Chem Soc Rev

Discovery, Synthesis, and Understanding of 2D Materials Enabled by Machine Learning

Journal:	*Chemical Society Reviews*
Manuscript ID	CS-SYN-05-2021-000503.R2
Article Type:	Tutorial Review
Date Submitted by the Author:	30-Jan-2022
Complete List of Authors:	Ryu, Byunghoon; Argonne National Laboratory Wang, Luqing; Northwestern University, Pu, Haihui; The University of Chicago Chan, Maria; Argonne National Laboratory, Center for Nanoscale Materials Chen, Junhong; The University of Chicago, Pritzker School of Molecular Engineering

Understanding, Discovery, and Synthesis of 2D Materials Enabled by Machine Learning

Byunghoon Ryu[1], Luqing Wang[2,3], Haihui Pu[1,4], Maria K. Y. Chan[2,5], and Junhong Chen[1,4,+]

[1]Chemical Sciences and Engineering Division, Physical Sciences and Engineering Directorate, Argonne National Laboratory, Lemont, Illinois 60439, USA.
[2]Center for Nanoscale Materials, Argonne National Laboratory, Lemont, IL 60439, USA.
[3]The Materials Research Center, Northwestern University, Evanston, Illinois 60208, USA.
[4]Pritzker School of Molecular Engineering, University of Chicago, Chicago, Illinois 60637, USA.
[5]Northwestern Argonne Institute of Science and Engineering, Evanston, Illinois 60208, USA.

Abstract

Machine learning (ML) is becoming an effective tool for studying 2D materials. Taking as input computed or experimental materials data, ML algorithms predict the structural, electronic, mechanical, and chemical properties of 2D materials that have yet to be discovered. Such predictions expand investigations on how to synthesize 2D materials and use them in various applications, as well as greatly reduce the time and cost to discover and understand 2D materials. This tutorial review focuses on the understanding, discovery, and synthesis of 2D materials enabled by or benefiting from various ML techniques. We introduce the most recent efforts to adopt ML in various fields of study regarding 2D materials and provide an outlook for future research opportunities. The adoption of ML is anticipated to accelerate and transform the study of 2D materials and their heterostructures.

Key learning points

1. Case studies for using ML toward the understanding, discovery, and synthesis of 2D materials.
2. The generation and gathering of training data for ML toward the discovery of 2D materials.
3. The key descriptors among a large number of characteristics of 2D materials for ML algorithms.
4. The usability of ML techniques in various fields of study related to 2D materials.
5. The applicability of ML techniques in future 2D materials research.

1. Introduction

Since the discovery of graphene, two-dimensional (2D) materials have been considered wonder materials that can lead to significant advancements in applications such as photovoltaics, semiconductors, catalysts, and sensors. Due to this great expectation, a large number of new 2D materials, such as transition metal dichalcogenides (TMDs), carbides/nitrides/carbonitrides (MXene), and borides (MBene), have been discovered and appended to the 2D materials family.[1] However, only the tip of the iceberg has been revealed. According to a recent study performed using density functional theory (DFT), there are nearly 2000 2D materials which may be exfoliated from their bulk-layered counterparts.[2] Furthermore, van der Waals heterostructures intentionally made up of combinations of stacked 2D materials significantly increase the total number of possible candidates in the 2D materials family.

Unfortunately, conventional experimental and computational approaches can scarcely keep up with the rapidly growing demands in the study of 2D materials. In addition, both experimental methods and computational simulations using first-principles calculations, such as high-

throughput DFT, require considerable time and cost, which slows progress in 2D materials study. In recent years, machine learning (ML) has become an effective tool for studying a wide variety of materials, and 2D materials are no exception. Using the structural, electrical, thermodynamical, and chemical features of already-known or simulated 2D materials, ML algorithms intelligently interpret complicated interconnections and correlations among such features, in an attempt to make predictions of unknown characteristics of new 2D materials.[3] Moreover, once trained, ML models can make very rapid predictions, making ML a promising tool for evaluating a large number of 2D materials. Therefore, the adoption of ML efficiently enables a wide variety of studies, including the understanding, discovery, and synthesis of 2D materials.

There are a few recent reviews about ML applications in materials research,[4-6] but they tend to cover the general use of ML in materials in general and mainly focus on the prediction of materials properties, which is one of the most common applications of ML. Therefore, a comprehensive review specifically concentrated on ML-enabled studies of 2D materials is still much needed. In this tutorial review, we condense and introduce the recent adoptions of ML in the field of 2D materials. Focusing on the understanding, discovery, and synthesis of 2D materials, we provide a comprehensive description and future outlooks. Starting from the review of various ML algorithms for supervised, unsupervised, and semisupervised learning, we describe how these algorithms have been applied to the specific study of 2D materials. More specifically, we discuss how various regression and classification algorithms can process labeled and unlabeled data and extract meaningful predictions that are otherwise difficult to detect. Such predictions enabled by ML algorithms are beneficial for understanding the mechanical, electrical, and chemical properties of 2D materials and their heterostructures that have yet to be discovered.

Furthermore, ML accurately identifies the layer thickness and size of 2D materials prepared by the mechanical transfer method[7-12] and predicts the synthesis probability of 2D materials,[13] which contributes to innovative synthesis approaches. Applicational studies that have benefitted from the use of 2D materials such as sensing[14, 15] and catalysis[16, 17] are also presented. Finally, we present future opportunities that could be an excellent starting point for researchers seeking to use ML in the field of 2D materials. It is highly anticipated that ML could significantly promote the study of 2D materials. Therefore, this review not only can inspire novice researchers but also guide mature researchers who are interested in applying ML in their studies.

2. Machine learning (ML) for 2D materials research

2.1. Fundamental ML algorithms

Machine learning (ML) approaches, as a subset of artificial intelligence (AI), are a group of algorithms which seek to determine the underlying connectivity among data. This process is referred to as "learning", in which ML algorithms are trained to review a specific data set and predict reliable outcomes upon new incoming data. Different from physics-based modeling, ML prediction is a type of data-driven decision-making process that can be self-improved by experiencing more data sets without direct reprogramming. Before discussing specific research examples that used ML to explore 2D materials, we first provide an overview of the ML techniques, followed by illustrative examples. An ML approach can be performed by four steps: data preparation, model selection, training, and evaluation. In data preparation, the dataset is collected, normalized or standardized, subjected to outlier removal (if appropriate), and split into training and testing, or training, validation, and testing subsets. Training consists of determining parameters to give functions that map inputs to outcomes in the training data (i.e., ML models).

Cross validation is performed by calculating the prediction errors of the ML models on validation data, in order to adjust and optimize the hyperparameters used in ML model training. Often, N-fold cross validation is used, in which the training/validation data is split into N subsets, and the training is performed on N-1 subsets while validation is performed on the last subset, and the whole process is repeated 10 times. Testing is using the ML models on data set aside, i.e., are not used in training or validation, to determine the accuracy of prediction. Evaluation can be performed by using the various measures of error which compare the original and predicted data. Depending on how data is handled, ML techniques mainly comprise three categories: supervised, unsupervised, and semisupervised learning, in which specific tasks such as regression, classification, clustering, and feature dimensionality reduction are performed. **Fig. 1a** and **Table 1** illustrate the different types of ML techniques and representative ML algorithms that have been widely used for 2D materials research.

The first and most common type of ML is supervised learning, which requires a large volume of pre-labeled data. The labeled data implies that outcomes for given inputs are correctly defined, which trains ML algorithms in the way that a teacher who already knows the answers (labeled data) teaches students (ML algorithms). Conceptually, supervised learning is the process of finding a mapping function, $f(\{x\})$, that can give outputs close to the labeled values of the original datasets for given inputs, $\{x\}$. The trained ML algorithms in supervised learning make predictions by "classification" or "regression" using known data. In classification, ML tries to find the best category ("class") to which a given input dataset likely belongs. Such classification models are beneficial for answering "yes or no" or discrete questions, such as "the feasibility of synthesizing 2D materials" or "the layer number of 2D materials". Regression, on the other hand, predicts continuous outputs from a given dataset. In other words, supervised regression results in a specific

numerical output that is as close as possible to labeled outputs in the training data instead of producing discrete results. Therefore, regression is suitable for predicting the properties of 2D materials with specific values such as band gaps, mechanical modulus, and formation energies.

A few ML algorithms commonly used for supervised learning to study 2D materials are support vector machine (SVM), least absolute shrinkage and selection operator (LASSO), and random forest (RF). SVM is one of the most robust ML training algorithms for handling classification[18] and regression (i.e., support vector regression (SVR)) problems. Based on statistical learning frameworks, SVM optimizes the process (**Fig. 1b**) to determine the hyperplane (red line) that bisects the maximized margin (the distance between the black dashed lines) and separates the datasets into different classes by choosing the appropriate support vectors (circular and star-shaped dots overlapped by dashed lines). Mathematically, the hyperplane is a set of point **x** that satisfies $\mathbf{w}^{\mathrm{T}}\mathbf{x} - \mathrm{b} = 0$, where **w** is the normal vector to the hyperplane and b is the half distance of the margin.

LASSO is a type of linear regression algorithm that uses regularization to shrink the data values towards a central point, such as the mean (**Fig. 1c**). Its objective is to find the coefficients of a fitting curve, which minimizes a loss function (least square errors plus the L1 norm, which is defined as the sum of the absolute values of the coefficients). In other words, for input vector **x** and target vector **y**, LASSO minimizes $\left\{||\boldsymbol{y} - w^T\mathbf{x}||_2^2 + \alpha||\mathbf{w}||_1\right\}$, where **w** is the vector of fitted coefficients, in contrast with ordinary least squares which minimizes $\left\{||\boldsymbol{y} - w^T\mathbf{x}||_2^2\right\}$. The parameter α is used for regularization and is typically determined by minimizing cross validation errors. LASSO is advantageous in regression problems in its ability to avoid overfitting, which enhances the generalizability of the regression model and thus increases prediction accuracy. Furthermore, LASSO performs variable (coefficient) selection, making the coefficients of trivial

variables zero and automatically ranking the remaining variables, and thus showing which variables are more important than others. This variable selection capability is beneficial for increasing accuracy and providing interpretability in many ML problems with a large number of possible input variables (input features).

Similar to LASSO, ridge regression (RR) and kernel ridge regression (KRR) are also widely used and robust supervised algorithms. RR estimates the coefficients of multiple-regression models and is used where independent variables are highly correlated. In RR, the quantity $\left\{||\boldsymbol{y} - w^T\mathbf{x}||_2^2 + \alpha||\mathbf{w}||_2\right\}$ is minimized. The key difference between RR and LASSO is that RR uses the L2 norm for regularization whereas LASSO uses the L1 norm, which results in small coefficients being more strongly penalized by LASSO than by RR. Therefore, while both RR and LASSO prevents overfitting, LASSO is more effective in model reduction. KRR combines RR with the kernel trick, which means learning a function in the space induced by the respective kernel. Instead of explicitly transforming data in raw representation into feature vector representations, KRR simply computes the inner products between the images of all pairs of data in the feature space. The amount of calculation significantly reduces for only access to the kernel and avoiding explicit computation of the coordinates.

Like SVM, random forest (RF) is a robust algorithm widely used for both classification and regression. RF is an ensemble of decision trees constructed by random samples selected from the original training data set (**Fig. 1d**). This random sampling process, known as "bagging (or bootstrapping)", repeatedly replaces the training set of decision trees. The RF fits the decision trees to the repeatedly updated samples and outputs the class most frequently selected by decision trees for the classification. In regression, an RF model outputs the mean or average of predictions returned by each tree. Additionally, the RF helps to rank the importance of variables by the order

of nodes and correct overfitting issues observed in a decision tree. Besides, more advanced methods have been developed to improve RF, such as gradient boosting machine (GBM) and extreme gradient boosting (XGBoost). GBM is a single strong prediction model in the form of an ensemble of weak prediction decision trees (i.e., shallow trees having high bias), which trains and improves the ensemble by iteratively adding previous weak trees, eventually reducing the high bias. Building upon GBM, XGBoost uses a more regularized model formalization that prunes the trees, which lowers the variance, thereby preventing overfitting. Furthermore, XGBoost builds trees in parallel, while GBM is sequential, resulting in faster predictions than GBM.

Furthermore, K-nearest neighbor (KNN) is a non-linear classifier that finds decision boundaries from training data and sorts testing data into various categories. Once a constant K is initially defined, the algorithm examines the K-nearest data around each testing data. Subsequently, testing data is assigned to a category where a majority among K-nearest data belongs. In addition, the neural network (NN) shown in **Fig. 1e** consists of input, hidden, and output layers and adaptively learns highly complex non-linear relationships between the input features and target outputs (i.e., labels of the original data). The input layer delivers input features to hidden layers comprised of neurons. Each neuron is connected to all the neurons from the previous layers and it adds up the input features multiplied by weights. The weighted sum of the input features is further delivered to the output layer, where the activation function determines the predicted variables. A series of such processes is repeated until the NN finds the optimal weights that minimizes the difference between prediction and labels (i.e., target outputs). With many hidden layers, the deep neural network (DNN) incrementally correlates input features with desired outcomes. There are several advantages of the DNN: (1) Automatic extraction of features from inputs without human intervention (i.e., DNN does not require additional labor to assign labels),

(2) ability to handle non-linear and complex problems, and (3) high predictive accuracy by increasing learning epochs, neurons, and hidden layers.

Unlike supervised learning in which training data are labeled in advance, unsupervised learning attempts to determine new patterns and distribution from unlabeled data. For example, unsupervised clustering divides data into individual groups with similar features. For materials research, data belonging to the same group potentially can be considered to have similar characteristics in material properties or synthesizability. Prime examples of unsupervised learning are K-means clustering and principal component analysis (PCA). The K-means algorithm works by finding mutual similarity between samples and clustering them into groups. Its goal is to achieve high similarity within-cluster while keeping low similarity inter-cluster. Technically, the K-means algorithm iteratively and continuously updates the centers (or centroids) of clusters until the variances of each cluster are minimized. For example, as illustrated in **Fig. 1f**, to reach the clustered result, K-means first generates three centroids of clusters at random locations of the data space and assigns data points to the nearest cluster based on the Euclidean distance ($d=\sqrt{x^2+y^2+z^2}$) to the centroids. Once the initial clusters are established, the mean variance (sum of d^2) of each cluster is evaluated and the centroids updated. These steps are iterated until the mean value of the variances saturates, indicating the formation of optimal clusters.

PCA is a nonparametric statistical technique most widely employed to reduce the dimension of a large data set in exploratory data analysis while minimizing the loss of information. Such data reduction is realized by computing the principal components that constitute a set of orthonormal bases on the data, where only the first few principal components are significant and the rest is ignored. For example, the 2D data points shown in **Fig. 1g** tend to be on a straight line (y=x), which implies that such data contain redundant features. To reduce redundancy (lower the

dimensionality), PCA first looks for a new axis (PC1) that maximizes the variance of projected data, and then another new axis (PC2) orthogonal to it. The data points show minimal variance along PC2, which means that PC2 is not as important as PC1 to represent the data. In this example, the dimension of data is lowered from 2-D (x, y) to 1-D (PC1). By ignoring PC2, PC1 becomes the new axis representing the original data without losing too much information. Oftentimes, PCA is used prior to supervised regression or classification, in which the reduced data dimensionality afforded by PCA can lead to more compact supervised ML models with less overfitting and better generalizability.

Another straightforward yet powerful unsupervised learning method is hierarchical clustering, which is an algorithm that iteratively groups similar objects into multi-level clusters.[19] It starts from merging two most similar objects, and proceeds through an iterative process that identifies and merges the two most similar clusters until the final state. The final state is a set of clusters in which each cluster is distinct from other clusters and the objects within each cluster are similar. The distance between any two clusters is called the linkage distance. The linkage criteria determine from where distance is computed. Single-linkage computes the minimum distance between two objects from two clusters, while complete-linkage takes the maximum distance, and mean or average-linkage takes the mean distance.

Semisupervised learning poses in between supervised and unsupervised learnings, playing its role when handling datasets in which only a few are labeled (supervised), but the rest is unlabeled (unsupervised). Such semisupervised learning is a practical ML model for dealing with many real-world classification and clustering problems associated with predicting outcomes based on a dearth of correct information. In 2D materials research, for example, semisupervised learning can be used to classify and predict the synthesizability of 2D materials where there are a few 2D materials

known to be synthesizable (labeled as "Yes") and many unknown 2D materials yet to be synthesized. Semisupervised learning is the most challenging ML approach compared with others and thus many algorithms are under development in this field. Specifically, the SVM shown in **Fig. 1b** can be applied to classify mixed datasets (labeled and unlabeled). The SVM first classifies such datasets using only labeled data, and predicts the probability of unlabeled data belonging to each class. Such probabilities in SVM are estimated using Platt scaling, which converts the outputs from the classification model into a probability distribution. Subsequently, unlabeled data showing high probability (i.e., a high confidence level) at a specific class is considered pseudo-labeled and added to the original training data. Finally, the augmented dataset is used for training the SVM.

Another algorithm widely used in semisupervised learning is positive and unlabeled (PU) learning. PU learning is a binary classifier that deals with two sets of data: the positive set P (labeled) and a mixed set U (unlabeled). In PU learning, various techniques that are used for a supervised classifier can be adopted. The PU algorithm first randomly selects a few unlabeled data and considers them as positive. Then, such positive and pseudo-positive data are classified using classifiers, and the probability of the pseudo-positive data being positive is evaluated. By repeating this process, PU learning probabilistically classifies mixed datasets into two classes (positive or not).

Note that in most cases, it cannot be determined *a priori* which ML algorithm would be best for specific tasks in 2D materials research. A common approach is to train several typical ML models and judge their performance by the errors and corresponding uncertainties. In the following, we present detailed discussions on which ML algorithms introduced above are applicable to a specific task, how they work, and how to determine their hyperparameters, along with actual examples.

2.2. Performing ML and validation

In applying ML to the study of 2D materials, a series of steps including data preparation, model selection, training, cross validation, and testing should be successively carried out to build an ML model. It is worth noting that ML study on 2D materials often suffers from the lack of data, because data acquisition processes are limited to time-consuming experiments and computations. To address this problem, material databases that provide comprehensive information about 2D materials can be used to gather ML data. **Table 2** shows a list of material databases, including the structural, electronic, elastic, thermodynamic, and optical properties of 2D materials obtained from previously performed experiments and computations. After accumulating the data, an ML model suitable for a specific study of 2D materials should be determined. Depending on the available data and the goal of the study, ML algorithms belonging to either supervised, unsupervised, or semisupervised learning should be considered for regression, classification, or clustering. For example, supervised regression can be used for predicting the properties of 2D materials. Furthermore, semisupervised classification can be considered for investigating the synthesizability of 2D materials. More examples of the application of ML models for specific studies are introduced in the following sections. Once an ML model is selected and trained, cross-validation is performed from validation dataset withheld from training to determine the accuracy of the model and adjust the hyperparameters. Afterwards, the trained and cross-validated ML model can generate predictions using test datasets. The predictions are further compared by labels from yet another test datasets, and their accuracy (*i.e.,* error) is evaluated using various statistical measures.[18]

Typically, there are two types of prediction errors in validating the ML model: variance and bias. **Fig. 2a** shows variance and bias errors that can respectively induce overfitting and underfitting of the model if not balanced. High variance error implies that the model captures too

many details in datasets, including unnecessary noise, making the model less generalizable and unable to predict beyond the original datasets. Many ML algorithms, e.g., decision trees, support vector machines, and k-nearest neighbors, can suffer from overfitting issues. To address such overfitting, several approaches such as regularization (e.g., using LASSO), removing features (using LASSO or PCA), ensemble (using RF), and cross-validation can be used.

On the other hand, high bias error originates from the model capturing the datasets too sparsely, resulting in an over-simplified model that does not include important details in datasets. Such an underfitting issue gives rise to inaccurate predictions and can be reduced by increasing the complexity of the model, the number of features, and the number of learning iterations.

Fig. 2b suggests useful statistical measures to evaluate such prediction errors and validate ML models. In regression models, RMSE, MAE, MAPE, and R^2 shown below are the most popular metrics.

$$\text{Root mean squared error (RMSE)} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}$$

$$\text{Mean absolute error (MAE)} = \frac{1}{N}\sum_{i=1}^{N}|Y_i - \hat{Y}_i|$$

$$\text{Mean absolute percentage error (MAPE)} = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i - \hat{Y}_i|}{Y_i} \times 100$$

$$\text{Coefficient of determination, } R^2 = 1 - \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}$$

where Y_i are the original labeled data, $\hat{Y}_i$ are the predicted outcomes from the trained ML model, and $\overline{Y}$ is the mean of the original labeled data. Such metrics evaluate how much the regression curve fits the original data, determining the performance of the regression model.

Precision, recall, accuracy, and area under curve (AUC) for receiver operating characteristic (ROC) curve are widely used to evaluate the performance of classification models. To understand such metrics, basic terminologies such as true positive (TP), true negative (TN), false positive

(FP), and false negative (FN) should be demonstrated. TP and TN represent number of items for which the classifier correctly predicts the class of data, while FP (FN) shows the number of items for which the classifier incorrectly predicts the data belonging to a negative (positive) class to a positive (negative) class. Using such terminologies, precision is defined as Precision = TP / (TP + FP), showing the ratio of correctly predicted positives to all the predicted positives. On the other hand, recall, defined as Recall = TP / (TP + FN), shows the ratio of correctly predicted positives to all true positives. Such precision and recall metrics are used to evaluate how many incorrectly classified FP and FN the classifier produces, respectively. Combining these two metrics, F1-score, defined as F1-score = 2 * ((Precision * Recall) / (Precision + Recall)), or the accuracy, (TP + TN) / (TP + TN + FP + FN), are generalized scores to evaluate the overall performance of the classifier.

ROC curve is a useful way to visualize the performance of a classifier, defined as the ratio of true positive rate (*i.e.,* TPR or recall) to false positive rate (FPR = FP / (FP + TN)). The FPR shows how often the classifier incorrectly predicts the data as positive out of all true negatives. ROC curve plots the TPR or recall as a function of the FPR, indicating a better performance as the curve is closer to the top-left corner. Additionally, AUC calculates the area under the ROC curve, scoring the performance of the classifier between 0 (bad) and 1 (good).

Effective metrics widely applied to clustering models are the rand index (RI) and gap statistics. The RI, similar to the accuracy discussed above, is calculated by RI = (TP + TN) / (TP + TN + FP + FN), showing how similarly the model predicts the data compared to the ground truth. Besides, the RI scores the similarity between clustered datasets using two different clustering models, validating the new model on the basis of the already verified model.

In clustering models, choosing an optimal number of clusters (*i.e.,* hyperparameter, K) is imperative to maximize the performance of the models. The simplest way to determine K is to use

the elbow method that plots within-cluster distances with respect to K, which determines the K, the starting point of significant drop. The elbow method is straightforward but often ineffective if the curve does not show a noticeable drop (*i.e.,* elbow shape). The gap statistics method depicted below is an alternative and advanced strategy to address this issue.

$$Gap(K) = \log\left(W_K^{ref}\right) - \log\left(W_K^{orig}\right)$$

$$S_{K+1} = S_K\sqrt{1 + 1/N}$$

where $W_K^{orig}$and W_K^{ref} are within-cluster distances obtained from the original dataset and the reference dataset generated by uniform sampling when the number of clusters is K. S_K is the standard deviation of within-cluster distances, $\log\left(W_K^{ref}\right)$, calculated from reference datasets obtained from N times sampling. Finally, the optimal number of clusters is the smallest K that satisfies the following relationship.

$$Gap(K) \geq Gap(K+1) - S_{K+1}$$

In short, the gap statistics compares the within-cluster distances of the original dataset with that of the reference dataset, thus finding a K value exhibiting the most significant gap between them.

Such evaluation metrics introduced above can be appropriately applied to various studies in 2D materials. For example, metrics for the regression and classification models can be used in the studies, such as investigating the properties and correlating such properties with molecular structures. Furthermore, metrics for the classification and clustering models can be applied when identifying the thickness and size of 2D materials and studying their synthesizability. Detailed usage of such metrics is introduced in the following sections.

3. Understanding and designing 2D materials using machine learning

2D materials have given rise to countless possibilities for applications due to their interesting two-dimensional planar atomic structures that contribute to promising mechanical, electrical, and chemical properties. Considering the tunability of 2D materials by composition tuning, defect engineering, surface doping, and the formation of heterostructures, the design space for potential 2D materials is staggeringly large. Recent studies introduced below illustrate how the use of ML takes advantage of materials data to significantly enhance the speed and lower the cost of studying 2D materials.

3.1. ML-enabled study on the mechanical properties of 2D materials

Unlike 3D bulk materials in which chemical bonds extend to three dimensions, bonding in 2D layered materials is strongly localized in-plane, resulting in weak out-of-plane stacking by van der Waals (vdW) forces. While this weak layer-stacking facilitates the separation of atomically thin 2D materials, the strong in-plane bonds generally endow the isolated 2D materials with promising mechanical stiffness and strength, such as high elastic modulus and tensile/compressive strengths. Furthermore, the atomic-scale thickness of 2D materials gives rise to their superior flexibility compared to their 3D bulk counterparts made up of the same atoms. For example, graphene, the most renowned and first-isolated 2D material, has been ranked as the strongest material ever discovered, exhibiting high levels of stretchability and flexibility that exceed its 3D allotrope, diamond. Graphene's intrinsic tensile strength, Young's modulus, and stretchability are evaluated at 130 GPa, 1 TPa, and ~20 % (from its initial lateral dimension), respectively[20], surpassing those of diamond (125 GPa, 1.1 TPa, and ~ 13.4 %, respectively, when the diamond is in the form of a 60 nm-diameter nanoneedle engineered to achieve its maximum tensile strength and strain).[21]

Besides graphene, other 2D materials such as hexagonal boron nitride (h-BN) and molybdenum disulfide (MoS_2), which exhibit insulating and semiconductive behaviors, respectively, show excellent Young's modulus and tensile strength of 270 GPa/22 GPa and 865 GPa/70.5 GPa, along with high flexibility.[22] The promising mechanical robustness of 2D materials triggers the rapid development of strong composite materials and flexible/wearable electronics. Additionally, 2D materials show interesting mechanical behaviors, such as easy interlayer fracture and shear due to the weak vdW forces, making them promising lubrication additives for controlling friction and wear.

The prevailing methods for evaluating the mechanical behaviors of 2D materials are direct measurements and computational simulations. In a direct measurement, target 2D materials are carefully transferred to cover an etched hole on a substrate, forming suspended 2D materials. These materials are then subjected to several nanoindentation tests using an atomic force microscopy (AFM) tip, resulting in force-displacement (F-D) curves. The F-D curves are subsequently analyzed to calculate Young's modulus (from the slope) and tensile strength (from the maximum stress point). The AFM can also be used to evaluate the tribological characteristics (i.e., friction and wear) of 2D materials. Specifically, lateral force microscopy (LFM) detects how much the AFM tip twists while traveling across the surface of 2D materials and converts the torsions to friction forces, thus evaluating nanoscopic friction behaviors. Yet such experimental approaches require intensive labor to prepare experimental setups and perform multiple tests. Computationally, MD and DFT simulate experimental configurations such as nanoindentation, bi/uni-axial stretching, and friction tests to calculate the mechanical properties of 2D materials; however, these methods can be computationally expensive, and thus are far from an efficient approach for studying numerous 2D materials.

To support such conventional methods, various ML techniques can be introduced to effectively study the mechanical properties of 2D materials. For example, SVM can be adopted to predict the fracture strength of graphene by identifying the monolayer and sampling its distribution of lateral sizes.[12] The optical micrograph (OM) of graphene layers transfer-printed on an Si/SiO_2 substrate shows different colors due to the thickness dependence of thin-film optical interference, and the intensities of red, green, and blue components from the image are extracted and used as input features for ML. In detail, each pixel on the OM can be represented by the combination of red, green, and blue (R, G, B) color values ranging from black (0, 0, 0) to white (255, 255, 255), which are strongly associated with the thickness of the 2D materials.

Fig. 3a shows the SVM classified graphene layers according to their thickness using labeled data (i.e., input features: intensities of red and green, labels: the thickness of graphene). The lateral sizes of the identified monolayer graphene layers were further determined by counting the number of pixels occupied in the OM. Compared with visual inspection, the accuracy of the ML-assisted identification of monolayer graphene and their sizes was 98.2% (**Fig. 3b**). It is noteworthy that the amount of time spent identifying the sizes of the monolayer graphene was ~136 sec, much faster than manual inspection. According to the Weibull strength theory, the size distribution of mechanically transferred graphene is dependent on the fracture behavior of graphene. Together with other materials properties such as shear modulus, Young's modulus, and thickness, the fracture strength of a graphene layer can be predicted (**Fig. 3c**).

In another example, the mechanical properties (fracture strain, strength, and Young's modulus) for tungsten disulfide (WS_2) were evaluated using an RF regression algorithm.[23] To obtain sufficient data for ML, uniaxial tensile tests were performed using MD simulations under various conditions. Specifically, the input conditions (e.g., chirality, strain rate, and density of defects) of

the MD simulations and the corresponding outputs (e.g., fracture strain, strength, and Young's modulus) served as input features and labeled outputs, respectively. **Fig. 3d** shows the schematic of a regression tree comprised of multiple nodes including root, decision, and leaf nodes. Input features from collected data occupy root and decision nodes, where branch splitting occurs according to if-else statements. A regression tree divides labeled data into two categories at every node and evaluates the error using the metrics of RMSE, MAE, MAPE, or R^2. If the error is minimized and meets the requirement, the tree splitting stops and reaches the leaf node. Finally, the data settled at the leaf nodes are averaged and used for prediction. The complexity of the RF algorithm is controlled by two hyperparameters, max-depth and n-estimator. The max-depth is related to the number of splits from a single tree, and n-estimator indicates the number of trees in the RF. The higher number of hyperparameters can lead to better prediction accuracy in the training dataset but at the cost of overfitting. To reduce overfitting, it is very important to choose an optimal feature at the root node. In selecting the optimal parameter, the RF generally calculates the Gini index of each feature, which counts the number of splits. The lower Gini index of the feature indicates more efficient data splitting, thus showing higher importance. However, in this study, an alternative metric that calculates feature importance was introduced. **Fig. 3e** shows the Pearson's correlation evaluating a linear relationship between the input features and the target outputs. Five input features (type, chirality, temperature, strain rate, and defect) were correlated to three target outputs (fracture strain, strength, and Young's modulus). In this study, "defect" was chosen for the feature at the root node because it showed a strong connection to the target outputs. **Fig. 3f** plots the prediction results of Young's modulus of WS_2 using the trained RF algorithm. The prediction error was 3.8 GPa (cf. 117.8 GPa Young's modulus for WS_2), revealing that the prediction using RF was highly accurate.

Moreover, the nanoscopic friction behaviors of 2D materials were predicted using Bayesian learning in a recent study.[24] Specifically, the inter-layer friction that occurs when two layers slide relative to each other was considered. Since the inter-layer friction occurs by overcoming the maximum energy barrier (MEB) of the potential energy surface (PES), the MEB is the core parameter for understanding the friction behavior of 2D materials. In this study, the structural, electronic, chemical, and thermal properties affecting the PES were used as the input features, and the DFT and MD-calculated MEB values of five different 2D materials found in the literature were used as the target outputs (i.e., labels). The low volume of labeled data (only five) would very likely end up overfitting many other ML models, and thus a probabilistic Bayesian algorithm was adopted in this study to handle such sparse data. Bayesian learning is based on Bayes' theorem (**Fig. 4a**), in which the posterior probability of A given B (P(A|B)) can be calculated with the knowledge of the likelihood of B given A (P(B|A)), prior of A (P(A)), and evidence of B (P(B)). Namely, it makes a prediction (posterior) by adjusting the likelihood in consideration of other probabilities (e.g., prior, evidence) associated with the target outputs. In this study, the Bayesian algorithm predicted MEB values of ten new 2D materials to be close to the known MEB values from five 2D materials in probability according to their similarity. **Fig. 4b** shows the correlation coefficients among 15 different 2D materials obtained by comparing their input features. A similar approach was used in the Bayesian algorithm, and, consequently, the MEB values of 15 different 2D materials, including five already-known values, were predicted, as shown in **Fig. 4c**.

The examples discussed above indicate that ML algorithms can produce fast and accurate predictions when synergistically combined with computational simulations and previously reported data, thereby reducing the time and cost to investigate the mechanical properties of 2D materials. Notably, conventional approaches to studying 2D materials are usually used for

accumulating data, while ML subsequently infers outputs by determining the correlations among data and applying these correlations to other materials.

3.2. ML-enabled study on the electronic properties of 2D materials

Other appreciated characteristics of 2D materials are their fascinating electronic and optical properties resulting from the confinement of electrons in a 2D plane. Contrary to 3D bulk materials, the crystal structure of 2D materials loses its periodicity along the direction normal to the plane, generating interesting band structures. For example, several 2D materials such as graphene, silicene, germanene, and graphynes (sp-sp^2 allotropes) show a Dirac-cone band structure which gives rise to massless Fermions, resulting in ultrahigh carrier mobility that is more than 100 times higher than that of silicon.[25] Additionally, the band structures of most 2D materials are highly dependent on the thickness (i.e., the number of layers) from bulk to monolayer, providing the tunability of band gaps. Such tunable band gaps combined with high carrier mobility could enable the development of next-generation optoelectronic, semiconductor, and sensor devices. Moreover, the dimensional confinement of 2D materials reduces the dielectric screening effect between the electrons and the holes, thus increasing the Coulomb interactions and exciton binding energy. As a result, excitons found in 2D materials are more tightly bound and stable than those in bulk materials, which leads to strong light-matter interactions. More interestingly, naturally existing conducting (e.g., graphene), semiconducting (e.g., MoS_2), and insulating (e.g., h-BN) 2D materials present many advantages, thus opening new opportunities to design electronic devices in which all the components (e.g., semiconducting channel, metallic electrode, and insulating dielectric) consist of atomically thin 2D materials. Such a combination benefitting from the variety in the

band gap of 2D materials is expected to overcome the scaling limit issue in current semiconductor devices.

In order to investigate these electronic properties of 2D materials, sophisticated experimental approaches and expensive computational calculations are required. For example, absorption spectroscopy is widely used to extract band gap and exciton energies by analyzing the absorption spectrum representing the absorption intensity of 2D materials to the incident radiation as a function of wavelength. In addition, carrier mobility can be experimentally inferred from the transfer characteristic curve of field-effect transistors (FETs) or Hall measurement. In a computational approach, DFT is used to calculate the band structure of 2D materials, identifying electronic characteristics such as the band gap and the effective mass, and with some approximations also the mobility of the charge carrier and conductivity. However, DFT says nothing of the experimental methods and some of the calculations require a significant amount of time and resources, even with the use of high-performance computers.

In recent years it has been shown that ML supports experimental and computational approaches, thus enabling productive studies on the electronic properties of 2D materials. For example, ML has been employed to unearth 2D MXenes with band gaps ranging from 0.5 to 2 eV and thus hold significant potential for various applications.[26] Additionally, the band gaps of the discovered MXenes were accurately predicted within seconds with the aid of ML. A series of ML steps for the above tasks is depicted in **Fig 5a**. Initially, 23,870 functionalized MXenes with a structure described as MM'XTT', where M and M' are elements in groups IIIB to VIB, X represents either C or N, and T and T' are either single elements (H, F, Cl, Br, O) or groups (CN, NO, PO, OH, OCl, OBr, OCN, SCN, NCS), were considered as subjects of the study. Afterwards, the evaluation of the Perdew-Burke-Ernzerhof (PBE) band gaps was carried out using DFT for

7,200 MXenes randomly chosen from the total of 23,870. The calculation of PBE band gaps is faster but significantly underestimated compared with GW band gaps (which currently are closest to experimental values among first-principles band gap prediction approaches for solids). However, such PBE band gaps are enough to be used as data labels for qualitatively building a metal-semiconductor classification model. In a decision tree (DT) classification model, a single tree from the RF method described in Section 3.1 was used to divide the MXenes into metallic and semiconducting. For the DT model, the PBE band gaps served as labels (i.e., target outputs) and the corresponding input features were adopted from materials databases.[26] As mentioned earlier, ML study on 2D materials often suffers from the lack of data, because data acquisition processes are limited to experiments and computations. Here, 643 MXenes randomly selected from 7,200 MXenes with pre-evaluated PBE band gaps were used to train the DT model. This trained model successfully screened out the metallic MXenes from all 23,870 MXenes, with an accuracy greater than 94%. After the classification, it was found that MXenes based on Sc and Y have band gaps between 0.5 and 2 eV, which is suitable for electronic and catalytic applications. Although the DT classification model successfully discovered promising semiconducting MXenes buried in a large number of MXenes, the predicted band gaps were underestimated because the model was trained with PBE band gaps. Therefore, additional ML was carried out to evaluate the accurate band gaps of classified MXenes. Here, 70 randomly chosen MXenes among as-classified MXenes underwent further high-accuracy band gap calculations based on the GW approximation. The GW band gaps along with 47 primary features of MXenes (e.g., Volume per atom: V, Lattice parameter: a, Phase of MXene: c, Boiling point: T_M, etc.) collected from material databases were used for the labels (Y) and input features ({x}), respectively. The 47 input features were subsequently reduced to 15

features that strongly correlated to GW band gaps, resulting in an efficient ML model with high accuracy and a low chance of overfitting.

The feature reduction was performed using the LASSO algorithm, and the 15 features (e.g., Vacuum potential for lower surface: $\emptyset_L$, Standard meting point: T_M^{STD}, etc.) having non-zero correlation to the GW band gap are shown in **Fig. 5b**. The LASSO described in Section 2.1 is one of the regression algorithms and expressed as $L(\beta) = \|Y - X\beta\|_2^2 + \alpha\|\beta\|_1$ (summation of L1 and L2 norms), where β is coefficients of the regression model, Y is target outputs (here, GW band gaps), and α is the coefficient to control the penalty. Since the constraint boundary of the L1 norm of the $L(\beta)$ is a diamond shape, which makes zero coefficients β at vertices, input features not directly involved (i.e., irrelevant) to the regression model can be removed.

Finally, supervised learning with the Gaussian process regression (GPR) algorithm was performed using the input features and GW band gaps as the data set {X, Y}. The GPR is a kernel-based stochastic process which is highly accurate even when using a low volume of data. In GPR, the regression model (i.e., regression function) does not predict deterministic outputs, but rather outputs with stochastic randomness because the model consists of Gaussian distributions, $N(\mu,\sigma^2)$. The following equation implies the concept of the GPR, $y' = \sum_{i=1}^{N} w(x',x_i)y_i$, where w are the weights, x_i and y_i are original data, and x' and y' are missing data points and corresponding outputs, respectively. In the GPR the w is represented by a kernel, and learning is the process to tune the kernel in order that y' reflects y_i more as x' is closer to x_i. GPR has a strong advantage of accurate prediction and provides the uncertainty of predictions. **Fig 5c** shows the predicted band gaps of MXenes using GPR, which is highly consistent with GW gaps with a RMSE of 0.14 eV. The trained model could successfully predict accurate band gaps of MXenes within seconds, even though the model did not include PBE band gaps ET_g^{PBE} as one of the input features. This indicates

there is no need to perform DFT calculations for extracting the GW band gaps of MXenes, as illustrated in **Fig 5d**.

Another recent study demonstrates that ML combined with DFT calculations can also significantly contribute to predicting the band gaps of 2D heterostructures.[18] Since 2D heterostructures consist of different 2D materials stacked vertically or stitched laterally, the use of ML dramatically enhances the efficiency of studying countless combinations of them. In this study, 21 non-metallic 2D materials such as MoS_2, HfS_2, BN, and CdO were used to construct 210 2D heterostructures consisting of two different monolayers stacked vertically. Subsequently, DFT calculations were carried out to collect the band gaps from 49 heterostructures that were randomly selected from 210. For data preparation, the property-labeled materials fragments (PLMF) method, which extracts a materials' characteristics from a graph representing a crystal structure, was used. In the PLMF method, the adjacency matrix is widely used to represent the crystal structures of materials.[27] **Fig. 5f** shows an example of the adjacency matrix constructed from MoS_2. Each atom in the MoS_2 has numbered labels and the connectivity is represented as a matrix, where 1 indicates the existence of a bond between atoms and 0, otherwise. The adjacency matrix does not necessarily reflect only the connectivity using either 1 or 0; instead, components of the matrix can include the structure and properties of materials such as bond lengths, angles, and charges, among others, which can be found in the materials databases shown in **Table 2**. Therefore, the PLMF method using the adjacency matrix can produce a considerable number of input features that encode the topology and properties of the corresponding materials. Furthermore, in this study, 1,529 input features were obtained using the PLMF method and then reduced to 11 significant features using LASSO.

Using such features and band gaps obtained from DFT, a neural network (NN) was trained and used to predict the band gaps from all 210 2D heterostructures. **Fig 5g** shows the band gaps predicted from the trained NN for a test set of 2D heterostructures have a linear correlation with the gaps from DFT calculations, with an MSE of 0.047 eV^2. Additionally, the trained NN successfully predicted the band gaps of all possible 210 2D heterostructures, revealing the powerful and promising advantage of ML for the study of 2D materials.

Lastly, the exciton valley polarization landscape of monolayer WSe_2 was predicted using the RF algorithm.[28] Typically, the exciton valley polarization of 2D materials can be observed using a low-temperature photoluminescence (PL) measurement, which requires high-end experimental apparatus. In this study, RF correlates the PL spectra obtained from 300K with those acquired from 15K, allowing the prediction of the exciton valley polarization landscape of WSe_2 without performing onerous low-temperature PL. For the ML, the polarization and position-resolved PL spectra from nine WSe_2 were measured under 300K (for input features) and 15K (for target outputs or labels). **Fig. 6a** shows the intensity, energy, full-width at half-maximum (FWHM) and the trion-exciton intensity ratio (T-X ratio) spectra obtained from the 300 K experiment that were used for the input features. The trained RF algorithm successfully predicted the exciton valley polarization of other WSe_2 using PL spectra measured from 300K, thereby mitigating the experimental complexity and cost required to perform low-temperature measurements. **Fig. 6b** displays a strong correlation between the predicted and experimentally measured exciton valley polarizations with a correlation coefficient (R^2) of 0.97.

In summary, supervised learning that correlates the input features with target outputs has been adopted to predict the electronic properties of 2D materials. In addition to conventional

experiments and simulations, materials databases provide comprehensive data for ML, which is expected to promote many more studies on the various properties of 2D materials.

3.3. ML-enabled design of 2D materials

2D materials design and engineering reveal fantastic prospects as well as significant challenges for fully leveraging quantum confinement effects. Novel 2D materials can be designed through various routes, such as defect engineering, the adsorption of atoms or molecules, and heterostructures, among others. Defect engineering is an attractive option, and studies have shown that purposefully-designed defects in 2D materials could exhibit exciting performance for novel applications such as single-photon emission, resistive switching, and neuromorphic computing. However, the diversity and complexity of defects makes their control challenging, and experimental screening and exploration is very slow. Therefore, the rapid prediction of defect properties in 2D materials through highly efficient methods, such as ML, is crucial.

A recent study has employed ML to rapidly predict defects in 2D materials for quantum emission and neuromorphic computing.[29] In this work, the most promising 2D material hosts for point defects were first identified through deep learning (DL), and then defects in these 2D material hosts for quantum emission and neuromorphic computing were predicted by the RF algorithm, as illustrated in **Fig 7a**. To identify the 2D material hosts, a dataset of 4,000 2D materials is available from the Computational 2D Materials Database (C2DB)[30], but this amount of training data is still insufficient for deep neural networks (DNNs). Thus, the DNN was pretrained on a data set of $10^4 \sim 10^5$ bulk materials from the Materials Project[31] database. Three models of graph networks as implemented in MatErials Graph Network (MEGNet)[32] were used to map the input structure graphs to the output target properties. The input graph representations were characterized by the

atomic numbers of the constituent elements and the spatial distance (bond lengths) between atoms. The output targets were formation energy, band gap, and Fermi energy. As an example of transfer learning, these pretrained models were then trained on the 2D materials data set C2DB ($\sim 10^3$), and the model weights were fine-tuned for 2D cases. **Fig 7b** shows the good performance of the deep-learning model on the formation energy prediction, as the R^2 is 0.98 and the MAE is 0.06 eV/atom on the test data. The accuracy of the metal versus nonmetal classifier is 0.84, with an F_1 score (which measures a combination of precision and recall) of 0.88 and 0.73 for metals and nonmetals, respectively. Though the performance of the 2D band gap model is worse than the others with an R^2 of 0.73, its MAE (0.36 eV) is similar to that of the bulk model (0.33 eV).[32]

For applications in quantum emission and neuromorphic computing, a good host material should have a wide band gap for isolating deep defect levels and small spin-orbit coupling (SOC). To identify optimal host 2D materials for these applications, the screening criteria were set as screening for nonmagnetic materials with band-gaps greater than 2 eV calculated with the GW approximation. Here, 158 candidate wide band gap (WBG) semiconductor 2D materials were identified. Screening out compounds with heavy elements to reduce the effects of SOC, 150 WBG candidates were obtained. Next, potential defects for quantum emission and neuromorphic computing in these 2D material hosts were explored by ML. To build the ML model, a data set with more than 1,000 quantum point defects (QPDs) was generated. The model started from the combination of the 150 optimal 2D material hosts identified from the first step through DL and 70 defects in 2D materials containing all possible vacancies, divacancies, antisites, and common dopants, which yielded more than 10,000 defect structures. These candidate defects were funneled into a subset for electronic structure calculations, which were then used to test ML models, as illustrated in **Fig 7c**. Relaxed defect geometries and band structures were computed for more than

1,000 QPDs and for 140 substitutional metal defects in the atomically thin resistive memory materials MX_2 (M = Mo, W; X = S, Se, Te) and h-BN. To identify promising defects, two ML models were built: one classifier to identify the deep center defects and one regressor for predicting defect formation energies. The output targets for ML prediction were DFT-computed band structures for the classifier, and the neutral defect formation energies E_f for the regressor. For the classifier, ΔVB and ΔCB were defined, which are the energy differences between the defect level and the valence band maximum and conduction band minimum, respectively. The threshold for a deep center defect was set as $\Delta CB > k_BT$ and $\Delta VB > k_BT$ at room temperature. For both the classifier and regressor, the input features for the ML approach were the structural and chemical properties of the host material and defect that were obtained from C2DB databases or the first-step deep transfer learning model predictions. The defect descriptors were normalized as percent differences between the corresponding descriptor for the bulk structure and the unrelaxed defect structure. The RF algorithm from *scikit-learn* was used for both models, and some descriptors were generated with *matminer*[33] and *automatminer*. Here, 90% of the data was split into a training set and 10% was held as a test set. **Fig 7d** shows the RF model for E_f prediction has R^2 of 0.74 and MAE of 0.67 eV on the test set. From the permutation feature importance in **Fig 7e**, the chemical potential of the defect species is directly related to E_f among the most important features. The linear Pearson correlations of individual features is lower than 0.3, implying the invalidity of a simple linear model to predict defect properties. Indeed, the performance of the nonlinear RF model on E_f prediction is much better than previously reported linear models such as LASSO and Ridge regression[34]. The RF model performed even better for the deep-center classifier, with F_1 of 0.92 on the test set, and 442 deep-center QPDs were identified. The most important feature is the lowest unoccupied molecular orbital (LUMO) energy.

Finally, to identify optimal defect candidates, a defect score metric that represents the fitness as a deep-center defect for quantum emission was defined as

$$S = \frac{1}{N}\left(E_{bg}^{GW} + \left(\frac{1}{2}s_d + \frac{1}{2}s_t\right) - A - E_f\right)$$

where E_{bg}^{GW}, s_d, s_t, A, and E_f are the GW band gap, dynamic stability, thermodynamic stability, maximum atomic number in the host, and defect formation energy, respectively. N is an overall normalization factor. Higher scores indicate the optimal defect candidates with larger band gaps, greater stability, smaller defect formation energies, and smaller SOC. The top 100 defect scores are plotted in **Fig 7f**, with the top 10 highlighted in the inset. Furthermore, a subset of substitutional metal defects in TMDs and h-BN with defects were identified for their potential nonvolatile resistance switching (NVRS) applications for information storage and neuromorphic computing. They were screened out based on Δz and E_{BE}, which are the change in out-of-plane distance relative to the equilibrium distance between the TM plane and the chalcogen plane, and the defect-binding energy of a metallic dopant, respectively. **Fig 7g** shows the highest five and lowest five defects by the maximum binding energy E_{BE}. The highest binding energy defects are of interest in memory applications for their assumed stability, while the lower binding energy defects require small switching voltages that are useful for neuromorphic architectures. These identified optimal defect candidates may find applications in quantum emission, resistive switching, and neuromorphic computing.

Adsorption is another effective way to engineering and design 2D materials. The intriguing success of 2D TMDs synthesis achieved through various methods such as mechanical exfoliation, chemical exfoliation, physical vapor deposition, and solution synthesis has led to intensive study of their potential applications. The adsorption of alkali metal atoms on 2D TMDs plays a crucial role in their performance as batteries, catalysts, and sensors. A recent work[35] has used a linear

regression ML model to investigate the characteristic energetic factors that determine the adsorption energy of lithium on 2D TMDs. The work demonstrated and was supported by ML that the lowest unoccupied states E_{LUS} is a novel efficient descriptor for predicting adsorption energies, due to the linear correlation.

In the ML process for this work, 112 cases were considered through the combination of seven transition metals (Ti, Hf, V, Nb, Ta, Mo, and W), two chalcogens (S and Se), two phases (the 2H stable semiconducting phase and 1T metastable metallic phase of TMDs), and four adsorbed alkali metals (Li, Na, K, and Rb). The input features were the DFT-calculated lowest unoccupied states E_{LUS} and the cohesive energy E_{coh}, and the ionization energy of the adsorbate E_{IE} from the literature. The output target was the DFT-calculated adsorption energy. A linear regression was performed, taking the ordinary least squares (OLS) scheme as implemented in the scikit-learn package, to determine the values of the parameters x_0, a, b, and c in the equation $E_{ads} \approx x_0 + a \cdot E_{LUS} + b \cdot E_{coh} + c \cdot E_{IE}$. The six-fold cross-validation scheme was used, i.e. the data set was randomly divided into six sets, five of each was used for training and one for testing. This procedure was repeated for all six sets and the performance on the test set was averaged and reported. **Fig 7h** shows the OLS-predicted adsorption energies versus the DFT-calculated results. To assess the regression model, the R^2, RMSE, and the MAE were calculated to be 0.968, 0.012 eV and 0.080 eV, respectively, and thus confirmed the validity of the model. The ML-trained parameter for E_{LUS} was 0.974, implying a linear correlation of E_{ads} and E_{LUS}. This result suggests that the lowest unoccupied state energy E_{LUS} can be used as a descriptor and further assist a high-throughput scanning of materials with desired adsorption properties.

4. Strategies for producing 2D materials using machine learning

The two most common methods for producing 2D materials are top-down and bottom-up. Simply, 2D materials can be transferred from their corresponding bulk crystals to target substrates using scotch tape in the top-down method. The transferred 2D materials can then be placed on top of other 2D materials to form 2D heterostructures. Recent deterministic methods using viscoelastic (e.g., Polydimethylsiloxane (PDMS)) or sacrificial polymers (e.g., polymethyl methacrylate (PMMA)) as carrying layers place the target 2D materials precisely on the designated locations on a substrate using control apparatus equipped with a motorized stage and an optical microscope. Such methods have opened a facile pathway to fabricate various types of 2D heterostructures, but the use of heterostructures produced via these methods is still limited to research, because their production requires extensive pre-processing time to identify the optimum 2D materials (e.g., ideal size and thickness) to be transferred. To expand the usability, this tedious pre-process should be automated, and the latest studies using ML give hints for coping with the issue. Another top-down method is liquid exfoliation, which delaminates the monolayers of 2D materials from their corresponding bulk crystals using etching or ion intercalation to weaken the van der Waals forces and expand interlayer spacing. Even though these methods have successfully synthesized many TMDs and some MXenes, numerous newly proposed 2D materials have yet to be synthesized. Liquid exfoliation also requires lots of time to find and transfer optimal 2D materials with the ideal size and thickness.

With bottom-up approaches such as chemical vapor transport (CVT) and chemical vapor deposition (CVD), precursor molecules are supplied to a heating tube and 2D materials are synthesized as a result of chemical reactions. Recently, these synthesis approaches combined with theoretical calculations have been efficiently performed by reducing potentially fruitless attempts

to synthesize new 2D materials. The theoretical calculations can screen out a large number of 2D materials that are less likely to be synthesized by investigating the critical materials' features, such as lattice parameters, formation energies, and cohesive energies that affect the stability of 2D materials in the ambient. A recent described in Section 4.1 correlates the critical features of such materials with the synthesizability and successfully predicts the most synthesizable 2D materials rapidly using ML, which could be an essential strategy in producing new 2D materials.

4.1. ML-enabled automatic identification of exfoliated 2D materials

The most critical weakness of the current mechanical exfoliation and transfer method is the lack of controllability in the size, thickness, and location of transferred 2D materials. Even though several previous studies demonstrated that pre-patterned 2D materials stamps or the use of adhesion layers (e.g., Au, Ni)[36] can control the size and thickness, respectively, of 2D materials after transferring, they are still premature and have a low rate of success. Therefore, identifying 2D materials with the optimal size and thickness among numerous other randomly distributed 2D materials transferred together is still required before moving on to the next step of the study of 2D materials. The identification process performed under the optical microscope is usually tedious and time-consuming, but strongly required to design functional devices with the desired size and thickness of 2D materials or heterostructures.

While observed under the optical microscope, 2D materials transferred on an Si/SiO_2 substrate have different colors depending on their layer thickness, which results from the thin-film optical interference. Once the incident visible light to the 2D materials on the substrate reflects at interfaces (e.g., air/2D materials, 2D materials/SiO2), reflected lights produce constructive or destructive interference depending on their phase (i.e., optical path) difference modulated by the

different thickness of the 2D materials. As the 2D materials get thinner, reflected lights with shorter wavelengths corresponding from yellowish to greenish tend to sequentially generate constructive interference (i.e., Bragg's law), resulting in a color code used to estimate the thickness of 2D materials. Recent studies have demonstrated that ML can precisely and promptly distinguish such a subtle color change and thus automatically identify the thickness and size of 2D materials with high accuracy and reliability.[7, 8, 11]

Fig. 8a shows the use of the K-means clustering algorithm to identify mechanically transferred graphene. At first, an OM of graphene was analyzed using image processing tools such as MATLAB, Python, and ImageJ, followed by extracting color features (R, G, B) from all pixels comprising the OM. Afterward, the K-means algorithm clustered the color features into sub-groups according to their mutual similarity. In **Fig. 8a**, the color features were grouped into four clusters and the thickness value was subsequently assigned to each cluster using the AFM measurement. The labeled clusters were used to determine the thickness of graphene very quickly.

A similar approach was used to identify the thickness of MoS_2, as shown in **Fig. 8b**.[8] The 3D plot shows the distribution of the color features obtained from the OM of MoS_2, displaying wide-spreading features with a rod shape. Based on the shape of the distribution of features, it is necessary to consider other ML algorithms, because the K-means algorithm performs best when the distribution of features have a round shape with roughly equal sizes/density clusters. In this study, SVM was applied to classify color features into sub-groups implying different thicknesses. The color features of the OM were used as a feature vector, x = (R, G, B), and the thickness of the MoS_2 investigated using AFM and Raman was added to the vector as labels. Finally, a training data set, (x, y) = (R, G, B, y), was acquired and used to train the SVM classification model.

The performance of the classification models explained above depends on the quality of the OM. In other words, OMs for training and testing should be taken under consistent conditions such as optical contrast, color temperature, and balance to acquire reliable classification results with high accuracy. Such requirements need an optical microscope well-suited for the classification models, which results in low accessibility. This limitation was resolved in a recent study employing a DNN. **Fig. 8c** illustrates how a DNN works to identify 2D materials that possess mono- or bi-layer thicknesses. In this study, 24 images for MoS_2 were initially obtained, which then were increased to 960 images by augmentation processes such as randomly cropping, rotating, changing color, and changing the HSV (hue, saturation, value) from the original images. The augmentation process is expected to impart the DNN with high robustness in variations on input images, thereby improving the generalizability and increasing the accessibility of the model. The DNN learned by using cross-entropy, softmax, and stochastic gradient descent (SGD) as loss, activation functions, and solver, respectively. **Fig. 8d** shows a segmented image using the trained DNN. The DNN was trained for solving a multi-classification problem that classifies 2D materials shown in the image into monolayer, bilayer, or nothing. It was reported that the optimized DNN algorithm could distinguish mono- and bi-layer MoS_2 from bulk MoS_2 and graphene with an accuracy of 70 ~ 80%.

As another top-down approach, liquid exfoliation holds great promise to realize the industrialization of various 2D materials in the form of dispersions in solution. For example, 2D materials dispersions can be applied to high throughput manufacturing technologies such as spin and spray coating, and inkjet printing, possibly enabling mass production of 2D materials. However, the most significant challenge in such dispersions is to control the quality, guaranteeing the industrially-required ratio of successfully exfoliated to un- or partially exfoliated sheets. In a

recent study shown in **Fig. 8e** and **f**, K-means clustering combined with advanced optical microscopy, such as quantitative polarized light microscopy (qPOM), has efficiently evaluated the quality of graphene-based dispersions.[37] **Fig. 8e** shows the clustering process of graphene oxides, unexfoliated and partially exfoliated graphite oxides in a dispersion after the liquid exfoliation process. Two optical parameters, brightness and retardance, extracted from brightfield microscopy and qPOM, respectively, were used as principal components for ML datasets. It is noted that the retardance ($R=\Delta n \times t$, where Δn is a birefringence and t is the layer thickness) derived from the anisotropy in the refractive indices along the in-plane and out-of-plane directions of 2D materials, is strongly related to the layer thickness. Such brightness and retardance at each pixel in the original image were normalized to the range between 0 and 1, clustered into three sub-groups using the K-means algorithm. The optimal number of clusters (i.e., hyperparameter, K) was determined by the aforementioned gap-statistics, i.e. finding a K value that results in the largest gap between within-cluster distance curves both obtained from the original dataset and a reference dataset distributed with no apparent clustering. Finally, **Fig. 8f** shows the quantified fraction of GO sheets (i.e., successfully exfoliated) from the dispersion. Within 30 minutes, K-means algorithm along with qPOM quantified unexfoliated graphite oxide (uGtO, 13.9%), partially exfoliated graphite oxide (pGtO, 13.6%), and graphene oxide (GO, 72.6%) in the dispersion, which can significantly advance the evaluation process of 2D materials dispersions manufactured by the liquid exfoliation.

The identification of optimal 2D materials occupies a substantial fraction of the current top-down methods such as mechanical transfer and liquid exfoliation processes. Therefore, the strategies for replacing manual laboratory work with an ML-based system could significantly enhance the efficiency and yield in producing 2D materials. Furthermore, the ML approaches

described in this section could be combined with a modern robotic system, resulting in a fully automated identification and transferring system for producing 2D materials and heterostructures.

4.2. ML-enabled prediction of the synthesizability of 2D materials

Even though countless 2D materials have been predicted to exist, only very few have been demonstrated experimentally due not only to technical difficulties but also to fundamental limitations. Such limitations originate from the fact that 2D materials which have been predicted to exist may not always be synthesizable, which renders many experimental efforts unsuccessful. For example, only approximately 20 out of many MXenes whose existence had been predicted by theoretical (e.g., DFT, MD) calculations have been successfully synthesized.[1] To cope with this problem, ML can be used to predict synthesizable candidates among numerous 2D materials, which could be an efficient strategy for saving resources by minimizing trials and errors for the synthesis. In a recent study, synthesizable MXenes were predicted in the order of the highest probabilities of being synthesized using ML.[13] From the perspective of ML, there was a small amount of positively labeled data (i.e., already proved to be synthesized) and a large amount of unlabeled data (i.e., MXenes to be tested for their synthesizability), which produced imbalanced data that required an advanced ML algorithm. In this study, the positive and unlabeled (PU) learning algorithm in semisupervised learning was adopted and trained to tackle such an imbalance.

Fig. 9a shows schematics of material search space and the PU learning algorithm. Considering 11 transition metal M, 12 A group elements, two X (carbon or nitrogen), and n = 1, 2, or 3 (number of layers of X), a total of 792 MAX and 66 MXenes were considered as the initial materials search space. MAX, a bulk phase of MXene, was also added to the search space because the

corresponding MXene could potentially be produced from MAX via the liquid exfoliation process as long as there is synthesizable MAX. Therefore, synthesizable MAX and MXene were independently predicted among the search space in this study.

Input features including structural, thermodynamic, electronic, and elemental information were gathered from DFT simulations and materials databases to train the ML algorithm. In the process of predicting synthesizable MXenes, a total of 66 MXenes, including 10 experimentally synthesized MXenes such as Hf_3C_2, Mo_2C, and Ti_2C (true positives), and 56 unlabeled MXenes were applied to the PU learning algorithm using the bootstrapping method. Bootstrapping is typically used for augmenting the original dataset by sampling with replacement. For example, k-times repeated bootstrapping produces k number of bootstrapped datasets that complement the deficiency of the original dataset. Such augmented datasets help avoid overfitting from insufficient original data and enhance the stability of the ML algorithm. Before applying the PU algorithm, the bootstrapped data were divided into training sets (90%) and test sets (10%). In **Fig. 9a**, the PU algorithm first randomly chooses some of the unlabeled data (blue squares) and labels them as "negative" (green squares, not synthesizable). Subsequently, a classification algorithm finds a hyperplane that classifies the dataset as positive (red circles, synthesizable) or negative. Thereafter, the trained classifier determines if individual unlabeled data not chosen in the first step belongs to "positive" or "negative". Such processes are repeated for all bootstrapped datasets, which results in k-times repetitions. In this study, a decision tree was used as a classifier and a "synthesizability score" was defined to sort MXenes in the order of high synthesizability. The synthesizability score can be described as the ratio of the number of times an unlabeled MXene is classified as positive out of k-repetitions. Specifically, k-repetitions of the PU algorithm with k-augmented datasets result in a k-trained decision tree. If t-number of decision trees out of k classify a given unlabeled

MXene as "positive", the synthesizability score of the specific MXene is evaluated as r/k. In the study, 18 out of 56 unlabeled MXenes were predicted as synthesizable because their high synthesizability score exceeded 0.5. With the same approach applied to MAX, 111 out of 729 unlabeled MAX were suggested as synthesizable.

Fig. 9b shows the probability of synthesizing MAX and MX with respect to transition metals M and A group elements. It was predicted that MAX comprised of Zr, Ti for M and Al, and Ga for A exhibits the highest possibility of being synthesized, and MX based on Hf and C as M and X, respectively, displays the most promise. Since most MXenes are synthesized by etching the A layer from the corresponding MAX, synthesizability predicted separately for MAX and MXene may not necessarily guarantee the true synthesizability. For example, although a MAX is predicted to be synthesizable, it could not produce the corresponding MXene if it requires high etching energy. Similarly, an MXene predicted to be synthesizable could not be synthesized if there is no synthesizable corresponding MAX, and thus it is important to investigate the synthesizability of MAX and MX pairs as combined synthesizability. **Fig. 9c** illustrates a schematic of the process to evaluate the combined synthesizability of MAX and MX pairs. First, the synthesizability scores of 111 pairs of MAX and the corresponding MX were obtained from the PU learning algorithm, and then the etching energy (E_{etch}) of 111 MAX was calculated. Subsequently, the 111 MAX and MX pairs were plotted as functions of the synthesizability of MAX and MX and the etching energy. Finally, the k-means algorithm was used to cluster 111 pairs and found the top 20 pairs with the highest combined synthesizability scores, as shown in **Fig. 9d**. These MAX/MX pairs include Zr_2GaC/Zr_2C, Nb_3AlC_2/Nb_3C_2, and Ti_4AsC_3/Ti_4C_3, which have yet to be synthesized but the outlook is promising.

Such predictions of synthesizability using ML could be an essential pre-process for efficiently synthesizing numerous new 2D materials, because the ML algorithm can filter out only those 2D materials likely to be synthesized successfully, thereby accelerating the synthesis process. Furthermore, the ML combined with DFT calculations can help understand the fundamentals in the synthesis of 2D materials by revealing input features most relevant to the synthesizability.

5. ML-assisted applications of 2D materials

Due to their excellent properties, 2D materials have had a significant impact on applications such as transistors, optoelectronics, sensors, and catalysts. Recent studies have shown that ML can be an effective tool for studying such applications[14, 15, 17, 38, 39] because it can determine the complex connectivity and relationships between numerous data and draw meaningful results beyond human intuition. The studies introduced in the following section show how ML can be used for applications based on 2D materials.

5.1. ML-enabled application of 2D materials in chemical sensing

2D materials have pushed the boundary of detection in sensing applications because their excellent structural, electrical, optical, and electrochemical properties enable them to outperform conventional 3D sensing materials. Specifically, the large surface area, tunable band gap, high electron mobility, electrically low-noise, long-lived plasmons, high stability, and low toxicity of 2D materials have been exploited to design electrochemical, electrical, and optical-based sensing schemes.[40] In an electrochemical sensing scheme, 2D materials are designed for a working electrode (WE) of electrochemical sensors that measure the change in the Faradaic current (using amperometry or voltammetry) or interfacial impedance (using electrochemical impedance

spectroscopy (EIS)) by reduction-oxidation (redox) reactions upon the adsorption of target analytes to the WE. Field-effect transistors (FETs) made of 2D materials are used for the sensing channels in electrical sensors to assess the change in channel resistance by the gate potential modulation due to the binding reactions between the analytes and receptors grafted on the surface of the 2D material. Additionally, optical systems such as surface plasmon resonance (SPR) and surface-enhanced Raman spectroscopy (SERS) measure the change in the local refractive index resulting from the adsorption of analytes on the sensor surface made of 2D materials. Sensors consisting of 2D materials as a sensing membrane show unprecedented detection capabilities that conventional 3D sensing materials cannot achieve; for example, a single molecule of NO_2 gas was detected using a graphene FET in 2007. Other 2D materials such as MoS_2, phosphorene, and $Ti_3C_2T_x$ have been demonstrated as sensitive bio- and environmental sensors, exhibiting a low signal-to-noise ratio, high sensitivity, and low limit-of-detection.[40]

Recently, new trials adopting ML techniques have been reported for improving the superior potential of 2D materials in sensing applications. For example, the atomically thin layer of 2D materials has enabled the sensitive analysis of the sequence of bases in DNA by detecting corresponding amino acids.[14] This sensing technology, known as nanopore-sequencing, takes advantage of the conductance change while a strand of DNA passes through a nanopore in 2D materials, as shown in **Fig. 10a**. However, the conductance change from the single nanopore of 2D materials is very subtle, measuring in the pico-ampere (pA) range, which requires a well-controlled experimental setup and analysis to discern the sensor signals from background noises. In a recent study, ML was used to identify sensor signals coupled with noises for detecting amino acids using MoS_2 with a nanopore.[14] Once a chain of amino acids travels through a nanopore of MoS_2, the change in ionic current originating from the conductance change is observed, depending

on the type of amino acids comprising the chain. Each amino acid has various functional groups, such as amino (-NH_2) and carboxyl (-COOH), molecular sizes, and weights, which show different interactions with the nanopore and result in varying residence times. Therefore, the ionic current and residence time associated with amino acids can serve as critical sensor signals during the nanopore-sequencing process. In the study, the ionic currents and residence times from 20 standard amino acids moving through a nanopore in MoS_2 were calculated using MD simulation with 100 repetitions. **Fig. 10b** plots the ionic currents and residence times that form scattered clusters. This plot was further used as training data for ML algorithms such as KNN and RF. **Fig. 10c** and **d** show the decision boundaries obtained from the KNN (k = 3) and RF (n-estimator = 9), respectively. Different color regions shown in **Fig. 10c** and **d** are associated with the individual amino acid, enabling the classification of testing data depending on their two features (i.e., ionic current and residence time). These decision boundaries trained by the KNN and RF efficiently classified the testing data with an accuracy of 94.6% and 99.6%, respectively. Furthermore, the trained KNN and RF algorithms were further used to identify a chain of amino acids, reflecting a more realistic sensing problem because amino acids favorably form a chain. **Fig. 10e** shows a chain of 16 amino acids used as test data and a plot of the corresponding ionic current with respect to the residence time. The plot obtained from a chain of amino acids includes a high degree of fluctuations originating from background noises, and thus it is unlikely that the sensor signals corresponding to amino acids can be easily deciphered. ML, given this condition, can be a powerful tool to analyze sensor signals buried in noises. In this study, two characteristic features were extracted by averaging ionic currents and residence times obtained from 10 repetition tests using the same chain of amino acids. Following that, predictions of amino acids were performed by applying such features to trained KNN and RF algorithms. The RF predicted better than the KNN,

and the overall accuracy was 62.5%. It should be noted that only two features were used for training the ML models in this study, which could be one of the reasons for such relatively low accuracy. Therefore, capturing multiple characteristic features related to amino acids could further increase the accuracy.

In another recent study, material databases beneficial to construct materials data for ML were used to discover 2D materials suitable for adsorbing and detecting airborne mercury (Hg^0) through a series of screening processes.[15] Materials databases (pymatgen[41] and AFLOW[42] listed in **Table 2**) and a thermochemical software (FactSage) were used to screen a number of 2D materials, as shown in **Fig. 10f**. First, stable TMDs were investigated using Pymatgen. Subsequently, easily synthesizable TMDs selected from previously chosen TMDs were identified using FactSage. **Fig. 10g** shows phase diagrams of WS_2 obtained from pymatgen and FactSage. The red dots in the pymatgen phase diagram show the stable compounds that can be obtained from the combinations between transition metals and chalcogens, and WS_2 was found as a stable TMD. Moreover, the FactSage diagram displays synthesizable chemical compounds from precursors. For example, type-I (pure TMD, WS_2) occupies a large portion of the diagram, indicating that 2D WS_2 is likely to be synthesized compared with other partial TMDs (from type-II to VI, W_xO_y-H_2S-H_2). In the last screening, the AFLOW database was used to confirm the atomic structure of previously screened TMDs and discover only 2D TMDs. Throughout such screening processes, TiS_2, NiS_2, ZrS_2, MoS_2, PdS_2, and WS_2 survived as promising TMDs for Hg^0 detection. Finally, DFT calculations confirmed that PdS_2 is the most suitable TMD because of its high Hg-uptake capacity and high charge density change under Hg adsorption. This study used open-source online databases that provide ML-based predictions to discover the best TMDs for the Hg sensor, which is beneficial for designing sensors with maximum sensing performance. Additionally, with the use

of the databases, this study could be carried out with minimal knowledge of ML and thus can serve as an excellent example for novice researchers.

The studies introduced in this section show that ML and material databases can be used to obtain meaningful findings from noisy sensing signals and find optimal 2D sensing channels for detecting a specific target analyte, thus improving the resolution and sensitivity of the 2D materials-based sensor. ML could also be considered for calibrating 2D materials-based sensors and compensating for the drift of sensor signals. A brief discussion on potential research ideas on these topics is e provided in Section 6.

5.2. ML-enabled application of 2D materials in catalysis

As the worldwide demand for energy continues to rise, the exploration of electrocatalysis, such as the hydrogen evolution reaction (HER), oxygen evolution reactions (OER), and nitrogen reduction reaction (NRR), is flourishing as it plays a central role in clean, effective, and sustainable energy conversion. Electrocatalysis has even been accelerated by 2D materials and single-metal-atom doping, as the former has a large surface area for reaction and the latter introduces more active-sites. However, as an experimental approach, electrocatalysis is time-consuming and expensive. A recent work has used ML to rapidly and accurately screen out excellent HER catalysts from MBenes and MXenes.[17] In this work, bare MBenes and MXenes were first compared to understand their differences in HER activity, then bare and single-atom doped MBenes were investigated extensively to predict the ideal HER catalysts. In both cases, simple structural and elemental descriptors were used as input features in the ML model. These descriptors can be grouped as DFT-calculated and elemental. As a major indicator of the catalytic activity, the Gibbs free energy of hydrogen adsorption ΔG_{H^*}[43] (with optimal value of ~0 eV) was calculated by DFT

and used as ML output targets. For bare MBenes and MXenes, 66 MXenes and 46 MBenes were geometrically optimized and used as the dataset. **Fig. 11a** shows example structures. A simple linear regression algorithm was applied in the ML model. **Fig. 11b** shows descriptor-predicted and ΔG_{H^*}. It was found that MBenes have much better HER performance than MXenes, as the ranges of ΔG_{H^*} were -0.4 to 0.4 eV for the former and −1.2 to −0.5 eV for the latter, respectively. Pd_2B_1 and Co_2B_2 were selected as potentially promising HER catalysts with ΔG_{H^*} of only −0.04 and −0.05 eV. Next, bare and single-atom doped MBenes were explored, since single-metal-atom doping may introduce more active-sites and improve HER performance. A workflow for the ML screening process of combined bare and single-atom doped MBenes is illustrated in **Fig. 11c.** A dataset containing 180 MBenes was generated. It started from the combinations of 19 bare MBenes, 23 metal elements of dopant, and n = 1, 2 of the layer ratio in M_2B_n, yielding 874 potential candidates. Then, 70 candidates were randomly selected and combined with 110 pre-existing candidates to generate a diverse dataset.

The dataset was divided, randomly placing 75% of the candidates in a training set and 25% in a testing set. Four ML algorithms were employed and compared to predict ΔG_{H^*}: least absolute shrinkage and selection operator (LASSO), random forest (RFR), kernel ridge (KRR) and support vector (SVR) regression. LASSO, KRR, and SVR paradoxically exhibit predictions with lower RMSE for the testing set than for the training set. **Fig. 11d** shows that the performance of SVR was quite good, with RMSE and R^2 of 0.12/0.09 eV and 0.85/0.91 for training/testing data. Considering the distribution of RMSE and R^2 values from more than 100 random trials, LASSO and SVR show better prediction of ΔG_{H^*} through the preliminary evaluation. To accelerate the training process, the correlated features, especially the DFT-calculated features, which are relatively expensive, should be removed to further speed up the predicted process. Using the

Pearson correlation coefficient (PCC), reducing the number of the features from 21 to 16 led to a 0.02 eV increase of the testing RMSE for LASSO, while the RMSE of SVR decreased by 0.01 eV. Thus, SVR is the best of the four models in this work. Then, an additional 25 single-atom doped MBenes were randomly selected and trained by the SVR model using 13 simple features. The value of RMSE is 0.15 eV, as shown in **Fig. 11e**. Based on the criteria that ΔG_{H^*} should be in the range of −0.25 to 0.25 eV and the cohesive and substitution energies should be less than −5.02 (MoS_2) and 0 eV, respectively, five MBenes (Co/Ni_2B_2, Pt/Ni_2B_2, Co_2B_2, Os/Co_2B_2 and Mn/Co_2B_2) were determined using the SVR model and DFT calculations to be promising HER catalysts among 205 MBenes and 66 MXenes, as shown in **Fig. 11f**. To ensure accurate screening, accurate DFT calculations, including spin-polarization, vdW-interaction, and PBE+U, were calculated for the final candidates. The changes in the predicted ΔG_{H^*} were small, indicating that reliable screening can be obtained by ML models trained on less accurate DFT calculations. Furthermore, the phonon dispersion curves and dynamic stability were calculated and considered. Finally, Co_2B_2 and Mn/Co_2B_2 were predicted as excellent HER catalysts, with $|\Delta G_{H^*}| < 0.15$ eV among bare and single-atom doped MBenes.

2D transition metal dichalcogenides (TMDs) have been reported as very promising catalysts, but the performance of intrinsic TMDs for electrocatalysis processes such as water splitting is inadequate.[44] However, different 2D TMDs can be stacked to form heterojunction materials with novel properties. A recent study has applied the LASSO ML approach to predict the vertical stacking heterostructures of 2D TMDs as bifunctional electrocatalysts for HER and OER.[39] In consideration of the stacking rotation angles, the study predicted that $MoTe_2/WTe_2$ with a rotation of 300° is the best electrocatalyst for water splitting, exhibiting an overpotential of 0.03 V for HER and 0.17 V for OER, respectively. The catalytic performance can be estimated through reaction

free energy ΔG. The mechanism for HER is $H^+ + e^- + * \rightarrow *H$, whereas for OER four steps are involved:

$H_2O(g) + * \rightarrow *OH + H^+ + e^-$

$*OH \rightarrow *O + H^+ + e^-$

$*O + H_2O(g) \rightarrow *OOH + H^+ + e^-$

$*OOH \rightarrow * + O_2 + H^+ + e^-$

The ideal HER catalyst should have ΔG_{*H} near 0 eV, while the ideal OER catalyst should have similar reaction free energies in those four charge transfer steps at zero potential (4.92 eV/4 = 1.23 eV). In this work, the overpotential of HER η_{HER} and OER η_{OER} were calculated to estimate the catalytic performance, where η_{HER} is $|\Delta G_{*H}|/e$ for HER, and η_{OER} is determined by the potential limiting step as $\eta_{OER} = \Delta G_{max}/e - 1.23$. For the ML approach, 48 systems were optimized, which were constructed by combining eight heterostructures (MoS_2/WS_2, $MoSe_2/WSe_2$, MoS_2/WSe_2, $MoSe_2/WS_2$, $MoTe_2/WTe_2$, MoS_2/WTe_2, $MoTe_2/WS_2$, and $MoTe_2/WSe_2$), and six rotation angles (0°, 60°, 120°, 180°, 240°, and 300°). The input variables were the rotational angle θ, the distance d between two TMDs, the average bond length ℓ, and the ratio (λ) of the bandgaps of two component materials. Output targets were the reaction overpotential η_{HER} and η_{OER}. First, 257,703 possible descriptors were generated by combining one or more input variables through operations including addition, subtraction, multiplication, division, absolute value, square, and square root. A linear regression LASSO algorithm was applied and repeated 50 times. The best-fit descriptor expression PL(λ, θ, d, ℓ) was selected. For OER, a good linear relationship between the best-fit descriptor PL (x) and the catalytic performance (y) was found to be $y = 1.04x + 0.6$ with an R^2 of 0.83, as shown in **Fig. 12a**. For HER, the same descriptor PL led to $y = -1.73x + 0.18$ with an R^2 of 0.80. Then, these equations from the LASSO regression were used to predict the overpotentials

η_{HER} and η_{OER}. **Fig. 12b** and **c** shows the relationships of the rotational angle and the overpotentials. It was concluded that $MoTe_2/WTe_2$ with a rotational angle of 300° had the best overall performance for HER and OER, with an overpotential of 0.03 V for HER and 0.17 V for OER.

The nitrogen reduction reaction (NRR) on the transition metals (TMs) is promising, but the efficiency was low in most cases. Recently, boron(B)-doped graphene (B-Gr) exhibited a higher efficiency of NRR than most TMs.[45] A recent study has designed a DNN to predict efficient electrocatalysts for NRR among B-doped graphene single-atom catalysts (SACs).[38] Three candidates were selected as very promising catalysts for NRR, especially B-Gr, with CrB_3C exhibiting a minimal overpotential of 0.13 V. In this work, the metric for a good NRR catalyst is determined by the adsorption energy ΔE_{N2} of N_2 and the hydrogenation free energy ΔG for each reaction step, as $\Delta E_{N2} < -0.50$ eV, $\Delta G_{N2\text{-}N2H} < 0.55$ eV, and $\Delta G_{NH2\text{-}NH3} < 0.7$ eV. To construct the ML dataset, 182 structures of B-doped graphene with single-metal atoms were considered by combining 26 transition metals and seven different types of coordination in single-atom catalysts (SACs), as shown in **Fig. 12d**. **Fig. 12e** shows the designed DNN model through the Keras library. A Coulomb matrix with components of the atomic number position was used as an input descriptor for representing atomic structures. PCA was used to reduce the dimensions of the Coulomb matrix into one axis (PC1). The adsorption energies and free energies of some intermediate steps involved in the NRR were predicted by using the Light Gradient Boosting Machine (LGBM) model, as shown in **Fig. 12f**. The output of the DNN is the probability of efficient catalysts. After screening, three B-Gr SACs were proposed to be very promising for NRR: B-Gr with CrB_3C, TcB_3C, and $HfBC_2$.

6. Conclusion and outlook

Over the last decade, the rapidly growing number of 2D materials and their heterostructures have surpassed the capacity that conventional experimental and computational approaches can handle. In recent years, ML has been on the rise as a powerful tool to support such conventional methods, thus bringing new opportunities to study them in intelligent ways. Harvested from materials databases and experimental and computational observations, the characteristics of 2D materials serve as input features to train various ML algorithms belonging to supervised, unsupervised, and semisupervised learning approaches. By understanding the intricate interrelationships among input features or correlating input features with target outputs, such trained ML algorithms result in new insights from accurate predictions, enabling the understanding, discovery, and synthesis of 2D materials.

This tutorial review has introduced recent efforts that seek to understand how ML can contribute to the study of 2D materials. The early and frequent adoption of ML is for predicting their properties. ML algorithms have quickly and accurately predicted the properties of numerous undiscovered 2D materials and heterostructures which otherwise would have required considerable time and resources. A relatively recent application of ML is for synthesis. ML algorithms, trained by using the features from optical images, are able to identify mechanically transferred 2D materials with the optimal size and thickness. Furthermore, ML predicts their synthesizability, thus significantly enhancing productivity by pre-screening countless candidates that are unlikely to have been synthesized. Finally, ML has been adopted to study the applications of 2D materials, which opens new opportunities such as ultrasensitive sensing and discovering the most competent catalysts. The recent studies performed using ML are organized and listed in **Table 3**.

Yet despite such great opportunities, there are several challenges that must be kept in mind in order to apply ML. First, it is rather difficult to obtain a sufficiently large volume of data to use ML to study newly discovered 2D materials. For example, as described in Section 4.2, only 20 MXenes were used as labeled data to predict the synthesizability of numerous others because MXenes are relatively new whose synthesizability has not been thoroughly investigated. Although advanced ML techniques such as PU learning can be applied to handle such sparse data, a minimal volume of data for the algorithms should be accumulated in advance through experimental or computational methods. Furthermore, such a low volume of data often cannot impartially represent the characteristics of 2D materials, resulting in biased predictions.

Second, ML requires not only a large quantity of but also high quality data to produce accurate predictions. Since ML is a stochastic process, prediction accuracy depends heavily on how well the data is used to train ML algorithms. In the study of 2D materials, the data source is limited to databases, experiments, and computations. Moreover, such data related to the structural, electronic, chemical, and thermal characteristics are usually represented as floating numbers with an error range, and thus suitable data normalization and averaging techniques should be applied to such data for accurate predictions. Moreover, repeated experiments or computationally expensive simulations are required to acquire reliable data with low variances.

Lastly, predictions from ML should be carefully interpreted and validated using experiments or computational simulations, because such predictions do not come from understanding the underlying physics of 2D materials. Instead, ML correlates the input features with target outputs and makes predictions based on those correlations. For example, NN generates complex interconnections between numerous nodes in hidden layers without considering any theoretical backgrounds in 2D materials, which could produce theoretically wrong correlations. However,

such challenges in ML for the study of 2D materials could be resolved in the near future, as researchers worldwide are collaborating to put together vast and accurate 2D materials libraries. Furthermore, state-of-the-art ML studies are incorporating physics-informed constraints into ML algorithms, thus enabling more theoretically reasonable predictions.

Leveraging ML can open new research opportunities. First, an automated system for producing an array of 2D materials and their heterostructures could be designed using ML and robotic technologies. For example, a recent study demonstrated that a robotic arm equipped with a Bernoulli gripper can successfully transfer individual 2D materials onto a target substrate, producing heterostructures.[46] This robotic system could be synergistically integrated with ML algorithms that identify optimal materials (discussed in Section 4.1), resulting in an intelligent production system. Second, in sensing applications, ML can be used for optimizing various sensors made of 2D materials. For example, ML can be used to correct the drift of sensor signals. The drift is a natural process of changing sensor signals with time due to environmental change or aging sensors. Since ML specializes in predicting expected trends based on historical data, it can generate a predicted drift curve that could be subtracted from the original sensor signal to remove the drift. Furthermore, ML could be beneficial for calibrating sensors that exhibit a relatively high device-to-device variation due to the lack of uniform manufacturing methods. By correlating the initial properties of 2D materials with sensor outputs, ML can accurately predict the concentration of target species at a given sensor output, which leads to calibration curves that compensate for device-to-device variations of sensors. Finally, ML could be adopted to design an all-in-one system that includes a series of "Discovery-Understanding-Screening-Synthesis-Application". In this concept, ML first predicts new 2D materials and reveals the materials' properties; among these, a few that have optimal properties for a specific application are screened, and, subsequently, the

synthesizability of them can be evaluated. Finally, applications using the synthesized 2D materials are optimized using ML.

In conclusion, ML has become an essential tool for supporting a series of studies in 2D materials from fundamentals to applications, which significantly accelerates the development of 2D materials and their heterostructures and thus opens numerous opportunities for applying them to more practical and real-world applications.

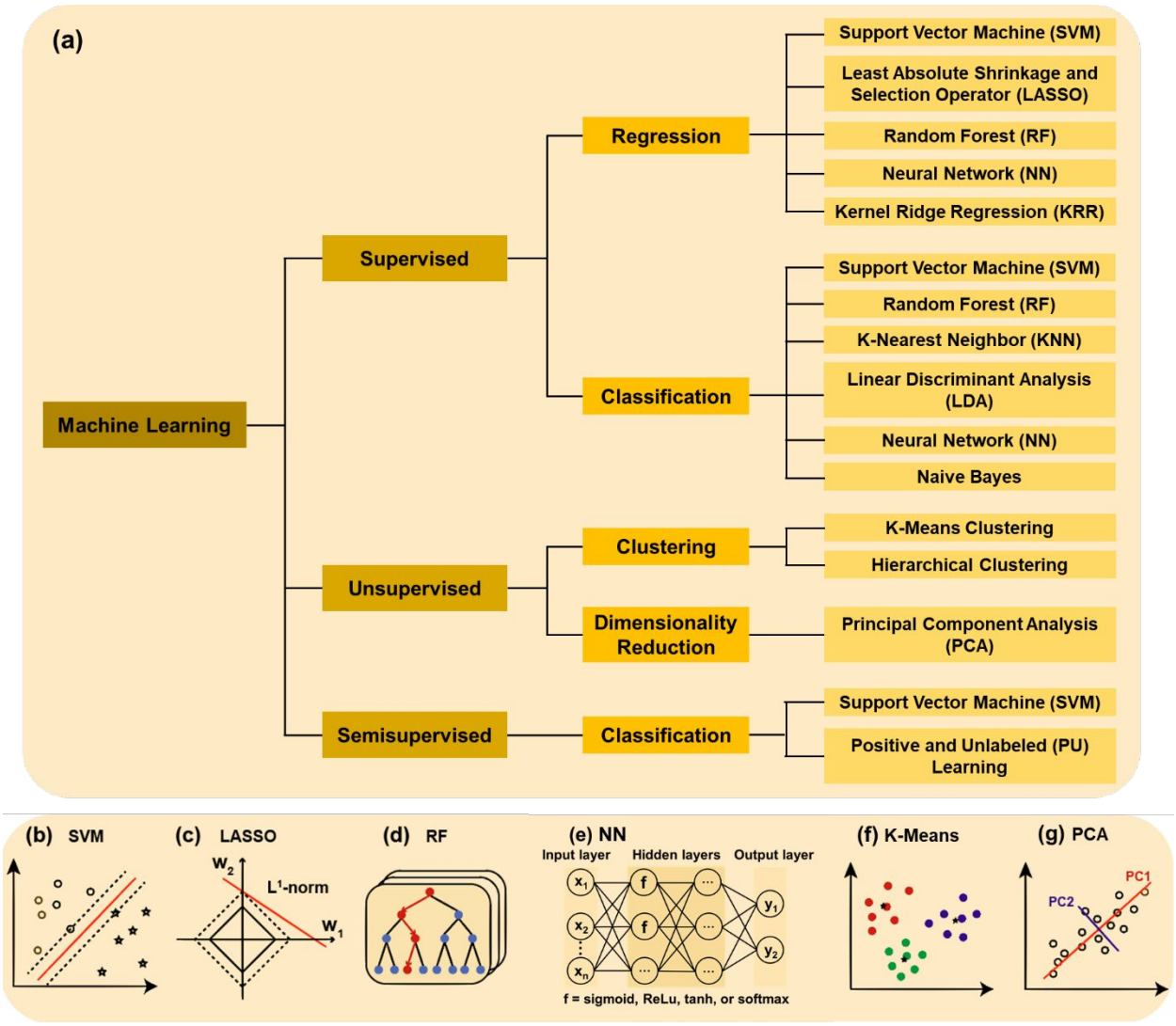

Figure 1. Common ML algorithms for the study of 2D materials. (a) Types of ML. (b-g) Representative ML algorithms including SVM, LASSO, RF, NN, K-Means, and PCA, respectively.

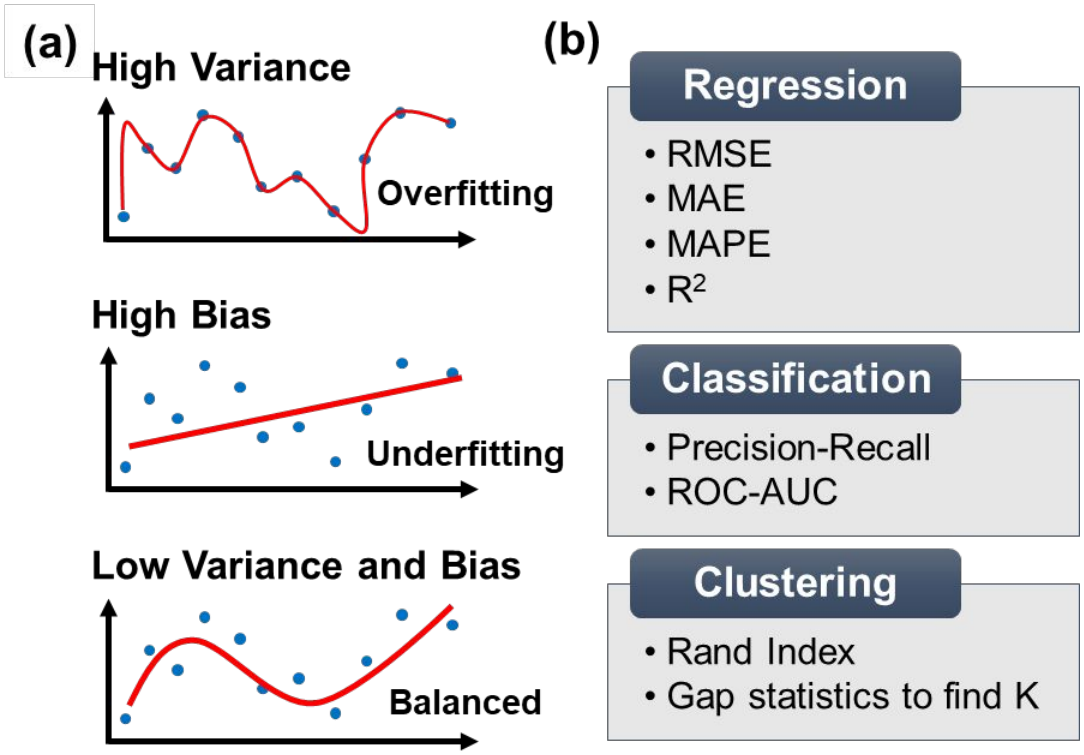

Figure 2. Prediction errors and validation metrics for ML models. (a) Variance and bias errors. (b) Useful evaluation metrics for regression, classification, and clustering models.

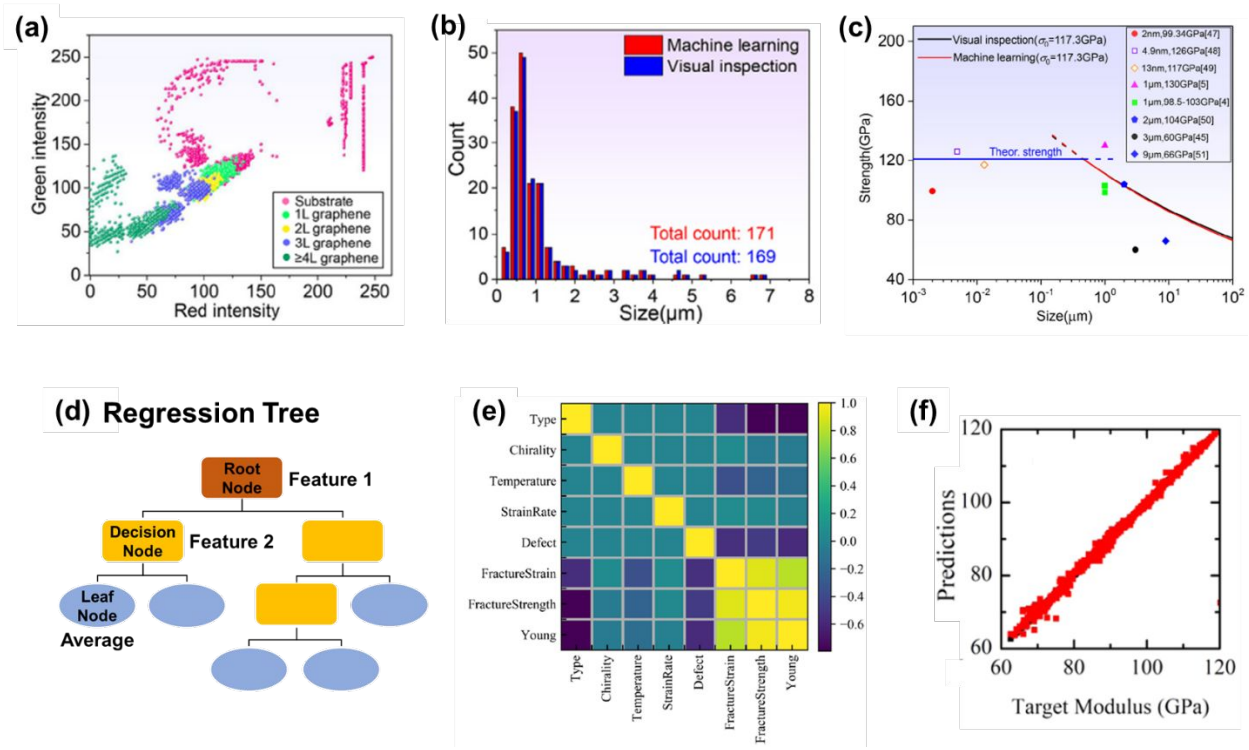

Figure 3. ML-enabled prediction of the mechanical properties of graphene and WS_2. (a) A plot of input features (the intensities of green and red) from the OM of graphene, classified by the SVM. (b) The accuracy of ML-enabled classification of the size of graphene compared with manual inspection. (c) ML-enabled prediction of the strength of graphene as a function of layer size. (d) A schematic of the regression tree. (e) Pearson's correlations between material features and target outputs of WS_2. (f) A comparison plot of Young's moduli obtained from ML prediction vs. the corresponding values from MD simulations. Panels (a), (b), and (c) are adapted from ref. 12 with permission from Elsevier, copyright 2020. Panels (e) and (f) are adapted from ref. 23 with permission from American Chemical Society, copyright 2019.

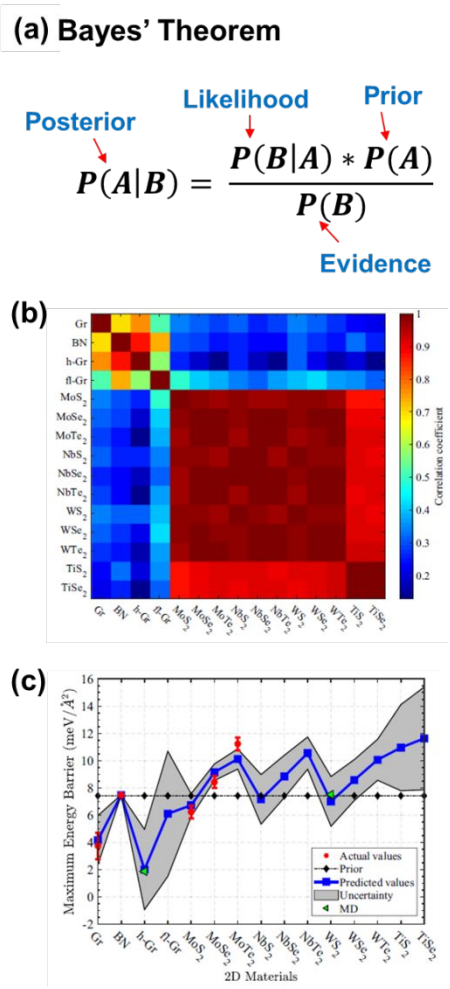

Figure 4. Prediction of the nanoscopic friction of 2D materials. (a) A concept of Bayes' theorem that is the foundation of the Bayesian algorithm. (b) Visualization of the correlation coefficients among 15 different 2D materials. (c) Predicted MEB from 15 2D materials compared with previously reported and MD-calculated values. Panels (b) and (c) are adapted from ref. 24 with permission from Springer, copyright 2020.

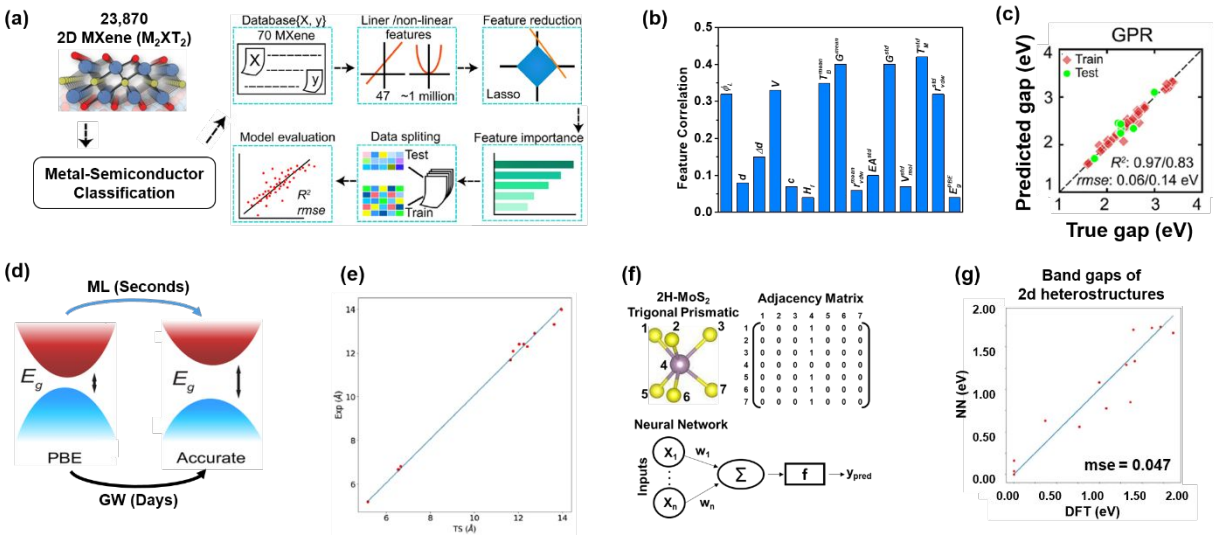

Figure 5. Prediction of the band gaps of 2D MXenes and 2D heterostructures. (a) A schematic of the working flow for ML processes. (b) A feature correlation plot of material features (i.e., input features) extracted from the bagging method. (c) Comparison of the predicted gap (from ML) with the true gap (GW). (d) A diagram showing the advantage of the ML approach over DFT simulation for predicting the band gaps of 2D materials. (e) Comparison between the lattice parameter predicted using the Tkatchenko–Scheffler (TS) vdW functional and experimental values. (f) Schematics of the adjacency matrix for extracting input features and neural networks for predicting the band gaps from 2D heterostructures. (g) A correlation plot between the predicted band gaps (from ML) and the calculated band gaps (from DFT) of 2D heterostructures. Panels (a) - (d) are reproduced from ref. 26 with permission from American Chemical Society, copyright 2018. Panels (e) and (g) are adapted from ref. 18 with permission from Wiley-VCH, copyright 2019.

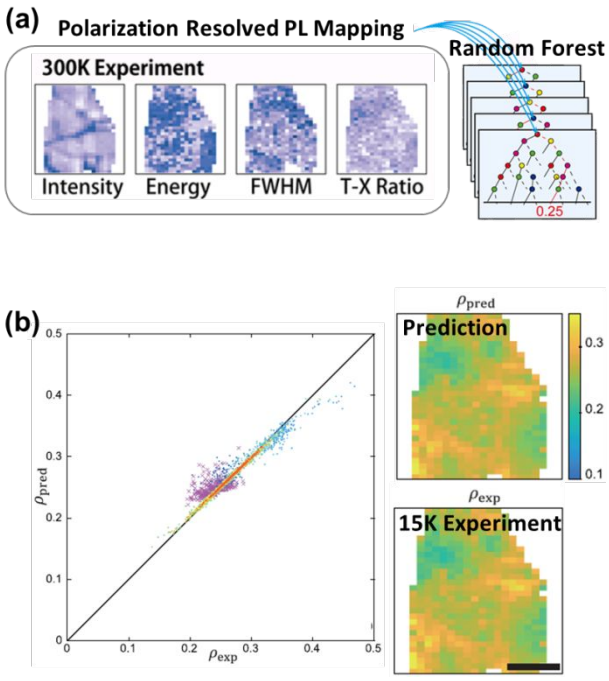

Figure 6. Prediction of the exciton valley polarization of monolayer WSe_2. (a) A schematic of the training process for the RF algorithm. Input features and labels (i.e., target outputs) were obtained from PL spectra performed under 300K and 15K, respectively. (b) Comparison of the RF-predicted exciton valley polarization of WSe_2 with the one directly measured under a low-temperature PL experiment. Panels (a) and (b) are reproduced from ref. 28 with permission from American Chemical Society, copyright 2019.

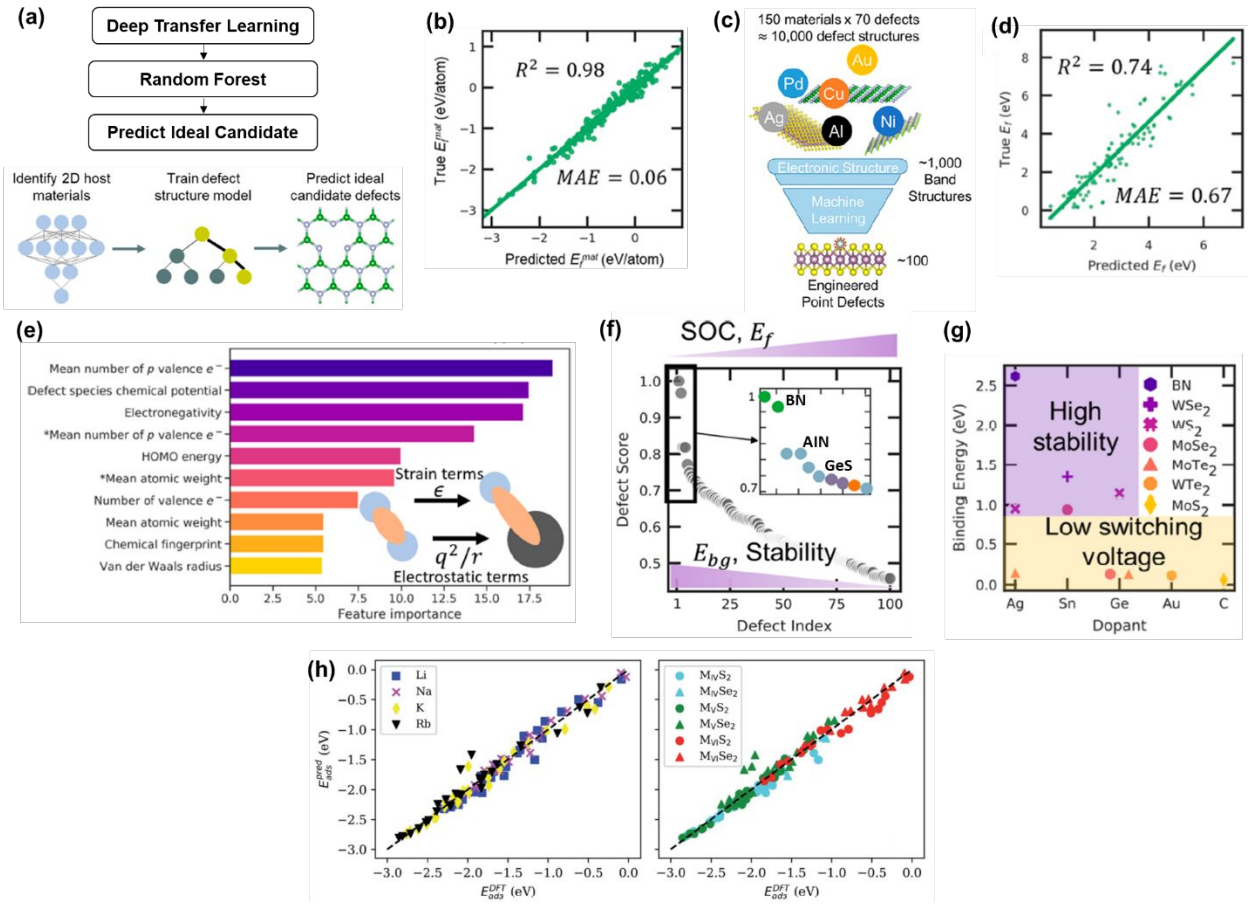

Figure 7. Designing point defects in 2D materials using ML predictions. (a) A schematic of the work flow for the ML process. (b) Comparison of the DFT-calculated true formation energy of 2D host materials with the predicted formation energy obtained by deep transfer learning. (c) A schematic of the dataset generation for ML models. (d) Comparison of the DFT-calculated true formation energy of point defects versus the predicted formation energy. (e) Feature importance for predicting the formation energy. (f) A plot of defect scores sorted in high order for 100 defects. (g) Defects with the highest and lowest maximum binding energy for resistive switching. (h) Predicted vs. DFT-calculated adsorption energies for Li, Na, K, and Rb adsorbed on the TMDs. Panels (a) – (g) are reproduced from ref. 29 with permission from American Chemical Society, copyright 2020. A panel (h) is adapted from ref. 35 with permission from Royal Society of Chemistry, copyright 2020.

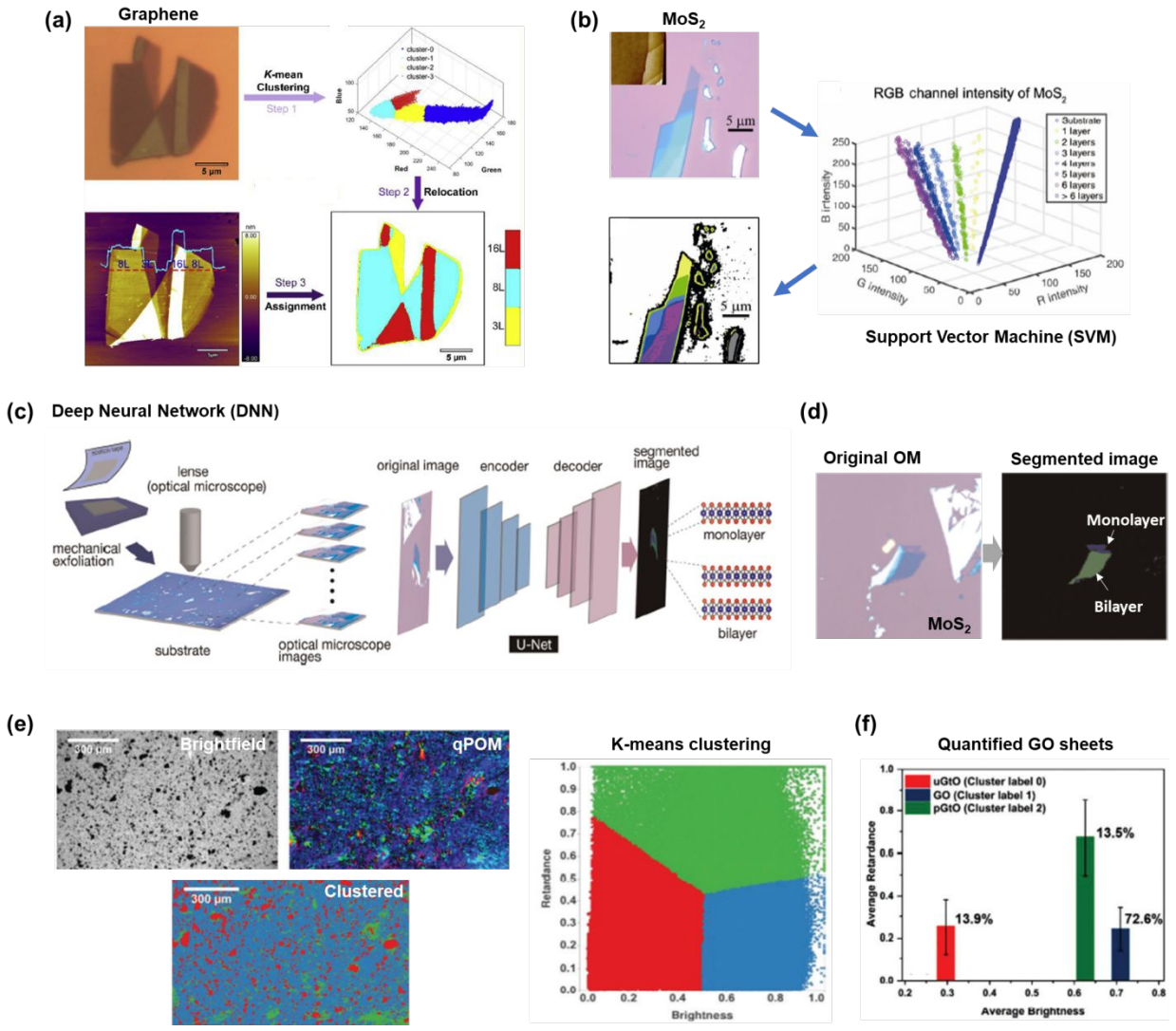

Figure 8. Identification and thickness analysis of 2D materials. (a) Thickness profile of graphene using an OM image and the K-mean clustering algorithm. (b) Identified thickness of MoS_2 flakes using an OM image and the SVM algorithm. (c) Neural network for predicting the thickness of MoS_2 flakes. (d) OM image of MoS_2 flakes and recognized thickness through NN. (e) K-means clustering using brightfield and quantitative polarized optical microscope (qPOM) images for quantifying each type of GO sheets in a dispersion (f) Quantified uGtO, GO, and pGtO in the dispersion. Panels (a) and (b) are adapted from ref. 7 with permission from Elsevier, copyright 2019 and ref. 8 with permission from Springer, copyright 2018, respectively. Panels (c) – (f) are adapted from open-

access articles ref. 11 with permission from Nature Publishing Group, copyright 2019 and ref. 37 with permission from Wiley-VCH, copyright 2020.

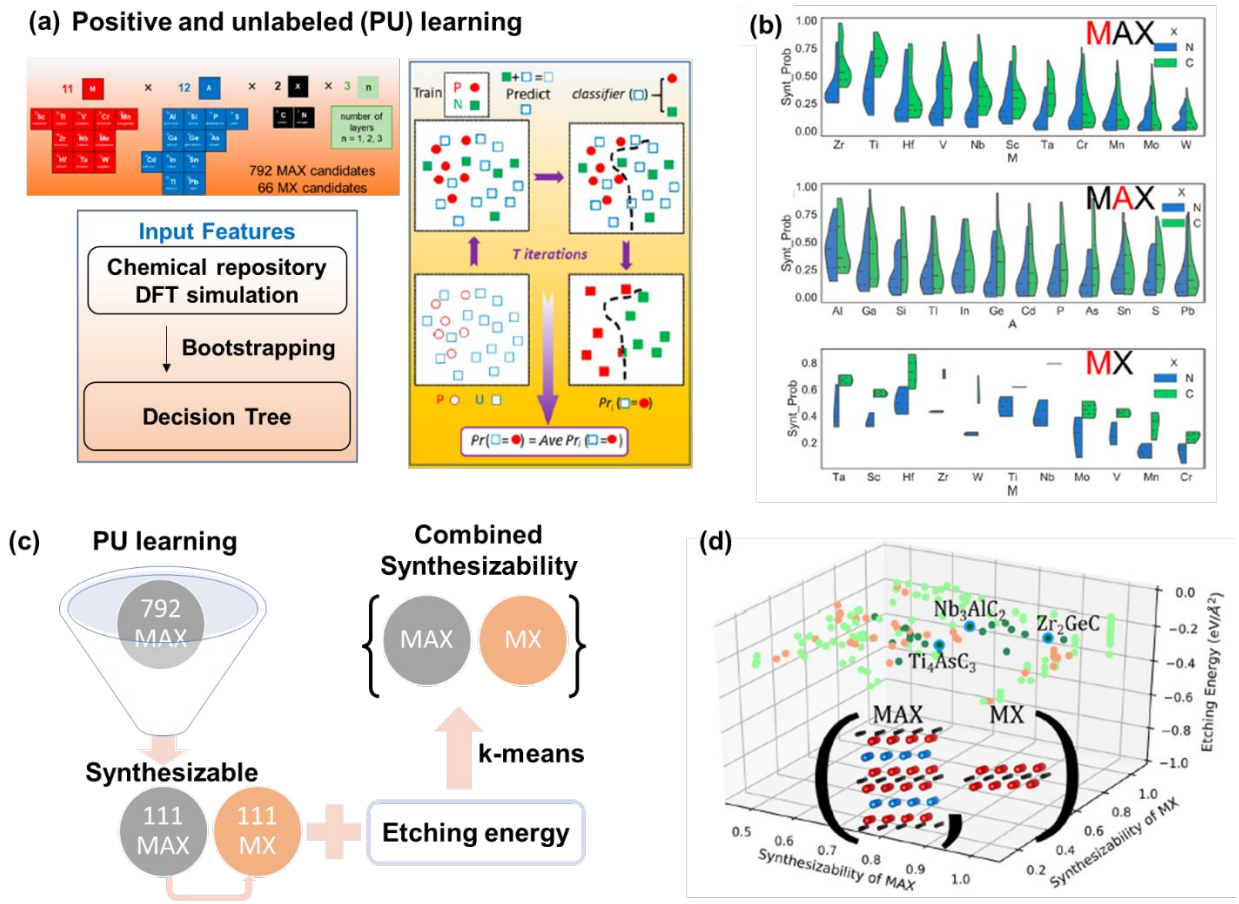

Figure 9. Prediction of the synthesizability of 2D MXenes. (a) Schematics of material search space (792 MAX and 66 MX) and the positive and unlabeled (PU) learning algorithm. (b) Extracted synthesis probability of MAX and MX with respect to composing atomic species. (c) A workflow for finding synthesizable (MAX, MX) pairs. (d) K-means clustering of (MAX, MX) pairs as a function of synthesizability and etching energy. Panels (a), (b), and (d) are reproduced from ref. 13 with permission from American Chemical Society, copyright 2019.

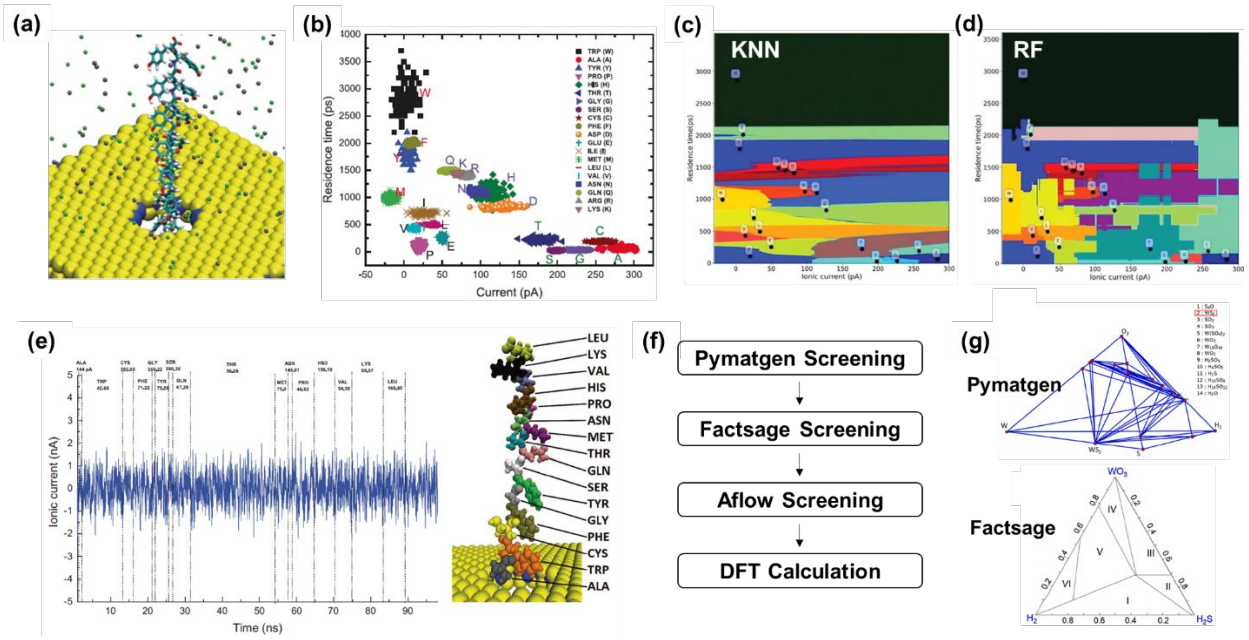

Figure 10. Identification of amino acids with a sensitive nanoporous MoS_2 and ML classifications. (a) A schematic illustration of nanoporous MoS_2 and a chain of amino acids. (b) Ionic currents and residence times obtained from 20 standard amino acids traveling through an MoS_2 nanopore. (c-d) Classified amino acids with respect to current and residence time using KNN and RF, respectively. (e) A plot of ionic current as a function of time, obtained from a chain of amino acids. (f) A series of ML screening steps for discovering optimal 2D materials for Hg sensing. (g) Representative phase diagrams obtained from Pymatgen and Factsage, predicting the stability and synthesizability of WS_2. Panels (a) – (e) are adapted from an open-access article ref. 14 with permission from Nature Publishing Group, copyright 2018. A panel (g) is reproduced from ref. 15 with permission from Elsevier, copyright 2019.

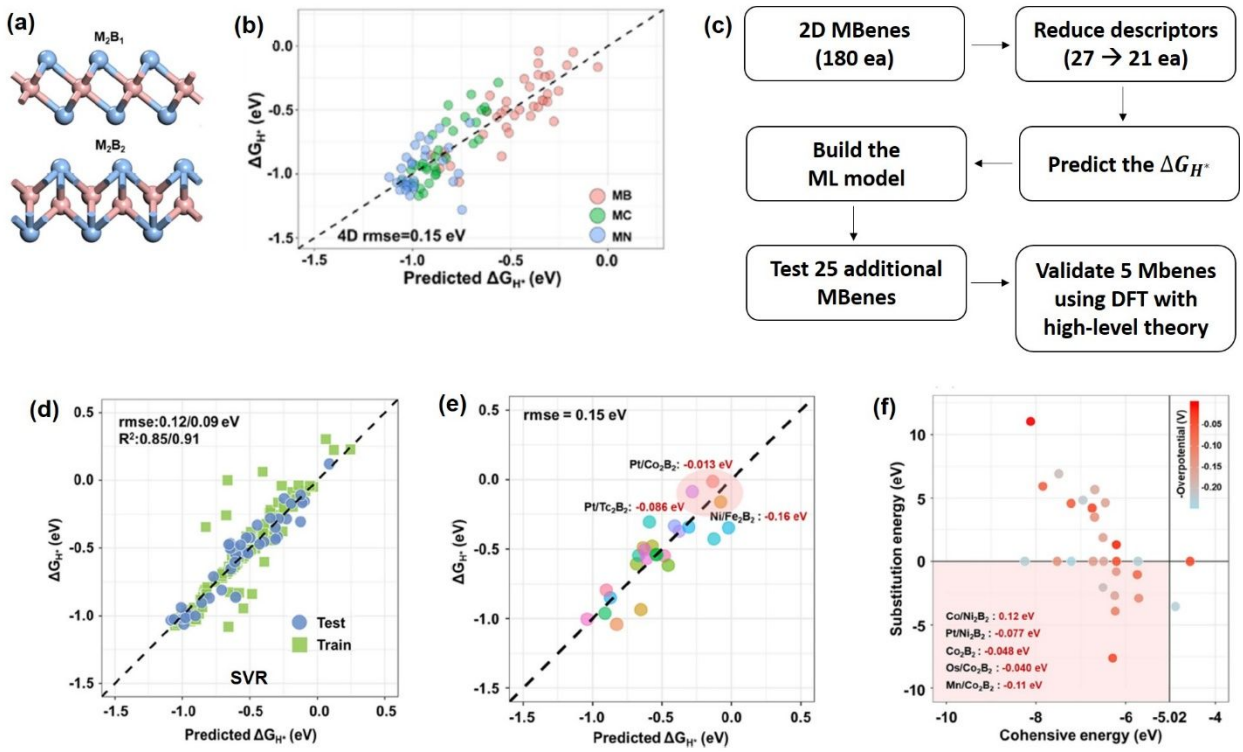

Figure 11. ML-assisted screening of 2D MBenes to find an optimal hydrogen evolution catalyst. (a) The structures of MBenes and MXenes. (b) Comparison of descriptor-predicted and DFT-calculated hydrogen adsorption ΔG_{H^*} for bare MBenes and MXenes. (c) A workflow for the ML screening process. (d) Comparison of predicted Gibbs free energy of ΔG_{H^*} with DFT-calculated ΔG_{H^*} for bare and single-atom doped MBenes. (e) Additional comparison plot of ΔG_{H^*} by using 25 new MBenes to evaluate the trained ML model. (f) DFT-calculated substitution and cohesive energy plot of 28 promising HER catalysts screened by using ML. Panels (a), (b), (d), (e), and (f) are adapted from ref. 17 with permission from Elsevier, copyright 2020.

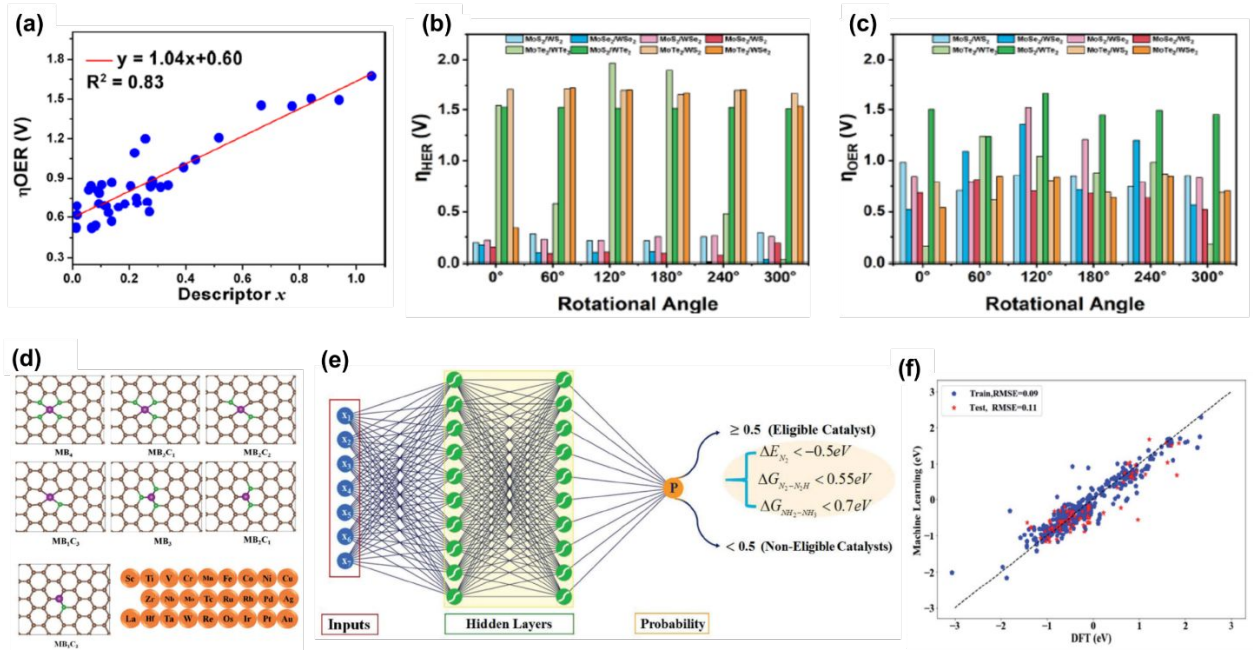

Figure 12. (a) The relationship between the descriptor and the catalytic performance for OER. (b, c) Relationships of the rotational angle with overpotential η_{HER} and η_{OER}. (d) Structures of B-doped graphene with single-metal atoms. (e) Deep neural network (10 neurons in each hidden layer) architecture. (f) Prediction performance plot between DFT-calculations and machine-learning outputs. Panels (a) – (c) are adapted from ref. 16 with permission from American Chemical Society, copyright 2020. Panels (d) – (f) are adapted from ref. 38 with permission from Royal Society of Chemistry, copyright 2020.

Table 1. ML types and description.

ML Class	Task	Method	Description
Supervised	Regression	Support Vector Machine (SVM)	It optimizes the process to determine the hyperplane that bisects the maximized margin and separates the datasets into different classes by choosing the appropriate support vectors.
		Least Absolute Shrinkage and Selection Operator (LASSO)	A linear regression that uses regularization to reduce the number of fitted coefficients, with advantage of avoiding overfitting.
		Random Forest (RF)	An ensemble of decision trees. It helps to rank the importance of variables by the order of nodes and correct overfitting in one decision tree.
		Neural Network (NN)	NN learns complex non-linear relationships between the features and target with the advantages of (1) automatic extraction of features from inputs without human intervention, (2) ability to handle non-linear and complex problems, and (3) high predictive accuracy by increasing learning epochs, neurons, and hidden layers.
		Kernel Ridge Regression (KRR)	It combines ridge regression with the kernel trick which learns a function in the space induced by the respective kernel. It simply computes the inner products between the images of all pairs of data in the feature space.
	Classification	Support Vector Machine (SVM)	See above.

		Random Forest (RF)	See above.
		K-Nearest Neighbor (KNN)	A non-linear classifier that finds decision boundaries and sorts data into various categories.
		Linear Discriminant Analysis (LDA)	It classifies the data by creating an axis that maximizes the distance between the means of categories while minimizing the scatter. Like PCA, it can reduce the dimension of the data.
		Neural Network (NN)	See above.
		Naive Bayes	It is a probabilistic classifier considering all input features independently. Thus, each feature equally contributes to drawing the estimation. It runs well with only a small number of training data.
Unsupervised	Clustering	K-Means Clustering	Clusters samples into K groups based on distances.
		Hierarchical Clustering	Starts from merging two most similar objects, and proceeds through an iterative process that identifies and merges the two most similar clusters until the final state, in which each cluster is distinct from other clusters.
	Dimensionality Reduction	Principal Component Analysis (PCA)	Widely employed to reduce the dimension of a large data set by computing the principal components that constitute a set of orthonormal bases on the data. Typically the first few to few dozen principal components, which explain most of the variance in the data, are taken as input and the rest is ignored.
Semisupervised	Classification	Support Vector Machine (SVM)	It can first classify only the labeled data, and then predict the probability for unlabeled data.

	Positive and Unlabeled (PU) Learning	A binary classifier that deals with two sets of data, the positive set P (labeled) and a mixed set U (unlabeled).

Table 2. List of open-source materials databases and software libraries.

Name	Description	URL
C2DB	Computational database of 2D materials	https://cmr.fysik.dtu.dk/c2db/c2db.html
ICSD	Experimental and computational database of inorganic materials	http://www2.fiz-karlsruhe.de/icsd_home.html
MaterialsCloud	Computational database of materials	https://materialscloud.org/discover
AFLOWlib	Computational database of materials	http://aflowlib.org
MaterialsProject	Computational database of materials	https://materialsproject.org
1D and 2D Materials Database	599 1D vdW and 1755 2D vdW solids	https://reedgroup.stanford.edu/databases.html
CMR	Computational database of materials	https://cmr.fysik.dtu.dk
MaterialsWeb	Computational database of materials	https://materialsweb.org
COD	Database of organic and inorganic materials searched from previous journal publications	http://crystallography.net
Pymatgen	Python library for materials analysis	https://pymatgen.org
Matminer	Python library for data mining the properties of materials	https://hackingmaterials.lbl.gov/matminer
aNANt	A functional materials database	http://anant.mrc.iisc.ac.in

Table 3. Summary of machine-learning-assisted studies of 2D materials.

Research Area	2D Materials	Source of Descriptors/Targets	Predicted Features	ML Algorithms	Ref
Prediction of materials' properties	▪ Mechanical				
	Graphene	OM image (R, G, B)	Material strength	SVM	12
	WS_2	Molecular dynamics (MD) simulation	Fracture strain, strength Young's modulus	RF	23
	MX_2 M (Mo, Nb, W, Ti) X (S, Se, Te)	Previous literature (Structural, electrical properties)	Nanoscale friction	Bayesian model	24
	▪ Electronic				
	MXenes	DFT, Chemical structures	Band gap	KRR, SVR, GPR	26
	Many 2D materials	Computational 2D materials database (C2DB)	Band gap, Exciton binding energy	LASSO	47
	▪ Optoelectronic				
	1L-WSe_2	Polarization resolved PL mapping image	Exciton valley polarization landscape	RF	28
	▪ Thermodynamic				
	Many 2D materials	C2DB	Thermodynamic stability	XGBoost, SISSO	48
Production of 2D materials	▪ Top-down approaches (Mechanical printing, Liquid exfoliation)				
	Graphene, MoS_2	Optical microscope image	Thickness of materials	K-means clustering	7
	Graphene, hBN, WTe_2, MoS_2	Optical microscope image	Thickness and Position of materials	Convolution neural network (CNN)	10
	Graphene, MoS_2	Optical microscope image	Layer number (Mono-, Bi-layer)	CNN	11
	Graphene, MoS_2	Optical microscope image (RGB)	Layer numbers	SVM	8
	Graphene	Optical microscope image	Layer numbers (Mono-, Bi-, Tri-layer)	Bayesian gaussian mixture model (BGMM)	9
	Graphene	Optical microscope image (R-G plane)	Layer number, Flake size	SVM	12
	Graphene dispersions	Optical microscope image (Brightfield) Quantitative polarized optical microscope (qPOM)	Layer number, Flake size	K-means clustering	37
	▪ Synthesizability				
	MXenes	DFT, Elemental features	Synthesizability	PU learning	13

	▪ Materials discovery				
	2D topological insulator	DFT, 2DMatPedia database	Discover 12 new 2D topological insulators	SISSO, XGBoost	[49]
Applications	▪ Sensor				
	2D TMDs	pymatgen, FactSage, AFLOW material database	Good TMDs for Hg^0 sensing	Material screening based on material databases	[15]
	Nanoporous MoS_2	MD simulation (Residence time, Ionic current)	Distinguish sensor readings of amino acids	Logistic regression (LR), KNN, RF	[14]
	▪ Catalyst				
	MBenes	DFT, Elemental features	Predict good MBenes for HER	LASSO, RFR, KRR, SVR	[17]
	2D TMD Heterostructures	DFT	Predict a good TMD heterostructure for HER and OER	LASSO	[16]
	▪ Photovoltaic (not discussed)				
	Many 2D materials	DFT, Inorganic crystal structure database (ICSD)	Predict good 2D photovoltaic materials	Gradient Boosting Classifier (GBC), LR, SVM, RF etc.	[50]

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

J. H. Chen acknowledges the financial support from the US National Science Foundation (NSF) Scalable Nanomanufacturing Program (NSF CMMI-1727846 and CMMI-2039268) and the NSF Future Manufacturing Program (NSF CMMI-2037026). This work is also supported by the Laboratory Directed Research and Development (LDRD) program from Argonne National Laboratory, provided by the Director, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-06CH11357. L. Wang and M. K. Y. Chan acknowledge funding from the National Science Foundation MRSEC program under grant number DMR-1720139. Use of the Center for Nanoscale Materials, an Office of Science user facility, was supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. DE-AC02-06CH11357.

References

1. B. Anasori, M. R. Lukatskaya and Y. Gogotsi, *Nat. Rev. Mater.*, 2017, **2**, 1-17.
2. N. Mounet, M. Gibertini, P. Schwaller, D. Campi, A. Merkys, A. Marrazzo, T. Sohier, I. E. Castelli, A. Cepellotti and G. Pizzi, *Nat. Nanotechnol.*, 2018, **13**, 246-252.
3. S. Das, H. Pegu, K. K. Sahu, A. K. Nayak, S. Ramakrishna, D. Datta and S. Swayamjyoti, in *Synthesis, Modeling, and Characterization of 2D Materials, and Their Heterostructures*, Elsevier, 2020, pp. 445-468.
4. J. Cai, X. Chu, K. Xu, H. Li and J. Wei, *Nanoscale Adv.*, 2020, **2**, 3115-3130.
5. J. Schmidt, M. R. Marques, S. Botti and M. A. Marques, *npj Comput. Mater.*, 2019, **5**, 1-36.
6. A. Y.-T. Wang, R. J. Murdock, S. K. Kauwe, A. O. Oliynyk, A. Gurlo, J. Brgoch, K. A. Persson and T. D. Sparks, *Chem. Mater.*, 2020, **32**, 4954-4965.
7. Y. Li, Y. Kong, J. Peng, C. Yu, Z. Li, P. Li, Y. Liu, C.-F. Gao and R. Wu, *J. Materiomics*, 2019, **5**, 413-421.
8. X. Lin, Z. Si, W. Fu, J. Yang, S. Guo, Y. Cao, J. Zhang, X. Wang, P. Liu and K. Jiang, *Nano Res.*, 2018, **11**, 6316-6324.
9. S. Masubuchi and T. Machida, *npj 2D Mater. Appl.*, 2019, **3**, 1-7.
10. S. Masubuchi, E. Watanabe, Y. Seo, S. Okazaki, T. Sasagawa, K. Watanabe, T. Taniguchi and T. Machida, *npj 2D Mater. Appl.*, 2020, **4**, 1-9.
11. Y. Saito, K. Shin, K. Terayama, S. Desai, M. Onga, Y. Nakagawa, Y. M. Itahashi, Y. Iwasa, M. Yamada and K. Tsuda, *npj Comput. Mater.*, 2019, **5**, 1-6.
12. J. Yang and H. Yao, *Extreme Mech. Lett.*, 2020, **39**, 100771.

13. N. C. Frey, J. Wang, G. I. n. Vega Bellido, B. Anasori, Y. Gogotsi and V. B. Shenoy, *ACS Nano*, 2019, **13**, 3031-3041.

14. A. B. Farimani, M. Heiranian and N. R. Aluru, *npj 2D Mater. Appl.*, 2018, **2**, 1-9.

15. H. Zhao, C. I. Ezeh, W. Ren, W. Li, C. H. Pang, C. Zheng, X. Gao and T. Wu, *Appl. Energy*, 2019, **254**, 113651.

16. L. Ge, H. Yuan, Y. Min, L. Li, S. Chen, L. Xu and W. A. Goddard III, *J. Phys. Chem. Lett.*, 2020, **11**, 869-876.

17. X. Sun, J. Zheng, Y. Gao, C. Qiu, Y. Yan, Z. Yao, S. Deng and J. Wang, *Appl. Surf. Sci.*, 2020, **526**, 146522.

18. S. A. Tawfik, O. Isayev, C. Stampfl, J. Shapter, D. A. Winkler and M. J. Ford, *Adv. Theor. Simul.*, 2019, **2**, 1800128.

19. J. H. Ward Jr, *J. Am. Stat.*, 1963, **58**, 236-244.

20. C. Lee, X. Wei, J. W. Kysar and J. Hone, *Science*, 2008, **321**, 385-388.

21. A. Nie, Y. Bu, P. Li, Y. Zhang, T. Jin, J. Liu, Z. Su, Y. Wang, J. He and Z. Liu, *Nat. Commun.*, 2019, **10**, 1-7.

22. C. Androulidakis, K. Zhang, M. Robertson and S. Tawfick, *2D Mater.*, 2018, **5**, 032005.

23. X. Wang, D. Han, Y. Hong, H. Sun, J. Zhang and J. Zhang, *ACS Omega*, 2019, **4**, 10121-10128.

24. B. S. Baboukani, Z. Ye, K. G. Reyes and P. C. Nalam, *Tribol. Lett.*, 2020, **68**, 1-14.

25. G. W. Semenoff, *Phys. Rev. Lett.*, 1984, **53**, 2449.

26. A. C. Rajan, A. Mishra, S. Satsangi, R. Vaish, H. Mizuseki, K.-R. Lee and A. K. Singh, *Chem. Mater.*, 2018, **30**, 4031-4038.

27. O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo and A. Tropsha, *Nature Commun.*, 2017, **8**, 1-12.

28. K. Tanaka, K. Hachiya, W. Zhang, K. Matsuda and Y. Miyauchi, *ACS Nano*, 2019, **13**, 12687-12693.

29. N. C. Frey, D. Akinwande, D. Jariwala and V. B. Shenoy, *ACS Nano*, 2020, **14**, 13406-13417.

30. S. Haastrup, M. Strange, M. Pandey, T. Deilmann, P. S. Schmidt, N. F. Hinsche, M. N. Gjerding, D. Torelli, P. M. Larsen, A. C. Riis-Jensen, J. Gath, K. W. Jacobsen, J. Jørgen Mortensen, T. Olsen and K. S. Thygesen, *2D Mater.*, 2018, **5**, 042002.

31. A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. A. Persson, *APL Mater.*, 2013, **1**, 011002.

32. C. Chen, W. Ye, Y. Zuo, C. Zheng and S. P. Ong, *Chem. Mater.*, 2019, **31**, 3564-3572.

33. L. Ward, A. Dunn, A. Faghaninia, N. E. R. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K. A. Persson, G. J. Snyder, I. Foster and A. Jain, *Comput. Mater. Sci.*, 2018, **152**, 60-69.

34. A. E. Hoerl and R. W. Kennard, *Technometrics*, 1970, **12**, 55-67.

35. M. Dou and M. Fyta, *J. Mater. Chem. A*, 2020, **8**, 23511-23518.

36. J. Shim, S.-H. Bae, W. Kong, D. Lee, K. Qiao, D. Nezich, Y. J. Park, R. Zhao, S. Sundaram and X. Li, *Science*, 2018, **362**, 665-670.

37. M. J. Abedin, T. Barua, M. Shaibani and M. Majumder, *Adv. Sci.*, 2020, **7**, 2001600.

38. M. Zafari, D. Kumar, M. Umer and K. S. Kim, *J. Mater. Chem. A*, 2020, **8**, 5209-5216.

39. L. Ge, H. Yuan, Y. Min, L. Li, S. Chen, L. Xu and W. A. Goddard, *J. Phys. Chem. Lett.*, 2020, **11**, 869-876.

40. C. Anichini, W. Czepa, D. Pakulski, A. Aliprandi, A. Ciesielski and P. Samorì, *Chem. Soc. Rev.*, 2018, **47**, 4860-4908.

41. S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson and G. Ceder, *Comput. Mater. Sci.*, 2013, **68**, 314-319.

42. S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. Hart, S. Sanvito and M. Buongiorno-Nardelli, *Comput. Mater. Sci.*, 2012, **58**, 227-235.

43. G. Gao, A. P. O'Mullane and A. Du, *ACS Catal.*, 2017, **7**, 494-500.

44. J. Liu, Y. Liu, N. Liu, Y. Han, X. Zhang, H. Huang, Y. Lifshitz, S.-T. Lee, J. Zhong and Z. Kang, *Science*, 2015, **347**, 970-974.

45. Y. Jiao, Y. Zheng, K. Davey and S.-Z. Qiao, *Nat. Energy*, 2016, **1**, 16130.

46. S. Masubuchi, M. Morimoto, S. Morikawa, M. Onodera, Y. Asakawa, K. Watanabe, T. Taniguchi and T. Machida, *Nat. Commun.*, 2018, **9**, 1-12.

47. J. Liang and X. Zhu, *J. Phys. Chem. Lett.*, 2019, **10**, 5640-5646.

48. G. R. Schleder, C. M. Acosta and A. Fazzio, *ACS Appl. Mater. Interfaces*, 2019, **12**, 20149-20157.

49. G. R. Schleder, B. Focassio and A. Fazzio, *Appl. Phys. Rev.*, 2021, **8**, 031409.

50. H. Jin, H. Zhang, J. Li, T. Wang, L. Wan, H. Guo and Y. Wei, *J. Phys. Chem. Lett.*, 2020, **11**, 3075-3081.