ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Machine Learning-Based Prediction of Crystal Systems and Space Groups from Inorganic Materials Compositions

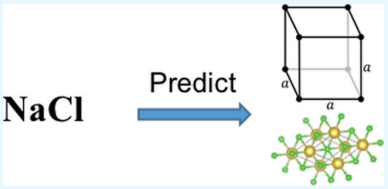
Yong Zhao, Yuxin Cui, Zheng Xiong, Jing Jin, Zhonghao Liu, Rongzhi Dong, and Jianjun Hu*

Cite This: *ACS Omega* 2020, 5, 3596−3606

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Structural information of materials such as the crystal systems and space groups are highly useful for analyzing their physical properties. However, the enormous composition space of materials makes experimental X-ray diffraction (XRD) or first-principle-based structure determination methods infeasible for large-scale material screening in the composition space. Herein, we propose and evaluate machine-learning algorithms for determining the structure type of materials, given only their compositions. We couple random forest (RF) and multiple layer perceptron (MLP) neural network models with three types of features: Magpie, atom vector, and one-hot encoding (atom frequency) for the crystal system and space group prediction of materials. Four types of models for predicting crystal systems and space groups are proposed, trained, and evaluated including one-versus-all binary classifiers, multiclass classifiers, polymorphism predictors, and multilabel classifiers. The synthetic minority over-sampling technique (SMOTE) is conducted to mitigate the effects of imbalanced data sets. Our results demonstrate that RF with Magpie features generally outperforms other algorithms for binary and multiclass prediction of crystal systems and space groups, while MLP with atom frequency features is the best one for structural polymorphism prediction. For multilabel prediction, MLP with atom frequency and binary relevance with Magpie models are the best for predicting crystal systems and space groups, respectively. Our analysis of the related descriptors identifies a few key contributing features for structural-type prediction such as electronegativity, covalent radius, and Mendeleev number. Our work thus paves a way for fast composition-based structural screening of inorganic materials via predicted material structural properties.



INTRODUCTION

Computational material screening based on high-speed machine-learning algorithms has become a reality, as shown by a growing number of related works.[1−5] There are two types of screenings: one for screening known materials with desired properties[3,6,7] and one for screening hypothetical materials that have not been discovered or synthesized and usually have only composition information available.[1,2,4] Usually, some kind of enumeration procedures[8] or generative machine-learning models[9] can be used to generate many (millions) of hypothetical material compositions as the combinations of selected set of elements, which requires fast algorithms to evaluate their stability,[10] to predict their crystal structures[11] or physical properties of interest.[12]

The crystal structure plays a critical role in determining the properties of materials. Knowing how the atoms of a material are arranged in the space helps understand its properties.[13] The structural information such as atomic coordinates or space groups can then be incorporated into the advancement of material design. Takahashi L. and Takahashi L.[14] use the Gaussian mixture model to reveal two data clusters, and then, Random Forest (RF) is used to classify the crystal structures using eight descriptors. Further, first-principles calculations are performed to confirm the stability of predicted materials. However, predicting the atomic coordinates of a crystalline only from its composition using crystal structure prediction algorithms such as USPEX[15] (Universal Structure Predictor: Evolutionary Xtallography) is challenging and time-consuming, as expensive density functional theory (DFT) simulations are needed.[16] In this case, prediction of the space groups or other structural information (such as atomic bonding angles and relative distances) of materials that have no crystal structure information can be useful to understand their physicochemical properties. For example, Ward et al.[16] use composition-based features of elemental properties and the Voronoi tessellation of the materials' crystal structure as inputs to ML to predict formation energies in their work.[16]

Conventionally, the crystal structures of materials can be determined experimentally by the X-ray diffraction (XRD) technique in which X-ray beams are used to hit nanoparticles and the scattered intensity of the beams are observed and measured. Novel materials can be unveiled by mapping XRD patterns to the measured or simulated XRD patterns of known materials. This method has led to the determination of a huge number of crystal structures, as deposited in databases such as Materials Project[17] and ICSD.[18,19] A large number of methods

Received: November 26, 2019
Accepted: January 31, 2020
Published: February 13, 2020

https://dx.doi.org/10.1021/acsomega.9b04012
ACS Omega 2020, 5, 3596−3606

have been developed to analyze the XRD data such as programs for indexing and space-group determination. ITO,[20] TREOR,[21] DICVOL,[22] McMaille,[23] EXPO,[24] and X-CELL[25] are part of cutting-edge software packages for indexing and space group determination. Space groups of materials can also be determined using machine-learning methods from their XRD data. Recently, Park et al.[26] showed that deep learning techniques can outperform rule-based programs without human involvement for space group determination from XRD data. The successful prediction of the crystal system of two novel inorganic compounds further confirms the potential of their method[26] in crystal structure determination. Another deep neural network algorithm[5] is proposed by Oviedo et al. to predict the space group and crystal dimensionality of materials through a limited number of experimental thin-film XRD data. This method augments small datasets based on physics knowledge, and their deep neural networks achieved high accuracy among other machine-learning algorithms.

Despite the success of XRD-based methods for material structure determination, this is not a feasible solution for high throughput material screening in which millions of possible elemental compositions need to be evaluated, which makes experimental method expensive, time-consuming, or just infeasible.[1] Next, the success of XRD method is heavily reliant on the quality of XRD results, which is not always easy to achieve.[27] It also takes hours to acquire and analyze XRD data to recognize the crystal structure for each material.[5]

Theoretically, given the chemical composition of a material, computational prediction of its crystal structure is possible. A couple of works utilize evolutionary algorithms or particle swarm optimization (PSO) and DFT to determine crystal structures.[15,27] USPEX[15] leverages the evolutionary algorithm to find the most stable crystal structures, of which local optimization, spatial heredity, and lattice mutation are three key components to minimize the free energy. Wang et al.[27] proposed an algorithm to search the free-energy space using the PSO algorithm within the evolutionary scheme together with ab initio structural optimization, symmetry constraint, and the geometrical structure parameter technique. Gator[28] uses various settings of first principles calculation and genetic algorithms to increase the chance of locating the numerous low-energy minima. Despite their powerful prediction abilities, these first-principles-based approaches are computationally demanding, which makes it impossible to perform high-speed screening for novel material discovery. For example, it is shown that it takes tens of thousands of CPU hours to calculate 45 DFT calculations of formation energy.[29]

In this paper, we propose a machine learning-based method for predicting the space group and the crystal system for an inorganic material, given only its composition information. Such models allow us to conduct fast screening of millions of potential chemicals as done in ref 1. We evaluate three types of features/descriptors: Magpie,[30] atom vector,[31] and one hot encoding (atom frequency) as the inputs of our machine-learning algorithms. Neither XRD data nor DFT calculation is involved in feature calculations. Because of the fact that one composition may correspond to multiple crystal structures, four classifiers are developed to predict material structures in terms of the crystal system and space group: one-versus-all classifiers, multiclass classifiers, polymorphism classifiers, and multilabel classifiers. We leverage multi layer perceptron (MLP) and RF to analyze how those feature sets can help determine the crystal structure using 10 fold cross-validation. By evaluating with different combinations of feature sets and machine-learning techniques, we find that RF with Magpie features are the best in one-versus-all classification of space groups; one hot encoding is better than other two when classifying the multistructure polymorphism and multiple space group labeling. Moreover, because most of the materials have a single crystal system or space group, we apply RF and MLP to assign these two labels to such materials. Our results indicate that RF with Magpie performs the best in determining the single crystal system or space group.

RESULTS AND DISCUSSION

Crystal System Prediction. *Material Crystal System Prediction from Composition Using One-Versus-All Binary Classifiers.* For each of the seven crystal systems, we train an one-versus-all binary classifier with the formulas of the selected crystal systems set as positive samples and all other samples as negative ones. The sample distribution for all crystal systems is shown in Figure 3b. Tables 1 and 2 show the F1 scores and Matthews correlation coefficient (MCC) for predicting crystal systems using RF and MLP, respectively. First, we find that RF achieves the highest performance with F1 scores ranging from 0.723 to 0.844 for all crystal systems except triclinic, for which the RF + atom frequency encoding achieves the best performance with F1 score of 0.704 and MCC of 0.434. In comparison, the atom vector encoding works the worst among all three encoding methods with RF.

Table 1. Performance of RF for Predicting Crystal Systems

crystal system	Magpie (F1-score/MCC)	atom vector (F1-score/MCC)	atom frequency (F1-score/MCC)
Cubic	**0.844/0.698**	0.753/0.538	0.775/0.457
Hexagonal	**0.794/0.618**	0.647/0.374	0.704/0.433
Monoclinic	**0.736/0.482**	0.670/0.360	0.730/0.467
orthorhombic	**0.729/0.485**	0.611/0.297	0.705/0.425
Tetragonal	**0.797/0.623**	0.654/0.388	0.723/0.477
Triclinic	0.686/0.412	0.644/0.337	**0.704/0.434**
Trigonal	**0.723/0.498**	0.616/0.320	0.703/0.436

Table 2. Performance of MLP for Predicting Crystal Systems

crystal system	Magpie (F1-score/MCC)	atom vector (F1-score/MCC)	atom frequency (F1-score/MCC)
Cubic	0.815/0.632	0.805/0.612	**0.830/0.660**
Hexagonal	0.774/0.553	0.741/0.486	**0.781/0.566**
Monoclinic	0.699/0.399	0.698/0.396	**0.732/0.465**
orthorhombic	0.692/0.385	0.689/0.380	**0.731/0.463**
Tetragonal	0.767/0.536	0.743/0.488	**0.773/0.548**
Triclinic	0.663/0.331	0.676/0.353	**0.709/0.421**
Trigonal	0.701/0.409	0.705/0.412	**0.743/0.489**

When we compare the performance of MLP with three encoding methods, it is found that the atom frequency encoding achieves the best performance for all the seven crystal systems. Comparing the best combination of RF with Magpie with the best combination of MLP with atom frequency, the F1 score of RF with Magpie is slightly better than the MLP with atom frequency in predicting some crystal systems (e.g., cubic, hexagonal, monoclinic, and tetragonal). For predicting orthorhombic, triclinic, and trigonal, MLP with atom frequency outperforms RF with Magpie slightly. However, RF with Magpie is better than MLP with atom frequency

https://dx.doi.org/10.1021/acsomega.9b04012
ACS Omega 2020, 5, 3596−3606

overall in terms of the MCC. Indeed, we find that atom vectors and atom frequency using MLP outperform their counterparts using RF and that RF and MLP using Magpie have close performance among all seven crystal systems. The possible reason is that atom vectors and atom frequency encode the internal connections inside a formula. Nonlinear operations by MLP help discriminate objects well. Plus, MLP can efficiently learn the mappings between the inputs and their labels. Because an F1-score of 0.844 is a relatively high score, this shows that the machine-learning algorithms have done a good job in materials crystal system prediction from the compositions.

The results by over-sampling are shown in Tables 3 and 4. It can be found that with over-sampling, the best performance of

Table 3. Performance of RF for Predicting Crystal Systems by Over-Sampling

crystal system	Magpie (F1-score/MCC)	atom vector (F1-score/MCC)	atom frequency (F1-score/MCC)
Cubic	0.846/0.693	0.779/0.557	0.777/0.556
Hexagonal	0.808/0.622	0.714/0.428	0.674/0.361
Monoclinic	0.750/0.500	0.707/0.418	0.725/0.450
orthorhombic	0.739/0.485	0.667/0.336	0.698/0.405
Tetragonal	0.803/0.613	0.720/0.441	0.695/0.409
Triclinic	0.714/0.429	0.690/0.383	0.707/0.419
Trigonal	0.742/0.494	0.680/0.360	0.693/0.402

Table 4. Performance of MLP for Predicting Crystal Systems by Over-Sampling

crystal system	Magpie (F1-score/MCC)	atom vector (F1-score/MCC)	atom frequency (F1-score/MCC)
Cubic	0.806/0.613	0.788/0.575	0.820/0.640
Hexagonal	0.752/0.507	0.717/0.435	0.759/0.518
Monoclinic	0.701/0.410	0.696/0.393	0.731/0.463
orthorhombic	0.682/0.365	0.678/0.358	0.727/0.454
Tetragonal	0.749/0.501	0.725/0.450	0.757/0.517
Triclinic	0.677/0.368	0.682/0.366	0.705/0.410
Trigonal	0.683/0.372	0.686/0.372	0.722/0.445

RF has not been improved by a large margin and instead scores of MLP is decreased. The possible reason is that the ratios of positive and negative labels of each dataset are between 1/11 and 1/4, which is acceptable to machine-learning algorithms. On the contrary, one interesting finding is that the performance of RF with atom vectors is improved by SMOTE significantly. On average, both F1-score and MCC are increased by 0.05.

To show what features contribute most to the prediction of crystal systems, we calculate and rank the top 20 features by their feature importance scores for each crystal system when the RF with Magpie (the best classifier) is applied for classification. The results are shown in Figure 1 (i.e., from subfigure 1a to subfigure 1g). We find that shared important features include the following: mean and average deviation of melting temperature, mean and average deviation of the Mendeleev number, mean and average deviation of covalent radius, mean, and average deviation of GS volume per pa, mean and average deviation of electronegativity, mean atom number, mean atomic weight, and mean N_p valence. These features describe physical properties which are known to be involved in crystal system formation.

Crystal System Prediction Using Multiclass Prediction Models. As shown in Figure 3a, 88.3% formulas (53,532 in total) have a unique crystal system. It is reasonable to develop a single classifier to assign the crystal system for a given composition, which is much more efficient than predicting its crystal system by running through seven binary classifiers. Here, we train one single RF classifier and MLP classifier for material crystal system prediction for each encoding approach. We only choose the formulas with a single crystal system. A stratified 10-fold cross-validation is used here for evaluating the classifiers. The 10-fold cross-validation results are shown in Table 5. Again, we find that RF with Magpie achieves the strongest performance with F1 score and MCC of 0.650 and 0.591 against other combinations of models and feature sets. Compared with RF, we find that MLP is inferior for all three feature types. It is interesting that again, for MLP, the best encoding is atom frequency rather than Magpie which achieves the best performance with RF. It should be noted that while we have spent sufficient effort for tuning the MLP model parameters to maximize its performance, we find it is not easy to further significantly improve the MLP performances here by simple parameter tuning or structure modification. New descriptors and machine-learning methods may be needed to improve the predictive performance. In addition, SMOTE shows inferior results across all combinations of learning methods and feature sets except for RF with atom vectors. The improvement for RF with atom vectors is only marginal.

Crystal System Polymorphism Prediction Using Binary Classifiers. Knowing whether or not a formula/composition can form compounds of multiple crystal systems is interesting to the material community. Here, we select all formula with multiple crystal systems as positive samples (7104 samples in total) while the remaining samples as negative ones (53,532 samples in total). Then, we train two binary classifiers using RF and MLP, respectively, to predict whether a given material composition can form multiple crystal systems or not. The 10-fold cross-validation results are shown in Table 6.

First, we find that MLP with atom frequency encoding achieves the best performance with an F1 score of 0.704 and MCC of 0.409. The RF with atom frequency is the second with an F1 score of 0.668 and MCC of 0.354. In comparison, the MLP and RF with Magpie and MLP with an atom vector achieve similar performance and are both much lower than those of RF and MLP + atom frequency. Over-sampling increases the F1 score of RF with all feature sets slightly but decreases the MCC of them. However, over-sampling decreases both F1 score and MCC of MLP with all feature sets.

Figure 1h shows the top 20 important features for crystal system polymorphism prediction. The shared features of mean and avg_devMendeleevNumber, mean and avg_dev GSvolume_pa, mean and avg_dev electronegativity, mean and avg_dev melting T, mean and avg_dev covalent radius, mean number, and mean atomic weight with one-versus-all case are keys for predicting the crystal system.

Crystal System Prediction Using Multilabel Classifiers. It is known that many materials with different crystal systems can share the same formula or composition. Therefore, the crystal system prediction problem can be formulated as a multilabel classification problem. Here, we apply multilabel classifiers to explore how machine-learning algorithms perform as regard to crystal system prediction from composition. We evaluate four multilabel prediction algorithms, each with three encoding.

https://dx.doi.org/10.1021/acsomega.9b04012
ACS Omega 2020, 5, 3596−3606

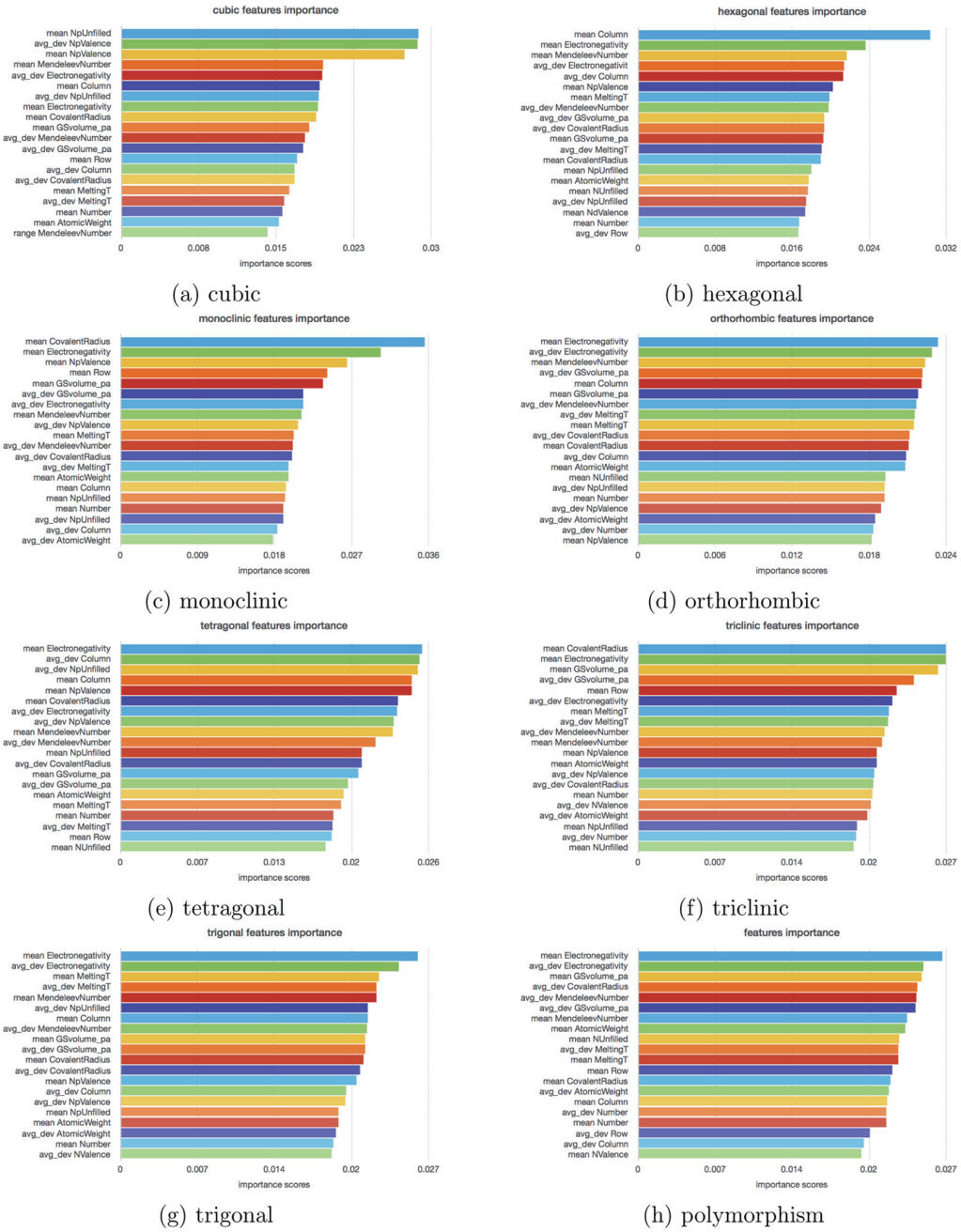


Figure 1. Ranking of Magpie Features for crystal system prediction.

Table 5. Performance for Multiclass Prediction of the Crystal System

	Magpie (F1-score/MCC)	atom vector (F1-score/MCC)	atom frequency (F1-score/MCC)
RF	**0.650/0.591**	0.511/0.445	0.575/0.511
MLP	0.559/0.486	0.559/0.489	0.615/0.551
RF-oversample	0.644/0.585	0.524/0.448	0.562/0.494
MLP-oversample	0.509/0.424	0.541/0.469	0.598/0.533

The algorithms include MLP and three transformation algorithms (BinaryRelevance,[32] ClassifierChain,[33] and LabelPowerset[32]) for multilabel classification, all using the RF as the base classifier. Tenfold cross-validation is applied for performance evaluation. Table 7 shows the best results for each evaluated algorithm.

We find that LabelPowerset with Magpie and MLP with atom frequency achieve close performance, and they are much better than other two transformation methods. LabelPowerset has the best performance with Exact MR of 0.598, accuracy of 0.638, precision of 0.673, recall of 0.649, and F1 score of 0.652, of which recall is 0.010 lower than MLP with atom frequency.

https://dx.doi.org/10.1021/acsomega.9b04012
ACS Omega 2020, 5, 3596−3606

Table 6. Performance for Crystal System Polymorphism Prediction

	Magpie (F1-score/MCC)	atom vector (F1-score/MCC)	atom frequency (F1-score/MCC)
RF	0.652/0.350	0.610/0.293	0.668/0.354
MLP	0.646/0.308	0.642/0.289	**0.704/0.409**
RF-oversample	0.670/0.343	0.636/0.272	0.672/0.348
MLP-oversample	0.646/0.304	0.633/0.267	0.699/0.399

Table 7. Performance for Multilabel Crystal System Prediction[a]

	AF + MLP	Magpie + BR	Magpie + CC	Magpie + LP
exact MR	0.579	0.469	0.534	**0.598**
accuracy	0.631	0.504	0.568	**0.638**
precision	0.660	0.531	0.601	**0.673**
recall	**0.659**	0.516	0.574	0.649
F1-score	0.650	0.516	0.580	**0.652**

[a]AF = atom frequency, BR = BinaryRelevance, CC = ClassifierChain, LP = LabelPowerset.

BinaryRelevance has the worst results, which is reasonable, because MLP, ClassifierChain, and LabelPowerset take the internal label relationships into account in the label space. Instead, BinaryRelevance assumes an independent classifier for each label.

Space Group Prediction. Determining the space group for a given material composition tells a lot of information about its physical properties. However, compared to 7 crystal systems, there are 223 space groups in total in the Materials Project dataset, which makes it much more challenging to build the prediction models. Here, we select top 18 space groups, each having more than 1000 compositions for exploring four classifiers for space group prediction from composition. We show the results of machine-learning models for space group prediction as evaluated via 10-fold cross-validation.

Materials Space Group Prediction from Composition Using One-Versus-All Binary Classifiers. For each of the 18 space groups and with selected machine learning algorithm (RF or MLP) and selected encoding methods, we train 10 binary classifiers for 10 fold cross-validation. Together, we have trained 180 space group classifiers. Therefore, instead of reporting the classifier performances for each of the space groups, we merely calculate the average F1 score and MCC for the 10 fold cross-validation performances of each space group over all space group categories, and the results are shown in Table 8.

Table 8 shows the average F1 score and MCC over 18 space groups using RF and MLP with different material encoding. Without oversampling, we can find that RF with Magpie and MLP with atom frequency are the best combinations for predicting the space groups of inorganic materials using composition. The MCC of RF with Magpie is slightly better than that of MCC of MLP with atom frequency. However, the F1 score of RF with Magpie is slightly worse than MCC of MLP with atom frequency. These scores are considerably lower compared with the performance scores for predicting crystal systems because there are much more categories of space groups than crystal systems. Both the F1 score and MCC of RF are improved by oversampling. The best combination becomes RF with Magpie after oversampling. The scores for MLP are decreased slightly by oversampling for all feature sets. For instance, F1 score and MCC of MLP with atom frequency are decreased to 0.753 and 0.508, respectively. The biggest improvement achieved by SMOTE is RF with atom vectors, whose F1 score and MCC are increased by 0.077 and 0.089, respectively.

Space Group Prediction Using Multiclass Prediction Models. Instead of building 18 binary classifiers for space group prediction, here, we build a multiclass predictor for determining the space group, given a material composition. We focus on the materials with a single space group. The stratified 10 fold cross-validation results are shown in Table 9.

Similar to multiclass prediction for crystal systems, the combination of RF and Magpie features has the best performance with the F1 score and MCC of 0.652 and 0.627, as shown in Table 9, respectively. In this case, the performance of each case is worse than the counterparts in the multiclass prediction of crystal systems. The possible reason is that the number of space groups is larger than the number of crystal systems so that the samples in each group are more sparse compared to crystal systems. Again, oversampling slightly decreases the performance of all combinations except for RF with an atom vector.

Space Group Polymorphism Prediction Using Binary Classifiers. Here, we develop algorithms for predicting whether a material composition can form materials of multiple space groups. We set the compositions with multiple space groups in the dataset as positive samples and the remaining as negative ones. Then, RF- or MLP-based predictors combined with one of three encoding methods are evaluated in terms of their prediction power. The 10 fold cross-validation results are shown in Table 10.

It is found that MLP with atom frequency achieves the best result for predicting space group polymorphism with an F1 score and MCC of 0.670 and 0.342, respectively. RF with Magpie features by oversampling achieves comparable but slightly lower performance (F1-score 0.651) as MLP with atom frequency. SMOTE improves the performance of all cases other than the combination MLP with frequency. MLP with Magpie and RF with the atom vector have the largest improvement by SMOTE. Both scores are improved by 0.05

Table 8. Average Performance for Predicting the Space Group Using RF and MLP

	Magpie (F1-score/MCC)	atom vector (F1-score/MCC)	atom frequency (F1-score/MCC)
RF	0.765/**0.566**	0.649/0.365	0.722/0.470
MLP	0.751/0.507	0.729/0.461	**0.768**/0.540
RF-oversample	0.787/0.579	0.726/0.454	0.725/0.459
MLP-oversample	0.743/0.493	0.718/0.437	0.753/0.508

https://dx.doi.org/10.1021/acsomega.9b04012
ACS Omega 2020, 5, 3596−3606

Table 9. Performance for Multi-Class Prediction of Space Groups

	Magpie (F1-score/MCC)	atom vector (F1-score/MCC)	atom frequency (F1-score/MCC)
RF	**0.652/0.627**	0.519/0.501	0.576/0.556
MLP	0.571/0.540	0.540/0.517	0.616/0.591
RF-oversample	0.643/0.619	0.531/0.505	0.566/0.543
MLP-oversample	0.557/0.528	0.525/0.502	0.597/0.573

Table 10. Performance for Space Group Polymorphism Prediction

	Magpie (F1-score/MCC)	atom vector (F1-score/MCC)	atom frequency (F1-score/MCC)
RF	0.610/0.273	0.540/0.147	0.614/0.253
MLP	0.582/0.205	0.591/0.190	**0.670/0.342**
RF-oversample	0.651/0.305	0.607/0.218	0.635/0.275
MLP-oversample	0.626/0.267	0.597/0.198	0.663/0.326

on average. Figure 2 shows the top 20 important features for space group polymorphism prediction. It is interesting, but as

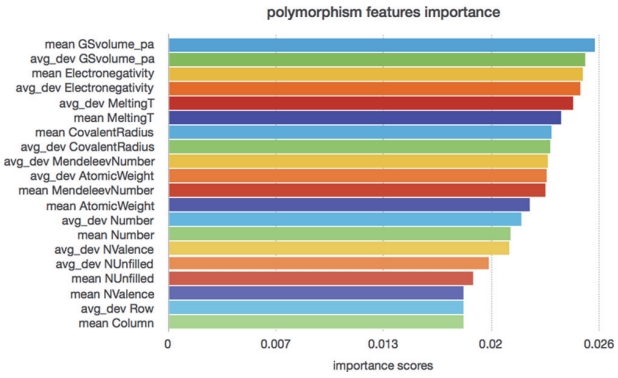


Figure 2. Magpie feature importance ranking for space group polymorphism prediction.

expected, that the top features here overlap a lot with those top 20 important features for predicting crystal systems. It means these features such as electronegativity, GSVolume, and Mendeleev Number play a critical role in predicting the crystal symmetry.

Space Group Prediction Using Multilabel Classifiers. Because each elemental composition may form materials of multiple different space groups, here, we evaluate how current machine-learning algorithms can predict all the multiple space groups for a given composition. Similar to multilabel predictions for crystal systems, we use BinaryRelevance,[32] ClassifierChain,[33] and LabelPowerset[32] plus MLP as multilabel classifiers, each evaluated with three features sets. Tenfold cross-validation results for the best combinations of algorithms and features sets are shown in Table 11. We can find that the

Table 11. Performance for Multilabel Space Group Prediction Using MLP[a]

	AF + MLP	Magpie + BR	Magpie + CC	Magpie + LP
exact MR	0.569	0.446	0.472	**0.597**
accuracy	0.612	0.467	0.491	**0.626**
precision	0.633	0.485	0.510	**0.651**
recall	0.634	0.472	0.495	**0.636**
F1-score	0.626	0.474	0.498	**0.637**

[a]AF = atom frequency, BR = BinaryRelevance, CC = ClassifierChain, LP = LabelPowerset.

performance of multilabel predictors for space group prediction is slightly inferior to their counterparts in the multilabel classification of crystal systems, which is expected because of the large number of space groups compared to the number of samples. Similar observations apply to space group predictions. LabelPowerset with Magpie has the best learning power and MLP with atom frequency comes next with close performance. BinaryRelevance is still the worst one because of the assumed independence of binary classifiers.

CONCLUSIONS

We propose and evaluate machine-learning algorithms for predicting the crystal systems and space groups of materials merely from their compositions. Two popular machine-learning algorithms including RFs and multilayered perceptron neural networks combined with three material representations are evaluated for four types of structure classification problems for both crystal system prediction and space group prediction: one-vs-all binary classification, multiclass classification, polymorphism prediction, and multilabel classification. Our extensive experiments show that the RF with Magpie features achieves the highest performance for one-vs-all binary classification, multilabel prediction, and multiclass classification of both crystal systems and space groups. In contrast, RF with atom frequency obtains the best results for polymorphism prediction of both crystal systems and space groups. However, the modest MCC scores of 0.591 and 0.627 for the multiclass crystal system and space group prediction shows current machine-learning algorithms and descriptors are far from achieving satisfactory performance, which calls for development of more advanced algorithms. One possible reason is that some artificial compounds have very high energy above hull, which might lead to unreasonable and misleading prediction over the crystal system and space group. In the future work, we may try to set a formation energy threshold to filter out those materials in the Materials Project dataset. In addition, our feature importance analysis shows that electronegativity, covalent radius, Mendeleev number, melting temperature, GAS volume pascal, and mean atomic weight are crucial factors for predicting the crystal system and space group for a given material composition. We also found that the ML performance for space group prediction is much lower than that of materials crystal system prediction given their composition. That is because the data is distributed more unevenly over 18 space groups in our study, which may call for more advanced techniques to address this issue.

https://dx.doi.org/10.1021/acsomega.9b04012
ACS Omega 2020, 5, 3596−3606

Our prediction models for crystal systems and space groups pave a way for performing large-scale fast structural screening of materials when only compositions are available. This is especially true when compared to XRD data and DFT-based approaches for space group determination, which is too expensive or slow for large-scale screening.

METHODS

Datasets. We describe how we create the datasets for training and evaluating our prediction models. Our material samples are extracted from the Materials Project,[17] which is an extensive database that deposits the properties (e.g., crystallographic parameters, formation energy, and band gap) of all known inorganic materials.[17] It is continuously growing and when we started this work, it contained 86,106 compounds in total. Table 12 summarizes the distribution of compounds as regard to the number of elements existent in the compounds. We find that the number of composition elements ranges from 2 to 8 and materials with 2, 3, 4, and 5 elements occupy 98.9% of the database (we exclude those materials of a single element).

Table 12. Distribution of Materials with Respect to the No. of Elements

no. of elements	no. of compounds
2	14,026
3	41,751
4	22,798
5	6585
6	874
7	67
8	5

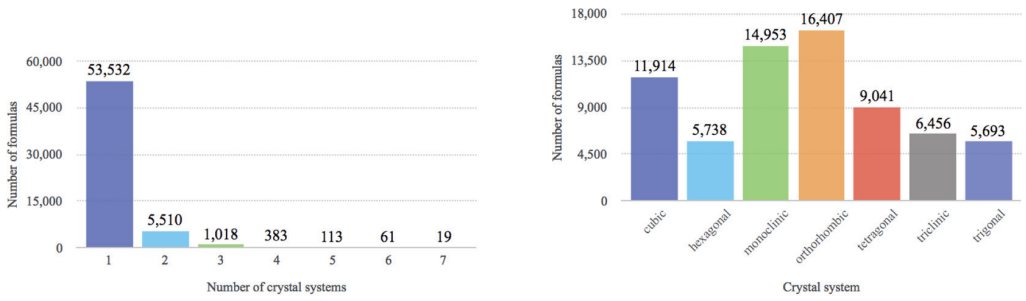
We eliminate duplicate entries with identical formulas and space group information by keeping one sample for each such group. In addition, we remove a material ($HeSiO_2$) that has no values in its Magpie features.[30] After this preprocessing, the total number of samples in our dataset is 60,636. These materials can be classified into 7 crystal systems and 223 space groups which we aim to predict. For each material, we generate three types of features merely based on its composition including Magpie, atom vector,[31] and one-hot encoding (atom frequency), which are detailed in the next section.

The goal of this paper is to develop classification algorithms to predict the crystal systems and space groups from material compositions. Because each inorganic compound formula might correspond to materials with multiple different crystal systems or space groups, the crystal system/space group assignment problem can be mapped as a multilabel classification problem. To understand the distribution of samples in these crystal systems and space groups, Figures 3 and 4 show the distributions of samples in these categories. Figure 3a shows that most of the formulas in the dataset (88.3% or 53,532 formulas) have unique crystal systems. Similar observation applies to the space group (85.7% or 51,988 formulas), as shown in Figure 4a. Among those formulas having multiple crystal systems, the number of 2-element formula is the largest, 3-element formula is the second, and few formulas have more than 3 elements. Figure 3b shows the distribution of materials in each crystal systems. We can find that the number of formulas is above 10,000 in orthorhombic, monoclinic, and cubic systems. In addition, the number is close to 10,000 for the tetragonal system. For the other three remaining systems, they have around 5000 formulas.

There are 223 unique space groups in our dataset. Some formulas may correspond to materials with multiple space groups. Figure 4a shows the distribution of formula with different numbers of space groups. It shows that a majority of compositions (51,993) only exist with one space group, and 5977 formulas have two space groups. In Figure 4b, we can find that most of the space groups have a number of formulas less than 1000. In our space group classification problems, we only consider those space groups that have more than 1000 formulas, and the total number of space groups is 18. The space group symbols are $Fm\bar{3}m$, $P2\_1/c$, $Pnma$, $P\bar{1}$, $P1$, $C2/c$, $C2/m$, $Immm$, $Pm\bar{3}m$, $I4/mmm$, $P6\_3/mmc$, $Ccmm$, $P4/mmm$, $R\bar{3}m$, $Cm2m$, $P2\_1/m$, Cm, and $F\bar{4}3m$. From the space group and crystal system classification system, we find that 8 out of these top 10 space groups belong to the top 4 crystal systems:[34] 2647 $Immm$ and 3891 $Pnma$ belong to the orthorhombic crystal system, 5220 $P2\_1/c$, 2707 $C2/c$, and 2647 $C2/m$ belong to the monoclinic crystal system, 6171 $Fm\bar{3}m$ and 2142 $Pm\bar{3}m$ belong to the cubic crystal system, and 2124 $I4/mmm$ belong to the tetragonal crystal system.

We develop four types of classifiers to predict the crystal systems and space groups of the materials:

- One-versus-all classifier, which predicts whether a given composition/formula can form compounds of a specific crystal system or space group. We need to train one classifier for each crystal system or space group.

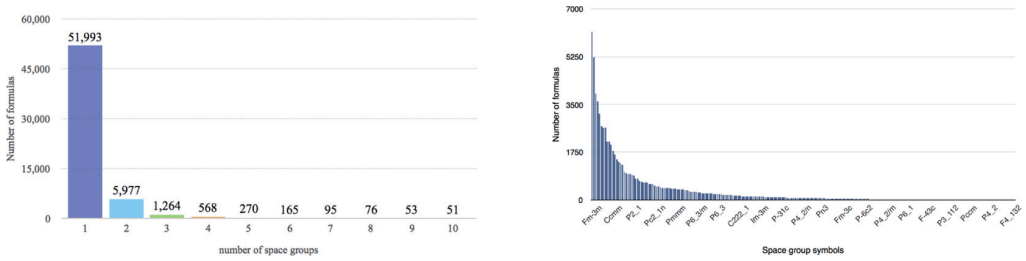


(a) The distribution of the number of crystal systems

(b) The number of formulas in each crystal system

Figure 3. Distribution of crystal systems in the dataset.

https://dx.doi.org/10.1021/acsomega.9b04012
ACS Omega 2020, 5, 3596−3606



(a) Top 10 space group polymorphism

(b) The number of formulas in each space-group

Figure 4. Distribution of space groups in the dataset.

- Multiclass classifier, which determines the single label for the materials with unique crystal system or space group. We only need to train one classifier for crystal system prediction and another classifier for space group prediction.
- Polymorphism classifier, which predicts whether a composition can form compounds of multiple (≥2) crystal systems or space groups.
- Multilabel classifier, which predicts with what crystal systems or space groups a composition can form compounds.

Descriptors. The machine-learning classifiers that we aim to develop are based on combinations of different machine-learning algorithms and feature encodings. In this paper, we explore three kinds of features for predicting the crystal system and space group from material compositions: Magpie,[30] atom vector,[31] and one-hot encoding (atom frequency). These features depend only on materials compositions or the formula themselves. In other words, we will not use any other structure information or physical properties calculated from the first principle.

Magpie Features. Magpie (Materials-Agnostic Platform for Informatics and Exploration)[30] is an extensive set of features related to the constituent elements in materials. The set covers a broad range of physical and chemical properties that fall into four different categories: stoichiometric features, elemental property statistics, electronic structure features, and ionic compound features. Stoichiometric features only contain the number of elements in the compound and their several L^p norms. Elemental property statistics are calculated by computing several statistics (i.e., average, minimum, maximum, range, mean absolute deviation, and mode) of 22 different elemental properties. Electronic structure features are the average fraction of s, p, d, and f valence electrons.[35] Ionic compound features are the possibility of forming an ionic compound when we assume all elements present in a single oxidation state and two adaptions for calculating the fractions of a compound based on electronegativity.[36] Compared to the atom vector and one-hot encoding, Magpie is a general-purpose feature type that can be used to predict the properties of materials based on their formulas, for example, it can describe the difference of heavy atoms and light atoms in a compound and link it to, for example, thermal conductivity prediction. Matminer[37] is used to retrieve the features, and we remove the features with respect to the crystal space group.

Atom2vec. Atom2vec[31] is a representation scheme for elements, which is calculated based on learning the relationship of elements among known materials. These learned properties are represented in terms of high-dimensional vectors for all elements. Atom-environment pairs are invented. The model maps the collection of all atoms in the environment to a feature vector for the composition. Suppose a $n \times d$ matrix $V = [v_1, v_2, ..., v_n]$ is given, where n is the number of atoms and d is the dimension of atom vector, assume that the environment contains k atoms, then, the environment vector can be represented as follows

$$E = C(v_1, v_2, ..., v_k)$$

where C is the summation over all atoms in our work. A score function (i.e., normalization score) is defined as $S(v_i,E)$ which evaluates the likelihood of the target atom v_i and it appears with the environment E. Atom vectors are trained by maximizing the normalization score over the whole dataset. Compared to other representations, Atom2vec represents atoms in terms of high-dimensional vectors that capture how atoms relate to each other in a high-dimensional space. Based on the atom/element vectors calculated by Atom2vec for all elements, for a given formula, we sum up all the atom vectors for the elements in the formula as the representation vector for the material.

One-Hot Encoding (Atom Frequency). This encoding approach represents each compound by a vector of atom numbers of each element. We first count the frequency of atoms for each element in the given inorganic compound. Then, a vector with 87 values is used to represent a formula because there are 87 unique elements in our dataset. Each component of the vector stores the frequency of a given element that exists in the formula or set to zero if a specific element is not available. Despite its simplicity,[1] it is shown that with large dataset and powerful models such as deep neural networks, even one-hot encoding can achieve highly predictive models.

Machine-Learning Methods. Two widely used machine-learning algorithms including MLP and RF and three multilabel learning algorithms are evaluated in this study:

- We design two MLP structures. The first structure is for one-versus-all, multiclass and polymorphism classifiers. It has 11 layers, and the numbers of nodes on hidden layers are 1024, 1024, 1024, 512, 512, 512, 256, 128, 64, 32, respectively. The second structure is only for multilabel classifiers. It has 13 layers, and the numbers of nodes on hidden layers are 4096, 2048, 2048, 2048, 1024, 1024, 512, 512, 256, 256, 128, 32, respectively. The number of neurons on last layers of both structures is decided by the specific classifier. For instance, the last layer of the first structure has seven neurons in

https://dx.doi.org/10.1021/acsomega.9b04012
ACS Omega 2020, 5, 3596−3606

predicting the multilabel crystal structure. ReLU[38] is used to activate neurons except for the last output layers. The activation function on last layers depends on the classification problem. We use sigmoid for one-versus-all, multilabel, and polymorphism classifiers and softmax normalization for the multiclass classifier. It should be noted that two basic deep fully connected MLP architectures are used here because of their demonstrated performance in material property prediction.[1] While more advanced deep neural networks such as convolutional or recurrent neural networks may be used and explored for each predicting task, however, tuning of model hyperparameters and architectures is left for future work.

- RF[39] is a popular machine-learning algorithm widely used in material informatics because of its robustness and capability to train with large datasets.[2,38,40] As an ensemble algorithm, RF aggregates the results of different decision trees (in our work, the number of decision tree is set as 50) to make more accurate models. Each decision tree is trained with a randomly selected subset of samples and features. The output of the final model either votes or averages the output of each decision tree depending on the specific task of regression or classification.
- BinaryRelevance (BR)[41−43] is considered as the most intuitive solution for multilabel classification. It transforms a multilabel problem into multiple independent binary-learning problems. The number of independent binary classifiers is reliant on the number of unique labels in the dataset. Each binary classifier corresponds to one label in the label space. All binary classifiers are trained on the decomposed dataset. For example, we have 7 crystal systems, and the dataset is decomposed into 7 datasets, of which labels in each dataset belong to one crystal system or not.
- ClassifierChain (CC)[33] is also a binary relevance method. However, CC differs from BR in that the feature space is augmented by the predictions of all previous binary classifiers in the chain. The added label information allows CC to take into account correlations among labels. If strong correlations exist in the label space, CC gives each base binary classifier relatively more predictive power.
- LabelPowerset[32] transforms the multilabel problem into a multiclass problem with one multiclass trained on all unique label combinations formed in the dataset. In other words, it considers each combination in the power set as a single label in the dataset. This technique needs the worst case of 2^L classifiers, where L is the number of labels in the label space. When L increases, the distinct label combinations can grow exponentially, which leads to memory and computing time explosion easily.

In addition, because of the imbalanced datasets, we investigate whether the over-sampling method [i.e., Synthetic Minority Over-sampling Technique (SMOTE)[44] improves the performance in predicting crystal systems and space groups using one-versus-all, multiclass, and polymorphism classifiers. To illustrate how SMOTE works, we take Magpie features as an example. For the minority class (e.g. cubic), we take a sample from the dataset and consider its k nearest neighbors in the Magpie feature space. To synthesize a new sample, we take one sample from current sample and its k nearest neighbors. Then, we multiply the Magpie feature with a random real number between 0 and 1.

Evaluation Metrics. Because our datasets are imbalanced, we use F1-score and MCC to evaluate the performances of one-vs-all classifiers and polymorphism classifiers. F1-score is the harmonic mean of precision and recall with the maximum value of 1 and minimum value of 0 as the worst. MCC is also used to measure the quality of binary and multiclass classifiers, which takes into account the balance ratios of true positive, true negative, false positive, and false negative of the predictions. A MCC of 1 means prefect prediction, 0 is an average random guess, and −1 is inverse prediction.

In multilabel classification problems, a sample can be labeled with one or more categories. The predicted labels for each sample can thus be fully correct, partially correct, or fully incorrect. Traditional evaluation metrics such as precision or recall no longer apply to multilabel classifiers for performance evaluation. Thus, we redefined the accuracy, precision, recall, and F1 score to evaluate the performance according to ref 45 In addition, we add the Exact Match Ratio[45] as one additional performance measure. Assuming n is the number of samples and T_i and P_i are real and predicted labels that the sample i have, the precision, recall, F1 score, and ExactMatchRatio can be calculated as follows

$$\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n}\frac{|T_i \cap P_i|}{|T_i \cup P_i|}$$

$$\text{Precision} = \frac{1}{n}\sum_{i=1}^{n}\frac{|T_i \cap P_i|}{|P_i|}$$

$$\text{Recall} = \frac{1}{n}\sum_{i=1}^{n}\frac{|T_i \cap P_i|}{|T_i|}$$

$$\text{F1} - \text{score} = \frac{1}{n}\sum_{i=1}^{n}\frac{2 \times |T_i \cap P_i|}{|T_i| + |P_i|}$$

$$\text{ExactmatchRatio} = \frac{1}{n}\sum_{i=1}^{n} I(T_i = P_i)$$

where accuracy is the intersection over union between real and predicted labels. Precision is the average of the ratio of predicted correct labels over the total number of real labels. Recall is the average of the ratio of predicted correct labels over the total number of predicted labels. F1 score is the harmonic mean of precision and recall mentioned above. The Exact Math Ratio is the proportion of entirely correct predictions over the total number of samples, where I is the indicator function.

We use 10 fold cross-validation to evaluate the performance of all classifiers composed of different machine learning algorithms and features sets. This evaluation strategy randomly splits the whole dataset into 10 equal partitions. Then, for each fold, training of a classification model over 9 of the 10 partitions and testing of the model over the remaining partition are done. The process is repeated until all 10 partitions are used as test sets once for each. The final performance is aggregated as the average performance over the whole dataset.

https://dx.doi.org/10.1021/acsomega.9b04012
ACS Omega 2020, 5, 3596−3606

AUTHOR INFORMATION

Corresponding Author

Jianjun Hu − *Department of Computer Science and Engineering, University of South Carolina, Columbia 29208, South Carolina, United States; School of Mechanical Engineering, Guizhou University, Guiyang 550025, China;* orcid.org/0000-0002-8725-6660; Phone: +1 (803) 7777304; Email: jianjunh@cse.sc.edu; Fax: +1 (803) 7773767

Authors

Yong Zhao − *Department of Computer Science and Engineering, University of South Carolina, Columbia 29208, South Carolina, United States;* orcid.org/0000-0002-6762-266X

Yuxin Cui − *Department of Computer Science and Engineering, University of South Carolina, Columbia 29208, South Carolina, United States*

Zheng Xiong − *Department of Computer Science and Engineering, University of South Carolina, Columbia 29208, South Carolina, United States*

Jing Jin − *Department of Computer Science and Engineering, University of South Carolina, Columbia 29208, South Carolina, United States*

Zhonghao Liu − *Department of Computer Science and Engineering, University of South Carolina, Columbia 29208, South Carolina, United States*

Rongzhi Dong − *School of Mechanical Engineering, Guizhou University, Guiyang 550025, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.9b04012

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Research reported in this work was supported in part by the NSF and SC EPSCoR Program under award number (NSF award #OIA-1655740 and SC EPSCoR grant 19-GC02). The views, perspective, and content neither necessarily represent the official views of the SC EPSCoR Program nor those of the NSF. This work was also partially supported by NSF under grant and 1940099 and 1905775.

REFERENCES

(1) Jha, D.; Ward, L.; Paul, A.; Liao, W.-k.; Choudhary, A.; Wolverton, C.; Agrawal, A. ElemNet: Deep Learning the Chemistry of Materials From Only Elemental Composition. *Sci. Rep.* **2018**, *8*, 17593.

(2) Kim, K.; Ward, L.; He, J.; Krishna, A.; Agrawal, A.; Wolverton, C. Machine-learning-accelerated high-throughput materials screening: Discovery of novel quaternary Heusler compounds. *Phys. Rev. Mater.* **2018**, *2*, 123801.

(3) Sendek, A. D.; Yang, Q.; Cubuk, E. D.; Duerloo, K.-A. N.; Cui, Y.; Reed, E. J. Holistic computational structure screening of more than 12000 candidates for solid lithium-ion conductor materials. *Energy Environ. Sci.* **2017**, *10*, 306−320.

(4) Olsthoorn, B.; Geilhufe, R. M.; Borysov, S. S.; Balatsky, A. V. Band gap prediction for large organic crystal structures with machine learning. **2018**, arXiv preprint arXiv:1810.12814.

(5) Oviedo, F.; Ren, Z.; Sun, S.; Settens, C.; Liu, Z.; Hartono, N. T. P.; Savitha, R.; DeCost, B. L.; Tian, S. I.; Romano, G.; Kusne, A. G.; Buonassisi, T. Fast classification of small X-ray diffraction datasets using data augmentation and deep neural networks. **2018**, arXiv preprint arXiv:1811.08425.

(6) Zhang, Y.; He, X.; Chen, Z.; Bai, Q.; Nolan, A. M.; Roberts, C. A.; Banerjee, D.; Matsunaga, T.; Mo, Y.; Ling, C. Unsupervised discovery of solid-state lithium ion conductors. *Nat. Commun.* **2019**, *10*, 1−7.

(7) He, X.; Bai, Q.; Liu, Y.; Nolan, A. M.; Ling, C.; Mo, Y. Crystal Structural Framework of Lithium Super-Ionic Conductors. *Adv. Energy Mater.* **2019**, *9*, 1902078.

(8) Davies, D. W.; Butler, K. T.; Jackson, A. J.; Morris, A.; Frost, J. M.; Skelton, J. M.; Walsh, A. Computational screening of all stoichiometric inorganic materials. *Chem* **2016**, *1*, 617−627.

(9) Dan, Y.; Zhao, Y.; Li, X.; Li, S.; Hu, M.; Hu, J. Generative adversarial networks (GAN) based efficient sampling of chemical space for inverse design of inorganic materials. **2019**, arXiv preprint arXiv:1911.05020.

(10) Jha, D.; Choudhary, K.; Tavazza, F.; Liao, W.-k.; Choudhary, A.; Campbell, C.; Agrawal, A. Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning. *Nat. Commun.* **2019**, *10*, 5316.

(11) Oganov, A. R.; Pickard, C. J.; Zhu, Q.; Needs, R. J. Structure prediction drives materials discovery. *Nat. Rev. Mater.* **2019**, *4*, 331−348.

(12) Goodall, R. E.; Lee, A. A. Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. **2019**, arXiv preprint arXiv:1910.00617.

(13) Ward, L.; O'Keeffe, S. C.; Stevick, J.; Jelbert, G. R.; Aykol, M.; Wolverton, C. A machine learning approach for engineering bulk metallic glass alloys. *Acta Mater.* **2018**, *159*, 102−111.

(14) Takahashi, K.; Takahashi, L. Creating Machine Learning-Driven Material Recipes Based on Crystal Structure. *J. Phys. Chem. Lett.* **2019**, *10*, 283−288.

(15) Oganov, A. R.; Glass, C. W. Crystal structure prediction using ab initio evolutionary techniques: Principles and applications. *J. Chem. Phys.* **2006**, *124*, 244704.

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

(17) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, K. A. P. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **2013**, *1*, 011002.

(18) Bergerhoff, G.; Hundt, R.; Sievers, R.; Brown, I. D. The inorganic crystal structure data base. *J. Chem. Inf. Model.* **1983**, *23*, 66−69.

(19) Belsky, A.; Hellenbrandt, M.; Karen, V. L.; Luksch, P. New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design. *Acta Crystallogr., Sect. B: Struct. Sci.* **2002**, *58*, 364−369.

(20) Visser, J. W. A fully automatic program for finding the unit cell from powder data. *J. Appl. Crystallogr.* **1969**, *2*, 89−95.

(21) Werner, P. E.; Eriksson, L.; Westdahl, M. TREOR, a semi-exhaustive trial-and-error powder indexing program for all symmetries. *J. Appl. Crystallogr.* **1985**, *18*, 367−370.

(22) Boultif, A.; Louër, D. Indexing of powder diffraction patterns for low-symmetry lattices by the successive dichotomy method. *J. Appl. Crystallogr.* **1991**, *24*, 987−993.

(23) Le Bail, A. Monte carlo indexing with mcmaille. *Powder Diffr.* **2004**, *19*, 249−254.

(24) Altomare, A.; Camalli, M.; Cuocci, C.; Giacovazzo, C.; Moliterni, A.; Rizzi, R. EXPO2009: structure solution by powder data in direct and reciprocal space. *J. Appl. Crystallogr.* **2009**, *42*, 1197−1202.

(25) Neumann, M. A. X-Cell: a novel indexing algorithm for routine tasks and difficult cases. *J. Appl. Crystallogr.* **2003**, *36*, 356−365.

(26) Park, W. B.; Chung, J.; Jung, J.; Sohn, K.; Singh, S. P.; Pyo, M.; Shin, N.; Sohn, K.-S. Classification of crystal structure using a convolutional neural network. *IUCrJ* **2017**, *4*, 486−494.

https://dx.doi.org/10.1021/acsomega.9b04012
ACS Omega 2020, 5, 3596−3606

(27) Wang, Y.; Lv, J.; Zhu, L.; Ma, Y. Crystal structure prediction via particle-swarm optimization. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2010**, *82*, 094116.
(28) Curtis, F.; Li, X.; Rose, T.; Vázquez-Mayagoitia, Á.; Bhattacharya, S.; Ghiringhelli, L. M.; Marom, N. GAtor: A First-Principles Genetic Algorithm for Molecular Crystal Structure Prediction. *J. Chem. Theory Comput.* **2018**, *14*, 2246−2264.
(29) Ziletti, A.; Kumar, D.; Scheffler, M.; Ghiringhelli, L. M. Insightful classification of crystal structures using deep learning. *Nat. Commun.* **2018**, *9*, 2775.
(30) Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2016**, *2*, 16028.
(31) Zhou, Q.; Tang, P.; Liu, S.; Pan, J.; Yan, Q.; Zhang, S.-C. Learning atoms for materials discovery. *Proc. Natl. Acad. Sci. U.S.A.* **2018**, *115*, E6411−E6417.
(32) Tsoumakas, G.; Katakis, I.; Vlahavas, I. *Data Mining and Knowledge Discovery Handbook*; Springer, 2009; pp 667−685.
(33) Read, J.; Pfahringer, B.; Holmes, G.; Frank, E. Classifier chains for multi-label classification. *ECML PKDD*, 2009; pp 254−269.
(34) Space group. Space group—Wikipedia, The Free Encyclopedia. 2019, https://en.wikipedia.org/wiki/Space_group (accessed Jan 7, 2019).
(35) Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J. E.; Doak, J.; Thompson, A.; Zhang, K.; Choudhary, A.; Wolverton, C. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *89*, 094104.
(36) Callister, W. D.; Rethwisch, D. G. *Materials Science and Engineering: an Introduction*; John Wiley & Sons: New York, 2007; Vol. *7*.
(37) Ward, L.; Dunn, A.; Faghaninia, A.; Zimmermann, N. E. R.; Bajaj, S.; Wang, Q.; Montoya, J.; Chen, J.; Bystrom, K.; Dylla, M.; Chard, K.; Asta, M.; Persson, K. A.; Snyder, G. J.; Foster, I.; Jain, A. Matminer: An open source toolkit for materials data mining. *Comput. Mater. Sci.* **2018**, *152*, 60−69.
(38) Nair, V.; Hinton, G. E. Rectified linear units improve restricted boltzmann machines. *ICML-10*, 2010; pp 807−814.
(39) Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5−32.
(40) Stanev, V.; Oses, C.; Kusne, A. G.; Rodriguez, E.; Paglione, J.; Curtarolo, S.; Takeuchi, I. Machine learning modeling of superconducting critical temperature. *npj Comput. Mater.* **2018**, *4*, 29.
(41) Tsoumakas, G.; Katakis, I. Multi-label classification: An overview. *Int. J. Data Warehous. Min.* **2007**, *3*, 1−13.
(42) Godbole, S.; Sarawagi, S. Discriminative methods for multi-labeled classification. *PAKDD*, 2004; pp 22−30.
(43) Zhang, M.-L.; Zhou, Z.-H. A k-nearest neighbor based algorithm for multi-label classification. *2005 IEEE International Conference on Granular Computing*, 2005; Vol. *5*, pp 718−721.
(44) Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321−357.
(45) Sorower, M. S. *A Literature Survey on Algorithms for Multi-label Learning*; Oregon State University: Corvallis, 2010; p 18.

https://dx.doi.org/10.1021/acsomega.9b04012
ACS Omega 2020, 5, 3596−3606