Methodological framework for materials discovery using machine learning

Eng Hock Lee,[1,*] Wei Jiang,[1,2,†] Hussain Alsalman,[1,3,‡] Tony Low,[1,§] and Vladimir Cherkassky[1,‖]

[1]*Department of Electrical and Computer Engineering, University of Minnesota, 200 Union St SE, Minneapolis, Minnesota 55455, USA*
[2]*School of Physics, Beijing Institute of Technology, Beijing 100081, China*
[3]*King Abdulaziz City for Science and Technology (KACST), Riyadh 6086-11442, Kingdom of Saudi Arabia*

(Received 29 August 2021; accepted 21 March 2022; published 14 April 2022)

Traditionally, materials discovery has been guided by basic physical rules, and such rules embody the basic understanding of the physical characteristics of interest of the material. However, the discovery of physical rules remains a challenging task due to the inherent difficulty in recognizing patterns in the high-dimensional and highly nonuniform distributed materials space. The standard data analytics approach using machine learning (ML) may fall short in producing meaningful results due to fundamental differences between the underlying assumptions and goals of ML vs materials discovery. ML is mainly focused on estimating complex black-box predictive models (that are nonlinear and multivariate), whereas in materials discovery, the goal is to come up with interpretable data-driven physical rules. Here, we attempt to tackle this problem by proposing a robust data analytics framework that allows us to derive basic physical rules from data. We introduce the concept of global and local modeling, utilizing both supervised and unsupervised learning, for highly nonuniformly distributed materials data. To enhance the model interpretation, we also introduce a model-independent interpretation technique to assist human experts in extracting useful physical rules. The proposed framework for extracting data-derived physical rules at the global and local level is illustrated using two case studies: (1) classification of van der Waals (vdW) and non-vdW (nvdW) materials and (2) classification of wide bandgap and non-wide bandgap vdW materials.

DOI: 10.1103/PhysRevMaterials.6.043802

I. INTRODUCTION

Data analytic approach for materials discovery involves application of machine learning (ML) methods to database of materials with known properties [1–10]. These include prediction of materials properties [11–18], searching for optimal materials structures [19,20], finding new materials compositions [21–23], and extraction of physically meaningful representation of input features (also known as descriptors) [24–32].

Next, we briefly review several representative studies. Umehara *et al.* [16] perform gradient analysis as a postprocessing step to extract local input feature importance from convolutional neural network models for predicting photoelectrochemical power of materials. Kunkel *et al.* [9] use an active ML (AML) approach to explore the materials space to discover organic semiconductor materials. At each iteration, the AML finds the highly promising molecules based on a fitness function (that balances model exploration and exploitation), which is then validated via first-principles *ab initio* simulation and then merged into the available data for the next iteration. Similarly, Del Cueto and Troisi [10] introduce an extrapolative search strategy that uses ML methods to find materials with high figures of merit. However, we stress that these studies are fundamentally different from this paper, where our objective is to extract physical rules at the global and local level.

On the other hand, many methodological aspects of this data-driven discovery process are not clearly understood due to different methodological assumptions made in ML and materials discovery. We argue that understanding these differences is critical for successful application of ML methods. Moreover, these differences (in underlying assumptions) should be formalized before data analytics modeling. That is, different assumptions and goals (of modeling) should be properly reflected in a modeling procedure. In contrast, many existing papers point out some of these differences and attempt to address them in an *ad hoc* manner, by suggesting heuristic modifications of existing ML algorithms [17,18]. For instance, Sutton *et al.* [17] identifies a subclass of materials using domains of applicability method to predict the formation energy of transparent conducting oxide materials with high accuracy, even though the overall accuracy for all available materials is poor. This method attempts to overcome an important difference between underlying assumptions used in ML (that training and test data have similar statistical distributions) and materials discovery, where future (test) data may be different from past data. Ward *et al.* [11] and Kailkhura *et al.* [18] propose a general-purpose ML pipeline aimed at achieving both high predictability and interpretability. The ML pipeline is designed to deal with the problem of imbalanced and nonuniform distribution data and allows interpretation at the model and decision levels. Their approach is to partition the target output into several subregions (based on

[*] leex7132@umn.edu
[†] jiangw@bit.edu.cn
[‡] halsalma@umn.edu
[§] tlow@umn.edu
[‖] cherk001@umn.edu

domain knowledge) and then perform ML modeling for each corresponding subregion. However, ML modeling and interpretation of models for subregions (formed by this method) may not be meaningful when a subregion contains materials from several different classes.

In materials discovery, available data are a set of materials with known properties, and the goal is to discover material(s) with useful target properties. Each known material (in a database) is represented as a set of values ($\mathbf{x}$, y), where input vector $\mathbf{x}$ is a set of physical parameters, and an output y encodes some (desired) target property. It is assumed that input characteristics (or $\mathbf{x}$ values) determine the output (target) property, and this unknown dependency $y = f(\mathbf{x})$ can be estimated from an available dataset using ML methods. The estimated dependency is then used to predict materials with useful properties. The notion of a useful target property is formalized as a set (or range) of y values that are specified *a priori*. For real-valued outputs, target outputs correspond to a range of y values, and estimation of dependency $y = f(\mathbf{x})$ is known as a regression problem. For categorical outputs, different types of materials (i.e., useful vs not useful) correspond to different classes, under classification formulation.

For both classification and regression estimation problems, the goal is to estimate a model that provides small prediction error (for new test inputs), where an error is the discrepancy between y values predicted by the model and true outputs for test inputs. This discrepancy is quantified via some loss function, given *a priori*. For regression problems, the typical loss is squared error, and for classification, it is misclassification error or the number of mispredictions for test inputs.

Even though both ML and materials discovery attempt to estimate unknown mapping $\mathbf{x} \rightarrow y$ from labeled training samples ($\mathbf{x}$, y), there are certain differences in underlying assumptions. These differences are discussed next.

(1) *Similarity between training and test data distributions*. This assumption is common in ML [33], but it may not hold in materials discovery. That is, training and test data distributions are nonuniform and may not be similar.

(2) *Goal of modeling and causality*. Under the ML setting, the goal of modeling is good prediction, i.e., minimization of average prediction error (for test data). In materials discovery, the goal is finding a good set of $\mathbf{x}$ values, i.e., regions in the input $\mathbf{x}$ space with desired properties (y values). Further, in materials discovery, an estimated model $f(\mathbf{x})$ is interpreted as causal dependency between input characteristics ($\mathbf{x}$ values) and output properties. However, there is no assumption of causality in ML. Hence, there is a growing realization that data analytics models can be used for prediction but not for estimation of causal dependencies [34–36]. Note that, in most ML applications, a data analytics model is always conditioned on underlying distribution of observational data, whereas true causal models do not depend on the distribution of ($\mathbf{x}$, y) data. Therefore, ML methods can estimate, at best, causal models only in the regions (in $\mathbf{x}$ space) where the training data ($\mathbf{x}$, y) is available.

(3) *Quality of estimated ML model.* Note that most ML methods adopt standard loss functions, such as the mean squared error or classification error. However, such loss functions may not be the best choice in materials discovery.

(4) *Interpretation vs prediction.* An important requirement for materials discovery using ML is good interpretability of estimated models. This contrasts with most ML methods, where the main goal is prediction, and interpretation is sometimes a secondary consideration. Good interpretability is difficult for two reasons. First, ML models are usually nonlinear and high dimensional, but good interpretability is typically possible only for low-dimensional problems (with just a few input variables). The second problem is that data analytics modeling depends on distribution of observational data, so such models cannot be regarded as interpretable causal models, as discussed in Point (2) above. Therefore, interpretation of ML models is also data dependent.

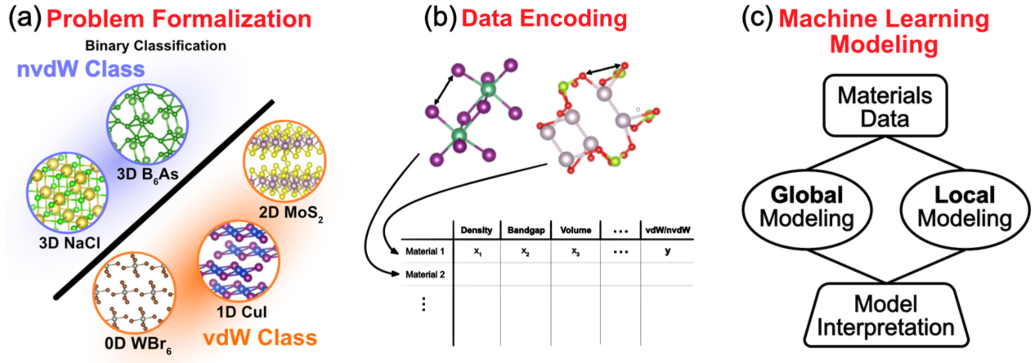
II. GLOBAL AND LOCAL MODELING APPROACHES FOR MATERIALS DISCOVERY USING ML

For materials discovery, available data are a set of materials with known properties, represented as a set of values ($\mathbf{x}$, y), and the goal is to discover material(s) with useful target properties (output y). Examples of such useful target properties could be: Does the material belong to a van der Waals (vdW) or non-vdW (nvdW) class, or is it wide or non-wide bandgap? As discussed earlier, there are two goals of modeling, i.e., prediction and interpretation. For interpretation, the goal is to derive simple interpretable rules for the estimated model $f(\mathbf{x})$. Note that achieving both objectives is (usually) not possible. For example, complex black box ML models, such as neural networks and support vector machines [33,37], lack good interpretation. Thus, the main point here is that data analytics modeling in material discovery usually involves a tradeoff between prediction and interpretation, whereas in most traditional ML applications, the main goal is prediction.

In materials discovery, estimating dependency $y = f(\mathbf{x})$ should also reflect highly nonuniform distribution of materials data. That is, nonuniform distribution of $\mathbf{x}$ values often results in several distinct data clusters in a multivariate input space, so that different clusters may exhibit different (local) dependencies $y = f(\mathbf{x})$. Such different local models will also have different interpretation. Considering nonuniform distribution of materials data leads to the following local modeling strategy: first, partition available training data into several clusters, based on similarity of $\mathbf{x}$ values only, and second, estimate a local model $y = f(\mathbf{x})$ for each cluster, using only ($\mathbf{x}$, y) samples from that cluster.

Note that estimated clusters effectively partition the input space into several disjoint regions (where each region corresponds to one cluster). For this partitioning, samples in each region can be further analyzed regarding their output property (y value), resulting in several local models. Using ML terminology, the first (clustering) step is known as *unsupervised* learning, and second step corresponds to a supervised learning task (such as classification or regression).

In contrast to local modeling, it is also possible to estimate a single global model using all available training data. Note that for the same training data, global and local modeling result in different estimated models. In terms of interpretation, global modeling results in a set of rules describing a single model (for all possible input values). In contrast, local modeling, based on unsupervised learning approach, yields



	Density	Bandgap	Volume	...	vdW/nvdW
Material 1	x_1	x_2	x_3	...	y
Material 2					
⋮					

FIG. 1. General overview. (a) Problem formalization as binary classification of van der Waals (vdW) vs non-vdW (nvdW) materials. (b) Each material is represented using five features (selected based on domain knowledge). (c) General flowchart of applying machine learning (ML) for global and local modeling and interpretation.

several local models (for different regions of the input space). Then each local model is interpreted as a set of rules. Notably, local and global approaches result in two different types of interpretations—both of which may be useful for materials discovery. In other words, global and local modeling represent two different views (and interpretations) of the same data. Arguably, large discrepancy between these two views can be expected if characteristically different materials classes are present in a dataset. Qualitatively speaking, local modeling enables discovering higher resolution knowledge hidden in the data.

Next, we present two case studies to demonstrate the proposed global and local modeling framework.

III. CASE STUDY 1: MODELING OF VDW VS NVDW MATERIALS

Materials discovery is usually an *ad hoc* process guided by physical intuition, physical rules of thumb, and experimental data, among many other factors. For example, when designing the bandgap of binary semiconducting systems, one can be guided by Vegard's law [38], which says that the bandgap of the binary semiconductor alloy can be linearly interpolated between that of its constituent bandgaps; when designing magnetic materials, the magnetic ground state can be estimated using Hund's rule [39], which states that the valence electrons arrange to maximize the total spin and angular momentum quantum number; when searching for flat band materials, one typically seeks high-symmetric geometric structures, e.g., kagome or Lieb lattice [40–42].

Here, we develop a framework for deriving the input feature importance for discriminating between vdW and nvdW materials [43]. Several input features that represent key defining features of vdW materials are discussed later in the Feature selection subsection.

Global and local modeling of vdW and nvdW materials includes several methodological steps, shown in Fig. 1. Specifically, the problem of estimating a mapping between physical characteristics of materials (or input features $\mathbf{x}$) and the target property y (vdW vs nvdW) is formalized as a binary classification problem, shown in Fig. 1(a). The data encoding step is shown in Fig. 1(b). This step also involves several other preprocessing steps, as detailed later. The data modeling step includes global or local modeling, as well as model interpretation, shown in Fig. 1(c).

Next, we describe application of global and local modeling approaches to a dataset of materials obtained from the Materials Project database [44]. This section is organized as follows. First, we introduce ML methods used for global and local modeling. Second, we describe the data preprocessing step. Finally, we present and discuss modeling results, including physical interpretation of estimated data analytics models.

A. ML methods

The following ML methods are used in this study: decision tree methods for supervised learning [45,46] and self-organizing maps (SOMs) for clustering [47]. Next, we provide a brief description and the motivation for using these methods.

Decision tree methods allow simple interpretation in the form of decision rules for partitioning the input space into regions corresponding to different classes. The classification and regression tree (CART) method [45] results in a single decision tree model which is highly interpretable. However, CART is not robust with respect to small variations in the training data. Thus, in this paper, we use the random forest (RF) method [46], which first estimates many different decision trees (of the same size) and then forms the final model by their averaging.

Local modeling includes a clustering step followed by estimation of a local model (for each cluster). Clustering is performed using only $\mathbf{x}$ values of the training data, and it results in partitioning of all materials into several clusters corresponding to materials with similar input properties. Note that the number of clusters (or regions in the input space) should be small for good interpretability. Then materials in each cluster are analyzed to estimate the relationship between materials with similar input properties and their target properties (output class labels). In this paper, the SOM method is used for clustering due to its advantages over traditional clustering methods [33,37]. That is, the SOM method not only performs clustering of multivariate data but also shows similarity between clusters in a low-dimensional structure. The SOM representation is particularly useful for interpretation of clusters formed by high-dimensional data. Typically, the SOM method is a set of units (clusters) arranged in a two-dimensional (2D) map topology.

B. Data preprocessing step

The raw materials data obtained from the Materials Project database [44] undergo extensive preprocessing steps before modeling using ML methods. These steps include data encoding and selection of input features, data filtering, and removing artificial (or heterostructure) materials from available data. These steps are discussed next.

1. Data encoding and feature selection

This includes proper encoding of inputs (physical parameters) and outputs (target properties) as real valued, categorical, or ordinal variables and selection of input features (usually a subset of all possible inputs) used for ML modeling. For this study, the goal of modeling is to estimate a predictive model for classifying a given material (with known physical properties) as vdW or nvdW. For this purpose, we first determine the dimensionality of the materials using the modified breadth-first-search algorithm [48]. Then we label a material as vdW if it is either zero dimensional (0D), one dimensional (1D), or 2D and nvdW if it is three dimensional (3D), as shown in Fig. 1(a).

Feature selection is an important part of the modeling process because chosen features need to be informative for discriminating between different target outputs (y) and provide meaningful interpretation of the estimated model. In this study, it means that the input features should contain *macroscopic* information about the material properties (e.g., bandgap, formation energy (FE) per atom) rather than *microscopic* information (e.g., distance between two atoms) that leads to ambiguous model interpretation. We selected five macroscopic properties encoded as real-valued input features: FE per atom, unit cell density (denoted as density), bandgap, unit cell volume (denoted as volume), and maximum lattice constant (Max ABC) of a unit cell lattice (with parameters a, b, and c). These features form the five-dimensional (5D) input vector $\mathbf{x}$. Additional motivation for selecting these input features is based on physical understanding of vdW and nvdW materials, as explained next:

(1) *Higher FE per atom for vdW.* vdW materials tend to have a smaller number of chemical bonds than nvdW materials because of the larger ratio of surface atoms mediated by vdW forces, which leads to overall higher FE per atom. We note that FE per atom is negative for stable materials, so the term higher FE per atom denotes less negative values.

(2) *Higher bandgap for vdW.* vdW materials have reduced dimensionality, hence stronger quantum confinement effect, which leads to generally larger bandgaps.

(3) *Higher Max ABC, lower density, higher volume for vdW.* In vdW materials, due to weak vdW interaction, the lattice constant should be larger than in nvdW materials with mostly strong covalent bond lengths. This automatically leads to higher Max ABC, lower density, and higher volume of vdW materials.

2. Data filtering

Following Cheon *et al.* [49], we removed any metastable materials by filtering them out with formation enthalpy energy above the convex hull $> 0.1\,\mathrm{eV/atom}$. Next, we removed materials containing elements from the lanthanoid and actinoid series of the periodic table since these elements are not commonly found in nature. Further, we removed materials containing elements from the noble gases group, as such material compounds are not likely to exist in nature. Finally, we removed potential organic materials that contain O-H and N-H clusters.

3. Removing artificially designed (heterostructure) materials

The available materials dataset contains many artificial materials. They are 2D heterostructure materials designed by researchers and uploaded into the Materials Project database. Such materials will be classified as vdW, but they are not relevant to our problem, as they do not represent naturally occurring vdW materials. It is difficult to remove these artificial heterostructure materials from the dataset automatically unless one exhaustively performs manual screening to filter them out. Fortunately, these artificial heterostructure materials contain similar lattice cell structure that is different from most vdW materials. This is because the atomic model construction of these artificial materials is represented in the so-called supercell, where a vacuum layer is added to the out-of-plane direction [50]. Therefore, we can apply a data analytics approach for identifying artificial materials, under the assumption that statistical distribution of five input features for such materials is different from distribution in natural vdW materials.

Based on this assumption, we performed clustering of all vdW materials (using 3×3 SOM modeling) followed by analysis of univariate histograms of input feature values in each cluster. Visual analysis of histograms of feature values shows that distribution of feature values in one SOM unit is very different from all other units. This outlier unit is likely to contain mostly artificial materials, and it has been confirmed by additional examination of materials from this cluster. Figure 2(a) shows the result of SOM modeling along with the number of vdW materials in each unit. The unit with distinctly different histograms of feature values is circled in red, and it contains 677 vdW materials identified as artificial. These materials are removed from the dataset. Some examples of the identified artificial materials are also presented in Fig. 2(b). Later, we also found out that most of the artificial materials identified here are in fact generated from Ref. [19].

The final number of materials used for modeling is 55 792.

C. Modeling results

This section describes data analytics modeling of the materials dataset. Application of ML methods using global and local modeling approaches results in different interpretation of available materials data. For example, interpretation of a global model may help to understand (relative) importance of input parameters for discovering vdW materials, whereas local modeling may help to identify important input parameters specific to local regions of the input space.

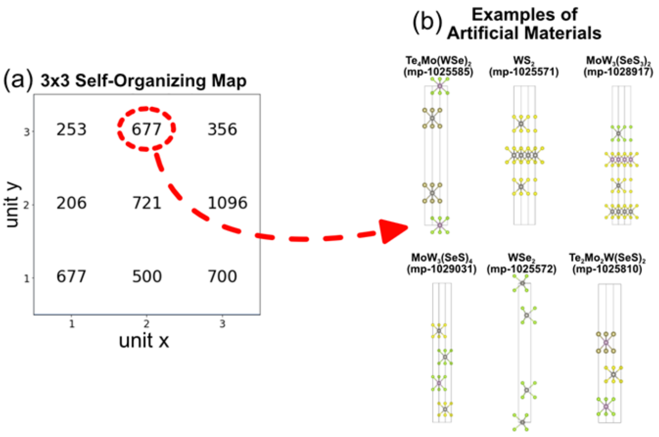

(a) 3x3 Self-Organizing Map
253
677
356
206
721
1096
677
500
700
unit y
unit x
1
2
3
1
2
3
(b) Examples of Artificial Materials
$Te_4Mo(WSe)_2$ (mp-1025585)
WS_2 (mp-1025571)
$MoW_3(SeS_3)_2$ (mp-1028917)
$MoW_3(SeS)_4$ (mp-1029031)
WSe_2 (mp-1025572)
$Te_2Mo_2W(SeS)_2$ (mp-1025810)

FIG. 2. (Self-organizing map SOM) method for removing artificial materials. (a) 3 × 3 SOM model trained on all vdW data. Visual analysis of feature histograms indicates that distribution of features for materials in unit (2, 3), marked in red, is significantly different from distribution in other units. (b) Examples of artificial materials from this cluster.

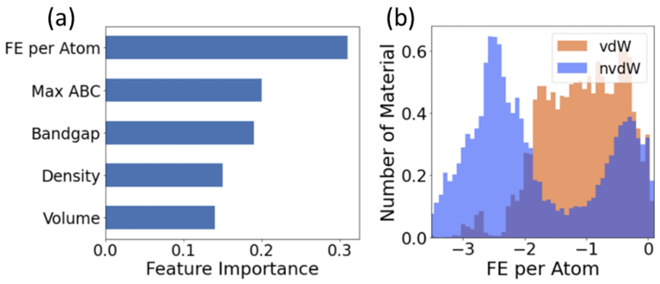


FIG. 3. Analysis of important features based on global modeling. (a) Ranking of feature importance using the random forest (RF) method. (b) Histogram of formation energy (FE) per atom from estimated model. Note that the histogram [for van der Waals (vdW) and non-vdW (nvdW)] are shown in the same scale (on the y axis) to account for class imbalance. From this histogram, we extract a simple rule: FE per atom > − 2 eV for vdW materials.

1. Global modeling using the RF method

The RF algorithm is applied to preprocessed data to estimate the classification model where binary class labels indicate vdW vs nvdW materials. This dataset is unbalanced, and the percentage of vdW and nvdW samples (materials) data is 8 and 92%, respectively. To address this class imbalance problem, we use unequal misclassification costs during training [33,37], so that misclassification errors for the minority class (vdW) are penalized more heavily than errors for the majority class (nvdW). These two types of errors are false negative (FN) and false positive (FP) errors, and their relative importance is specified by a predefined parameter R, or the ratio of misclassification costs [33,37]. For unbalanced data, the value of R is commonly defined by the class imbalance ratio, e.g.,

$$R = \frac{n_{\text{nvdW}}}{n_{\text{vdW}}}, \tag{1}$$

where n_{nvdW} and n_{vdW} denote the number of nvdW and vdW samples in the training data.

Further, prediction performance or prediction accuracy should be also properly adjusted or normalized to account for unbalanced data [33,37,51]. The normalized prediction accuracy (for test data) is defined as

$$A_{\text{norm}} = \frac{(\text{TP} \times R) + \text{TN}}{(\text{TP} + \text{FN}) \times R + \text{TN} + \text{FP}}, \tag{2}$$

where TP/TN denote true positive/negative accuracy, and FP/FN denote false positive/negative accuracy. According to Eq. (2), $A_{\text{norm}} = 0.5$ corresponds to a random guess, and $A_{\text{norm}} = 1$ corresponds to a 100% accurate prediction.

The procedure for training the RF model is discussed in Appendix A. The final estimated RF model has optimal *maxDepth* = 9 and *maxFeatures* = 4. For this model, prediction accuracy for test data is $A_{\text{norm}} = 0.751$, and its feature importance ranking is shown in Fig. 3(a). Based on RF modeling, the most important feature is FE per atom, followed by Max ABC, bandgap, density, and volume.

Next, we describe a simple technique to extract physical rules of thumb for input features using the estimated ML model. This technique is model independent (can be used for any ML model), and it yields a qualitative relationship between input features and the classification decision (vdW vs nvdW) predicted by a model. This procedure is illustrated next using the RF model shown in Fig. 3: (1) estimate RF model from training data and (2) plot univariate histograms for input feature values (of training samples) separately for each output class predicted by the model. For example, Fig. 3(b) shows the histograms of FE per atom (most important feature) for two output classes predicted by the global RF model. These histograms effectively reflect different distribution (of feature values) for two output classes. (3) Visual analysis of histograms [estimated in Step (2)] can be used for deriving simple IF-THEN-ELSE rules that qualitatively describe the effect of an input feature on classification decision. For example, from the histogram in Fig. 3(b), we can extract the rule: FE per atom > − 2 eV for vdW materials (or equivalently, FE per atom < − 2 eV for nvdW materials).

Further, it may be possible to relate these data analytics rules to first-principles knowledge. For example, the rule shown above agrees with physical knowledge that vdW materials tend to have relatively higher FE energies (vs nvdW materials) because of the weak vdW interactions. The more negative FE values indicate stronger chemical bonds formed in the materials. Therefore, FE per atom is the most important feature in the overall materials database, and it is found to be more important than the structural feature Max ABC.

Note that, for some (nonimportant) input features, it is possible that the two histograms (for two output classes) in Step (3) may be highly overlapping. This indicates that such features are not important, and the rules of thumb cannot be derived.

Finally, we point out several critical issues for interpretation of global models. All global models estimated by ML methods are optimized for average prediction performance; that is, performance index is averaged over (unknown) distribution of input features. As discussed in the Introduction section, such average prediction performance may not be a suitable index for materials discovery, where the goal is to

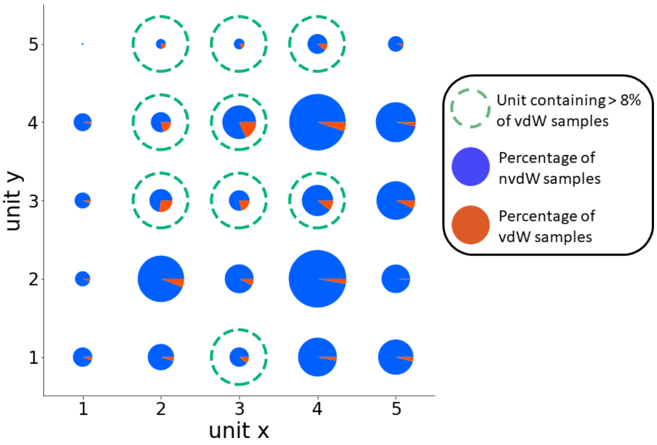


FIG. 4. Trained 5 × 5 self-organizing map (SOM) model showing 25 local regions (units) in five-dimensional feature space. The size of each unit corresponds to the number of materials in this unit, and the percentage of van der Waals (vdW) and non-vdW (nvdW) materials is represented as a pie chart. The units highlighted by green dashed circles contain more than 8% of the vdW samples, and they are selected for local modeling.

understand and/or predict materials in a local region of the input parameter space. For example, in this case study, the distribution of five input features is highly nonuniform, and all features have (unknown) nonlinear correlations. In addition, the target class distribution is unbalanced, i.e., there is only a small number of vdW materials relative to other (nvdW) materials. Moreover, the available materials dataset is collected from the public Materials Project database, with no restriction on the types of materials uploaded into the database. Hence, the global model trained on the whole dataset is biased by statistical properties of uploaded data. In this project, we partially overcome this bias by removing heterostructure materials from the dataset. However, many of these issues can be alleviated by local modeling, that is, estimating ML models for subsets of available data. This local modeling approach results in estimating local models for different (local) regions of the input feature space.

2. *Local modeling using SOM and RF*

Local modeling relies on partitioning of available data into several subsets (or clusters) based on similarity of input feature values or input ($\mathbf{x}$) values of the training data. In this study, we use the SOM method for clustering and RF for estimating local classifiers for each cluster.

Before SOM modeling, the available data are normalized so that each real-valued input is prescaled to a similar range (zero mean and unit standard deviation). Such prescaling of input features (to the same range) is necessary to ensure that all inputs contribute equally to distance calculation during clustering. (Recall that the raw input feature represents different physical variables that may have vastly different ranges of values.) Details of SOM modeling are discussed in Appendix B.

Clustering is performed using a 5 × 5 SOM so that the trained SOM divides the 5D input space into 25 local regions represented by SOM units. Figure 4 displays the trained SOM model in a graphical form as a collection of units (clusters). These SOM units are arranged in a 2D structure so that adjacent units in this structure indicate most similar clusters [33,51]. Each local region or cluster contains a subset of the materials data, and their statistics are represented as a pie chart in Fig. 4. The size of a pie chart corresponds to the number of samples in a cluster (unit), and the pie chart displays the percentage of vdW and nvdW materials in that cluster. Note high variability of cluster sizes in Fig. 4, reflecting highly nonuniform distribution of data samples (materials) in 5D feature space. In this study, we are interested in clusters where their percentage of vdW samples is higher than the percentage of vdW in the whole materials dataset. Clusters with a green dashed circle contain at least 8% of the vdW samples, and these nine clusters are selected for local modeling, as discussed next.

Local modeling is performed by training the RF classifier using labeled materials data for each of the nine selected clusters. Then each local classifier can be used for predicting the output class (vdW vs nvdW) for test inputs that fall into one of these clusters. Note that the labeled dataset within each cluster is highly unbalanced, as evident from Fig. 4. This class imbalance is handled using different misclassification costs, in the same manner as discussed earlier for global modeling using RF.

Figure 5(a) shows the results of local modeling for each selected cluster. Each cluster in Fig. 5(a) shows the label indicating the position of the corresponding SOM unit in the 5 × 5 map shown in Fig. 4 and summary statistics of a local classifier, including its prediction accuracy (for test data) and the top three input features. The relative importance of input features (for predicting vdW material) is shown in parentheses. Note that different local models in Fig. 5(a) may have different important features; that is, local modeling results in different rules of thumb for different local regions of the input space. This confirms our earlier discussion about the limitations of global modeling.

To understand these results better, we include additional analysis on their elemental distributions (also known as periodic table heatmap) and the number of 0D, 1D, and 2D materials in each of the nine clusters, shown in Figs. 5(b) and 5(c), respectively. Figure 5(d) show the histogram of the top two important features from all estimated local models. This additional information is used, along with the local modeling results in Fig. 5(a), to identify similar local models that describe the same type of materials. Then combining such local models into a small number of groups will improve model interpretation.

First, note that clusters (2,3), (2,4), (2,5), and (3,3) consist mostly of 0D materials, which may explain the finding that all four local models (for these clusters) have the same important features, i.e., FE per atom, bandgap, and density, albeit with different ordering of importance [note that volume is ranked fourth in the cluster (3,3) and is not shown in Fig. 5(a)]. Therefore, these three features are key characteristic features of 0D vdW materials. From the periodic table heatmap analysis, we find a high percentage of the chlorine element in all these clusters (roughly 40% of materials contain the chlorine element in each cluster). One possible explanation is that halogen atoms can react with metals to form metal halides,

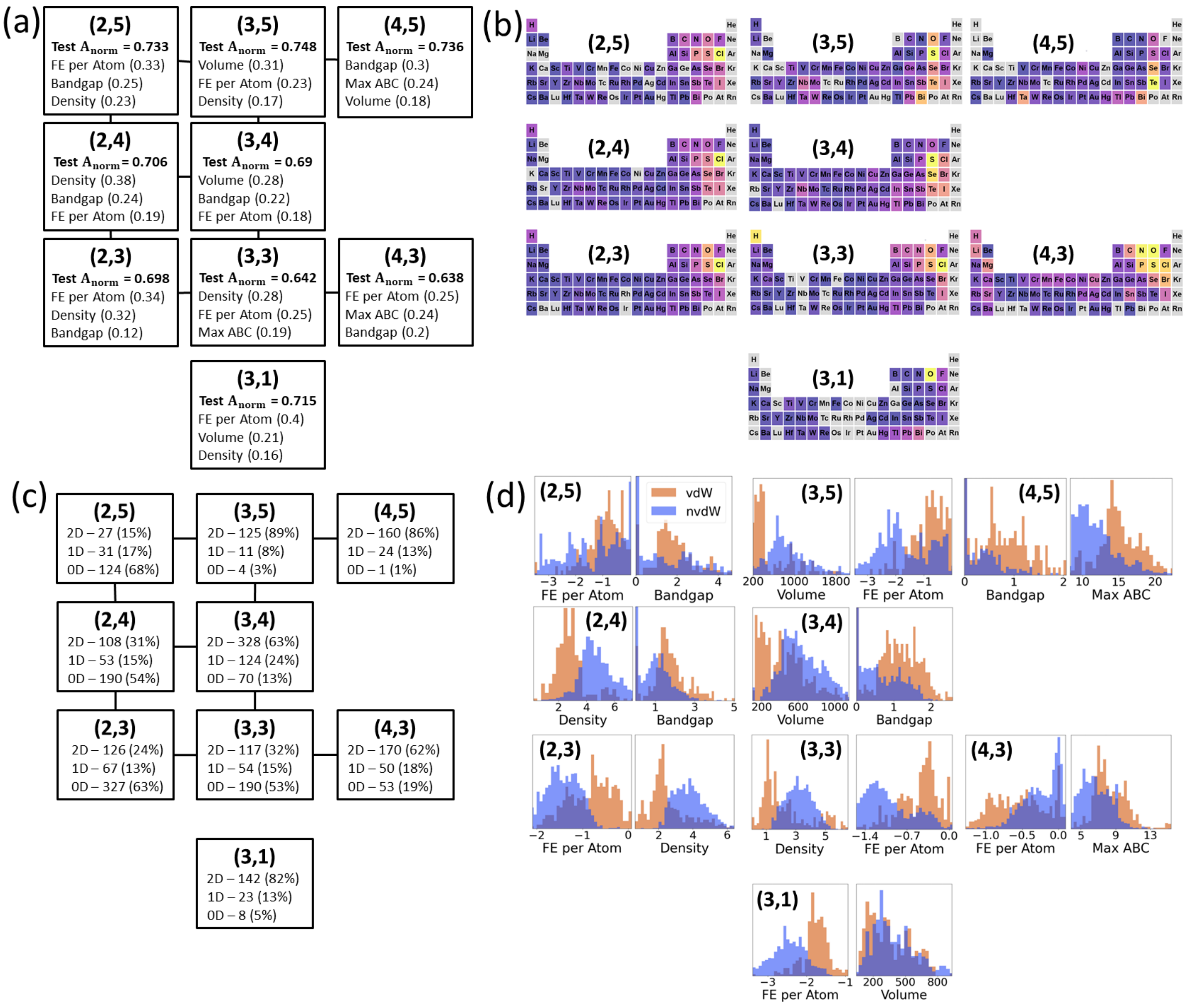


FIG. 5. Results of local modeling using the random forest (RF) method. (a) Summary statistics for local classification models estimated for each of the nine clusters. Each square box represents a unit (or cluster) of a 5 × 5 self-organizing map (SOM) model (shown in Fig. 4). Similarity between clusters is indicated by links connecting neighboring units. (b) Periodic table heatmap for interpretation of local models. (c) The number (and percentage) of zero-dimensional (0D), one-dimensional (1D), and two-dimensional (2D) materials in each SOM unit (cluster). (d) The histograms of the two most important feature values in each cluster.

and this helps to passivate unpaired electrons. They form one of the largest discrete molecules in addition to the organic molecules [52].

In what follows, we focus our attention on the 2D vdW materials instead since these materials are most important for device applications. We note that clusters (3,4), (3,5), and (4,5) contain 62, 77, and 87% of 2D metal chalcogenides (MX, M = metallic element, X = O, S, Se, or Te element), respectively. On top of that, roughly 23% of all 2D metal chalcogenide materials in cluster (4,5) contain the Bi element, which forms one of the most intensively studied topological insulators, such as Bi_2Se_3 [53] and WSe_2 [54], and most of them have much larger bandgap values than those nvdW materials in the same cluster. On the other hand, we notice that both clusters (3,4) and (3,5) have volume as the most important feature. By checking the materials within those clusters, we find most of these transition metal chalcogenide vdW materials contain either elongated or primitive lattice cell structure with no vacuum layer in the out-of-plane direction [see example materials in Fig. 6(a)], leading to very small in-plane area and thus a relatively smaller total volume. This is evident from the Fig. 5(d), where the volume (the most important feature) for vdW materials is much smaller than the nvdW materials counterpart. Based on this analysis, we can provide the same interpretation for clusters (3,4), (3,5), and (4,5).

In addition, we found that clusters (3,3) and (4,3) contain 33 and 21% metal halides (MX, M = metal, X = F, Cl, Br, or I) respectively, followed by 7 and 13% MXene, respectively. In cluster (4,3), the vdW materials have more negative FE per atom than the nvdW materials, as shown in Fig. 5(d). This is reasonable, as halogen atoms form much stronger covalent bonds, which tend to lower the FE. On the other hand, though it contains a substantial amount of metal halides and MXene, most materials in cluster (3,3) are 0D materials, as previously stated. Therefore, the density becomes the most important feature due to the lower dimension of those materials. This is also consistent with the histogram analysis shown in Fig. 5(d).

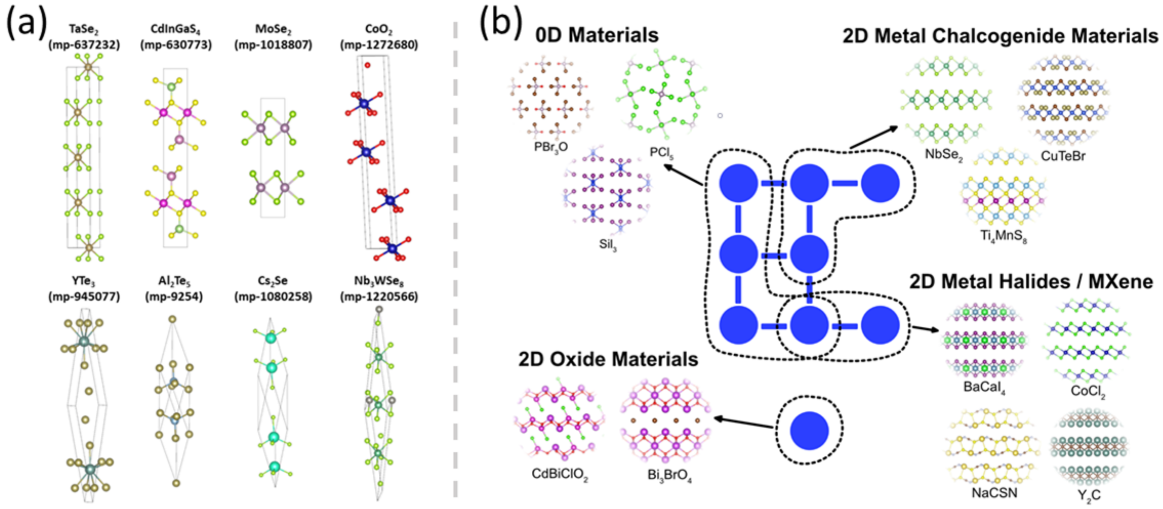


FIG. 6. Interpretation of local models as known classes of materials. (a) Examples of the transition metal chalcogenide van der Waals (vdW) materials found in cluster (3,4) and (3,5). (b) Interpretation of self-organizing map (SOM) units (shown in blue) containing different groups of materials.

Hence, we can provide the same interpretation for clusters (3,3) and (4,3).

Finally, we note that the SOM unit (3,1) has no neighboring SOM units in Fig. 5(a). Therefore, we conclude that its materials are not similar to other units in terms of their input feature values. This may lead to conjecture that it has target properties different from all other local models. Detailed analysis of the local model for cluster (3,1) shows that it has 82% of 2D materials, and among them, 88% of the 2D materials are oxide materials, which explains why FE per atom is identified as the most important feature. Additionally, within these majority oxides materials, there are negligible halide elements, which form even stronger bonds than oxygen atoms. Thus, the FE per atoms for vdW materials are less negative than those nvdW materials, which is mostly due to the weak vdW interaction of those vdW materials. This is consistent with the histogram analysis shown in Fig. 5(d).

Groups of clusters with the same interpretation are shown in Fig. 6(b). Note that each group includes clusters corresponding to neighboring SOM units. This observation confirms the hypothesis that materials with similar input features have similar target properties.

Comparison of the top three features for global and local models (for the same dataset) clearly shows their similarities and differences. For all local models, the FE per atom and bandgap are most likely to be selected among the three most important features (for classification of vdW materials). This agrees well with the global model interpretation, based on general physical properties of vdW materials, as discussed earlier, see Fig. 3(a). On the other hand, for different types of vdW materials, e.g., metal chalcogenide materials, layered topological materials, and oxide materials, the most important features selected by local modeling are quite different. These results demonstrate how the local modeling approach can extract higher resolution physical rules of thumb for different material families and differentiate the nuances in interpretability of these rules. In contrast, the global modeling approach is not capable of estimating specific models and rules of thumb for different materials families.

IV. CASE STUDY 2: MODELING OF WIDE VS NON-WIDE BANDGAP VDW MATERIALS

To further illustrate interpretation of global and local modeling approaches, we present another case study, using modeling of wide bandgap vs non-wide bandgap vdW semiconductor materials. Materials that exhibit large bandgaps are essential for high-temperate and high-power device applications due to their ability to maintain electronic functionalities at high ambient temperature environments [55–58]. Furthermore, multiple studies of vdW materials in the past decade have led to a recent surge of interest in wide bandgap vdW materials for next-generation electronic devices [59–64]. Like our previous vdW vs nvdW study, our goal is to demonstrate the ability of the global/local modeling framework to derive important features for discovery of wide bandgap vdW materials. This case study follows the same methodological approach as Case Study 1 and applies the same ML methods (RF and SOM) and the same techniques for deriving rules of thumb from data analytics models.

A. Data preprocessing step

The dataset of wide and non-wide bandgap vdW semiconductor materials is extracted from the Materials Project database, following the same preprocessing steps (as in the vdW vs nvdW case study). This includes removing irrelevant materials (such as metastable materials, materials containing lanthanoid and actinoids elements, etc.), as well as removing artificial vdW materials and the nvdW materials. The encoding of target output is described next. In the literature, the term *wide bandgap* usually refers to bandgap values >2–3 eV. However, the exact threshold is not well defined. In this study, the threshold value 3 eV is used to define wide bandgap materials, which is chosen based on the energy spectrum of the ultraviolet light. There are four input features used for modeling, i.e., FE per atom, density, volume, and Max ABC. In summary, the final dataset contains 4493 vdW materials, including 858 wide bandgap materials.

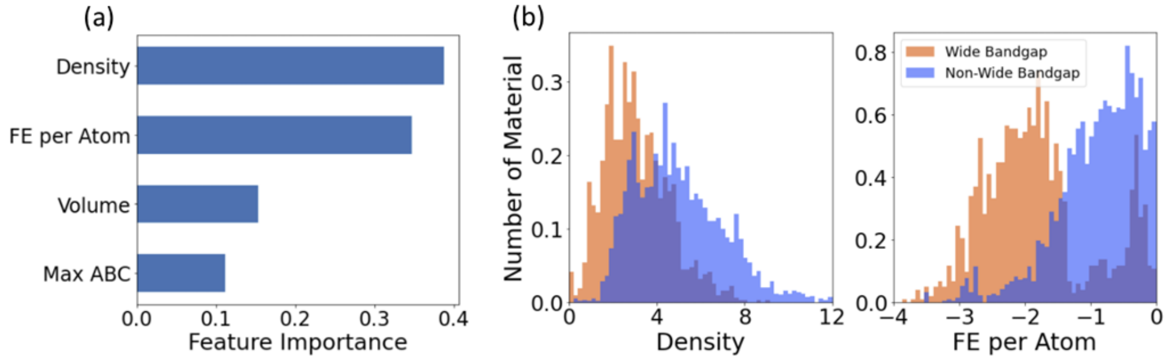


FIG. 7. Global modeling results using random forest (RF) methods for classification of wide vs non-wide bandgap van der Waals (vdW) materials. (a) Feature importance for the RF method. (b) Histogram of the two most important input features, the density and formation energy (FE) per atom, for each output class.

For this dataset, we estimate a binary classifier (for predicting wide bandgap materials), using global and local modeling, as described next.

B. Modeling results

1. Global modeling results

Global modeling for estimating a classifier (from the 4493 available vdW materials) is performed using the RF learning method. Like the earlier study of vdW vs non-VdW materials, we address the class imbalance during training using Eq. (1) and the performance index for test data as in Eq. (2). The training procedure is detailed in Appendix A. The prediction accuracy of the final model (estimated using test data) is 0.782, and its feature importance ranking is shown in Fig. 7(a). Based on Fig. 7(a), the two most important features are density and FE per atom (as both have very similar feature importance), followed by less important features volume and Max ABC. The physical basis for the importance of these features to wide bandgap materials can be understood from materials science. The first scenario corresponds to the case of 0D materials, where molecular orbitals are energetically isolated. One typical feature of 0D materials is their small density. The second scenario corresponds to strongly bonded materials that contain elements with high electronegativity, which is usually manifested as a stronger FE (more negative). The global modeling results show that density and FE per atom are almost equally important, indicating that the global database contains large portions of 0D materials and high electronegativity-element based materials. Further, Fig. 7(b) shows the histogram of density and FE per atom features for the global model. We can clearly see that smaller density values correspond to wide bandgap vdW materials. Similarly, we can see that more negative FE per atom corresponds to wide bandgap vdW materials. From the figures, we can extract a simple rule of thumb, i.e., density $<$ $4\,\mathrm{g/cm^3}$ and FE per atom $< -1.5\,\mathrm{eV}$ for wide bandgap vdW materials for the global modeling. However, as we demonstrated in the first case study, variations in rules of thumb can arise across different local materials clusters, which are explored next.

2. Local modeling results

For local modeling, we use SOM modeling to partition the available dataset into several clusters. The preprocessing and training procedure for SOM modeling is the same as described in the vdW vs nvdW case study, except that a 3 × 3 SOM is chosen (due to smaller sample size in this study). Hence, SOM modeling results in nine local regions (SOM units) in the four-dimensional input space, and these local regions are represented as pie charts that display the percentage of wide bandgap and non-wide bandgap materials in each unit, as shown in Fig. 8(a). For each unit, we estimate a classifier using the RF learning method. Note that unit (3,3) contains only non-wide bandgap materials, and therefore, applying local modeling to this unit is not necessary. Since this study focuses on discovering wide bandgap materials, any unit containing only non-wide bandgap materials can be discarded and not used for further analysis. Figure 8(b) shows the results of local modeling for each of eight units. Based on these results, we can combine neighboring units that have similar feature importance in their local models. Specifically, we identified the following groups of similar units: (a) groups containing units (1,2) and (1,3); (b) groups containing units (2,1) and (3,1); (c) groups containing units (2,2), (2,3), and (3,2); and (d) groups containing units (1,1), as highlighted by colored ellipses in Fig. 8.

Additional analysis of local models like the analysis presented for the vdW vs nvdW case study includes the periodic table heatmap, the number of 0D, 1D, and 2D materials, and the histogram of input features obtained from estimated local models. This analysis, not presented here due to space constraints, indicates that most of the wide bandgap materials in groups (a) and (b) are 0D materials. This can be explained by the fact that the most important feature in these four groups of units is density, as 0D materials tend to have smaller density due to weak vdW interaction between molecules in the unit cell. Indeed, the histogram of density for wide bandgap vdW materials has much smaller values than for non-wide bandgap vdW materials. This is consistent with our global model analysis that indicates large concentration of 0D materials.

For group (c), units (2,2), (2,3), and (3,2) have 90, 80, and 53% 2D metal oxides, respectively; the most important feature is FE per atom. Also, the histograms of FE per atom

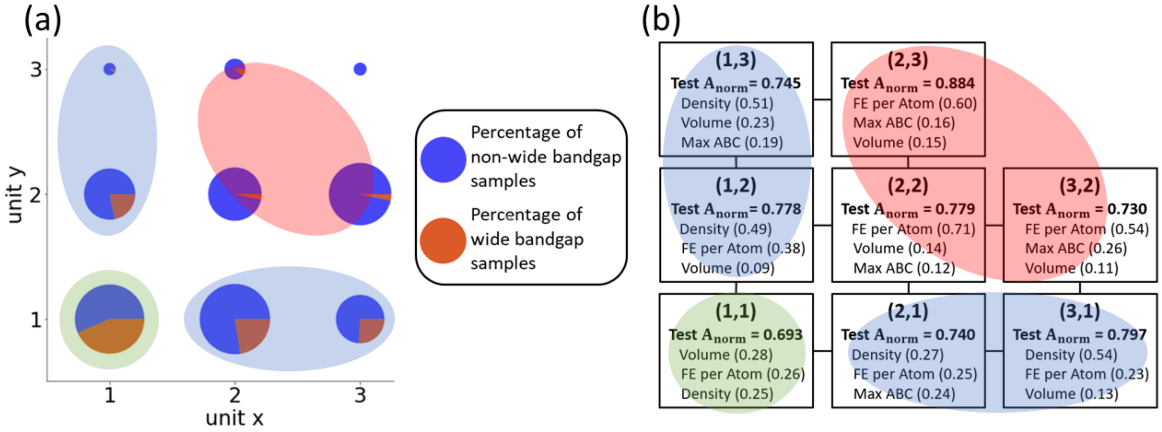


FIG. 8. Local modeling results on the classification of wide vs non-wide bandgap van der Waals (vdW) materials. (a) 3 × 3 self-organizing map (SOM) modeling to partition the dataset into nine clusters, represented as pie charts. (b) Summary statistics for local classification models using the random forest (RF) method.

for all units indicate that FE per atom for wide bandgap vdW materials is much smaller (more negative) than for non-wide bandgap materials. This can be explained by the fact that these materials tend to have higher electronegativity, due to the large electronegativity of oxygen and its tendency to form strong covalent bond, which results in larger bandgap.

Finally, unlike all other units that show a single dominant important feature, for unit (1,1), all four features have very similar importance, i.e., volume: 0.28, FE per atom: 0.26, density: 0.25, and Max ABC: 0.21. This suggests that this unit contains different types of materials. Indeed, further analysis of unit (1,1) shows that it contains 34% of 0D and 44% of 2D materials. In addition, it also contains 44% of metal oxide, 53% of metal fluoride, and 14% of metal oxide fluoride materials.

V. SUMMARY

This paper described methodological issues arising in application of ML for materials discovery. This includes similarities and differences between modeling assumptions used in ML and the objectives of materials discovery. ML modeling depends on underlying (unknown) distribution of observational data, whereas in materials discovery, the ultimate goal is to find causal relationships that can be related (interpreted) to first-principles physical laws.

We presented a framework for materials discovery, using local modeling, where all available data are first partitioned into several subsets of similar materials (via some clustering algorithm), and then a local model is estimated for each subset of the data. We argue that, for materials discovery, local modeling is more effective than the global modeling approach (when a single ML model is estimated for all available data). Local modeling enables better interpretation, as there may be different models (and their interpretations) for different families of materials. On the other hand, a global approach tends to suffer from the arbitrariness of the different weights from materials classes, which is more subjective to the current trends in materials research which populate the database than the true representation of their material sizes in nature.

The advantages of local modeling are demonstrated using two case studies: a classification model for vdW vs nvdW materials, and classification of wide vs non-wide bandgap vdW materials. Both case studies follow the same methodological approach and demonstrate the advantage of local modeling for modeling and interpretation of data analytics models. For both case studies, the problem is formalized as binary classification, and local modeling involves clustering using the SOM method followed by estimation of local classification models (for each cluster) using the RF method. We also introduced two techniques for interpretation of local models estimated from data. The first one is selection (ranking) of input features, according to their importance for discovering materials with desired target properties. This ranking of important features is specific to the RF method. The second technique enables derivation of simple rules of thumb (extracted from histograms of feature values) that describe the effect of an input feature on classification decision. The proposed approach for extracting rules of thumb (from the ML model) can be used with any ML method.

Case studies presented in this paper demonstrate that meaningful interpretation of data analytics models depends on both: (1) properties of estimated data analytics models, reflecting ML aspects of modeling, and (2) physical properties of materials data, reflecting first-principles knowledge.

ACKNOWLEDGMENT

This paper was supported by the National Science Foundation (NSF) under Contract No. DMREF-1921629.

APPENDIX A: RF IMPLEMENTATION

To implement RF, we use the CART algorithm from the SCIKIT-LEARN library [65] with modifications to account for the class imbalance. For both global and local modeling, the same RF algorithm is used, as described next. The parameter *numTrees* (number of trees) is set to 200 and 500 for global and local modeling, respectively. There are two tuning parameters that control the complexity of individual CART trees: $maxFeatures$, the maximum number of randomly selected input features, and $maxDepth$, the maximum depth of tree for each CART tree during model training. For the vdW vs nvdW materials case study, the range of possible values for $maxFeatures$ is set to [2,3,4] for both local and global RF modeling, while the range of $maxDepth$ is set to [7, 9, . . ., 15] and [3, 4, . . ., 8] for global and local modeling, respectively. Similarly, for the wide vs non-wide bandgap vdW materials case study, $maxFeatures$ and $maxDepth$ are set to [2,3] and [1, 2, . . ., 10], respectively, for both global and local modeling. Optimal tuning of these parameters (aka model selection) is performed via tenfold cross-validation (of the training data) for both studies.

For both global and local modeling, prediction accuracy is estimated via stratified resampling technique. That is, available data are randomly split into 80% training and 20% test data. Model estimation (training) is performed using training data, and then prediction performance is evaluated using test data. Stratification is used to ensure that each (randomly chosen) training and test dataset has the same class imbalance as the original data [66].

APPENDIX B: SOM IMPLEMENTATION

In this paper, we used the batch version of the SOM algorithm [33]. We use the 2D topological map, and its size is chosen loosely based on the number of training samples. Another factor is that a smaller map size tends to be more suitable for interpretation (of local models) because it may be difficult to interpret many local models. On the other hand, using just one unit (or one cluster) is not a good choice because it results in a single global model for all available data.

In this paper, a small 3×3 map size is selected for the artificial materials removal step and the local modeling step in the wide vs non-wide bandgap vdW materials case study, where only the vdW samples were used (∼4000 vdW materials). For the local modeling step in the vdW vs nvdW case study, a larger 5×5 map was used for training using all data (∼50 000 materials).

[1] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim, Machine learning in materials informatics: Recent applications and prospects, npj Comput. Mater. **3**, 54 (2017).

[2] Y. Liu, T. Zhao, W. Ju, and S. Shi, Materials discovery and design using machine learning, J. Materiomics **3**, 159 (2017).

[3] A. Jain, G. Hautier, S. P. Ong, and K. A. Persson, New opportunities for materials informatics: Resources and data mining techniques for uncovering hidden relationships, J. Mater. Res. **31**, 977 (2016).

[4] J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, Recent advances and applications of machine learning in solid-state materials science, npj Comput. Mater. **5**, 83 (2019).

[5] O. A. Von Lilienfeld and K. Burke, Retrospective on a decade of machine learning for chemical discovery, Nat. Commun. **11**, 4895 (2020).

[6] Z. Lu, Computational discovery of energy materials in the era of big data and machine learning: A critical review, Materials Rep.: Energy. **1**, 100047 (2021).

[7] J. Saal, A. Oliynyk, and B. Meredig, Machine learning in materials discovery: Confirmed predictions and their underlying approaches, Annu. Rev. Mater. Res. **50**, 49 (2020).

[8] G. R. Schleder, A. C. M. Padilha, C. M. Acosta, M. Costa, and A. Fazzio, From DFT to machine learning: Recent approaches to materials science—a review, J. Phys. Mater. **2**, 032001 (2019).

[9] C. Kunkel, J. T. Margraf, K. Chen, H. Oberhofer, and K. Reuter, Active discovery of organic semiconductors, Nat. Commun. **12**, 2422 (2021).

[10] M. Del Cueto and A. Troisi, Determining usefulness of machine learning in materials discovery using simulated research landscapes, Phys. Chem. Chem. Phys. **23**, 14156 (2021).

[11] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials, npj Comput. Mater. **2**, 16028 (2016).

[12] Y. Saad, D. Gao, T. Ngo, S. Bobbitt, J. R. Chelikowsky, and W. Andreoni, Data mining for materials: Computational experiments with AB compounds, Phys. Rev. B **85**, 104104 (2012).

[13] A. M. Deml, R. O'Hayre, C. Wolverton, and V. Stevanović, Predicting density functional theory total energies and enthalpies of formation of metal-nonmetal compounds by linear regression, Phys. Rev. B **93**, 085142 (2016).

[14] Y. Zhuo, A. M. Tehrani, and J. Brgoch, Predicting the band gaps of inorganic solids by machine learning, J. Phys. Chem. **9**, 1668 (2018).

[15] C. Nyshadham, M. Rupp, B. Bekker, A. V. Shapeev, T. Mueller, C. W. Rosenbrock, G. Csányi, D. W. Wingate, and G. L. W. Hart, Machine-learned multi-system surrogate models for materials prediction, npj Comp. Mater. **5**, 51 (2019).

[16] M. Umehara, H. S. Stein, D. Guevarra, P. F. Newhouse, D. A. Boyd, and J. M. Gregoire, Analyzing machine learning models to accelerate generation of fundamental materials insights, npj Comput Mater. **5**, 34 (2019).

[17] C. Sutton, M. Boley, L. M. Ghiringhelli, M. Rupp, J. Vreeken, and M. Scheffler, Identifying domains of applicability of machine learning models for materials science, Nat. Commun. **11**, 4428 (2020).

[18] B. Kailkhura, B. Gallagher, S. Kim, A. Hiszpanski, and T. Y. J. Han, Reliable and explainable machine-learning methods for accelerated material discovery, npj Comput. Mater. **5**, 108 (2019).

[19] L. Bassman, P. Rajak, R. K. Kalia, A. Nakano, F. Sha, J. Sun, D. J. Singh, M. Aykol, P. Huck, K. Persson, and P. Vashishta, Active learning for accelerated design of layered materials, npj Comput. Mater. **4**, 74 (2018).

[20] C. C. Fischer, A machine learning approach to crystal structure prediction, Doctoral dissertation, Massachusetts Institute of Technology, 2007.

[21] C. L. Phillips and G. A. Voth, Discovering crystals using shape matching and machine learning, Soft Matter **9**, 8552 (2013).

[22] G. Hautier, C. C. Fischer, A. Jain, T. Mueller, and G. Ceder, Finding nature's missing ternary oxide compounds using machine learning and density functional theory, Chem. Mater. **22**, 3762 (2010).

[23] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton, Combinatorial screening for new materials in unconstrained composition space with machine learning, Phys. Rev. B **89**, 094104 (2014).

[24] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, Big Data of Materials Science: Critical Role of the Descriptor, Phys. Rev. Lett. **114**, 105503 (2015).

[25] N. Artrith, A. Urban, and G. Ceder, Efficient and accurate machine-learning interpolation of atomic energies in compositions with many species, Phys. Rev. B **96**, 014112 (2017).

[26] M. Rupp, A. Tkatchenko, K. R. Müller, and O. A. Von Lilienfeld, Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning, Phys. Rev. Lett. **108**, 058301 (2012).

[27] K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller, and E. K. Gross, How to represent crystal structures for machine learning: Towards fast prediction of electronic properties, Phys. Rev. B **89**, 205118 (2014).

[28] L. Ward, R. Liu, A. Krishna, V. I. Hegde, A. Agrawal, A. Choudhary, and C. Wolverton, Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations, Phys. Rev. B **96**, 024104 (2017).

[29] E. Swann, B. Sun, D. M. Cleland, and A. S. Barnard, Representing molecular and materials data for unsupervised machine learning, Mol. Simul. **44**, 905 (2018).

[30] H. Huo and M. Rupp, Unified representation of molecules and crystals for machine learning, arXiv:1704.06439.

[31] O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, and A. Tropsha, Universal fragment descriptors for predicting properties of inorganic crystals, Nat. Commun. **8**, 15679 (2017).

[32] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. von Lilienfeld, K. R. Müller, and A. Tkatchenko, Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space, J. Phys. Chem. **6**, 2326 (2015).

[33] V. Cherkassky and F. M. Mulier, *Learning from Data: Concepts, Theory, and Methods* (John Wiley & Sons, Hoboken, 2007).

[34] B. Schölkopf, Causality for machine learning, arXiv:1911.10500.

[35] V. Vovk, H. Papadopoulos, and A. Gammerman, *Measures of Complexity* (Springer, New York, 2015), Chap. 19.

[36] T. Liu, L. Ungar, and K. Kording, Quantifying causality in data science with quasi-experiments, Nat. Comput. Sci. **1**, 24 (2021).

[37] V. Cherkassky, Predictive Learning (www.VCTextbook.com, 2013).

[38] A. R. Denton and N. W. Ashcroft, Vegard's law, Phys. Rev A **43**, 3161 (1991).

[39] M. Koskinen, M. Manninen, and S. M. Reimann, Hund's Rules and Spin Density Waves in Quantum Dots, Phys. Rev. Lett. **79**, 1389 (1997).

[40] W. Jiang, D. J. P. de Sousa, J. P. Wang, and T. Low, Giant Anomalous Hall Effect Due to Double-Degenerate Quasiflat Bands, Phys. Rev. Lett. **126**, 106601 (2021).

[41] W. Jiang, M. Kang, H. Huang, H. Xu, T. Low, and F. Liu, Topological band evolution between Lieb and kagome lattices, Phys. Rev. B **99**, 125131 (2019).

[42] W. Jiang, X. Ni, and F. Liu, Exotic topological bands and quantum states in metal-organic and covalent-organic frameworks, Acc. Chem. Res. **54**, 416 (2021).

[43] F. Oviedo, Z. Ren, S. Sun, C. Settens, Z. Liu, N. T. P. Hartono, S. Ramasamy, B. L. DeCost, S. I. P. Tian, G. Romano, A. G. Kusne, and T. Buonassisi, Fast and interpretable classification of small x-ray diffraction datasets using data augmentation and deep neural networks, npj Comput. Mater. **5**, 60 (2019).

[44] 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, Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, APL Mater. **1**, 011002 (2013).

[45] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees* (Routledge, New York, 1984).

[46] L. Breiman, Random forests, Mach. Learn. **45**, 5 (2001).

[47] T. Kohonen, The self-organizing map, Proc. IEEE **78**, 1464 (1990).

[48] P. M. Larsen, M. Pandey, M. Strange, and K. W. Jacobsen, Definition of a scoring parameter to identify low-dimensional materials components, Phys. Rev. Materials **3**, 034003 (2019).

[49] G. Cheon, K. A. N. Duerloo, A. D. Sendek, C. Porter, Y. Chen, and E. J. Reed, Data mining for new two- and one-dimensional weakly bonded solids and lattice commensurate heterostructures, Nano Lett. **17**, 1915 (2017).

[50] V. O. Özcelik, J. G. Azadani, C. Yang, S. J. Koester, and T. Low, Band alignment of two-dimensional semiconductors for designing heterostructures with momentum space matching, Phys. Rev. B **94**, 035125 (2016).

[51] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining, 2nd edition* (Pearson, New York, 2019).

[52] P. Auffinger, F. A. Hays, E. Westhof, and P. S. Ho, Halogen bonds in biological molecules, Proc. Natl. Acad. Sci. **101**, 16789 (2004).

[53] W. Zhang, R. Yu, H. J. Zhang, X. Dai, and Z. Fang, First-principles studies of the three-dimensional strong topological insulators Bi_2Te_3, Bi_2Se_3 and Sb_2Te_3, New J. Phys. **12**, 065013 (2010).

[54] S. Y. Xu, Q. Ma, H. Shen, V. Fatemi, S. Wu, T. R. Chang, G. Chang, A. M. M. Valdivia, C. K. Chan, Q. D. Gibson, J. Zhou, Z. Liu, K. Watanabe, T. Taniguchi, H. Lin, R. J. Cava, L. Fu, N. Gedick, and P. Jarillo-Herrero, Electrically switchable berry curvature dipole in the monolayer topological insulator WTe_2, Nat. Phys. **14**, 900 (2018).

[55] D. Garrido-Diez and I. Baraia, *Review of Wide Bandgap Materials and their Impact in New Power Devices*, Proc. 2017 IEEE Int. Workshop of ECMSM (IEEE, Donostia, 2017).

[56] F. Omnès, E. Monroy, E. Muñoz, and J. L. Reverchon, *Wide Bandgap UV Photodetectors: A Short Review of Devices and Applications*, Proc. SPIE, 6473E (SPIE, San Jose, 2007).

[57] H. Jin, L. Qin, L. Zhang, X. Zeng, and R. Yang, Review of wide band-gap semiconductors technology, MATEC Web Conf. **40**, 01006 (2016).

[58] P. G. Neudeck, R. S. Okojie, and L. Y. Chen, High-temperature electronics-a role for wide bandgap semiconductors? Proc. IEEE **90**, 1065 (2002).

[59] A. Chaves, J. G. Azadani, H. Alsalman, D. R. da Costa, R. Frisenda, A. J. Chaves, S. H. Song, Y. D. Kim, D. He, J. Zhou, A. Castellanos-Gomez, F. M. Peeters, Z. Liu, C. L. Hinkle, S. Oh, P. D. Ye, S. J. Koester, Y. H. Lee, P. Avouris, X. Wang, and T. Low, Bandgap engineering of two-dimensional semiconductor materials, npj 2D Mater. Appl. **4**, 29 (2020).

[60] Y. Lu and J. H. Warner, Synthesis and applications of wide bandgap 2D layered semiconductors reaching the green and blue wavelengths, ACS Appl. Electron. Mater. **2**, 1777 (2020).

[61] S. Jiang, J. Li, W. Chen, H. Yin, G. P. Zheng, and Y. Wang, InTeI: A novel wide-bandgap 2D material with desirable stability and highly anisotropic carrier mobility, Nanoscale **12**, 5888 (2020).

[62] Y. Yan, W. Xiong, S. Li, K. Zhao, X. Wang, J. Su, X. Song, X. Li, S. Zhang, H. Yang, X. Liu, L. Jiang, T. Zhai, C. Xia, J. Li, and Z. Wei, Direct wide bandgap 2D $GeSe_2$ monolayer toward anisotropic UV photodetection, Adv. Opt. Mater. **7**, 1900622 (2019).

[63] S. Weng, W. Zhen, Y. Li, X. Yan, H. Han, H. Huang, L. Pi, W. Zhu, H. Li, and C. Zhang, Air-stable wide-bandgap 2D semiconductor $ZnIn_2S_4$, Phys. Status Solidi RRL **14**, 2000085 (2020).

[64] B. Mortazavi and T. Rabczuk, Boron monochalcogenides; stable and strong two-dimensional wide band-gap semiconductors, Energies **11**, 1573 (2018).

[65] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and J. Vanderplas, scikit-learn: Machine learning in Python, J. Mach. Learn. Res. **12**, 2825 (2011).

[66] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevMaterials.6.043802 for the data and scripts used to produce the results in this paper.