COMPOSITION BASED CRYSTAL MATERIALS SYMMETRY PREDICTION USING MACHINE LEARNING WITH ENHANCED DESCRIPTORS

A PREPRINT

Yuxin Li
School of Mechanical Engineering
Guizhou University
Guiyang China 550025

Rongzhi Dong, Wenhui Yang
School of Mechanical Engineering
Guizhou University
Guiyang China 550025

Jianjun Hu [*]
Department of Computer Science and Engineering
University of South Carolina
Columbia, SC, 29201, USA
jianjunh@cse.sc.edu

August 18, 2021

ABSTRACT

Geometric information such as the space groups and crystal systems plays an important role in the properties of crystal materials. Prediction of crystal system and space group thus has wide applications in crystal material property estimation and structure prediction. Previous works on experimental X-ray diffraction (XRD) and density functional theory (DFT) based structure determination methods achieved outstanding performance, but they are not applicable for large-scale screening of materials compositions. There are also machine learning models using Magpie descriptors for composition based material space group determination, but their prediction accuracy only ranges between 0.638 and 0.907 in different kinds of crystals. Herein, we report an improved machine learning model for predicting the crystal system and space group of inorganic materials using only the formula information. Benchmark study on a dataset downloaded from Materials Project Database shows that our random forest models based on our new descriptor set, achieve significant performance improvements compared with previous work with accuracy scores ranging between 0.712 and 0.961 in terms of space group classification. Our model also shows large performance improvement for crystal system prediction. Trained models and source code are freely available at `https://github.com/Yuxinya/SG_predict`

***K*eywords** crystal system prediction · space group prediction · materials informatics · machine learning

1 Introduction

According to the degree of geometric form symmetry, the crystals can be divided into different crystal systems and space groups[1, 2]. Determining the symmetry information, particularly the space group, provides wide applications for crystal material property prediction[3] and crystal structure prediction[4, 5]. Recently, we proposed the method of knowledge-rich approach for crystal structure prediction[6, 7, 8, 9], which is inspired by the recent advances in protein structure prediction (PSP)[10, 11] which predicts protein structures using the predicted distance matrix. In our approach, various global optimization algorithms[6] such as genetic algorithm[7] and differential evolution algorithm[9] have

[*] Corresponding author: J.H. (webpage)

been used to reconstruct the crystal structure atom coordinates. However, it is indispensable to provide the space group information for a given material composition before predicting their structures.

Several machine learning models have been proposed to predict the space groups of crystals. Several studies developed machine learning approaches for space group classification by the X-ray diffraction[12] data of materials. Suzuki et al.[13] emphasised on demonstrating the potential of simple machine learning techniques suitable for knowledge discovery and real-world experiments. Their tree-ensemble-based machine learning model works with over 90% accuracy for crystal system classification based on powder X-ray diffraction patterns. Park et al.[14] developed three convolutional neural networks (CNN) for the space group, extinction group and crystal system classification of 150,000 powder XRD patterns, which returned test accuracy of 81.14, 83.83 and 94.99% respectively. Vecsei et al.[15] studied the problem of space group determination from powder X-ray diffraction patterns by using fully connected neural networks and convolutional neural networks and then tested those two models on the other database. Oviedo et al.[16] proposed a supervised machine learning framework for rapid crystal structure identification of novel materials from thin-film XRD measurements. Chakraborty et al.[17] performed augmentation of thin filmed X-ray diffraction patterns and developed a high accuracy model for lattice classification from X-ray diffraction. Zaloga et al.[18] identified crystal systems and symmetry space groups by full-profile X-ray diffraction patterns using convolutional neural networks and explored the factors that affect the classification performance. Ziletti et al.[19] represented crystals by calculating a diffraction image, then constructed a deep learning neural network model for classification which achieves robust performance even in the presence of highly defective structures.

Although the XRD based symmetry prediction algorithms can achieve good performance in space group classification, they have several limitations. The performance of XRD based methods is frequently influenced by low-quality X-ray diffraction data[20]. Moreover, it is time-consuming to acquire and analyze XRD data to recognize the crystal structure for each material[16]. There are other machine learning models, different from XRD based methods, applied for space groups and crystal systems classification. Liu et al.[21] trained a convolutional neural network model by atomic pair distribution function for 45 most heavily represented space groups with an accuracy of 0.70. Kaufmann et al.[22] used a machine learning–based approach and developed a general methodology for rapid and autonomous identification of the crystal symmetry from electron backscatter diffraction (EBSD) patterns. However, those methods are inconvenient for predicting thousands of space groups since the input pictures of materials should be provided. Theoretically, given the chemical composition of a material, computational prediction of its crystal structure is possible[23]. Several studies determine crystal structures by combining global optimization with DFT calculations[23]. These methods have demonstrated successes in a variety of cases. However, the DFT based methods generally require thousands of CPU hours and can only be applied to predict structures of relative small systems[19], which is not suitable for large-scale material space group determination.

Recently, there emerged several crystal structure prediction methods that start with a seed structure generated by the symmetry-restricted procedure [5, 8, 24]. In these algorithms, usually for a given composition, a space group is specified to generate some random structures that satisfy the symmetry constraints of the space group. There is thus a need to predict the space group for a given composition. Several machine learning algorithms have been proposed for material crystal system and space group prediction quickly using composition information alone[25, 26]. In order to get the best classification performance, Zhao et al.[25] uses two machine learning algorithms, random forest and multiple layer perceptron neural network models, combined with three kinds of descriptors, atom vector, one-hot encoding and Magpie, for the crystal system and space group classification. For the Material Project database they used, there are only 18 space groups selected, each has more than 1000 samples for multi-class classification which achieve a performance of 0.652 and 0.637 in terms of the F1-scores. Liang et al.[26] proposed Cryspnet for Bravais lattice, space group and lattice constants prediction of crystal materials using deep neural networks. However, the accuracy scores only range from 0.638 and 0.907 for space group classification in the fourteen Bravais lattice categories.

In this work, we present an improved machine learning model for the crystal system and space group prediction by inputting the formulas of the crystal materials. Magpie descriptors are used as the basic descriptor set which is combined with a new descriptor set that we proposed, to train our machine learning models. To ensure that our classification algorithm can predict all types of crystal materials, we trained and validated the models on all kinds of entries from the Material Project Database of September 2020. Compared with previous works based only on the Magpie descriptors, the addition of the new descriptors enables our models to achieve significantly improved performance on crystal system and space group prediction. For example, the space group prediction of the cubic crystal system, which consists of 18,325 materials, has an accuracy score of 0.961. Our models can simultaneously make crystal system and space group predictions for a large number of hypothetical materials, which can make contribution to our knowledge-rich approach for crystal structure prediction. Our algorithm is also useful for downstream tasks such as the exploration of the structure and properties of new materials.

Our contributions can be summarized as follows:

- We propose a new descriptor set for crystal system and space group prediction of crystal materials which achieves significant performance improvements compared to prior studies. We also identified what features contribute most in our experiments for the classification performance
- We remove the duplication caused by the isomer, and build ML models for multi-class classification for space group and crystal system of crystal materials.
- We build ML models for multi-label classification for space group prediction and crystal system prediction.
- We conduct extensive experiments with different machine learning algorithms. Our experiments show that our algorithm based on random forest achieves high performance in crystal system and space group prediction.

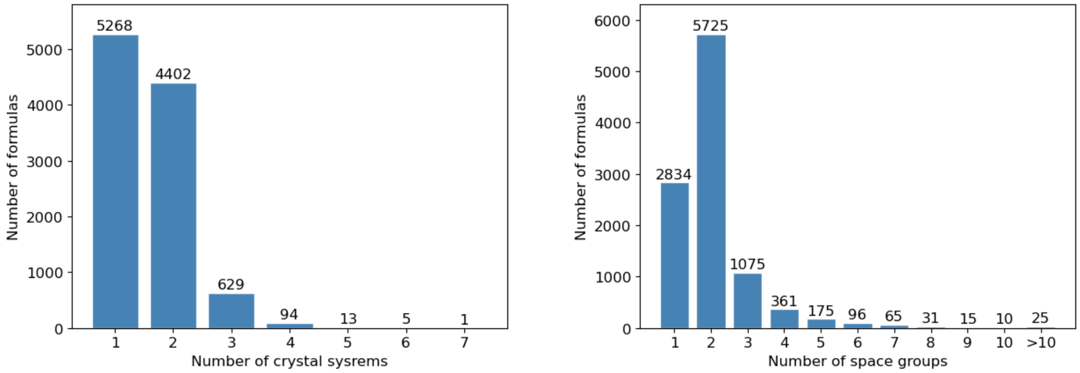
2 Materials and Methods

2.1 Datasets

Table 1: lattice parameter relationships for materials of different lattice systems.

Crystal system	Edge lengths	Axial angles	Space groups	Amount
Cubic	$a = b = c$	$\alpha = \beta = \gamma = 90$	195-230	18324
Hexagonal	$a = b$	$\alpha = \beta = 90, \gamma = 120$	168-194	9243
Trigonal	$a = b \neq c$	$\alpha = \beta = 90, \gamma = 120$	143-167	11086
Tetragonal	$a = b \neq c$	$\alpha = \beta = \gamma = 90$	75-142	14654
Orthorhombic	$a \neq b \neq c$	$\alpha = \beta = \gamma = 90$	16-74	26800
Monoclinic	$a \neq c$	$\alpha = \gamma = 90, \beta \neq 90$	3-15	29872
Triclinic	all other cases	all other cases	1-2	15297

There are 125,276 inorganic material items used in our experiment, which are extracted from the Materials Project[27], an extensive material database which includes the properties of all known inorganic materials. According to the degree of geometric forms symmetry, those crystals are divided into seven categories, namely cubic system, hexagonal system, tetragonal system, trigonal system, orthorhombic system, monoclinic system and triclinic system. The amount of those crystal systems and some details of geometric forms symmetry are shown in Table 1. For those crystals, they are divided into 230 different combinations of symmetrical elements. Each space group has a unique crystal system corresponding to it.



(a) The relationship of molecular formula quantity and crystal system quantity.

(b) The relationship of molecular formula quantity and space group quantity.

Figure 1: The statistics of the formulas which have isomers.

However, for a given crystal composition, there may be multiple structural isomers which share the same chemical formula. There are a total of 10,412 unrepeated formulas that have isomers in the dataset we used. As is shown in Figure 1, we show several statistics of these 10,412 formulas. We can see that there are some formulas that have only one crystal system and even one space group even though they have isomers. This is because the isomers of above formulas belong to the same crystal system or space group. In addition, there are no more than two crystal systems for most materials formula isomers, but there is a formula that has even 7 crystal systems. For material formulas with isomers, more than half of the formulas have two space groups. In order to make each formula match only one crystal system and space group, we calculate the retain score by Formula 1, which is also used in Cryspnet[26], and keep the material structure with the highest score among the isomers while dropping out other materials. Then, our machine learning model has a unique crystal system or space group target for a formula to achieve multi-class classification. Moreover, we also trained machine learning models for multi-label classification[25] for the 10,412 crystals which have isomers.

$$Score\left(f,s\right)=\frac{Abundance\left(f,s\right)}{E_{hull}\left(f,s\right)+\alpha} \tag{1}$$

In this formula, α is tunable for balancing the formation energy term and the abundance count from the Materials Project dataset, we set it as 0.1 in our work, s and f represent the space group and chemical formula. The $Abundance(s, f)$ means the number of records which have the same formula and space group. The $E_{hull}(s, f)$ is used to find the lowest formation energy above the convex hull by the given composition and space group. For the formulas that have isomers, we choose the material with the highest $Score(s, f)$ for our benchmark experiments. Then, there are 102,528 materials used in our experiment for multiclass classification.

2.2 Descriptors

Magpie descriptors[28] are a set of composition based materials attributes that calculate the statistics of stoichiometric attributes, elemental properties, electronic structure attributes and ionic compound attributes. It has been widely used for building machine learning models for composition based materials property predictions. In our work, the elemental property statistics have been used as the baseline data set for space group and crystal system prediction. There are 22 kinds of features in Magpie element property statistics including Atomic Number, Mendeleev Number, Atomic Weight, Melting Temperature, Periodic Table Row and Column, Covalent Radius, Electronegativity, the number of Valence e in each Orbital(s, p. d, f, total), the number of unfilled e in each orbital (s, p. d, f, total), Ground State Volume, Ground State Band Gap Energy, Ground State Magnetic Moment, and the Space Group Number of elements. The main features of the Magpie feature set are obtained by calculating the mean, average deviation, range, mode, minimum, and maximum of above elemental properties (weighted by the fraction of each element in the composition) to transform raw materials data into a form compatible with machine learning. We also add some other Magpie properties, which are used in Cryspnet[26] for structure information prediction, include Stoichiometry p-norm (p=0,2,3,5,7), Elemental Fraction, Fraction of Electrons in each Orbital, Band Center, Ion Property (possible to form ionic compound, ionic charge) to improve classification performance.

Additionally, we propose a set of new descriptors to improve crystal systems and space groups prediction performance. Our descriptors are related to the number of various atoms in a crystal structure. For a crystal, the numbers of atoms is

Table 2: Descriptors.

Element Property statistics of Magpie	Additional Predictors of Magpie	Added Predictors in this work
Atomic Number	Stoichiometry p-norm (p=0,2,3,5.7)	Total Atom Number
Mendeleev Number	Elemental Fraction	Maximum Atom Number
Atomic Weight	Fraction of Electrons in each Orbital	Minimum Atom Number
Melting Temperature	Band Center	Average Atom Number
Periodic Table Row and Column	lon Property (possible to form ionic compound, ionic charge)	Specific Value
Covalent Radius		Atom Number Variance
Electronegativity		
The number of Valence e in each orbital(s, p. d, f, total)		
The number of unfilled e in each orbital (s, p. d, f, total)		
Ground State Volume		
Ground State Band Gap Energy		
Ground State Magnetic Moment		
Space Group Number of elements		

different for different elements within a unit cell and our new descriptor is obtained by (for an element) calculating the maximum atom number, the minimum atom number, the total atom number, the average atom number, the specific value (ratio of pretty formula to full formula) and the atom number variance of all the elements within a crystal structure. All the features used in our experiments are shown in Table 2

2.3 Machine learning models: Random forest(RF), Extreme Gradient Boosting(XGBoost) and Deep Neural Networks(DNN)

In this study, we use ensemble machine learning algorithms and neural networks to find out the best model for crystal system and space group prediction.

2.3.1 Random Forest

Random forest (RF)[29], a popular ensemble machine learning algorithm, is one of the major bagging machine learning models. Its driving principle in classification is to build several estimators independently and each estimator gets the probability of possible output labels. Then the average of the predicted probabilities of all estimators are calculated as outputs. Labels with the highest average probability are used as the output results. In our random forest classification models, we choose 'entropy' as our criterion. The two important hyper-parameters, the number of trees and max features, are set to be 100 and 80 respectively. The max depth is None. The min samples leaf and min samples split are sat as 1 and 2. This algorithm was implemented by the Scikit-Learn[30] library in Python 3.6

2.3.2 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost)[31] is an excellent boosting algorithm which is widely used in machine learning classification tasks. As an integrated algorithm, XGBoost has outstanding performance in classification with excellent generalization performance. In addition, it is optimized by a series of methods, such as supporting regularization and using second-order Taylor expansion for loss function. Similar to traditional boosting algorithms which are composed of several weak algorithms, when training the XGBoost model, each weak algorithm tries to correct the error of the previous algorithms. In our work, we chose the 'gbtree' as our booster whose max depth was set as 6. As an important parameter, the number of the trees is set as 180. The learning rate, alpha, gamma, lambda are set as 0.3, 0, 0 ,1 respectively. For triclinic, which has only two kinds of space groups, the objective (used to specify the learning task and the corresponding learning objective) and evaluation metrics are set as softmax and multiclass logloss. For the rest experiments of the XGBoost model, we choose the logistic and logloss as the objective and evaluation metrics. We use the XGBoost library in Python 3.6 to implement this algorithm.

2.3.3 Deep Neural Networks

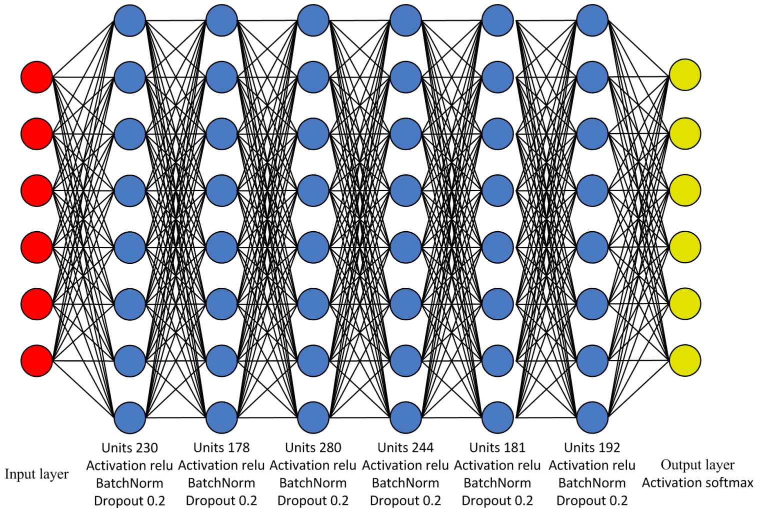


Figure 2: Architecture of the deep neural network.

The deep neural networks (DNN), which plays an important role in material performance prediction and discovery of new materials, is also used in our work. The structure of our DNN model, shown in Figure 2, is composed of 7 fully connected layers, and the numbers of the hidden nodes are 230, 178, 280, 244, 181, 192 respectively. Relu[32] is used as the action function for those layers. And after each layer except the last one, two strategies, Dropout and BatchNorm, are used to prevent overfitting. The loss function is set as the categorical cross entropy loss. The Adam optimizer is used in model training. The learning rate, epochs and batch size are set as 0.001, 2000 and 255 respectively. In order to prevent overfitting and improve performance of the DNN, we use the early stopping strategy where the monitor is 'loss' and patience is 30. Our DNN model is implemented on TensorFlow2.4.

2.4 Evaluation criteria

We use K-fold cross-validation in our work to assess the performance of classification models of different algorithms. The process of K-fold cross-validation strategy is that it randomly splits the initial sample set into K sub-sample sets, taking one of them as the test set, the rest as the training set. After the initial sample set split into K sub-sample sets, the cross-validation is repeated K times, each sub-sample set is verified once, then the results of K times are averaged to finally obtain a single performance estimate. The 10-fold cross-validation method is chosen to evaluate the classification performance of different algorithms in our work. Due to the uneven distribution of samples, the following performance criteria in this work are used, including accuracy, Matthews correlation coefficient (MCC), weighted precision, weighted recall and weighted F1 score[33, 34].

3 Results and Discussion

3.1 Multi-class classification

3.1.1 Prediction performance of multi-class classification

Table 3: The classification performance of RF algorithm for all data removed the duplication caused by isomers.

	Accuracy	MCC	Precision	Recall	F1 score
Crystal system prediction	0.816±0.005	0.779±0.006	0.818±0.005	0.816±0.005	0.816±0.005
Space group prediction	0.729±0.004	0.721±0.004	0.734±0.004	0.729±0.004	0.725±0.004

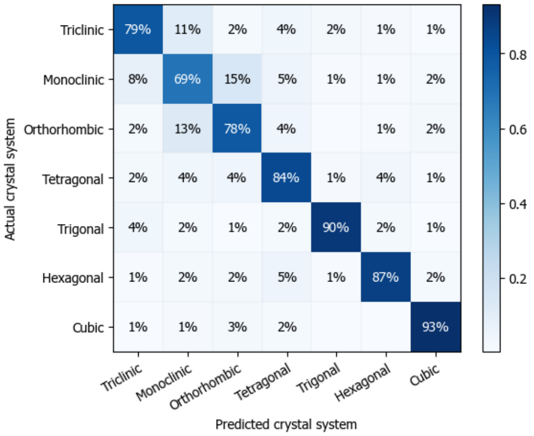


Figure 3: Confusion matrix of crystal system classification.

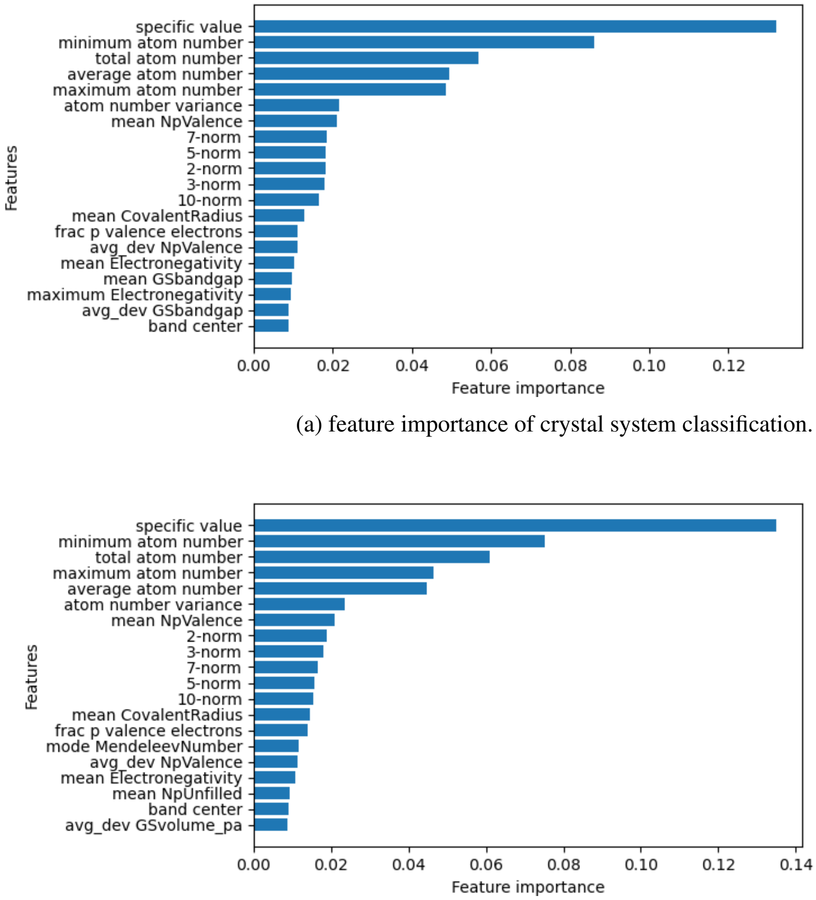
We trained the multi-class prediction models using the data sets after removing the duplications caused by isomers. Then we use the 10-fold cross-validation evaluation approach to assess our model performance. The crystal system and space group classification performance for all data are shown in Table 3. Our random forest model achieves an accuracy score of 0.816 and F1 score of 0.812 for crystal systems prediction, and its confusion matrix is shown in Figure 3 which shows that the more regular of the crystal material structures, the better classification result the model has in general. For space groups prediction with around 230 categories, the accuracy and MCC scores are 0.729 and 0,721 by just using the formula as the input to the machine learning models.

The results of space group classification in each crystal system are shown in Table 4, the accuracy score of our RF model reaches 0.961 in space group classification for cubic materials. Although the crystal structures in the cubic

Table 4: The performance of space group classification in different crystal systems.

	data set size	Accuracy	MCC	Precision	Recall	F1 score
Cubic	17367	0.961±0.006	0.945±0.008	0.960±0.005	0.961±0.006	0.959±0.006
Hexagonal	8201	0.909±0.008	0.888±0.010	0.908±0.008	0.909±0.008	0.906±0.008
Trigonal	9429	0.824±0.012	0.797±0.014	0.823±0.013	0.824±0.012	0.818±0.012
Tetragonal	12675	0.849±0.013	0.832±0.015	0.846±0.013	0.849±0.013	0.840±0.014
Orthorhombic	22392	0.755±0.005	0.729±0.006	0.759±0.005	0.755±0.005	0.746±0.006
Monoclinic	23024	0.712±0.009	0.647±0.011	0.715±0.010	0.712±0.009	0.703±0.010
Triclinic	9440	0.835±0.013	0.665±0.026	0.835±0.013	0.835±0.013	0.834±0.013

crystal system have the highest symmetry compared to other crystal systems, the cubic crystal system has 36 space groups, from the space group of 195 to 230. It is exciting to achieve such high space group classification performance with so many categories. However, the classification performances in orthorhombic and monoclinic for space group prediction are much lower with accuracy scores neither reaching 0.8. This is because these two crystal systems have complex crystal structures, and the input of our machine learning models is only the crystal formula. It is difficult to get the complex crystal structures information merely by the formulas of materials, which can however be ameliorated with large dataset. While the top-1 accuracy scores may not be ideal, it is possible to use the top-k prediction results of our models in downstream tasks to improve the hit rate. For other crystal systems, the performances in space groups classification have scores above 0.80 in terms of accuracy, MCC and F1 scores.



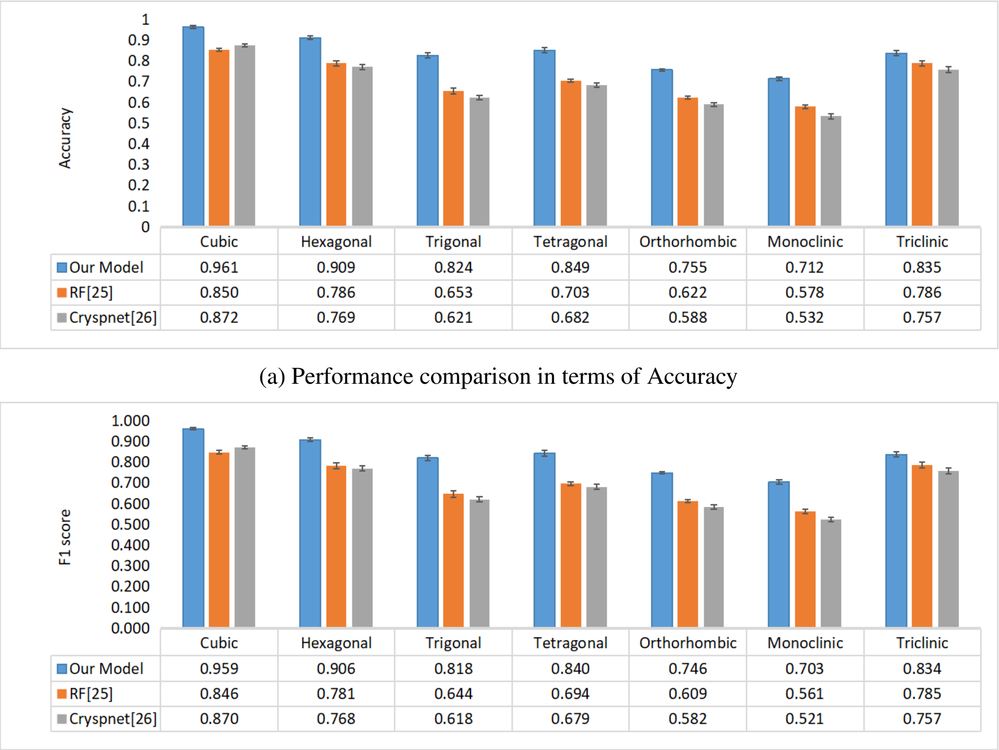
(a) feature importance of crystal system classification.

(b) feature importance of space group classification.

Figure 4: Feature importance of crystal system and space group classification.

In order to find out which features most affect the performance of classification, we calculated the feature importance scores to sort the top 20 features which are shown in Figure 4. We find that the specific value, minimum atom number, total atom number, average atom number, maximum atom number and atom number variance have major impact, which explains why our model is better than previous Magpie features based machine learning methods.

3.1.2 Performance comparison with previous works



	Cubic	Hexagonal	Trigonal	Tetragonal	Orthorhombic	Monoclinic	Triclinic
🟦 Our Model	0.961	0.909	0.824	0.849	0.755	0.712	0.835
🟧 RF[25]	0.850	0.786	0.653	0.703	0.622	0.578	0.786
⬜ Cryspnet[26]	0.872	0.769	0.621	0.682	0.588	0.532	0.757

(a) Performance comparison in terms of Accuracy

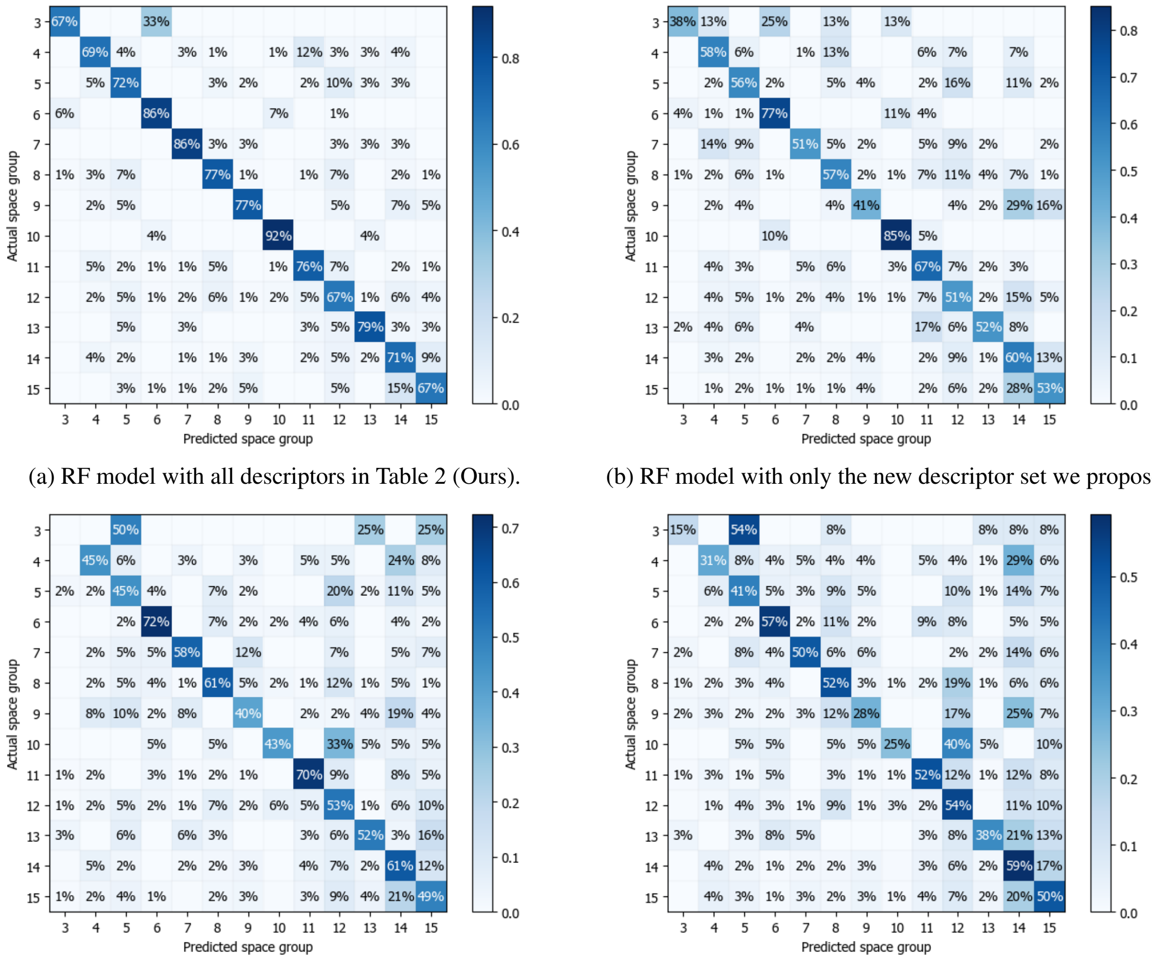
	Cubic	Hexagonal	Trigonal	Tetragonal	Orthorhombic	Monoclinic	Triclinic
🟦 Our Model	0.959	0.906	0.818	0.840	0.746	0.703	0.834
🟧 RF[25]	0.846	0.781	0.644	0.694	0.609	0.561	0.785
⬜ Cryspnet[26]	0.870	0.768	0.618	0.679	0.582	0.521	0.757

(b) Performance comparison in terms of F1 score

Figure 5: Performance comparison with previous works for space group prediction. Note that the results in RF[25] and Cryspnet[26] are based on our implementations of the corresponding algorithms described in the literature.

For multi-class classification, we select all kinds of space groups available in Materials Project in order that our classification algorithm can predict all types of crystal materials. However, in ref.[25], there are only 18 space groups selected for experiment, each having more than 1000 compositions for space group prediction. In addition, we use Formula 1 to tackle the duplicate formula caused by isomers, but in paper[25], there is no information about the isomer processing. In Magpie descriptors based space group classification, compared with Cryspnet[26], we use random forest algorithm which has better performance than the neural network[25]. And most importantly, we propose a new descriptor set, which greatly improves the performance of our model compared to previous works on space group prediction for inorganic materials[25, 26]. As is shown in Figure 5, we make comparisons with two previous works for space group prediction. Our model is the random forest in this work, which is trained using all the descriptors in Table 2. RF[25] is the random forest algorithm trained by Element Property statistic descriptors of Magpie. Cryspnet[26] is the neural networks framework of Cryspnet trained by Element Property statistic descriptors and additional Predictors of Magpie. The three models are trained using the same dataset of Materials Project processed by Formula 1. As we can see that compared with the previous works of Cryspnet and ref.[25], the space group prediction performance in terms of accuracy scores is improved by about 0.14 and 0.16 on average respectively in the six crystal systems except the triclinic. For F1 scores, the improvements are similar. Moreover, the higher the symmetry of the crystal structures, the better prediction performance our RF can achieve for space group classification in the three algorithms.

In particular, we show the confusion matrices in Figure 6 to explore some classification details for the space group prediction performance in the monoclinic system which has 13 space groups and relatively complex crystal structures. For our RF model, confusion matrices shown in Figure 6(a), the classification performance is better than previous two works whose confusion matrices are shown in Figure 6(c) and Figure 6(d). For more than half of the space groups, there



(a) RF model with all descriptors in Table 2 (Ours).

(b) RF model with only the new descriptor set we proposed.

(c) RF model in reference[25].

(d) NN of Cryspnet[26].

Figure 6: Performance of space group prediction for monoclinic materials using different algorithms. All those performances are trained by the same dataset. Note that the results in (c) and (d) are based on our implementations of the corresponding algorithms described in the literature.

are about 0.2 improvements in terms of accuracy. And the improvement of the space group of 3 and 10 are obvious which has about 0.5 improvement. In addition, the prediction results of our model are more concentrated, for example, for the actual space group of 6, the classification result concentrates on the space group of 3, 6, 10 and 12. However, in previous works[25, 26], particularly the Cryspnet[26], the classification results are scattered. To illustrate the validity of our descriptor, we make classification by the combination of random forest algorithm and the new descriptors we proposed, the performance is shown in Figure 6(b). The model even has a better performance than the two previous works, which indicates that the descriptors related to the number of various atoms in a crystal structure are effective for space group classification.

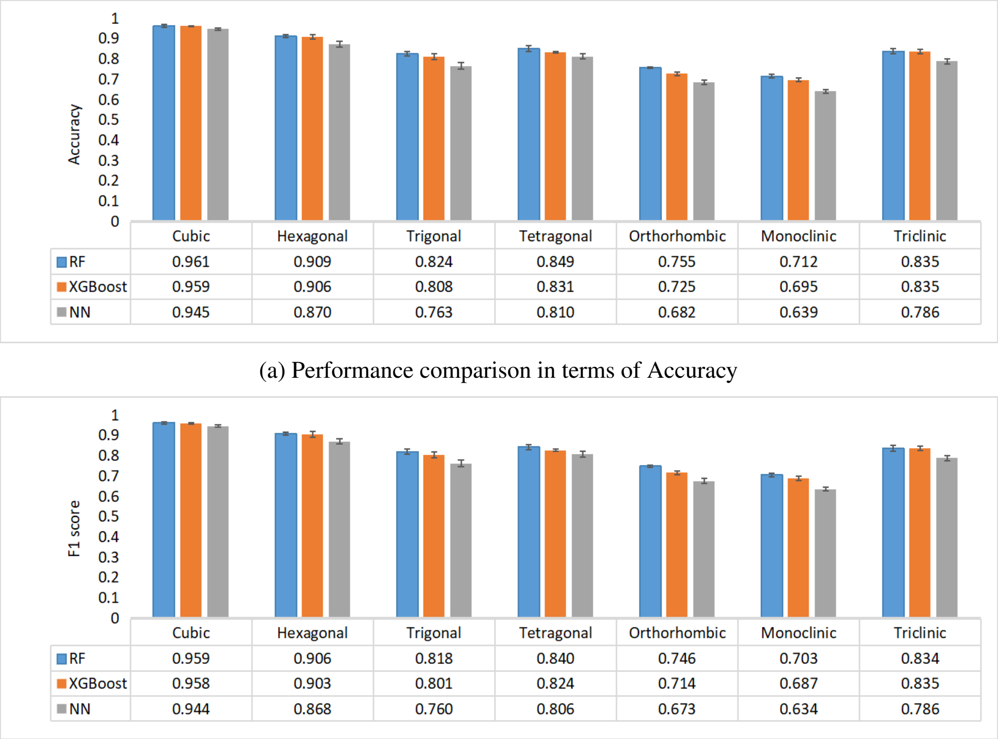
3.1.3 Performance comparison of different algorithms

Various machine learning algorithms have been used for structural information prediction. Here, we evaluate the performance of three powerful machine learning methods, the DNN, RF, and XGBoost, in space group and crystal system classification. As is shown in Figure 7, the RF model performances in both accuracy and F1 score are slightly better than those of XGBoost. And in our experiment, which uses the physical and chemical properties as descriptors, the performance of the two kinds of ensemble tree algorithms is better than that of DNN. In this work, RF has demonstrated better performance than DNN in making crystal systems and space groups classification, which is consistent with the results in previous study[25]. However, there are many works indicating that neural networks has excellent performance

Table 5: The performance of multi-label classification.

	Precision	Recall	F1 score
Crystal system prediction	0.813±0.011	0.706±0.013	0.751±0.010
Space group prediction	0.764±0.020	0.452±0.018	0.547±0.017

compared with other machine learning models in the field of material informatics[13, 16]. Therefore, for different problems, choosing suitable machine learning algorithms is needed to achieve the best results.



	Cubic	Hexagonal	Trigonal	Tetragonal	Orthorhombic	Monoclinic	Triclinic
RF	0.961	0.909	0.824	0.849	0.755	0.712	0.835
XGBoost	0.959	0.906	0.808	0.831	0.725	0.695	0.835
NN	0.945	0.870	0.763	0.810	0.682	0.639	0.786

(a) Performance comparison in terms of Accuracy

	Cubic	Hexagonal	Trigonal	Tetragonal	Orthorhombic	Monoclinic	Triclinic
RF	0.959	0.906	0.818	0.840	0.746	0.703	0.834
XGBoost	0.958	0.903	0.801	0.824	0.714	0.687	0.835
NN	0.944	0.868	0.760	0.806	0.673	0.634	0.786

(b) Performance comparison in terms of F1 score

Figure 7: Performance comparison of different algorithms in our work for space group prediction

3.2 Multi-label classification

Considering the fact that some formulas can correspond to multiple structures of different crystal systems or space groups, the prediction problem of crystal systems and space groups can also be mapped as a multi-label classification problem[25]. Here we train machine learning models for multi-label classification for the 10,412 crystals which have isomers. The crystal system classification performance is shown in Table 5 which has the recall score of 0.706, the precision score of 0.813, and the F1 score is 0.751. For space group classification, the performance is inferior to the crystal system prediction. The precision, recall and F1 scores are 0.764, 0.452 and 0.547 respectively. This is because the crystal structures have 230 space groups, which are far more diverse than the seven types of crystal systems.

4 Conclusion

Computational prediction of space groups and crystal systems plays an important role in analyzing crystal material structures and their physical and chemical properties and is useful for crystal structure prediction. While there are various methods to classify the space groups and crystal systems of crystal materials in previous works, this study proposes an efficient and easy-to-use method to achieve more accurate classification by introducing a new set of materials composition based descriptors. When combined with the Magpie descriptors, our algorithms' classification performance have improved significantly compared to previous algorithms that use only the magpie descriptors. Our trained models and source code are freely available at `https://github.com/Yuxinya/SG_predict`, which should be helpful for downstream works such as material property exploration, material screening, and crystal structure prediction. To further improve the performance of composition based space group and crystal system prediction, more advanced machine algorithms and deep learning based representation learning are two most promising directions. For example, graph neural networks with attention layers can be used for composition based representation learning. Moreover, more physics based heuristic descriptors can also potentially improve the performance.

5 Contributions

Conceptualization, J.H. and Y.L.; methodology, Y.L. and J.H.; software, Y.L. and J.H.; validation, Y.L. and J.H.;investigation, Y.L.,J.H., R.D., and W.Y.; resources, J.H.; writing–original draft preparation, J.H. and Y.L.; writing–review and editing, J.H and Y.L.; visualization, Y.L., R.D; supervision, J.H.; funding acquisition, J.H

6 Data Availability

The data that support the findings of this study are openly available in Materials Project database[27] at `http://www.materialsproject.org`.

7 Acknowledgement

Research reported in this work was supported in part by NSF under grant and 1940099 and 1905775. The views, perspective,and content do not necessarily represent the official views of the NSF

References

[1] Peter Paufler and Stanislav K Filatov. Es fedorov promoting the russian-german scientific interrelationship. *Minerals*, 10(2):181, 2020.

[2] Yuxin Li, Wenhui Yang, Rongzhi Dong, and Jianjun Hu. Mlatticeabc: Generic lattice constant prediction of crystal materials using machine learning. *ACS Omega*, 0(0):null, 0.

[3] Tian Xie and Jeffrey C Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, 120(14):145301, 2018.

[4] Hoi Chun Po, Ashvin Vishwanath, and Haruki Watanabe. Symmetry-based indicators of band topology in the 230 space groups. *Nature communications*, 8(1):1–9, 2017.

[5] Nobuya Sato, Tomoki Yamashita, Tamio Oguchi, Koji Hukushima, and Takashi Miyake. Adjusting the descriptor for a crystal structure search using bayesian optimization. *Physical Review Materials*, 4(3):033801, 2020.

[6] Jianjun Hu, Wenhui Yang, Rongzhi Dong, Yuxin Li, Xiang Li, Shaobo Li, and Edirisuriya MD Siriwardane. Contact map based crystal structure prediction using global optimization. *CrystEngComm*, 23(8):1765–1776, 2021.

[7] Jianjun Hu, Wenhui Yang, and Edirisuriya M Dilanga Siriwardane. Distance matrix-based crystal structure prediction using evolutionary algorithms. *The Journal of Physical Chemistry A*, 2020.

[8] Jianjun Hu, Yong Zhao, Wenhui Yang, Yuqi Song, Edirisuriya Siriwardane, Yuxin Li, and Rongzhi Dong. Alphacrystal: Contact map based crystal structure prediction using deep learning. *arXiv preprint arXiv:2102.01620*, 2021.

[9] Wenhui Yang, Edirisuriya M. Dilanga Siriwardane, Rongzhi Dong, Yuxin Li, and Jianjun Hu. Crystal structure prediction of materials with high symmetry using differential evolution, 2021.

[10] Wei Zheng, Yang Li, Chengxin Zhang, Robin Pearce, SM Mortuza, and Yang Zhang. Deep-learning contact-map guided protein structure prediction in casp13. *Proteins: Structure, Function, and Bioinformatics*, 87(12):1149–1164, 2019.

[11] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

[12] Cameron F Holder and Raymond E Schaak. Tutorial on powder x-ray diffraction for characterizing nanoscale materials, 2019.

[13] Yuta Suzuki, Hideitsu Hino, Takafumi Hawai, Kotaro Saito, Masato Kotsugi, and Kanta Ono. Symmetry prediction and knowledge discovery from x-ray diffraction patterns using an interpretable machine learning approach. *Scientific reports*, 10(1):1–11, 2020.

[14] Woon Bae Park, Jiyong Chung, Jaeyoung Jung, Keemin Sohn, Satendra Pal Singh, Myoungho Pyo, Namsoo Shin, and K-S Sohn. Classification of crystal structure using a convolutional neural network. *IUCrJ*, 4(4):486–494, 2017.

[15] Pascal Marc Vecsei, Kenny Choo, Johan Chang, and Titus Neupert. Neural network based classification of crystal symmetries from x-ray diffraction patterns. *Physical Review B*, 99(24):245120, 2019.

[16] Felipe Oviedo, Zekun Ren, Shijing Sun, Charles Settens, Zhe Liu, Noor Titan Putri Hartono, Savitha Ramasamy, Brian L DeCost, Siyu IP Tian, Giuseppe Romano, et al. Fast and interpretable classification of small x-ray diffraction datasets using data augmentation and deep neural networks. *npj Computational Materials*, 5(1):1–9, 2019.

[17] Abhik Chakraborty and Raksha Sharma. See deeper: Identifying crystal structure from x-ray diffraction patterns. In *2020 International Conference on Cyberworlds (CW)*, pages 49–54. IEEE, 2020.

[18] Alexander N Zaloga, Vladimir V Stanovov, Oksana E Bezrukova, Petr S Dubinin, and Igor S Yakimov. Crystal symmetry classification from powder x-ray diffraction patterns using a convolutional neural network. *Materials Today Communications*, 25:101662, 2020.

[19] Angelo Ziletti, Devinder Kumar, Matthias Scheffler, and Luca M Ghiringhelli. Insightful classification of crystal structures using deep learning. *Nature communications*, 9(1):1–10, 2018.

[20] Yanchao Wang, Jian Lv, Li Zhu, and Yanming Ma. Crystal structure prediction via particle-swarm optimization. *Physical Review B*, 82(9):094116, 2010.

[21] C-H Liu, Yunzhe Tao, Daniel Hsu, Qiang Du, and Simon JL Billinge. Using a machine learning approach to determine the space group of a structure from the atomic pair distribution function. *Acta Crystallographica Section A: Foundations and Advances*, 75(4):633–643, 2019.

[22] Kevin Kaufmann, Chaoyi Zhu, Alexander S Rosengarten, Daniel Maryanovsky, Tyler J Harrington, Eduardo Marin, and Kenneth S Vecchio. Crystal symmetry determination in electron diffraction using machine learning. *Science*, 367(6477):564–568, 2020.

[23] Artem R Oganov and Colin W Glass. Crystal structure prediction using ab initio evolutionary techniques: Principles and applications. *The Journal of chemical physics*, 124(24):244704, 2006.

[24] In-Ho Lee and KJ Chang. Crystal structure prediction in a continuous representative space. *Computational Materials Science*, 194:110436, 2021.

[25] Yong Zhao, Yuxin Cui, Zheng Xiong, Jing Jin, Zhonghao Liu, Rongzhi Dong, and Jianjun Hu. Machine learning-based prediction of crystal systems and space groups from inorganic materials compositions. *ACS omega*, 5(7):3596–3606, 2020.

[26] Haotong Liang, Valentin Stanev, A Gilad Kusne, and Ichiro Takeuchi. Cryspnet: Crystal structure predictions via neural networks. *Physical Review Materials*, 4(12):123802, 2020.

[27] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1):011002, 2013.

[28] Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2(1):1–7, 2016.

[29] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[30] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[31] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[32] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[33] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[34] Qiuling Tao, Pengcheng Xu, Minjie Li, and Wencong Lu. Machine learning for perovskite materials design and discovery. *npj Computational Materials*, 7(1):1–18, 2021.