; Molina-Mendoza, A. J.; Mueller, T. Ultrafast machine vision with 2D material neural network image sensors. *Nature* **2020**, *579*, 62−66.
(208) Bengio, Y.; Courville, A.; Vincent, P. Representation Learning: A Review and New Perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798−1828.
(209) Shrestha, A.; Mahmood, A. Review of Deep Learning Algorithms and Architectures. *IEEE Access* **2019**, *7*, 53040−53065.
(210) Liu, Q.; Fang, L.; Yu, G.; Wang, D.; Xiao, C.-L.; Wang, K. Detection of DNA base modifications by deep recurrent neural network on Oxford Nanopore sequencing data. *Nat. Commun.* **2019**, *10*, 2449.
(211) Längkvist, M.; Karlsson, L.; Loutfi, A. A review of unsupervised feature learning and deep learning for time-series modeling. *Pattern Recognit. Lett.* **2014**, *42*, 11−24.
(212) Jurtz, V. I.; Johansen, A. R.; Nielsen, M.; Almagro Armenteros, J. J.; Nielsen, H.; Sønderby, C. K.; Winther, O.; Sønderby, S. K. An introduction to deep learning on biological sequence data: examples and solutions. *Bioinformatics* **2017**, *33*, 3685−3690.
(213) Kermany, D. S.; Goldbaum, M.; Cai, W. J.; Valentim, C. C. S.; Liang, H. Y.; Baxter, S. L.; McKeown, A.; Yang, G.; Wu, X. K.; Yan, F. B.; et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell* **2018**, *172*, 1122−1131.
(214) Hosny, A.; Parmar, C.; Quackenbush, J.; Schwartz, L. H.; Aerts, H. J. W. L. Artificial intelligence in radiology. *Nat. Rev. Cancer* **2018**, *18*, 500−510.
(215) Esteva, A.; Kuprel, B.; Novoa, R. A.; Ko, J.; Swetter, S. M.; Blau, H. M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115−118.
(216) Shen, D.; Wu, G.; Suk, H.-I. Deep Learning in Medical Image Analysis. *Annu. Rev. Biomed. Eng.* **2017**, *19*, 221−248.
(217) Alom, M. Z.; Taha, T. M.; Yakopcic, C.; Westberg, S.; Sidike, P.; Nasrin, M. S.; Hasan, M.; Van Essen, B. C.; Awwal, A. A. S.; Asari, V. K. A State-of-the-Art Survey on Deep Learning Theory and Architectures. *Electronics* **2019**, *8*, 292.
(218) Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, June 7−12, 2015; IEEE: New York, 2015; pp 3156−3164.
(219) Makin, J. G.; Moses, D. A.; Chang, E. F. Machine translation of cortical activity to text with an encoder-decoder framework. *Nat. Neurosci.* **2020**, *23*, 575−582.
(220) Ziletti, A.; Kumar, D.; Scheffler, M.; Ghiringhelli, L. M. Insightful classification of crystal structures using deep learning. *Nat. Commun.* **2018**, *9*, 2775.
(221) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning Representations by Back-Propagating Errors. *Nature* **1986**, *323*, 533−536.
(222) Burden, F.; Winkler, D. Bayesian Regularization of Neural Networks. In *Artificial Neural Networks: Methods and Applications*; Livingstone, D. J., Ed.; Humana Press: Totowa, NJ, 2009; pp 23−42.
(223) Burden, F. R.; Winkler, D. A. Robust QSAR Models Using Bayesian Regularized Neural Networks. *J. Med. Chem.* **1999**, *42*, 3183−3187.
(224) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436−444.
(225) Winkler, D. A.; Le, T. C. Performance of Deep and Shallow Neural Networks, the Universal Approximation Theorem, Activity Cliffs, and QSAR. *Mol. Inform.* **2017**, *36*, 1600118.
(226) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **2017**, *8*, 13890.
(227) Jha, D.; Choudhary, K.; Tavazza, F.; Liao, W.-k.; Choudhary, A.; Campbell, C.; Agrawal, A. Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning. *Nat. Commun.* **2019**, *10*, 5316.
(228) Kim, B.; Lee, S.; Kim, J. Inverse design of porous materials using artificial neural networks. *Sci. Adv.* **2020**, *6*, No. eaax9324.
(229) Wang, S.-H.; Pillai, H. S.; Wang, S.; Achenie, L. E. K.; Xin, H. Infusing theory into deep learning for interpretable reactivity prediction. *Nat. Commun.* **2021**, *12*, 5288.
(230) Ding, R.; Chen, Y.; Chen, P.; Wang, R.; Wang, J.; Ding, Y.; Yin, W.; Liu, Y.; Li, J.; Liu, J. Machine Learning-Guided Discovery of Underlying Decisive Factors and New Mechanisms for the Design of Nonprecious Metal Electrocatalysts. *ACS Catal.* **2021**, *11*, 9798−9808.
(231) Yang, Q.; Xu, R.; Wu, P.; He, J.; Liu, C.; Jiang, W. Three-step treatment of real complex, variable high-COD rolling wastewater by rational adjustment of acidification, adsorption, and photocatalysis using big data analysis. *Sep. Purif. Technol.* **2021**, *270*, 118865.

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

(232) Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278−2324.
(233) Aguiar, J. A.; Gong, M. L.; Unocic, R. R.; Tasdizen, T.; Miller, B. D. Decoding crystallography from high-resolution electron imaging and diffraction datasets with deep learning. *Sci. Adv.* **2019**, *5*, No. eaaw1949.
(234) Lee, J.-W.; Park, W. B.; Lee, J. H.; Singh, S. P.; Sohn, K.-S. A deep-learning technique for phase identification in multiphase inorganic compounds using synthetic XRD powder patterns. *Nat. Commun.* **2020**, *11*, 86.
(235) Back, S.; Yoon, J.; Tian, N.; Zhong, W.; Tran, K.; Ulissi, Z. W. Convolutional Neural Network of Atomic Surface Structures To Predict Binding Energies for High-Throughput Screening of Catalysts. *J. Phys. Chem. Lett.* **2019**, *10*, 4401−4408.
(236) Wang, Z.; Wang, Q.; Han, Y.; Ma, Y.; Zhao, H.; Nowak, A.; Li, J. Deep learning for ultra-fast and high precision screening of energy materials. *Energy Storage Mater.* **2021**, *39*, 45−53.
(237) Zhou, B.; Khosla, A.; Lapedriza, À.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* **2016**, 2921−2929.
(238) Samek, W.; Montavon, G.; Lapuschkin, S.; Anders, C. J.; Müller, K. R. Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. *Proc. IEEE* **2021**, *109*, 247−278.
(239) Ribeiro, M. T.; Singh, S.; Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: San Francisco, CA, 2016; pp 1135−1144.
(240) Zeiler, M. D.; Fergus, R. In *Visualizing and Understanding Convolutional Networks*; Springer International Publishing: Cham, 2014; pp 818−833.
(241) Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; Wattenberg, M. SmoothGrad: removing noise by adding noise. *arxiv* **2017**, 1706.03825.
(242) Sundararajan, M.; Taly, A.; Yan, Q. Q. Axiomatic Attribution for Deep Networks. In *34th International Conference on Machine Learning*; Precup, D., Teh, Y. W., Eds.; Sydney, Australia, August 06−11, 2017; JMLR, 2017.
(243) Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; Samek, W. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLoS One* **2015**, *10*, No. e0130140.
(244) Dietterich, T. G. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Mach. Learn.* **2000**, *40*, 139−157.
(245) Riley, P. Three pitfalls to avoid in machine learning. *Nature* **2019**, *572*, 27−29.
(246) Deng, C.; Su, Y.; Li, F.; Shen, W.; Chen, Z.; Tang, Q. Understanding activity origin for the oxygen reduction reaction on bi-atom catalysts by DFT studies and machine-learning. *J. Mater. Chem. A* **2020**, *8*, 24563−24571.
(247) Kronberg, R.; Lappalainen, H.; Laasonen, K. Hydrogen Adsorption on Defective Nitrogen-Doped Carbon Nanotubes Explained via Machine Learning Augmented DFT Calculations and Game-Theoretic Feature Attributions. *J. Phys. Chem. C* **2021**, *125*, 15918−15933.
(248) Ying, Y.; Fan, K.; Luo, X.; Qiao, J.; Huang, H. Unravelling the origin of bifunctional OER/ORR activity for single-atom catalysts supported on C2N by DFT and machine learning. *J. Mater. Chem. A* **2021**, *9*, 16860−16867.
(249) Freund, Y.; Schapire, R.; Abe, N. A short introduction to boosting. *J. Jap. Soc. Artificial Intel.* **1999**, *14*, 1612.
(250) Rätsch, G.; Onoda, T.; Müller, K. R. Soft Margins for AdaBoost. *Mach. Learn.* **2001**, *42*, 287−320.
(251) Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189−1232.
(252) Ahmad, Z.; Xie, T.; Maheshwari, C.; Grossman, J. C.; Viswanathan, V. Machine Learning Enabled Computational Screening of Inorganic Solid Electrolytes for Suppression of Dendrite Formation in Lithium Metal Anodes. *ACS Cent. Sci.* **2018**, *4*, 996−1006.
(253) Davies, D. W.; Butler, K. T.; Walsh, A. Data-Driven Discovery of Photoactive Quaternary Oxides Using First-Principles Machine Learning. *Chem. Mater.* **2019**, *31*, 7221−7230.
(254) Zhang, X.; Chen, A.; Chen, L.; Zhou, Z. 2D Materials Bridging Experiments and Computations for Electro/Photocatalysis. *Adv. Energy Mater.* **2021**, 2003841.
(255) Pankajakshan, P.; Sanyal, S.; de Noord, O. E.; Bhattacharya, I.; Bhattacharyya, A.; Waghmare, U. Machine Learning and Statistical Analysis for Materials Science: Stability and Transferability of Fingerprint Descriptors and Chemical Insights. *Chem. Mater.* **2017**, *29*, 4190−4201.
(256) Timoshenko, J.; Frenkel, A. I. Inverting" X-ray Absorption Spectra of Catalysts by Machine Learning in Search for Activity Descriptors. *ACS Catal.* **2019**, *9*, 10192−10211.
(257) Schlexer Lamoureux, P.; Winther, K. T.; Garrido Torres, J. A.; Streibel, V.; Zhao, M.; Bajdich, M.; Abild-Pedersen, F.; Bligaard, T. Machine Learning for Computational Heterogeneous Catalysis. *ChemCatChem.* **2019**, *11*, 3581−3601.
(258) Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K.-R. Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies. *J. Chem. Theory Comput.* **2013**, *9*, 3404−3419.
(259) Xu, J. H.; Zhang, X. G.; Li, Y. D. Kernel MSE algorithm: A unified framework for KFD, LS-SVM and KRR. *International Joint Conference on Neural Networks (IJCNN 01)*, Washington, DC, July 15−19; IEEE: New York, 2001; pp 1486−1491.
(260) scikit-learn.org Comparison of kernel ridge regression and SVR. https://scikit-learn.org/stable/auto_examples/miscellaneous/plot_kernel_ridge_regression.html.
(261) Pilania, G.; Whittle, K. R.; Jiang, C.; Grimes, R. W.; Stanek, C. R.; Sickafus, K. E.; Uberuaga, B. P. Using Machine Learning To Identify Factors That Govern Amorphization of Irradiated Pyrochlores. *Chem. Mater.* **2017**, *29*, 2574−2583.
(262) von Lilienfeld, O. A.; Ramakrishnan, R.; Rupp, M.; Knoll, A. Fourier series of atomic radial distribution functions: A molecular fingerprint for machine learning models of quantum chemical properties. *Int. J. Quantum Chem.* **2015**, *115*, 1084−1093.
(263) Noh, J.; Back, S.; Kim, J.; Jung, Y. Active learning with non-ab initio input features toward efficient CO_2 reduction catalysts. *Chem. Sci.* **2018**, *9*, 5152−5159.
(264) Le, T. C.; Winkler, D. A. Discovery and Optimization of Materials Using Evolutionary Approaches. *Chem. Rev.* **2016**, *116*, 6107−6132.
(265) Hinton, G.; Nowlan, S. How Learning Can Guide Evolution. *Complex Systems* **1987**, *1*, 495−502.
(266) Smith, J. M. When learning guides evolution. *Nature* **1987**, *329*, 761−762.
(267) Umasankar, Y.; Ting, T.-W.; Chen, S.-M. Characterization of Poly(brilliant cresyl blue)-Multiwall Carbon Nanotube Composite Film and Its Application in Electrocatalysis of Vitamin B9 Reduction. *J. Electrochem. Soc.* **2011**, *158*, K117.
(268) Vilhelmsen, L. B.; Walton, K. S.; Sholl, D. S. Structure and Mobility of Metal Clusters in MOFs: Au, Pd, and AuPd Clusters in MOF-74. *J. Am. Chem. Soc.* **2012**, *134*, 12807−12816.
(269) Chung, Y. G.; Gómez-Gualdrón, D. A.; Li, P.; Leperi, K. T.; Deria, P.; Zhang, H.; Vermeulen, N. A.; Stoddart, J. F.; You, F.; Hupp, J. T.; et al. In silico discovery of metal-organic frameworks for precombustion CO_2 capture using a genetic algorithm. *Sci. Adv.* **2016**, *2*, No. e1600909.
(270) Sun, G.; Sautet, P. Metastable Structures in Cluster Catalysis from First-Principles: Structural Ensemble in Reaction Conditions and Metastability Triggered Reactivity. *J. Am. Chem. Soc.* **2018**, *140*, 2812−2820.

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

(271) Chang, T.; Lu, J.; Shen, Z.; Huang, Y.; Lu, D.; Wang, X.; Cao, J.; Morent, R. Simulation and optimization of the post plasma-catalytic system for toluene degradation by a hybrid ANN and NSGA-II method. *Appl. Catal. B Environ.* **2019**, *244*, 107−119.
(272) Talwar, S.; Verma, A. K.; Sangal, V. K. Modeling and optimization of fixed mode dual effect (photocatalysis and photo-Fenton) assisted Metronidazole degradation using ANN coupled with genetic algorithm. *J. Environ. Manag.* **2019**, *250*, 109428.
(273) Karimi Estahbanati, M. R.; Feilizadeh, M.; Babin, A.; Mei, B.; Mul, G.; Iliuta, M. C. Selective photocatalytic oxidation of cyclohexanol to cyclohexanone: A spectroscopic and kinetic study. *Chem. Eng. J.* **2020**, *382*, 122732.
(274) Zhang, Z.; Wang, Y.-G. Molecular Design of Dispersed Nickel Phthalocyanine@Nanocarbon Hybrid Catalyst for Active and Stable Electroreduction of CO2. *J. Phys. Chem. C* **2021**, *125*, 13836−13849.
(275) He, H.; Garcia, E. A. Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263−1284.
(276) Fronzi, M.; Isayev, O.; Winkler, D. A.; Shapter, J. G.; Ellis, A. V.; Sherrell, P. C.; Shepelin, N. A.; Corletto, A.; Ford, M. J. Active Learning in Bayesian Neural Networks for Bandgap Predictions of Novel Van der Waals Heterostructures. *Adv. Intell. Syst.* **2021**, *3*, 2100080.
(277) Kurilovich, A. A.; Alexander, C. T.; Pazhetnov, E. M.; Stevenson, K. J. Active learning-based framework for optimal reaction mechanism selection from microkinetic modeling: a case study of electrocatalytic oxygen reduction reaction on carbon nanotubes. *Phys. Chem. Chem. Phys.* **2020**, *22*, 4581−4591.
(278) Li, X.; Chiong, R.; Hu, Z.; Page, A. J. Low-Cost Pt Alloys for Heterogeneous Catalysis Predicted by Density Functional Theory and Active Learning. *J. Phys. Chem. Lett.* **2021**, *12*, 7305−7311.
(279) Golbraikh, A.; Tropsha, A. Beware of q2! *J. Mol. Graph. Model.* **2002**, *20*, 269−276.
(280) Alexander, D. L. J.; Tropsha, A.; Winkler, D. A. Beware of R^2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *J. Chem. Inf. Model.* **2015**, *55*, 1316−1322.
(281) Lipovetsky, S. Pareto 80/20 law: derivation via random partitioning. *International Journal of Mathematical Education in Science and Technology* **2009**, *40*, 271−277.
(282) Konovalov, D. A.; Llewellyn, L. E.; Vander Heyden, Y.; Coomans, D. Robust Cross-Validation of Linear Regression QSAR Models. *J. Chem. Inf. Model.* **2008**, *48*, 2081−2094.
(283) Wang, S.; Zhang, Z.; Dai, S.; Jiang, D.-e. Insights into CO_2/N_2 Selectivity in Porous Carbons from Deep Learning. *ACS Mater. Lett.* **2019**, *1*, 558−563.
(284) Dondapati, J. S.; Chen, A. Quantitative structure-property relationship of the photoelectrochemical oxidation of phenolic pollutants at modified nanoporous titanium oxide using supervised machine learning. *Phys. Chem. Chem. Phys.* **2020**, *22*, 8878−8888.
(285) Bruix, A.; Margraf, J. T.; Andersen, M.; Reuter, K. First-principles-based Multiscale Modelling of Heterogeneous Catalysis. *Nat. Catal.* **2019**, *2*, 659−670.
(286) Foscato, M.; Jensen, V. R. Automated in Silico Design of Homogeneous Catalysts. *ACS Catal.* **2020**, *10*, 2354−2377.
(287) Gu, G. H.; Choi, C.; Lee, Y.; Situmorang, A. B.; Noh, J.; Kim, Y.-H.; Jung, Y. Progress in Computational and Machine-Learning Methods for Heterogeneous Small-Molecule Activation. *Adv. Mater.* **2020**, *32*, 1907865.
(288) Erdem Günay, M.; Yıldırım, R. Recent advances in knowledge discovery for heterogeneous catalysis using machine learning. *Catal. Rev.* **2021**, *63*, 120−164.
(289) Ooka, H.; Wintzer, M. E.; Nakamura, R. Non-Zero Binding Enhances Kinetics of Catalysis: Machine Learning Analysis on the Experimental Hydrogen Binding Energy of Platinum. *ACS Catal.* **2021**, *11*, 6298−6303.
(290) Gamler, J. T. L.; Ashberry, H. M.; Skrabalak, S. E.; Koczkur, K. M. Random Alloyed versus Intermetallic Nanoparticles: A Comparison of Electrocatalytic Performance. *Adv. Mater.* **2018**, *30*, 1801563.
(291) Rößner, L.; Armbrüster, M. Electrochemical Energy Conversion on Intermetallic Compounds: A Review. *ACS Catal.* **2019**, *9*, 2018−2062.
(292) Liang, J.; Ma, F.; Hwang, S.; Wang, X.; Sokolowski, J.; Li, Q.; Wu, G.; Su, D. Atomic Arrangement Engineering of Metallic Nanocrystals for Energy-Conversion Electrocatalysis. *Joule* **2019**, *3*, 956−991.
(293) Li, J.; Sun, S. Intermetallic Nanoparticles: Synthetic Control and Their Enhanced Electrocatalysis. *Acc. Chem. Res.* **2019**, *52*, 2015−2025.
(294) Zhou, M.; Li, C.; Fang, J. Noble-Metal Based Random Alloy and Intermetallic Nanocrystals: Syntheses and Applications. *Chem. Rev.* **2021**, *121*, 736−795.
(295) Liu, B.; Zhou, M. X.; Chan, M. K. Y.; Greeley, J. P. Understanding Polyol Decomposition on Bimetallic Pt-Mo Catalysts-A DFT Study of Glycerol. *ACS Catal.* **2015**, *5*, 4942−4950.
(296) Fournier, R.; Mohareb, A. Optimizing molecular properties using a relative index of thermodynamic stability and global optimization techniques. *J. Chem. Phys.* **2016**, *144*, 024114.
(297) Oliynyk, A. O.; Mar, A. Discovery of Intermetallic Compounds from Traditional to Machine-Learning Approaches. *Acc. Chem. Res.* **2018**, *51*, 59−68.
(298) Vojvodic, A.; Nørskov, J. K. New design paradigm for heterogeneous catalysts. *Natl. Sci. Rev.* **2015**, *2*, 140−143.
(299) Green, I. X.; Tang, W.; Neurock, M.; Yates, J. T. Insights into Catalytic Oxidation at the Au/TiO_2 Dual Perimeter Sites. *Acc. Chem. Res.* **2014**, *47*, 805−815.
(300) Li, Z.; Wang, S.; Chin, W. S.; Achenie, L. E.; Xin, H. High-throughput screening of bimetallic catalysts enabled by machine learning. *J. Mater. Chem. A* **2017**, *5*, 24131−24138.
(301) Toyao, T.; Suzuki, K.; Kikuchi, S.; Takakusagi, S.; Shimizu, K.-i.; Takigawa, I. Toward Effective Utilization of Methane: Machine Learning Prediction of Adsorption Energies on Metal Alloys. *J. Phys. Chem. C* **2018**, *122*, 8315−8326.
(302) Zhong, M.; Tran, K.; Min, Y.; Wang, C.; Wang, Z.; Dinh, C.-T.; De Luna, P.; Yu, Z.; Rasouli, A. S.; Brodersen, P.; et al. Accelerated discovery of CO2 electrocatalysts using active machine learning. *Nature* **2020**, *581*, 178−183.
(303) Kim, M.; Yeo, B. C.; Park, Y.; Lee, H. M.; Han, S. S.; Kim, D. Artificial Intelligence to Accelerate the Discovery of N2 Electroreduction Catalysts. *Chem. Mater.* **2020**, *32*, 709−720.
(304) Zhang, Y.; Zuo, T. T.; Tang, Z.; Gao, M. C.; Dahmen, K. A.; Liaw, P. K.; Lu, Z. P. Microstructures and properties of high-entropy alloys. *Prog. Mater. Sci.* **2014**, *61*, 1−93.
(305) George, E. P.; Raabe, D.; Ritchie, R. O. High-entropy alloys. *Nat. Rev. Mater.* **2019**, *4*, 515−534.
(306) Huang, L.; Zaman, S.; Tian, X.; Wang, Z.; Fang, W.; Xia, B. Y. Advanced Platinum-Based Oxygen Reduction Electrocatalysts for Fuel Cells. *Acc. Chem. Res.* **2021**, *54*, 311−322.
(307) Tomboc, G. M.; Kwon, T.; Joo, J.; Lee, K. High entropy alloy electrocatalysts: a critical assessment of fabrication and performance. *J. Mater. Chem. A* **2020**, *8*, 14844−14862.
(308) Pedersen, J. K.; Batchelor, T. A. A.; Bagger, A.; Rossmeisl, J. High-Entropy Alloys as Catalysts for the CO_2 and CO Reduction Reactions. *ACS Catal.* **2020**, *10*, 2169−2176.
(309) Batchelor, T. A. A.; Pedersen, J. K.; Winther, S. H.; Castelli, I. E.; Jacobsen, K. W.; Rossmeisl, J. High-Entropy Alloys as a Discovery Platform for Electrocatalysis. *Joule* **2019**, *3*, 834−845.
(310) Lu, Z.; Chen, Z. W.; Singh, C. V. Neural Network-Assisted Development of High-Entropy Alloy Catalysts: Decoupling Ligand and Coordination Effects. *Matter* **2020**, *3*, 1318−1333.
(311) Rück, M.; Garlyyev, B.; Mayr, F.; Bandarenka, A. S.; Gagliardi, A. Oxygen Reduction Activities of Strained Platinum Core-Shell Electrocatalysts Predicted by Machine Learning. *J. Phys. Chem. Lett.* **2020**, *11*, 1773−1780.
(312) Xiang, C.; Suram, S. K.; Haber, J. A.; Guevarra, D. W.; Soedarmadji, E.; Jin, J.; Gregoire, J. M. High-Throughput Bubble Screening Method for Combinatorial Discovery of Electrocatalysts for Water Splitting. *ACS Comb. Sci.* **2014**, *16*, 47−52.

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

(313) Hong, W. T.; Risch, M.; Stoerzinger, K. A.; Grimaud, A.; Suntivich, J.; Shao-Horn, Y. Toward the Rational Design of Non-precious Transition Metal Oxides for Oxygen Electrocatalysis. *Energy Environ. Sci.* **2015**, *8*, 1404−1427.
(314) Zhang, C.; Fagan, R. D.; Smith, R. D. L.; Moore, S. A.; Berlinguette, C. P.; Trudel, S. Mapping the Performance of Amorphous Ternary Metal Oxide Water Oxidation Catalysts Containing Aluminium. *J. Mater. Chem. A* **2015**, *3*, 756−761.
(315) Gawande, M. B.; Goswami, A.; Felpin, F.-X.; Asefa, T.; Huang, X.; Silva, R.; Zou, X.; Zboril, R.; Varma, R. S. Cu and Cu-Based Nanoparticles: Synthesis and Applications in Catalysis. *Chem. Rev.* **2016**, *116*, 3722−3811.
(316) Hwang, J.; Rao, R. R.; Giordano, L.; Katayama, Y.; Yu, Y.; Shao-Horn, Y. Perovskites in catalysis and electrocatalysis. *Science* **2017**, *358*, 751−756.
(317) Li, M.; Garg, S.; Chang, X.; Ge, L.; Li, L.; Konarova, M.; Rufford, T. E.; Rudolph, V.; Wang, G. Toward Excellence of Transition Metal-Based Catalysts for CO_2 Electrochemical Reduction: An Overview of Strategies and Rationales. *Small Methods* **2020**, *4*, 2000033.
(318) Xu, W.; Andersen, M.; Reuter, K. Data-Driven Descriptor Engineering and Refined Scaling Relations for Predicting Transition Metal Oxide Reactivity. *ACS Catal.* **2021**, *11*, 734−742.
(319) Yu, N.; Lu, X.; Song, F.; Yao, Y.; Han, E. Electrocatalytic Degradation of Sulfamethazine on IrO_2-RuO_2 Composite Electrodes: Influencing Factors, Kinetics and Modeling. *J. Environ. Chem. Eng.* **2021**, *9*, 105301.
(320) Yang, X.-F.; Wang, A.; Qiao, B.; Li, J.; Liu, J.; Zhang, T. Single-Atom Catalysts: A New Frontier in Heterogeneous Catalysis. *Acc. Chem. Res.* **2013**, *46*, 1740−1748.
(321) Zhu, C.; Fu, S.; Shi, Q.; Du, D.; Lin, Y. Single-Atom Electrocatalysts. *Angew. Chem., Int. Ed.* **2017**, *56*, 13944−13960.
(322) Peng, Y.; Lu, B.; Chen, S. Carbon-Supported Single Atom Catalysts for Electrochemical Energy Conversion and Storage. *Adv. Mater.* **2018**, *30*, 1801995.
(323) Wu, L.; Guo, T.; Li, T. Machine learning-accelerated prediction of overpotential of oxygen evolution reaction of single-atom catalysts. *iScience* **2021**, *24*, 102398.
(324) Anasori, B.; Lukatskaya, M. R.; Gogotsi, Y. 2D metal carbides and nitrides (MXenes) for energy storage. *Nat. Rev. Mater.* **2017**, *2*, 16098.
(325) Zheng, J.; Sun, X.; Qiu, C.; Yan, Y.; Yao, Z.; Deng, S.; Zhong, X.; Zhuang, G.; Wei, Z.; Wang, J. High-Throughput Screening of Hydrogen Evolution Reaction Catalysts in MXene Materials. *J. Phys. Chem. C* **2020**, *124*, 13695−13705.
(326) Schultz, D. M.; Yoon, T. P. Solar Synthesis: Prospects in Visible Light Photocatalysis. *Science* **2014**, *343*, 1239176.
(327) Wang, W.; Tade, M. O.; Shao, Z. P. Research progress of perovskite materials in photocatalysis- and photovoltaics-related energy conversion and environmental treatment. *Chem. Soc. Rev.* **2015**, *44*, 5371−5408.
(328) Zhang, G.; Liu, G.; Wang, L. Z.; Irvine, J. T. S. Inorganic Perovskite Photocatalysts for Solar Energy Utilization. *Chem. Soc. Rev.* **2016**, *45*, 5951−5984.
(329) Hou, X.; Mu, L.; Chen, F.; Hu, X. Emerging Investigator Series: Design of Hydrogel Nanocomposites for the Detection and Removal of Pollutants: from Nanosheets, Network Structures, and Biocompatibility to Machine-Learning-Assisted Design. *Environ. Sci. Nano* **2018**, *5*, 2216−2240.
(330) Agarwal, A.; Goverapet Srinivasan, S.; Rai, B., Data-Driven Discovery of 2D Materials for Solar Water Splitting. *Front. Mater.* **2021**, *8*. DOI: 10.3389/fmats.2021.679269.
(331) Wang, X.; Maeda, K.; Thomas, A.; Takanabe, K.; Xin, G.; Carlsson, J. M.; Domen, K.; Antonietti, M. A metal-free polymeric photocatalyst for hydrogen production from water under visible light. *Nat. Mater.* **2009**, *8*, 76−80.
(332) Wang, Y.; Vogel, A.; Sachs, M.; Sprick, R. S.; Wilbraham, L.; Moniz, S. J. A.; Godin, R.; Zwijnenburg, M. A.; Durrant, J. R.; Cooper, A. I.; et al. Current understanding and challenges of solar-driven hydrogen generation using polymeric photocatalysts. *Nat. Energy* **2019**, *4*, 746−760.
(333) Rahman, M. Z.; Kibria, M. G.; Mullins, C. B. Metal-free photocatalysts for hydrogen evolution. *Chem. Soc. Rev.* **2020**, *49*, 1887−1931.
(334) Kosco, J.; Bidwell, M.; Cha, H.; Martin, T.; Howells, C. T.; Sachs, M.; Anjum, D. H.; Gonzalez Lopez, S.; Zou, L.; Wadsworth, A.; et al. Enhanced photocatalytic hydrogen evolution from organic semiconductor heterojunction nanoparticles. *Nat. Mater.* **2020**, *19*, 559−565.
(335) Calegari Andrade, M. F.; Ko, H.-Y.; Zhang, L.; Car, R.; Selloni, A. Free energy of proton transfer at the water-TiO_2 interface from ab initio deep potential molecular dynamics. *Chem. Sci.* **2020**, *11*, 2335−2341.
(336) Estahbanati, M. R. K.; Feilizadeh, M.; Iliuta, M. C. Photocatalytic valorization of glycerol to hydrogen: Optimization of operating parameters by artificial neural network. *Appl. Catal. B Environ.* **2017**, *209*, 483−492.
(337) Kaneko, M.; Fujii, M.; Hisatomi, T.; Yamashita, K.; Domen, K. Regression model for stabilization energies associated with anion ordering in perovskite-type oxynitrides. *J. Energy Chem.* **2019**, *36*, 7−14.
(338) Can, E.; Yildirim, R. Data mining in photocatalytic water splitting over perovskites literature for higher hydrogen production. *Appl. Catal. B Environ.* **2019**, *242*, 267−283.
(339) Mahmoud, M. S.; Mahmoud, A. S. Wastewater Treatment Using Nano Bimetallic Iron/Copper, Adsorption Isotherm, Kinetic Studies, and Artificial Intelligence Neural Networks. *Emerg. Mater.* **2021**, *4*, 1455−1463.
(340) Dehghani, M. H.; Hassani, A. H.; Karri, R. R.; Younesi, B.; Shayeghi, M.; Salari, M.; Zarei, A.; Yousefi, M.; Heidarinejad, Z. Process optimization and enhancement of pesticide adsorption by porous adsorbents by regression analysis and parametric modelling. *Sci. Rep-Uk* **2021**, *11*, 11719.
(341) Mohammed, N.; Palaniandy, P.; Shaik, F. Pollutants removal from saline water by solar photocatalysis: a review of experimental and theoretical approaches. *Int. J. Environ. Anal. Chem.* **2021**, 1−21.
(342) Tabatabai-Yazdi, F.-S.; Ebrahimian Pirbazari, A.; Esmaeili Khalil Saraei, F.; Gilani, N. Construction of Graphene Based Photocatalysts for Photocatalytic Degradation of Organic Pollutant and Modeling Using Artificial Intelligence Techniques. *Phys. B: Condens Matter* **2021**, *608*, 412869.
(343) Vasseghian, Y.; Berkani, M.; Almomani, F.; Dragoi, E.-N. Data mining for pesticide decontamination using heterogeneous photocatalytic processes. *Chemosphere* **2021**, *270*, 129449.
(344) Norouzi, M.; Karimian, A.; Dehghani, H.; Rezvan Leylan, S. A. Photocatalytic degradation of 2,4,6-trinitrotoluene (TNT) in the presence of ZnS, NiS and ZnS/NiS supported Clinoptilolite under UV irradiation: experimental and neural network modelling. *Int. J. Environ. Anal. Chem.* **2021**, 1−25.
(345) Caglar Gencosman, B.; Eker Sanli, G. Prediction of Polycyclic Aromatic Hydrocarbons (PAHs) Removal from Wastewater Treatment Sludge Using Machine Learning Methods. *Water, Air, Soil Pollut.* **2021**, *232*, 87.
(346) Ayodele, B. V.; Alsaffar, M. A.; Mustapa, S. I.; Cheng, C. K.; Witoon, T. Modeling the effect of process parameters on the photocatalytic degradation of organic pollutants using artificial neural networks. *Process Saf. Environ. Prot.* **2021**, *145*, 120−132.
(347) Pelalak, R.; Alizadeh, R.; Ghareshabani, E.; Heidari, Z. Degradation of sulfonamide antibiotics using ozone-based advanced oxidation process: Experimental, modeling, transformation mechanism and DFT study. *Sci. Total Environ.* **2020**, *734*, 139446.
(348) Jiang, Z.; Hu, J.; Zhang, X.; Zhao, Y.; Fan, X.; Zhong, S.; Zhang, H.; Yu, X. A generalized predictive model for TiO2-Catalyzed photo-degradation rate constants of water contaminants through artificial neural network. *Environ. Res.* **2020**, *187*, 109697.
(349) Pellegrino, F.; Isopescu, R.; Pellutiè, L.; Sordello, F.; Rossi, A. M.; Ortel, E.; Martra, G.; Hodoroaba, V.-D.; Maurino, V. Machine learning approach for elucidating and predicting the role of synthesis

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

parameters on the shape and size of TiO2 nanoparticles. *Sci. Rep-Uk* **2020**, *10*, 18910.
(350) Yang, Y.; Guan, C.; Chen, S. Structural Characterization and Catalytic Sterilization Performance of a TiO2 Nano-Photocatalyst. *Food Sci. Nutr.* **2020**, *8*, 3638−3646.
(351) Amani-Ghadim, A. R.; Dorraji, M. S. S. Modeling of photocatalyatic process on synthesized ZnO nanoparticles: Kinetic model development and artificial neural networks. *Appl. Catal. B Environ.* **2015**, *163*, 539−546.
(352) Mohammadzadeh Kakhki, R.; Mohammadpoor, M.; Faridi, R.; Bahadori, M. The development of an artificial neural network - genetic algorithm model (ANN-GA) for the adsorption and photocatalysis of methylene blue on a novel sulfur-nitrogen co-doped Fe2O3 nanostructure surface. *RSC Adv.* **2020**, *10*, 5951−5960.
(353) Smaali, A.; Berkani, M.; Merouane, F.; Le, V. T.; Vasseghian, Y.; Rahim, N.; Kouachi, M. Photocatalytic-persulfate- oxidation for diclofenac removal from aqueous solutions: Modeling, optimization and biotoxicity test assessment. *Chemosphere* **2021**, *266*, 129158.
(354) Baaloudj, O.; Nasrallah, N.; Kebir, M.; Guedioura, B.; Amrane, A.; Nguyen-Tri, P.; Nanda, S.; Assadi, A. A. Artificial neural Network Modeling of Cefixime Photodegradation by Synthesized $CoBi_2O_4$ Nanoparticles. *Environ. Sci. Pollut. Res.* **2021**, *28*, 15436−15452.
(355) Lin, C.-C.; Chang, C.-W.; Kaun, C.-C.; Su, Y.-H. Stepwise Evolution of Photocatalytic Spinel-Structured (Co,Cr,Fe,Mn,Ni)3O4 High Entropy Oxides from First-Principles Calculations to Machine Learning. *Crystals* **2021**, *11*, 1035.
(356) Low, J. X.; Yu, J. G.; Jaroniec, M.; Wageh, S.; Al-Ghamdi, A. A. Heterojunction Photocatalysts. *Adv. Mater.* **2017**, *29*, 1601694.
(357) Eskandarloo, H.; Badiei, A.; Behnajady, M. A. Study of the Effect of Additives on the Photocatalytic Degradation of a Triphenylmethane Dye in the Presence of Immobilized TiO_2/NiO Nanoparticles: Artificial Neural Network Modeling. *Ind. Eng. Chem. Res.* **2014**, *53*, 6881−6895.
(358) Kassahun, S. K.; Kiflie, Z.; Kim, H.; Baye, A. F. Process Optimization and Kinetics Analysis for Photocatalytic Degradation of Emerging Contaminant Using N-doped TiO_2-SiO_2 Nanoparticle: Artificial Neural Network and Surface Response Methodology Approach. *Environ. Technol. Innov.* **2021**, *23*, 101761.
(359) Gupta, B.; Gupta, A. K.; Tiwary, C. S.; Ghosal, P. S. A multivariate modeling and experimental realization of photocatalytic system of engineered S-C3N4/ZnO hybrid for ciprofloxacin removal: Influencing factors and degradation pathways. *Environ. Res.* **2021**, *196*, 110390.
(360) Sheydaei, M.; Ayoubi-Feiz, B.; Abbaszade-Fakhri, G. A visible-light active g-C3N4/Ce-ZnO/Ti nanocomposite for efficient photo-electrocatalytic pharmaceutical degradation: Modelling with artificial neural network. *Process Saf. Environ. Prot.* **2021**, *149*, 776−785.
(361) Ataei, A.; Mehrizad, A.; Zare, K. Photocatalytic degradation of cefazoline antibiotic using zeolite-supported CdS/CaFe2O4 Z-scheme photocatalyst: Optimization and modeling of process by RSM and ANN. *J. Mol. Liq.* **2021**, *328*, 115476.
(362) Grills, D. C.; Ertem, M. Z.; McKinnon, M.; Ngo, K. T.; Rochford, J. Mechanistic aspects of CO2 reduction catalysis with manganese-based molecular catalysts. *Coord. Chem. Rev.* **2018**, *374*, 173−217.
(363) Kim, W.; McClure, B. A.; Edri, E.; Frei, H. Coupling carbon dioxide reduction with water oxidation in nanoscale photocatalytic assemblies. *Chem. Soc. Rev.* **2016**, *45*, 3221−3243.
(364) Li, X.; Yu, J.; Jaroniec, M.; Chen, X. Cocatalysts for Selective Photoreduction of CO2 into Solar Fuels. *Chem. Rev.* **2019**, *119*, 3962−4179.
(365) Huynh, M. T.; Mora, S. J.; Villalba, M.; Tejeda-Ferrari, M. E.; Liddell, P. A.; Cherry, B. R.; Teillout, A.-L.; Machan, C. W.; Kubiak, C. P.; Gust, D.; et al. Concerted One-Electron Two-Proton Transfer Processes in Models Inspired by the Tyr-His Couple of Photosystem II. *ACS Cent. Sci.* **2017**, *3*, 372−380.
(366) Odella, E.; Wadsworth, B. L.; Mora, S. J.; Goings, J. J.; Huynh, M. T.; Gust, D.; Moore, T. A.; Moore, G. F.; Hammes-Schiffer, S.; Moore, A. L. Proton-Coupled Electron Transfer Drives Long-Range Proton Translocation in Bioinspired Systems. *J. Am. Chem. Soc.* **2019**, *141*, 14057−14061.
(367) Mora, S. J.; Odella, E.; Moore, G. F.; Gust, D.; Moore, T. A.; Moore, A. L. Proton-Coupled Electron Transfer in Artificial Photosynthetic Systems. *Acc. Chem. Res.* **2018**, *51*, 445−453.
(368) Goings, J. J.; Hammes-Schiffer, S. Nonequilibrium Dynamics of Proton-Coupled Electron Transfer in Proton Wires: Concerted but Asynchronous Mechanisms. *ACS Cent. Sci.* **2020**, *6*, 1594−1601.
(369) Schneider, J.; Bahnemann, D. W. Undesired Role of Sacrificial Reagents in Photocatalysis. *J. Phys. Chem. Lett.* **2013**, *4*, 3479−3483.
(370) Ruff, L.; Kauffmann, J. R.; Vandermeulen, R. A.; Montavon, G.; Samek, W.; Kloft, M.; Dietterich, T. G.; Müller, K. R. A Unifying Review of Deep and Shallow Anomaly Detection. *Proc. IEEE* **2021**, *109*, 756−795.
(371) Qian, L.; Winfree, E.; Bruck, J. Neural network computation with DNA strand displacement cascades. *Nature* **2011**, *475*, 368−372.
(372) Costantino, F.; Kamat, P. V. Do Sacrificial Donors Donate H2 in Photocatalysis? *ACS Energy Lett.* **2022**, *7*, 242−246.
(373) Das, R.; Chakraborty, S.; Peter, S. C. Systematic Assessment of Solvent Selection in Photocatalytic CO2 Reduction. *ACS Energy Lett.* **2021**, *6*, 3270−3274.
(374) Hao, Y.-C.; Guo, Y.; Chen, L.-W.; Shu, M.; Wang, X.-Y.; Bu, T.-A.; Gao, W.-Y.; Zhang, N.; Su, X.; Feng, X.; et al. Promoting Nitrogen Electroreduction to Ammonia with Bismuth Nanocrystals and Potassium Cations in Water. *Nat. Catal.* **2019**, *2*, 448−456.
(375) Mai, H.; Lu, T.; Li, Q.; Liu, Z.; Li, Y.; Kremer, F.; Li, L.; Withers, R. L.; Wen, H.; Liu, Y. Above-Band Gap Photoinduced Stabilization of Engineered Ferroelectric Domains. *ACS Appl. Mater. Interfaces* **2018**, *10*, 12781−12789.
(376) Di, J.; Xia, J.; Chisholm, M. F.; Zhong, J.; Chen, C.; Cao, X.; Dong, F.; Chi, Z.; Chen, H.; Weng, Y.-X.; et al. Defect-Tailoring Mediated Electron-Hole Separation in Single-Unit-Cell Bi3O4Br Nanosheets for Boosting Photocatalytic Hydrogen Evolution and Nitrogen Fixation. *Adv. Mater.* **2019**, *31*, 1807576.
(377) Selcuk, S.; Selloni, A. Facet-dependent trapping and dynamics of excess electrons at anatase TiO2 surfaces and aqueous interfaces. *Nat. Mater.* **2016**, *15*, 1107−1112.
(378) Pronobis, W.; Schütt, K. T.; Tkatchenko, A.; Müller, K.-R. Capturing Intensive and Extensive DFT/TDDFT Molecular Properties with Machine Learning. *Eur. Phys. J. B* **2018**, *91*, 178.
(379) Ramakrishnan, R.; Hartmann, M.; Tapavicza, E.; von Lilienfeld, O. A. Electronic Spectra from TDDFT and Machine Learning in Chemical Space. *J. Chem. Phys.* **2015**, *143*, 084111.
(380) Luedtke, A.; Carone, M.; Simon, N.; Sofrygin, O. Learning to learn from data: Using deep adversarial learning to construct optimal statistical procedures. *Sci. Adv.* **2020**, *6*, No. eaaw2140.
(381) Mai, H.; Le, T. C.; Hisatomi, T.; Chen, D.; Domen, K.; Winkler, D. A.; Caruso, R. A. Use of Meta Models for Rapid Discovery of Narrow Bandgap Oxide Photocatalysts. *iScience* **2021**, *24*, 103068.
(382) Olier, I.; Orhobor, O. I.; Dash, T.; Davis, A. M.; Soldatova, L. N.; Vanschoren, J.; King, R. D. Transformational machine learning: Learning how to learn from many related scientific problems. *Proc. Natl. Acad. Sci. U. S. A.* **2021**, *118*, No. e2108013118.
(383) Burden, F. R.; Winkler, D. A. Optimal Sparse Descriptor Selection for QSAR Using Bayesian Methods. *QSAR Comb. Sci.* **2009**, *28*, 645−653.
(384) Winkler, D. A. Role of Artificial Intelligence and Machine Learning in Nanosafety. *Small* **2020**, *16*, 2001883.
(385) Batra, R.; Dai, H.; Huan, T. D.; Chen, L.; Kim, C.; Gutekunst, W. R.; Song, L.; Ramprasad, R. Polymers for Extreme Conditions Designed Using Syntax-Directed Variational Autoencoders. *Chem. Mater.* **2020**, *32*, 10489−10500.
(386) Dai, H.; Tian, Y.; Dai, B.; Skiena, S.; Song, L. Syntax-Directed Variational Autoencoder for Structured Data. *arxiv* **2018**, 1802.08786.
(387) Yao, Z.; Sánchez-Lengeling, B.; Bobbitt, N. S.; Bucior, B. J.; Kumar, S. G. H.; Collins, S. P.; Burns, T.; Woo, T. K.; Farha, O. K.; Snurr, R. Q.; et al. Inverse Design of Nanoporous Crystalline

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX

Reticular Materials with Deep Generative Models. *Nat. Mach. Intell.* **2021**, *3*, 76−86.
(388) Kim, S.; Noh, J.; Gu, G. H.; Aspuru-Guzik, A.; Jung, Y. Generative Adversarial Networks for Crystal Structure Prediction. *ACS Cent. Sci.* **2020**, *6*, 1412−1420.
(389) Noh, J.; Kim, J.; Stein, H. S.; Sanchez-Lengeling, B.; Gregoire, J. M.; Aspuru-Guzik, A.; Jung, Y. Inverse Design of Solid-State Materials via a Continuous Representation. *Matter* **2019**, *1*, 1370−1384.
(390) Li, J.; Tu, Y.; Liu, R.; Lu, Y.; Zhu, X. Toward "On-Demand" Materials Synthesis and Scientific Discovery through Intelligent Robots. *Adv. Sci.* **2020**, *7*, 1901957.
(391) Stach, E.; DeCost, B.; Kusne, A. G.; Hattrick-Simpers, J.; Brown, K. A.; Reyes, K. G.; Schrier, J.; Billinge, S.; Buonassisi, T.; Foster, I.; et al. Autonomous experimentation systems for materials development: A community perspective. *Matter* **2021**, *4*, 2702−2726.
(392) Tao, H.; Wu, T.; Aldeghi, M.; Wu, T. C.; Aspuru-Guzik, A.; Kumacheva, E. Nanoparticle synthesis assisted by machine learning. *Nat. Rev. Mater.* **2021**, *6*, 701−716.
(393) Li, J.; Li, J.; Liu, R.; Tu, Y.; Li, Y.; Cheng, J.; He, T.; Zhu, X. Autonomous discovery of optically active chiral inorganic perovskite nanocrystals through an intelligent cloud lab. *Nat. Commun.* **2020**, *11*, 2046.
(394) Burger, B.; Maffettone, P. M.; Gusev, V. V.; Aitchison, C. M.; Bai, Y.; Wang, X.; Li, X.; Alston, B. M.; Li, B.; Clowes, R.; et al. A mobile robotic chemist. *Nature* **2020**, *583*, 237−241.
(395) Pan, Z.; Zhou, Y.; Zhang, L. Photoelectrochemical Properties, Machine Learning, and Symbolic Regression for Molecularly Engineered Halide Perovskite Materials in Water. *ACS Appl. Mater. Interfaces* **2022**, *14*, 9933−9943.
(396) Zheng, J.; Sun, X.; Hu, J.; Wang, S.; Yao, Z.; Deng, S.; Pan, X.; Pan, Z.; Wang, J. Symbolic Transformer Accelerating Machine Learning Screening of Hydrogen and Deuterium Evolution Reaction Catalysts in MA2Z4 Materials. *ACS Appl. Mater. Interfaces* **2021**, *13*, 50878−50891.
(397) Xie, J.; Zhang, L. Machine learning and symbolic regression for adsorption of atmospheric molecules on low-dimensional TiO2. *Appl. Surf. Sci.* **2022**, *597*, 153728.

https://doi.org/10.1021/acs.chemrev.2c00061
Chem. Rev. XXXX, XXX, XXX−XXX