Machine Learning Prediction of Thermodynamic Stability and Electronic Properties of 2D Layered Conductive Metal-Organic Frameworks

Zeyu Zhang,* Yuliang Shi, and Farnaz A. Shakib*

Department of Chemistry and Environmental Science, New Jersey Institute of Technology, Newark 07102, NJ United States

Received April 1, 2024; E-mail: zz389@njit.edu; shakib@njit.edu

Abstract: 2D layered metal-organic frameworks (MOFs) are a new class of multifunctional materials that can provide electrical conductivity on top of the conventional structural characteristics of MOFs, offering potential applications in electronics and optics. Here, for the first time, we employ Machine Learning (ML) techniques to predict the thermodynamic stability and electronic properties of layered electrically conductive (EC) MOFs, bypassing expensive ab initio calculations for the design and discovery of new materials. Proper feature engineering is a very important factor in utilizing ML models for such purposes. Here, we show that a combination of elemental features, using generic statistical reduction methods and crystal structure information curated from the recently introduced EC-MOF database, leads to a reasonable prediction of the thermodynamic and electronic properties of EC MOFs. We utilize these features in training a diverse range of ML classifiers and regressors. Evaluating the performance of these different models, we show that an ensemble learning approach in the form of stacking ML models can lead to higher accuracy and more reliability on the predictive power of ML to be employed in future MOF research.

1. Introduction

Technological advancements in recent years, along with geopolitical scarcity of resources, impose the need for the design and synthesis of complex multifunctional materials. Perpetual miniaturization of electronic devices, for example, has called for a transition from pure metal oxide semiconductors to hybrid organic-inorganic materials that can bring additional properties and complementary functionalities to the table. Due to their exceptionally large surface area, high porosity, and abundant void space, metal-organic frameworks (MOFs) have emerged as a new class of such multifunctional materials with applications in gas adsorption, separation, water treatment, and heterogeneous catalysis.[1–5] The advent of electrically-conductive (EC) MOFs in 2012[6] expanded the horizons of MOF research to electronics including their utilization as electrode materials in batteries[7] and supercapacitors[8] as well as chemiresistive sensors.[9] Nevertheless, the chemical versatility of MOFs, which is the result of the self-assembly of molecular building blocks via strong coordinative bonds,[10] has led to an increasingly large number of reported MOFs on a regular basis.[11] The unlimited and underexplored MOF universe makes the classical trial-and-error approach impractical for the design and discovery of targeted materials.

High-throughput screening (HTS) is a valuable tool for discovering targeted materials from existing databases, often through implementing quantum mechanical calculations. Multiple such databases exist for MOFs, e.g., CoRE MOF database,[12] hypothetical MOF database,[13] and EC-MOF database.[14] The latter has been developed very recently to assist the design and discovery of low-dimensional EC MOFs for use in next-generation electronic devices. EC-MOF is the only database dedicated to potentially conductive MOFs, which provides not only the DFT-minimized structures but also other relevant DFT-calculated properties of 1072 bulk and monolayer systems. This collection allows searching, for example, for a thermally stable metallic or semiconductor MOF for a desired application in an HT manner. However, currently, the EC-MOF database provides HTS opportunities only for 2D layered planar single-metal node EC MOFs. Expanding it to other classes, while desirable for a thorough EC MOF discovery, will inevitably lead to astronomical computational expenses due to the sheer number of potential EC MOFs,[15] and their intricate structures, which require careful benchmarking and care.[16–19] Hence, an imperative need arises for a more efficient and simplified methodology to compute and screen properties across various classes of EC MOFs.

Employing big data analysis and constructing predictive models through the application of machine learning (ML)[20,21] techniques is certainly an option worth exploring for this purpose. ML techniques have attracted much attention in screening conventional 3D MOFs for various purposes, including gas adsorption and storage,[22–28] however, their application for predicting the electronic properties of conductive MOFs is fewer and far between. In 2018, He *et al.*[29] applied ML techniques for identifying metallic MOF crystal structures for the first time. Due to the absence of a dedicated database for EC MOFs at the time, they trained four different ML algorithms using $\sim 52,300$ inorganic materials from the Open Quantum Materials Database[30] and used them to screen 2,932 3D structures in CoRE MOF database, identifying six metallic crystal structures. Following DFT calculations on the selected MOFs revealed an accuracy of 67% for their transfer learning approach. In 2021, Rosen *et al.*[31] introduced the QMOF database containing quantum-mechanical properties of 14,000 experimentally synthesized crystal structures. DFT-calculated band gaps showed an insulator nature for most of the structures, with a band gap peak around 2.9 eV, and a semi-conductor nature for a smaller subset of MOFs containing open-shell metal centers, with a band gap peak around 0.9 eV. By training different ML models on DFT-calculated band gaps using descriptors derived from the un-relaxed experimental crystal structures, they were able to predict band gap values with the Mean Absolute Error (MAE) as low as 0.274 eV using crystal graph convolutional neural network[32] (CGCNN).[31] The introduction of the QMOF database presented the research community with the opportunity to develop and eval-

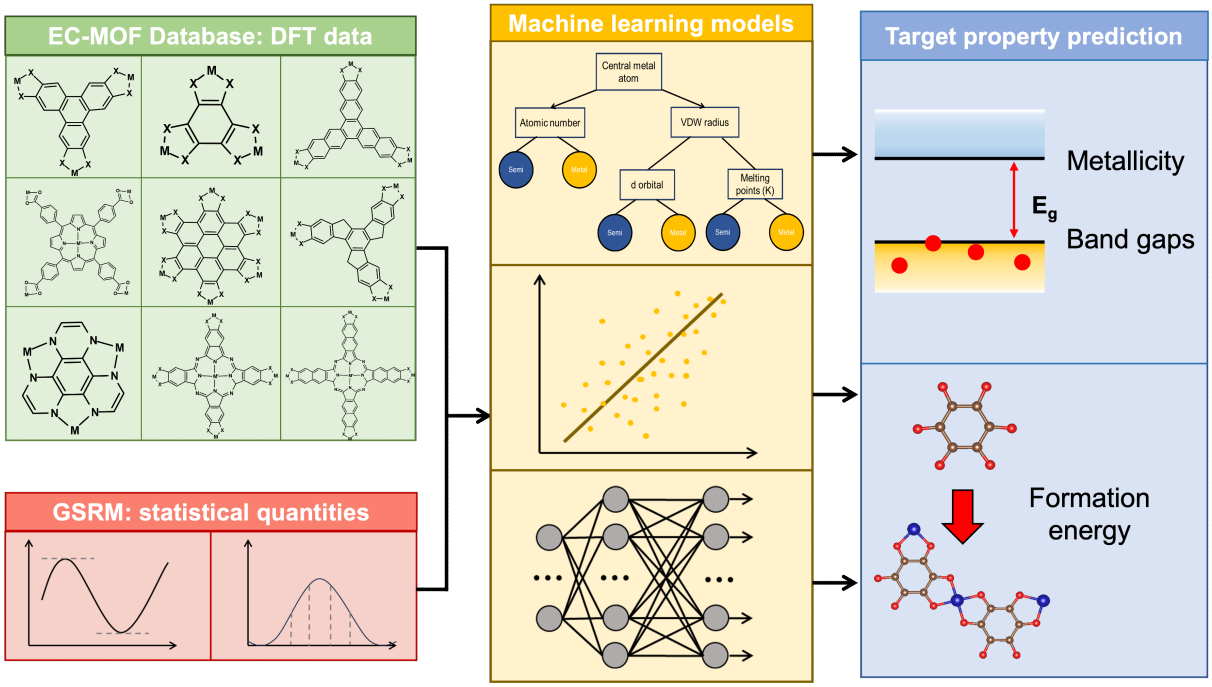

Figure 1. The overall schematics of the methodology employed in this work. A diverse set of linear, tree, and ensemble ML models takes DFT data from the EC-MOF database and statistical GSRM features as inputs. The trained model is then used for both classification and regression purposes.

uate novel ML algorithms to reduce the need for otherwise expensive DFT calculations. Very recently, Cao *et al.*[33] and Kang *et al.*[34] utilized the pre-training Transformer architecture,[35] to develop MOFormer and MOFTransformer, respectively, and used them to predict DFT band gaps gathered in the QMOF database among other properties. MOFormer, which only relies on a text string representation of MOFs, introduced as MOFid,[36] predicted band gaps with an MAE of 0.387 eV, which is considerably higher than the original CGCNN training by Rosen *et al.* The reason is the lack of information about the chemical environment of atoms in MOFormer. Pre-training by CGCNN to encode crystal data only slightly improved the performance of MOFormer, with an MAE of 0.367 eV. This, however, does not diminish the importance of predicting properties of interest using only a text string, which circumvents the need for obtaining 3D structures. On the other hand, MOFTransformer, which was pre-trained on one million hypothetical MOFs and then fine-tuned by training on a smaller number of 5,000-20,000 MOFs, slightly outperformed CGCNN in predicting DFT band gaps.[34]

Compared to conventional 3D MOFs, developing and applying ML techniques to predict the intrinsic properties of 2D EC MOFs is a challenging task. Not only does the system size and number of chemical species increase the training time exponentially, but also the intricate nature of their structures, which includes weak van der Waals interactions between the layers, makes this even more challenging. Furthermore, due to the high conductivity of the π-stacked layers, they show no or small band gaps ranging from 0 to ∼0.9 eV values calculated at the DFT level.[14] This makes the classification of EC MOFs or accurate prediction of their band gaps even more formidable. Here, we pursue targeted EC MOF design by employing ML prediction models trained using a combination of DFT data from our EC-MOF database and the generic statistical reduction methods (GSRM),[37] Figure 1. We conduct a thorough investigation using diverse traditional ML algorithms for the purpose of both classification, i.e., identifying metallic vs. semiconductor EC MOFs, and regression, i.e., calculating formation energies and electronic band gaps. Our results show that all employed ML models perform better in predicting structural and electronic properties when the DFT data from the EC-MOF database are utilized. Furthermore, we demonstrate that the ML property prediction can be considerably improved following a model stacking strategy. In the remainder of this work, we first provide the details of our ML training and, specifically, feature engineering on monolayer systems collected from the EC-MOF database. In section 3, we present the predicted properties from different individual ML models as well as stacked models for classification and regression tasks. Section 4 outlines future directions and concluding remarks.

2. Methodology

Feature engineering

To train an accurate ML model for specific purposes, it is critical to first find the appropriate features that sufficiently and adequately describe the studied systems so that the targetted properties can be reasonably deduced.[15] In this work, we create features and representations to utilize ML models in predicting metallicity, band gap values (E_g), and formation energies (E_f) of the 2D MOFs in the EC-MOF database. The schematic representation of different steps of our property prediction strategy is shown in Figure 1. First, we apply generic statistical reduction methods (GSRM)[37] to create representations of the unit cells based on elemental properties as shown in the Supporting Information (SI), Table S1. GSRM can generate representations regardless of the size or the elemental diversity of unit cells. A total number

of 17 elemental properties is selected, and 5 statistic quantities (standard mean, geometric mean, standard deviation, maximum value, and minimum value) are calculated for each of the properties.[38] Although it had been shown before that GSRM features perform well in predicting metallicity for inorganic materials with small unit cells,[38] they may not be sufficient for EC MOFs which contain more than 100 atoms per unit cell. Therefore, we complement GSRM features with the crystal structure information extracted from our EC-MOF database, Figure 1. A total number of 8 database features are curated, including six lattice parameters (a, b, c, α, β, and γ), cell volume, and total number of atoms in the unit cell. Additionally, a series of new features related to the electrical conductivity properties are created which are not a part of the published database. These features include bond ratios among different atoms, the total number of bonds, the distance between nearest metal atoms, and the number of d electrons; more details can be found in the SI, Section S1. As a result, the training set employed here consists of 102 columns of GSRM+Database (GD) features. These features are used to train various ML models for predicting target properties, including metallicity, class 0 for metal and class 1 for semiconductor, E_g, and E_f values, Figure 1. In predicting thermodynamic stability, we utilized a special feature engineering process called factor analysis. This method is able to reduce a large number of variables into a smaller number of factors while the variations of the whole data set are preserved. A series of tests between the number of factors and the coefficient of determination (R^2) during training is implemented. As a result, in predicting E_f values, the 102 feature columns are reduced to 50 new components to maximize the training score, R^2. Then, the 50 components are kept consistent in all ML regressors for E_f.

ML training based on the generated features

The data set used in this work is comprised of a matrix of 524 mono-layer structures from the EC-MOF database times 102 features for each system. It should be noted that the 102 columns of features are not in the same order of magnitude; for example, a feature like a row in the periodic table is a single digit, but another feature like the melting point could be up to three digits. Thus, features should be appropriately scaled into a reasonable range to eliminate any bias induced by different units or magnitudes. A data set of the target properties also consists of 524 rows according to the number of mono-layer structures and three target properties, metallicity (0 or 1), which is a classification task, and predicting E_g and E_f values, which is a regression task. The ratio of training/test set is set to 90%/10% to get a better training performance. In each case of training, various ML models are tested in order to investigate the best model for the specific task. It should be noted that there is no universal ML model for all the materials and properties. This highlights the importance of benchmarking different classes of ML models, i.e., linear, tree, and ensemble, as shown in Figure 1, for a new class of materials like 2D EC MOFs considered in this work. A 10-fold cross-validation (CV) test is adopted in all models for the training set. The model with the best performance is re-trained using the whole data set and is used for predictions related to hypothetical structures. Implementation of ML is carried out using the scikit-learn package (version 1.2.2).[39] The random state in all cases is set to 1 for reproducible results.

3. Results and Discussion

Thermodynamic stability of EC MOFs: Prediction of formation energies

Formation energy (E_f), which is an important indicator of the thermodynamic stability of the material and one of the most sought-after parameters for prediction,[40,41] can be computed using the following equation

$$E_f = E_{tot} - \frac{1}{N}\sum_{i=1}^{N} x_i \mu_i \quad (1)$$

where E_{tot} is the calculated total energy of the material, N is the total number of atoms with x_i and μ_i being the number and chemical potential of element i in the system. A negative value of E_f indicates a stable crystal structure, while a positive value will be an indicator of thermodynamic instability. The 2D crystal structures gathered in the EC-MOF database are composed of 13 different π-conjugated ligands coordinated to eight different metal ions with +2 oxidation states (Mn, Fe, Co, Ni, Cu, Zn, Pd, and Pt) via three possible functional groups (-NH, -O, and -S).[14] The DFT E_f values of all bulk and mono-layer structures are calculated at the level of Perdew-Burke-Ernzenhof (PBE) functional[42] with Grimme's damped D3 dispersion correction.[43] Hubbard U parameters[44] are adopted for a better description of the d and f electrons. Details of our DFT calculations can be found in Ref. 14. 96.06% of the bulk and 92.75% of the mono-layer structures in the EC-MOF database have negative E_fs. Overall, in each linker family, regardless of the metal type, EC MOFs with -NH functional groups possess the most negative E_fs, i.e., higher thermodynamic stability, whereas the ones with -S functional groups possess the least negative or, in some cases, even positive values.[14]

Using the DFT calculated E_f values of mono-layers in the EC-MOF database as target properties, we train a diverse collection of regressors among linear models, tree models, and ensemble models, as well as a few commonly employed complex models, to predict the thermodynamic stability of EC MOFs. For the linear regression models, we used Ridge Regressor (RR), Linear Regression (LIR), Passive Aggressive Regressor (PAR), and Stochastic Gradient Descent (SGD). These models utilize linear decision boundaries for regression and, obviously, are well-suited for linear problems. Our tree models include Extra Tree Regressor (ETR) and Random Forest Regressor (RFR), which leverage tree structures to model non-linear relationships, providing great flexibility in capturing the complex relationship between the data. Ensemble models include Gradient Boosted Decision Trees Regressor (GBR), Bagging Regressor (BR), and AdaBoost Regressor (ABR). These models are employed as they are suitable for processing data with highly non-linear and complex structures. Additionally, widely used models such as Support Vector Machine Regressor (SVR), k-Nearest Neighbors Regressor (KNR), and Neural Network Regressor (NNR) have been tested due to their demonstrated great performance in diverse ML tasks before. For every model, optimal settings listed in the SI, Table S3, are chosen based on the mean training accuracy of the 10-fold CV.

Supervised regression is employed to predict the formation energy of the mono-layer EC MOFs. To avoid the unbalanced weight of different features, a pre-process of scaling

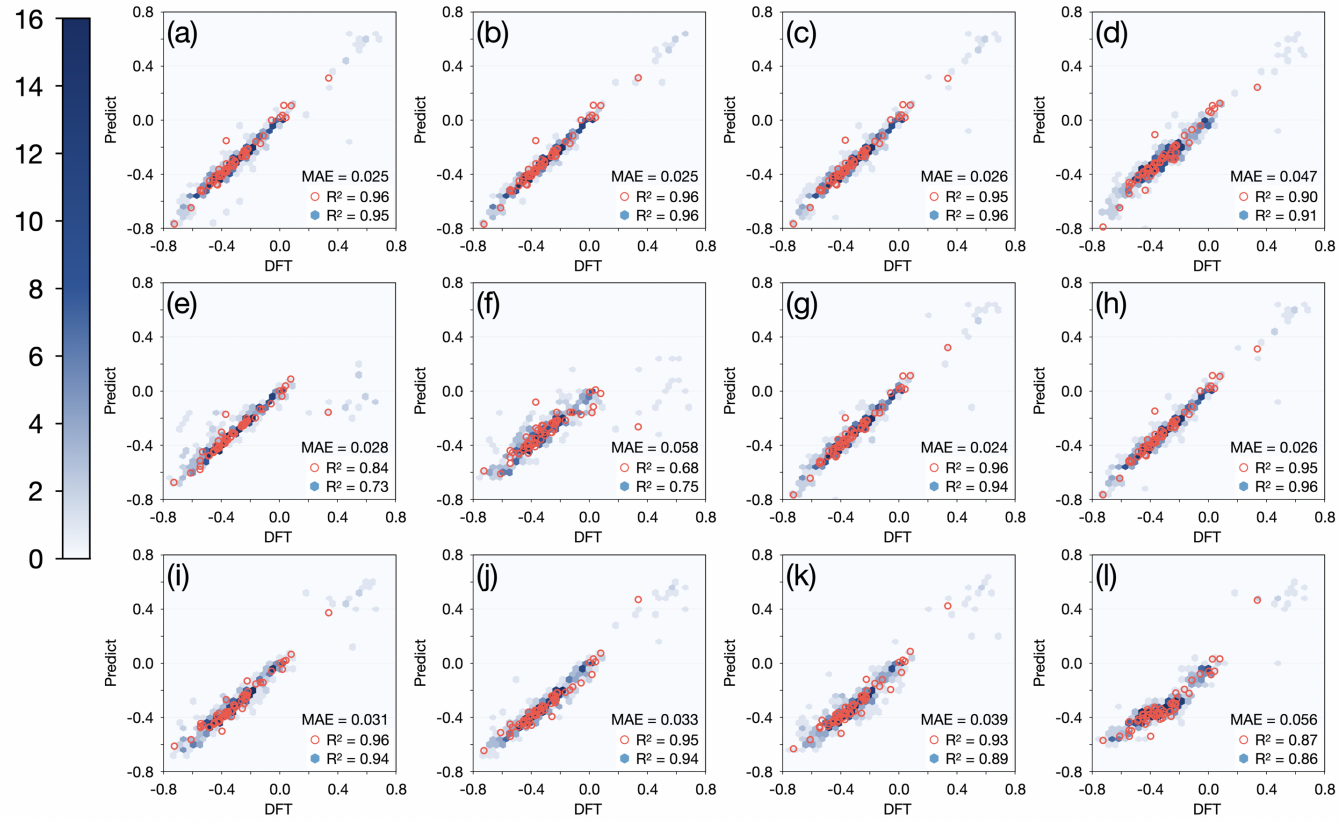

Figure 2. Parity plots of ML predicted E_f values vs. DFT calculated ones as obtained from different classes of regressors with the number of points located in one bin indicated by the color bar on the left. Coefficient of determination (R^2) in training (blue hexagonal bins) and test sets (red circles) for RR (a), LIR (b), SGD (c), PAR (d), SVR (e), KNR (f), NNR (g), RFR (h), ETR (i), GBR (j), BR (k) and ABR (l). The unit for E_f and MAEs is eV/atom. All models are trained based on GD features, i.e., elemental features from GSRM, together with crystal structure information from the EC-MOF database.

the data is implemented as dividing the data by the maximum value of the column ($x/|Max|$) so that all data drops into a reasonable range, $[-1, 1]$ without changing the sparsity. As stated in the section on "Feature engineering", an extra process of factor analysis is implemented to maximize the training score. Factor analysis method, which is a simple linear generative model with Gaussian latent variables,[45] is adopted for this purpose.

Figure 2 demonstrates the parity plots of the ML predicted E_f values vs. the reference DFT calculated ones as obtained from different classes of regressors using GD features, which contain both elemental and structural information. Predicted points are represented by cumulative hexa-bins in blue for the training set and in red for the test set. The number of points located in one bin is indicated by the color bar on the left. The coefficient of determination (R^2) for both training and the test set, as well as the mean absolute error (MAE) for the test set, is given for each model. All plots show a nearly linear relationship between DFT-calculated E_f values and their ML-predicted ones. Three out of the four linear regression models (RR, LIR, and SGD) perform best in this task with a test set R^2 of ≥ 0.95 and an MAE as low as 0.026 eV/atom. The remaining linear model PAR, although it performs poorer than the other three, still shows a good R^2 of ≥ 0.90 and a relatively low MAE of 0.047 eV/atom. Similarly, the ETR and RFR tree models produce an excellent test set R^2 of ≥ 0.95 and an MAE as low as 0.026 eV/atom. The ensemble regressors show a varied behavior with the test set R^2 covering a range from 0.86 in ABR to 0.94 in GBR. Among the three commonly used regressors, only NNR has as high an accuracy as linear and tree models, while SVR and kNR show the lowest test set R^2 of 0.73 and 0.75 among the studied models.

A discussion on feature importance

Overall, the high test set R^2 values and low test set MAE values obtained in most of the ML models can be credited to the created GD features that properly describe the systems and distinguish the differences among them. To investigate the importance of the chemical composition features in GSRM vs. the crystal structure features curated from the EC-MOF database in predicting the thermodynamic stability of EC MOFs, all models were trained using the GSRM features only. A complete comparison between training/test set R^2 and MAE values of the ML models trained with GD vs. GSRM features are presented in the SI, Table S5, and S6. Utilizing GSRM features only, one can produce a maximum R^2 value of ∼0.85 for the three linear models (RR, LIR, and SGD) as well as the tree model RFR and the commonly-used model NNR. The rest of the models show an average R^2 value of ∼0.56 with the lowest value of 0.27 calculated for the ensemble model ABR. Generally, ML models trained on GSRM features show a lower R^2 and a higher MAE than those trained on GD features. This emphasizes the importance of including crystal structure information for creating predictive ML models for the thermodynamic

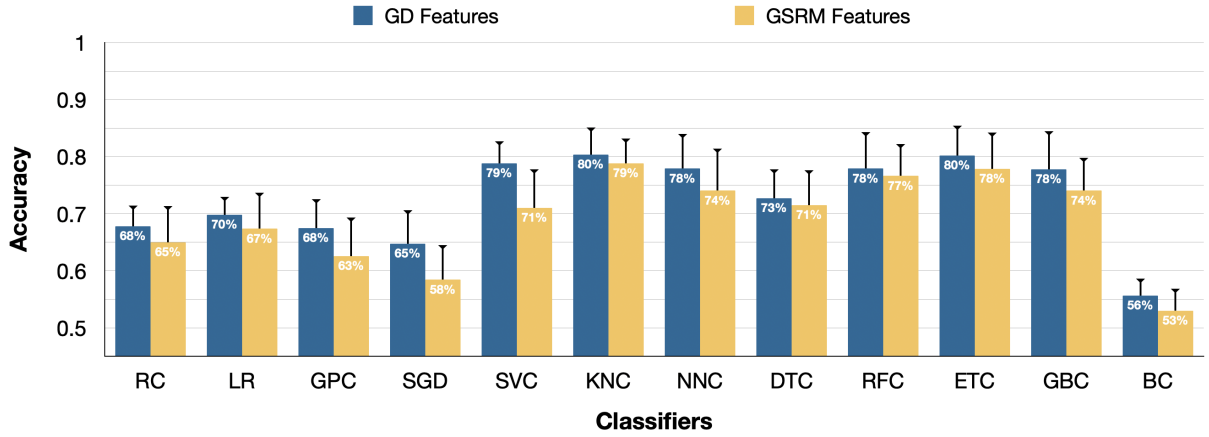

Figure 3. Accuracy and standard deviation of the tested ML classification models using GSRM+Database (GD) features (dark blue) and GSRM features only (yellow). The accuracy is defined as the percentage of the correctly predicted metallic structures in the entire training set with respect to the reference DFT data. The standard deviation is shown as error bars.

properties of EC MOFs.

Electrical conductivity nature of EC MOFs

For the classification of EC MOFs as metallic and semiconductors, we evaluated the performance of linear Ridge Classification (RC), Logistic Regression (LR), Gaussian Process Classification (GPC), and Stochastic Gradient Descent (SGD). Tree models include Decision Trees classifier (DTC), Extra Tree classifier (ETC), and Random Forest Classifier (RFC). Finally, the ensemble models considered in this work include the Bagging Classifier (BC) and Gradient Boosted Decision Trees Classifier (GBC). As before, commonly used Support Vector Machine Classifier (SVC), k-Nearest Neighbors classifier (KNC), and Neural Network Classifier (NNC) were also adopted to test their performance. For every model, the optimal setting listed in the SI, Table S4, is chosen based on the mean training accuracy of the 10-fold CV. The scaling process is implemented as before. By observing the distribution of the *whole data set*, there are 165 metallic MOFs and 359 semiconductors, indicating a slightly imbalanced distribution. An over-sampling strategy, adaptive synthetic sampling, is adopted to balance the *training set* and avoid any bias from the imbalanced distribution.[46] This is done using the imbalanced-learn package.[47] The over-sampling strategy results in a near-even ratio of 337 metallic and 319 semiconductor structures.

While half of the bulk layered structures in the EC-MOF database show a metallic character, nano-structuration leads to creation of semiconductors with narrow band gaps of up to $\leq$ 0.9 eV.[14] Figure 3 shows the accuracy of all chosen classifiers, trained on GD or GSRM features, in determining the metallic nature of EC MOFs with respect to the DFT reference data. The standard deviation (SD) among 10 folds of the CV process is calculated and represented as an error bar for each ML model. As can be seen, all trained classifiers show an accuracy of higher than 50% compared to the DFT classification in the EC-MOF database. Overall, training using GD features shows an accuracy range of 56% (in BC) to 80% (in KNC and ETC). These values are only slightly higher than the accuracy achieved by training ML models using GSRM features, which is 53% (in BC) to 79% (in KNC). Trained ML models based on GD features generally have slightly smaller errors than the GSRM ones. Overall, the similar accuracy achieved by utilizing GD or GSRM features shows that the chemical composition has a more pronounced effect on inducing electronic properties than the crystal structure. However, one should remember that training and prediction are carried out for mono-layer EC MOFs here. In the case of a bulk layered system, the $\pi - \pi$ interaction between layers creates a dominant charge transfer pathway, drastically affecting the nature of electrical conductivity in EC MOFs. Hence, the GD features are still recommended over GSRM for reaching a universally reliable ML prediction. The results of the classification of the test set were not satisfactory with any of the trained ML models. Hence, extra measures, as explained in the next section, should be taken to increase the accuracy and create a predictive tool for practical applications.

Stacking ML models to create a predictive tool

To create a predictive ML tool, we adopt a form of *ensemble learning* approach, which entails stacking different ML classifiers to construct the final model. This idea was first reported in 1992 under the name stacking generalization.[48] Once some first-level classifiers are trained using the corresponding training set, a set of predictions can be collected from these basic classifiers. Based on this predicted set of properties along with the original training set, a second-level classifier will be trained, which is expected to have better performance than all first-level classifiers. Such a process also reduces biases and avoids overfitting. Hence, we expect a better performance from the stacking approach than the statistical multivoting used before,[38] which reached an accuracy of 67% in the classification of 3D EC MOFs to metallic and nonmetallic structures.

Here, we evaluate the performance of four distinct stacking models (SC) in predicting the electrically conductive nature of 2D MOFs. Initially, we combine RC, KNC, GBC, and NNC models (SC1) to ensure model diversity and adaptability. Among the four classifiers, RC can handle linear problems well and contributes to lightweight SC. KNC, GBC, and NNC are, respectively, grounded in distance metrics, gradient boosting decision trees, and multilayer perception algorithms. It is worth checking to see whether stacking these different classes can deal with diverse problems more effectively. The accuracy results are reported in Figure 4 for

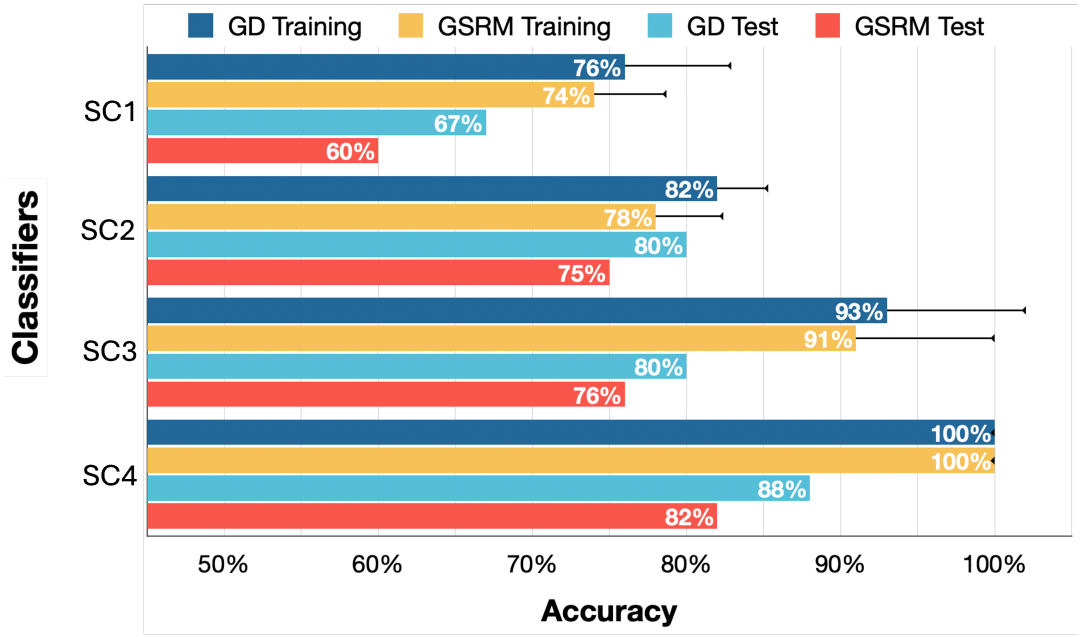

Figure 4. Performance of different combinations of stacking classifiers in training and test sets.

both training and test sets based on GD features, or GSRM features only. With 76% accuracy in classifying the training set using GD features, SC1 does not show an improvement over its constituent models RC, KNC, NNC, and GBC, each showing 68%, 80%, 78%, and 78% accuracy, respectively. On the other hand, the SC1 model using GD features reaches an underwhelming accuracy of 67% in classifying the test set as metallic or semiconductor. Finally, SC1 shows a high standard deviation for the training set; see the error bars in Figure 4.

The second SC model that we explore is based on combining the SGD, ETC, and SVC classifiers (SC2). ETC and SVC are tailored for non-linear scenarios, while SGD is used to handle linear problems. Stacking these three classifiers enhances the robustness of the model, while the inclusion of SVC augments its performance in high-dimensional spaces. As shown in Figure 4, SC2 provides an accuracy of 82% for the training set, which is similar to the accuracy of its non-linear constituents, i.e., 80% for ETC and 79% for SVC and considerably improved upon its liner component, i.e., 65%. Most importantly, it shows a good accuracy of 80% for the test set, which is 23% higher than what SC1 achieved. The standard deviation is considerably lower than the SC1. Overall, these results show the robustness of the SC2 model compared to the SC1.

To achieve even higher accuracies, we combined three ensemble models, GBC, RFC, and BC (SC3). This stacking leads to an increase of the training accuracy to 93%, which is the highest among the three SC models at the expense of the highest standard deviation as well. It is a stark improvement in comparison to the BC model, which produced an underwhelming accuracy of 56% by itself. Similar to SC2, SC3 reaches a good accuracy of 80% for the test set. Overall, the results of stacking up till now emphasize that our data relationships are mainly non-linear in high-dimensional space, and it is difficult for linear models to effectively capture the complex nature of the data set. To describe high-latitude nonlinear relationships more effectively, we include SVC, ETC, RFC, and DTC in the fourth SC model (SC4). We expect that SC4 will ensure not only the robustness of the model but also the diversity of the algorithm to better handle complex data relationships. Figure 4 shows that with SC4, the training accuracy reaches 100% regardless of the choice of input features, whether it be GD features or GSRM alone. However, the performance on the test set differs in that the SC trained by GD features results in 88% accuracy compared to the 82% of the SC trained on GSRM features only. Overall, we show that stacking different ML models is a formidable strategy for creating a predictive tool for discerning the electrical conductivity nature of EC MOFs, even in the face of a database with a moderate size.

Prediction of electronic band gaps

After the classification of EC MOFs based on their metallicity, we use different ML models to predict the band gap values (E_g) of the semiconductor materials. E_g values are significantly important descriptors to be determined for semiconductors as they affect the charge transport mechanism and the overall electrical conductivity of the material. We enforce positive E_g values by assigning "Positive" keyword as "True", if applicable. Notably, the range of DFT-calculated E_g distribution among mono-layer structures from the EC-MOF database is relatively narrow, i.e., less than 1 eV with ∼60% of all mono-layer structures having E_g values less than 0.3 eV. Considering the accuracy of DFT calculations with common functionals is $\sim 2-3$ kcal/mol ($\sim 0.087-0.130$ eV/particle),[49] it is unnecessary (or rather meaningless) for the regression model to be perfectly linear between the ML predicted values and DFT reference data. In this case, the MAE is a more appropriate evaluator of regression tasks, which is an unambiguous measurement of the average error magnitude. The optimal settings for ML training, reported in the SI Table S5, are chosen according to the MAE in eV for training and test sets. The 359 semiconductors form another 359×102 training matrix for regression models. The same scaling process as the classification task is implemented to avoid possible bias. The

Table 1. Metrics and scoring in predicting band gap values for the mono-layer structures in the EC-MOF database.

	Trainig MAE (eV)	Test MAE (eV)	Training R^2	Test R^2
SR1	0.031	0.065	0.905	0.515
SR2	0.027	0.068	0.919	0.482
SR3	0.029	0.065	0.909	0.541

MAE of training and test sets among all tested regressors, as well as the coefficients of determination R^2, are reported in the SI, Table S8. The best regressor concluded from the table is ETR with 0.069/0.063 eV MAEs and 0.511/0.542 R^2 values for training/test sets, respectively. The R^2 values of all other regressors are much lower than 0.5, even zero in some cases, which indicates a weak learning rate and predictive ability in such tasks. Hence, to have a reliable tool, regardless of the type of ML model, we restore the strategy of stacking ML models for predicting the very small E_gs in EC MOFs.

Firstly, we choose the single regressors with the highest R^2 in training. These include ETR, RFR, GBR, BR, and ABR (SR1 in Table 1), which have an average R^2 of 0.45 in training. The performance during training is greatly improved, with an R^2 of 0.905, showing better learning ability of the stacking regressor (SR). The test score reaches 0.515 with an MAE of 0.065 eV. Considering the limits of DFT in evaluating band gaps, as mentioned before, as well as experimentally measured values, which can vary by several tenths of an electronvolt, an MAE of less than 0.1 eV is very promising for predicting hypothetical EC MOF band gaps. Secondly, we adopt different types of regressors, including RR (a linear model), SVR (a support vector machine model), KNR (a nearest-neighbor model), NNR (a neural network), and GBR (a tree model). The results are collected as SR2 in Table 1. The first four regressors have an underwhelming average R^2 value of 0.137 in training, while the last one scores 0.474. SR2 reaches an R^2 value of 0.919 in training and 0.482 in the test set, which is a drastic improvement over its constituent models. Similar to SR1, it shows a very promising MAE of 0.068 eV for deciphering structure-function relationships in future EC MOF studies. The last SR model, SR3, comprises three linear models, LR, RR, and PAR, which are by no means promising regressors based on their individual performances, i.e., they have scored as low as zero in training or test set and produced the highest MAEs individually. Surprisingly, SR3 gives the highest training R^2, 0.966, and a comparative R^2 in the test set, 0.504. Such improvement in the case of the unproductive single regressors serves to recognize the prowess of the stacking strategy.

Property predictions for a hypothetical MOF

Implementing ML within the database is the first step in validating the effectiveness of the created features and ML models for future investigations on EC MOFs. Naturally, the final step is to implement the developed ML models to predict important target properties of hypothetical systems without the need for expensive DFT calculations. To investigate whether the ML models used here are transferable to the MOFs not included in the EC-MOF database, we create a class of hypothetical MOFs based on the organic linker depicted in Figure 5. This organic linker is a modified form of 1,4,5,8,9,12-hexaazatriphenylene (HAT) linker that has already been used to synthesize EC MOFs in 2022.[50]

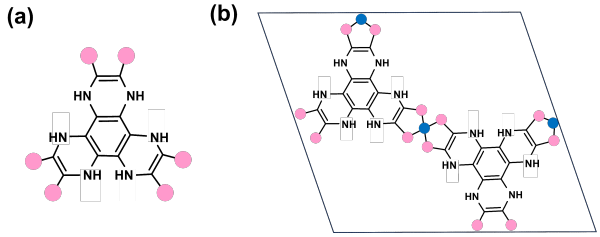

Figure 5. The structures of the hypothetical organic linker (a) and the hypothetical MOFs (b). The pink spheres represent different functional groups, -O, -NH, and -S, while the blue spheres represent Mn, Fe, Co, Ni, Cu, Zn, Pd, and Pt metal nodes, all in +2 oxidation states.

The pink spheres represent functional groups, -O, -NH, and -S, while the blue spheres represent Mn, Fe, Co, Ni, Cu, Zn, Pd, and Pt metal nodes, all in +2 oxidation states. We utilize Crystal Structure Producer (CrySP) developed in our previous study[14] to survey all possible combinations among the building blocks. As a result, a total of 24 hypothetical EC MOFs are built. For the input features of the hypothetical MOFs, 102 columns of features are generated in the same manner as mentioned before based on the optimized structures. Metallicity, E_g, and E_f are our target properties. The best-trained models, SC4, SR2, and SGD regressor, are chosen to make predictions on metallicity, E_g and E_f, respectively. The SC classifier predicts that among the 24 hypothetical MOFs, six are metallic, and 18 are semiconductor systems. We carried out reference DFT calculations via an HTS approach where one metallic and 23 semiconductor MOFs were found. As a result, the accuracy of metallicity classification was found to be 79%. Figure 6 shows the E_g values of the hypothetical MOFs obtained from the SR2 as well as E_f values predicted by the SGD regressor, both trained using the GD features. As can be seen from the color bar, the ML predicted E_g values range from 0 to 0.230 eV. For comparison, the DFT calculated results are also shown alongside the predicted results by the ML models. Details of DFT calculations are available in the SI, section S5. The MAE of the ML predictions is 0.102 eV, which is still within the accuracy of DFT calculations, indicating the acceptable transferability of our ML models. Regarding the formation energies, E_f values of all hypothetical structures were predicted by SGD regressor using the same dimensional reduction procedure as mentioned before. The MAE of E_f prediction was found to be 0.087 eV/atom. The pattern of formation energies among different functional groups deduced from the ML predicted E_f values is that systems with imino and thio groups correspond to the most and least stable MOFs, respectively. Such a pattern is also confirmed by the DFT calculations here. It is worth mentioning that this is the same pattern as reported for the EC-MOF database. Overall, these results show that we can predict the properties of hypothetical structures using our created GD features and ML models with acceptable accuracy and without extra DFT calculations.

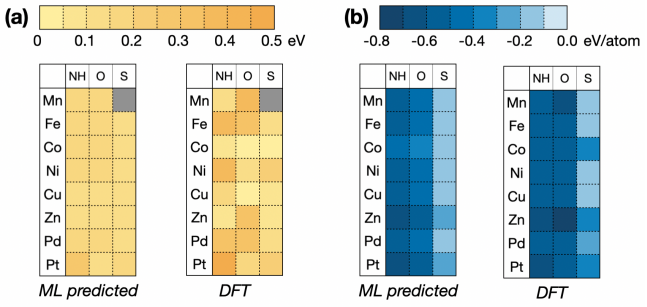

Figure 6. Comparison of band gap values (a) and formation energies (b) of hypothetical MOFs calculated by DFT and ML models.

4. Concluding Remarks

In this article, for the first time, we explored the application of ML techniques for predicting different properties of less explored 2D layered EC MOFs. With a focus on the importance of feature engineering, we combined elemental features from GSRM with DFT-calculated crystal structure properties gathered in the EC-MOF database, collectively coined GD features. This recently developed database, which is the first of its kind, contains the geometries as well as structural and electronic properties of 1057 bulk and mono-layer EC MOFs. Using the created GD features for mono-layer systems, three machine learning (ML) tasks, i.e., classification of metallicity as well as prediction of E_g and E_f values, are carried out with multiple ML models. Overall, the GD features universally improved the performance of all tested ML models. Even though the GSRM features alone perform well in predicting the electronic properties of monolayer systems, they cannot be generalized to bulk materials where electronic properties will be drastically affected by the interlayer van der Waals interactions. As such, GD features are overall recommended for achieving universally good predictive performances for a range of different properties. We have also explored the strategy of stacking ML models to create a more reliable predictive tool. The results show high accuracy for metallicity classification, with the stacking classifiers producing an accuracy of up to 100% for the training set and 88% for the test set. The mean absolute errors (MAE) of ML predicted E_g values of less than 0.07 eV for the test set are smaller than the accuracy limit of DFT calculations. The R^2 values of E_f predicting are higher than 0.95 in both training and test sets for most of the ML models. All the results indicate that by integrating the features from the database and GSRM, accurate predictions of electronic and energetic properties can be achieved using a limited amount of data and simple ML models compared to other more complex models, such as convolutional neural networks. Additionally, 24 hypothetical MOFs are built and are subjected to the best ML models to predict target properties. The trained ML models show promising transferability to the MOFs that are not part of the EC-MOF database and were not introduced in the training process. This proof of concept paves the way for the application of ML models as promising tools for future accelerated EC MOF design and discovery.

Future work should be directed toward improving the transferability of the ML models. This can be pursued via (i) enhanced feature engineering so a specific ML model shows excellent performance for predicting a full spectrum of structural, electronic, and optical properties; or (ii) inclusion of bulk structures in the training set which needs further feature creations for describing interlayer chemistry.

Supporting Information Available

The details of feature generations and the parameters utilized in training ML models are available in the supporting information accompanying this manuscript. DFT-calculated properties of EC MOFs can be found in our online EC-MOF database at https://ec-mof.njit.edu.

Acknowledgement This research was supported by the National Science Foundation via award no. CBET-2302617. Technical support and computing resources provided by the HPC center at NJIT are gratefully acknowledged.

References

(1) Farha, O. K.; Yazaydın, A. Ö.; Eryazici, I.; Malliakas, C. D.; Hauser, B. G.; Kanatzidis, M. G.; Nguyen, S. T.; Snurr, R. Q.; Hupp, J. T. De novo synthesis of a metal–organic framework material featuring ultrahigh surface area and gas storage capacities. **2010**, *2*, 944–948.

(2) Farha, O. K.; Eryazici, I.; Jeong, N. C.; Hauser, B. G.; Wilmer, C. E.; Sarjeant, A. A.; Snurr, R. Q.; Nguyen, S. T.; Yazaydın, A. Ö.; Hupp, J. T. Metal–organic framework materials with ultrahigh surface areas: is the sky the limit? *J. Am. Chem. Soc.* **2012**, *134*, 15016–15021.

(3) Chen, Z.; Li, P.; Anderson, R.; Wang, X.; Zhang, X.; Robison, L.; Redfern, L. R.; Moribe, S.; Islamoglu, T.; Gómez-Gualdrón, D. A., et al. Balancing volumetric and gravimetric uptake in highly porous materials for clean energy. *Science* **2020**, *368*, 297–303.

(4) Yang, F.; Du, M.; Yin, K.; Qiu, Z.; Zhao, J.; Liu, C.; Zhang, G.; Gao, Y.; Pang, H. Applications of Metal-Organic Frameworks in Water Treatment: A Review. *Small* **2022**, *18*, 2105715.

(5) Furukawa, H.; Cordova, K. E.; O'Keeffe, M.; Yaghi, O. M. The chemistry and applications of metal-organic frameworks. *Science* **2013**, *341*, 1230444.

(6) Hmadeh, M. et al. New porous crystals of extended metal-catecholates. *Chem. Mater.* **2012**, *24*, 3511–3513.

(7) Cai, D.; Lu, M.; Li, L.; Cao, J.; Chen, D.; Tu, H.; Li, J.; Han, W. A highly conductive MOF of graphene analogue Ni3 (HITP) 2 as a sulfur host for high-performance lithium–sulfur batteries. *Small* **2019**, *15*, 1902605.

(8) Muzaffar, N.; Afzal, A. M.; Hegazy, H.; Iqbal, M. W. Recent advances in two-dimensional metal-organic frameworks as an exotic candidate for the evaluation of redox-active sites in energy storage devices. *J. Energy Storage* **2023**, *64*, 107142.

(9) Campbell, M. G.; Sheberla, D.; Liu, S. F.; Swager, T. M.; Dincă, M. Cu3 (hexaiminotriphenylene) 2: an electrically conductive 2D metal–organic framework for chemiresistive sensing. *Angew. Chem. Int. Ed.* **2015**, *54*, 4349–4352.

(10) Yaghi, O. M.; O'Keeffe, M.; Ockwig, N. W.; Chae, H. K.; Eddaoudi, M.; Kim, J. Reticular synthesis and the design of new materials. *Nature* **2003**, *423*, 705–714.

(11) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr. B: Struct. Sci. Cryst. Eng. Mater.* **2016**, *72*, 171–179.

(12) Chung, Y. G.; Camp, J.; Haranczyk, M.; Sikora, B. J.; Bury, W.; Krungleviciute, V.; Yildirim, T.; Farha, O. K.; Sholl, D. S.; Snurr, R. Q. Computation-Ready, Experimental Metal–Organic Frameworks: A Tool to Enable High-Throughput Screening of Nanoporous Crystals. *Chem. Mater.* **2014**, *26*, 6185–6192.

(13) Boyd, P. G.; Chidambaram, A.; García-Díez, E.; Ireland, C. P.; Daff, T. D.; Bounds, R.; Gładysiak, A.; Schouwink, P.; Moosavi, S. M.; Maroto-Valer, M. M., et al. Data-driven design of metal–organic frameworks for wet flue gas CO2 capture. *Nature* **2019**, *576*, 253–256.

(14) Zhang, Z.; Valente, D. S.; Shi, Y.; Limbu, D. K.; Momeni, M. R.; Shakib, F. A. In Silico High-Throughput Design and Prediction of Structural and Electronic Properties of Low-Dimensional Metal–Organic Frameworks. *ACS Appl. Mater. Interfaces* **2023**, *15*, 9494–9507.

(15) Chong, S.; Lee, S.; Kim, B.; Kim, J. Applications of machine learning in metal-organic frameworks. *Coord. Chem. Rev.* **2020**, *423*, 213487.

(16) Sakaida, S.; Otsubo, K.; Sakata, O.; Song, C.; Fujiwara, A.; Takata, M.; Kitagawa, H. Crystalline coordination framework endowed with dynamic gate-opening behaviour by being downsized to a thin film. *Nature Chemistry* **2016**, *8*.

(17) Sakaida, S.; Haraguchi, T.; Otsubo, K.; Sakata, O.; Fujiwara, A.; Kitagawa, H. Fabrication and Structural Characterization of an Ultrathin Film of a Two-Dimensional-Layered Metal–Organic Framework, Fe(py) 2 [Ni(CN) 4] (py = pyri-

dine). *Inorganic Chemistry* **2017**, *56*.
(18) Liu, C.-M.; Zhang, D.-Q.; Hao, X.; Zhu, D.-B. A Chinese Pane-Like 2D Metal- Organic Framework Showing Magnetic Relaxation and Luminescence Dual-Functions. *Scientific Reports* **2017**, 7.
(19) Koitz, R.; Iannuzzi, M.; Hutter, J. Building Blocks for Two-Dimensional Metal-Organic Frameworks Confined at the Air-Water Interface: An Ab Initio Molecular Dynamics Study. *The Journal of Physical Chemistry C* **2015**, *119*, 150204162325009.
(20) Dral, P. O. Quantum chemistry in the age of machine learning. *J. Phys. Chem. Lett* **2020**, *11*, 2336–2347.
(21) Shen, B.; Gnanasambandam, R.; Wang, R.; Kong, Z. J. Multi-task Gaussian process upper confidence bound for hyperparameter tuning and its application for simulation studies of additive manufacturing. *IISE Trans.* **2023**, *55*, 496–508.
(22) Ohno, H.; Mukae, Y. Machine Learning Approach for Prediction and Search: Application to Methane Storage in a Metal–Organic Framework. *J. Phys. Chem. C* **2016**, *120*, 23963–23968.
(23) Anderson, R.; Rodgers, J.; Argueta, E.; Biong, A.; Gómez-Gualdrón, D. A. Role of Pore Chemistry and Topology in the CO2 Capture Capabilities of MOFs: From Molecular Simulation to Machine Learning. *Chem. Mater.* **2018**, *30*, 6325–6337.
(24) Bucior, B. J.; Bobbitt, N. S.; Islamoglu, T.; Goswami, S.; Gopalan, A.; Yildirim, T.; Farha, O. K.; Bagheri, N.; Snurr, R. Q. Energy-based descriptors to rapidly predict hydrogen storage in metal–organic frameworks. *Mol. Syst. Des. Eng.* **2019**, *4*, 162–174.
(25) Dureckova, H.; Krykunov, M.; Aghaji, M. Z.; Woo, T. K. Robust Machine Learning Models for Predicting High CO2 Working Capacity and CO2/H2 Selectivity of Gas Adsorption in Metal Organic Frameworks for Precombustion Carbon Capture. *J. Phys. Chem. C* **2019**, *123*, 4133–4139.
(26) Wang, R.; Zhong, Y.; Bi, L.; Yang, M.; Xu, D. Accelerating discovery of metal–organic frameworks for methane adsorption with hierarchical screening and deep learning. *ACS Appl. Mater. Interfaces* **2020**, *12*, 52797–52807.
(27) Sheng, L.; Wang, Y.; Mou, X.; Xu, B.; Chen, Z. Accelerating Metal–Organic Framework Selection for Type III Porous Liquids by Synergizing Machine Learning and Molecular Simulation. *ACS Appl. Mater. Interfaces* **2023**, *15*, 56253–56264.
(28) Wang, Y.; Jiang, Z.-J.; Wang, D.-R.; Lu, W.; Li, D. Machine Learning-Assisted Discovery of Propane-Selective Metal–Organic Frameworks. *J. Am. Chem. Soc* **2024**, *146*, 6955–6961.
(29) He, Y.; Cubuk, E. D.; Allendorf, M. D.; Reed, E. J. Metallic Metal–Organic Frameworks Predicted by the Combination of Machine Learning Methods and Ab Initio Calculations. *The Journal of Physical Chemistry Letters* **2018**, *9*, 4562–4569, PMID: 30052453.
(30) Saal, J. E.; Kirklin, S.; Aykol, M.; Meredig, B.; Wolverton, C. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM* **2013**, *65*, 1501–1509.
(31) Rosen, A. S.; Iyer, S. M.; Ray, D.; Yao, Z.; Aspuru-Guzik, A.; Gagliardi, L.; Notestein, J. M.; Snurr, R. Q. Machine learning the quantum-chemical properties of metal-organic frameworks for accelerated materials discovery. *Matter* **2021**, *4*, 1578–1597.
(32) Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Physical Review Letters* **2018**, *120*, Cited by: 1052; All Open Access, Green Open Access.
(33) Cao, Z.; Magar, R.; Wang, Y.; Barati Farimani, A. MOFormer: Self-Supervised Transformer Model for Metal–Organic Framework Property Prediction. *J. Am. Chem. Soc.* **2023**, *145*, 2958–2967.
(34) Kang, Y.; Park, H.; Smit, B.; Kim, J. A multi-modal pre-training transformer for universal transfer learning in metal–organic frameworks. *Nat. Mach. Intell.* **2023**, *5*, 309–318.
(35) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, 5998–6008.
(36) Bucior, B. J.; Rosen, A. S.; Haranczyk, M.; Yao, Z.; Ziebel, M. E.; Farha, O. K.; Hupp, J. T.; Siepmann, J. I.; Aspuru-Guzik, A.; Snurr, R. Q. Identification schemes for metal–organic frameworks to enable rapid search and cheminformatics analysis. *Cryst. Growth Des.* **2019**, *19*, 6682–6697.
(37) Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *Npj Comput. Mater.* **2016**, *2*, 1–7.
(38) He, Y.; Cubuk, E. D.; Allendorf, M. D.; Reed, E. J. Metallic metal–organic frameworks predicted by the combination of machine learning methods and ab initio calculations. *J. Phys. Chem. Lett* **2018**, *9*, 4562–4569.
(39) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
(40) Rasmussen, F. A.; Thygesen, K. S. Computational 2D materials database: electronic structure of transition-metal dichalcogenides and oxides. *J. Phys. Chem. C* **2015**, *119*, 13169–13183.
(41) Mao, Y.; Yang, H.; Sheng, Y.; Wang, J.; Ouyang, R.; Ye, C.; Yang, J.; Zhang, W. Prediction and classification of formation energies of binary compounds by machine learning: an approach without crystal structure information. *ACS Omega* **2021**, *6*, 14533–14541.
(42) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865.
(43) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate Ab initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *Chem. Phys.* **2010**, *132*, 154104.
(44) Anisimov, V. I.; Zaanen, J.; Andersen, O. K. Band theory and Mott insulators: Hubbard U instead of Stoner I. *Phys. Rev. B* **1991**, *44*, 943–954.
(45) Barber, D. *Bayesian reasoning and machine learning*; Cambridge University Press, 2012.
(46) He, H.; Bai, Y.; Garcia, E. A.; Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). 2008; pp 1322–1328.
(47) Guillaume, L.; Fernando, N.; Christos, K. A. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *J. Mach. Learn. Res.* **2017**, *18*, 1–5.
(48) Wolpert, D. H. Stacked generalization. *Neural Netw.* **1992**, *5*, 241–259.
(49) Bogojeski, M.; Vogt-Maranto, L.; Tuckerman, M. E.; Müller, K.; Burke, K. Quantum chemical accuracy from density functional approximations via machine learning. *Nat. Commun.* **2020**, *11*, 5223.
(50) Lyu, C.; Gao, Y.; Gao, Z.; Mo, S.; Hua, M.; Li, E.; Fu, S.; Chen, J.; Liu, P.; Huang, L.; Lin, N. Synthesis of Single-Layer Two-Dimensional Metal–Organic Frameworks $M_3(HAT)_2$ (M= Ni, Fe, Co, HAT= 1, 4, 5, 8, 9, 12-hexaazatriphenylene) Using an On-Surface Reaction. *Angew. Chem. Int. Ed.* **2022**, *134*, e202204528.

Graphical TOC Entry